ExPRSweb: An online repository with polygenic risk scores for common health-related exposures

Authors

Ying Ma, Snehal Patil, Xiang Zhou, Bhramar Mukherjee, Lars G. Fritsche

Correspondence

larsf@umich.edu

A wide range of health-related, lifestyle, and environmental variables impact disease risks; however, they are often unmeasured. By generating and analyzing polygenic scores that captured these factor's heritable component, we highlight their roles across medical phenotypes of two large biobanks and showcase the scores' ability to improve disease prediction models.

Ma et al., 2022, The American Journal of Human Genetics *109*, 1742–1760
October 6, 2022
https://doi.org/10.1016/j.ajhg.2022.09.001

CellPress

ARTICLE

ExPRSweb: An online repository with polygenic risk scores for common health-related exposures

Ying Ma,[1,2] Snehal Patil,[1,2] Xiang Zhou,[1,2,3] Bhramar Mukherjee,[1,2,3,4,5,6,8] and Lars G. Fritsche[1,2,3,5,7,8,*]

Summary

Complex traits are influenced by genetic risk factors, lifestyle, and environmental variables, so-called exposures. Some exposures, e.g., smoking or lipid levels, have common genetic modifiers identified in genome-wide association studies. Because measurements are often unfeasible, exposure polygenic risk scores (ExPRSs) offer an alternative to study the influence of exposures on various phenotypes. Here, we collected publicly available summary statistics for 28 exposures and applied four common PRS methods to generate ExPRSs in two large biobanks: the Michigan Genomics Initiative and the UK Biobank. We established ExPRSs for 27 exposures and demonstrated their applicability in phenome-wide association studies and as predictors for common chronic conditions. Especially the addition of multiple ExPRSs showed, for several chronic conditions, an improvement compared to prediction models that only included traditional, disease-focused PRSs. To facilitate follow-up studies, we share all ExPRS constructs and generated results via an online repository called ExPRSweb.

Introduction

A central challenge in genetics is to understand inherited factors underlying complex traits and disorders. Substantial efforts in the past decade, especially genome-wide association studies (GWASs), have successfully uncovered genetic variants associated with a plethora of traits.[1] However, translating these to disease etiology or to predict outcomes is not straightforward. Most genetic risk variants have weak and sparse marginal effects, accounting for only a small fraction of the phenotypic variation, even for highly heritable traits.[2–4] Consequently, incorporating information across genetic variants is necessary for assessing the predisposition of complex traits.

The construction of a polygenic risk score (PRS) is among the widely used approaches to translate genetic information into a disease risk.[5,6] A PRS is formed as a summation of an individual's risk alleles, weighted by the effect sizes obtained from an external GWAS. PRS methods rely on the polygenicity of complex traits and vary in data input, model assumptions, validation procedures, and whether functional annotations or pleiotropic information are incorporated.[7]

In addition to genetic risk factors, a wide range of health-related biomarkers, intermediate traits, lifestyle, and environmental variables—in this study broadly summarized as "exposures"—can impact disease risks. For example, high body mass index, smoking, blood lipid levels, and pre-existing type 2 diabetes (T2D) were recognized as prominent risk factors for cardiovascular disease,[8] respiratory diseases,[9] and cancers.[10,11] Given the relevance for these often modifiable risk factors for morbidity and mortality, exposure information is pivotal for precision prevention.[10] However, data on even common exposures are not always available, especially when using electronic health records (EHRs). Furthermore, data can be prone to measurement error, bias, and non-random missingness.[12,13] Yet, some exposures have a heritable component identifiable through GWASs[14,15] and thus offer the opportunity to construct exposure PRSs (ExPRSs).

As genetic proxies at the individual level, ExPRSs have been used in many applications, e.g., risk prediction and stratification,[16–18] predicting exposures,[19] instruments for Mendelian randomization analyses, or phenome-wide association studies (PheWASs).[20–23] Including ExPRSs to prediction models could improve disease diagnosis, screening, therapeutic interventions, and precision medicine approaches. PheWASs with ExPRSs may identify clinical phenotypes associated with a modifiable exposure and thereby highlight diseases whose onset might be influenced by early intervention or behavioral/lifestyle modification.[20] In contrast, ExPRSs for unmodifiable exposures, e.g., height or age at menarche, will not be amenable to individualized interventions. Of note, ExPRSs capture the genetic predisposition of an exposure assigned at birth but not the environmental influence, thus leaving a large proportion

