scientific reports

OPEN

Machine learning models for predicting blood pressure phenotypes by combining multiple polygenic risk scores

Yana Hrytsenko[1,2,3,37], Benjamin Shea[3,37], Michael Elgart[1,2], Nuzulul Kurniansyah[1], Genevieve Lyons[4], Alanna C. Morrison[5], April P. Carson[6], Bernhard Haring[7,8], Braxton D. Mitchell[9], Bruce M. Psaty[10,11,12,13], Byron C. Jaeger[14], C. Charles Gu[15], Charles Kooperberg[16], Daniel Levy[17,18], Donald Lloyd-Jones[19], Eunhee Choi[20], Jennifer A. Brody[10,12], Jennifer A. Smith[21,22], Jerome I. Rotter[23], Matthew Moll[1,2,24,33], Myriam Fornage[5,25], Noah Simon[26], Peter Castaldi[1,2], Ramon Casanova[14], Ren-Hua Chung[27], Robert Kaplan[7,16], Ruth J. F. Loos[28,29], Sharon L. R. Kardia[21], Stephen S. Rich[30], Susan Redline[1,2,31], Tanika Kelly[32], Timothy O'Connor[9,35,36], Wei Zhao[21,22], Wonji Kim[33], Xiuqing Guo[23], Yii-Der Ida Chen[23], The Trans-Omics in Precision Medicine Consortium* & Tamar Sofer[1,2,3,4,34]✉

We construct non-linear machine learning (ML) prediction models for systolic and diastolic blood pressure (SBP, DBP) using demographic and clinical variables and polygenic risk scores (PRSs). We developed a two-model ensemble, consisting of a baseline model, where prediction is based on demographic and clinical variables only, and a genetic model, where we also include PRSs. We evaluate the use of a linear versus a non-linear model at both the baseline and the genetic model levels and assess the improvement in performance when incorporating multiple PRSs. We report the ensemble model's performance as percentage variance explained (PVE) on a held-out test dataset. A non-linear baseline model improved the PVEs from 28.1 to 30.1% (SBP) and 14.3% to 17.4% (DBP) compared with a linear baseline model. Including seven PRSs in the genetic model computed based on the largest available GWAS of SBP/DBP improved the genetic model PVE from 4.8 to 5.1% (SBP) and 4.7 to 5% (DBP) compared to using a single PRS. Adding additional 14 PRSs computed based on two independent GWASs further increased the genetic model PVE to 6.3% (SBP) and 5.7% (DBP). PVE differed across self-reported race/ethnicity groups, with primarily all non-White groups benefitting from the inclusion of additional PRSs. In summary, non-linear ML models improves BP prediction in models incorporating diverse populations.

[1]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [2]Department of Medicine, Harvard Medical School, Boston, MA, USA. [3]CardioVascular Institute (CVI), Beth Israel Deaconess Medical Center, Boston, MA, USA. [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [5]Department of Epidemiology, School of Public Health, Human Genetics Center, The University of Texas Health Science Center at Houston, Houston, TX, USA. [6]Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA. [7]Department of Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, NY, USA. [8]Department of Medicine III, Saarland University, Homburg, Saarland, Germany. [9]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. [10]Department of Medicine, University of Washington, Seattle, WA, USA. [11]Department of Epidemiology, University of Washington, Seattle, WA, USA. [12]Cardiovascular Health Research Unit, University of Washington, Seattle, WA, USA. [13]Health Systems and Population Health, University of Washington, Seattle, WA, USA. [14]Department of Biostatistics and Data Science, Wake Forest University School of Medicine, Winston-Salem, NC, USA. [15]The Center for Biostatistics and Data Science, Washington University, St. Louis, USA. [16]Division of Public Health Sciences, Fred Hutchinson Cancer

Center, Seattle, WA, USA. [17]The Population Sciences Branch of the National Heart, Lung and Blood Institute, Bethesda, MD, USA. [18]The Framingham Heart Study, Framingham, MA, USA. [19]Department of Preventive Medicine, Northwestern University, Chicago, IL, USA. [20]Columbia Hypertension Laboratory, Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA. [21]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA. [22]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA. [23]Department of Pediatrics, The Institute for Translational Genomics and Population Sciences, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA. [24]VA Boston Healthcare System, West Roxbury, MA, USA. [25]Brown Foundation Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX, USA. [26]Department of Biostatistics, School of Public Health, University of Washington, Seattle, WA, USA. [27]Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Taipei City, Taiwan. [28]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [29]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty for Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [30]Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA, USA. [31]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA, USA. [32]Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA. [33]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, USA. [34]Center for Life Sciences CLS-934, 3 Blackfan St., Boston, MA 02115, USA. [35]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA. [36]Program in Health Equity and Population Health, University of Maryland School of Medicine, Baltimore, MD, USA. [37]These authors contributed equally: Yana Hrytsenko and Benjamin Shea. [*]A list of authors and their affiliations appears at the end of the paper. ✉email: tsofer@bidmc.harvard.edu

Polygenic (risk) scores (PRSs) summarize information from many genetic variants across the genome. PRSs are being increasingly developed for risk prediction and for quantifying the inherited predisposition for a given trait or condition. The number/dosage of associated alleles, typically weighted according to effect size estimates from a genome-wide association study (GWAS) for a given phenotype, are summed to produce a PRS for an individual[1]. Commonly, PRS studies involve testing for the association between a PRS and a trait in the target data and estimating its effect. However, PRSs rely on linear relationship between allele counts and the outcome[2] and do not account for potential interactions between SNPs[3] and non-linear associations between genetic variants and the outcome of interest. Linear prediction models that rely on standard PRS therefore usually only explain a fraction of observed genetic variance. Recently, we developed a non-linear machine learning (ML) model that incorporated both individual SNPs and a PRS for predicting predisposition to a certain trait. We showed that it improves Percent Variance Explained (PVE) in an independent test dataset over the standard approach, in which a single PRS is incorporated into a linear prediction model, in a dataset comprising diverse individuals from multiple self-reported race/ethnic groups[4]. However, due to their potential large number, inclusion of individual SNPs may lead to both high computational burden and to model overfitting to the training dataset, where a model performs poorly on a new dataset (i.e., data that were not used in the training dataset). While feature selection tools may be applied to reduce the number of SNPs, e.g., using least absolute shrinkage and selection operator (LASSO[5]; used in Elgart et al.[4]), these tools may be limited by incorporating an assumption of linearity.

Other PRS prediction approaches improve upon the single-PRS models by employing multiple PRSs calculated from several GWASs, also known as "multi-PRS" approaches[6]. The goal of incorporating multiple PRSs in the model is to utilize the discoveries of multiple GWASs (multi-trait, multi-ancestry) and thus boost the model's performance. Increase in PVE using multi-PRS model compared with the best single-score predictions was reported by[7] in the context of using PRSs of multiple traits. Other studies also reported improvement in association analysis when utilizing multiple PRSs compared with a single PRS for a single trait association, with and without a PRS selection step[7–12]. Overall, studies comparing "single-PRS" approaches and "multi-PRS" approaches showed higher performance of multiple PRS models[6,13–15]. While some multi-PRS models combine PRSs based on different GWASs, other multi-PRS models construct several PRSs from the same GWAS, usually based on multiple p-value (significance) thresholds—typically when using the clump & threshold methodology. With the clump & threshold methodology, PRS construction requires setting a p-value parameter, for which a set of optimal SNPs is selected to calculate the score. However, there is no single optimal p-value threshold that is known a priori. Thus, one strategy for multiple clump & threshold PRSs is to construct PRS for several different p-value thresholds and then include all PRSs in the analysis[16]. Coombes et al.[17] proposed to perform a principal component analysis over a set of PRSs calculated for a range of clump & threshold parameter settings and then using the first "PRS-PC" for the association testing. The main motivation behind the method is that the largest amount of variation in the computed PRSs is captured by the first PRS-PC, thus potentially improving discrimination of the phenotype tested. Thus, multi-PRS approaches combining PRSs from multiple GWAS, and approaches combining multiple PRSs from the same GWAS, have been shown to improve PRS models, where the first approach (multiple GWAS PRSs) has been particularly useful for improving PRS models in diverse populations.

Blood pressure (BP) is highly polygenic and, when elevated, is one of the primary risk factors for the development of several cardiovascular diseases such as coronary artery disease and stroke[18,19]. Nearly half of the adults in the U.S have hypertension[20], with higher prevalence among adults who are Black, compared to other subgroups[21,22]. PRSs have been developed to predict BP phenotypes across the lifespan[23–26]. In prior work, we showed that combining multiple PRSs based on a few GWASs, from populations of differences ancestral make-ups, improves BP PRS models in ancestrally and race/ethnicity diverse population[13]. Our work also noted that PRS effect sizes and performance vary by strata defined by important clinical BP predictors, such as age groups, biological sex, and obesity. Non-linear ML models more naturally account for differences in relationship between variables across population subsets. Thus, BP phenotypes may especially benefit from non-linear ML models. Finally, another important motivation for the development of non-linear ML models that include multiple PRSs is the potential for improved prediction accuracy across diverse populations, in a single model. The hope is that the developed model will use both clinical- and genetic ancestry-related characteristics, including their interaction, as needed.

Here, we develop multi-PRS non-linear ML models and assess their association with systolic and diastolic BP. For each of the two BP outcomes, we develop an ensemble machine learning model that makes a BP prediction based on demographic (age, sex, self-reported race/ethnic background, and study center), BMI, and SBP/DBP PRSs. The ensemble model consists of two consecutive components—a baseline model and a genetic model. The baseline model predicts a phenotype using the set of covariates without the genetic component, and the genetic model further explains the residuals from the baseline model. We evaluate the use of a linear versus a non-linear model at both the baseline and the genetic model level and assess the improvement in performance when incorporating multiple PRSs based on several GWAS and p-value thresholds. We compare the ensemble model's performance on the held-out test dataset stratified by self-reported race/ethnic groups. In secondary analyses, we also developed PRSs specific to linkage disequilibrium (LD) regions, referred to as local-PRSs, and evaluated the possibility of using multi-local-PRSs.

