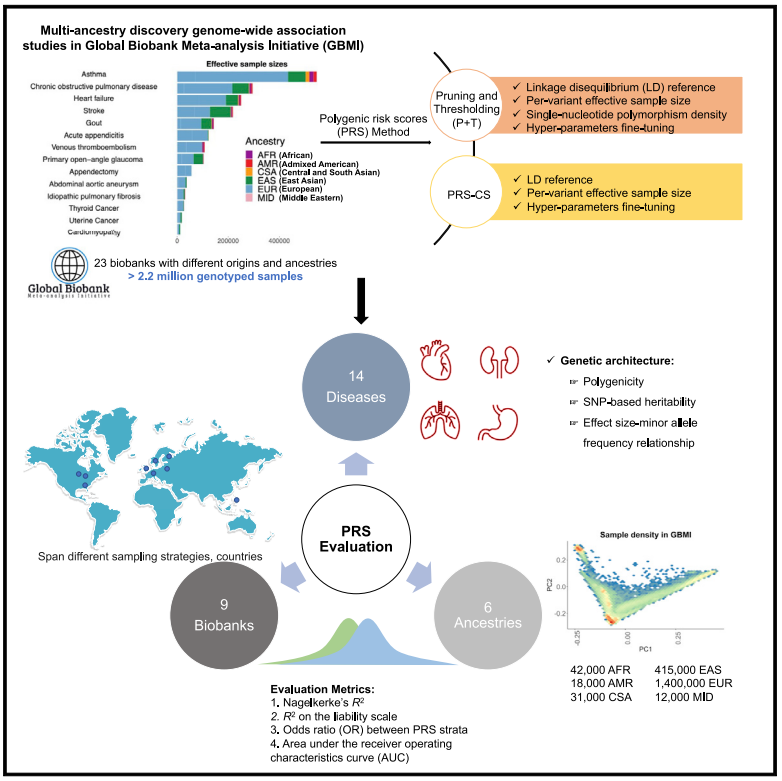
Cell Genomics

Article

Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts

Graphical abstract



Highlights

- PRS accuracy is heterogeneous across disease endpoints, ancestries, and biobanks
- Larger sample sizes and greater diversity of GBMI improves PRS accuracy
- Lessons and guidelines for developing PRS with multi-ancestry GWASs are provided

Authors

Ying Wang, Shinichi Namba, Esteban Lopera, ..., Yukinori Okada, Alicia R. Martin, Jibril Hirbo

Correspondence

yiwang@broadinstitute.org (Y.W.), armartin@broadinstitute.org (A.R.M.), jibril.hirbo@vumc.org (J.H.)

In brief

Wang et al. used the unique resource from Global Biobank Meta-analysis Initiative to develop and evaluate PRSs for 14 disease endpoints with varying genetic architectures and prevalences. They developed guidelines regarding the effects of multi-ancestry and heterogeneous GWASs, trait-specific genetic architecture, and PRS methods on prediction performance across diverse populations.

Wang et al., 2023, Cell Genomics 3, 100241
January 11, 2023
https://doi.org/10.1016/j.xgen.2022.100241

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts

Ying Wang,[1,2,30,*] Shinichi Namba,[3] Esteban Lopera,[4] Sini Kerminen,[5] Kristin Tsuo,[1,2] Kristi Läll,[6] Masahiro Kanai,[1,2,3,7] Wei Zhou,[1,2] Kuan-Han Wu,[8] Marie-Julie Favé,[9] Laxmi Bhatta,[10] Philip Awadalla,[9,11] Ben Brumpton,[10,12,13] Patrick Deelen,[4,14] Kristian Hveem,[10,12] Valeria Lo Faro,[15,16,17] Reedik Mägi,[6] Yoshinori Murakami,[18] Serena Sanna,[4,19]

(Author list continued on next page)

[1]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA
[2]Stanley Center for Psychiatric Research and Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[3]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[4]Department of Genetics, UMCG, University of Groningen, Groningen, the Netherlands
[5]Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland
[6]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia
[7]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[8]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48103, USA
[9]Ontario Institute for Cancer Research, Toronto, ON, Canada
[10]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, 7030 Trondheim, Norway
[11]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
[12]HUNT Research Centre, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, 7600 Levanger, Norway
[13]Clinic of Medicine, St. Olav's Hospital, Trondheim University Hospital, 7030 Trondheim, Norway
[14]Oncode Institute, Utrecht, the Netherlands
[15]Department of Ophthalmology, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
[16]Department of Clinical Genetics, Amsterdam University Medical Center (AMC), Amsterdam, the Netherlands
[17]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden
[18]Division of Molecular Pathology, Institute of Medical Science, the University of Tokyo, Tokyo, Japan

(Affiliations continued on next page)

SUMMARY

Polygenic risk scores (PRSs) have been widely explored in precision medicine. However, few studies have thoroughly investigated their best practices in global populations across different diseases. We here utilized data from Global Biobank Meta-analysis Initiative (GBMI) to explore methodological considerations and PRS performance in 9 different biobanks for 14 disease endpoints. Specifically, we constructed PRSs using pruning and thresholding (P + T) and PRS-continuous shrinkage (CS). For both methods, using a European-based linkage disequilibrium (LD) reference panel resulted in comparable or higher prediction accuracy compared with several other non-European-based panels. PRS-CS overall outperformed the classic P + T method, especially for endpoints with higher SNP-based heritability. Notably, prediction accuracy is heterogeneous across endpoints, biobanks, and ancestries, especially for asthma, which has known variation in disease prevalence across populations. Overall, we provide lessons for PRS construction, evaluation, and interpretation using GBMI resources and highlight the importance of best practices for PRS in the biobank-scale genomics era.

INTRODUCTION

Population- and hospital-based biobanks are increasingly coupling genomic and electronic health record data at sufficient scale to evaluate the potential of personalized medicine.[1] The growth of these paired datasets enables genome-wide association studies (GWASs) to estimate increasingly precise genetic effect sizes of variants that contribute to disease risk. In turn,

Jordan W. Smoller,[20] Jasmina Uzunovic,[9] Brooke N. Wolford,[8,10] Global Biobank Meta-analysis Initiative, Cristen Willer,[10,21,22] Eric R. Gamazon,[23,24,25] Nancy J. Cox,[23,25] Ida Surakka,[21] Yukinori Okada,[3,26,27,28] Alicia R. Martin,[1,2,29,*] and Jibril Hirbo[23,25,29,*]

[19]Institute for Genetics and Biomedical Research (IRGB), National Research Council (CNR), 09100 Cagliari, Italy
[20]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
[21]Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109, USA
[22]Department of Biostatistics and Center for Statistical Genetics, and Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA
[23]Department of Medicine, Division of Genetic Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA
[24]MRC Epidemiology Unit, University of Cambridge, Cambridge, UK
[25]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA
[26]Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
[27]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC) and Center for Infectious Disease Education and Research (CiDER), Osaka University, Suita 565-0871, Japan
[28]Department of Genome Informatics, Graduate School of Medicine, the University of Tokyo, Tokyo 113-0033, Japan
[29]These authors contributed equally
[30]Lead contact
*Correspondence: yiwang@broadinstitute.org (Y.W.), armartin@broadinstitute.org (A.R.M.), jibril.hirbo@vumc.org (J.H.)
https://doi.org/10.1016/j.xgen.2022.100241

GWAS summary statistics can be used to aggregate the effects of many genetic markers (usually in the form of single-nucleotide polymorphisms [SNPs]) to estimate individuals' genetic predispositions for complex diseases via polygenic risk scores (PRSs). As GWAS power has increased, PRS accuracy has also improved, with PRSs for some traits having comparable accuracies to independent biomarkers already routinely used in clinical risk models.[2] Consequently, several areas of medicine have already begun investigating the potential for integrating PRSs alongside other biomarkers and information currently used in clinical risk models.[3–5] However, evidence of clinical utility for PRSs across disease areas is currently limited or inconsistent.[2,6–8] Furthermore, many methods have been developed to compute PRSs, each with different strengths and weaknesses.[9–11] Thus, guidelines that delineate best practices while considering a range of real-world healthcare settings and disease areas are critically needed.

Best practices for PRSs are critical but lacking for a range of considerations that have been shown to contribute to variability in accuracy and interpretation. These include guidance for variable phenotype definitions and precision for both discovery and target populations, which varies with cohort ascertainment strategy, geography, environmental exposures, and other common covariates.[12–14] Other considerations include varying genetic architectures, statistical power of the discovery GWAS, and PRS methods, which vary in which variants are included and how weights are calculated.[9,15] A particularly pernicious issue requiring best practices is regarding maximizing generalizability of PRS accuracy among ancestry groups.[16,17] Developing best practices for PRSs therefore requires harmonized genetic data spanning diverse phenotypes, participants, and ascertainment strategies.

To facilitate the development of best practices, we evaluate several considerations for PRS in Global Biobank Meta-analysis Initiative (GBMI). GBMI brings together 23 population- and hospital-based biobanks developed in countries spanning different continents (as of April 2022). GBMI aggregates paired genetic and phenotypic data from >2.2 million individuals across diverse ancestries, including ~1.4 million Europeans (EURs); ~18,000 admixed Americans (AMRs); ~12,000 Middle Easterners (MIDs); ~31,000 Central and South Asians (CSAs); ~415,000 East Asians (EASs); and ~42,000 Africans (AFRs). Biobanks have collated phenotype information through different sources including electronic health records (EHRs), self-report data from epidemiological survey questionnaires, billing codes, doctors' narrative notes, and death registries. A detailed description of each biobank is found in Zhou et al.[18]

Here, we outline a framework for PRS analyses of multi-ancestry GWASs across multiple biobanks, as shown in Figure 1. The endpoints examined are asthma; chronic obstructive pulmonary disease (COPD); heart failure (HF); stroke; acute appendicitis (AcApp); venous thromboembolism (VTE); gout; appendectomy; primary open-angle glaucoma (POAG); uterine cancer (UtC); abdominal aortic aneurysm (AAA); idiopathic pulmonary fibrosis (IPF); thyroid cancer (ThC); and hypertrophic or obstructive cardiomyopathy (HCM), for which the phenotype definitions can be found in Zhou et al.[18] Those 14 endpoints represent the pilot effort of GBMI, which greatly vary in disease prevalence. It ranges from <1% for AAA, IPF, ThC, and HCM to ~6% for COPD and ~9% for asthma. Some endpoints (for example, appendectomy, which can be extracted from EHR procedure codes) have not been broadly studied in previous GWASs. By evaluating PRSs across 14 endpoints and 9 biobanks, we review and explore practical considerations for three steps: genetic architecture estimation, PRS method optimization and selection, and evaluation of PRS accuracy. Our framework applies to biobank-scale resources with both homogeneous and diverse ancestries.

