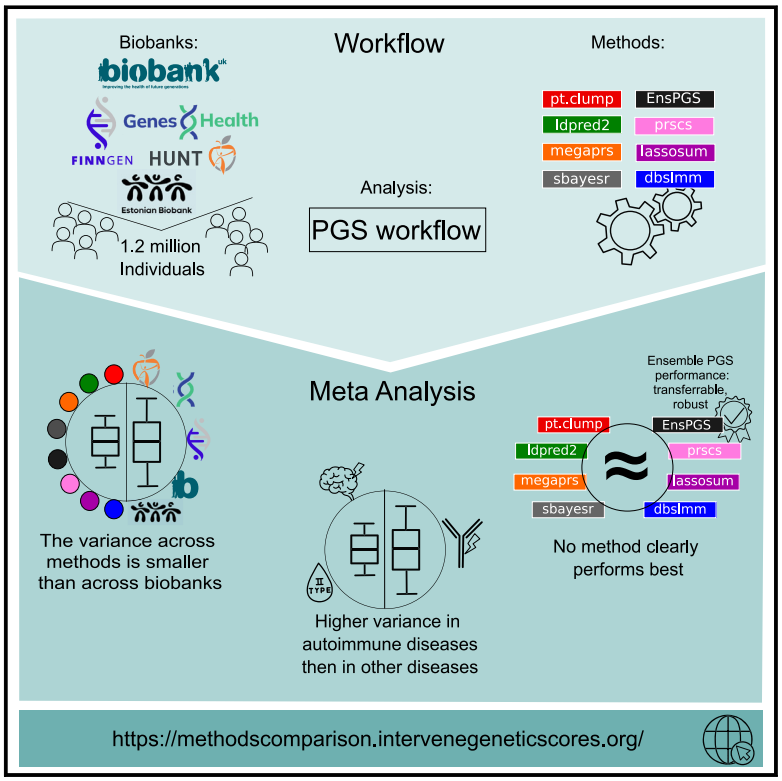
Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and finds benefits of ensemble learning

Monti, Remo; Eick, Lisa; Hudjashov, Georgi; Läll, Kristi; Kanoni, Stavroula ...

2024-06-21
Cell Press
http://hdl.handle.net/10138/589593
Monti, R, Eick, L, Hudjashov, G, Läll, K, Kanoni, S, Wolford, B N, Wingfield, B, Pain, O, Wharrie, S, Jermy, B, McMahon, A, Hartonen, T, Heyne, H, Mars, N, Lambert, S, Hveem, K, Inouye, M, van Heel, D A, Mägi, R, Marttinen, P, Ripatti, S, Ganna, A & Lippert, C 2024, 'Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and finds benefits of ensemble learning', American Journal of Human Genetics, vol. 111, no. 7, pp. 1431-1447. https://doi.org/10.1016/j.ajhg.2024.06.003

Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and finds benefits of ensemble learning

Graphical abstract



Authors

Remo Monti, Lisa Eick, Georgi Hudjashov, ..., Samuli Ripatti, Andrea Ganna, Christoph Lippert

Correspondence

christoph.lippert@hpi.de

Systematic evaluation of polygenic scoring methods in 1.2 million individuals across five biobanks finds that no single method performs best. Performance varied more between biobanks than between methods, suggesting that future research should address between-biobank variability. Ensembles provided high, robust, and transferable performance. Workflow and results browser are open source.

Monti et al., 2024, The American Journal of Human Genetics *111*, 1–17
July 11, 2024
https://doi.org/10.1016/j.ajhg.2024.06.003

Please cite this article in press as: Monti et al., Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and..., The American Journal of Human Genetics (2024), https://doi.org/10.1016/j.ajhg.2024.06.003

ARTICLE

Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and finds benefits of ensemble learning

Remo Monti,[1,2,25] Lisa Eick,[3,25] Georgi Hudjashov,[4] Kristi Läll,[4] Stavroula Kanoni,[5] Brooke N. Wolford,[6] Benjamin Wingfield,[7] Oliver Pain,[8] Sophie Wharrie,[9] Bradley Jermy,[3] Aoife McMahon,[7] Tuomo Hartonen,[3] Henrike Heyne,[1] Nina Mars,[3,10,11] Samuel Lambert,[12,14,15,17] Genes and Health Research Team, Kristian Hveem,[6,12] Michael Inouye,[13,14,15,16,17,18] David A. van Heel,[19] Reedik Mägi,[4] Pekka Marttinen,[9] Samuli Ripatti,[3,20,20] Andrea Ganna,[3,21] and Christoph Lippert[1,22,23,24,*]

Summary

Methods of estimating polygenic scores (PGSs) from genome-wide association studies are increasingly utilized. However, independent method evaluation is lacking, and method comparisons are often limited. Here, we evaluate polygenic scores derived via seven methods in five biobank studies (totaling about 1.2 million participants) across 16 diseases and quantitative traits, building on a reference-standardized framework. We conducted meta-analyses to quantify the effects of method choice, hyperparameter tuning, method ensembling, and the target biobank on PGS performance. We found that no single method consistently outperformed all others. PGS effect sizes were more variable between biobanks than between methods within biobanks when methods were well tuned. Differences between methods were largest for the two investigated autoimmune diseases, seropositive rheumatoid arthritis and type 1 diabetes. For most methods, cross-validation was more reliable for tuning hyperparameters than automatic tuning (without the use of target data). For a given target phenotype, elastic net models combining PGS across methods (ensemble PGS) tuned in the UK Biobank provided consistent, high, and cross-biobank transferable performance, increasing PGS effect sizes (β coefficients) by a median of 5.0% relative to LDpred2 and MegaPRS (the two best-performing single methods when tuned with cross-validation). Our interactively browsable online-results and open-source workflow prspipe provide a rich resource and reference for the analysis of polygenic scoring methods across biobanks.

Introduction

Polygenic scores (PGSs), also referred to as polygenic risk scores (PRSs), have become a major application of genome-wide association studies (GWASs). PGSs are constructed by scoring individuals on the basis of their genotype and thereby adding up the effects of many genetic variants genome-wide. They can improve existing disease risk models that rely on family history and established biomarkers,[1–4] and individuals in the upper tail of the PGS distribution have an elevated disease risk similar to that caused by rare damaging monogenic mutations for some diseases.[5] PGSs have received attention in areas ranging from disease prevention to clinical trials as a result of their wide applicability to personalized medicine.[6–9]

Various methods to derive PGS weights from GWAS summary statistics (effect sizes and their correlation structure) have been developed. These methods are of particular interest as they do not rely on access to individual-level data, which is typically restricted. Furthermore, the largest GWAS are meta-analyses, for which direct access to all individual-level source data is not feasible.

[1]Hasso Plattner Institute, University of Potsdam, Digital Engineering Faculty, Potsdam, Germany; [2]Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association, Berlin Institute for Medical Systems Biology, Berlin, Germany; [3]Institute for Molecular Medicine Finland, Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland; [4]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia; [5]William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK; [6]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Faculty of Medicine and Health, Norwegian University of Science and Technology, Trondheim, Norway; [7]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK; [8]Maurice Wohl Clinical Neuroscience Institute, Department of Basic and Clinical Neuroscience; Institute of Psychiatry, Psychology and Neuroscience; King's College London, London, UK; [9]Aalto University, Department of Computer Science, Espoo, Finland; [10]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; [11]Stanley Center for Psychiatric Research and Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; [12]Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway; [13]Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; [14]Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, VIC, Australia; [15]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; [16]Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge, UK; [17]British Heart Foundation Cambridge Centre of Research Excellence, School of Clinical Medicine, University of Cambridge, Cambridge, UK; [18]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK; [19]Blizard Institute, Queen Mary University of London, London, UK; [20]Department of Public Health, University of Helsinki, Helsinki, Finland; [21]Massachusetts General Hospital and Broad Institute of MIT and Harvard, Cambridge, MA, USA; [22]Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [23]Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [24]Department of Diagnostic, Molecular, and Interventional Radiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[25]These authors contributed equally
*Correspondence: christoph.lippert@hpi.de
https://doi.org/10.1016/j.ajhg.2024.06.003.

The construction of a PGS from summary statistics can be divided into two main stages: a public stage, which relies only on publicly available data and tools, and a private stage, which requires access to individual-level target data, i.e., genotypes and phenotypes. The public stage uses variant correlation (linkage disequilibrium; LD) from reference panels that are matched in ancestry to the GWAS sample to adjust the marginal-effect-size estimates of genetic variants and derive the per-variant PGS weights. These adjustments include frequentist shrinkage,[10] Bayesian approaches,[11–15] or other strategies such as thresholding, which depend on one or more hyperparameters (e.g., *p*-value thresholds, heritability estimates, or shrinkage parameters).

Many methods allow automatically setting suitable parameters without the use of phenotype data (we refer to this generally as automatic tuning). Alternatively, target data can be used to empirically determine hyperparameters on the basis of, for example, cross-validation (CV). The adjusted variant effect sizes (PGS weights) are used in the private stage for scoring individuals on the basis of their genotypes according to a linear additive model, i.e., for calculating their PGS.

Authors of PGS methods usually claim superior performance to other methods. However, comparisons are often limited to a small number of methods, traits, or target datasets. Furthermore, the input summary statistics used in those comparisons might not reflect the properties of (messy) real-world data, especially those from meta-analyses. In practice, other factors, such as ease of use and documentation, also affect performance. A few studies have compared a large number of PGS methods.[16–18] Yet, evaluation either only covered few traits in specialized cohorts[17] or was largely limited to within-biobank comparisons.[16,18]

The INTERVENE consortium[19] seeks to develop risk scoring methods that integrate PGS with other health-related information. For this reason, we compared summary-statistics-based PGS methods. Building on an updated version of the GenoPred suite that was originally introduced by Pain et al. and that implements different PGS methods in a reference standardized framework,[16] we developed prspipe, a snakemake[20] workflow that runs seven polygenic scoring methods. A full evaluation including hyperparameter tuning with CV was performed in the UK Biobank[21] (UKBB) and replicated in FinnGen,[22] Estonian Biobank[23] (EBB), the Trøndelag Health Study[24] (HUNT), and Genes & Health[25] (GNH). In total, we meta-analyzed performances for ten harmonized binary disease traits and six quantitative traits in two replicated ancestry groups, European (EUR) and South Asian (SAS). Replication in multiple biobanks allowed us to estimate how much PGS effect sizes vary within biobanks (between methods) and how this compares to the variation between biobanks.

We are publishing our workflow, summary data, and PGS weights, allowing others to replicate analyses, e.g., for methods comparisons or the development of new polygenic scores from summary statistics. The results of this analysis are made available in a browsable online resource (see data and code availability).

Methods

Participating studies

Data from five biobanks were considered: The UK Biobank,[21] FinnGen,[22] Estonian Biobank,[23] Trøndelag Health Study (HUNT)[24] and Genes & Health.[25] All biobanks independently performed genotyping, imputation, and variant quality control (supplemental methods).

Selection and processing of GWAS summary statistics

We selected summary statistics from the GWAS catalog for eight binary traits and for five continuous traits. Table 1 shows GWAS catalog study identifiers and traits. Where available, we directly used the pre-harmonized summary statistics provided by the GWAS catalog. For GWAS catalog studies GCST90013445[26] type 1 diabetes (T1D), GCST008972[27] (urate), GCST007954[28] glycated hemoglobin (HbA1c) and GCST004773[29] type 2 diabetes (T2D), we used the MungeSumstats R package[30] (version 1.0.1) to retrieve missing fields (e.g., variant positions). GWAS variants were matched to the HapMap3-1KG variants on the basis of positions and allele codes and renamed accordingly. Other quality-control steps include the flipping of variants to match the HapMap3-1KG reference; variant frequency filtering (>1%); and removal of variants with invalid *p* values (>1 or <0), ambiguous variants, variants with missing data, duplicate variants, and variants with sample size more than three standard deviations away from the median per-variant sample size (if available), as previously described.[16]

We selected GWAS studies with discovery samples from predominantly European ancestry discovery because the evaluated biobanks primarily contain individuals of European ancestry. Because we use subsets of the UKBB for evaluation and tuning, we selected for studies that had large sample sizes and that preferably did not include the UKBB in the discovery sample. Yet, the selected summary statistics for Alzheimer disease (AD) and height came from a GWAS that included the UKBB-EUR sample. Therefore, we did not use the UKBB-EUR sample for tuning or evaluation in these phenotypes.