[1]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA; [2]Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA; [3]Center for Precision Health Data Science, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA; [4]Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA; [5]University of Michigan Rogel Cancer Center, University of Michigan, Ann Arbor, MI 48109, USA; [6]Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI 48109, USA
[7]Present address: Department of Biostatistics, School of Public Health University of Michigan, 1415 Washington Heights, SPH Tower Room 4636, Ann Arbor, MI 48109, USA
[8]These authors contributed equally
*Correspondence: larsf@umich.edu
https://doi.org/10.1016/j.ajhg.2022.09.001.

of the exposure's variance unexplained. Still, the identification of associations between diseases and ExPRSs may help to tease apart the interplay of genetic and environmental pathways through which they influence disease risk.

The emerging utility of PRSs is evidenced via the accumulation of more than 1,000 PRS-related articles indexed in PubMed since 2009[24] and spurred by significant advances in PRS methods.[7] Despite the rise in popularity, their transition into clinical settings is often limited by lack of transparency, compatibility, and reproducibility across cohorts. Therefore, an ExPRS resource that integrates adequate information for constructing, evaluating, and utilizing ExPRSs to accelerate ExPRS-related research is desirable and necessary. Recently, we established "Cancer PRSweb," an interactive, online repository with cancer PRSs for 35 common cancer traits.[20] Building upon our previous work, we present ExPRSweb, a uniform analytic framework and an extension of PRSweb that specifically focusses on ExPRSs for 28 common exposures.

By using available exposure GWAS summary statistics and two large biobanks, the Michigan Genomics Initiative (MGI) and the UK Biobank (UKB), we generated ExPRSs with four methods varying in complexity and modeling (i.e., linkage disequilibrium clumping and p value thresholding [C + T], Lassosum, deterministic Bayesian sparse linear mixed model [DBSLMM], and PRS-CS, a Bayesian method with continuous shrinkage priors).[25–30] We also highlight ExPRS applications including PheWAS, risk stratification, and prediction of common chronic conditions. For the latter, we evaluated the predictive performance of single and multiple ExPRSs when combined with disease-specific PRSs and could show substantial improvement for several traits. We also contrasted these predictors with "poly-exposure risk scores" (PXSs), which integrate multiple measured exposures. In absence of high-quality exposure data on many individuals, ExPRSs can serve as surrogates if one has genotype data on a larger and more representative sample. Our repository ExPRSweb unlocks access to over 300 ExPRSs for 27 different exposures and facilitates scientific collaboration to strengthen their future application.

Subjects and methods

Michigan Genomics Initiative (MGI)

MGI cohort

Adult participants aged between 18 and 101 years at enrollment were recruited through the Michigan Medicine health system between 2012 and 2020. Participants have consented to allow research on both their biospecimens and EHR data as well as linking their EHR data to national data sources such as medical and pharmaceutical claims data. Participants were primarily recruited through the MGI - Anesthesiology Collection Effort (n = 51,160) while awaiting a diagnostic or interventional procedure either at a preoperative appointment or on the day of their operative procedure at Michigan Medicine. Additional participants were recruited through the Michigan Predictive Activity and Clinical Trajectories (MIPACT, n = 2,685) Study, the Mental Health Biobank (MHB2, n = 617), and the Michigan Genomics Initiative-Metabolism, Endocrinology, and Diabetes (MGI-MEND, n = 2,522) Study. The data used in this study included diagnoses coded with the Ninth and Tenth Revision of the International Statistical Classification of Diseases (ICD9 and ICD10) with clinical modifications (ICD9-CM and ICD10-CM), laboratory measurements, anthropometrics (height, thinness, and body mass index [BMI]), vitals (systolic and diastolic blood pressure [SBP and DBP, respectively]), health behavior (alcohol amount, smoker, and drinker), sex, precomputed principal components (PCs), genotyping batch, recruitment study, and age. Data were collected according to the Declaration of Helsinki principles. MGI study participants' consent forms and protocols were reviewed and approved by the University of Michigan Medical School Institutional Review Board (IRB ID HUM00099605 and HUM00155849). Opt-in written informed consent was obtained. Additional details about MGI can be found online (see web resources).