RESULTS

The diverse ancestries included in GBMI accounted for different proportions ranging from ~76.4% for EUR, 0.1% for MID, 1.0% for AMR, 1.7% for CSA, 18.9% for EAS, and 1.8% for AFR. We

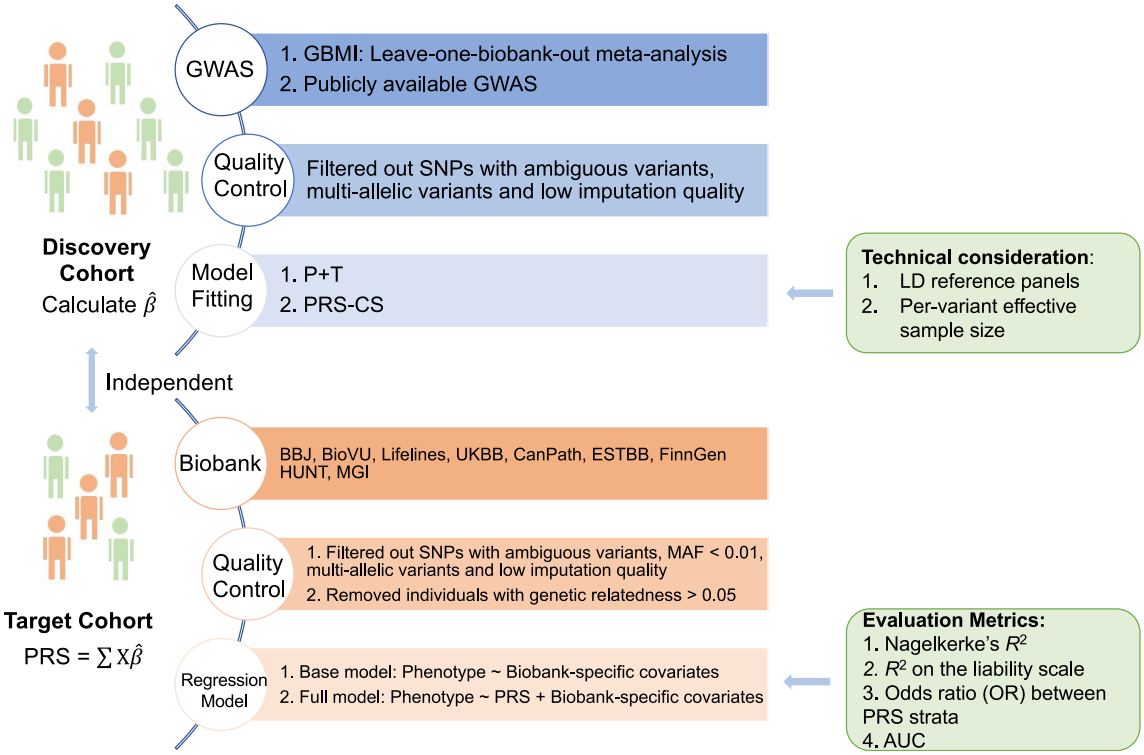


Figure 1. Overview of the study framework

explored the genetic architecture of 14 endpoints using GWAS summary statistics from all ancestries and EUR only in GBMI.[18] We used leave-one-biobank-out meta-analyzed GWASs in GBMI as our primary discovery datasets for the following PRS analyses. The ancestry compositions of discovery GWASs used in this study can be found in Table S1.

Genetic architecture of 14 endpoints in GBMI

We first estimated the genetic architecture of 14 endpoints based on HapMap3 SNPs (STAR Methods). Different prediction methods vary in which SNPs are selected and which effect sizes are assigned to them. Thus, understanding the genetic architecture of complex traits along with sample size and ancestry composition of the discovery GWAS is critical for choosing optimal prediction methods. For example, the SNP-based heritability (h^2_{SNP}) bounds PRS accuracy. We used SBayesS[19] to estimate h^2_{SNP}, polygenicity (the proportion of SNPs with nonzero effects), and the relationship between minor allele frequency (MAF) and SNP effects (i.e., a metric of negative selection, hereafter denoted as S) for the 14 endpoints. In addition to presenting results using EUR only GWAS summary statistics (EUR GWAS), we also reported estimates using meta-analysis from all ancestries on 18 biobanks (multi-ancestry GWAS). Using the attenuation ratio statistic estimated from linkage disequilibrium score regression (LDSC),[20] we found that the LD of multi-ancestry GWASs in GBMI can be reasonably approximated using the EUR-based LD reference panel (STAR Methods).

We observed that the estimates were overall higher using multi-ancestry GWASs compared with EUR GWASs (Figure 2). The SBayesS model failed to converge for HCM, likely because its estimated h^2_{SNP} was found to be not significantly different from 0 using LDSC. This could be ascribed to its known predisposing monogenic mutations, low disease prevalence, and heterogeneous subtypes.[18] Therefore, this endpoint was dropped from downstream analyses. Overall, the median estimates of polygenicity across 13 endpoints were 0.34% for multi-ancestry GWASs and 0.14% for EUR GWASs (p = 0.002, paired Wilcoxon signed-rank test), respectively. The corresponding median estimates for h^2_{SNP} were 0.051 for multi-ancestry GWASs and 0.043 for EUR GWASs (p = 0.002, paired Wilcoxon signed-rank test), respectively. The largest difference of 0.06 was found in gout. This could be due to higher h^2_{SNP} estimated in non-EUR GWASs. For example, the estimates for h^2_{SNP} using EUR and EAS GWASs was 0.051 (SE = 0.0027) and 0.088 (SE = 0.005), respectively. Moreover, we have also found that the estimated effect sizes of two gout-associated loci (close to genes *ALDH16A1* and *SLC2A9*) were different across ancestries.[18] Specifically, we observed that a few top gout-associated variants showed much higher allele frequencies in EAS compared with EUR, thus resulting in larger variance explained (Figure S1).

Polygenicity and h^2_{SNP} estimates varied greatly among different endpoints. Specifically, the h^2_{SNP} estimates were highest for asthma (h^2_{SNP} = 0.085, SE = 0.0011) and gout (h^2_{SNP} = 0.111, SE = 0.0024) using multi-ancestry GWASs, while asthma was

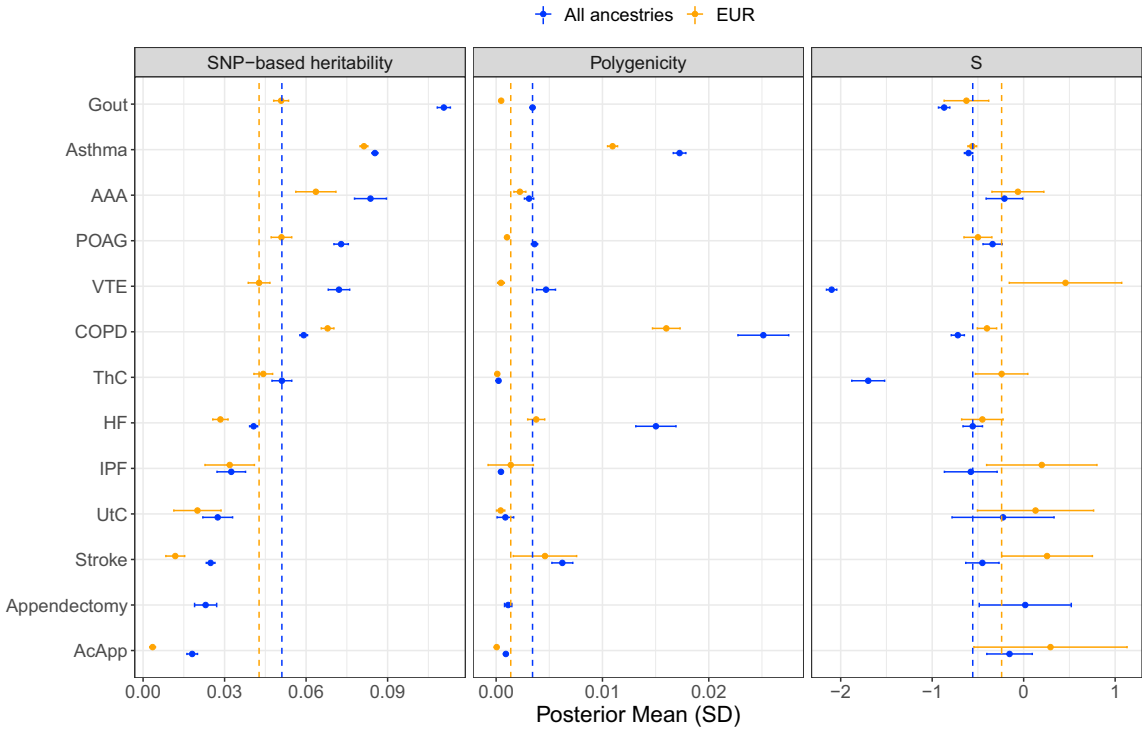


Figure 2. Genetic architecture of endpoints in GBMI

We reported the estimates from using meta-analyzed GWASs from all ancestries (All ancestries) and European only (EUR), respectively. The error bars are standard deviations (SD). Phenotypes are ranked based on SNP-based heritability on the liability scale estimates using all ancestries. The vertical dashed lines in each panel indicate the corresponding median estimates across 13 endpoints. The results for hypertrophic or obstructive cardiomyopathy are not presented. COPD, chronic obstructive pulmonary disease; HF, heart failure; AcApp, acute appendicitis; VTE, venous thromboembolism; POAG, primary open-angle glaucoma; UtC, uterine cancer; AAA, abdominal aortic aneurysm; IPF, idiopathic pulmonary fibrosis; ThC, thyroid cancer.

found to be much more polygenic than gout. We caution that the numeric interpretation of polygenicity depends on various factors and cannot be interpreted as the number of causal variants. For example, larger and more powerful GWASs tend to discover more trait-associated variants and thus appear to have higher polygenicity. Because we used the same set of SNPs in SBayesS analyses for all endpoints, we hence used the results as a relative measurement of the degree of polygenicity. We observed that the estimate of polygenicity for UtC using multi-ancestry GWASs was not statistically different from 0 (Wald test, $p > 0.05/13$) due to limited power observed as relatively low h^2_{SNP}. Overall, COPD and asthma were estimated to be the most polygenic traits, followed by HF and stroke, whereas AcApp, UtC, and ThC were the least polygenic. Lastly, we observed signals of negative selection for traits including asthma ($S = -0.56$, $SE = 0.05$), COPD ($S = -0.40$, $SE = 0.11$), and POAG ($S = -0.50$, $SE = 0.15$) when using EUR GWASs, consistent with empirical findings of negative selection explaining extreme polygenicity of complex traits.[21]

In summary, we observed largely varied key parameters of genetic architecture among 13 endpoints using multi-ancestry and EUR GWASs. We found that asthma and COPD had the highest h^2_{SNP} as well as polygenicity. We excluded HCM in our subsequent prediction analyses due to lower evidence of polygenicity and its nonsignificant h^2_{SNP}.

Optimal prediction performance using heuristic methods depends on phenotype-specific genetic architecture

We first evaluated the pruning and thresholding (P + T) method (p value thresholds ranged from 5×10^{-8} to 1) using the EUR-based LD reference panel for all 13 endpoints in the UK Biobank (UKBB)[22] and Biobank Japan (BBJ),[23] respectively, given its widespread use and relative simplicity. We further explored how different factors impact the prediction performance of P + T in diverse ancestry groups, including LD parameters (window sizes and r^2 thresholds), LD reference panels (ancestry composition, sample size, and SNP density), and per-variant effective sample size (N_{eff}) and MAF (STAR Methods).