Reference-genotype harmonization

We constructed our own definition of the HapMap3-variants[41] to avoid favoring one of the definitions used by the PGS methods. We retrieved HapMap3 variant rsIDs and downloaded genotypes for the 1000 Genomes reference from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140708_previous_phase3/v5_vcfs/ (v5). We retrieved updated rsIDs for all 1000 Genomes variants by using the Bioconductor SNPlocs.Hsapiens.dbSNP144.GRCh37 R package[42] (the latest version for the GRCh37 genome build at the time) with the GRCh37 variant positions and allele codes and intersected them with the HapMap3 variants on the basis of rsIDs. We then mapped these variants from GRCh37 to GRCh38 by using liftOver[43] and retrieved rsIDs in that genome build too, on the basis of location and allele codes, by using the SNPlocs.Hsapiens.dbSNP151.GRCh38 R package[42] (the latest version for the GRCh38 genome build at the time). We retained variants with an allele frequency of at least 1% in any of the 1000 Genomes superpopulations. These variants (HapMap3-1KG, $n = 1{,}330{,}821$) form the basis for subsequent analyses.

Please cite this article in press as: Monti et al., Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and..., The American Journal of Human Genetics (2024), https://doi.org/10.1016/j.ajhg.2024.06.003

Table 1. GWAS summary statistics used to derive PGS weights

Study	GWAS trait	n_{cas}	n_{con}	$n_{variants}$	Target traits
GCST005838[31]	stroke	67,162	454,450	1,121,867	stroke
GCST90012877[32]	AD or family history of AD	53,042	355,900	1,136,233	AD
GCST90013534[33]	RA	22,628	288,664	778,275	RA
GCST004773[29]	T2D	26,676	132,532	1,071,786	T2D
GCST004988[34]	breast cancer	76,192	63,082	1,137,481	breast cancer
GCST006085[35]	prostate cancer	79,148	61,106	1,139,693	prostate cancer
GCST90013445[26]	T1D	22,153	37,374	63,204	T1D
GCST004131[36]	IBD	25,042	34,915	1,103,333	IBD
GCST008059[37]	eGFR	567,460	–	1,141,659	CKD, eGFR
GCST90018959[38]	height	525,444	–	1,119,889	height
GCST008972[27]	urate levels	457,690	–	1,005,478	gout, urate levels
GCST002783[39]	BMI	236,781	–	1,039,042	BMI
GCST007140[40]	HDL	94,674	–	1,138,452	HDL
GCST007954[28]	HbA1c	88,355	–	1,009,664	HbA1c

Entries are ordered by the total sample size and type of trait (binary or continuous). From left to right: GWAS catalog study identifiers (study), the respective reported GWAS traits, number of cases (n_{cas}) and controls (n_{con}), the number of variants after intersection with HapMap3-1KG and quality control ($n_{variants}$), and the evaluated target traits. Scores constructed from urate and eGFR summary statistics were also evaluated for gout and CKD, respectively. The GWAS for T1D considered only a small panel of variants, of which 84% remained after intersection and QC.

The list of variants, along with GRCh37 (hg19) and GRCh38 (hg38) coordinates, rsIDs, and allele frequencies in the 1000 Genomes superpopulations, is available on Github (see data and code availability section). Scripts to reproduce these steps are available as part of the prspipe workflow (see data and code availability). The filtered and intersected 1000 Genomes genotypes are provided as a separate resource. Variants are further filtered during the construction of polygenic scores, as described below.

Target-genotype harmonization

Target genotype data were intersected with the HapMap3-1KG variants on the basis of positions and allele codes, renamed, and converted to PLINK1 format. The harmonized data served as input to all subsequent analyses involving target genetic data, i.e., ancestry reference matching and polygenic scoring. Target harmonization is part of the prspipe workflow, and corresponding steps are defined in https://github.com/intervene-EU-H2020/prspipe/blob/main/workflow/rules/genotype_harmonization.smk.

Binary disease-phenotype harmonization

We used expert-curated ICD-code-based definitions[22,44] developed at FinnGen to define binary disease traits (referred to as endpoints). Individuals were counted as cases for a specific endpoint if they satisfied ICD-9- or ICD-10-code-based inclusion or exclusion criteria (Table S13). The remaining (non-matching) individuals for that endpoint were counted as controls. All data used to define binary disease endpoints were registry based. Breast cancer was only evaluated when the reported sex was female, and prostate cancer only when the reported sex was male.

For the UKBB, we considered both main (data fields 41202 and 41203) and secondary (data fields 41204 and 41205) ICD-9 and ICD-10 diagnosis codes derived from hospital inpatient admissions.

Continuous-trait definitions

For the UKBB, we used the following data fields to define continuous traits: 50 for height, 21001 for body mass index (BMI), 30700 for creatinine, 30750 for HbA1c, and 30880 for urate.

For GNH, we considered all instances where a continuous trait was measured per individual through their primary and secondary health records. We removed outliers on the basis of a 6SD deviation per trait and calculated the mean value per trait per individual to use in the analysis.

For HUNT, the latest value was chosen when continuous traits were measured at more than one baseline enrollment or sub-study screening over three recruitment waves since 1984. Standard quality-assessment measures were taken across variables and are described at the HUNT Databank (https://hunt-db.medisin.ntnu.no/hunt-db/variablelist). BMI was defined by height and weight measured at screening. High-density lipoprotein (HDL) and creatinine were measured from serum in non-fasting individuals. HbA1C was measured in mmol/mol according to the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) standard.

For Estonian Biobank, the earliest values available were chosen for BMI and height, as some individuals are repeatedly measured. Height values larger than 260 cm or smaller than 100 cm were omitted. Similarly, BMI values less than 10 kg/m^2 or larger than 200 kg/m^2 were discarded. Metabolic profiles for HDL and creatinine were obtained with NMR for a random subset from the Estonian Biobank ($n = 10{,}681$).

For each biobank in which creatinine measurements were available, we calculated the estimated glomerular filtration rate (eGFR) based on the diet according to the renal-disease study equation,[45] as follows:

$$eGFR = \alpha \times S_{cr}^{-1.154} \times age^{-0.203} \times \sigma$$

Where α is 30,849 if creatinine was measured in μmol/L or 175 if measured in mg/dL, S_{cr} is the serum creatinine measurement, and σ is 0.742 if the reported sex is female or 1 if the sex is male. We did not include the multiplier for "ethnicity" because we only perform comparisons within ancestry-matched populations. During evaluation, all continuous traits are standardized to mean 0 and unit standard deviation.

Polygenic-score weight derivation

We derived polygenic scoring weights with pT + clump, lassosum, PRScs, SBayesR (robust parameterization), LDpred2, and DBSLMM by using the settings described previously.[16] For MegaPRS, we used the author-recommended BLD-LDAK heritability model and specified "–model mega" to fit many different scores with different tools (lasso, bolt, ridge, bayesr) and included the HLA region, as recommended. Software versions and sources for each tool are listed in Table S14.

Besides letting methods determine suitable hyperparameters on the basis of the summary statistics alone (automatic tuning), we generated scores over grids of hyperparameters for target-data-based tuning with 10-fold CV (see below).

We used European-ancestry LD reference panels for all analyses, as the selected GWAS were performed in samples of mostly European ancestry. In contrast to Pain et al., we used the PGS-method author-provided LD-references for DBSLMM, lassosum, LDpred2, SBayesR, and PRScs. DBSLMM and lassosum LD references are based on the 1000 Genomes data. For LDpred2, SBayesR, and PRScs, they are based on UKBB data. We used the 1000 Genomes EUR-subset to calculate LD when running pT + clump and MegaPRS. Scripts to download PGS-method software and data are part of the prspipe workflow. The workflow uses GenoPred scripts to generate PLINK2-compatible scoring files.

Ancestry matching and removal of genetic outliers

Rather than directly inferring genetic ancestry, we scored individuals according to their similarity with groups defined in the 1000 Genomes reference [46]. We used GenoPred Ancestry_identifier.R to project target genetic data into the 1000 Genomes genetic principal-component space and match individuals to one of the five 1000 Genomes superpopulations (AFR, AMR, EAS, EUR, or SAS). Both the target and 1000 Genomes genotype data are filtered to variants available in both samples (subset of HapMap3-1KG) with allele frequencies above 5%, missingness below 2%, and variants that do not violate Hardy-Weinberg equilibrium ($p < 1 \times 10^{-6}$). Regions of long LD were excluded,[47] and variants were LD-pruned on the basis of the 1000 Genomes reference (with PLINK "–indep-pairwise 1000 5 0.2"). Genetic PCs were derived on the basis of the filtered variants, and an elastic net classifier is fit with 5-fold CV so that individuals were placed into one of the five groups on the basis of 100 PCs. Target genotype data were projected into the same PC space with PLINK, and the classifier was used to predict the most highly matched superpopulation for all individuals.

Additionally, we used GenoPred Population_outlier.R to remove extreme outliers within the assigned ancestry-matched groups in the target data on the basis of the first eight genetic principal components constructed within those groups. We used the same variant filters described above (except that LD pruning was now performed within the target data), calculate genetic PCs within the assigned groups, and define up to ten centroids in the PC space by using R NbClust[48] (distance = 'euclidian,' method = 'kmeans'). For each centroid, the Euclidian distance of individuals to the center is calculated, and those with distances that are larger than the 75th percentile +30 IQR (i.e., extreme outliers) are removed. The UKBB was the only biobank that had more than one ancestry group well represented. Our analyses focus on the replicated groups EUR (UKBB, EBB, HUNT, and FinnGen) and SAS (UKBB and GNH).

Polygenic scoring

We performed polygenic scoring for scores derived by single methods with PLINK2 and GenoPred Scaled_polygenic_scorer_plink2.R. Polygenic scoring is part of the prspipe workflow. For the evaluation of the ensemble PGS, we performed scoring with PLINK2. In both cases, missing genotypes are imputed via the 1000-Genomes-matched-superpopulation allele frequencies as previously described.[16]

Hyperparameter tuning and ensemble PGS

For the methods that generate scores over a range of hyperparameters (pT + clump, lassosum, PRScs, LDpred2), we used 10-fold CV to select the score with the largest correlation with the trait, as described previously.[16] Where available, we included scores produced by methods' automatic settings in the selection process. We perform CV by using 80% of the UKBB EUR data and retain 20% for evaluation (we used different subsamples for each trait in order to perform stratified sampling).

For pT + clump, we defined the score given by the *p*-value threshold of 1×10^{-8} as the automatically tuned score. SBayesR and DBSLMM only use automatic settings, i.e., they produce just a single set of weights and are not tuned with CV.

To fit the ensemble PGS, we included all scores from all methods across hyperparameters and used 10-fold CV to determine suitable shrinkage parameters for an elastic net model combining the different scores with the caret R package[49] (which relies on glmnet[50]). For tuning of the ensemble PGS, we used non-nested scores for pT + clump, i.e., scores with disjoint variant sets corresponding to 10 *p*-value bins. These steps were performed with GenoPred Model_builder_V2.R and are part of the prspipe workflow.

To score other biobanks with the UKBB ensemble PGS, we generated PLINK2-compatible scoring files by multiplying the PGS weights of every variant with their corresponding weights in the ensemble PGS model and adding them (yielding a single weight for each variant).

Performance evaluation within biobanks

All PGSs were standardized to mean zero and unit standard deviation within biobanks and ancestries for performance evaluation. We calculated the following metrics for binary disease traits: β coefficients, i.e., the change in log-odds ratios per PGS standard deviation; the change in odds ratio per PGS standard deviation (OR = exp(β)); the fraction of variance explained on the observed scale (r^2_{obs}); and the area under the receiver operating characteristic curve (AUROC). To calculate the variance explained on the liability scale (r^2_{liab}) from r^2_{obs},[51] we used the median prevalence within ancestries as the population prevalence estimate (Table S1). We retrieved DeLong 95% confidence intervals[52] for the AUROC by using the ci.auc-function in the pROC R package. Confidence intervals for r^2_{obs} were derived from 1000 bootstrap samples of r_{obs} (the Pearson correlation on the observed scale) for binary traits.

