Bioinformatics, 38(7), 2022, 1938–1946
https://doi.org/10.1093/bioinformatics/btac024
Advance Access Publication Date: 10 January 2022
Original Paper

OXFORD

Genetics and population analysis

A data-adaptive Bayesian regression approach for polygenic risk prediction

Shuang Song[1,2], Lin Hou [1,2,3,*] and Jun S. Liu[4,*]

[1]Center for Statistical Science, Tsinghua University, Beijing 100084, China, [2]School of Life Sciences, Department of Industrial Engineering, Tsinghua University, Beijing 100084, China, [3]MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China and [4]Department of Statistics, Harvard University, Cambridge, MA 02138, USA

*To whom correspondence should be addressed.
Associate Editor: Russell Schwartz

Received on July 13, 2021; revised on December 21, 2021; editorial decision on January 5, 2022; accepted on January 9, 2022

Abstract

Motivation: Polygenic risk score (PRS) has been widely exploited for genetic risk prediction due to its accuracy and conceptual simplicity. We introduce a unified Bayesian regression framework, NeuPred, for PRS construction, which accommodates varying genetic architectures and improves overall prediction accuracy for complex diseases by allowing for a wide class of prior choices. To take full advantage of the framework, we propose a summary-statistics-based cross-validation strategy to automatically select suitable chromosome-level priors, which demonstrates a striking variability of the prior preference of each chromosome, for the same complex disease, and further significantly improves the prediction accuracy.

Results: Simulation studies and real data applications with seven disease datasets from the Wellcome Trust Case Control Consortium cohort and eight groups of large-scale genome-wide association studies demonstrate that NeuPred achieves substantial and consistent improvements in terms of predictive r^2 over existing methods. In addition, NeuPred has similar or advantageous computational efficiency compared with the state-of-the-art Bayesian methods.

Availability and implementation: The R package implementing NeuPred is available at https://github.com/shuangsong0110/NeuPred.

Contact: houl@tsinghua.edu.cn or jliu@stat.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) of human complex diseases have identified tens of thousands of associated genetic variants (Jostins and Barrett, 2011), providing novel insights about disease mechanisms and revealing extensive polygenic genetic architectures. In clinical translation of GWAS discoveries, polygenic risk score (PRS), which quantifies genetic risks via aggregation of risk alleles, has emerged as a promising tool to stratify patients for precision prevention, screening and diagnosis and treatments (Allen *et al.*, 2010; Consortium *et al.*, 2009; Ripke *et al.*, 2011). PRS calculates a weighted sum of the number of risk alleles carried in a personal genome, and finding a good weighting strategy is key to the success of a PRS tool.

PRS methods differ by their selection of risk loci and estimation of effect sizes. An early PRS approach, Pruning and Thresholding (P + T), first selects a subset of significant and approximately independent single nucleotide polymorphisms (SNPs) via linkage disequilibrium (LD) clumping and *P*-value thresholding, and then calculates PRS based on the selected SNPs. Instead of using individual-level genotype data, P + T only requires GWAS summary statistics to construct PRS, which is attractive because of the data sharing concerns and privacy policies. However, this simple construction discards potentially useful information due to the *ad hoc* nature of their aggregation of marginal effects of the selected SNPs, which hurts its prediction accuracy. A main challenge in constructing a good PRS lies in the high dimensionality of genetic variants and the complex LD structure between them, which complicates risk variant selection and effect size estimation. Advanced statistical techniques in high-dimensional data analysis are particularly helpful in this respect.

A recent trend in PRS research is to leverage high-dimensional techniques in variable selection and shrinkage estimation. Some methods leverage the marginal estimator of variant effect sizes and infer the posterior distribution of true effect sizes through Bayesian [LDpred (Vilhjálmsson *et al.*, 2015) and the updated version LDpred2 (Privé *et al.*, 2021)] or empirical Bayes methods [EB-PRS (Song *et al.*, 2020)]. Both approaches enforce sparsity and shrinkage

in effect size estimation by utilizing spike-and-slab priors. Other methods employ high-dimensional regression analysis to jointly estimate the effect sizes of risk variants, and incorporate various penalty terms to shrink the linear coefficients. For example, PANPRS (Chen *et al.*, 2020) use L_1 penalty, TlpSum takes a truncated LASSO penalty (Pattee and Pan, 2020), and LassoSum (Mak *et al.*, 2017) and ElastSum (Pattee and Pan, 2020) use a combination of L_1 and L_2 penalties. In comparison to penalized regression approaches, Bayesian high-dimensional regression methods bring additional flexibility by allowing for a wide range of priors to model the polygenic structure of complex diseases. Specifically, RSS (Zhu and Stephens, 2017) and SBayesR (Lloyd-Jones *et al.*, 2019) employ finite normal mixture distributions as the prior, while PRS-CS (Ge *et al.*, 2019) uses the Strawderman-Berger prior (Berger, 1980; Strawderman, 1971), to characterize the distribution of genetic effects. DBSLMM assumes that all SNPs have non-zero effects on the phenotype, but that some SNPs have larger effect sizes than the others (Yang and Zhou, 2020).

Existing methods have unambiguously demonstrated benefits of high-dimensional statistical methods in PRS construction. However, there is a lack of guidance on how to determine the optimal penalties or the class of prior distributions for a trait of interest. Intuitively, the relative performance of different PRS methods depends on how well their internal model assumptions match the underlying genetic architectures. The true effect size distributions of human diseases are diverse and complex (Park *et al.*, 2010; Zhang *et al.*, 2018), and most importantly, unknown. The genetic architecture of a certain disease may also vary from chromosome to chromosome (Moser *et al.*, 2015). Moreover, subtle tweaks in penalty terms and prior distributions could raise completely new computational challenges in the corresponding optimization algorithms and Markov chain Monte Carlo (MCMC) sampling strategies.

In light of the aforementioned limitations, we propose NeuPred, a Bayesian PRS framework that selects prior classes and hyperparameters in a data adaptive and computationally effective fashion. Our main contributions are two: (i) a general Bayesian framework built upon the recently introduced 'neuronized prior' for Bayesian regression (Shin and Liu, 2021); (ii) a flexible summary-statistics-based cross-validation (CV) strategy to select suitable priors. With a unified formulation and efficient MCMC computations, neuronized priors cover diverse types of sparse and shrinkage priors commonly used in Bayesian linear regressions, such as continuous and discrete spike-and-slab priors, Laplace priors, Cauchy priors, horseshoe priors, etc. NeuPred searches in a wide class of tunable priors, ranging from conjugate to non-conjugate, from discrete mixture to continuous hierarchical, and from heavy-tailed to light-tailed, and uses the proposed CV strategy to select a suitable one. Note that it is straightforward to conduct CV when individual data are available, but is not obvious how to do CV when only summary statistics for associations are available.

Simulations and real data applications on seven Wellcome Trust Case Control Consortium (WTCCC) traits and another eight groups of large-scale GWAS datasets demonstrate that NeuPred achieves substantial and consistent improvements in prediction accuracy compared to existing approaches due to its adaptability to varying genetic architectures. Furthermore, NeuPred is robust when the LD matrices are externally estimated based on genotype data of a reference panel with relevant ancestry, and is also computationally more efficient than many existing Bayesian PRS algorithms.