First of all, we selected the optimal p value threshold (the p value threshold with the highest prediction accuracy, as measured by R^2 on the liability scale, $R^2_{liability}$, if not specified) in the tuning cohorts and evaluated the accuracies in the test cohorts. Specifically, we found that for UKBB with diverse ancestries, using ancestry-specific tuning cohorts provided better prediction performance compared with using EUR-based tuning cohorts (Figure S2). We found that the optimal p value threshold differed considerably between various endpoints (Figure S3; Table S2). This pattern is found to be related to polygenicity of studied endpoints, but it is also due to a combination of factors such as GWAS discovery cohort sample size, disease

prevalence, trait-specific genetic architecture, and genetic and environmental differences between discovery and target ancestries.[24] For example, when the optimal p value was determined in the UKBB-EUR subset, the less polygenic traits of ThC (106 variants) and AcApp (17 variants) showed the highest accuracy at p value thresholds of 5×10^{-5} and 5×10^{-7}, respectively, while the more polygenic traits of stroke (115,609 variants), HF (115,741 variants), asthma (7,858 variants), and COPD (29,751 variants) achieved the highest accuracy when including SNPs with p values less than 1, 1, 0.01, and 0.1, respectively. To investigate whether ancestries affect the optimal p value threshold, we replicated our analysis in the BBJ (Figure S3). In the BBJ, p value thresholds of 5×10^{-5}, 0.01, and 5×10^{-5} presented the best performance for gout, stroke, and HF, respectively. Consistent with previous studies, these results suggest that optimal prediction parameters (here, p value threshold specifically) for P + T appear to be dependent on the ancestry of the target data among other factors.[25,26] Further, we found that for more polygenic traits including asthma, COPD, stroke, and HF, prediction was more accurate with more variants in the PRS (i.e., a less significant threshold) than using the genome-wide significance threshold ($p < 5 \times 10^{-8}$). On the contrary, less polygenic traits showed no or modest improvement with less stringent p value thresholds, especially for traits such as gout, which has trait-associated SNPs with large effects. However, these trends were less obvious in the BBJ, which might be attributed to the small proportion of EAS included in the discovery GWAS. One caveat we noted was that fixed LD parameters of P + T were used; thus, the results might be impacted by additional optimization of other parameters, which we will further explore below.

We found that further optimizing LD parameters, including window size and r^2 thresholds, of P + T did not contribute to significant improvement of accuracy across endpoints. Specifically, we observed that the median accuracies with versus without LD parameter optimization were 0.018 and 0.015, respectively (Figure S4). However, there was slight but statistically significant accuracy improvement in EUR for asthma (~0.006). This might be due to more stratified signals being tagged, which results in noise reduction of the predictor. Compared with using fixed LD parameters, we found similar relationships between polygenicity and optimal p value thresholds when optimizing LD parameters in the UKBB. Specifically, the optimal p value thresholds were overall less stringent for more polygenic traits and more stringent for less polygenic traits. For example, the accuracy using LD parameter optimization in the UKBB-EUR was highest with the p value thresholds of 0.5, 1, 0.1, and 0.2 for the highly polygenic traits of stroke, HF, asthma, and COPD, respectively. In contrast, the optimal p value thresholds of 5×10^{-5} and 5×10^{-7} were observed for less polygenic traits of ThC and AcApp, respectively. To balance the computational burden and signal-to-noise ratio, we used an LD window size of 250 Kb and an LD r^2 of 0.1 as before. We repeated our analyses using genome-wide common SNPs and compared the prediction accuracy with that using HapMap3 SNPs only (Figure S4; Table S2). There were no significant improvements in prediction accuracies using a denser SNP set, which suggests that the HapMap3 SNP set represents genome-wide common SNPs well. Specifically, we found the accuracies in EUR for the most polygenic traits, asthma (~0.006), COPD (~0.005), and HF (~0.004), to be slightly improved using HapMap3 SNPs. Moreover, we found that the sample sizes of the LD reference panel had little impact on P + T performance (Figure S5), but the parameters described above including LD window sizes and LD r^2 thresholds had a larger impact on accuracy. We also showed that using EUR samples from the 1000 Genome Project[27] (1KG-EUR) as the LD reference panel performed well compared with using other ancestral populations with similar sample sizes in the 1KG dataset, which could be explained by the overrepresentation of EUR participants (~76.4%) in GBMI (Figure S6; Table S2). Similar to previous findings, we found that even in the leave-UKBB-out GWAS with the lowest EUR proportion (Table S1), its LD information can be well approximated using the EUR reference panel, which was reflected by the values of attenuation ratio not statistically larger than 0.2 and not statistically different from EUR GWASs in GBMI. We therefore used 1KG-EUR as the LD reference panel for all subsequent P + T analyses, but the choice of an external LD reference panel for multi-ancestry GWASs needs further exploration, especially when the discovery GWAS becomes more diverse.

Finally, we investigated the impact of per-variant effective sample size heterogeneity. Since GBMI consists of a number of biobanks with diverse ancestries, the number of samples used for meta-analysis was notably heterogeneous among the variants; the majority of the variants in the GWAS meta-analysis had only a limited number of N_{eff} (Figure 3A). Therefore, although sample size heterogeneity is not usually considered for PRSs, it may confound the PRS prediction accuracy in the case of global biobank collaborations. By filtering the variants according to N_{eff} per variant (i.e., N_{eff} larger than 50% or 80% thresholds of the maximum N_{eff} of the trait of interest; STAR Methods), we observed that the $R^2_{liability}$ increased substantially for less stringent thresholds ($p > 5 \times 10^{-5}$) in the UKBB (Figure S7A). As a representative example, the largest $R^2_{liability}$ (0.034) was obtained for asthma when the p value threshold was 5×10^{-3}, whereas the $R^2_{liability}$ was 6.6×10^{-3} at the threshold without N_{eff} filtering (Figure 3B; Table S3). Next, we investigated whether N_{eff} filtering could be substituted by other filtering criteria. Although excluding variants with MAFs less than 0.1 partially compensated for PRS transferability, the improvement of N_{eff} filtering in $R^2_{liability}$ was still observed (Figure S7B). Heterogeneity in N_{eff} might be confounding especially in multi-ancestry meta-analyses because it can be distorted by heterogeneous allele frequencies and imputation quality spectra among ancestries. Indeed, as rarer variants tend to be more ancestry specific, variants with low N_{eff} tend to be unique to specific ancestries (Figure 3C). Of note, the dependency of $R^2_{liability}$ on the N_{eff} was, however, largely rectified for most of the traits by using only HapMap3 SNPs (Figure S7C). Given that the $R^2_{liability}$ for HapMap3 SNPs was comparable to that for genome-wide SNPs (Figure S4), restricting to HapMap3 SNPs might be suitable for meta-analysis of diverse populations. On the other hand, HapMap3 SNPs generally have good imputation quality, although a recent study shows that relaxing imputation INFO score from 0.9 to 0.3 has negligible impacts on prediction accuracy.[9] Our findings in the BBJ mirror those in the UKBB (Figures S7D–S7F).

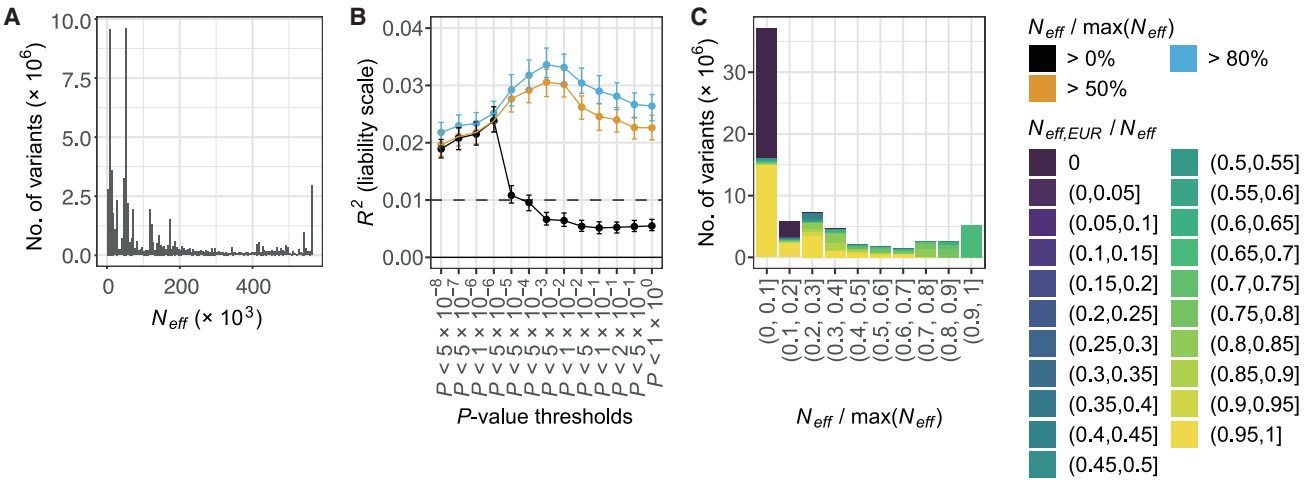


Figure 3. Sample size heterogeneity affects PRS prediction accuracy for P + T

(A) The distribution of effective sample sizes (N_{eff}).

(B) Predictive performance of P + T for EUR in the UK Biobank.

(C) The ratio of N_{eff} of EUR compared with N_{eff} of all samples. Asthma is shown as a representative result. The error bars represent the 95% confidence intervals (CIs). Full results are shown in Figure S7 and Table S3.

Overall, we found the prediction performance of P + T to be affected by a combination of factors, with p value thresholds showing larger effects compared with other parameters, such as LD window sizes, LD r^2 thresholds, and variant filtering by N_{eff} or MAF. Moreover, the optimal p value threshold varied substantially between different endpoints in GBMI. We also demonstrated that restricted use of HapMap3 SNPs showed comparable or better prediction accuracy relative to using genome-wide common SNPs for P + T, particularly for GWASs from diverse cohorts as in GBMI, with genetic variants showing considerable heterogeneity in effective sample sizes.

Bayesian approaches for calculating PRSs improve accuracy

We also evaluated fully genome-wide PRSs by first fine-tuning the parameters in PRS-continuous shrinkage (CS). We ran PRS-CS using both the grid model and automated optimization model (referred to as auto model), the former of which specifies a global shrinkage parameter (phi, in which smaller values indicate less polygenic architecture, and vice versa for larger values), with 1KG-EUR as the LD reference panel. We note that the optimized phi parameter with highest prediction accuracy in the grid model differed among traits (Figure S8). Specifically, we found that for more polygenic traits (as estimated using SBayesS) including asthma, COPD, and stroke (Figure 2), the optimal phi parameter was 1×10^{-3} in EUR (Figure S8). There was no significant difference between prediction accuracy using the optimal grid model versus the auto model (Figure S8), which suggests that PRS-CS can learn the phi parameter from discovery GWASs well when its sample size is considerably large. Therefore, we hereafter used the auto model because of its computational efficiency. Across target ancestral populations in the UKBB, PRSs from EUR-based LD reference panels showed significantly higher or comparable prediction accuracies compared with PRSs using other ancestry-based LD reference panels (Figure S9A). Note that we also compared the prediction accuracy of LD reference panels derived from UKBB-EUR, which has a much larger sample size, against 1KG-EUR and found no significant difference (Figure S9B). These results suggest that it is reasonable to use a EUR-based LD reference in GBMI and that PRS-CS is not sensitive to the sample size of the LD reference, which are consistent with previous findings.[28,29]

We then compared the optimal prediction accuracy of P + T versus the PRS-CS auto model in the UKBB and BBJ and found that PRS-CS showed overall better prediction performance for traits with higher h^2_{SNP} but no or slight improvements for traits with lower h^2_{SNP} (Figure 4). Specifically, the highest significant improvement of PRS-CS relative to that of P + T in EUR was observed for HF (60.9%), followed by COPD (53.2%) and asthma (48.8%). Substantial increments were observed for HF (105.2%), COPD (102.5%), and asthma (60.9%) in EAS. 45.8% and 48.1% improvements were shown for asthma in CSA and AFR, respectively. P + T saw better prediction performance over PRS-CS for a few trait-ancestry comparisons; however, such improvement was not statistically significant. Compared with P + T, which requires tuning p value thresholds and is affected by variant-level quality controls such as N_{eff}, there is no need to tune prediction parameters using the PRS-CS auto model, thus reducing the computational burden.