For continuous traits, we calculated β coefficients, i.e., the change in standard deviations of the trait per standard deviation of the PGS and the fraction of variance explained (r^2_{obs}).

When comparing two effect sizes of scores $\beta_{a,i}$ and $\beta_{b,i}$ within biobank "I," we used the two-sided z test and adjusted for the correlation between scores, with the following test statistic:

$$z = \frac{\beta_{a,i} - \beta_{b,i}}{\sigma}$$

where

$$\sigma = \sqrt{\sigma^2_{a,i} + \sigma^2_{b,i} - 2\rho_{(a,b),i}\sigma_{a,i}\sigma_{b,i}}$$

$\sigma_{a,i}$ and $\sigma_{b,i}$ denote the standard deviations of $\beta_{a,i}$ and $\beta_{b,i}$, respectively, and $\rho_{(a,b),i}$ denotes the correlation between scores "a" and "b" measured in biobank "i."

Data exclusions before meta-analyses

None of the scores evaluated in GNH-SAS for T1D reached nominal significance ($p < 0.05$, two-sided z test) for association with the endpoint (Table S1), and all effect sizes were close to zero. We removed these data from further analysis. We found strongly reduced effect sizes of scores for HbA1c in GNH-SAS compared to UKBB-SAS (Table S2; Figure S1) and decided not to include these data for meta-analyses (it was unclear whether reduced performance was due to a phenotyping issue). We found low effect sizes in comparison to those obtained from other biobanks for T1D in HUNT (Figure S1), determined it was most likely due to a phenotyping issue, and excluded those data from meta-analyses.

Meta-analysis for method comparisons

All meta-analyses were performed in R (version 4.1.1) with the metafor package (rma.mv function, version 3.8–1); the V-argument was used to account for the dependence of effect sizes within biobanks (see below), and models were fit with REML.

We meta-analyzed the β coefficients of scores across biobanks and within ancestries and traits by using meta-analytic mixed-effects models. The observed β coefficients are modeled as follows:

$$\beta_{s,b} = \boldsymbol{\alpha}\mathbf{w}_s + \zeta_b + \epsilon_{s,b},$$

where $\beta_{s,b}$ is the observed coefficient for PGS "s" in biobank "b" for a specific trait. $\beta_{s,b}$ is modeled as a combination of fixed effects (moderators) with realizations $\mathbf{w}_s$ and parameters $\boldsymbol{\alpha}$ (bold characters indicate vectors) and two error terms: the sampling error $\epsilon_{b,s}$ and a biobank-specific random intercept ζ_b (shared by all observed coefficients in that biobank). $\tau^2_{biobank} = var(\zeta)$ is the random effect, where var(ζ) denotes the variance of the biobank-specific random intercepts ζ.

For every trait, we meta-analyzed up to 13 PGSs in the same model. PGS choice is modeled with the fixed effects, i.e., $\mathbf{w}_s$ only contains a single non-zero entry of 1, indicating which PGS "s" produced $\beta_{s,b}$. With this parameterization, parameters in $\boldsymbol{\alpha}$ directly correspond to the meta-analyzed effect sizes for the different PGSs after inverse variance weighting, and the formula above can equivalently be written as follows:

$$\beta_{s,b} = \beta_{s*} + \zeta_b + \epsilon_{s,b},$$

where β_s* is the average effect size for score "s" across all biobanks "*." To test whether two meta-analyzed effect sizes are significantly different, we compared parameters α_a and α_b with H_0: $\alpha_a - \alpha_b = 0$ by using the z test. We also report results for the t test (produced by the anova function applied to metafor rma.mv objects) in Tables S5 and S6.

We retrieved 95% likelihood-based confidence intervals for $\tau^2_{biobank}$ by using the confint function (all values are reported in Table S4). In Tables S5 and S6, we further report meta-analyzed AUROC, r^2_{obs}, and r^2_{liab} values that we produced by weighting studies by their effective sample size.[53]

Method ranking

To rank methods across traits, we considered just the traits for which CV tuning and ensemble PGS were available (i.e., all except height and AD) and ranked scores on the basis of their meta-analyzed effect sizes β_s* (see definition above). To avoid counting scores produced by the same summary statistics twice for eGFR and CKD and for urate and gout (e.g., for the ranks shown in Figure 3A), we applied the following rule: If the continuous phenotype was available in the same number of biobanks as its binary counterpart, we used the continuous phenotype (higher power), otherwise we used the binary phenotype (larger target diversity). This led to consideration of eGFR for SAS, CKD in EUR, urate for SAS, and gout in EUR. We applied the same reasoning when calculating mean and median values of method performances across all traits.

3-Level meta-analytic random effects models

For the 3-level meta-analysis, the observed effect-sizes are modeled as follows:

$$\beta_{b,m} = \mu_\beta + \zeta_{(2)b,m} + \zeta_{(3)b} + \epsilon_{b,m},$$

where $\beta_{b,m}$ is the observed β coefficient for method "m" in biobank "b," μ_β is the mean of the distribution of true effect sizes across biobanks and methods, $\zeta_{(2)b,m}$ is the within-biobank random intercept due to the choice of method (level 2), and $\zeta_{(3)b}$ is the random intercept due to target biobank (level 3, shared by all observations in the biobank). The estimated parameter $\tau^2_{biobank} = \tau^2_{(3)} = var(\boldsymbol{\zeta_{(3)}})$ quantifies the heterogeneity of effect sizes due to the target biobank, and $\tau^2_{method} = \tau^2_{(2)} = var(\boldsymbol{\zeta_{(2)}})$ quantifies the heterogeneity of effect sizes due to the choice of method within the biobank.[54] In contrast to the model introduced in the previous section, method effects are considered nested within biobanks (independent between biobanks).

All models were fit with restricted maximum-likelihood with the metafor package in R (rma.mv function), and use of the V argument accounted for the dependence of effect sizes measured within the same biobanks (see below). We retrieved 95% likelihood-based confidence intervals for $\tau^2_{biobank}$ and τ^2_{method} by using the confint function. To calculate $I^2_{biobank} = I^2_{(3)}$ and $I^2_{method} = I^2_{(2)}$, we used the implementation provided in dmetar[54] (var.comp/mlm.variance.distribution function, https://github.com/MathiasHarrer/dmetar/blob/master/R/mlm.variance.distribution.R, commit 21bde652cbae5677b56b0ff848eb96c9bea877d8) on the basis of the three-level extension of the I^2 metric.[55] $I^2_{biobank}$ captures the fraction of the overall variance in effect sizes (including sampling error) attributable to the biobank (level 3), and I^2_{method} captures the fraction of the overall variance in effect sizes attributable to methods within biobanks (level 2).

Accounting for dependent effect sizes in meta-analytic models

Within each biobank, ancestry, and trait, we calculated pairwise correlations between polygenic scores on the basis of up to

50,000 randomly sampled individuals. We used the resulting score-score correlation matrix R_b (where "b" indicates the biobank) to estimate V_b, the variance-covariance matrix capturing the dependency of errors of effect-size estimates for biobank "b" as follows:

$$V_b = S_b R_b S_b,$$

where S_b is a diagonal matrix containing the standard errors of the estimated effect sizes corresponding to the rows and columns of R_b. The effect sizes measured in different biobanks are considered independent; therefore, the full matrix V supplied to the rma.mv function is a block diagonal matrix containing all V_b values for the different biobanks b from 1 to n on the diagonal

$$V = \begin{matrix} V_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & V_n \end{matrix},$$

where 0 denotes a matrix of zeros with the same shape as the different V_b.

Calculating PGS variance in the HLA region

For the phenotypes T1D and rheumatoid arthritis (RA), we scored individuals in the 1000 Genomes EUR subset by using either all PGS variants or only variants contained in the HLA region (defined as the interval 28,000,000–34,000,000 on chromosome 6). We then computed the fraction of variance by dividing the variance of HLA-only PGS by that of the full PGS.

Results

Prspipe workflow and experimental setup

We created a snakemake workflow prspipe to run different polygenic-risk scoring methods based on GWAS summary statistics. Prspipe makes it possible to automate the within-biobank analyses from Pain et al.[16] on the the basis of the GenoPred suite of scripts (see web resources). Notable differences include an updated set of methods (*p*-value thresholding and clumping (pT + clump), lassosum,[10] PRScs,[11] LDpred2,[13] DBSLMM,[14] SBayesR[12] (robust parameterization), and MegaPRS[15]), the use of LD reference panels provided by the methods' authors, and software managed partially with containers. We used prspipe to derive PGS weights by using both methods' automatic settings (auto) and grids of hyperparameters (MegaPRS, LDpred2, PRScs, lassosum, and pT + clump).[16] For the baseline method pT + clump, we considered the score with the most stringent *p*-value threshold ($p < 1 \times 10^{-8}$, i.e., keeping only highly significant variants) as the automatically tuned score.

The workflow defines steps to set up PGS methods, download and process summary statistics from the NHGRI-EBI GWAS Catalog,[56] run PGS methods (i.e., the derivation of PGS weights), target genotype harmonization, run ancestry matching based on the 1000 Genomes superpopulations,[46] and target polygenic scoring with PLINK2.[57] Because PGS performance depends on the genetic similarity of the target and GWAS samples,[58] performance evaluation is stratified according to the matched superpopulation. Using CV, one fits elastic net models combining scores from different methods (ensemble PGSs), and the best single PGS weights are selected for each method (hyperparameter tuning).[16]

We applied this workflow to 14 sets of summary statistics from the GWAS Catalog to derive a PGS and predict six continuous traits and 10 binary disease traits derived from harmonized ICD-code-based definitions[22] (methods, Table 1). Our main analyses focus on the two replicated ancestry-reference-matched superpopulations: EUR and SAS. The number of cases used for performance evaluation across biobanks ranged from 5,384 (T1D) to 81,487 (type 2 diabetes; T2D) for EUR and 60 (RA, available in GNH only) to 8,696 (T2D) for SAS ancestry-matched target data. The total sample size for performance evaluation for continuous traits ranged from 85,973 (urate, available in UKBB only) to 524,056 (height) for EUR target data and from 13,572 (urate) to 43,197 (height) for SAS target data (Table 2).

Using 80% of the UKBB EUR target data (training set), we selected the best-performing weights for each method and fit the ensemble PGS (full workflow). PGS weights were shared with other biobanks, in which we still performed target data harmonization, ancestry matching, and polygenic scoring steps needed for performance evaluation (Figure 1).

Browsable results, meta-analysis and ranking

As outlined in Figure 2 for T2D, we calculated PGS effect sizes (Figure 2A) for continuous and binary traits across all target biobanks and ancestries (Tables S1–S3; Figures S1–S5) and performed mixed-model meta-analyses within ancestry groups to determine the best-performing PGS (the one with the largest effect size) for each trait across biobanks (Figures 2B and 2C; Tables S4–S6). Additionally to scores produced by single methods, we evaluated the UKBB-tuned ensemble PGSs in other biobanks after projecting them back to the variant-level (see methods). For each trait, we meta-analyzed β coefficients (i.e., the change in the trait per PGS standard deviation, on the log-odds scale for binary traits, in standard deviations for continuous traits) of up to 13 PGSs corresponding to different tuning types (auto or CV) for seven methods and the UKBB-EUR-tuned ensemble PGS (Figure S6).

Other than the ensemble PGS, we found that CV-tuned PGSs from LDpred2 and MegaPRS ranked highly across traits (Figures 3A and 3B). The median relative increase in PGS effect size over CV-tuned pT + clump was 29.2% for CV-tuned LDpred2 (mean 30.9% ± 10.3 SD, $n = 12$, EUR) and 29.9% for CV-tuned MegaPRS (mean 31.2% ± 12.6 SD, $n = 12$, EUR), showing overall comparable performance (the median relative difference between the two was 0.1% in favor of MegaPRS).