2 Materials and methods

2.1 Method overview

We provide two algorithms: NeuPred, which targets cases when only GWAS summary statistics are provided, and NeuPred-I, which works with individual-level genotype data. Due to potential privacy and data sharing concerns, NeuPred can be more extensively applied than NeuPred-I. As the key innovations are similar in the two methods, our discussions focus on NeuPred in the main text. NeuPred is based on the Bayesian linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y}$ denotes the vector of phenotypes of the n individuals and $\mathbf{X}$, an $n \times p$ matrix, denotes genotypes of the n individuals at p SNPs. The regression coefficient vector $\boldsymbol{\beta}$ is of p-dimensional, and the error term $\boldsymbol{\varepsilon} \sim N(0, \sigma_\epsilon^2 \mathbf{I}_n)$ accounts for environmental effects. We assume that both $\mathbf{y}$ and $\mathbf{X}$ are standardized. When only GWAS summary statistics are provided, we only have access to $\hat{\boldsymbol{\beta}}_{\text{marg}} = \mathbf{X}^\top \mathbf{y}/n$. Let the in-sample linkage-disequilibrium (LD) matrix be $\mathbf{R} = \mathbf{X}^T\mathbf{X}/n$, which may be available in some cases or estimated from other sources. Multiplying $\mathbf{X}^T/n$ to both sides of Equation (1) we obtain

$$\hat{\boldsymbol{\beta}}_{\text{marg}} = \mathbf{R}\boldsymbol{\beta} + n^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}. \tag{2}$$

Consider the eigen-decomposition $\mathbf{R} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, where $\mathbf{U}$ and $\mathbf{D}$ are orthogonal and diagonal matrices, respectively. Multiplying both sides of (2) by $\sqrt{n}\mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T$, we obtain a new regression equation

$$\mathbf{y}' = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}', \tag{3}$$

where the error term satisfies the i.i.d. Gaussian condition, i.e. $\text{Var}(\boldsymbol{\varepsilon}') = \sigma_\epsilon^2 \mathbf{I}_p$. If $\mathbf{R}$ is known exactly, the OLS estimate based on (3) is exactly the same as that based on (1). But (3) enables conduction of regularized regression when only $\mathbf{R}$ is available and the feature matrix $\mathbf{X}$ is not.

To overcome the winner's curse in high-dimensional analysis, Bayesian approaches resort to a specific class of shrinkage priors, such as point-normal or normal mixtures (LDpred, RSS, SBayesR and EB-PRS), and Strawderman-Berger prior (PRS-CS) to induce shrinkage estimation in linear coefficients. However, choosing a prior that can lead to optimal prediction accuracy is a non-trivial task. NeuPred is based the *neuronized prior* introduced by Shin and Liu (2021), which postulates that each regression coefficient can be represented *a priori* as

$$\beta_j := T(\alpha_j - \alpha_0)\omega_j, \tag{4}$$

where T is a non-decreasing activation function, $\alpha_j \sim N(0,1)$, and $\omega_j \sim N(0, \tau_\omega^2)$, for $j = 1, \ldots, p$. The hyperparameter α_0 can be either specified to a fixed value or updated with Gibbs sampling (Liu and Sabatti, 2000; Shin and Liu, 2021). This formulation enables a unified implementation of various classes of shrinkage priors by simply changing the activation function. This is desirable when coping with genetic data of a specific disease with unknown genetic architectures, as it can be implemented efficiently via one common MCMC algorithm. Details about the MCMC algorithm and the selection of hyperparameters are provided in Supplementary Note S1.1.

To take full advantage of the neuronized priors, we design a summary-statistics-based CV strategy to automatically select a suitable prior for each chromosome (see Section 2.2). PRS is estimated subsequently as the posterior mean effect size under the selected prior. Three built-in neuronized priors are considered in our R package NeuPred, including the spike-and-slab Laplace (Neu-SpSL-L), spike-and-slab Cauchy (Neu-SpSL-C) and horseshoe (Neu-HS) priors, which appear to be sufficient for most genetic data applications we have tested.

For all reported simulation experiments and real data applications, we trained summary-statistics-based methods with only GWAS summary statistics, and evaluated the performances of PRS in an independent validation dataset. The prediction accuracy is evaluated with respect to the predictive r^2 and the area under the receiver operating characteristic (ROC) curve (AUC) in the validation dataset.

2.2 A cross-validation strategy for prior selection

To adapt to varying genetic architectures, we wish to implement CV to select an appropriate neuronized prior for each chromosome. If individual-level genotype data are available, we can directly apply

fivefold CV to the training data and choose the best performing neuronized prior according to the predictive r^2. When only GWAS summary statistics are provided, we design the following two-step CV procedure based on the post-transformation data, $\mathcal{D}' = (\mathbf{y}', \mathbf{X}')$ as in (3) of Section 2.1.

Step 1: For each prior choice (i.e. Neu-SpSL-L, Neu-SpSL-C and Neu-HS) and each chromosome, we calculate the Pearson correlation coefficient (PCC) between the observed $\mathbf{y}'$ and its prediction $\hat{\mathbf{y}}'$ derived from fivefold CV based on data $\mathcal{D}'$; we test whether the prediction accuracy of a prior is significantly higher than the other two (one-tailed test, P-value < 0.05) after the Fisher transformation of the PCC values and choose the one if it is significantly better than the other two. If no prior significantly outperforms the other two in terms of PCC, we conduct Step 2 focusing on the robustness and similarity between $\mathbf{y}'$ and $\hat{\mathbf{y}}'$.

Step 2: We compute the Kolmogorov–Smirnov (KS) test statistic between empirical distributions of the observed $\mathbf{y}'$ and the predicted $\hat{\mathbf{y}}'$, and choose the prior with the minimum KS statistic if we have not made a decision in Step 1.

We stress that the proposed CV procedure only requires GWAS summary statistics and a LD reference panel, and does not need any external information from a validation dataset.

2.3 Simulation settings

2.3.1 Neuronized priors for varying genetic architectures

We first conducted simulations based on generated genotypes to explore the performances of NeuPred under three neuronized priors for varying genetic architectures. The minor allele frequencies (MAF) were sampled from Unif[0.1, 0.5]. The numbers of SNPs p and individuals n were fixed at 1000 and 2000, respectively. The genotypes with correlated LD structures were generated from binomial distributions with AR(1) correlation matrix, under which the correlation decreases with increasing spatial distances, using the R package CorBin (Jiang *et al.*, 2021), and the correlation coefficient was fixed at 0.1. The proportion of causal SNPs κ took values in $\{0.01, 0.05, 0.1, 0.3, 1\}$, the true effect sizes for causal SNPs were sampled from $N(0, \frac{h^2}{\kappa p})$. The quantitative phenotypes were generated by Equation (1), and the error term $\boldsymbol{\varepsilon} \sim N\left(0, (1-h^2)\mathbf{I}_n\right)$.