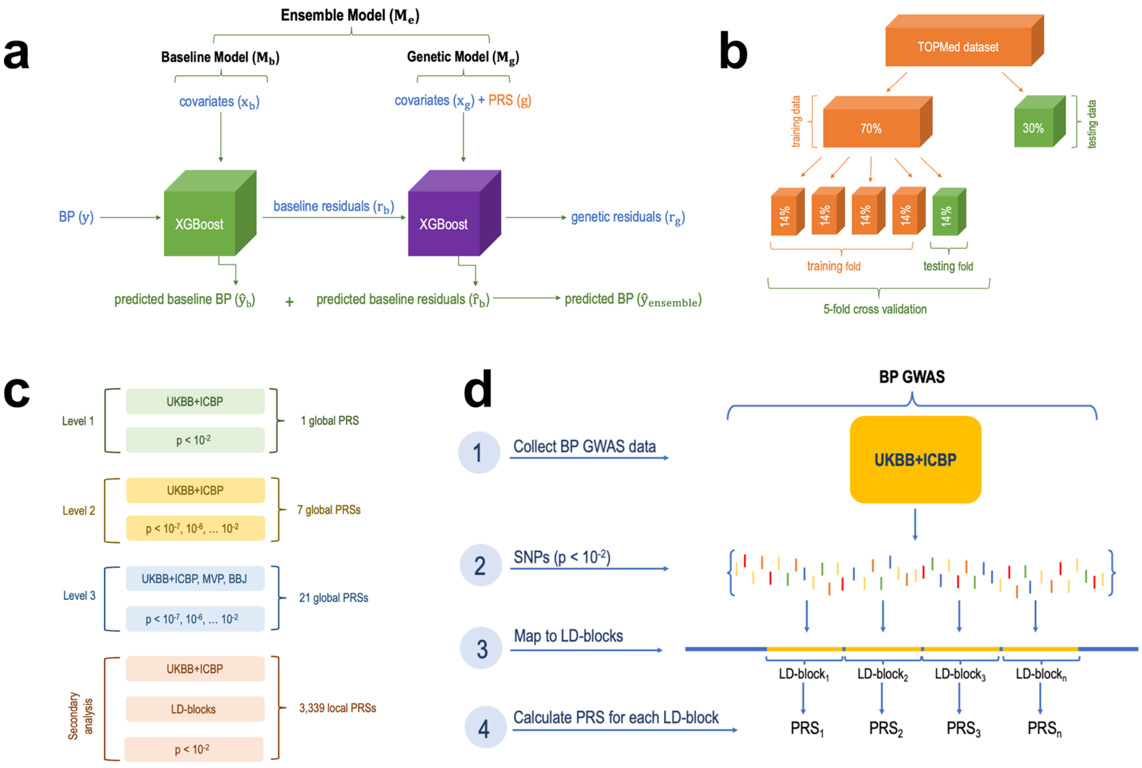

Figure 1. Study design. (**a**) The proposed ensemble model framework. The ensemble is composed of two models. The baseline model, trained on covariates (X_b) only for prediction of SBP and DBP ($\widehat{y}_b$). To assess the accuracy of the baseline model we calculated the residuals (baseline residuals r_b) by subtracting the predicted value of SBP/DBP from the actual value of SBP/DBP. The genetic model was trained on a subset of the covariates, and genetic components (global PRSs) for prediction of the baseline model residuals r_b. We measured the accuracy of the genetic model by subtracting predicted genetic residuals $\widehat{r}_g$ from baseline residuals r_b. The overall prediction of BP by the ensemble model is the sum of the predicted baseline BP $\widehat{y}_b$ (by the baseline model) and the predicted baseline residuals $\widehat{r}_b$ (by the genetic model). The accuracy of the ensemble model was assessed by calculating percent variance explained (PVE) by two models jointly. (**b**) The split of the primary, TOPMed dataset, into training and testing sets followed by the fivefold cross validation procedure where the training dataset is further split into 5 equal parts with one part designated for testing (repeated 5 times with 1/5 of the training data being designated at random for testing at each iteration). (**c**) Increasing levels of genetic models' complexity where each new model included additional PRSs. (**d**) The process of calculating local PRSs per LD-blocks (secondary analysis). *BBJ* BioBank Japan, *BP* blood pressure, *GWAS* genome wide association study, *LD* linkage disequilibrium, *Level* model complexity level, *MVP* Million Veteran Program, *P* p-value threshold, *PRS* polygenic risk score, *SNPs* single-nucleotide polymorphisms, *TOPMed* Trans-Omics for Precision Medicine, *UKBB + ICBP* UK Biobank and International Consortium for Blood Pressure.

Results

Figure 1 visualizes the study design and major steps in development and assessment of the ensemble model. The ensemble model included two components: a baseline and a genetic model, where we compared multiple constructions of genetic models. We used cross validation to tune model parameters in the training dataset and evaluated the models in the independent test dataset. Models were trained using 70% of the available data from our multi-ethnic dataset and tested in the remaining 30%.

TOPMed participant characteristics

We used a multi-ethnic dataset from the TOPMed consortium (freeze 8 release) to train non-linear ML models using PRSs for systolic and diastolic BP (SBP and DBP, respectively). The dataset included 62,295 unrelated participants from fifteen U.S.- and Taiwan-based studies. Participant characteristics are provided in Table 1. Individuals self-identified according to categories of race and ethnicity: there were 14,587 Black participants, 30,668 White participants, 4655 Asian participants, 11,904 Hispanic/Latino participants and 481 participants of "Other" or "Unknown" decent. Descriptions of each of the contributing TOPMed studies are provided in Supplementary Note 1.

The TOPMed dataset was randomly split into a training dataset (70% of the individuals) in which fivefold cross-validation was used to choose tuning parameters for models, and a held-out test dataset (30% of the individuals) in which the trained models' predictions were evaluated. Supplementary Table 1 characterizes the training TOPMed dataset. Characteristics of the held-out test TOPMed dataset are provide in Supplementary Table 2. Characteristics broken down by specific TOPMed studies are provided in Supplementary Table 3. In brief, due to the random split to the training and test datasets, both datasets have similar characteristics, including about 49% self-reported White, 23% Black, 19% Hispanic/Latino, 7% Asian individuals and < 1% of Other or Unknown race/ethnicity. The mean age is approximately 55 years, 63% of the participants are female, and 59% are hypertensive, with hypertension being more common in self-reported Black individuals and least common in self-reported Asian individuals.

Non-linear ML models improve modelling of covariate effects compared to linear models

For each BP outcome, we trained ensemble models using the TOPMed training dataset. The first model in the ensemble, referred to as the baseline model, included the covariates age, sex, BMI, race/ethnic background, and study center, where the latter is a dataset-specific variable. Figure 2 visualizes the phenotypic independent test dataset PVE obtained from the baseline model when fitted using a non-linear ML model (a gradient boosting trees model fitted using the XGBoost package) and using a linear model (see Supplementary Table 4 for baseline model complete results). The non-linear ML model had higher PVE for both SBP and DBP, both when evaluated over the complete dataset and by groups defined by self-reported race/ethnic background. Therefore, we proceeded with ensemble models with non-linear ML baseline model. Complete results, including from cross-validated PVE in the training datasets, are provided in Supplementary Table 5. Surprisingly, we observed slightly lower performance in models trained using global SBP/DBP PRSs developed using Bayesian approach, PRS-CSx (complete results are reported in the Supplementary Table 6). The hyperparameters for the XGBoost model and their values selected after tuning are listed in Supplementary Table 7. For the baseline non-linear ML model, the phenotypic PVE was higher for SBP prediction (30% PVE in the race/ethnicity combined dataset) than for DBP (17.4% phenotypic PVE), as reported in other PRS studies of the two phenotypes. When tested on the testing set stratified by race/ethnicity, the PVE was highest in the group of Black individuals, for both SBP and DBP (33.4% and 22.4%, respectively). PVE was lowest in the group of Asian individuals (19.7% and 11.9% for SBP and DBP, respectively).

In Supplementary Fig. 1, we also report results from a secondary analysis comparing baseline models with and without inclusion of genetic PCs, demonstrating that PCs inclusion does not improve the baseline model PVE. Thus, and given that use of PCs challenges the transferability of prediction models between datasets, we did not include them.

Characteristic	White	Black	Hispanic/Latino	Asian	Other/unknown
N	30,668	14,587	11,904	4655	481
Gender (N (%))					
Female	20,218 (66%)	9165 (63%)	7108 (60%)	2376 (51%)	198 (41%)
Male	10,450 (34%)	5422 (37%)	4796 (40%)	2279 (49%)	283 (59%)
Age, years (median, IQR))	60 (50, 69)	55 (47, 64)	52 (43, 61)	48 (40, 57)	59 (50, 67)
SBP, mmHg (median, IQR))	126 (113, 142)	133 (119, 150)	126 (113, 144)	122 (110, 138)	134 (116, 152)
DBP, mmHg (median, IQR))	75 (68,83)	80 (72, 89)	76 (68, 84)	75 (68, 85)	76 (66, 86)
BMI, kg/m^2 (median, IQR))	26.4 (23.5, 30.0)	29.0 (25.1, 34.0)	29.0 (26.0, 33.0)	23.9 (21.8, 26.2)	27.3 (24.3, 31.5)
Hypertensive (N (%))	17,274 (56%)	10,103 (69%)	6733 (57%)	2381 (51%)	323 (67%)

Table 1. TOPMed dataset (training and testing sets combined) characteristics aggregated over the studies, stratified by self-reported race/ethnic background. Hypertension was defined as SBP ≥ 130, DBP ≥ 80, or use of antihypertensive medications. *IQR* interquartile range.

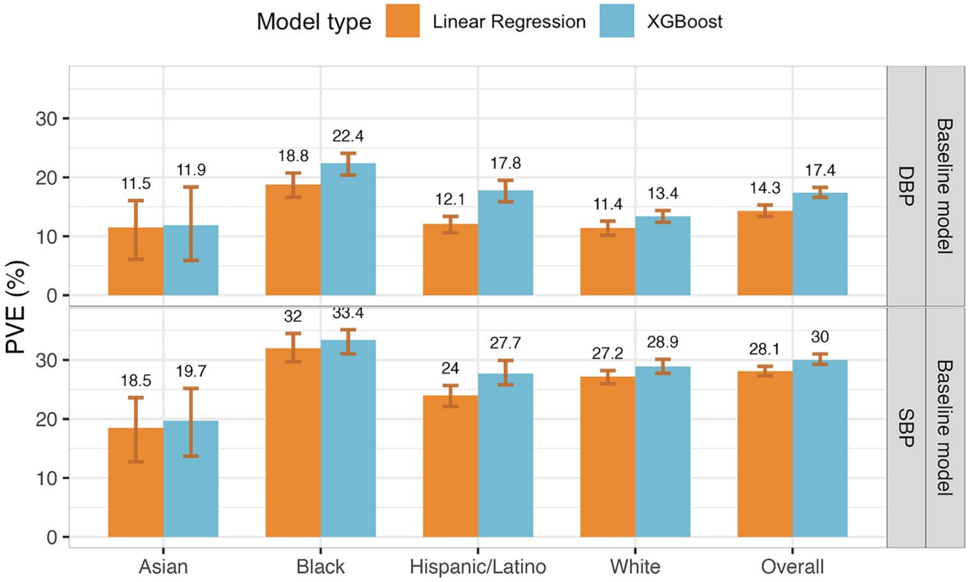

Figure 2. Estimated phenotypic PVE of baseline models fitted using non-linear ML and linear models. Estimated PVEs in the TOPMed test dataset for baseline model performance for prediction of SBP and DBP in the overall test dataset and stratified by self-reported race/ethnicity (White N = 10,877, Hispanic/Latino N = 3831, Black N = 3657, Asian N = 403 for DBP; White N = 10,823, Hispanic/Latino N = 3877, Black N = 3674, Asian N = 374 for SBP). The visualized 95% confidence intervals were computed as the 2.5% and 97.5% percentiles of the bootstrap distribution of the PVEs estimated over the test dataset. *PVE* percent variance explained, *TOPMed* Trans-Omics for Precision Medicine, *SBP* systolic blood pressure, *DBP* diastolic blood pressure.