MGI genotype data

DNA from 56,984 blood samples was genotyped on customized Illumina Infinium CoreExome-24 bead arrays and subjected to various quality control filters, resulting in a set of 502,255 polymorphic variants. PCs and European/non-European ancestry were estimated by projecting all genotyped samples into the space of the PCs of the Human Genome Diversity Project reference panel with PLINK (938 individuals).[31,32] To further characterize inferred non-European ancestry individuals, we used 938 unrelated individuals of the Human Genome Diversity Panel (HGDP) as reference panel for ADMIXTURE (v1.3.0) to estimate for each non-European MGI individual their ancestry fraction of African (AFR), Central/South Asian (CSA), East Asian (EAS), European (EUR), Native American (AMR), Oceanian (OCE), or West Asian (WAS) ancestral HGDP continental populations.[33] We used majority global ancestry, the largest ancestry fraction, to define additional non-EUR ancestry groups (AFR, AMR, CSA, EAS, and WAS); no individual with majority OCE ancestry was found (details can be found elsewhere[34]). We assessed pairwise kinship with the software KING,[35] and we used the software FastIndep to reduce each ancestry group to a maximal subset that contained no pairs of individuals with third-or-closer degree relationship.[36] We removed participants without diagnosis data. The main analytical sample included 46,782 EUR individuals, while additional auxiliary samples (non-EUR samples with $n \geq 500$) included 3,012 AFR, 919 EAS, and 606 CSA individuals. The remaining non-EUR samples AMR and WAS had fewer than 500 individuals and were not included in any analyses. Additional genotypes were obtained with the Haplotype Reference Consortium reference panel of the Michigan Imputation Server[37] and included over 24 million imputed variants with $R^2 \geq 0.3$ and minor allele frequency (MAF) $\geq 0.01\%$.

MGI phenome

The MGI phenome was based on ICD9-CM and ICD10-CM code data for 46,782 unrelated, genotyped individuals of recent European ancestry. Longitudinal time-stamped diagnoses were recoded to indicators for whether a patient ever had given a diagnosis code recorded by Michigan Medicine. These ICD9-CM and ICD10-CM codes were aggregated to form up to 1,814 PheCodes with the PheWAS R package. In short, ICD codes that map to a phenotype concept (PheCode) were used as inclusion criteria for cases, while individuals whose ICD codes map to a set of related PheCodes were excluded as controls. Gender-specific exclusions were applied

if necessary. All remaining individuals were considered as controls (further details are described elsewhere[20,37,38]). To minimize differences in age and sex distributions, avoid extreme case-control ratios, and reduce the computational burden, we matched up to ten controls to each case by using the R package "MatchIt."[38] Nearest neighbor matching was applied for age and the first four PCs of the genotype data (PC1–4) via Mahalanobis distance with a caliper/width of 0.25 standard deviations. Exact matching was applied for sex and genotyping array. A total of 1,685 case-control studies with >50 cases were used for our analyses of the MGI phenome.

MGI common chronic conditions

We used the CCW Condition Algorithms (rev. 02/2021) from the CMS Chronic Condition Warehouse (CCW; see web resources) to define 27 common chronic conditions in MGI. In short, like the PheCode system, the CCW algorithms are based on ICD-9-CM- and ICD-10-CM-based inclusion and exclusion criteria. Here, we were interested in any observation of such conditions and disregarded the algorithms' stated reference period or the required numbers/types of qualifying claims for Medicare or Medicaid. The resulting 27 case-control studies were labeled CCW01–CCW27 and are listed in Table S12.

UK Biobank (UKB) cohort

UKB cohort

UKB is a population-based cohort collected from multiple sites across the United Kingdom and includes over 500,000 participants aged between 40 and 69 years when recruited in 2006–2010.[39] The open-access UK Biobank data used in this study included genotypes, ICD9 and ICD10 codes, biomarker data, anthropometrics, vitals, women's health, health behavior, inferred sex, inferred White British ancestry, kinship estimates down to third degree, birth year, genotype array, and precomputed PCs of the genotypes. UK Biobank received ethical approval from the NHS National Research Ethics Service North West (11/NW/0382).