2.3.2 Effectiveness of the cross-validation strategy for prior selection

We fixed the number of SNPs to be 1000 and the heritability to be 0.5, and let the proportion of causal SNPs take values from $\{0.01, 0.05, 0.1, 0.3, 1\}$. We considered three scenarios: n is smaller than p ($n = 500$); n is equal to p ($n = p = 1000$); and n is larger than p ($n = 2000$). The block size was set to be 50. In each block, we randomly sampled MAF from Unif[0.1, 0.5], and generated genotypes under an AR(1) correlation structure. The correlation coefficient in each block was randomly sampled from Unif[0.1, $\rho_{\max}$], where $\rho_{\max}$ is the Prentice constraint imposed on marginal expectations and correlation coefficients (Prentice, 1988). We used the two CV procedures (fivefold) to tune parameter λ in LASSO regression to minimize the MSEs, and compared their estimates including and the sets of variables selected, and the MSEs and predictive r^2 in an independent test dataset of sample size 1000. The Jaccard index was employed for comparing two sets A and B, which is defined as $J(A, B) = |A \cap B|/|A \cup B|$.

2.3.3 Comparison between NeuPred and other PRS methods

We conducted a simulation study based on the real genotype data of chromosome 22 (4097 SNPs) from the WTCCC rheumatoid arthritis (RA) study on 4685 individuals. Markers with MAF smaller than 0.005 or genotyping failure rate larger than 0.05 or significant Hardy–Weinberg equilibrium (HWE) with $P < 10^{-5}$ in PLINK 1.9 (Chang *et al.*, 2015) were removed. Samples with more than 10% missing were also removed. A method's performance is evaluated by the fivefold CV: four-fifths of the samples were used to calculate the GWAS summary statistics to train PRS models and the remaining one-fifth were reserved as test data.

We also performed simulations with WTCCC genotypes at a larger scale, corresponding to 15 860 individuals and 260 243 SNPs after overlapping with HapMap 3 SNPs and quality control. The effect sizes were simulated from a point-normal distribution $(1-\kappa)\delta_0 + \kappa N(0, \frac{h^2}{\kappa p})$, with $\kappa = 0.1\%, 1\%, 10\%$ and 100%. Four scenarios representing different effective sample sizes were considered: (i) all SNPs (chromosomes 1–22, 260 243 SNPs), (ii) chromosomes 1–4 (78 067 SNPs), (iii) chromosomes 1 and 2 (42 984 SNPs) and (iv) chromosome 1 (21 007 SNPs). The effective sample sizes are defined as the sample size that maintains the same n/p ratio if all SNPs are used, i.e. $n_{\text{eff}} = (n/p_{\text{sim}})p_{\text{all}}$. Here p_{sim} is the actual number of SNPs used in each simulation, and p_{all} is the number of all autosomal SNPs (Vilhjálmsson *et al.*, 2015).

2.4 Compared methods

We compared NeuPred with 12 state-of-the-art summary-statistics-based PRS methods, including unadjusted PRS (unadj PRS), P + T, LDpred-inf, LDpred (Vilhjálmsson *et al.*, 2015), SBayesR (Lloyd-Jones *et al.*, 2019), SBayesC (Habier *et al.*, 2011), RSS (Zhu and Stephens, 2017), PRS-CS-auto, PRS-CS (Ge *et al.*, 2019), LDpred2-inf, LDpred2-auto and LDpred2-grid (Privé *et al.*, 2021) (detailed in Supplementary Note S1.2), and an individual-level-data-based method, BayesR (Moser *et al.*, 2015) was used as benchmarks. For each compared method, we used the default setting in its provided software. For RSS, we set the number of MCMC iterations to be 10^6 with 2×10^5 burn-in iterations. There were cases that long MCMC chains of RSS encountered computational instabilities; and we halved the number of iterations in such cases.

Four methods, P + T, LDpred, PRS-CS and LDpred2-grid, need further parameter calibrations when applied to a new dataset. However, reporting the highest accuracy with *post hoc* parameter tuning makes comparisons with other methods unfair. To alleviate the concern, in simulations and WTCCC studies, we carved out one-fifth of the test data to tune parameters and evaluated prediction accuracy on the remaining four-fifth (this is still a bit unfair to other methods). For real data applications with independent test datasets, we tuned the parameters with an external validation dataset (see Supplementary Table S1). We also provide the results with the best *post hoc* tuning parameters for the four methods in Supplementary Materials.

2.5 Reference LD matrix construction

The in-sample LD matrix was estimated from GWAS samples. An external LD matrix was estimated via a non-linear shrinkage method (Ledoit and Wolf, 2015, 2017) based on the reference panel of the 1000 Genomes Project (1000G, henceforth), which contains 489 Europeans with 9 997 231 SNPs after quality control. We partitioned the genome into 1703 independent blocks using LDetect (Berisa and Pickrell, 2016), based on the 1000G reference panel with European ancestry (https://bitbucket.org/nygcresearch/ldetect-data/src/master/), and performed Bayesian regression in each LD block. For SBayesR and RSS, we used the gctb software (https://cnsgenomics.com/software/gctb/) to shrink the off-diagonal entries of the sample LD matrix toward zero. The gctb software is implemented by SBayesR (Lloyd-Jones *et al.*, 2019), which is also a C++ port from that provided with the RSS software (Zhu and Stephens, 2017). We also used an LD matrix estimated from the UK Biobank (UKBB) samples of European ancestry, which is available at https://pan.ukbb.broadinstitute.org with Pan-UKB Team (2020). Note that, several methods have built-in options to use UKBB LD matrices, namely, SBayesR, SBayesC, PRS-CS and LDpred2. For these methods, we directly downloaded and used their internal LD estimation.

2.6 Computation time

We compared the CPU time of the Bayesian PRS methods including NeuPred, LDpred, SBayesR, SBayesC, RSS, PRS-CS and LDpred2.

Numbers of MCMC iterations for these methods were chosen according to their respective default settings, i.e. 10^4, 60, 10^4, 10^4, 10^6, 10^3 and 500, with the corresponding burn-in iterations 2×10^3, 5, 2×10^3, 2×10^3, 2×10^5, 500 and 10^3, respectively. LDpred, PRS-CS and LDpred2 all require grid search for finding the best tuning parameters. The time we reported is only for one specific parameter setting, and the computation time for estimating LD matrices is negligible. The computation time for each of the seven methods is for simulations based on the WTCCC RA dataset, with an Intel Xeon processor with 2.50 GHz and 48 cores. Among the methods, NeuPred, RSS, PRS-CS and LDpred2 were run with all 48 cores, while SBayesR, SBayesC and LDpred did not have parallel computing capacity and used 1 CPU core only.