Using multiple PRSs from the same and from multiple GWASs improves model performance in Asian, Black, and Hispanic participants

The second part of the ensemble model is the genetic model, which included covariates that can be used across datasets, including age, sex, BMI, self-reported race/ethnic background, and SBP/DBP PRS measures. The ensemble model used non-linear ML baseline model (because it performed better than linear regression), and genetic models fitted using either non-linear ML or using conventional linear regression. Further, genetic models were of increasing complexity where we included one or more PRSs according to the following logic: model complexity 1 included a single PRS based on the clump & threshold methodology using p-value threshold of 10^{-2} and using summary statistics from the BP GWAS of the UK Biobank and the International Consortium of Blood Pressure (UKBB + ICBP), which yielded the most powerful single-GWAS PRSs in a past paper developing BP PRSs[13]. An analysis using the TOPMed training datasets, comparing the inclusion of only a single PRS in the genetic model, further suggested that this PRS yields the highest PVE in models that use all individuals and just a single one of the UKBB-ICBP based PRSs, however, notably, the training dataset optimal PRS threshold differed by race/ethnic group (Supplementary Table 8 summarizes the clump & threshold p-value threshold that yielded the highest PRS for each group, and Supplementary Table 9 provides the PVE values for each PRS and group). Model complexity 2 included 7 PRSs based on the clump & threshold methodology each using a different p-value threshold for SNP inclusion, and using the same UKBB + ICBP GWAS. Model complexity 3 included 21 PRSs, 7 PRSs from each of the UKBB + ICBP, Million Veteran Program (MVP), and Biobank Japan (BBJ). PRSs based on all GWAS were constructed using the same clump & threshold approach with the same p-value thresholds used in model complexity 2.

Supplementary Table 5 reports the complete performance results as attained PVEs from the genetic models (PVEs of predicting residuals from the baseline model) and ensemble models (PVEs of predicting the raw trait) estimated in cross validation on the training dataset, and from the independent test dataset. Genetic models fitted using the non-linear ML approach tended to have similar or better performance than genetic model fitted using linear regression (see Supplementary Fig. 2). Two exceptions were linear model performed better at prediction of DBP in Black individuals (2.2% PVE by model complexity level 3 linear regression compared with 1.7% PVE of the ML model) and prediction of SBP in Hispanic/Latino individuals (5.7% versus 4.6% in linear regression versus ML level 3 models). We therefore here focus on the genetic models (non-linear ML fitted using XGBoost).

Figure 3 visualizes the genetic model performance, measured by PVEs in prediction of residuals from the baseline model, and the ensemble model performance, measured by PVEs at the phenotypic level. Genetic model performance improved with the addition of PRSs, with improvement being large for the non-White groups, and low and almost not existing, for the group of self-reported White individuals. This is likely because the self-reported White individuals are mostly of European genetic ancestry, closely matching the genetic ancestry of the population participating in the UKBB + ICBP GWAS. Concretely, genetic model performance in the White group were 7.7%, 7.6%, and 7.8% for the three increasing complexity levels for SBP, and 7.3%, 6.7%, and 7.2% for

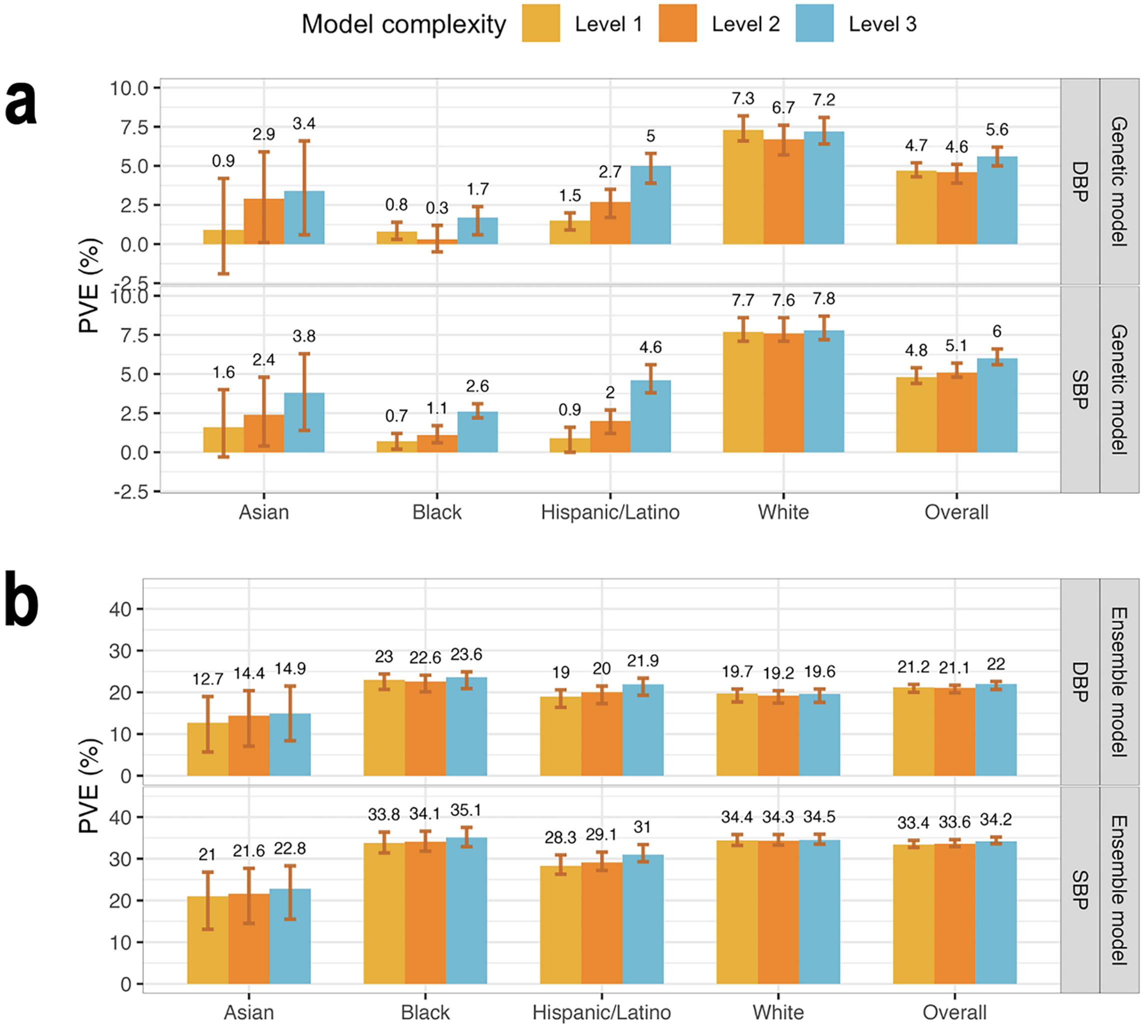

Figure 3. Comparison of genetic and ensemble model performance in TOPMed test dataset. (**a**) Estimated PVEs in the TOPMed test dataset obtained by genetic models incorporating one or more PRSs according to the three complexity levels. Level 1: a single PRS based on the UKBB + ICBP GWAS. Level 2: PRSs based on the UKBB + ICBP GWAS based on seven p-value thresholds. Level 3: 21 PRSs, 7 PRSs based on each of the UKBB + ICBP, MVP, and BBJ GWAS. PVE is reported for predicting residuals from the baseline model, where the baseline model was a non-linear ML model and only used non-genetic covariates. (**b**) Estimated PVEs in the TOPMed test dataset for ensemble model at the raw phenotypic level. PVEs are reported for models of SBP and DBP, in the overall test dataset and stratified by self-reported race/ethnicity (White N = 10,877, Hispanic/Latino N = 3831, Black N = 3657, Asian N = 403 for DBP; White N = 10,823, Hispanic/Latino N = 3877, Black N = 3674, Asian N = 374 for SBP). The visualized 95% confidence intervals were computed as the 2.5% and 97.5% percentiles of the bootstrap distribution of the PVEs estimated over the test dataset. *PVE* percent variance explained, *TOPMed* Trans-Omics for Precision Medicine, *SBP* systolic blood pressure, *DBP* diastolic blood pressure, *PRS* polygenic risk score.

DBP. In other individuals the improvement was strongly apparent. In Hispanic/Latino individuals and prediction of SBP, the PVEs were 0.9%, 2%, and 4.6%, and for DBP they were 1.5%, 2.7%, and 5%. In the Asian group the SBP PVEs were 1.6%, 2.4%, and 3.8%, and for DBP they were 0.9%, 2.9%, and 3.4%. Finally, Black individuals had the lowest genetic model performance with PVEs by complexity for SBP being 0.7%, 1.1%, and 2.6%, and for DBP 0.8%, 0.3%, and 1.7%. Therefore, including PRSs based on non-European GWAS summary statistics substantially contributed to the model performance in non-White individuals but only to a small extent, and only for SBP, for White individuals.

At the phenotypic, ensemble-model level, the improvement in PVE achieved by increasing the complexity of the genetic models is less impressive, because the covariates explained the lion's share of the phenotypic variance. Interestingly, while the genetic model had the highest level of PVE in the group of White individuals, the ensemble model, as a whole, had the highest PVE in the group of Black individuals. This was true for both SBP and DBP, and is already seen when using model complexity level 2 (multiple PRSs based only on the UKBB + ICBP GWAS). Supplementary Table 7 provides the tuning parameters selected by the cross validation for each of the models. Supplementary Table 10 reports timing and RAM use comparison between the models. Level 3 non-linear models fitted using XGBoost required the most time and memory, as expected, with 979 mebibytes and