Overall, after examining 13 disease endpoints, these results favor the use of PRS-CS for developing PRS from multi-ancestry GWAS of primarily EUR samples, which is also consistent with previous findings that Bayesian methods generally show better prediction accuracy over P + T across a range of different traits.[9,28] The practical considerations about the two models, PRS-CS and P + T, used in this study are shown in Table S4.

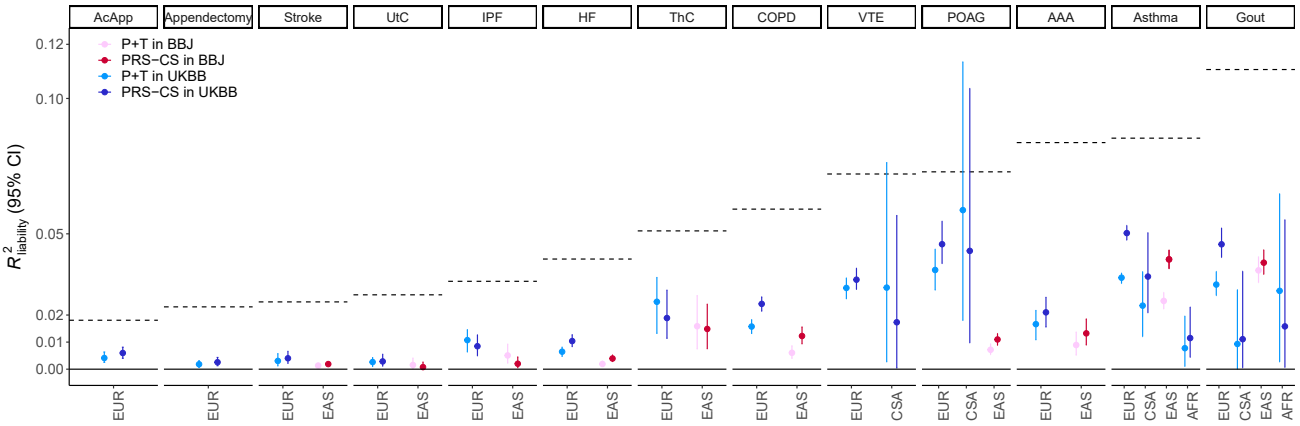


Figure 4. Prediction performance using P + T versus that using PRS-CS

Phenotypes are ranked based on SNP-based heritability (indicated by the dashed line). Only trait-ancestry pairs with significant accuracies in both P + T and PRS-CS are presented. Prediction accuracy in P + T was based on the optimal p value thresholds. The auto model was used for PRS-CS. The error bars represent the 95% confidence intervals (CIs). EUR, Europeans; CSA, Central and South Asians; EAS, East Asians; AFR, Africans; COPD, chronic obstructive pulmonary disease; HF, heart failure; AcApp, acute appendicitis; VTE, venous thromboembolism; POAG, primary open-angle glaucoma; UtC, uterine cancer; AAA, abdominal aortic aneurysm; IPF, idiopathic pulmonary fibrosis; ThC, thyroid cancer.

PRS accuracy is heterogeneous across ancestries and biobanks

For each of the participating biobanks, we used leave-one-biobank-out meta-analysis as the discovery GWAS to estimate the prediction performance of PRSs in that specific biobank. The disease prevalence and effective sample size of each biobank is shown in Figure S10. Generally, the PRS prediction accuracy of different traits increased with larger h^2_{SNP} (Figure S11;Table S5). For example, the average $R^2_{liability}$ across biobanks (hereafter denoted as $\overline{R^2_{liability}}$; STAR Methods) in EUR ranged from <1% for AcApp, appendectomy, stroke, UtC, and IP, 1% for HF, and ~2.2% for COPD and ThC to 3.8% for gout and 4.6% for asthma. Notably, accuracy was sometimes heterogeneous across biobanks within the same ancestry for some traits. Specifically, the $R^2_{liability}$ for asthma in Estonian Biobank (ESTBB)[30] and BioVU[31] was significantly lower than $\overline{R^2_{liability}}$, which might be attributable to between-biobank differences such as recruitment strategy, phenotyping, disease prevalence, and environmental factors. The prediction accuracy was generally lower in non-EUR ancestries compared with EUR ancestries, especially in AFR ancestry, which is mostly consistent with previous findings[32–34] with a few exceptions. For example, we observed comparable prediction accuracy for gout in EAS relative to that in EUR, which could be reflected by large effective sample sizes and some gout-associated SNPs with large effects exhibiting higher allele frequencies in EAS (Figure S1). For example, the MAFs of the top gout-associated SNP rs4148157 were 0.073 in 1KG-EUR and 0.25 in 1KG-EAS, respectively, and the phenotypic variance explained by that SNP in EAS (8.3%) was more than twice as high as that in EUR (3.0%). The accuracy of PRSs to predict asthma risks in AMR was found to be significantly higher than that in EUR, which could be due to the small sample size in AMR (Table S5). Thus, further validation is needed in larger AMR population cohorts.

The ability of PRSs to stratify individuals with higher disease risks was also found to be heterogeneous across biobanks and ancestries, as shown in Figure 5 and Table S6. We showed that the PRS distribution across different biobanks slightly varied. Specifically, we calculated the absolute difference of median PRSs in each decile for each endpoint between biobanks for cases and controls, separately, and found that the largest absolute differences were 0.06 and 0.21 for stroke controls and stroke cases, respectively (Figure S12). This justifies the comparison of odds ratios (ORs) in terms of relative risks. The ORs between the top 10% and bottom 10% were more heterogeneous between biobanks and also higher relative to other comparisons (e.g., top 10% vs. middle and other strata). This is consistent with previous studies where OR reported between tails of the PRS distribution is generally inflated relative to those between top-ranked PRSs and general populations.[11] We measured the variation of ORs between biobanks using the coefficient of variation of OR ($CoeffVar_{OR}$; STAR Methods). The largest $CoeffVar_{OR}$ in EUR was observed for ThC (0.46) between the top 10% and bottom 10% compared with 0.27 and 0.23 for the top 10% vs. middle and other, respectively. We recapitulated the findings using $R^2_{liability}$ that ORs were overall higher for traits with higher h^2_{SNP} and also higher in EUR than non-EUR ancestries, which is expected as the two accuracy metrics are interrelated. For example, the averaged ORs across biobanks weighted by the inverse variance in EUR (STAR Methods) for gout were 4.6, 2.4, and 2.2 for the top 10% vs. bottom 10%, middle, and other strata, separately. The corresponding estimates in EUR for stroke were 1.6, 1.3, and 1.3, respectively. Across ancestries, the average OR of asthma between the top 10% and bottom 10% ranged from 4.1 in EUR to 2.4 in AFR.

Overall, the predictive performance of PRSs measured by $R^2_{liability}$ and OR was found to be heterogeneous across ancestries. This heterogeneity was also presented across biobanks

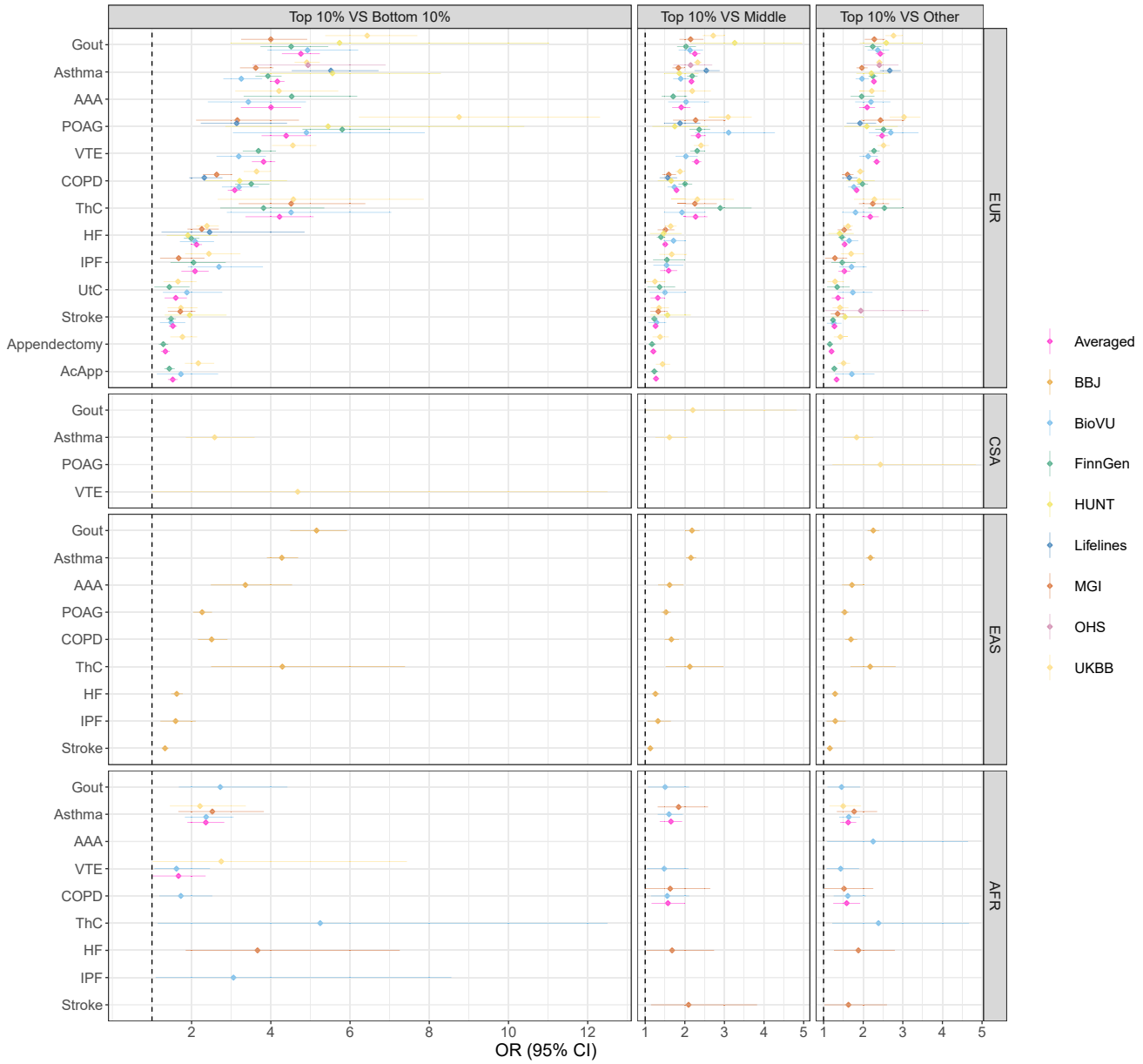


Figure 5. The odds ratio (OR) between different PRS strata for endpoints in GBMI

The dashed line indicates OR = 1. Only significant trait-ancestry specific OR was reported. The error bars represent the 95% confidence intervals (CIs). The full results are shown in Table S6. The averaged OR was calculated using the inverse-variance weighted method (STAR Methods). Phenotypes were ranked based on SNP-based heritability. EUR, Europeans; CSA, Central and South Asians; EAS, East Asians; AFR, Africans; COPD, chronic obstructive pulmonary disease; HF, heart failure; AcApp, acute appendicitis; VTE, venous thromboembolism; POAG, primary open-angle glaucoma; UtC, uterine cancer; AAA, abdominal aortic aneurysm; IPF, idiopathic pulmonary fibrosis; ThC, thyroid cancer.

for traits such as asthma, which is considered a syndrome comprising heterogeneous diseases.[35]