2.7 Genetic datasets analyzed

The WTCCC datasets on seven complex diseases (Wellcome Trust Case Control Consortium, 2007) were used in both simulations and real data analysis. As for the large-scale GWAS studies, we trained models with 4533 individuals with celiac disease (CEL) and 10 750 controls from Dubois' study (Dubois *et al.*, 2010). The test dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) CEL study (1716 cases and 530 controls, dbGaP: phs000274) (Garner *et al.*, 2014). For Crohn's disease (CD), we trained models using summary statistics from the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC; 15 056 cases and 6333 controls) (Franke *et al.*, 2010). Individuals from the WTCCC were removed from the meta-analysis and used as the test dataset (2891 cases and 1689 controls). For RA, we used summary statistics from the Stahl *et al.* (2010) (5539 cases and 20 169 controls), removing WTCCC samples for training, and used the WTCCC data as the test data. For type 2 diabetes (T2D), we trained models using summary statistics from the Diabetes Genetics Replication and Meta-analysis consortium (DIAGRAM, 56 862 cases and 12 171 controls) (Morris *et al.*, 2012), and tested the model on samples from the Northwestern NUgene Project (517 cases and 662 controls, dbGaP: phs000237).

We also trained PRS with the UKBB GWAS summary statistics, downloaded from Neale lab GWAS round 2 (http://www.nealelab.is/uk-biobank), and tested with the Genetic Epidemiology Research on Aging (GERA) summary statistics for asthma (ATH) (dbGaP: phs000674.v3.p3, http://cg.bsc.es/gera_summary_stats/), and with the WTCCC data for hypertension (HT). We tested with the GWAS summary statistics on 449 899 Europeans in Turcot *et al.* (2018) for body mass index (BMI) and 253 288 individuals of European ancestry in Wood *et al.* (2014) for height (HGT). As for the *post hoc* parameter tuning for P + T, LDpred, PRS-CS and LDpred2-grid, we used UKBB summary statistics for CD, CEL, RA, T2D, BMI and HGT; GWAS data from GABRIEL consortium for ATH (Moffatt *et al.*, 2010); and the FinnGen GWAS data for HT. More details about the datasets are provided in Supplementary Table S1.

3 Results

3.1 Simulation experiments

3.1.1 Neuronized priors for varying genetic architectures

We simulated datasets to investigate performances of NeuPred under different types of priors for various genetic architectures. The simulated genetic architectures vary in three aspects: LD structure, heritability and the sparsity of causal SNPs. Scenarios with both independent SNPs and arbitrary LD structures were tested (Supplementary Figs S1 and S2), with heritability level h^2 at 20%, 50% and 80%.

NeuPred with either Neu-SpSL-L or Neu-SpSL-C obtained similar predictive r^2, but performed worse under the Neu-HS prior when causal signals are sparse, regardless of the LD structures and heritability settings. On the other hand, the Neu-HS prior worked the best under more polygenic architectures. The results highlight that different underlying genetic architectures prefer different priors, often strongly, and thus the prediction accuracy can be potentially much improved with a proper selection of prior distributions.

3.1.2 Effectiveness of the cross-validation strategy for prior selection

When individual-level genotype data are available, it is straightforward to use CV for prior and other tuning parameter selections. When only GWAS summary statistics are provided, however, the standard individual-data-based CV is no longer applicable. In Section 2.2, we detail a summary-statistics-based CV approach useful for selecting prior for each chromosome. To demonstrate its effectiveness, we used LASSO as the regression tool and simulations to compare the new CV procedure with the standard one. Similar results are expected to hold true for more time-consuming Bayesian procedures.

We observed that MSEs and predictive r^2 reported by the two CV procedures were similar (Supplementary Fig. S3), especially when the prediction accuracy was high (under sparser settings). We further evaluate the similarity between the sets of variables selected by the two CV procedures by the Jaccard index (JI). We observed that JI between the two selected sets ranged from 0.725 to 0.885, under sparse settings (i.e. $\kappa = 0.01, 0.05, 0.1$), indicating a strong consistency between the two CV procedures. In polygenic cases ($\kappa = 0.3$), JI was reduced to 0.37 when $n = 500$, reflecting a high uncertainty in variable selection (Supplementary Table S2).

3.1.3 Prediction accuracy

We first applied all the considered PRS methods to a small-scale simulation based on the observed genotypes chromosome 22 on the WTCCC data (Section 2). Quantitative phenotypes were generated with three genetic models: (i) a point-normal distribution with 1% causal SNPs; (ii) a point-normal distribution with 10% causal SNPs; (iii) equal effect sizes for 1% causal SNPs. The heritability was fixed at 0.5. Two full individual-level-data-based methods, NeuPred-I and BayesR were also applied and their performances are listed as benchmarks. For summary-statistics-based methods, we evaluate their performances under both in-sample and external LD matrices. Since PRS-CS-auto and PRS-CS have built-in LD matrices estimated from the 1000G reference panel, their corresponding results with in-sample LD matrices are not available. For P + T, in-sample LD matrices were used to clump SNPs.

For setting (i) where the simulated genetic signals are sparse, NeuPred achieved a high prediction accuracy compared with other summary-statistics-based methods (Supplementary Fig. S4). In particular, NeuPred had performed comparably to NeuPred-I and BayesR, which used the full individual-level data. LDpred2-grid had the best performance among the summary-statistics-based methods in simulation setting (ii) where the genetic architecture were relatively polygenic ($\kappa = 0.1$). We note that the generating models of settings (i) and (ii) are consistent with the underlying models of LDpred2-auto, LDpred2-grid and SBayesC, and LDpred2-grid further tunes parameters among a grid size of 126. The setting (iii) assumes equal contribution from each causal SNP, which violates the normal assumption and results in less optimal performances for methods assuming point-normal or normal mixture priors. In contrast, NeuPred remained robustness and showed a better performance in this case.

We then applied the methods to whole-genome-scale simulations with a larger sample size, with varying proportion of causal SNPs (Section 2). We omitted the comparison with unadjusted PRS, LDpred and RSS since their performances were demonstrated to be inferior to the others in the literature (Privé *et al.*, 2021) and the small-scale simulation studies. Note that, the generating models in the simulation settings are the same as the underlying models (including priors) of LDpred2-auto, LDpred2-grid and SBayesC, and are similar to that of SBayesR. These four methods performed well in these simulations, especially LDpred2-auto and LDpred2-grid (Supplementary Fig. S5). LDpred2-inf had good performances under polygenic scenarios ($\kappa = 1$), but became less competitive under sparse settings. Even employed priors are different from that of the generating models, NeuPred remained to perform robustly under all sparse and polygenic simulation settings, especially when the effective sample sizes were large. This high robustness is consistent with our findings in the small-scale simulations and is likely due to its

ability in choosing one from three distinctive classes of priors adaptively via our new CV strategy.

In addition, we assessed the calibration of polygenic prediction methods by regressing the true phenotype onto the PRS predictor and inspecting the regression slope. A slope close to one often indicates the predictor is correctly calibrated (Vilhjálmsson *et al.*, 2015). We note that the usage of the calibration slope is being debated (Stevens and Poppe, 2020; Vach, 2013; Wang, 2020) and we suggest to interpret this statistic with caution. In general, the Bayesian approaches had calibration slopes closer to one than P + T, especially when the predictive accuracy was high (Supplementary Table S3). NeuPred was well calibrated under sparse settings with large effective sample sizes.