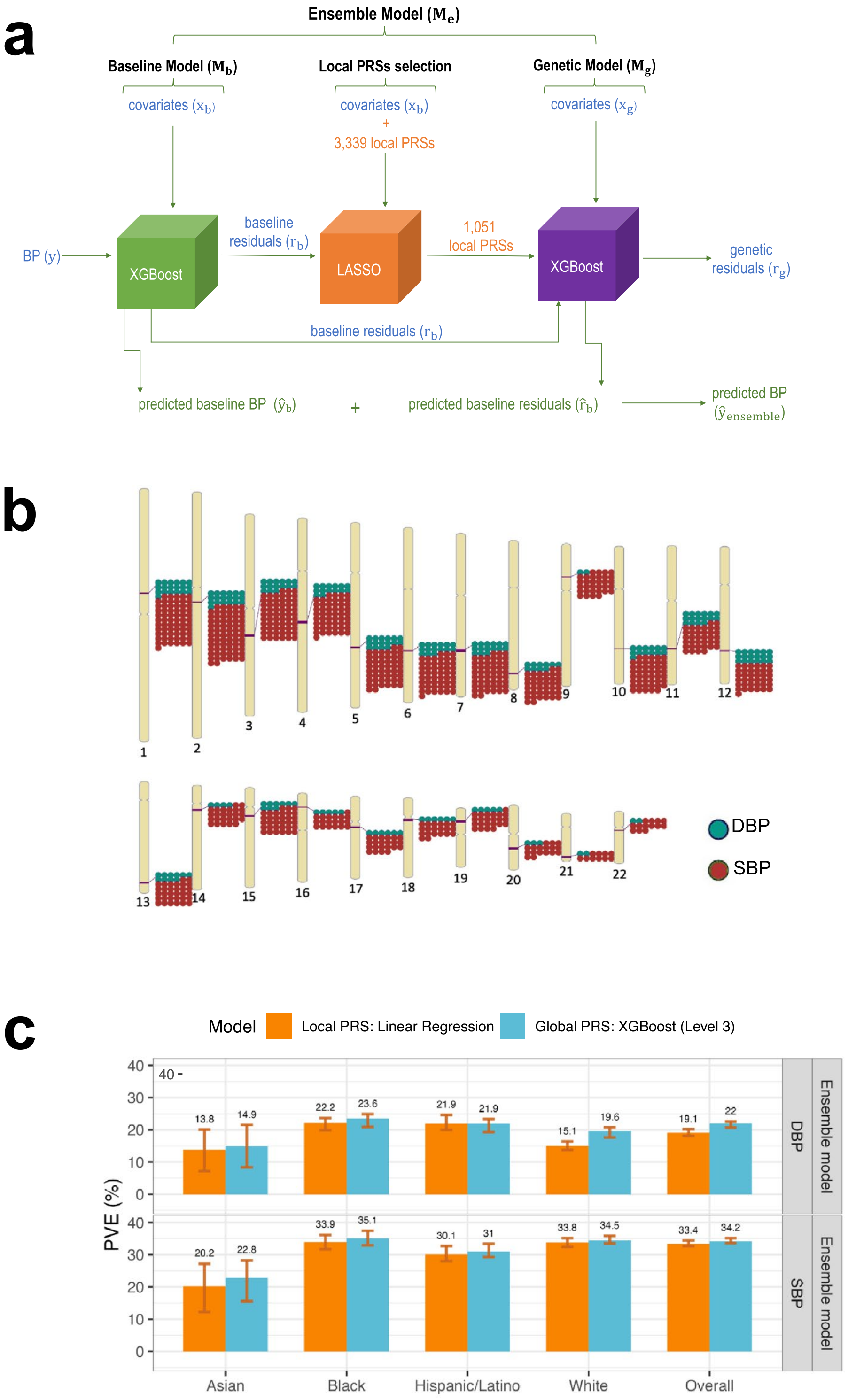

Figure 4. Integration of LASSO feature selection tool into the ensemble model workflow. (**a**) The workflow of the ensemble model with the integration of the LASSO variable selection tool. To include local PRSs in the ensemble model while attempting to avoid overfitting, we added a LASSO selection step to the ensemble model development. As visualized, the residuals of the baseline model were used as the outcome in LASSO penalized regression with the local PRSs as features. LASSO substantially reduced the number of local PRSs (to 827 for SBP and 224 for DBP). The local PRSs selected by LASSO were then used as an input into the genetic model for prediction of the baseline residuals ($\hat{r}_b$). (**b**) Genomic locations of local PRSs, calculated over predefined LD-regions, selected by LASSO for SBP and DBP. (**c**) Comparison between the estimated PVE in the TOPMed test dataset for ensemble model Level 3 using global PRSs and the ensemble model using Linear regression and local PRSs. PVEs are reported for models of SBP and DBP, in the overall test dataset and stratified by self-reported race/ethnicity (White N = 10,877, Hispanic/Latino N = 3831, Black N = 3657, Asian N = 403 for DBP; White N = 10,823, Hispanic/Latino N = 3877, Black N = 3674, Asian N = 374 for SBP). The visualized 95% confidence intervals were computed as the 2.5% and 97.5% percentiles of the bootstrap distribution of the PVEs estimated over the test dataset. *BP* blood pressure, *DBP* diastolic blood pressure, *LASSO* least absolute shrinkage and selection operator, *PRS* polygenic risk score, *SBP* systolic blood pressure.

close to 11 min (n ≈ 45 K individuals). Predictions from fitted models took less than a second for all models, and used up to 400 mebibytes (n ≈ 18.5 K).

Secondary analysis: using local PRSs in the genetic model

In secondary analysis, we developed a new model that included local PRSs (Fig. 4a), constructed based on the UKBB + ICBP GWAS, with each local PRS being based on summary statistics restricted to an LD-region, with regions defined according to European populations. These PRSs used the clump & threshold methodology with p-value threshold $< 10^{-2}$. Because gradient boosting trees models tend to overfit to the training dataset when many features are included, due to higher model complexity (see for instance the XGBoost tutorial https://xgboost.readthedocs.io/en/latest/index.html), we extended the development of the ensemble model to include a feature selection step using LASSO penalized regression, as described in Fig. 4a.

As before, we considered either non-linear ML or linear regression genetic models (using the LASSO-selected local PRSs). For comparison, we additionally used an ensemble model that used the fitted LASSO model itself as the genetic model. Note that the difference between the ensemble model with linear regression and with LASSO genetic model is that the linear regression model re-evaluated the coefficients of the local PRSs, while the ensemble model with the LASSO genetic model used the ℓ_1-penalized coefficients from the LASSO operation. We evaluated the SBP and DBP local PRS ensemble models' PVE. LASSO selected a subset of local PRSs which are most salient in model's prediction (i.e., with the lowest cost function). Specifically, out of 1670 local PRSs for SBP LASSO selected 827, and it selected 224 of the 1669 local PRSs for DBP. As shown in Fig. 4b, the selected local PRSs for SBP and DBP are concentrated in the same genomic regions, and, more generally, selected local PRSs are clustered in specific regions. As reported in Supplementary Table 10, the LASSO model took 18.6 min and 1735 mebibytes to tune and fit, and the subsequent non-linear ML model tunning and fitting took a little over 7.7 h and 1265 mebibytes of RAM.

Supplementary Figure 3 provides the PVE of the genetic and ensemble models implemented with the three models (non-linear ML, linear regression, and LASSO). Across self-reported race/ethnic groups the results varied and there is no one method that is always superior to others. Notably, non-linear ML genetic model is almost never the best model. Figure 4c compares the performance of the global PRS model (level 3) with the local PRS model that used a linear regression genetic model (because it tended to perform better than other genetic model specifications). The local PRS genetic model sometimes performed equally well compared to the global PRS model, despite using less information (local compared to genome-wide PRS). Potential reasons could be that the local PRS models allow for separately accounting for the contributions of different genomic LD-regions. Some regions may be more or less important than others in some individuals, based on genetics and/or covariate characteristics, so the flexibility of the local models may be useful especially in datasets including individuals representing diverse genetic backgrounds, lifestyles, and environmental exposures.

As the local PRSs are based on the UKBB + ICBP GWAS and were generated using the 10^{-2} p-value threshold, it is interesting to compare the genetic model that used them (results in Supplementary Fig. 3) to the genetic model with a single UKBB + ICBP PRS (model complexity level 1; results in Supplementary Fig. 2). Considering the combined TOPMed test dataset, for SBP, the genetic model using linear regression and local PRSs has better PVE (4.8%) compared to the linear regression genetic model with one global PRS (PVE = 4.4%). However, this was not true for DBP (linear regression local PRS model PVE = 2.6% versus 4.4% for global, single PRS linear regression). The non-linear ML global PRS model with complexity level 1 performed the same (for SBP) or better (DBP) than linear regression local PRS models. Surprisingly, when focusing on Hispanic/Latino individuals, local PRS models performed substantially better than global PRS complexity level 1 models: genetic local PRS model PVEs ranged from 1.6 to 5.7% (across methods and BP traits), while global PRS level 1 model PVEs were at most 1.5% (see Supplementary Fig. 2). Supplementary Table 11 reports complete performance results (attained PVEs) from the genetic models (PVEs of predicting residuals from the baseline model) and ensemble models (PVEs of predicting the raw trait) estimated in cross validation on the training dataset, and from the independent test dataset using local PRS in TOPMed dataset.

Challenges of model generalization to clinical data

Figure 5 visualizes workflow with the application of the model trained on the TOPMed dataset to the MGB Biobank (MGBB) data. For calibration, we first trained a baseline model on the MGBB data (covariates only) and calculated residuals. Next, we applied the genetic model trained on the TOPMed dataset to predict the residuals for the MGBB dataset. Because MGBB data is enriched with data related to hospital-based visits, we restricted the analysis to individuals with no antihypertensive related codes in the 2 years prior to data extraction (5/25/2021–5/25/2023), and only participants with data from this time period were used. There were 9,494 individuals meeting these inclusion criteria. Of these 7985 were White, 412 Black, and 200 Asian individuals. The mean age was 59, there are 63% female individuals. Supplementary Table 12 characterizes the study population.

Despite strict inclusion criteria, models performed less well on the MGBB dataset. The baseline model was fit in MGBB, for calibration purposes, i.e., to account for differences in populations between the TOPMed and MGBB datasets. Note that the MGBB baseline model was fit without a training–testing approach. We compared the approach of calibrating the baseline model to a new dataset. Figure 5b demonstrates the PVE of the TOPMed-trained genetic models with the baseline model being either TOPMed-trained or MGB-trained. One can see that the calibration approach worked well, in that genetic model performance was usually better when applied over the MGBB-trained baseline model. Supplementary Fig. 4 provides the cross-validated PVEs from the baseline non-linear ML model trained on the MGBB dataset. The cross-validated PVEs are often lower than the TOPMed test dataset PVEs (Fig. 2), other than in the Asian group, where the PVEs are similar. Supplementary Figure 5