GBMI facilitates improved PRS accuracy compared with previous studies

GBMI resources might be expected to improve prediction accuracy due to large sample sizes and the inclusion of diverse ancestries. To explore this, we compared the prediction accuracy achieved by GBMI versus previously published GWASs using the same pipeline to run PRS-CS. As shown in Figure 6, the accuracy improvements were most obvious for traits with larger h^2_{SNP}, but there was no or slight improvement for traits with lower h^2_{SNP}. Specifically, we calculated the absolute improvement of GBMI relative to that using previously published GWASs and found that on average across biobanks, the largest improvements of $R^2_{liability}$ in EUR were 0.033 for asthma, 0.031 for gout,

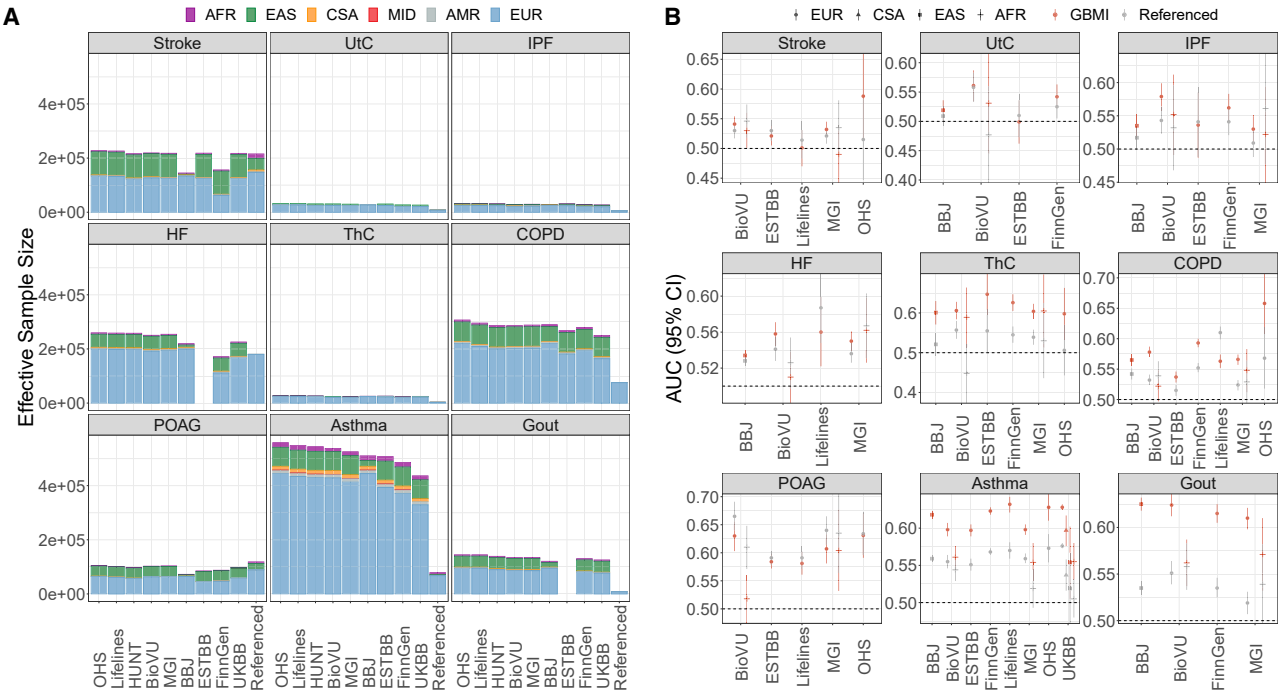


Figure 6. Prediction performance and ancestry compositions of GBMI versus previously published GWAS

(A) Ancestry compositions of GBMI and referenced GWAS. The label for biobanks in the x axis indicated the leave-one-biobank-out meta-analyzed GBMI GWAS. The previously published GWAS is labeled as referenced.

(B) Comparison of AUC between GBMI and referenced GWAS. AUC was calculated by fitting PRS only. The error bars represent the 95% confidence intervals (CIs). Full results are shown in Table S5. EUR, Europeans; CSA, Central and South Asians; EAS, East Asians; AFR, Africans; COPD, chronic obstructive pulmonary disease; HF, heart failure; AcApp, acute appendicitis; VTE, venous thromboembolism; POAG, primary open-angle glaucoma; UtC, uterine cancer; AAA, abdominal aortic aneurysm; IPF, idiopathic pulmonary fibrosis; ThC, thyroid cancer.

0.019 for ThC, and 0.017 for COPD (Figure S13), while the corresponding improvements of the area under the receiver operating characteristic curve (AUC) were 0.051, 0.078, 0.078, and 0.041, respectively. Substantial improvements were also observed for gout in EAS ($R^2_{liability}$: 0.037, AUC: 0.090); for asthma in CSA ($R^2_{liability}$: 0.026, AUC: 0.060), EAS ($R^2_{liability}$: 0.017, AUC: 0.047), and AFR ($R^2_{liability}$: 0.009, AUC: 0.034); and for ThC in EAS ($R^2_{liability}$: 0.014, AUC: 0.080) and AFR ($R^2_{liability}$: 0.016, AUC: 0.108). However, PRS accuracy was overall higher for published GWASs relative to the current GBMI for POAG in EUR and AFR and COPD in the specific case of LifeLines[36] (Table S5). We referred to the datasets included in the public GWASs of POAG and found that individuals from diverse datasets of EUR and AFR populations were also part of the discovery dataset, thus we cannot rule out the possibility of sample overlapping or relatedness between the discovery and target datasets for these populations (Table S7). Also, the phenotypes of POAG across different biobanks are likely more heterogeneous in GBMI than targeted case-control studies.[18,37] The meta-analysis of GBMI with International Glaucoma Genetics Consortium (IGGC) did not lead to substantially improved prediction performance.[37] Another concern might be the disproportional case/control ratio of POAG in GBMI (~27,000 cases and ~1.4 M controls), thus POAG-related phenotypes with shared genetics in the controls or possible uncontrolled ancestry differences between cases and controls might confound the GBMI GWASs. A very high heterogeneity for phenotype definitions is also found for COPD. However, this does not explain why one biobank alone presents this pattern; a specific environmental or population effect not considered in the broad analysis might affect this particular observation.

To boost statistical power, we can meta-analyze GBMI GWASs with other nonoverlapping cohorts as shown in other GBMI working groups.[37–39] However, we should note that more heterogeneity might be introduced from different resources such as population structure and phenotype definitions, which we cannot control with summary statistics data and that could exacerbate the heterogeneous performance of PRSs across target populations. On the other hand, GBMI is open to more cohorts and has been continuously working on integrating more datasets.

DISCUSSION

The GBMI resource is notable in its collection of phenotypes studied and its range of participating cohorts from multiple ancestry groups; it has therefore offered a unique opportunity to comprehensively evaluate and develop guidelines regarding the effects of multi-ancestry and heterogeneous GWAS discovery data, polygenicity, and PRS methods on prediction performance in diverse target cohorts. Indeed, we found overall across a range of phenotypes and ancestries that using the

large-scale meta-analysis from GBMI significantly improved PRS accuracy compared with previous studies with smaller sample sizes and less diverse cohorts. While some previous studies have benchmarked PRS methods and accuracies, most have been based on relatively homogeneous GWAS discovery cohorts or evaluated for specific phenotypes.[3,9,26,40] Even when assessing the portability of PRSs across ancestries, most evaluations have included ancestrally diverse target cohorts but still relatively homogeneous discovery cohorts.[12,13,41] Thus, based on the results of our analyses using GBMI, we have provided additional lessons and guidelines for developing PRSs with multi-ancestry discovery data for different endpoints (Figure S14). We have organized these best practices according to (1) characteristics of the discovery GWAS, (2) PRS model fitting, and (3) the target cohort.

First, GWAS discovery cohorts provide the prerequisite inputs for PRS calculations and interpretations, namely how phenotypes are ascertained and in which populations, which SNPs to include, and which effect sizes will be used. We recommend that standard quality controls should be performed with more caution when considering multi-ancestry discovery GWASs. Specifically, we suggest filtering variants based on per-variant N_{eff} and MAF as they might show considerable heterogeneity across datasets and ancestries. When we filtered out variants with extremely small N_{eff} in our P + T analyses, and in particular when using HapMap3 SNPs, PRS prediction performance improved. The allele frequencies of variants in GBMI GWASs were compared with those in Genome Aggregation Databases (gnomAD) using Mahalanobis distance and flagged if they were three standard deviations away from the mean.[18] We recommend computing such statistics and filtering with this information or, if infeasible, restricting to using only HapMap3 variants.

Given the significant improvements in PRS accuracy with GBMI discovery GWASs over previous studies with smaller sample sizes and less diversity, we recommend using the largest and most diverse GWAS discovery cohort available when constructing PRSs, even if it matches the ancestry composition of the target cohort slightly less well than a smaller GWAS. Overall, we showed here that traits with higher h^2_{SNP}, such as asthma and gout, showed greater improvement with the GBMI discovery data compared with those with lower h^2_{SNP}, such as AcApp. This indicates that PRS performance will continually benefit from larger sample sizes and more diverse populations. However, further research is needed to understand more concretely how the composition of under-represented populations, including specific ancestries and varying sample sizes, can be modeled alongside current Eurocentric GWASs to best facilitate PRS accuracy and generalizability.

Second, when fitting PRS models, important choices include which PRS methods to use, how to fine-tune hyper-parameters, and which LD reference panels to use. So far, PRS models that use GWAS summary statistics have been favored over those that use individual-level data due to their computational efficiency and data access restrictions. These models have been comprehensively reviewed recently.[10,42] We here explored the prediction performance of two widely used summary-level based PRS methods, P + T and PRS-CS. We paired the results of these methods with prior knowledge of trait-specific genetic architecture estimates from SBayesS. The best predictor for P + T is often obtained by fine-tuning the p value thresholds in a validation dataset, while other LD-related parameters, such as r^2 and window size, are usually arbitrarily specified. Here, we found that the prediction accuracy of P + T was much less sensitive to different LD-related parameters compared with various p value thresholds. Moreover, the optimal p value threshold varied across phenotypes, likely because of trait-specific genetic architecture, especially the degree of polygenicity. However, differences in discovery GWASs and target datasets such as sample sizes, phenotype definition, disease population prevalence, and population characteristics could also contribute to this variation. When analyzing PRS-CS results, we validated a previous finding that the auto model, which does not require post-hoc tuning of the phi parameter, showed similar prediction performance relative to the more computationally intensive grid model, which requires determining the optimal phi parameter in an independent tuning cohort.[28]

We also recommend using prior knowledge and empirical measurements of the genetic architecture of studied phenotypes to choose specific types of PRS models. Trait-specific architecture affected both the choice of method and optimal hyper-parameters. For example, extremely polygenic traits are more suitable for an infinitesimal model or Bayesian models that are adaptive to the trait genetic architecture. The specific model hyper-parameters are also affected by trait genetic architecture. For example, the optimal p value threshold of P + T might be more stringent for less polygenic traits but less stringent for highly polygenic traits.