Scores produced by automatic tuning appeared to be overall less reliable, especially for LDpred2 (see discussion section) and SBayesR (as previously described[16]). Although

Table 2. Target sample sizes across traits

	EUR total	SAS total	EUR EBB	EUR FinnGen	EUR HUNT	EUR UKBB (test)	EUR UKBB (train)	SAS GNH	SAS UKBB
AD	15,940	–	555	13,823	1,562	–	–	–	–
RA	13,060	60	2,384	9,332	1,139	205	820	60	–
Breast cancer	23,610	393	2,685	16,076	1,729	3,120	12,483	197	196
CKD	19,714	1,609	4,224	9,314	2,802	3,374	13,496	1,131	478
Gout	22,399	488	10,646	8,759	1,318	1,676	6,704	282	206
IBD	13,016	634	2,097	7,815	1,769	1,335	5,340	466	168
Prostate cancer	20,492	205	2,227	13,606	2,242	2,417	9,671	95	110
Stroke	37,920	635	4,515	26,166	5,204	2,035	8,142	424	211
T1D	5,384	443	501	4,286	396	201	804	443	–
T2D	81,487	8,696	12,344	59,345	3,861	5,937	23,748	6,630	2,066
BMI	346,290	42,243	189,651	–	66,663	89,976	359,913	33,146	9,097
HDL	139,248	37,693	10,642	–	49,824	78,782	315,135	29,628	8,065
HbA1c	120,242	21,696	–	–	34,192	86,050	344,209	12,948	8,748
Height	524,056	43,197	190,013	267,343	66,700	–	–	34,089	9,108
Urate	85,973	13,572	–	–	–	85,973	343,904	4,730	8,842
eGFR	152,793	38,916	–	–	66,759	86,034	344,140	3,061	8,855

For each trait and replicated ancestry group (EUR, SAS), the number of cases (binary disease traits) or the sample size are shown, either combined ("total," excluding UKBB training data) or separated by biobank. For the UKBB-EUR, data were split into training (80%, used to tune hyperparameters and ensemble PGS) and test sets (20%, used for evaluation and meta-analyses). UKBB EUR data were excluded for Alzheimer disease and height as a result of sample overlap and could therefore not be used for tuning (leaving 14 traits for a full evaluation). Dashes indicate the phenotype was unavailable.

automatic tuning typically outperformed the baseline method pT + clump (even when the latter was tuned with CV), we observed seemingly non-systematic cases of reduced relative performance (e.g., SBayesR for urate and gout, DBSLMM for Alzheimer disease, or LDpred2 for HbA1c or RA) (Figure S11). MegaPRS was the best automatically tuned method (median 23.3% relative increase over CV-tuned pT + clump, mean 27.4% ± 15.9 sd, EUR), yet PGS effect sizes were comparatively low for some continuous traits (e.g., BMI, HDL, and height, Figure S11).

Effect-size differences between the top PGSs as revealed by single methods were mostly not significant (Figure S7, FWER $\leq$ 0.05, two-sided z test). We also provide these data on the level of individual biobanks (Figures S9 and S10), revealing that the best single method for a given trait was not necessarily consistent between biobanks.

UKBB-tuned ensemble PGS outperforms other methods

The ensemble PGS ranked favorably for all traits in EUR- and SAS-matched target data (Figures 3A and 3B) except for T1D in EUR (driven by lower performance in FinnGen) and stroke in SAS (the trait with the overall lowest performance). For EUR target data, effect sizes were significantly greater than those of all other PGSs for 6/9 binary and 5/5 continuous traits (FWER $\leq$ 0.05, two-sided z test) and were the largest overall in 13/14 traits for which we fit the ensemble PGS. These results stood out in comparison to those for single methods, which did not produce a consistent best method and for which differences were mostly not significant.

Compared to the best single methods, the median relative increase in effect size was 3.7% over CV-tuned LDpred2 and 4.5% over CV-tuned MegaPRS for binary disease traits ($n = 9$). Median relative increases for continuous traits were larger (5.2% and 7.9%, respectively, $n = 5$). When measured in terms of variance explained, relative differences were larger. We observed median relative increases of 7.4% and 9% for binary traits (liability scale) and 10.7% and 16.1% for continuous traits, respectively (Figure S8; Table S7). Similar trends were observed for SAS target data, with the ensemble PGS having the largest effect size in 12/13 traits, albeit its effect size was only significantly larger than all others for continuous traits urate, eGFR and HDL (FWER $\leq$ 0.05, two-sided z test). We report relative effect sizes of all methods relative to the ensemble PGS in Table 3.

CV tuning increases PGS robustness

Hyperparameter tuning with CV using the UKBB EUR data was often beneficial and rarely harmful when evaluated on EUR target data (Figure 3C). Hyperparameter tuning with CV strongly increased effect sizes in a subset of traits for specific methods, rather than providing large benefits across traits (Figures S12 and S13). pT + clump benefited most from CV tuning when evaluated on EUR target data, i.e., selecting p value thresholds larger than the baseline 1×10^{-8} was always beneficial (median 12.8%

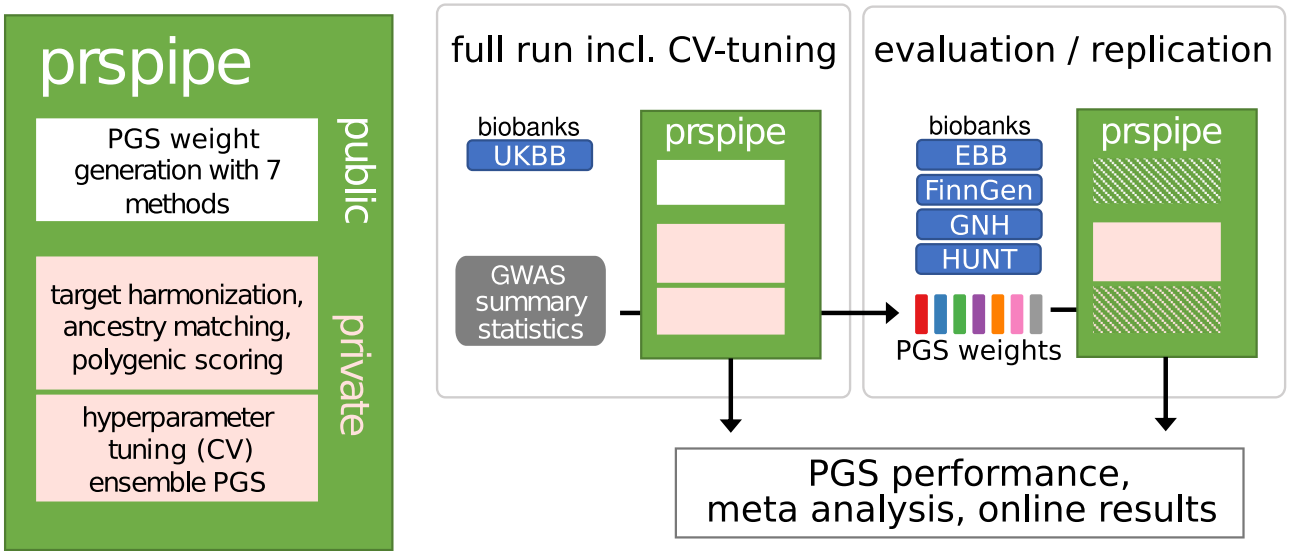


Figure 1. prspipe workflow and application

Prspipe, introduced by Pain et al.,[16] is a snakemake workflow that automates within-biobank method comparisons. The public stage uses only public data (e.g., summary statistics, ancestry reference, and PGS software) to derive PGS weights by using seven methods from GWAS summary statistics. The private stage requires access to target genotype and phenotype data and includes data harmonization, polygenic scoring, and PGS tuning via cross-validation (CV). We used prspipe to generate PGS weights and tune hyperparameters in the UKBB EUR data (full run). PGS weights were shared with other biobanks for evaluation and replication (skipping the public stage). Other biobanks were not used for hyperparameter tuning. Downstream analyses were conducted so that PGS performance could be determined with a meta-analytical framework (not part of the workflow), and results were published as an online resource at https://methodscomparison.intervenegeneticscores.org.

increase in effect size); lassosum (median 6.2% increase) and LDpred2 (median 4.1% increase) followed. MegaPRS and PRScs benefited the least (median 1.2% and 0.2% increase, respectively). For SAS target data, the median benefits were smaller, except for PRScs (Table S8), and were overall less consistent (Figure 3C). Mean increases were larger for all methods except for PRScs in EUR target data, often dominated by few instances in which automatic tuning had comparatively low performance.

The performance increases seen for CV tuning were by and large significant when evaluated in EUR target data (Figure 3C, FWER $\leq$ 0.05, two-sided z test), except for PRScs which only saw an improvement for two phenotypes. Significant negative effects of CV tuning for EUR data were only observed for RA (PRScs and MegaPRS, driven by FinnGen) and CKD (PRScs, Figure S12). For SAS target data, we observed fewer significant differences, and PRScs was the only method for which we observed a significant reduction in effect size (BMI, Figure S13). A more detailed description of these comparisons is provided in the supplemental results.

Tuned PGS performance varies more between biobanks than between methods within biobanks

We estimated PGS effect-size heterogeneity between biobanks and how it compares to the heterogeneity between methods within biobanks by using three-level meta-analytic random-effects models in EUR target data (methods, Figure 4). These nested models have two random-effect parameters: τ^2_{biobank} and τ^2_{method}. τ^2_{biobank} captures effect-size heterogeneity due to differences between biobanks, and τ^2_{method} captures heterogeneity due to differences between methods within biobanks (we report their square roots, τ_{biobank} and τ_{method}, because they are on the same scale as the PGS effect sizes). Additionally, we estimated I^2_{biobank} and I^2_{method}, which quantify the overall fraction of variance (between 0 and 1) in effect sizes attributable to the biobank or choice of method within the biobank, respectively.

We focused on scores selected via CV in the UKBB-EUR sample (if available) and excluded SBayesR scores that performed poorly in the UKBB-EUR 80% training data (RA, T1D, BMI, and urate/gout). We did not consider the ensemble PGS or baseline method pT + clump, meaning that up to six scores were considered per trait. We chose this setting to mimic the case in which multiple validated PGSs from standard methods are available.

We found significant heterogeneity of PGS effect sizes in all 13 traits replicated in at least two biobanks (FWER $\leq$ 0.05, Cochran's Q test, accounting for 13 tests). The target biobank had a larger influence on the PGS effect size than the choice of method within biobank across all traits (i.e., $\tau_{\text{method}} < \tau_{\text{biobank}}$, Figure 4; Table 4). When adjusting for covariates sex, age, and genetic PCs 1–10, we found that this effect was slightly reduced, but $\tau_{\text{method}} < \tau_{\text{biobank}}$ remained true for the majority of traits (10 out of 13; T1D, stroke, and T2D had $\tau_{\text{method}} > \tau_{\text{biobank}}$) (supplemental results, Table S16). However, likelihood-based 95% confidence intervals for τ_{biobank} were large and sometimes included the estimate for τ_{method} (RA, stroke, and T2D) and 0 (T1D and breast cancer). The variation in PGS effect sizes could

Please cite this article in press as: Monti et al., Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and..., The American Journal of Human Genetics (2024), https://doi.org/10.1016/j.ajhg.2024.06.003