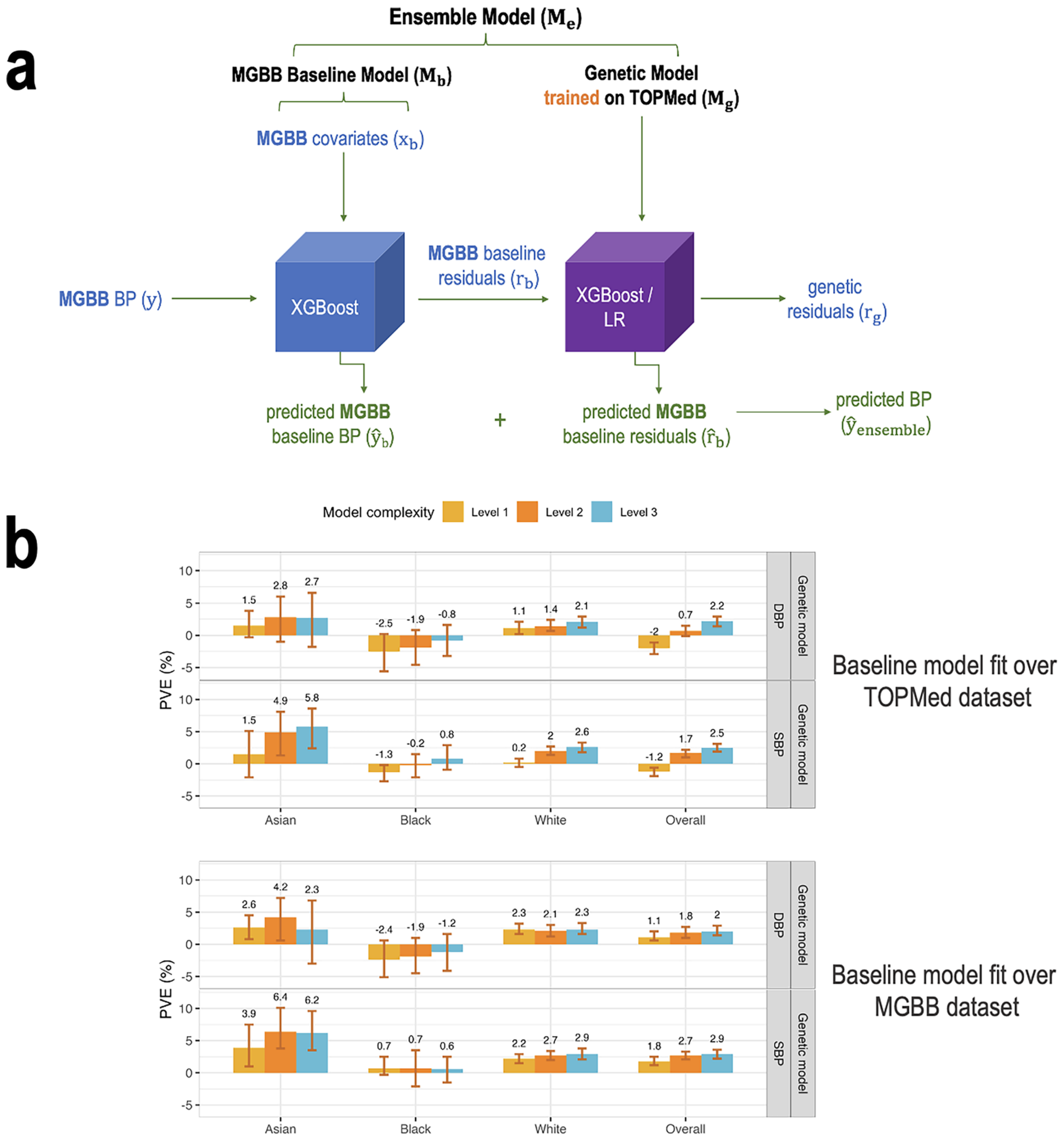

Figure 5. Application of the model trained on the TOPMed dataset to the MGBB data. (**a**) The figure visualizes the workflow of the Ensemble model with the baseline being trained on the MGBB dataset and application of the genetic models, trained on the TOPMed data, incorporating one or more PRSs according to the three model complexity levels. Level 1: a single PRS based on the UKBB + ICBP GWAS. Level 2: seven PRSs based on the UKBB + ICBP GWAS, difference p-value thresholds. Level 3: 21 PRSs, 7 PRSs based on each of the UKBB + ICBP, MVP, and BBJ GWAS. (**b**) Estimated PVE in the MGBB test dataset for XGBoost genetic models fitted on the TOPMed dataset of three levels of complexity with baseline model fitted using TOPMed baseline model weights (top) and using MGBB baseline model weights (bottom). PVEs are shown for the performance in prediction of the second order of residuals for SBP and DBP phenotypes in the overall test dataset and stratified by race/ethnicity (White N = 7985, Black N = 412, Asian N = 200). *BP* blood pressure, *MGBB* Mass General Brigham Biobank, *PRS* polygenic risk score, *TOPMed* trans-omics in precision medicine project.

provides the PVEs from both the TOPMed-trained genetic model applied on MGBB residuals (as in Fig. 5), and from the full calibrated ensemble (MGBB baseline model + TOPMed-trained genetic model).

Discussion

The objectives of this study were (a) to develop a non-linear ML-based prediction model for BP phenotypes that can more accurately predict BP phenotypes compared to standard linear regression model that includes a single genome-wide PRS model, (b) assess the usefulness of including multiple PRSs in the association model as an alternative to the inclusion of individual SNPs, as the latter option risks overfitting, and (c) examine an approach for calibrating a non-linear ML model to a new dataset. To accomplish these goals, we constructed an ensemble model, successively training first on a set of commonly available covariates (demographic variables and BMI), and then on genetic components (SBP/DBP PRSs), to directly evaluate the usefulness of a non-linear ML compared with a linear regression-based models at both steps. We further developed and compared polygenic models that employ a few levels of PRS inclusion: a single genome-wide PRS constructed based on a single powerful GWAS, multiple genome-wide PRSs constructed based on the same GWAS, and multiple genome-wide PRSs from multiple independent GWASs. We examined the improvement that such models confer to BP models in groups defined by self-reported race/ethnicity. We also explored the construction of genetic models that use local PRSs, constructed based on LD-regions, and, finally, studied the potential to apply the constructed models, and to calibrate model, to an Electronic Health Records (EHR)-based dataset using the MGB Biobank dataset.

We quantified model performance using PVEs computed on both the overall, phenotypic level, and at the level of the residuals from the baseline (non-genetic) model. Consistently with other studies, higher PVE was achieved for SBP compared with DBP. For both phenotypes, it was clearly beneficial to use a non-linear ML model at the baseline level (consistent with other studies[27]), while a non-linear ML genetic model performed better than a linear regression genetic model primarily in self-reported White individuals. Regardless of type of model, using multiple PRSs, including those constructed based on the same GWAS using different p-value thresholds, improved model PVE especially in non-White individuals. In self-reported White individuals using multiple PRSs had relatively little improvement in test dataset PVE. These findings suggest that (a) well-powered GWAS from European-ancestry populations is sufficient to construct good PRS in White individuals (who are primarily of European genetic ancestries); this may be true for individuals of other genetic ancestries, yet data is not yet available to prove it, and (b) GWAS from non-European ancestry populations are useful for PRS in non-White individuals, as is known, and (c) PRSs based on European ancestry GWAS are indeed useful in non-White individuals, and (d) flexibly allowing for potentially varying association effects of PRSs, each based on different p-value thresholds, is useful. Note that the usefulness of incorporating multiple PRSs based on the same GWAS is immediately apparent. On the face of it, one strong PRS based on a high p-value threshold should be sufficient. The usefulness of multiple PRSs based on the same GWAS resonates with the variability in the contribution of different genetic variants, beyond what is captured in GWAS. We hypothesized the local PRSs may capture this variability better, presumably because different genomic regions have potentially different associations with BP in different populations due to interactions with environment, lifestyle, and other factors. Yet, the local PRS model was less generally successful than global PRS models (with some exceptions, e.g., in Hispanic/Latino individuals), likely due to overfitting to the training dataset. In future work we will assess models using local PRSs in studies with larger sample sizes, which will potentially alleviate the overfitting problem. We also plan to study different potential constructions of local PRSs, in terms of both PRS derivation method, and definition of "locality". While we here used LD-regions as the local information, PRSs based on specific pathways, which have been recently studied[28–30], can also be viewed as local PRSs. Thus, both the definition of "local" and the method to construct local PRSs should be assessed.

We compared baseline models that included and not included genetic PCs, and ruled out using them in the model, as PCs did not improve model performance. This may not always be the case: other published work suggested that, for some phenotypes, using PCs improves genetic prediction, with and without inclusion of PRSs in the prediction models[31,32]. Therefore, it is important to keep evaluating the potential use of PCs in prediction models.

The TOPMed dataset that we used is diverse in terms of both race/ethnicity and genetic ancestry composition. It is important to put this work in the context of many publications about the limited generalizability of PRSs that were developed primarily (or only) in population of European ancestries to populations of other ancestries[33,34]. Here, we studied the accuracy of BP prediction models via the "complexity levels" of the genetic models. First, level 1, used only a single PRS based on a large dataset of European ancestries (UKBB + ICBP). As shown in Supplementary Table 8, the optimal p-value threshold when using a single PRS differed between self-reported race/ethnic group, due to either genetic ancestry effects, covariate distributions, or both. The addition of multiple PRSs constructed based on the same and from different GWAS, each one as a separate variable, allowed the non-linear ML model to utilize the different PRSs in a way that increase prediction accuracy across groups, but without making an explicit choice of a single PRS for any group. This approach is different than models that attempt to construct a better, single, PRS by leveraging information across genetic ancestries or by prioritizing variants[35–39]. While both approaches are useful, our proposed multi-PRS approach within a non-linear ML model addresses the issue of interactions—potential differences in impact of PRS in different genetic ancestries or by different level of covariates, and further, by genomic region. It is also important to note that we compared the non-linear ML models to those using ancestry-specific PRSs developed using PRS-CSx, which were then summed to generate a single PRS (which is a standard application of PRS-CSx). However, these models performed less well. This could due in part to our summation of the PRSs: we used an unweighted sum, while an ideal implementation would use another independent dataset to compute PRS summation weights. The goals of improving PRS generalizability and of accounting for interactions in prediction models are both important and more work is needed for incorporating both purposes in the same framework.

It is important to highlight the differences in results when considering genetic (residual prediction, after accounting for non-genetic covariates) versus ensemble (BP phenotype prediction) model performance: the highest PVE on the phenotypic level was achieved in Black individuals. On the other hand, at the genetic level, the highest PVE was reached in White individuals, and lowest in Black individuals. We interpret this as Black participants included in the TOPMed program having the most diverse distribution in some risk factors for BP, such that the variance was well explained by the baseline covariates. This is consistent with results we recently published, showing that BP PRS performance in All of Us dataset varied by strata of BP risk factors[13]. This further underscores the importance of incorporating multiple risk factors when developing genetic BP prediction models.

Calibration of prediction models to new populations is a well-recognized problem[40,41]. For example, in the context of atherosclerotic cardiovascular disease, many publications suggested specific ways to recalibrate the pooled cohort equations for specific or modern populations[42,43]. Including, recent literature studied the potential addition of a PRS for coronary artery disease to the pooled cohort equations, with model recalibration (for example[44,45]). In the context of non-linear ML models in the medical literature, model calibration is offered to address model "drift", where patient characteristics or prevalence/incidence of outcomes change over time[46]. Here, we attempted a new way to calibrate the models developed in TOPMed to the MGB Biobank dataset by a full refitting of the baseline model (only covariates), but with no update to the genetic model. Indeed, the (TOPMed-trained) genetic model performed better on the MGB Biobank dataset when the baseline model was fit on an MGB-trained baseline model. Still, more comprehensive work is needed to evaluate this and other calibration approaches in the context of genetic models, especially in the context of interactions. An important limitation of the assessment of prediction model over BP phenotypes is that BP has circadian rhythm[47], and while typically BP is measured in cohort studies early in the day, in the hospital settings it is measures at the time of the patient visit, adding noise to the MGB dataset. In our work, TOPMed-trained genetic models' performance was lower in MGB Biobank than in TOPMed, consistent with previous results studying hypertension PRS[48]. Surprisingly, models' performance in the subgroup of self-reported Asian individuals were high and similar to the performance in the TOPMed test dataset. However, as the number of Asian participants is low, the confidence intervals of the PVE estimates in this group are wide, limiting conclusions.