Another decision point in fitting PRS models is regarding which LD reference panel to use when multi-ancestry GWAS discovery and target populations are available. An in-sample LD reference panel that spans the full discovery cohort is optimal but rarely available. Here, we have shown that EUR-based LD reference panels can reasonably approximate the LD of GBMI multi-ancestry GWASs. However, choosing LD reference panels that mirror the ancestry composition of the discovery GWAS when in-sample LD reference panels are not available is ideal. For convenience, if one ancestry is dominant in the multi-ancestry GWAS, we suggest using that ancestry-matched reference panel. The attenuation ratio statistic estimated from LDSC can further be used as a measure to quantify the degree of LD mismatch between discovery GWASs and LD reference panels.[29] When ancestry proportions are relatively evenly distributed, using LD reference panels with ancestry proportions that match the discovery GWAS could provide better prediction performance, especially for less polygenic traits with large effect variants, such as lipid traits.[43] We also found that prediction performance can be improved when using ancestry-matched tuning cohorts for PRS construction to fine-tune hyper-parameters and avoid overfitting, while other studies have also explored options, such as pseudo-validation, when no additional tuning cohort is available.[44,45]

Third, the practical considerations for target populations involved in PRS analyses are quite consistent between using homogeneous GWASs and multi-ancestry GWASs. In this study, we used biobanks with various ancestry compositions and recruitment strategies as the target cohorts.[18] For example,

BBJ, BioVU, and MGI[46] are hospital-based biobanks, whereas others are population based or have mixed enrollment strategies, which can impact phenotype precision or ascertainment bias and therefore heritability. UKBB, MGI, and BioVU have diverse ancestries, while others primarily consist of one ancestry (either EUR or EAS participants). The performance of PRSs in different target populations can also be affected by the ancestry proportions in the discovery GWAS and precision of phenotype definition aside from biobank-specific factors (e.g., environmental factors), which warrants further exploration. We therefore recommend considering those factors and reporting PRS distribution statistics (e.g., median PRS) and accuracy metrics when benchmarking the prediction performance between different PRS predictors. More reporting standards about PRS models have been well documented in the PGS Catalog.[40]

Related to the target cohorts, we also found that the prediction performance showed great heterogeneity across biobanks and ancestries. Because PRSs are only intended to capture genetic factors, other considerations such as environmental exposures and demographic history may impact the predictive power of PRSs within and across ancestries, with recommendations for how to model these alongside PRS an open question for future research and methods development. For example, we found that $R^2_{liability}$ in the Ontario Health Study (OHS)[47] was overall higher than in other biobanks, which may be attributed to the more complex relatedness structure in this founder population. Notably, the phenotype definitions, recruitment strategy, and disease prevalence also vary to different extents across the biobanks studied here.

The GBMI resource constitutes remarkable progress in expanding the number of endpoints and ancestry groups studied, laying the groundwork for several future directions for exploration. For example, PRS methods that model GWAS summary statistics alongside LD information from multiple ancestries have shown promising accuracy improvements for some traits.[16,48] But statistical methods are insufficient for equitable accuracy without simultaneous progress in generating large-scale diverse data, as early investigation into one of these methods has yielded marginal improvement in both EUR and non-EUR ancestries for asthma in GBMI.[49] In addition to multi-ancestry GWASs, sex-stratified GWASs in GBMI also provide opportunities to explore the role of sex-specific effects as well as impacts from the sample size ratio of males/females on prediction performance of PRSs across biobanks. Beyond genetic effects, biobank-specific risk factors and environmental exposures provide further opportunities to better understand the heterogeneity in PRS accuracy that we have identified across biobanks and ancestries.[50,51] This will be extremely important as previous work has shown that prediction performance differences between target cohorts are not likely to be reduced using various PRS construction methods.[9] Finally, extending these collaboration efforts to more biobanks in the future, particularly those including recently admixed populations, will bring more resolution into those effects that are biobank and ancestry specific. Studies in recently admixed populations show that GWAS power can be improved by utilizing local ancestry-specific SNP effect estimates and thus have the potential to benefit genetic prediction accuracy and generalizability, particularly for less polygenic traits.[52–54] Altogether, these initiatives hold great promise for improving transferability of PRSs across biobanks and ancestries by harnessing the phenotypic richness and diversity present in different biobanks.

Limitations of the study

We note a few limitations in our study, which also serve as a future direction. First, we chose 1KG-EUR as the LD reference panel because data security practices often preclude the use of individual-level GWAS data across analytical teams. Although we have shown that the EUR-based LD reference panels can reasonably approximate the LD of GBMI GWASs, it still could affect SNP effect size estimates and thus prediction performance. Further efforts are required to provide more appropriate LD reference panels. For example, utilizing the large-scale UKBB with individual-level genotypes to construct a panel with matched ancestry proportions to the discovery GWAS has been used in a recent study.[43] Also, sharing LD matrices from participating biobanks without accessing individual-level data would be another alternative to construct an in-sample LD matrix. On the other hand, individual-level-based PRS methods across large-scale biobanks without relying on LD reference panels are also promising. Such methods could potentially benefit from secure large-scale GWASs across multiple datasets. For example, homomorphic encryption has been used to establish a privacy-preserving framework to perform GWASs and decrypt the results for sharing through a project coordinator.[55] Second, we have focused on common SNPs, specifically HapMap3 SNPs for PRS-CS. As a result, information from rarer variants missing in the LD reference panel was not captured in other non-EUR ancestries, which may explain a small fraction of the loss of accuracy across populations. Third, although a harmonized analysis framework was developed for GBMI, there remains a multitude of factors that may contribute to heterogeneous accuracy across both biobanks and ancestries. These include, but are not limited to, phenotype precision, cohort-level disease prevalence, and environmental factors. Fourth, we evaluated PRS predictive performance using multi-ancestry GWASs, but comparisons with single-ancestry GWASs at sufficient scale would enable us to better understand the specific contributions of ancestry diversity and increase sample size, especially for under-represented ancestries. Last, we are not benchmarking against all currently available PRS methods and caution that there is no one-size-fits-all method. Instead, choice of method should depend on various factors, especially trait-specific genetic architecture as we have shown. The disease endpoints in the present study are selected as GBMI pilot efforts considering their varying prevalence and smaller efforts by previous GWAS consortia compared with other exemplar endpoints commonly studied. We have shown their distinct genetic architecture and recommended general practice accounting for this in the context of multi-ancestry GWASs.

CONSORTIA

The members of GBMI are Wei Zhou, Masahiro Kanai, Kuan-Han H. Wu, Humaira Rasheed, Kristin Tsuo, Jibril B Hirbo, Ying Wang, Arjun Bhattacharya, Huiling Zhao, Shinichi Namba, Ida Surakka, Brooke N. Wolford, Valeria Lo Faro, Esteban A. Lopera-Maya,

Kristi Läll, Marie-Julie Favé, Sinéad B. Chapman, Juha Karjalainen, Mitja Kurki, Maasha Mutaamba, Juulia J. Partanen, Ben M. Brumpton, Sameer Chavan, Tzu-Ting Chen, Michelle Daya, Yi Ding, Yen-Chen A. Feng, Christopher R. Gignoux, Sarah E. Graham, Whitney E. Hornsby, Nathan Ingold, Ruth Johnson, Triin Laisk, Kuang Lin, Jun Lv, Iona Y. Millwood, Priit Palta, Anita Pandit, Michael H. Preuss, Unnur Thorsteinsdottir, Jasmina Uzunovic, Matthew Zawistowski, Xue Zhong, Archie Campbell, Kristy Crooks, Geertruida H. de Bock, Nicholas J. Douville, Sarah Finer, Lars G. Fritsche, Christopher J. Griffiths, Yu Guo, Karen A. Hunt, Takahiro Konuma, Riccardo E. Marioni, Jansonius Nomdo, Snehal Patil, Nicholas Rafaels, Anne Richmond, Jonathan A. Shortt, Peter Straub, Ran Tao, Brett Vanderwerff, Kathleen C. Barnes, Marike Boezen, Zhengming Chen, Chia-Yen Chen, Judy Cho, George Davey Smith, Hilary K. Finucane, Lude Franke, Eric R. Gamazon, Andrea Ganna, Tom R. Gaunt, Tian Ge, Hailiang Huang, Jennifer Huffman, Jukka T. Koskela, Clara Lajonchere, Matthew H. Law, Liming Li, Cecilia M. Lindgren, Ruth J.F. Loos, Stuart MacGregor, Koichi Matsuda, Catherine M. Olsen, David J. Porteous, Jordan A. Shavit, Harold Snieder, Richard C. Trembath, Judith M. Vonk, David Whiteman, Stephen J. Wicks, Cisca Wijmenga, John Wright, Jie Zheng, Xiang Zhou, Philip Awadalla, Michael Boehnke, Nancy J. Cox, Daniel H. Geschwind, Caroline Hayward, Kristian Hveem, Eimear E. Kenny, Yen-Feng Lin, Reedik Mägi, Hilary C. Martin, Sarah E. Medland, Yukinori Okada, Aarno V. Palotie, Bogdan Pasaniuc, Serena Sanna, Jordan W. Smoller, Kari Stefansson, David A. van Heel, Robin G. Walters, Sebastian Zöllner, BBJ, BioMe, BioVU, Canadian Partnership for Tomorrow's Health/OHS, China Kadoorie Biobank Collaborative Group, Colorado Center for Personalized Medicine, deCODE Genetics, ESTBB, FinnGen, Generation Scotland, Genes & Health, LifeLines, Mass General Brigham Biobank, Michigan Genomics Initiative, QIMR Berghofer Biobank, Taiwan Biobank, The HUNT Study, UCLA ATLAS Community Health Initiative, UKBB, Alicia R. Martin, Cristen J. Willer, Mark J. Daly, and Benjamin M. Neale.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Datasets and quality control
  - Genetic architecture of 14 endpoints in GBMI
  - PRS construction
  - LD reference panel
  - Evaluation of prediction performance

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2022.100241.

ACKNOWLEDGMENTS

A.R.M. is funded by K99/R00MH117229. E.L. is funded by the Colciencias fellowship ed.783. S.N. was supported by Takeda Science Foundation. Y.O. was supported by JSPS KAKENHI (22H00476) and AMED (JP21gm4010006, JP22km0405211, JP22ek0410075, JP22km0405217, and JP22ek0109594); JST Moonshot R&D (JPMJMS2021 and JPMJMS2024); Takeda Science Foundation; and Bioinformatics Initiative of Osaka University Graduate School of Medicine, Osaka University. E.R.G. is supported by NIH awards R35HG010718, R01HG011138, and R01GM140287 and NIH/NIA AG068026. V.L.F. was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 675033 (EGRET plus). L.B. and B.B. receive support from the K.G. Jebsen Center for Genetic Epidemiology funded by Stiftelsen Kristian Gerhard Jebsen; the Faculty of Medicine and Health Sciences, NTNU; the Liaison Committee for education, research and innovation in Central Norway; and the Joint Research Committee between St. Olavs Hospital and the Faculty of Medicine and Health Sciences, NTNU. K.L. and R.M. were supported by the Estonian Research Council grant PUT (PRG687) and by INTERVENE. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 101016775. W.Z. was supported by the NHGRI of the NIH under award number T32HG010464. The work of the contributing biobanks was supported by numerous grants from governmental and charitable bodies (Data S1).

AUTHOR CONTRIBUTIONS

Study design, A.R.M., J.H., Y.O., and Y.W.; data collection/contribution, L.B., P.A., B.B., P.D., K.H., R.M., Y.M., S.S., J.U., C.W., N.J.C., I.S., and J.H.; data analysis, Y.W., S.N., E.L., S.K., K.T., K.L., M.K., W.Z., K.-H.W., M.-J.F., L.B., V.L.F., and J.H.; writing, Y.W., S.N., E.L., Y.O., A.R.M., and J.H.; revision, Y.W., S.N., E.L., K.T., W.Z., S.S., J.W.S., B.N.W., C.W., E.R.G., N.J.C., Y.O., A.R.M., and J.H.

DECLARATION OF INTERESTS

E.R.G. received an honorarium from the journal *Circulation Research of the American Heart Association* as a member of the editorial board.