UKB genotype data

We used the UK Biobank imputed dataset (v3) and limited analyses to the documented 408,595 White British[40] individuals and 47,836,001 variants with imputation information score $\geq$ 0.3 and MAF $\geq$ 0.01%, of which 22,933,317 overlapped with the imputed MGI data (see above). Two random subsets of 5,000 and 10,000 unrelated White British individuals were used for linkage disequilibrium (LD) analyses of UKB-based summary statistics. Genotyping, quality control, and imputation are described in detail elsewhere.[41]

UKB phenome

The UK Biobank phenome was based on ICD9 and ICD10 code data of 408,595 White British,[40] genotyped individuals that were similarly aggregated to PheCodes as MGI (see above, also described elsewhere[42]). In contrast to MGI, there were many pairwise relationships reported for UKB participants.

To retain a larger effective sample size for each phenotype, we first selected a maximal set of unrelated cases for each phenotype (defined as no pairwise relationship of third degree or closer[36,43]) before selecting a maximal set of unrelated controls unrelated to these cases. Similar to MGI, we matched up to ten controls to each case by using the R package "MatchIt."[38] Nearest neighbor matching was applied for birth year (as proxy for age because age at diagnosis was not available to us) and PC1-4 (Mahalanobis-metric matching; matching window caliper/width of 0.25 standard deviations), and exact matching was applied for sex and genotyping array. A total of 1,419 matched case-control studies with >50 cases each were used for our analyses of the UK Biobank phenome.

Exposure data

For a set of 21 continuous and seven binary exposures for which we could find freely available and complete GWAS summary statistics (see exposure GWAS summary statistics below), we extracted the corresponding EHR data as described in Table S1. For the binary exposures that are common disorders (type 2 diabetes, hypertension, insomnia, and sleep apnea), we use the PheWAS code-based definitions (see MGI phenome and UKB phenome above; Table S7). Survey-based measures with multiple responses per person (never/past/current alcohol use and smoking status) were recoded to never/ever responses. For continuous exposures, we removed outliers by using the 1.5× interquartile range (IQR) rule, i.e., we removed measurements outside 1.5 times the IQR above the upper quartile and below the lower quartile of the exposure's distribution in the cohort. After removing outliers, we used the mean of any remaining multiple measurements per person. We found that only using the median without outlier removal was insufficient to reduce the impact of potential outliers. For the UKB cohort, we calculated the estimated glomerular filtration rate (eGFR) on the natural scale by using the harmonized serum creatinine values (data field 30700), race and sex information, and the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation.[44]

Exposure GWAS summary statistics

For each of the 28 exposures, we collected complete GWAS summary statistics from up to four different sources: (1) catalogued GWASs of the NHGRI EBI GWAS Catalog,[1] (2) GWASs from the FinnGen Consortium, (3) published GWAS meta-analyses, and (4) publicly available GWAS summary statistics of phenome × genome screening efforts of the UK Biobank data (Lee and Neale Lab, see Table S2 and web resources). We only included GWAS summary statistics of studies that analyzed broad European ancestry to match the ancestry of discovery GWASs and target cohorts (MGI and UKB).

If needed, we lifted over coordinates of GWAS summary statistics to human genome assembly GRCh37 (LiftOver, UCSC Genome Browser Store, see web resources). Entries with missing effect alleles or effect sizes were excluded. If effect allele frequency (EAF) was reported in the summary statistics, we also compared EAF between the discovery GWAS and the target dataset (MGI and/or UKB). If the proportion of likely flipped alleles (whose EAF deviated more than 15% between the datasets) was above 40%, we excluded the GWAS as source for PRS construction. These chosen thresholds were subjective and based on clear differentiation between correct and likely flipped alleles on the two diagonals, as noted frequently in GWAS meta-analyses quality control procedures.

Statistical methods

Heritability estimation

For each set of GWAS summary statistics from both UK Biobank and non-UK Biobank sources (e.g., FinnGen, GWAS catalog, large meta-analyses), we first estimated the SNP heritability to estimate the proportion of phenotypic variance explained by all measured SNPs based on summary statistics. The estimated SNP heritability represents the upper limit for the prediction performance of PRS methods and serves as an initial filtering criterion to validate the quality of the downloaded summary statistics. To do so, we applied the method MQS (MinQue for summary statistics), which was implemented in Gemma, to calculate the SNP heritability

estimate (see web resources).[45,46] MQS estimates the SNP heritability based on the minimal norm quadratic unbiased estimation (MINQUE)[47,48] criterion. Specifically, we first converted the p values into marginal Z scores, and then we used the Z scores as well as 5,000 randomly selected, unrelated samples (reference panel) as input to run Gemma. Finally, we obtained the proportion of variance in phenotypes explained (PVE) estimates from Gemma, which corresponds to the SNP heritability estimate. We further filtered out the summary statistics that had negative heritability estimates.