One of the strengths of this work is the use of primary, TOPMed dataset, which is comprised of large, racial/ethnically diverse and prospectively collected data used for both training and testing of the models. The dataset is of high-quality deep sequencing, joint allele-calling and the phenotypes were harmonized across studies. We used independent training and testing datasets, ensuring the reliability of model validation results. This study has some limitations as described above, including, limited sample size relative to the number of features when fitting a genetic model using local PRSs, MGB Biobank dataset is in general a "noisier" dataset, as it relies on data from health care visits and thus suffers from limitation of such datasets (BP measures may not follow best standards, may be measured using sick rather than healthy visits, etc.). Another limitation of this study is the development of PRSs based on the UKB + ICBP GWAS that partly overlapped with White TOPMed participants. We estimated the contribution of these participants to the BP GWASs by performing a GWAS using these individuals in TOPMed, and then applied a new algorithm to eliminate the contribution of this TOPMed White individuals-specific GWAS from the UKBB + ICBP summary statistics. While mathematically our algorithm is accurate, the contribution may not be entirely eliminated as the relevant summary statistics used by the UKBB + ICBP meta-analysis may have been based on a slightly different set of individuals (e.g., if not all participants from, say, Framingham Heart Study, are in this TOPMed data freeze). Thus, there may be some low levels of effects due to genetic relatedness between TOPMed participants and other participants in the UKBB + ICBP GWAS that were not included in the TOPMed White participants-specific GWAS. However, we think that this effect is likely very small. Finally, it should be noted that the Level 3 models included 21 PRSs, which are correlated to some extent, some highly correlated (based on the same GWAS but different thresholds). While this should not impact the predictions themselves, it would impact application of potential interpretation approaches for these models, including effect estimates and feature importance analyses[49].

In summary, we constructed and evaluated non-linear genetic association models with SBP and DBP, composed of sequentially-trained ensembles of a baseline and a genetic model. We showed that using multiple PRSs for the same trait based on the same GWAS improves the genetic model, and further including multiple PRSs based on the same trait based on multiple GWAS further improves the model. These improvements are mostly in non-White, i.e., self-reported Black, Asian, and Hispanic, populations. We also proposed a new way to leverage ensemble dataset to calibrate a model to a new dataset: by refitting one component of the model and using the other component as it was previously trained. Our results point to the promising potential of non-linear ML to combine traditional epidemiological risk factors for hypertension with genetic score for BP prediction.

Methods

We used two datasets. The primary dataset is from the Trans-Omics in Precision Medicine (TOPMed) consortium, which was used to train and evaluate multiple models. The second dataset is from the Mass General Brigham (MGB) Biobank and was used to evaluate selected model performance in an independent healthcare-system dataset. Below we describe the datasets and the steps for constructing and evaluating models. Figure 1 describes the framework for the development of multi-PRS non-linear ML model allowing for non-genetic components to be calibrated across datasets.

The TOPMed dataset

The TOPMed study population included 62,295 unrelated (3rd degree or less) participants from 15 studies based in the U.S. and in Taiwan. Data was extracted from the freeze 8 TOPMed dataset release. Information about the studies including ethics statements is provided in Supplementary Note 1. Blood pressure phenotypes were harmonized by the TOPMed Data Coordinating Center (DCC)[50] and included systolic blood pressure (SBP), diastolic blood pressure (DBP), and status of antihypertensive medication use (HTNMED_V1). Medication status was used to increase values of SBP and DBP by 15 and 10mmHg, respectively, to account for the expectations that their values would be higher if the corresponding individuals did not use antihypertensive medications.

Genotype data

We used whole genome sequencing (WGS) data from Trans-Omics for Prevision Medicine (TOPMed) program[51] Freeze 8 dataset. Genome samples were sequenced through TOPMed and the National Human Genome Research

Institute's Centers for Common Disease Genomics (CCDG) program and harmonized together via joint allele calling. The methods for TOPMed WGS data acquisition and quality control (QC) are described in https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8. Genetic Principal Components (PCs) and kinship coefficients were computed for the genetic data by the TOPMed DCC using the PC-Relate and PR-AiR algorithms[52,53] implemented in the GENESIS R package[54]. Based on the kinship coefficients, we identified related individuals and generated a dataset in which all individuals were degree-3 unrelated, i.e., all kinship coefficients were $< 2^{(-9/2)}$ (approximately 0.04). We extracted allele counts of variants that passed TOPMed quality control flags from GDS files using the SeqArray package version 1.28.1 and then further processed the genetic data using custom R version 3.6.2 and Python version 3.6.15 scripts. Only variants with minor allele frequency ≥ 0.01 in the TOPMed dataset and that passed TOPMed QC were used in this study. There were 63,093 participants and 12,341,303 variants available for PRS development.

Blood pressure phenotypes

We trained non-linear ML models to predict two phenotypes: systolic blood pressure (SBP) and diastolic blood pressure (DBP). SBP and DBP values were extracted, when available, from a harmonized datasets created by TOPMed[50], and for some TOPMed-parent studies they were prepared by study researchers. To address the existence of unrealistic values in the dataset, but without forcing a specific threshold, we removed individuals with outlying values defined by phenotypic values above the 99th quantile and values below the 1st quantile for each of the phenotypes. This amounted to 1047 and 1403 individuals from the analyses of SBP and DBP, respectively. The quantiles were computed over the complete dataset. Values for SBP and DBP for individuals that were taking antihypertensive medications were adjusted by increasing SBP and DBP values by 15 and 10 mmHg, respectively.

Summary statistics from published GWAS

We used summary statistics from published GWAS of SBP and DBP provided by BioBank Japan Project (BBJ), Million Veteran Program (MVP), as well as UK BioBank and the International Consortium of Blood-Pressure Genome Wide Associations Studies (UKBB + ICBP). Details are provided in Supplementary Table 12. Because some of TOPMed European ancestry participants also participated in the UKBB + ICBP meta-analysis, we performed GWAS of SBP and DBP using all self-reported White individuals in the TOPMed dataset (i.e. participants included in both TOPMed training and testing datasets), and then applied a procedure to "remove" the contribution of this GWAS from the overall UKBB + ICBP GWAS summary statistics[55] as described in Supplementary Note 2.

Polygenic risk scores

We developed two types of PRS—"global PRS", using SNPs from the entire genome, and, in secondary analysis, "local PRS", calculated from SNPs within LD-regions. In both cases SNPs were selected using the clump & threshold approach. We developed and constructed global PRSs using PRSice2 version 2.3.5[56], from the BP GWAS summary statistics described above. As tuning parameters, we set $R^2 = 0.1$, distance = 1000 kB, and several p-value thresholds: 5×10^{-8}, 10^{-7}, 10^{-6},…, 10^{-2}. We used the TOPMed data set as a reference panel for LD (used for clumping). In secondary analysis, we developed local PRSs. We used previously computed LD-regions[57] provided in BED files defining chromosomal segments (see Data Availability) based on a European reference panel to subset the UKBB + ICBP GWAS summary statistics to files consisting of variants falling within each LD-region (chromosomal segment). We developed local PRS based for each of these regions using the same clump & threshold approach using PRSice2, but now, due to the large number of segments and thus features, we used a single p-value threshold of 10^{-2}. For this secondary analysis we only used the UKBB + ICBP GWAS because we saw before that, although it is based on single genetic ancestry (European) PRS based on it perform well when evaluated in various self-reported race/ethnic groups[13]. In a Supplementary Note 3 we also describe the comparison of the models' performance using global SBP and DBP developed using PRS-CSx[58], which is an extension of the Bayesian polygenic prediction method PRS-CS[59], from the three GWAS summary statistics used in this study.

Non-linear ML model training and hyperparameter tuning

We used the python version 3.6.15 library xgboost version 1.5.2[60] to fit ensembles of gradient boosted trees. We performed hyperparameter tuning using Optuna[61] library version 3.0.6. Specifically, we split the training dataset at random into 5 independent datasets and performed a fivefold cross-validation procedure to select optimal values of relevant tuning parameters.

Ensemble model

As described in Fig. 1, the ensemble non-linear ML model consisted of two consecutive components—"baseline model" and "genetic model". The goal behind the two-model construction was (1) separately assess the benefit of non-linear modelling of non-genetic measures and genetic measures, (2) allow for flexible combination of the two models (e.g., linear model for covariates and non-linear model for PRSs, or the other way around), and (3) facilitate model calibration and generalizability between datasets while acknowledging that some covariates are potentially dataset-specific, and these are included only in the baseline model. The genetic model is expected to be fully transferable to a new dataset by using features that are comparable (or harmonized) across datasets.

We compared two potential constructions of both the baseline and the genetic models: non-linear ML (allowing for data-driven incorporation of interactions), and linear regression (without modeling interactions). We divided the dataset such that 70% of the data was used as training data and 30% of the data was held out as a validation set. First, we trained the baseline model using covariates only, including age, sex, self-reported race/

ethnicity, BMI, and study. The genetic model was trained on the same set of features as in the baseline model, other than study, which is dataset-specific, and it also included genetic components, i.e., PRSs, where we used the global PRSs described above, the local PRSs in secondary analysis (described later). The genetic model was trained to predict the residuals from the baseline model.

Model development using multiple PRSs

In primary analysis, we studied the use of multiple PRSs in the genetic model via multiple models of increased complexity (in the sense that they include higher number of PRSs). We refer to these models as Level 1, Level 2 and Level 3 with Level X being shorthand for Model complexity levels (Fig. 1c). Note that all genetic models included the same non-genetic covariates as described above, and they only differed by the inclusion of additional PRSs. Level 1 of the genetic model included a single PRS developed from the UKBB + ICBP GWAS summary statistics using the p-value threshold 10^{-2}. Level 2 included 7 PRSs, all from the UKBB + ICBP GWAS, and based on all considered p-value threshold: 5×10^{-8}, 10^{-7}, 10^{-6},…, 10^{-2}. Level 3 included PRSs constructed based on all considered GWAS (UKBB + ICBP, MVP, and BBJ) GWASs and using all p-value thresholds, i.e., it included 21 PRSs. Level 1 models included only a single PRS based on the 10^{-2} p-value threshold because we expected this PRS to have the best performance based on past work studying BP PRSs[13]. In secondary analysis, we also evaluated model performance, based on the training dataset only, when using each of the other UKBB + ICBP based PRSs (i.e. based on each of the 7 thresholds). The models were fit using the combined, multi-ethnic dataset, with training dataset cross-validation performance reported on the combined group and stratified by self-reported race/ethnicity. The goal was to assess whether the same or different thresholds are useful across groups, and thus whether multi-PRS models are useful because they potentially allow for different utilization of PRSs across individuals (i.e. due to interactions).