Received: December 1, 2021
Revised: August 28, 2022
Accepted: December 3, 2022
Published: January 4, 2023

REFERENCES

1. Abul-Husn, N.S., and Kenny, E.E. (2019). Personalized medicine and the power of electronic health records. Cell *177*, 58–69. https://doi.org/10.1016/j.cell.2019.02.039.
2. Inouye, M., Abraham, G., Nelson, C.P., Wood, A.M., Sweeting, M.J., Dudbridge, F., Lai, F.Y., Kaptoge, S., Brozynska, M., Wang, T., et al. (2018). Genomic risk prediction of coronary artery disease in 480, 000 adults: implications for primary prevention. J. Am. Coll. Cardiol. *72*, 1883–1893. https://doi.org/10.1016/j.jacc.2018.07.079.
3. Torkamani, A., Wineinger, N.E., and Topol, E.J. (2018). The personal and clinical utility of polygenic risk scores. Nat. Rev. Genet. *19*, 581–590. https://doi.org/10.1038/s41576-018-0018-x.
4. 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.
5. Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J.P., Chen, T.-H., Wang, Q., Bolla, M.K., et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. Am. J. Hum. Genet. *104*, 21–34. https://doi.org/10.1016/j.ajhg.2018.11.002.

6. Landi, I., Kaji, D.A., Cotter, L., Van Vleck, T., Belbin, G., Preuss, M., Loos, R.J.F., Kenny, E., Glicksberg, B.S., Beckmann, N.D., et al. (2021). Prognostic value of polygenic risk scores for adults with psychosis. Nat. Med. *27*, 1576–1581. https://doi.org/10.1038/s41591-021-01475-7.

7. Dudbridge, F., Pashayan, N., and Yang, J. (2018). Predictive accuracy of combined genetic and environmental risk scores. Genet. Epidemiol. *42*, 4–19. https://doi.org/10.1002/gepi.22092.

8. Craig, J.E., Han, X., Qassim, A., Hassall, M., Cooke Bailey, J.N., Kinzy, T.G., Khawaja, A.P., An, J., Marshall, H., Gharahkhani, P., et al. (2020). Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. Nat. Genet. *52*, 160–166. https://doi.org/10.1038/s41588-019-0556-y.

9. 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. Psychiatry *90*, 611–620. https://doi.org/10.1016/j.biopsych.2021.04.018.

10. Ma, Y., and Zhou, X. (2021). Genetic prediction of complex traits with polygenic scores: a statistical review. Trends Genet. *37*, 995–1011. https://doi.org/10.1016/j.tig.2021.06.004.

11. Kulm, S., Marderstein, A., and Mezey, J. (2021). A systematic framework for assessing the clinical impact of polygenic risk scores. Preprint at medRxiv. https://doi.org/10.1101/2020.04.06.20055574.

12. Majara, L., Kalungi, A., Koen, N., Zar, H., Stein, D.J., Kinyanda, E., Atkinson, E.G., and Martin, A.R. (2021). Low generalizability of polygenic scores in African populations due to genetic and environmental diversity. Preprint at bioRxiv. https://doi.org/10.1101/2021.01.12.426453.

13. 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.

14. Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J.K., and Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. Elife *9*, e48376. https://doi.org/10.7554/eLife.48376.

15. Martin, A.R., Daly, M.J., Robinson, E.B., Hyman, S.E., and Neale, B.M. (2019). Predicting polygenic risk of psychiatric disorders. Biol. Psychiatry *86*, 97–109. https://doi.org/10.1016/j.biopsych.2018.12.015.

16. 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.

17. Weissbrod, O., Kanai, M., Shi, H., Gazal, S., Peyrot, W.J., Khera, A.V., Okada, Y., Finucane, H.K., Price, A.L., Biobank Japan Project; and Martin, A.R., et al. (2022). Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. Nat. Genet. *54*, 450–458. https://doi.org/10.1038/s41588-022-01036-9.

18. Zhou, W., Kanai, M., Wu, K.-H.H., Rasheed, H., Tsuo, K., Hirbo, J.B., Wang, Y., Bhattacharya, A., Zhao, H., Namba, S., et al. (2022). Global biobank meta-analysis initiative: powering genetic discovery across human disease. Cell Genomics *2*, 100192. https://doi.org/10.1016/j.xgen.2022.100192.

19. Zeng, J., Xue, A., Jiang, L., Lloyd-Jones, L.R., Wu, Y., Wang, H., Zheng, Z., Yengo, L., Kemper, K.E., Goddard, M.E., et al. (2021). Widespread signatures of natural selection across human complex traits and functional genomic categories. Nat. Commun. *12*, 1164. https://doi.org/10.1038/s41467-021-21446-3.

20. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium; Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295. https://doi.org/10.1038/ng.3211.

21. O'Connor, L.J., Schoech, A.P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A.L. (2019). Extreme polygenicity of complex traits is explained by negative selection. Am. J. Hum. Genet. *105*, 456–476. https://doi.org/10.1016/j.ajhg.2019.07.003.

22. 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.

23. Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., et al. (2017). Overview of the BioBank Japan project: study design and profile. J. Epidemiol. *27*, S2–S8. https://doi.org/10.1016/j.je.2016.12.005.

24. Zhang, Y., Qi, G., Park, J.-H., and Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nat. Genet. *50*, 1318–1326. https://doi.org/10.1038/s41588-018-0193-x.

25. Ware, E.B., Schmitz, L.L., Faul, J., Gard, A., Mitchell, C., Smith, J.A., Zhao, W., Weir, D., and Kardia, S.L.R. (2017). Heterogeneity in polygenic scores for common human traits. Preprint at bioRxiv. https://doi.org/10.1101/106062.

26. Choi, S.W., Mak, T.S.-H., and O'Reilly, P.F. (2020). Tutorial: a guide to performing polygenic risk score analyses. Nat. Protoc. *15*, 2759–2772. https://doi.org/10.1038/s41596-020-0353-1.

27. 1000 Genomes Project Consortium; Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.

28. 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.

29. Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A.U., Jiang, Y., Raghavan, S., et al. (2022). A saturated map of common genetic variants associated with human height. Nature *610*, 704–712. https://doi.org/10.1038/s41586-022-05275-y.

30. 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.

31. Bowton, E.A., Collier, S.P., Wang, X., Sutcliffe, C.B., Van Driest, S.L., Couch, L.J., Herrera, M., Jerome, R.N., Slebos, R.J.C., Alborn, W.E., et al. (2015). Phenotype-driven plasma biobanking strategies and methods. J. Pers. Med. *5*, 140–152. https://doi.org/10.3390/jpm5020140.

32. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum. Genet. *100*, 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004.

33. Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., and Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. Nat. Commun. *10*, 3328. https://doi.org/10.1038/s41467-019-11112-0.

34. Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P.M., and Yengo, L. (2020). Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat. Commun. *11*, 3865. https://doi.org/10.1038/s41467-020-17719-y.

35. Borish, L., and Culp, J.A. (2008). Asthma: a syndrome composed of heterogeneous diseases. Ann. Allergy Asthma Immunol. *101*, 1–8. https://doi.org/10.1016/S1081-1206(10)60826-5.

36. Scholtens, S., Smidt, N., Swertz, M.A., Bakker, S.J.L., Dotinga, A., Vonk, J.M., van Dijk, F., van Zon, S.K.R., Wijmenga, C., Wolffenbuttel, B.H.R., et al. (2015). Cohort Profile: LifeLines, a three-generation cohort study and biobank. Int. J. Epidemiol. *44*, 1172–1180. https://doi.org/10.1093/ije/dyu229.

37. Lo Faro, V., Bhattacharya, A., Zhou, W., Zhou, D., Wang, Y., Läll, K., Kanai, M., Lopera-Maya, E., Straub, P., Pawar, P., et al. (2021). Genome-wide association meta-analysis identifies novel ancestry-specific primary open-angle glaucoma loci and shared biology with vascular mechanisms and cell proliferation. Preprint at bioRxiv. https://doi.org/10.1101/2021.12.16.21267891.

38. Surakka, I., Wu, K.-H., Hornsby, W., Wolford, B.N., Shen, F., Zhou, W., Huffman, J.E., Pandit, A., Hu, Y., Brumpton, B., et al. (2022). Multi-ancestry meta-analysis identifies 2 novel loci associated with ischemic stroke and reveals heterogeneity of effects between sexes and ancestries. Preprint at bioRxiv. https://doi.org/10.1101/2022.02.28.22271647.

39. Partanen, J.J., Häppölä, P., Zhou, W., Lehisto, A.A., Ainola, M., Sutinen, E., Allen, R.J., Stockwell, A.D., Leavy, O.C., Oldham, J.M., et al. (2022). Leveraging global multi-ancestry meta-analysis in the study of idiopathic pulmonary fibrosis genetics. Cell Genomics *2*, 100181. https://doi.org/10.1016/j.xgen.2022.100181.

40. Wand, H., Lambert, S.A., Tamburro, C., Iacocca, M.A., O'Sullivan, J.W., Sillari, C., Kullo, I.J., Rowley, R., Dron, J.S., Brockman, D., et al. (2021). Improving reporting standards for polygenic scores in risk prediction studies. Nature *591*, 211–219. https://doi.org/10.1038/s41586-021-03243-6.

41. Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., O'Reilly, P.F., and Vilhjálmsson, B.J. (2022). Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. Am. J. Hum. Genet. *109*, 373. https://doi.org/10.1016/j.ajhg.2022.01.007.

42. Wang, Y., Tsuo, K., Kanai, M., Neale, B.M., and Martin, A.R. (2022). Challenges and opportunities for developing more generalizable polygenic risk scores. Annu. Rev. Biomed. Data Sci. *5*, 293–320. https://doi.org/10.1146/annurev-biodatasci-111721-074830.

43. Graham, S.E., Clarke, S.L., Wu, K.-H.H., Kanoni, S., Zajac, G.J.M., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., Winkler, T.W., et al. (2021). The power of genetic diversity in genome-wide association studies of lipids. Nature *600*, 675–679. https://doi.org/10.1038/s41586-021-04064-3.

44. 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.

45. Miao, J., Guo, H., Song, G., Zhao, Z., Hou, L., and Lu, Q. (2022). Quantifying portable genetic effects and improving cross-ancestry genetic prediction with GWAS summary statistics. Preprint at bioRxiv. https://doi.org/10.1101/2022.05.26.493528.

46. Zawistowski, M., Fritsche, L.G., Pandit, A., Vanderwerff, B., Patil, S., Schmidt, E.M., VandeHaar, P., Brummett, C.M., Keterpal, S., Zhou, X., et al. (2021). The Michigan Genomics Initiative: a biobank linking genotypes and electronic clinical records in Michigan Medicine patients. Preprint at medRxiv. https://doi.org/10.1101/2021.12.15.21267864.

47. Dummer, T.J.B., Awadalla, P., Boileau, C., Craig, C., Fortier, I., Goel, V., Hicks, J.M.T., Jacquemont, S., Knoppers, B.M., Le, N., et al. (2018). The Canadian Partnership for Tomorrow Project: a pan-Canadian platform for research on chronic disease prevention. CMAJ *190*, E710–E717. https://doi.org/10.1503/cmaj.170292.

48. Márquez-Luna, C., Loh, P.-R., and SIGMA Type 2 Diabetes Consortium; and Price, A.L.; South Asian Type 2 Diabetes SAT2D Consortium (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet. Epidemiol. *41*, 811–823. https://doi.org/10.1002/gepi.22083.

49. Tsuo, K., Zhou, W., Wang, Y., Kanai, M., Namba, S., Gupta, R., Majara, L., Nkambule, L.L., Morisaki, T., Okada, Y., et al. (2022). Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity. Cell Genomics *2*, 100212. https://doi.org/10.1016/j.xgen.2022.100212.

50. Meisner, A., Kundu, P., and Chatterjee, N. (2019). Case-only analysis of gene-environment interactions using polygenic risk scores. Am. J. Epidemiol. *188*, 2013–2020. https://doi.org/10.1093/aje/kwz175.