3.1.4 Robustness to external LD information

We conducted simulations based on the RA data of WTCCC (Section 2), to evaluate robustness of NeuPred with respect to the LD information input under three prior choices. For each chromosome, we simulated the SNP effect sizes from a point-normal distribution, fixed the heritability at 0.5 and varied the proportion of causal SNPs. The LD matrices were either estimated from the simulated samples (in-sample LD) or from the 1000G reference panel with European ancestry (external LD). We can see that the influence of the LD input varies among chromosomes and prior choices, whereas the Neu-SpSL-L prior enabled the most robust performance under varying genetic architectures (Supplementary Fig. S6 and Supplementary Table S4). Therefore, we recommend to always use Neu-SpSL-L as the default setting when the LD information is from an external source, and to employ a shrinkage method to estimate the LD matrices (Ledoit and Wolf, 2015, 2017). The robustness of NeuPred is also evident in Supplementary Figure S4. Although performances of all methods modeling LD information deteriorated when external LD estimates were used in the place of the in-sample LD matrices, NeuPred appeared to be least affected.

3.2 Real data applications

3.2.1 Adaptability to varying genetic architectures

We tested NeuPred with the WTCCC datasets of seven complex diseases, including bipolar disorder (BD), coronary artery disease (CAD), T2D, HT, CD, RA and type 1 diabetes (T1D). To assess the added value of prior selection, NeuPred was implemented in two ways: (i) using one of the three default neuronized priors, namely, Neu-SpSL-L, Neu-SpSL-C and Neu-HS, versus (ii) using a data-adaptive prior selected (among the three) by the summary-statistics-based CV procedure. We used the in-sample LD information and conducted fivefold CV to evaluate predictive r^2 and AUC for each method and disease. In terms of predictive r^2, we observe that for the three immune-related diseases, CD, RA and T1D, using either Neu-SpSL-C or Neu-SpSL-L gave similar results and was superior to using Neu-HS (Fig. 1). In contrast, the Neu-HS prior was strongly favored by BD. The three priors worked comparably well for CAD, T2D and HT. Similar patterns were observed with respect to AUC (Supplementary Table S5).

Because the major histocompatibility complex region explains a large amount of the overall variance of autoimmune-related traits (Moser *et al.*, 2015; Vilhjálmsson *et al.*, 2015; Zhang *et al.*, 2018), the favorable performance of Neu-SpSL-C and Neu-SpSL-L in autoimmune diseases echos the simulation results that these priors are advantageous for less polygenic traits. Regardless of the underlying genetic architecture, the approach with adaptive prior selection consistently outperforms any single-prior approach. The selected prior for each chromosome is marked for CAD and HT (Fig. 1b) and others (Supplementary Fig. S7). For most cases, our CV strategy selected the best performing prior for each chromosome, which explains the substantial improvement in the overall prediction accuracy. For scenarios where Neu-SpSL-L and Neu-SpSL-C had better performance compared with Neu-HS, our CV strategy showed a slight preference to Neu-SpSL-L, which shrinks more strongly and is more conservative than Neu-SpSL-C. Although in general Neu-SpSL-L and Neu-SpSL-C are preferred by autoimmune diseases and Neu-HS is favored by polygenic genetic architectures, Figure 1b shows a significant variability among the chromosome-level prior preferences, which further highlights the importance of using data-adaptive priors. We also provide a comparison between our summary-statistics-based CV strategy and the pseudo-validation strategy proposed in the lassosum method (Mak *et al.*, 2017). Even if we gave an advantage to the pseudo-validation method by allowing it to use the test genotype data, our summary-statistics-based CV strategy showed more improvement in six of the seven traits (Supplementary Note S1.3 and Supplementary Table S6).

3.2.2 Prediction accuracy

To assess the performance of NeuPred and other summary-statistics-based PRS methodologies, we analysed seven WTCCC complex diseases and eight large-scale GWAS studies with independent test datasets (Section 2). For calibration, we provide the results of NeuPred-I and BayesR as benchmarks when possible as they leverage full individual-level genotype information. We also estimated the SNP-based heritability (on the observed scale) of each of the seven WTCCC traits using LDSC (Bulik-Sullivan *et al.*, 2015), which provides a theoretical upper bound for the predictive r^2 (Supplementary Table S7).

For the seven WTCCC traits based on CV, we observe that NeuPred consistently outperformed all other summary-statistics-based methods in terms of both predictive r^2 and AUC criteria (see details in Supplementary Note S1.4, Supplementary Fig. S8 and Supplementary Tables S8–S11). We note, however, that the extent of improvement by CV may not be extensible to other studies, due to the limited sample sizes of the WTCCC datasets and potential over-fitting problems. We further validated the predictors trained with the WTCCC data on the independent samples from the Northwestern NUgene Project of T2D. Although the prediction accuracy diminished for all tested methods, NeuPred still achieved the best performance, showcasing that the model fitted by NeuPred generalizes well in independent datasets (Supplementary Table S12).

We applied all the methods to four large-scale GWAS studies on CD, CEL, RA and T2D, and evaluated their performance with four independent test datasets (Fig. 2a, Table 1 and Supplementary Table S13). The hyperparameters for P + T, LDpred, PRS-SC and LDpred2-grid were tuned with UKBB GWAS data, which is independent of our training and test datasets. We also provide the results for the four methods with the best *post hoc* tuned parameters in test data in Supplementary Table S14. LDpred was relatively robust with parameters either optimized *post hoc* or tuned with external validation datasets, while the other three all performed much worse on two or more of the four traits without *post hoc* optimizing. We also notice that SBayesR and SBayesC had convergence issues and suboptimal performances on traits such as CEL and RA, which were also reported in other studies (Zhou and Zhao, 2021). Therefore, we used a shorter chain length (5000) if the two methods encountered a convergence problem. If the convergence issue remained, we substituted LD matrices with that estimated from the UKBB LD reference panel. We can see that NeuPred still consistently outperformed the others in terms of both predictive r^2 and AUC on all compared traits. We also provide the calibration slope results in Supplementary Table S15, showing that it was challenging for all methods to derive PRS with slope close to one with an independent test cohort.