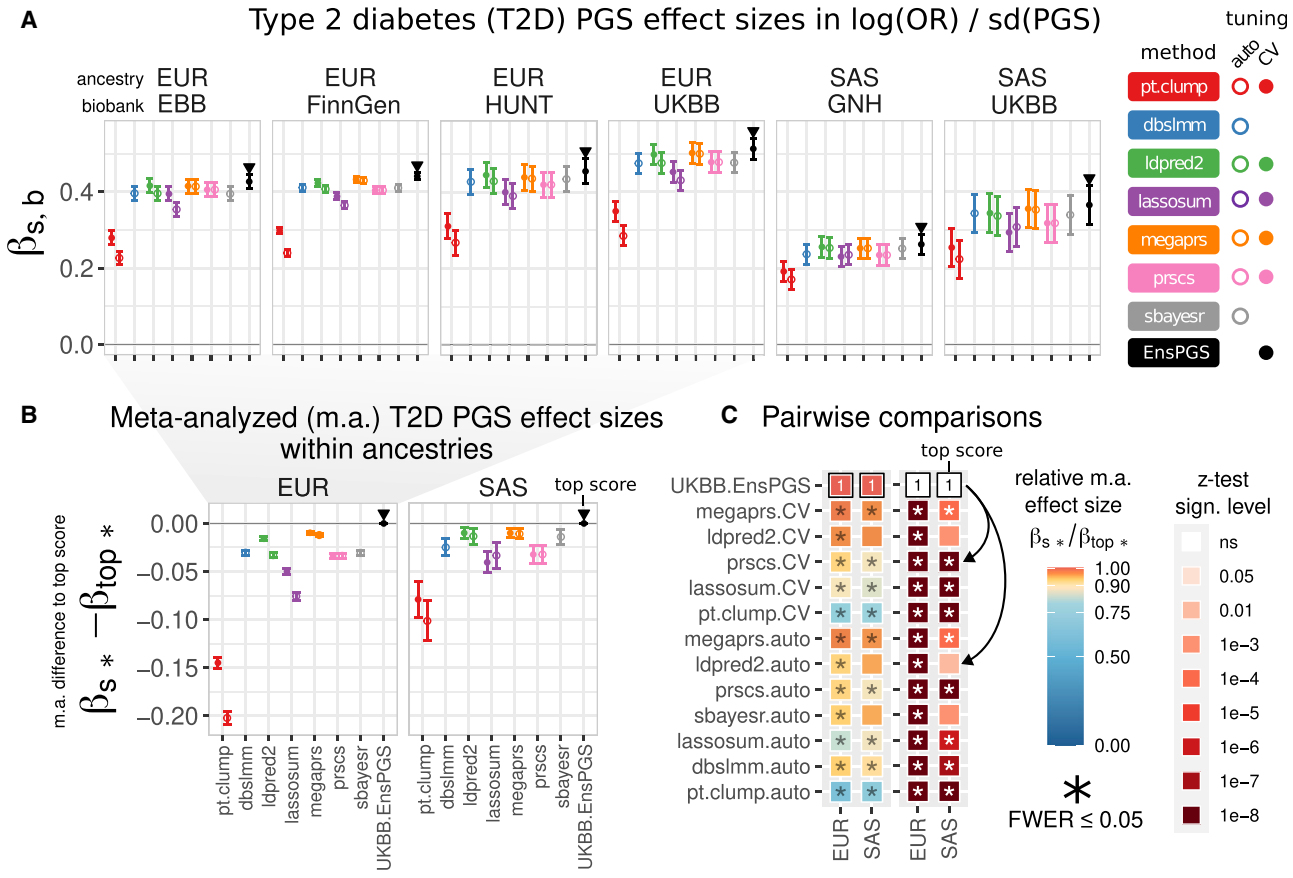


Figure 2. Meta-analysis workflow for method comparison

(A) PGS effect sizes $\beta_{s,b}$ (i.e., the change in log odds ratio per PGS standard deviation measured for scores "s" across biobanks "b," (see methods) with 95% confidence intervals for all PGS methods (x axis) stratified by biobank, replicated ancestries (EUR or SAS), and tuning types (auto or CV) serve as the inputs for the meta-analysis (type 2 diabetes is shown as an example trait). We evaluated scores for the seven methods shown on the right, as well as the UKBB-EUR-tuned ensemble PGS (EnsPGS). The largest effect size for each ancestry and biobank is marked with a triangle (given by the ensemble in all cases). $\beta_{s,b}$ values for all target data and traits are displayed in Figure S1 and are browsable online.

(B) PGS effect sizes are meta-analyzed within ancestries across biobanks (yielding a single β_s* for each score "s"). Effect-size differences relative to the largest meta-analyzed effect size (β_{top}*, given by the ensemble) and 95% confidence intervals are shown. All pairwise differences are available in Tables S5 and S6 and are browsable online.

(C) Meta-analyzed effect sizes β_s* are compared, and significance testing is performed. Heatmaps show both the effect size relative to the largest (β_s*/β_{top}*, left) as well as corresponding two-sided z test significance levels at which H_0: β_s*-β_{top}* = 0 can be rejected (right). Significant differences at an FWER ≤ 0.05 are marked with an asterisk (*), accounting for all 351 tests performed across traits and ancestries. The score against which comparisons are performed with effect size β_{top}* is marked with a "1" and black border. Arrows indicate two example comparisons: against PRScs-CV (significant difference in SAS and EUR) and LDpred2-auto (significant only in EUR). Data for all PGSs and traits are provided in Figure S6.

to a large degree be explained by heterogeneity between biobanks (average $I^2_{biobank}$ = 82.9% ± 14.3 SD, n = 13) and, to a lesser degree, by heterogeneity between methods (average I^2_{method} = 11.97% ± 12.4 SD, n = 13).

Effect sizes for inflammatory bowel disease and RA varied most between biobanks ($\tau_{biobank}$), including when we adjusted for the average effect size ($\tau_{biobank}/\mu_\beta$). Effect sizes for BMI varied the least between biobanks, both absolutely ($\tau_{biobank}$) and relative to the average effect size ($\tau_{biobank}/\mu_\beta$). For binary traits, effect sizes for breast cancer varied the least across biobanks, both absolutely and relative to the average effect size.

Across traits, τ_{method} was correlated with the average effect size (Pearson correlation 0.54, p = 0.0558, t statistic = 2.1377, 11 degrees of freedom), especially when T1D and RA were removed (Pearson correlation 0.85, p = 0.00094, t statistic = 4.83, 9 degrees of freedom), i.e., we found a linear relationship between the differences between methods and overall effect size, especially in the set of non-autoimmune traits.

$\tau_{biobank}$ was less correlated with the meta-analyzed average effect size (Pearson correlation 0.367, p = 0.219, t statistic = 1.30, 11 degrees of freedom), i.e., large PGS effect sizes weren't necessarily associated with higher variability between biobanks.

For SAS ancestry target data, we did not find significant heterogeneity of PGS effect sizes in CKD, stroke, prostate cancer, or breast cancer (FWER ≤ 0.05, Cochran's Q test, accounting for 11 tests), and $\tau_{biobank}$ could never reliably be estimated (Figure S14). Likelihood-based 95%

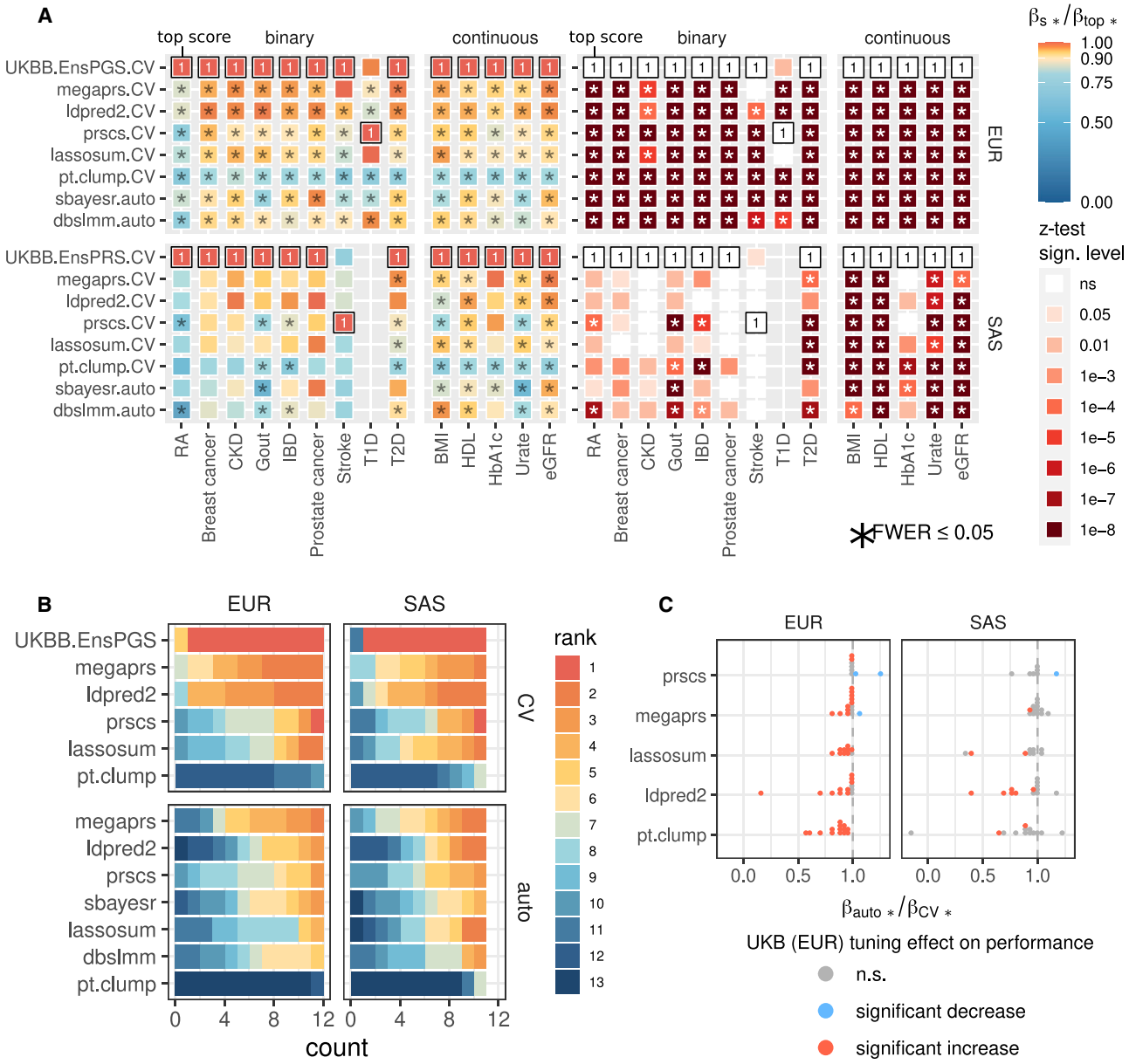


Figure 3. Relative meta-analyzed PGS effect sizes across 14 traits

(A) For the 14 traits for which we tuned hyperparameters by using CV (x axis), we show heatmaps of meta-analyzed β coefficients relative to the highest within traits (β_s*/β_{top}*, left) as well as significance levels for the two-sided z test (H_0: β_s* − β_{top}* = 0, right) stratified by ancestry (EUR or SAS). The top score with the largest effect size for each trait (β_{top}*) is marked with a "1" and black box. Differences significant at FWER ≤ 0.05 are marked with asterisks (*), accounting for all 351 pairwise tests performed across traits, replicated ancestries, and tuning types (auto or CV) (full data for all traits and scores are displayed in Figure S6).

(B) Bar plot counting PGS ranks across traits (1 is the highest), stratified by ancestry, method, and tuning-type (auto or CV). Methods are ordered by the average rank-sum across tuning types (highest to lowest).

(C) For each method (y axis), dot plots show the relative meta-analyzed effect size of the score derived with the methods' automatic settings against the CV-tuned scores (β_{auto}*/β_{CV}*). Colors denote the sign and significance of the two-sided z test (H_0: β_{CV}*-β_{auto}* = 0) at FWER ≤ 0.05 after 114 tests were accounted for across traits and ancestries (Figures S12 and S13). Methods are ordered by the median difference.

confidence intervals for $\tau_{biobank}$ included 0 for 9/11 replicated traits (all but T2D and eGFR).