For binary traits with potentially ascertained case-control data, we converted the heritability estimates from the observed scale to the liability scale by using the R package "PDRohde/ugnome" and reported population prevalence estimates (Table S3).[49]

ExPRS construction

We constructed the PRS for an individual j in the form $PRS_j = \sum_i \beta_i G_{ij}$ where i indexes the included variants for that trait, weight β_i is the log odds ratios retrieved from the external GWAS summary statistics for variant i, and G_{ij} is a continuous version of the measured dosage data for the risk allele of variant i in subject j. To construct a PRS, one must determine which genetic loci to include in the PRS and their relative weights. We have obtained GWAS summary statistics from several external sources, resulting in several sets of weights for each trait of interest. For each set of weights obtained from GWAS summary statistics from the above-mentioned sources and each trait, we generated for each exposure GWAS up to five different PRSs reflecting the five implementations of four different PRS methods: the C + T (both, best guess genotype [GT] and dosage [DS] version),[25–27] lassosum,[28] DBSLMM,[29] and PRS-CS[30] (Figure 1).

We summarized the statistical aspects of these construction methods in Table S22. The goal of this approach was to compare multiple PRS methods and find the method that works best for the various types of GWAS summary statistics.

LD clumping and p value thresholding (C + T). We performed linkage disequilibrium (LD) clumping/pruning of variants with p values below 0.1 by using the imputed allele dosages of 10,000 randomly selected samples and a pairwise correlation cut-off at $r^2 < 0.1$ within a 1 Mb window. We constructed many different PRSs across a fine grid of p value thresholds. We used the p value threshold with the highest pseudo-R^2 (binary trait) or highest R^2 (continuous traits) (see PRS evaluation below) to define the optimized "Clumping and Thresholding (C and S)" PRS. We applied two approaches for LD clumping: C + T (GT) and C + T (DS). Specifically, the "C + T (GT)" is implemented by plink-1.9 with the best-guess genotypes (GT, imputed genotype dosages are rounded to the next integer) for LD calculations, while "C + T (DS)" is implemented in R and considers the uncertainty of imputed genotypes by using the dosage data (DS).

Lassosum. Lassosum obtains PRS weights by applying elastic net penalization to GWAS summary statistics and incorporating LD information from a reference panel. Here, we used 5,000 randomly selected, unrelated samples as the LD reference panel. We applied an MAF filter of 1% and, in contrast to the previous two approaches, only included autosomal variants that overlap between summary statistics, LD reference panel, and target panel. Each "Lassosum" run resulted in up to 76 combinations of the elastic net tuning parameters s and λ, and consequently, in 76 SNP sets with corresponding weights used to construct. We then selected the PRS with the pseudo-R^2 (binary trait) or highest R^2 (continuous traits) to define the "Lassosum" PRS (see PRS evaluation below).

Deterministic Bayesian sparse linear mixed model (DBSLMM). DBSLMM assumes that the true SNP effect sizes derive from a mixture of normal distributions and relies on an efficient deterministic search algorithm for statistical inference. DBSLMM requires both GWAS summary statistics and LD information from a reference panel. Specifically, DBSLMM first selects SNPs with large effect in a deterministic fashion through the C + T procedure and then directly obtains both large SNP effect sizes and small SNP effect sizes through analytic forms. Here, we used 5,000 randomly selected unrelated samples as the LD reference panel. We applied an MAF filter of 1% and only included autosomal variants that overlap between summary statistics, LD reference panel, and target panel. Heritability estimates obtained from Gemma (see above-mentioned procedure) were used as the input of DBSLMM. All other parameters we used are the default parameters in the "DBSLMM" software. For example, we set the cutoff of SNPs clumping and pruning to be $r^2 < 0.1$ within a 1 Mb window and p value $< 1 \times 10^{-6}$, respectively. Each DBSLMM run resulted in one SNP set with corresponding weights to construct the PRS. We used the default version of DBSLMM, which does not require cross-validation and refer to the obtained PRS as "DBSLMM" PRS.