Secondary analysis using local PRSs

We considered using local PRSs instead of global PRSs because they may result in more interpretable models, i.e., where one could hopefully explain why different genomic regions may have different potential contributions to the BP model. Due to the large number of PRSs when computed over all LD-regions, potentially resulting in model overfitting, we augmented the ensemble model approach with a variable selection step. Here, we applied LASSO regression from the python library scikit-learn[62] version 0.24.2 using, as before, cross-validation for tuning parameter selection, on a linear model predicting the residuals from the baseline model using all constructed local PRSs. We next use the selected PRSs in the genetic model (see Fig. 4a for visualization of the Ensemble model with integrated LASSO step). We evaluated this model performance on the test dataset, and also, for comparison, of the model that uses only the LASSO (without the following non-linear XGBoost ML genetic model).

Assessment of computational needs

We assessed computational runtime and memory (RAM) usage of the non-linear ML models in comparison with the linear regression models using python *timeit* and *memit* modules. These were measured on a 13-inch 2020 MacBook Pro machine, with the Apple M1-chip, macOS Sonoma version 14.4.1, 16 Gb of RAM and 8 vCPUs. As the non-linear ML models require tuning parameters fitting, we measured their runtime and memory over the fivefold cross validation process. We also measured the runtime and memory for applying the various models for prediction over the test dataset. We performed each computing task 10 times and provide the average runtime and RAM measures, other than for local PRS models, which took longer time, and we therefore fit them once for this assessment. We performed this analysis only using SBP models, that used a slightly larger sample size compared to DBP.

Model performance evaluation

Models' performance was assessed on the test dataset (30% of the TOPMed dataset of unrelated individuals). We report two performance measures: percent variance explained (PVE) at both the residual (of the baseline model) and at the phenotypic (original BP phenotypes) level. To explain how each is computed we introduce some notation. Including, we distinguish between the predicted values of the baseline model, the genetic model, and the combined ensemble, and between the residuals of the baseline and the genetic models.

Let M_b and M_g denote a baseline and a genetic model, $\boldsymbol{x_b}$ and $\boldsymbol{x_g}$ the sets of covariates used by M_b, and M_g respectively, and $\boldsymbol{g}$ the set of PRSs used by M_g. Let y denote a BP outcome. A trained model M_b uses $\boldsymbol{x_b}$ to predict y, as:

$$M_b(\boldsymbol{x_b}) = \widehat{y}_b. \tag{1}$$

With $\widehat{y}_b$ being the prediction of M_b applied on $\boldsymbol{x_b}$. The residuals of model M_b are obtained as the difference between the observed and the predicted BP value, and are denoted by r_b:

$$r_b = y - \widehat{y}_b. \tag{2}$$

The genetic model is trained to predict r_b using $\boldsymbol{x_g}$ and PRSs $\boldsymbol{g}$. Thus

$$M_g\left(\boldsymbol{x_g}, \boldsymbol{g}\right) = \widehat{r}_b. \tag{3}$$

The residuals of M_g are given as the difference between the value it attempts to predict and its prediction:

$$r_g = r_b - \widehat{r}_b. \tag{4}$$

Noting based on Eqs. (2) and (4) that the observed BP measure y can be decomposed as:

$$y = \widehat{y}_b + r_b = \widehat{y}_b + \widehat{r}_b + r_g. \quad (5)$$

We denote the prediction BP based on the ensemble model M_e by $\widehat{y}_{ensemble}$ with:

$$\widehat{y}_{ensemble} = \widehat{y}_b + \widehat{r}_b. \quad (6)$$

With the residuals r_g of the genetic models being also the residuals of the ensemble model. For a given outcome out the PVE by the predicted outcome $\widehat{out}$ is defined as the percent reduction in the variance of out when accounting for $\widehat{out}$, using this formula:

$$PVE\left(out, \widehat{out}\right) = \left(1 - \frac{var(out) - var\left(\widehat{out}\right)}{var(out)}\right) \times 100\%. \quad (7)$$

Finally, we define the performance measures of the various models. To assess the performance of baseline models we compute phenotyping PVE, $PVE\left(y, \widehat{y}_b\right)$. To assess the performance of the genetic models we compute the PVE at the level of the baseline model residuals, i.e., $PVE(r_b, \widehat{r}_b)$. To assess the performance of the ensemble model we compute again PVE at the phenotypic level, i.e., $PVE\left(y, \widehat{y}_{ensemble}\right)$.

To further interpret results, we computed 95% confidence intervals for the estimated PVEs using bootstrap sampling from the test dataset, with 100 repetitions. Specifically, we sampled with replacement individuals from the test dataset using the same sample size of the test dataset relevant for each analysis, and applied the prediction models that were trained over the train dataset. When assessing genetic model, we applied both the baseline and the genetic model over the sampled individuals. The PVE was computed for each bootstrap sample, and the 95% confidence interval was derived using the percentile method, i.e. using the 2.5 and 97.5 percentiles of the bootstrap distributions.

The Mass General Brigham (MGB) Biobank dataset

To assess how the patterns observed in the TOPMed dataset generalize to a healthcare-based medical system, we implemented all fitted models on the MGB Biobank dataset (MGB Biobank). In MGB Biobank, phenotypes are available from electronic health records (EHR). Because blood pressure, BMI, and medication data are available sporadically depending on patient visits, we restricted the datasets health records from two years, 5/25/2021 to 5/25/2023, so that all time-varying variable correspond to approximately the same age and time. The initial MGB Biobank dataset included data for 142,476 individuals. Of these, N = 108,389 individuals did not take antihypertensive medications (dataset codes "antihypertensive medications", "antihypertensive-other", "beta blockers/related", "calcium channel blockers", "diuretics", "direct renin inhibitor", "antihypertensive combinations"). Next, we further filtered the dataset to only include individuals with available systolic or diastolic reading (N = 29,282). We then only included individuals with genetic data, who also are genetically unrelated to each other, resulting in 9494 individuals in the final dataset. For each individual, we extracted the median SBP, DBP, and BMI values from these years. Individuals self-identified with categories of race and ethnicity. Individuals with Hispanic ethnicity were set to the "Hispanic" category, and otherwise individuals with non-Hispanic ethnicity and with Black or African American race to Black, and those with non-Hispanic White or Asian race were set to White or Asian, respectively, and non-Hispanic individuals with more than one self-identified race or with "other" or "unknown" were set to "other". More details about the MGB Biobank dataset are provided in Supplementary Note 4.

To calibrate the ensemble model to the MGB Biobank dataset, we trained the baseline model on the set of covariates from MGB Biobank for prediction of the phenotype (SBP/DBP) and calculated the residuals. Next, we evaluated, separately, the genetic model (trained on the TOPMed training dataset) and the ensemble (MGB Biobank-trained baseline model + TOPMed trained genetic model) models. The genetic models included the three model complexity levels trained on the TOPMed dataset. To assess this calibration approach, we assessed the performance of the TOPMed-trained genetic model when applied over residuals from the MGB-trained and residuals from the TOPMed-trained baseline models.

Ethics statement

All methods were carried out in accordance with relevant guidelines and regulations. The presented analysis relies on observational data only, the experimental protocol is purely computational. This work was approved by the Mass General Brigham IRB (protocol #2021P001928) and by the Beth Israel Deaconess Medical Center Committee on Clinical Investigators (protocol #2023P000541). Informed consent was obtained from all participants, as described in Supplementary Note 1 (TOPMed participants) and Supplementary Note 4 (MGB Biobank participants).

Data availability

TOPMed freeze 8 WGS data and harmonized BP phenotypes are available by application to dbGaP according to the study specific accessions: Amish: "phs000956" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000956.v1.p1), ARIC: "phs001211" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001211.v4.p3), BioMe: "phs001644" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001644.v2.p2), CARDIA: "phs001612" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001612.v1.p1), CFS: "phs000954" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000954.v4.p2), CHS: "phs001368" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001368.v3.p2), COPDGene: "phs000951" (https://www.ncbi.nlm.nih.gov/projects/gap/

cgi-bin/study.cgi?study_id=phs000951.v5.p5), FHS: "phs000974" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000974.v4.p3), GENOA: "phs001345" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001345.v3.p1), GenSalt: "phs001217" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001217.v3.p1), HCHS/SOL: "phs001395" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001395.v2.p1), JHS: "phs000964" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000964.v5.p1), MESA: "phs001211" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001211.v4.p3), THRV: "phs001387" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001387.v3.p1), WHI: "phs001237" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001237.v3.p1). Summary statistics from MVP BP GWAS are available from dbGaP by application to study accession "phs001672" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001672.v4.p1). The summary statistics from the UKBB + ICBP BP GWAS are available at https://grasp.nhlbi.nih.gov/FullResults.aspx. MGB Biobank genotyping and phenotypic data are available to Mass General Brigham investigators with required approval from the Mass General Brigham Institutional Review board (IRB). Data needed to construct the selected BP PRSs generated in this study will become publicly available on a Zenodo repository upon paper acceptance and will include variants, alleles, and weights for each of the PRS based on GWAS of SBP and DBP. The BED files that define LD-regions used for construction of local PRSs are available under the Bitbucket repository in https://bitbucket.org/nygcresearch/ldetect-data/src/master/.