High variability between PGS methods for autoimmune diseases

Effect sizes were most variable between methods for autoimmune diseases T1D and RA (τ_{method}) (Figure S15), even when we accounted for the average effect size in those traits (τ_{method}/μ_β) or when we considered variation in these effect sizes relative to the total variation of effect sizes (I^2_{method}). The scores for T1D and RA also had the largest fraction of PGS variance originating in the HLA region (mean 0.7 ± 0.12 SD and 0.54 ± 0.21 SD, respectively) (Figure S16). For T1D, the method with the largest effect size appeared to be

Table 3. PGS-meta-analyzed β coefficients relative to those analyzed with the ensemble PGS ($\beta_s^*/\beta_{EnsPGS}^*$)

Method	Tuning type	Trait	N (EUR)	N (SAS)	Median (EUR)	Median (SAS)	Mean (EUR)	SD (EUR)	Mean (SAS)	SD (SAS)
ldpred2	CV	binary	9	8	0.965	0.963	0.943	0.045	0.972	0.127
megaprs	CV	binary	9	8	0.957	0.934	0.947	0.041	0.958	0.124
lassosum	CV	binary	9	8	0.921	0.914	0.913	0.061	0.920	0.111
prscs	CV	binary	9	8	0.903	0.900	0.896	0.100	0.909	0.259
pt.clump	CV	binary	9	8	0.735	0.734	0.721	0.077	0.748	0.186
megaprs	auto	binary	9	8	0.948	0.950	0.933	0.050	0.964	0.104
ldpred2	auto	binary	9	8	0.927	0.955	0.838	0.265	0.904	0.314
prscs	auto	binary	9	8	0.925	0.876	0.915	0.052	0.877	0.264
sbayesr	auto	binary	9	8	0.907	0.895	0.873	0.083	0.841	0.181
dbslmm	auto	binary	9	8	0.904	0.865	0.890	0.092	0.815	0.199
lassosum	auto	binary	9	8	0.891	0.870	0.861	0.103	0.753	0.262
pt.clump	auto	binary	9	8	0.629	0.627	0.607	0.098	0.527	0.302
ldpred2	CV	continuous	5	5	0.950	0.936	0.948	0.016	0.925	0.049
megaprs	CV	continuous	5	5	0.927	0.931	0.940	0.024	0.946	0.034
prscs	CV	continuous	5	5	0.923	0.926	0.909	0.031	0.875	0.090
lassosum	CV	continuous	5	5	0.906	0.920	0.914	0.026	0.916	0.021
pt.clump	CV	continuous	5	5	0.735	0.729	0.743	0.037	0.718	0.048
prscs	auto	continuous	5	5	0.923	0.928	0.907	0.028	0.903	0.076
dbslmm	auto	continuous	5	5	0.901	0.904	0.891	0.039	0.897	0.066
sbayesr	auto	continuous	5	5	0.887	0.862	0.868	0.075	0.806	0.184
megaprs	auto	continuous	5	5	0.883	0.907	0.885	0.067	0.922	0.055
lassosum	auto	continuous	5	5	0.873	0.890	0.877	0.021	0.901	0.054
ldpred2	auto	continuous	5	5	0.851	0.778	0.823	0.121	0.781	0.114
pt.clump	auto	continuous	5	5	0.643	0.614	0.606	0.088	0.635	0.112

For the 14 traits for which we tuned hyperparameters with CV, relative PGS effect sizes relative the ensemble PGS are shown ($\beta_s^*/\beta_{EnsPGS}^*$) and are stratified by PGS method, tuning type (CV or auto), ancestry (EUR or SAS), and type of trait (binary or continuous). The number of traits (N), medians, means and standard deviations (sd) are shown. Methods are ordered by the median EUR relative effect size within traits and tuning types.

biobank specific; FinnGen favored PRScs, whereas the UKBB and EBB had significantly larger effect sizes for LDpred2 and MegaPRS (forest plots for all three-level meta-analytic models are available in Figure S17). In contrast, effect sizes for BMI and HDL varied the least between methods.

Regarding SAS-matched target data, RA was only available in GNH with a limited number of cases (60) but displayed the highest heterogeneity of effect sizes due to method ($\tau_{method} = 0.137$, 95% CI: 0.046–0.379), consistent with the findings in EUR ancestry. T1D scores were not predictive in GNH (Figure S1) and were not evaluated in the UKBB because of the small sample size; therefore, we couldn't replicate the related findings from the EUR subset.

Discussion

With this study, we have provided a comprehensive systematic comparison of PGS methods by analyzing data from one million individuals across multiple biobanks.

By publishing our workflow, we aim to increase access to PGS methods and facilitate future research. We believe that PGS method software could be greatly improved by support for standard formats (e.g., those maintained by the GWAS Catalog and PGS Catalog[59]) alongside software containerization (containers were supported in all the research environments that contributed to this study).

Our analysis was based on a previously published framework[16] that we automated and whose application we expanded and evaluated in multiple biobanks. Recent methods explicitly tailored for diverse target populations or source GWAS[60–62] were missing in this framework, and diverse ancestries were not well represented in our target data, which provides a limitation of this study. PGS tuning was performed in one biobank (UKBB-EUR) relying largely on author-provided LD reference panels. This approach more closely resembles real-world PGS application and allowed us to harness the full sample sizes in other biobanks and ancestries to maximize statistical power and test transferability.

Please cite this article in press as: Monti et al., Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and..., The American Journal of Human Genetics (2024), https://doi.org/10.1016/j.ajhg.2024.06.003

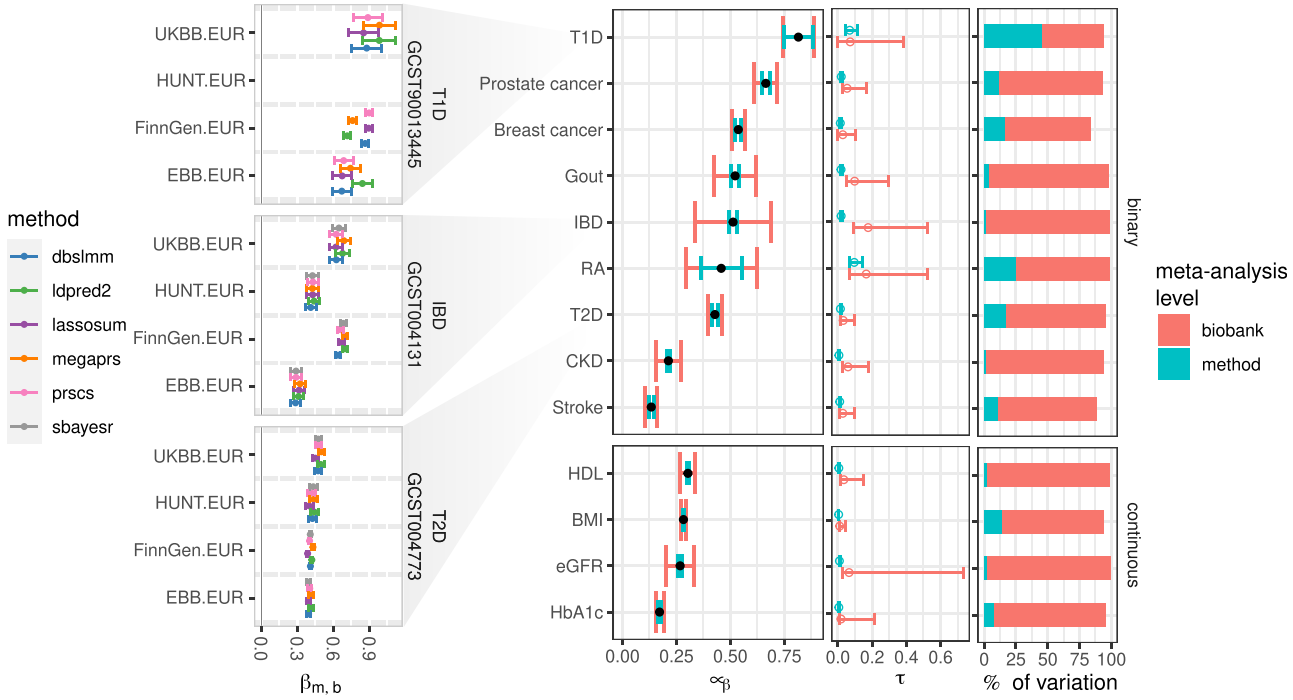


Figure 4. Three-level meta-analysis of PGS effect sizes in EUR target data
For all 13 traits replicated in at least two biobanks in EUR ancestry target data and CV-tuned in UKBB, from left to right: (1) PGS effect sizes (β coefficients, $\beta_{m,b}$) with 95% confidence intervals for three example traits within biobanks (T1D: high variability between methods, IBD: high variability between biobanks, T2D: intermediate to low variability between methods and biobanks), (2) the meta-analyzed average effect sizes across biobanks and methods (μ_β), with bars denoting the square roots of the variance components (τ), i.e., the standard deviations of the random intercepts for biobanks or methods, (3) τ-values with likelihood-based 95% confidence intervals, and (4) I^2 estimates, i.e., the fraction of variance of effect sizes explained by heterogeneity between biobanks or methods within biobanks. τ and I^2 are colored according to the levels of the meta-analytic three-level random-effects model (Methods).

Importantly, we were unable to identify a single method that consistently outperformed all others (not counting the ensemble PGS), and the two highest-performing methods (CV-tuned MegaPRS and LDpred2) were virtually tied. The best automatically tuned method was MegaPRS, albeit like other automatic methods, it suffered sporadic cases of comparatively lower performance. Which method performs best might vary based on the specifics of the GWAS summary statistics, trait, and target sample. Given that the best methods performed so similarly, other modeling choices not investigated here (for instance, the set of included variants and their availability in the target sample) could well tip the balance in favor of one or the other when researchers start from the same GWAS summary statistics. Based on our results, we recommend tuning with CV (with sufficiently large ranges of hyperparameters) instead of using methods' automatic settings, primarily to prevent cases of comparatively lower performance, rather than to provide large improvements across traits. These findings are in line with previous comparisons showing moderate gains when tuning and evaluation are performed within biobanks.[16,18]

One reason for the lower performance of automatic tuning could be model misspecification, e.g., mismatched LD references, or misreported fields in the input summary statistics. These inconsistencies might not be considered when tools are developed. The variable performance of LDpred2-auto stood out particularly against the high performance of CV-tuned PGS from same method. We note that LDpred2-auto has been updated at the time of writing and includes an optional new parameterization, which could affect its performance.[63] Limiting ourselves to the implementation of methods provided by GenoPred (which implements default method parameters) meant that we did not evaluate CV tuning for DBSLMM, which has since been recommended by the authors (performance gains over the default automatic version are about 1.13%[18]).

These cases highlight a challenge faced by any method comparison: The frequent emergence of new tools, methods, and related recommendations means that comparisons risk becoming outdated shortly after execution. Method evaluation across multiple biobanks can hardly match the pace of new developments. We therefore caution against using the results of this study to make definitive claims about relative method performance of actively developed methods. A more sustainable approach to method comparisons would be decentralized; researchers would individually submit performance estimates for published scores (starting from the same summary statistics and variants) to a central repository and receive credit when such submissions are referenced.

Table 4. Three-level meta-analytical random-effects model results (EUR)

trait	$n_{biobank}$	n_{method}	$\mu_\beta \pm$ sd	$\tau_{biobank}$ (95% CI)	τ_{method} (95% CI)	$I^2_{biobank}$ (%)	I^2_{method} (%)
T1D	3	5*	0.815 ± 0.05	0.072 (0–0.383)	0.069 (0.046–0.112)	48.8	45
Prostate cancer	4	6	0.664 ± 0.029	0.054 (0.024–0.167)	0.02 (0.015–0.029)	82.2	11
Breast cancer	4	6	0.537 ± 0.017	0.029 (0–0.1)	0.014 (0.012–0.02)	67.1	16.5
Gout	4	5*	0.522 ± 0.05	0.098 (0.049–0.294)	0.018 (0.014–0.028)	94.7	3.3
IBD	4	6	0.513 ± 0.089	0.177 (0.092–0.525)	0.019 (0.015–0.028)	97.8	1.1
RA	4	5*	0.458 ± 0.087	0.165 (0.067–0.522)	0.095 (0.068–0.144)	74.4	24.7
T2D	4	6	0.428 ± 0.017	0.031 (0.013–0.099)	0.015 (0.012–0.02)	77.8	17.2
CKD	4	6	0.213 ± 0.031	0.059 (0.028–0.18)	0.006 (0.006–0.01)	93.4	0.9
Stroke	4	6	0.133 ± 0.016	0.029 (0.01–0.099)	0.01 (0.01–0.016)	78.5	10.4
HDL	3	6	0.303 ± 0.02	0.035 (0.016–0.148)	0.005 (0.005–0.009)	96.2	2.3
BMI	3	5*	0.282 ± 0.006	0.01 (0.01–0.046)	0.004 (0.004–0.004)	80.2	13.8
eGFR	2	6	0.267 ± 0.046	0.065 (0.025–0.733)	0.009 (0.009–0.015)	97.9	1.8
HbA1c	2	6	0.172 ± 0.013	0.018 (0.011–0.211)	0.005 (0.005–0.009)	88.3	7.5

Table corresponding to Figure 4. From left to right, the target trait, the number of biobanks with the trait ($n_{biobank}$), the number of methods/scores considered (n_{method}), the meta-analyzed average PGS effect size across methods (or scores) and biobanks (μ_β) with standard deviation (SD), the standard deviation of the random intercepts specific to biobanks ($\tau_{biobank}$) and the 95% likelihood-based confidence intervals (95% CI), the standard deviation of the random intercepts specific to methods within biobanks (τ_{method}) and the 95% CI, the fraction of total effect-size variance due to heterogeneity between biobanks ($I^2_{biobank}$) as a percentage, and the fraction of the total effect size variance due to heterogeneity between methods (I^2_{method}) in %. Endpoints are ordered by type (binary or continuous) and μ_β. SBayesR was excluded for RA, T1D, gout, and BMI (*). Full results for EUR and SAS are given in Tables S9–S12.