51. Loika, Y., Irincheeva, I., Culminskaya, I., Nazarian, A., and Kulminski, A.M. (2020). Polygenic risk scores: pleiotropy and the effect of environment. Geroscience *42*, 1635–1647. https://doi.org/10.1007/s11357-020-00203-2.

52. Atkinson, E.G., Maihofer, A.X., Kanai, M., Martin, A.R., Karczewski, K.J., Santoro, M.L., Ulirsch, J.C., Kamatani, Y., Okada, Y., Finucane, H.K., et al. (2021). Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. Nat. Genet. *53*, 195–204. https://doi.org/10.1038/s41588-020-00766-y.

53. Cavazos, T.B., and Witte, J.S. (2021). Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. HGG Adv. *2*, 100017. https://doi.org/10.1016/j.xhgg.2020.100017.

54. Marnetto, D., Pärna, K., Läll, K., Molinaro, L., Montinaro, F., Haller, T., Metspalu, M., Mägi, R., Fischer, K., and Pagani, L. (2020). Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. Nat. Commun. *11*, 1628. https://doi.org/10.1038/s41467-020-15464-w.

55. Blatt, M., Gusev, A., Polyakov, Y., and Goldwasser, S. (2020). Secure large-scale genome-wide association studies using homomorphic encryption. Proc. Natl. Acad. Sci. USA *117*, 11608–11613. https://doi.org/10.1073/pnas.1918257117.

56. 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. https://doi.org/10.1186/s13742-015-0047-8.

57. Krokstad, S., Langhammer, A., Hveem, K., Holmen, T.L., Midthjell, K., Stene, T.R., Bratberg, G., Heggland, J., and Holmen, J. (2013). Cohort profile: the HUNT study, Norway. Int. J. Epidemiol. *42*, 968–977. https://doi.org/10.1093/ije/dys095.

58. 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.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
GWAS summary statistics	Zhou et al.[18]	https://www.globalbiobankmeta.org/resources
PRS weights	This paper	https://www.pgscatalog.org/publication/PGP000262/
1000 Genome Phase 3	Auton et al.[27]	ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp
Software and algorithms		
Plink	Chang et al.[56]	https://www.cog-genomics.org/plink/
PRS-CS	Ge et al.[28]	https://github.com/getian107/PRScs
SBayesS/GCTB	Zeng et al.[19]	https://cnsgenomics.com/software/gctb/
LD score regression (LDSC)	Bulik-Sullivan et al.[20]	https://www.nature.com/articles/ng.3211
Codes for this study	This paper	https://doi.org/10.5281/zenodo.7321467

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ying Wang (yiwang@broadinstitute.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The all-biobank and ancestry-specific GWAS summary statistics are publicly available for downloading at https://www.globalbiobankmeta.org/resources and browsed at the PheWeb Browser http://results.globalbiobankmeta.org/. The PRS weights re-estimated using PRC-CS-auto for multi-ancestry GWAS including all biobanks and leave-UKBB-out multi-ancestry GWAS have been uploaded to PGS Catalog (https://www.pgscatalog.org/) under the study ID PGP000262. 1000 Genome Phase 3 data can be accessed at NCBI FTP site: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp. We used UKB data via application 31063. The software used in this study can be found at: Plink (https://www.cog-genomics.org/plink/), PRS-CS (https://github.com/getian107/PRScs), and SBayesS/GCTB (https://cnsgenomics.com/software/gctb/). The codes used in this study have been deposited to Zenodo: https://doi.org/10.5281/zenodo.7321467. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Datasets and quality control

Discovery datasets: For each of 14 endpoints, we used GWAS summary statistics from both GBMI and public datasets with summary statistics available in GWAS Catalog if applicable (Table S1 and Table S7) as the discovery dataset. We filtered out SNPs with ambiguous variants, tri- and multi-allelic variants and low imputation quality (imputation INFO score <0.3). For the GBMI discovery datasets, leave-one-biobank-out meta-analysis using the inverse-variance weighted meta-analysis strategy was applied.[18]

Target datasets: We used 9 biobanks, i.e., BioBank Japan (BBJ),[23] BioVU,[31] Lifelines,[36] UK Biobank (UKBB),[22] Ontario Health Study (OHS),[47] Estonian Biobank (ESTBB),[30] FinnGen, Michigan Genomics Initiative (MGI)[46] and Trøndelag Health Study (HUNT),[57] as the target datasets, which were independent from the datasets included in the discovery GWAS. Brief descriptions about these biobanks can be found in Zhou et al.[18] We removed individuals with genetic relatedness larger than 0.05 and applied the same filters as the discovery GWAS for SNPs. In addition, only common SNPs with MAF >1% were retained.

Genetic architecture of 14 endpoints in GBMI

SBayesS is a summary-level based method utilizing a Bayesian mixed linear model, which can report key parameters describing the genetic architecture of complex traits.[19] It only requires GWAS summary statistics and LD correlation matrix estimated from a reference panel. We ran SBayesS using the GWAS summary statistics from all 14 endpoints in GBMI, including meta-analyses on all

ancestries and on EUR only in 19 biobanks.[18] We evaluated the SNP-based heritability (h^2_{SNP}), polygenicity (proportion of SNPs with nonzero effects) and the relationship between allele frequency and SNP effects (S). We used the shrunk LD matrix (i.e., an LD matrix ignoring small LD correlations due to sampling variance) on HapMap3 SNPs provided by GCTB software. The LD matrix was constructed based on 50K European individuals from UKBB. Note that we observed inflated SNP-based heritability estimates using effective sample size for each SNP and hence used the total GWAS sample size instead. We used other default settings in the software. We calculated the p value of each parameter using Wald test to evaluate whether it was significantly different from 0. The h^2_{SNP} was further transformed into liability-scale with disease prevalence approximated as the case proportions in the GWAS summary statistics.[58]

PRS construction

P + T: P + T is used to clump quasi-independent trait-associated loci within an LD window size using a specific LD r^2 threshold. We first ran P + T in the UKBB and BBJ using an LD r^2 threshold of 0.1 and an LD window (LD_{win}) of 250Kb. We performed the analysis on both HapMap3 SNPs and genome-wide SNPs. We constructed PRS using *–score* implemented in Plink v1.9[56] using 13 different p value thresholds (5×10^{-8}, 5×10^{-7}, 1×10^{-6}, 5×10^{-6}, 5×10^{-5}, 5×10^{-4}, 5×10^{-3}, 0.01, 0.05, 0.1, 0.2, 0.5, 1). We further explored how per-variant filtering based on effective sample sizes (N_{eff}) and MAF thresholds would affect the prediction performance. We used three thresholds to retain variants by their N_{eff}: >0%, >50%, and >80% of N_{eff} compared to the total ones and also three MAF filters: 0.01, 0.05 and 0.1. In the UKBB, we also explored the impact of optimizing LD parameters on prediction performance by using different combinations of LD_{win} (250, 500, 1000, and 2000Kb) and LD r^2 thresholds (0.01, 0.02, 0.05, 0.1, 0.2, and 0.05) with the following flags: *–clump-p1 1 –clump-p2 1 –clump-r2* LD_{win} *–clump-kb* r^2 in Plink v1.9. For each population in the specific biobank, we randomly split the individuals into two even parts. One part was used as a validation cohort to fine-tune the parameters and the other part was used as the test cohort to evaluate the performance of PRS. To explore the impact of tuning cohorts on target populations with diverse ancestries such as UKBB in this study, we also used 10,000 EUR samples, not included in the discovery GWAS and independent from the test cohort, as the tuning cohort.

PRS-CS: PRS-CS[28] is a Bayesian regression framework which enables continuous shrinkage priors on SNP effects to infer their posterior mean effects. We ran PRS-CS using both the grid and auto models in the UKBB. In the grid model, we used a series of global shrinkage parameters (phi = 1×10^{-6}, 1×10^{-5}, 1×10^{-4}, 1×10^{-3}, 0.01, 0.1, 1), with lower phi values suggesting less polygenic genetic architecture and vice versa for more polygenic genetic architecture. For the auto model, PRS-CS will learn the phi parameter from the discovery GWAS without requiring post-hoc tuning. We used both total GWAS sample size and effective sample size as input for PRS-CS and found little difference, suggesting that PRS-CS is insensitive to the input of GWAS sample size. We hence used the effective sample size for subsequent analyses in this study. We used the default settings for other parameters. We generalized the auto model for all endpoints in both UKBB and BBJ. When comparing the two models, we selected the optimal phi parameter from the grid model based on the highest prediction accuracy in the target population.

LD reference panel

Both P + T and PRS-CS are summary-level based PRS prediction methods, utilizing GWAS summary statistics and an LD reference panel. To explore the impact of LD reference panels on prediction performance, we used LD reference panels of different ancestral compositions, varying sample sizes and SNP density. Specifically, we used four global ancestry groups, i.e., European (EUR), South-Asian (SAS), East-Asian (EAS) and African (AFR), from 1000G Phase 3 (1KG)[27] as LD reference panels for P + T. Further, we randomly sampled a subset of individuals with sample sizes of 500, 5000, 10,000 and 50,000 from UKBB-EUR to analyze how the sample sizes of LD reference panel would affect prediction accuracy for P + T. Moreover, we ran P + T on both the HapMap3 SNP set and a denser SNP set with genome-wide SNPs. We ran PRS-CS with the LD matrix provided by PRS-CS software,[28] which are based on both 1KG and UKBB populations from those four ancestry groups and Admixed American population (AMR). We performed those analyses using leave-UKBB-out GWAS in GBMI and evaluated the prediction performance in diverse ancestry groups in the UKBB.

To further explore how well EUR-based LD reference approximated the LD of multi-ancestry GWAS in GBMI, we ran LD score regression (LDSC) to estimate the attenuation ratio statistic.[20] The values of attenuation ratio larger than 0.2 suggest a strong LD mismatch between GWAS summary statistics and LD reference panel. We performed LDSC analyses on different GWAS, including GBMI GWAS from meta-analyses on all ancestries (multi-ancestry GWAS), EUR only and leave-one-biobank-out. We found that the ratio of LDSC using the EUR LD reference panel for GBMI multi-ancestry GWAS was not statistically larger than 0.2. Also, the values were not statistically different from those achieved using GBMI EUR GWAS. This is consistent with a previous study which has found that EUR-based LD can reasonably approximate the LD in their multi-ancestry GWAS consisting of ~75% EUR individuals.[29]

Evaluation of prediction performance

After constructing PRS, we evaluated the prediction performance in the independent target datasets. We used a logistic regression to calculate the Nagelkerke's R^2 and variance on the liability-scale explained by PRS as described previously.[58] Area under the receiver operating characteristic curve (AUC) was also reported for full models with additional covariates and models including PRS only. We used bootstrap with 1000 replicates to estimate their corresponding 95% confidence intervals (CIs). Note that the proportion of cases in each ancestry in the target dataset was approximated as the disease population prevalence. The same covariates (usually age, sex and 20 genotypic principal components, PCs) used in the GWAS analyses were included in the full regression model as

phenotype ~ PRS + covariates. We also calculated the average R^2 on the liability scale across biobanks ($\overline{R^2_{liability}}$) in each ancestry by weighting the effective sample size of each biobank for each endpoint. Further, we divided the target individuals into deciles based on the ranking of PRS distribution. We compared the odds ratio (OR) of the top decile relative to those ranked as the bottom, the middle and the remaining, when using the first decile as the referenced group. For endpoints presented in two or more biobanks, we calculated the averaged OR using the inverse variance weighted method and the coefficient of variation of OR ($CoeffVar_{OR}$) as SD(OR)/mean(OR).