We further evaluated prediction accuracy of NeuPred, P + T, SBayesR, SBayesC, PRS-CS and LDpred2 with UKBB GWAS datasets on two binary traits, ATH and HT, and two quantitative traits, BMI and HGT. The LD matrices were estimated with UKBB individuals (Section 2). The evaluation is based on the square of the (quasi-) correlation in an independent test dataset (see Supplementary Note S1.5). We tuned parameters for P + T, PRS-CS and LDpred2-grid with an external validation dataset for ATH and HT, but used the optimal parameters in the training data for BMI and HGT (as no independent validation cohort was available). We also provide the results under the optimal *post hoc* tuned parameters (i.e. the parameter set that leads to the best testing result) for the four methods in Supplementary Table S16. Overall, NeuPred and LDpred2-auto performed

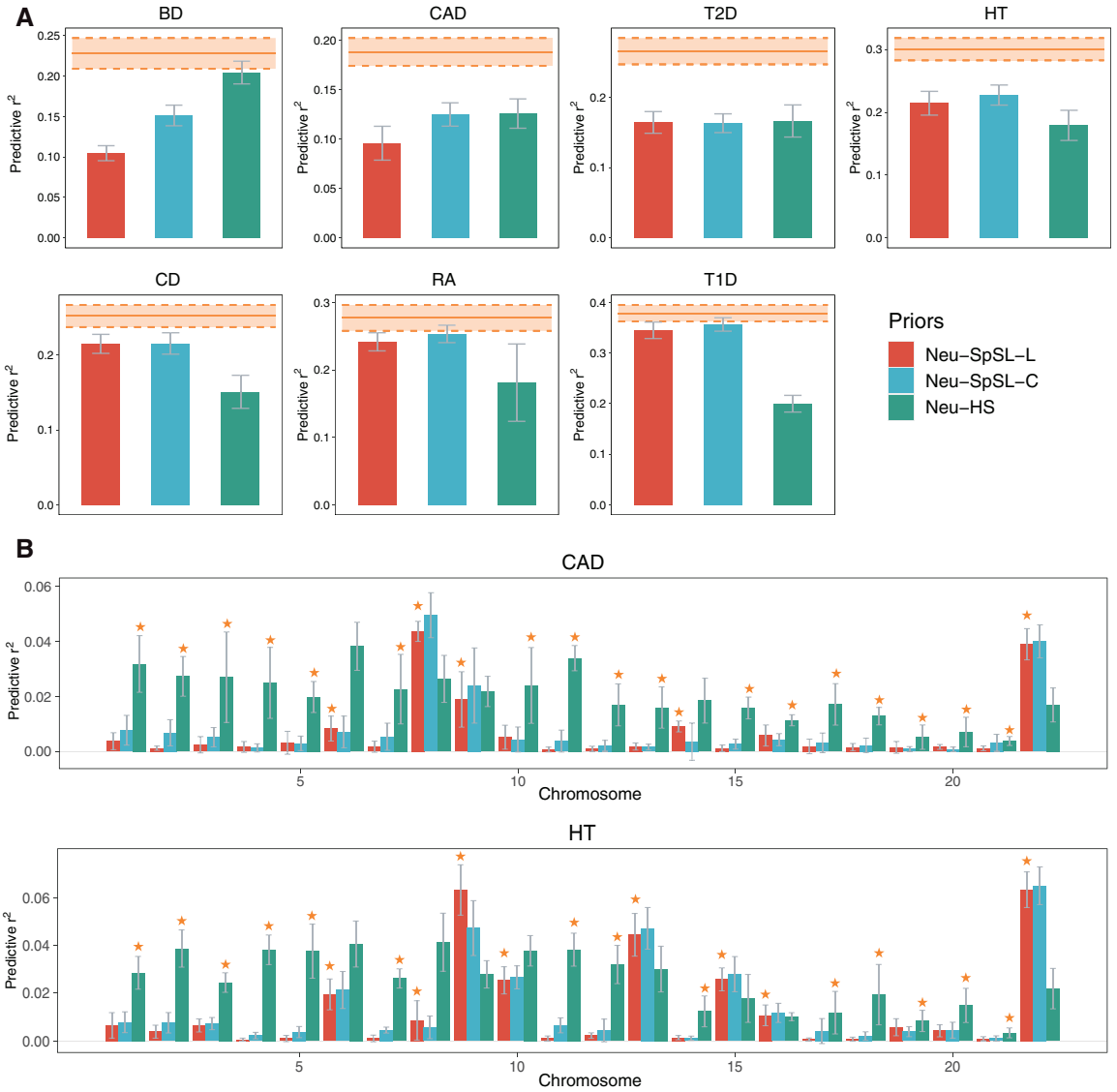

Fig. 1. Comparison of three neuronized priors applied to the seven WTCCC complex diseases. (**a**) The predictive r^2 is estimated from fivefold CV. The orange solid and dashed lines indicate the mean predictive r^2 and standard deviations, respectively, estimated by NeuPred with priors automatically selected. (**b**) The predictive r^2 of NeuPred under three neuronized priors for each chromosome for CAD and HT. In each CV, NeuPred automatically selected a prior, and the prior that is most frequently selected is marked with a star, which varies greatly among diseases and chromosomes

comparably and better than other methods (Fig. 2b), both requiring no *post hoc* tuning. LDpred2-grid performed competitively.

3.2.3 Computational efficiency

Supplementary Figure S9 and Supplementary Table S17 show the computation time of NeuPred (under three prior choices), LDpred, SBayesR, SBayesC, RSS, PRS-CS and LDpred2 for analyzing the simulated datasets based on chromosome 1 (26 799 SNPs) and chromosome 22 (4097 SNPs) of the WTCCC dataset. Compared with the other six Bayesian methodologies, NeuPred is very competitive in terms of time efficiency. NeuPred under the Neu-SpSL-L prior has the shortest mean running time with about 30-folds improvement over PRS-CS at the default setting.

4 Discussion

As GWAS sample sizes continue to grow, PRS methods are becoming more and more powerful in predicting complex diseases, and offer potentials for applications in clinical care and precision medicine (Chatterjee *et al.*, 2013; Dudbridge, 2013). Several Bayesian PRS methods have been developed to improve the prediction accuracy, each focusing on a specific class of priors, such as point-normal, normal mixtures and the Strawderman–Berger distributions. However, a unified framework is still lacking for accommodating diverse and unknown genetic architectures of complex traits. NeuPred described in this article fills in this gap by providing a unified treatment of priors, which allows one to test out, tune and combine multiple types of prior classes to accommodate different genetic architectures. NeuPred takes GWAS summary statistics and a LD matrix as input, and employs an efficient computational strategy to entertain a wide range of prior choices such as point-normal, horseshoe, Cauchy and point-Laplace. As a consequence, NeuPred substantially improves the prediction accuracy and computational stability compared with existing Bayesian PRS methods. In particular, NeuPred allows one to test out, tune and combine diverse classes of priors to accommodate different genetic architectures at chromosome level with no need of information from an external validation dataset, which is useful and flexible as a validation dataset is not always available in clinical applications.