Using meta-analytic mixed models, we found that the performances of well-tuned PGSs varied more between biobanks than within biobanks. This trend held true for most traits even when we included the covariates age, sex, and genetic PCs 1–10. This most likely reflects heterogeneity in phenotyping (e.g., disease diagnosis practices) rather than differences in population structure or genotyping. Effect sizes for BMI, which presumably is consistently measured, varied the least between biobanks, supporting this hypothesis. Yet, we cannot exclude a genetic contribution to the heterogeneity between biobanks because PGS performance has been shown to vary with the distance to the GWAS sample even within genetically similar groups matched to reference populations.[64] The variability between biobanks for some traits implies that scores need to be re-evaluated when researchers switch between different target data even during comparisons of ancestry-matched populations.

We note that the parameters by which we quantified variability are sensitive to which biobanks and PGSs are included. The setting we chose mimics the case in which multiple UKBB-EUR-validated PGSs are available. The variability between methods could increase if poorly performing (non-validated) scores are included in the analysis. On the other hand, the variability between biobanks could decrease if, e.g., phenotype definitions were further refined.

We found particularly large differences between methods for autoimmune diseases T1D and RA. This could be driven by the way methods handle the HLA region, as well as genotyping differences in the target biobanks. Our analyses highlight modeling of the HLA region as an area in which methods could potentially be improved.

One of the most useful insights from this study is that the ensemble PGS tuned in the UKBB-EUR sample provided consistently strong performance, albeit at the cost of higher computational demand during training. This shows that benefits seen within a target sample[16,18] can be transferred to other samples without re-tuning ensemble weights. We see this method as complimentary to cross-trait prediction strategies (MultiPGS)[65–67] that use PGS constructed from multiple sets of GWAS summary statistics (from different traits). Considering the small differences in performance we observed for well-tuned scores from single methods, we see ensemble PGS and MultiPGS as promising avenues for further improving PGS performances beyond what is currently possible with single methods. Future research needs to assess how well EUR-trained ensemble PGSs transfer to other genetic ancestries. It is possible that training needs to be performed in a population similar to the target population to ensure optimal performance and avoid exacerbating already existing issues with current PGSs.[7]

In summary, although no single method outperformed all others, method ensembles provided consistently strong performance (with few exceptions). PGS effect-size heterogeneity between biobanks was larger than between methods within biobanks, most likely pointing to challenges with phenotyping. Large heterogeneity between methods was observed for autoimmune diseases, indicating that special care should be taken for PGSs that rely heavily on the HLA region. Our open-source workflow, analyses framework, and

online results provide a rich ground for future method benchmarking and development.

Data and code availability

The prspipe workflow used for generating polygenic score weights and performing polygenic scoring and ancestry matching is available on GitHub (https://github.com/intervene-EU-H2020/prspipe). The list of HapMap3-1KG variants used to construct polygenic scores is available at https://github.com/intervene-EU-H2020/prspipe/blob/main/resources/1kg/1KGPhase3_hm3_hg19_hg38_mapping_cached.tsv.gz.

Non-sensitive experimental data exported from the biobanks are permissively licensed and deposited in an open data repository (https://zenodo.org/doi/10.5281/zenodo.10012995). Processed summary statistics are permissively licensed and hosted on GitHub and accessible through in an R data package (https://github.com/intervene-EU-H2020/pgsCompaR). A website containing an interactive results browser is permissively licensed and available on GitHub (https://github.com/intervene-EU-H2020/pgs-method-compare), hosted at https://methodscomparison.intervenegeneticscores.org/.

Polygenic score weights for scores that were at least nominally significantly associated with the phenotype ($p < 0.05$) in all EUR target data samples are made publicly available through the GWAS catalog (https://www.ebi.ac.uk/gwas/) with publication ID PGP000517. All evaluated scores except the one produced by LDpred2-auto for RA met this threshold. A list of PGS catalog score IDs is provided in Table S15.

Consortia

Genes & Health Research Team (in alphabetical order by surname): Shaheen Akhtar, Mohammad Anwar, Omar Asgar, Samina Ashraf, Saeed Bidi, Gerome Breen, James Broster, Raymond Chung, David Collier, Charles J Curtis, Shabana Chaudhary, Grainne Colligan, Panos Deloukas, Ceri Durham, Faiza Durrani, Fabiola Eto, Sarah Finer, Joseph Gafton, Ana Angel, Chris Griffiths, Joanne Harvey, Teng Heng, Sam Hodgson, Qin Qin Huang, Matt Hurles, Karen A Hunt, Shapna Hussain, Kamrul Islam, Vivek Iyer, Benjamin M Jacobs, Georgios Kalantzis, Ahsan Khan, Claudia Langenberg, Cath Lavery, Sang Hyuck Lee, Daniel MacArthur, Sidra Malik, Daniel Malawsky, Hilary Martin, Dan Mason, Rohini Mathur, Mohammed Bodrul Mazid, John McDermott, Caroline Morton, Bill Newman, Elizabeth Owor, Asma Qureshi, Shwetha Ramachandrappa, Mehru Raza, Jessry Russell, Nishat Safa, Miriam Samuel, Moneeza Siddiqui, Michael Simpson, John Solly, Marie Spreckley. Daniel Stow, Michael Taylor, Richard C Trembath, Karen Tricker, David A van Heel, Klaudia Walter, Caroline Winckley, Suzanne Wood, John Wright, Ishevanhu Zengeya, and Julia Zöllner.

Members of the HUNT All-In Research Team (in alphabetical order by surname): Bjørn Olav Åsvold, Ben Brumpton, Maiken Elvestad Gabrielsen, Kristian Hveem, Ida Surakka, Laurent Thomas, and Wei Zhou.

Estonian Biobank research team members are Andres Metspalu, Lili Milani, Tõnu Esko, Reedik Mägi, Mari Nelis, and Georgi Hudjashov.

The current members of FinnGen can be found in Table S17.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2024.06.003.

Acknowledgments

See supplemental information.

Author contributions

C.L. and A.G. conceptualized the study. R. Monti, S.W. and O.P. wrote the prspipe workflow. R. Monti, L.E., G.H., K.L., S.K. and B. Wolford performed analyses in biobanks. R. Monti and L.E. performed statistical meta-analyses. B. Wingfield implemented the companion website. R. Monti wrote the manuscript with assistance from all co-authors. L.E. lead revisions. L.E., R. Monti, G.H., K.L., S.K., and B.N.W. performed revisions. All authors contributed to regular discussions and provided critical feedback regarding the study design and results and contributed to review and editing of the manuscript.

Declaration of interests

M.I. is a trustee of the Public Health Genomics (PHG) Foundation, is a member of the Scientific Advisory Board of Open Targets, and has a AstraZeneca PLC research collaboration that is unrelated to this study. M.I. is supported by core funding from the British Heart Foundation (RG/18/13/33946) and NIHR Cambridge Biomedical Research Center (BRC-1215-20014; NIHR203312). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. K.L. has participated as an analyst in a collaboration research project at the Institute of Genomics, University of Tartu, which was funded by Geneto OÜ. O.P. provides consultancy services for UCB pharma company.

Received: November 20, 2023
Accepted: June 5, 2024
Published: June 21, 2024

Web resources

GenoPred, https://github.com/intervene-EU-H2020/GenoPred

References

1. Lee, A., Mavaddat, N., Wilcox, A.N., Cunningham, A.P., Carver, T., Hartley, S., Babb de Villiers, C., Izquierdo, A.,

Simard, J., Schmidt, M.K., et al. (2019). BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genet. Med. *21*, 1708–1718. https://doi.org/10.1038/s41436-018-0406-9.

2. Weale, M.E., Riveros-Mckay, F., Selzam, S., Seth, P., Moore, R., Tarran, W.A., Gradovich, E., Giner-Delgado, C., Palmer, D., Wells, D., et al. (2021). Validation of an Integrated Risk Tool, Including Polygenic Risk Score, for Atherosclerotic Cardiovascular Disease in Multiple Ethnicities and Ancestries. Am. J. Cardiol. *148*, 157–164. https://doi.org/10.1016/j.amjcard.2021.02.032.
3. Mars, N., Koskela, J.T., Ripatti, P., Kiiskinen, T.T.J., Havulinna, A.S., Lindbohm, J.V., Ahola-Olli, A., Kurki, M., Karjalainen, J., Palta, P., et al. (2020). Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat. Med. *26*, 549–557. https://doi.org/10.1038/s41591-020-0800-0.
4. Mars, N., Lindbohm, J.V., Della Briotta Parolo, P., Widén, E., Kaprio, J., Palotie, A., FinnGen, and Ripatti, S. (2022). Systematic comparison of family history and polygenic risk across 24 common diseases. Am. J. Hum. Genet. *109*, 2152–2162. https://doi.org/10.1016/j.ajhg.2022.10.009.
5. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224.
6. Lewis, C.M., and Vassos, E. (2020). Polygenic risk scores: From research tools to clinical instruments. Genome Med. *12*, 44. https://doi.org/10.1186/s13073-020-00742-5.
7. Adeyemo, A., Balaconis, M.K., Darnes, D.R., Fatumo, S., Granados Moreno, P., Hodonsky, C.J., Inouye, M., Kanai, M., Kato, K., Knoppers, B.M., et al. (2021). Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. Nat. Med. *27*, 1876–1884. https://doi.org/10.1038/s41591-021-01549-6.
8. Marston, N.A., Kamanu, F.K., Nordio, F., Gurmu, Y., Roselli, C., Sever, P.S., Pedersen, T.R., Keech, A.C., Wang, H., Lira Pineda, A., et al. (2020). Predicting Benefit From Evolocumab Therapy in Patients With Atherosclerotic Disease Using a Genetic Risk Score: Results From the FOURIER Trial. Circulation *141*, 616–623. https://doi.org/10.1161/CIRCULATIONAHA.119.043805.
9. Damask, A., Steg, P.G., Schwartz, G.G., Szarek, M., Hagström, E., Badimon, L., Chapman, M.J., Boileau, C., Tsimikas, S., Ginsberg, H.N., et al. (2020). Patients With High Genome-Wide Polygenic Risk Scores for Coronary Artery Disease May Receive Greater Clinical Benefit From Alirocumab Treatment in the ODYSSEY OUTCOMES Trial. Circulation *141*, 624–636. https://doi.org/10.1161/CIRCULATIONAHA.119.044434.
10. Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X., and Sham, P.C. (2017). Polygenic scores via penalized regression on summary statistics. Genet. Epidemiol. *41*, 469–480. https://doi.org/10.1002/gepi.22050.
11. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776. https://doi.org/10.1038/s41467-019-09718-5.
12. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun. *10*, 5086. https://doi.org/10.1038/s41467-019-12653-0.
13. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2020). LDpred2: Better, faster, stronger. Bioinformatics *36*, 5424–5431. https://doi.org/10.1093/bioinformatics/btaa1029.
14. 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. https://doi.org/10.1016/j.ajhg.2020.03.013.
15. Zhang, Q., Privé, F., Vilhjálmsson, B., and Speed, D. (2021). Improved genetic prediction of complex traits from individual-level data or summary statistics. Nat. Commun. *12*, 1–9. https://doi.org/10.1038/s41467-021-24485-y.
16. Pain, O., Glanville, K.P., Hagenaars, S.P., Selzam, S., Fürtjes, A.E., Gaspar, H.A., Coleman, J.R.I., Rimfeld, K., Breen, G., Plomin, R., et al. (2021). Evaluation of polygenic prediction methodology within a reference-standardized framework. PLoS Genet. *17*, e1009021–e1009022. https://doi.org/10.1371/journal.pgen.1009021.
17. Ni, G., Zeng, J., Revez, J.A., Wang, Y., Zheng, Z., Ge, T., Restuadi, R., Kiewa, J., Nyholt, D.R., Coleman, J.R.I., et al. (2021). A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. Biol. Psychiatr. *90*, 611–620. https://doi.org/10.1016/j.biopsych.2021.04.018.
18. Yang, S., and Zhou, X. (2022). PGS-server: accuracy, robustness and transferability of polygenic score methods for biobank scale studies. Briefings Bioinf. *23*, bbac039. https://doi.org/10.1093/bib/bbac039.
19. Jermy, B., Läll, K., Wolford, B., Wang, Y., Zguro, K., Cheng, Y., Kanai, M., Kanoni, S., Yang, Z., Hartonen, T., et al. (2023). A unified framework for estimating country-specific cumulative incidence for 18 diseases stratified by polygenic risk. Preprint at medRxiv. https://doi.org/10.1101/2023.06.12.23291186.
20. Köster, J., Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., et al. (2021). Sustainable data analysis with Snakemake. F1000Research *10*. https://doi.org/10.12688/f1000research.29032.2.
21. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z.
22. Kurki, M.I., Karjalainen, J., Palta, P., Sipilä, T.P., Kristiansson, K., Donner, K.M., Reeve, M.P., Laivuori, H., Aavikko, M., Kaunisto, M.A., et al. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. Nature *613*, 508–518. https://doi.org/10.1038/s41586-022-05473-8.
23. Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L., et al. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int. J. Epidemiol. *44*, 1137–1147. https://doi.org/10.1093/ije/dyt268.
24. Åsvold, B.O., Langhammer, A., Rehn, T.A., Kjelvik, G., Grøntvedt, T.V., Sørgjerd, E.P., Fenstad, J.S., Heggland, J., Holmen, O., Stuifbergen, M.C., et al. (2023). Cohort Profile Update: The HUNT Study, Norway. Int. J. Epidemiol. *52*, e80–e91. https://doi.org/10.1093/ije/dyac095.
25. Finer, S., Martin, H.C., Khan, A., Hunt, K.A., MacLaughlin, B., Ahmed, Z., Ashcroft, R., Durham, C., MacArthur, D.G.,