Received: 22 January 2024; Accepted: 22 May 2024
Published online: 30 May 2024

References

1. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**(9), 581–590 (2018).
2. Choi, S. W., Mak, T. S. & O'Reilly, P. F. Tutorial: A guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**(9), 2759–2772 (2020).
3. Ho, D. S. W. *et al.* Machine learning SNP based prediction for precision medicine. *Front. Genet.* **10**, 1 (2019).
4. Elgart, M. *et al.* Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Commun. Biol.* **5**(1), 856 (2022).
5. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **58**(1), 267–288 (1996).
6. Pain, O. *et al.* Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genet.* **17**(5), e1009021 (2021).
7. Krapohl, E. *et al.* Multi-polygenic score approach to trait prediction. *Mol. Psychiatry* **23**(5), 1368–1374 (2018).
8. Schoeler, T. *et al.* Multi-polygenic score approach to identifying individual vulnerabilities associated with the risk of exposure to bullying. *JAMA Psychiatry* **76**(7), 730–738 (2019).
9. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53**(2), 185–194 (2021).
10. Abraham, G. *et al.* Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat. Commun.* **10**(1), 5819 (2019).
11. Rodriguez, V. *et al.* Use of multiple polygenic risk scores for distinguishing schizophrenia-spectrum disorder and affective psychosis categories in a first-episode sample; the EU-GEI study. *Psychol. Med.* **53**(8), 3396–3405 (2023).
12. Meisner, A. *et al.* Combined utility of 25 disease and risk factor polygenic risk scores for stratifying risk of all-cause mortality. *Am. J. Hum. Genet.* **107**(3), 418–431 (2020).
13. Kurniansyah, N. *et al.* Evaluating the use of blood pressure polygenic risk scores across race/ethnic background groups. *Nat. Commun.* **14**(1), 3202 (2023).
14. Coombes, B. J. *et al.* Dissecting clinical heterogeneity of bipolar disorder using multiple polygenic risk scores. *Transl. Psychiatry* **10**(1), 314 (2020).
15. Xin, J. *et al.* Risk assessment for colorectal cancer via polygenic risk score and lifestyle exposure: A large-scale association study of East Asian and European populations. *Genome Med.* **15**(1), 4 (2023).
16. Collister, J. A., Liu, X. & Clifton, L. Calculating polygenic risk scores (PRS) in UK Biobank: A practical guide for epidemiologists. *Front. Genet.* **13**, 818574 (2022).
17. Coombes, B. J. *et al.* A principal component approach to improve association testing with polygenic risk scores. *Genet. Epidemiol.* **44**(7), 676–686 (2020).
18. Arvanitis, M. *et al.* Linear and nonlinear Mendelian randomization analyses of the association between diastolic blood pressure and cardiovascular events: The J-curve revisited. *Circulation* **143**(9), 895–906 (2021).
19. Wan, E. Y. F. *et al.* Blood pressure and risk of cardiovascular disease in UK Biobank: A Mendelian randomization study. *Hypertension* **77**(2), 367–375 (2021).
20. Tsao, C. W. *et al.* Heart disease and stroke statistics-2023 update: A report from the American Heart Association. *Circulation* **147**(8), e93–e621 (2023).
21. Mills, K. T. *et al.* Global disparities of hypertension prevalence and control: A systematic analysis of population-based studies from 90 countries. *Circulation* **134**(6), 441–450 (2016).
22. Jaeger, B. C. *et al.* Hypertension statistics for US adults: An open-source web application for analysis and visualization of national health and nutrition examination survey data. *Hypertension* **80**(6), 1311–1320 (2023).
23. Ference, B. A. *et al.* Clinical effect of naturally random allocation to lower systolic blood pressure beginning before the development of hypertension. *Hypertension* **63**(6), 1182–1188 (2014).
24. Niiranen, T. J. *et al.* Prediction of blood pressure and blood pressure change with a genetic risk score. *J. Clin. Hypertens.* **18**(3), 181–186 (2016).
25. Fujii, R. *et al.* Associations of genome-wide polygenic risk score and risk factors with hypertension in a Japanese population. *Circ. Genom. Precis. Med.* **15**(4), e003612 (2022).
26. Grinde, K. E. *et al.* Generalizing polygenic risk scores from Europeans to Hispanics/Latinos. *Genet. Epidemiol.* **43**(1), 50–62 (2019).
27. McCaw, Z. R. *et al.* DeepNull models non-linear covariate effects to improve phenotypic prediction and association power. *Nat. Commun.* **13**(1), 241 (2022).
28. Goodman, M. O. *et al.* Pathway-specific polygenic risk scores identify obstructive sleep apnea—Related pathways differentially moderating genetic susceptibility to coronary artery disease. *Circ. Genom. Precis. Med.* **15**(5), e003535 (2022).
29. Choi, S. W. *et al.* PRSet: Pathway-based polygenic risk score analyses and software. *PLoS Genet.* **19**(2), e1010624 (2023).
30. Darst, B. F. *et al.* Pathway-specific polygenic risk scores as predictors of amyloid-β deposition and cognitive function in a sample at increased risk for Alzheimer's disease. *J. Alzheimers Dis.* **55**(2), 473–484 (2017).

31. Naret, O. *et al.* Improving polygenic prediction with genetically inferred ancestry. *HGG Adv.* **3**(3), 100109 (2022).
32. Chen, C. Y. *et al.* Explicit modeling of ancestry improves polygenic risk scores and BLUP prediction. *Genet. Epidemiol.* **39**(6), 427–438 (2015).
33. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* **109**(2), 373 (2022).
34. Wang, Y. *et al.* Challenges and opportunities for developing more generalizable polygenic risk scores. *Annu. Rev. Biomed. Data Sci.* **5**, 293–320 (2022).
35. Zhao, Z. *et al.* The construction of cross-population polygenic risk scores using transfer learning. *Am. J. Hum. Genet.* **109**(11), 1998–2008 (2022).
36. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**(5), 573–580 (2022).
37. Hoggart, C. J. *et al.* BridgePRS leverages shared genetic effects across ancestries to increase polygenic risk score portability. *Nat. Genet.* **56**(1), 180–186 (2024).
38. Hu, X. *et al.* Polygenic transcriptome risk scores for COPD and lung function improve cross-ethnic portability of prediction in the NHLBI TOPMed program. *Am. J. Hum. Genet.* **109**(5), 857–870 (2022).
39. Weissbrod, O. *et al.* Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**(4), 450–458 (2022).
40. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**(29), 1925–1931 (2014).
41. Van Calster, B. *et al.* Calibration: The Achilles heel of predictive analytics. *BMC Med.* **17**(1), 230 (2019).
42. Cook, N. R. & Ridker, P. M. Calibration of the pooled cohort equations for atherosclerotic cardiovascular disease. *Ann. Intern. Med.* **165**(11), 786–794 (2016).
43. Emdin, C. A. *et al.* Evaluation of the pooled cohort equations for prediction of cardiovascular risk in a contemporary prospective cohort. *Am. J. Cardiol.* **119**(6), 881–885 (2017).
44. Khan, S. S. *et al.* Coronary artery calcium score and polygenic risk score for the prediction of coronary heart disease events. *JAMA* **329**(20), 1768–1777 (2023).
45. Mujwara, D. *et al.* Integrating a polygenic risk score for coronary artery disease as a risk-enhancing factor in the pooled cohort equation: A cost-effectiveness analysis study. *J. Am. Heart Assoc.* **11**(12), e025236 (2022).
46. Davis, S. E. *et al.* Calibration drift among regression and machine learning models for hospital mortality. *AMIA Annu. Symp. Proc.* **2017**, 625–634 (2017).
47. Zhang, J. *et al.* Circadian blood pressure rhythm in cardiovascular and renal health and disease. *Biomolecules* **11**, 6 (2021).
48. Kurniansyah, N. *et al.* A multi-ethnic polygenic risk score is associated with hypertension prevalence and progression throughout adulthood. *Nat. Commun.* **13**(1), 3549 (2022).
49. Toloşi, L. & Lengauer, T. Classification with correlated features: Unreliability of feature ranking and solutions. *Bioinformatics* **27**(14), 1986–1994 (2011).
50. Stilp, A. M. *et al.* A system for phenotype harmonization in the national heart, lung, and blood institute trans-omics for precision medicine (TOPMed) program. *Am. J. Epidemiol.* **190**(10), 1977–1992 (2021).
51. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**(7845), 290–299 (2021).
52. Conomos, M. P. *et al.* Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* **98**(1), 127–148 (2016).
53. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.* **39**(4), 276–293 (2015).
54. Gogarten, S. M. *et al.* Genetic association testing using the GENESIS R/bioconductor package. *Bioinformatics* **35**(24), 5346–5348 (2019).
55. Sofer, T. *tamartsi/Remove_overlap_GWAS_summary_stat: v1.0.0* (Zenodo, 2022).
56. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic risk score software. *Bioinformatics* **31**(9), 1466–1468 (2015).
57. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**(2), 283–285 (2016).
58. Ruan, Y. *et al.* Author Correction: Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**(8), 1259 (2022).
59. Ge, T. *et al.* Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**(1), 1776 (2019).
60. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016).
61. Akiba, T. *et al.* Optuna: A next-generation hyperparameter optimization framework. In *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2623–2631 (Association for Computing Machinery, 2019).
62. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

This study was performed as a collaboration of the NHLBI Trans-Omics in Precision Medicine (TOPMed) Consortium. We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed and CCDG. TOPMed and CCDG acknowledgements, as well as descriptions, acknowledgements, and ethics statements of contributing studies are provided in Supplementary Note 5. TOPMed consortium researchers and their affiliations are listed in Supplementary Note 6. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services. We thank Mass General Brigham Biobank for providing samples, genomic data, and health information data. We also thank the HPC support team of Enterprise Research Infrastructure & Services at Mass General Brigham for their support and for the provision of computational resources. This work was approved by the Mass General Brigham Institutional Review Board and by the Beth Israel Deaconess Medical Center Committee on Clinical Investigations. This study was supported by National Heart Lung and Blood Institute (NHLBI) Grant R01HL161012 to TS. MM was supported NHLBI Grant K08HL159318.

Author contributions

YH and TS drafted the manuscript. BS and ME developed PRSs and developed association models. YH and BS prepared figures and tables. NK, BS preprocessed and harmonized analytic datasets. GL advised the developed of ensemble models. TO, TK, MF, DL-J, JB, BMP, XG, WK, MM, PC, BCJ, EC, APC, SR, YDIC, CG, R-HC, CK, RC, BH, JS, SK, WZ, RJFL, BDM, RK, SSR, JIR, DL, ACM contributed to design and data curation in TOPMed studies they represent. BS, ME, NK, GL, TO, TK, MF, DL-J, JB, BMP, NS, XG, WK, MM, PC, BCJ, EC, APC, SR, YDIC, CG, R-HC, CK, RC, BH, JS, SK, WZ, RJFL, BDM, RK, SSR, JIR, DL, ACM critically reviewed the manuscript.

Competing interests

B Psaty serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. G Lyons is currently a full time employee of Alexion, AstraZeneca Rare Disease, and hold stock in the company, however, her contributions to the present manuscript were performed as part of her previous affiliation at the Harvard T.H. Chan School of Public Health and this work is not related to her current occupation and affiliation. M Moll has received grant funding from Bayer and consulting fees from TriNetX, 2ndMD, TheaHealth, Sitka, Verona Pharma, and Axon Advisors. All other authors report no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-62945-9.

Correspondence and requests for materials should be addressed to T.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The Trans-Omics in Precision Medicine Consortium

April P. Carson[6], Braxton D. Mitchell[9], Bruce M. Psaty[10,11,12,13], C. Charles Gu[15], Charles Kooperberg[16], Daniel Levy[17,18], Jennifer A. Brody[10,12], Jennifer A. Smith[21,22], Jerome I. Rotter[23], Matthew Moll[1,2,24,33], Myriam Fornage[5,25], Peter Castaldi[1,2], Ren-Hua Chung[27], Robert Kaplan[7,16], Ruth J. F. Loos[28,29], Sharon L. R. Kardia[21], Stephen S. Rich[30], Susan Redline[1,2,31], Timothy O'Connor[9,35,36], Wei Zhao[21,22], Wonji Kim[33], Xiuqing Guo[23], Yii-Der Ida Chen[23] & Tamar Sofer[1,2,3,4,34]✉