We provide three candidate neuronized priors for consideration by default: Neu-SpSL-L, Neu-SpSL-C and Neu-HS, which should

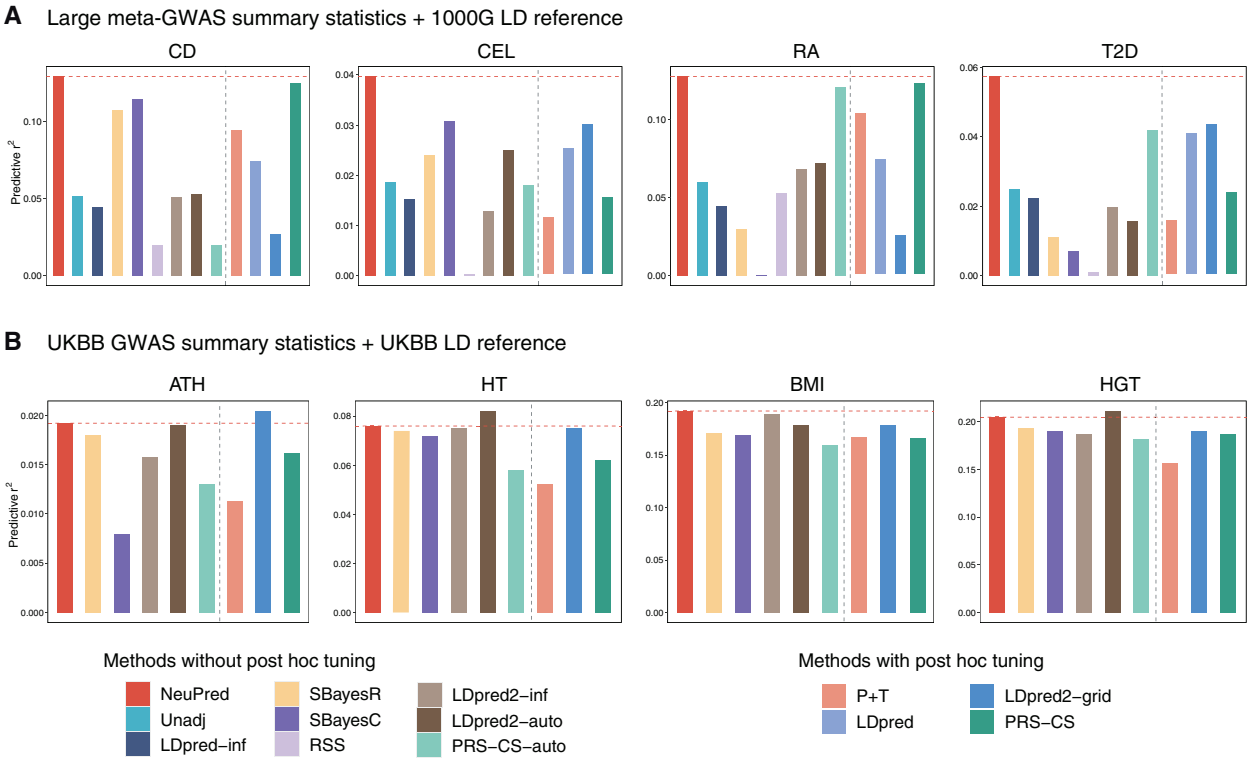

Fig. 2. Comparison of prediction accuracy among NeuPred and other 12 methods on real data experiments. (**a**) Predictive r^2 on the four diseases (CD, CEL, RA and T2D) with large-scale GWAS studies. PRSs were derived from summary statistics of GWAS studies, and the AUC was evaluated based on independent test datasets. The LD matrix was externally estimated from the 1000G reference panel. (**b**) Predictive r^2 on the four diseases (ATH, HT, BMI and HGT) with GWAS studies from UKBB. The AUC was evaluated based on independent test datasets. The LD matrix was estimated from the UKBB European individuals

Table 1. AUC of summary-statistics-based PRS methods for four diseases with large-scale GWAS studies and independent test data

Trait	Without *post hoc* tuning									With *post hoc* tuning			
	NeuPred	Unadj PRS	LDpred-inf	SBayesR	SBayesC	RSS	PRS-CS-auto	LDpred2-inf	LDpred2-auto	P + T	LDpred	PRS-CS	LDpred2-grid
CD	**0.712**	0.632	0.623	0.692	0.698	0.584	0.584	0.631	0.633	0.679	0.661	0.707	0.632
CEL	**0.630**	0.594	0.585	0.618	0.617	0.508	0.587	0.571	0.607	0.572	0.606	0.584	0.625
RA	**0.710**	0.645	0.625	0.598	0.608	0.636	0.706	0.654	0.656	0.688	0.662	0.704	0.596
T2D	**0.632**	0.587	0.581	0.604	0.619	0.523	0.616	0.575	0.565	0.567	0.614	0.584	0.614

Note: The four diseases are Crohn's disease (CD), celiac disease (CEL), rheumatoid arthritis (RA) and type 2 diabetes (T2D). The LD matrix was externally estimated from the 1000G. The UKBB data were used for *post hoc* parameter tuning for P + T, LDpred, PRS-CS and LDpred2-grid. The highest AUC is highlighted in boldface.

have covered most genetic structures in our experiments. We also propose a novel CV strategy for choosing the best performing prior when only the GWAS summary statistics are available in the training dataset, and show that its performances are comparable to the same CV-procedure applied to the corresponding individual-level genotypes when such data are available, especially when dimension p is high. Empirically, users may also specify a proper prior based on our knowledge about the target disease and use our software to derive posterior effect size estimates.

We observe that the Neu-HS prior performed the best for polygenic genetic architectures, in which a large number of SNPs are disease-associated but with very small effects. This is also consistent with the findings in Moser *et al.* (2015). The Neu-SpSL-C prior is powerful for capturing strong signals as its heavy tail helps keep large effects unaffected, while its spike part shrinks small coefficients to zero. Equipped with the CV procedure, NeuPred automatically selects the most appropriate prior for each chromosome. When estimating PRS with an external LD matrix, however, we recommend to use Neu-SpSL-L, since it is lighter-tailed, shrinks more strongly, and is more robust than Neu-SpSL-C to a potential mismatch between the target population and the LD reference panel.

In real data applications, we found that although sample sizes of GWAS studies were larger, the prediction accuracy on an independent test dataset was not as high as that for the WTCCC studies with in-sample LD reference. There are several reasons: (i) The genetic signals across training and test cohorts in the WTCCC studies are well matched, whereas there could be differences in sample ascertainment in an external validation dataset; (ii) The bias in an external LD reference may have also led to the decrease in prediction accuracy, especially for the 1000G reference panel, which contains only 489 individuals after quality control; (iii) The disparity of case/control ratios further made the predictive r^2 not comparable across cohorts. In theory, optimizing CV (for predictive likelihood) is equivalent to optimizing AIC asymptotically, thus tends to lead to a larger model than the true one (i.e. is not model consistent, whereas

the BIC procedure is model-selection consistent). However, for training a proper model and in the lack of future test data, the CV procedure still appears to be one of the best available general methods.

As mentioned earlier, the calibration slope should be used with caution. Vach (2013) commented that a calibration can look perfect even if all regression coefficients are underestimated, and introduced the concept of 'bias slope' ($\hat{\beta}_{\text{bias}}$), which is the slope for regressing the predicted scores against the true phenotype. Clearly, the product $\beta_{\text{calib}} \times \hat{\beta}_{\text{bias}}$ is equal to predictive r^2. In a clinical context, the bias perspective is more relevant as the aim of a prediction rule should be to inform a patient about the prognosis, while the calibration slope focuses on whether patients with a certain estimated probability can expect on average to experience an event rate equal to this value. Therefore, it is hard to tell whether a calibration slope close to one really comes from a well calibrated model, or from a sacrifice of the bias slope, especially when the predictive r^2 does not reach the level of SNP-based heritability. In real data applications, if calibration is really desired, a post-training adjustment could be adopted, such as multiplying the estimated regression vector by a constant, to make a better calibration.