McCarthy, M.I., et al. (2020). Cohort Profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. Int. J. Epidemiol. *49*, 20–21i. https://doi.org/10.1093/ije/dyz174.

26. Robertson, C.C., Inshaw, J.R.J., Onengut-Gumuscu, S., Chen, W.-M., Santa Cruz, D.F., Yang, H., Cutler, A.J., Crouch, D.J.M., Farber, E., Bridges, S.L., et al. (2021). Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. Nat. Genet. *53*, 962–971. https://doi.org/10.1038/s41588-021-00880-5.
27. Tin, A., Marten, J., Halperin Kuhns, V.L., Li, Y., Wuttke, M., Kirsten, H., Sieber, K.B., Qiu, C., Gorski, M., Yu, Z., et al. (2019). Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. Nat. Genet. *51*, 1459–1474. https://doi.org/10.1038/s41588-019-0504-x.
28. Wheeler, E., Leong, A., Liu, C.-T., Hivert, M.-F., Strawbridge, R.J., Podmore, C., Li, M., Yao, J., Sim, X., Hong, J., et al. (2017). Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. PLoS Med. *14*, e1002383. https://doi.org/10.1371/journal.pmed.1002383.
29. Scott, R.A., Scott, L.J., Mägi, R., Marullo, L., Gaulton, K.J., Kaakinen, M., Pervjakova, N., Pers, T.H., Johnson, A.D., Eicher, J.D., et al. (2017). An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes *66*, 2888–2902. https://doi.org/10.2337/db16-1253.
30. Skene, N.G., Bryois, J., Bakken, T.E., Breen, G., Crowley, J.J., Gaspar, H.A., Giusti-Rodriguez, P., Hodge, R.D., Miller, J.A., Muñoz-Manchado, A.B., et al. (2018). Genetic identification of brain cell types underlying schizophrenia. Nat. Genet. *50*, 825–833. https://doi.org/10.1038/s41588-018-0129-5.
31. Malik, R., Chauhan, G., Traylor, M., Sargurupremraj, M., Okada, Y., Mishra, A., Rutten-Jacobs, L., Giese, A.-K., van der Laan, S.W., Gretarsdottir, S., et al. (2018). Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat. Genet. *50*, 524–537. https://doi.org/10.1038/s41588-018-0058-3.
32. Schwartzentruber, J., Cooper, S., Liu, J.Z., Barrio-Hernandez, I., Bello, E., Kumasaka, N., Young, A.M.H., Franklin, R.J.M., Johnson, T., Estrada, K., et al. (2021). Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. Nat. Genet. *53*, 392–402. https://doi.org/10.1038/s41588-020-00776-w.
33. Ha, E., Bae, S.-C., and Kim, K. (2021). Large-scale meta-analysis across East Asian and European populations updated genetic architecture and variant-driven biology of rheumatoid arthritis, identifying 11 novel susceptibility loci. Ann. Rheum. Dis. *80*, 558–565. https://doi.org/10.1136/annrheumdis-2020-219065.
34. Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. Nature *551*, 92–94. https://doi.org/10.1038/nature24284.
35. Schumacher, F.R., Al Olama, A.A., Berndt, S.I., Benlloch, S., Ahmed, M., Saunders, E.J., Dadaev, T., Leongamornlert, D., Anokian, E., Cieza-Borrella, C., et al. (2018). Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat. Genet. *50*, 928–936. https://doi.org/10.1038/s41588-018-0142-8.
36. de Lange, K.M., Moutsianas, L., Lee, J.C., Lamb, C.A., Luo, Y., Kennedy, N.A., Jostins, L., Rice, D.L., Gutierrez-Achury, J., Ji, S.-G., et al. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat. Genet. *49*, 256–261. https://doi.org/10.1038/ng.3760.
37. Wuttke, M., Li, Y., Li, M., Sieber, K.B., Feitosa, M.F., Gorski, M., Tin, A., Wang, L., Chu, A.Y., Hoppmann, A., et al. (2019). A catalog of genetic loci associated with kidney function from analyses of a million individuals. Nat. Genet. *51*, 957–972. https://doi.org/10.1038/s41588-019-0407-x.
38. Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M., et al. (2021). A cross-population atlas of genetic associations for 220 human phenotypes. Nat. Genet. *53*, 1415–1424. https://doi.org/10.1038/s41588-021-00931-x.
39. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197–206. https://doi.org/10.1038/nature14177.
40. Hoffmann, T.J., Theusch, E., Haldar, T., Ranatunga, D.K., Jorgenson, E., Medina, M.W., Kvale, M.N., Kwok, P.-Y., Schaefer, C., Krauss, R.M., et al. (2018). A large electronic-health-record-based genome-wide study of serum lipids. Nat. Genet. *50*, 401–413. https://doi.org/10.1038/s41588-018-0064-5.
41. International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58. https://doi.org/10.1038/nature09298.
42. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. *5*, R80. https://doi.org/10.1186/gb-2004-5-10-r80.
43. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590–D598. https://doi.org/10.1093/nar/gkj144.
44. World Health Organization (2004). ICD-10 : International Statistical Classification of Diseases and Related Health Problems : Tenth Revision (World Health Organization).
45. Levey, A.S., Coresh, J., Greene, T., Marsh, J., Stevens, L.A., Kusek, J.W., Van Lente, F.; and Chronic Kidney Disease Epidemiology Collaboration (2007). Expressing the modification of diet in renal disease study equation for estimating glomerular filtration rate with standardized serum creatinine values. Clin. Chem. *53*, 766–772. https://doi.org/10.1373/clinchem.2006.077180.
46. Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.
47. Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., Taylor, K.D., et al.

(2008). Long-Range LD Can Confound Genome Scans in Admixed Populations. Am. J. Hum. Genet. *83*, 132–139. https://doi.org/10.1016/j.ajhg.2008.06.005.

48. Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. J. Stat. Software *61*, 1–36. https://doi.org/10.18637/jss.v061.i06.
49. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Software *28*, 1–26. https://doi.org/10.18637/jss.v028.i05.
50. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Software *33*, 1–22. https://doi.org/10.18637/jss.v033.i01.
51. Lee, S.H., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2012). A better coefficient of determination for genetic profile analysis. Genet. Epidemiol. *36*, 214–224. https://doi.org/10.1002/gepi.21614.
52. DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics *44*, 837–845. https://doi.org/10.2307/2531595.
53. Wang, Y., Namba, S., Lopera, E., Kerminen, S., Tsuo, K., Läll, K., Kanai, M., Zhou, W., Wu, K.-H., Favé, M.-J., et al. (2023). Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. Cell Genom. *3*, 100241. https://doi.org/10.1016/j.xgen.2022.100241.
54. Ebert, M.H., Cuijpers, P., and Toshi Furukawa, D. (2021). Doing Meta-Analysis with R: A Hands-On Guide (Chapman and Hall/CRC). https://doi.org/10.1201/9781003107347.
55. Cheung, M.W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. Psychol. Methods *19*, 211–229. https://doi.org/10.1037/a0032968.
56. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901. https://doi.org/10.1093/nar/gkw1133.
57. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7–015.
58. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591. https://doi.org/10.1038/s41588-019-0379-x.
59. Lambert, S.A., Gil, L., Jupp, S., Ritchie, S.C., Xu, Y., Buniello, A., McMahon, A., Abraham, G., Chapman, M., Parkinson, H., et al. (2021). The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat. Genet. *53*, 420–425.
60. Ruan, Y., Lin, Y.-F., Feng, Y.-C.A., Chen, C.-Y., Lam, M., Guo, Z., Stanley Global Asia Initiatives, He, L., Sawa, A., Martin, A.R., et al. (2022). Improving polygenic prediction in ancestrally diverse populations. Nat. Genet. *54*, 573–580. https://doi.org/10.1038/s41588-022-01054-7.
61. Cai, M., Xiao, J., Zhang, S., Wan, X., Zhao, H., Chen, G., and Yang, C. (2021). A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. Am. J. Hum. Genet. *108*, 632–655. https://doi.org/10.1016/j.ajhg.2021.03.002.
62. Hoggart, C.J., Choi, S.W., García-González, J., Souaiaia, T., Preuss, M., and O'Reilly, P.F. (2024). BridgePRS leverages shared genetic effects across ancestries to increase polygenic risk score portability. Nat. Genet. *56*, 180–186. https://doi.org/10.1038/s41588-023-01583-9.
63. Privé, F., Albiñana, C., Arbel, J., Pasaniuc, B., and Vilhjálmsson, B.J. (2023). Inferring disease architecture and predictive ability with LDpred2-auto. Am. J. Hum. Genet. *110*, 2042–2055. https://doi.org/10.1016/j.ajhg.2023.10.010.
64. Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., Privé, F., Vilhjálmsson, B.J., Olde Loohuis, L.M., and Pasaniuc, B. (2023). Polygenic scoring accuracy varies across the genetic ancestry continuum. Nature *618*, 774–781. https://doi.org/10.1038/s41586-023-06079-4.
65. Norland, K., Schaid, D.J., and Kullo, I.J. (2024). A linear weighted combination of polygenic scores for a broad range of traits improves prediction of coronary heart disease. Eur. J. Hum. Genet. *32*, 209–214. https://doi.org/10.1038/s41431-023-01463-0.
66. Krapohl, E., Patel, H., Newhouse, S., Curtis, C.J., von Stumm, S., Dale, P.S., Zabaneh, D., Breen, G., O'Reilly, P.F., and Plomin, R. (2018). Multi-polygenic score approach to trait prediction. Mol. Psychiatr. *23*, 1368–1374. https://doi.org/10.1038/mp.2017.163.
67. Albiñana, C., Zhu, Z., Schork, A.J., Ingason, A., Aschard, H., Brikell, I., Bulik, C.M., Petersen, L.V., Agerbo, E., Grove, J., et al. (2023). Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. Nat. Commun. *14*, 4702. https://doi.org/10.1038/s41467-023-40330-w.