For future directions, it is conceptually advantageous to jointly model multiple genetically correlated traits and functional annotations to further improve the prediction accuracy (Hu *et al.*, 2017a,b; Turley *et al.*, 2018). We believe that the concise and unified form of neuronized priors can also be used in such an attempt to bring in additional gains. An additional direction for further exploration is to discern small effects from the noise by taking further advantage of existing genetic knowledge. Similar to most PRS studies, we currently removed markers with MAF smaller than 0.005 and focused on the common variants. A better way to model rare variants awaits for a future careful investigation. Our current work imposes independence among the β_j's *a priori*. It may be of interest to induce correlations among the β_j's based on our genetic knowledge. This can be achieved in our NeuPred framework by, say, modeling the α_j's in Equation (4) as a Markov chain.

Acknowledgements

The authors thank Minsuk Shin for helpful discussions regarding neuronized priors, and the knowledgeable referees for their insightful and constructive comments and suggestions. This study makes use of data generated by the Wellcome Trust Case Control Consortium (2007). A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. This research has been conducted using the publicly available UK Biobank and FinnGen GWAS summary statistics released by Neale's lab. The author also thank the investigators of GERA, DIAGRAM, IIBDGC, NIDDK, GABRIEL and the NUgene project (www.nugene.org) for providing GWAS data. We want to acknowledge the participants and investigators of the FinnGen study.

Funding

This work was supported in part by National Science Foundation [DMS-1903139 and DMS-2015411 to J.S.L.]; and the National Natural Science Foundation of China [12071243 to L.H.].

Conflict of Interest: none declared.

Data availability statements

The data underlying this article are available in the article and in its online supplementary material.

References

Allen,H.L. *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**, 832–838.

Berger,J. (1980) A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Stat.*, **8**, 716.

Berisa,T. and Pickrell,J.K. (2016) Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, **32**, 283–285.

Bulik-Sullivan,B.K. *et al.*; , Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2015) LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, **47**, 291–295.

Chang,C.C. *et al.* (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 7.

Chatterjee,N. *et al.* (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.*, **45**, 400–405.

Chen,T.-H. *et al.* (2020) A penalized regression framework for building polygenic risk models based on summary statistics from genome-wide association studies and incorporating external information. *J. Am. Stat. Assoc.*, **116**, 1–11.

Consortium,I.S. *et al.* (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748.

Dubois,P.C. *et al.* (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.*, **42**, 295–302.

Dudbridge,F. (2013) Power and predictive accuracy of polygenic risk scores. *PLoS Genet.*, **9**, e1003348.

Franke,A. *et al.* (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.*, **42**, 1118–1125.

Garner,C. *et al.* (2014) Genome-wide association study of celiac disease in North America confirms FRMD4B as new celiac locus. *PLoS One*, **9**, e101428.

Ge,T. *et al.* (2019) Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.*, **10**, 1–10.

Habier,D. *et al.* (2011) Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, **12**, 186–112.

Hu,Y. *et al.* (2017a) Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS Genet.*, **13**, e1006836.

Hu,Y. *et al.* (2017b) Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput. Biol.*, **13**, e1005589.

Jiang,W. *et al.* (2021) A set of efficient methods to generate high-dimensional binary data with specified correlation structures. *Am. Stat.*, **75**, 310–322.

Jostins,L. and Barrett,J.C. (2011) Genetic risk prediction in complex disease. *Hum. Mol. Genet.*, **20**, R182–R188.

Ledoit,O. and Wolf,M. (2015) Spectrum estimation: a unified framework for covariance matrix estimation and PCA in large dimensions. *J. Multivar. Anal.*, **139**, 360–384.

Ledoit,O. and Wolf,M. (2017) Numerical implementation of the QuEST function. *Comput. Stat. Data Anal.*, **115**, 199–223.

Liu,J.S. and Sabatti,C. (2000) Generalised gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika*, **87**, 353–369.

Lloyd-Jones,L.R. *et al.* (2019) Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.*, **10**, 1–11.

Mak,T.S.H. *et al.* (2017) Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.*, **41**, 469–480.

Moffatt,M.F. *et al.*; GABRIEL Consortium. (2010) A large-scale, consortium--based genomewide association study of asthma. *N. Engl. J. Med.*, **363**, 1211–1221.

Morris,A.P. *et al.*; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.*, **44**, 981–990.

Moser,G. *et al.* (2015) Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet.*, **11**, e1004969.

Pan-UKB Team (2020) https://pan.ukbb.broadinstitute.org (29 October 2020, date last accessed).

Park,J.H. *et al.* (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.*, **42**, 570–575.

Pattee,J. and Pan,W. (2020) Penalized regression and model selection methods for polygenic scores on summary statistics. *PLoS Comput. Biol.*, **16**, e1008271.

Prentice,R.L. (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033–1048.

Privé,F. *et al.* (2021) LDpred2: better, faster, stronger. *Bioinformatics*, **36**, 5424–5431.

Ripke,S. *et al.* (2011) Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.*, **43**, 969.

Shin,M. and Liu,J.S. (2021) Neuronized priors for Bayesian sparse linear regression. *J. Am. Stat. Assoc.*, 1–43. https://doi.org/10.1080/01621459.2021.1876710.

Song,S. *et al.* (2020) Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. *PLoS Comput. Biol.*, **16**, e1007565.

Stahl,E.A. *et al.*; BIRAC Consortium. (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.*, **42**, 508–514.

Stevens,R.J. and Poppe,K.K. (2020) Validation of clinical prediction models: what does the "calibration slope" really measure? *J. Clin. Epidemiol.*, **118**, 93–99.

Strawderman,W.E. (1971) Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Stat.*, **42**, 385–388.

Turcot,V. *et al.*; CHD Exome+ Consortium. (2018) Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. *Nat. Genet.*, **50**, 26–41.

Turley,P. *et al.*; Social Science Genetic Association Consortium. (2018) Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.*, **50**, 229–237.

Vach,W. (2013) Calibration of clinical prediction rules does not just assess bias. *J. Clin. Epidemiol.*, **66**, 1296–1301.

Vilhjálmsson,B.J. *et al.*; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study. (2015) Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.*, **97**, 576–592.

Wang,J. (2020) Calibration slope versus discrimination slope: shoes on the wrong feet. *J. Clin. Epidemiol.*, **125**, 161–162.

Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661.

Wood,A.R., LifeLines Cohort Study. *et al.* (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.*, **46**, 1173–1186.

Yang,S. and Zhou,X. (2020) Accurate and scalable construction of polygenic scores in large biobank data sets. *Am. J. Hum. Genet.*, **106**, 679–693.

Zhang,Y. *et al.* (2018) Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.*, **50**, 1318–1326.

Zhou,G. and Zhao,H. (2021) A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics. *PLoS Genet.*, **17**, e1009697.

Zhu,X. and Stephens,M. (2017) Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann. Appl. Stat.*, **11**, 1561–1592.