PRS-CS. PRS-CS utilizes a Bayesian regression framework and assumes a continuous shrinkage (CS) prior on the effect sizes. Specifically, we applied the default "auto" version of PRS-CS that obtain weights through the Gibbs sampling algorithm. Here, PRS-CS-auto uses a precomputed LD reference panel based on external European samples of the 1000 Genomes Project ("EUR reference") to construct a PRS. We applied an MAF filter of 1% and only included autosomal variants that overlap between summary statistics, LD reference panel, and target panel. The obtained PRS is referred to as "PRS-CS" PRS.

For each trait and set of GWAS summary statistics, these approaches usually resulted in up to five PRSs. However, approaches that resulted in less than five weights/variants were excluded. Using the R package "Rprs" (see web resources), the value of each PRS was then calculated for each MGI participant and, if the GWAS source to the best of our knowledge did not include UKB samples, also for each UKB participant. For comparability of association effect sizes corresponding to the continuous PRS across exposures and PRS construction methods, we centered PRS values in MGI and UKB to a mean of 0 and scaled them to have a standard deviation of 1.

ExPRS evaluation

To assess the predictive performance of these generated PRSs, each PRS was assessed through cross-validation in either the MGI cohort or the UKB cohort: we split the data corresponding to each trait in training (50% of the samples with gender ratio unchanged) and test set (50% of the samples with gender ratio unchanged). We used the training set to determine the PRS-tuning parameter(s) and used the testing set to obtain performance metric for that PRS.

For the PRS evaluations, except for when computing the pseudo-R^2 for binary exposures (which is a measure of marginal association of the ExPRS with the exposure),[50] we fit the following model for each PRS and exposure adjusting for covariates:

$$g(E(\text{Exposure}|\text{PRS}, \text{Age}, \text{Sex}, \text{Array}, \text{PCs})) = \beta_0 + \beta_{PRS}\text{PRS} + \beta_{Age}\text{Age} + \beta_{Sex}\text{Sex} + \beta_{Array}\text{Array} + \beta_{PCs}\text{PCs}, \quad \text{(Equation 1)}$$

where $g(\cdot)$ is the link function (e.g., identity link function for continuous traits and logit link function for binary traits). PCs were the first four principal components obtained from the

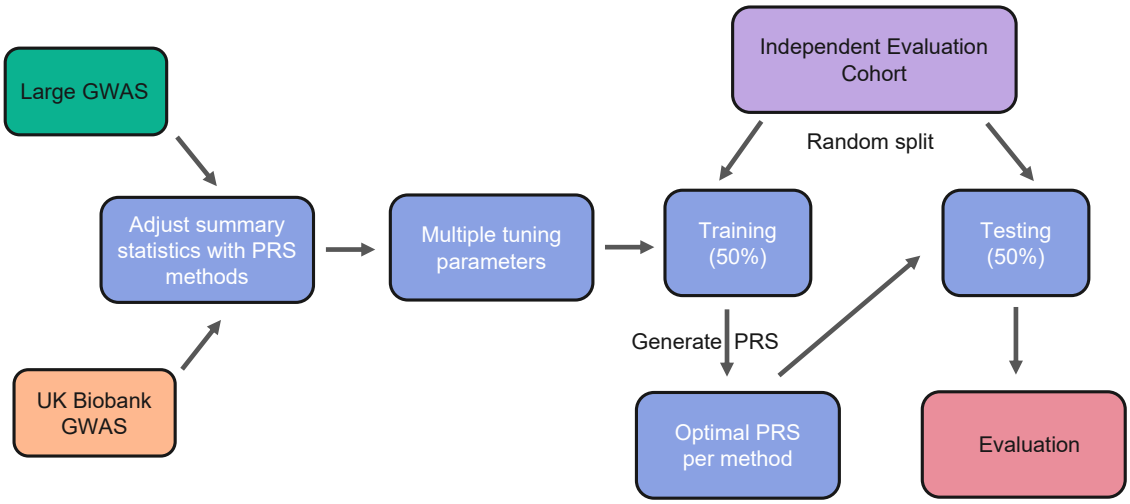

Figure 1. Flow chart of ExPRS construction, evaluation, and selection

principal-component analysis of the genotyped GWAS markers, where "Age" was the age at last observed diagnosis in MGI and birth year in UKB and where "Array" represents the genotyping array.

Binary traits. We used Nagelkerke's pseudo-R^2 to select the tuning parameters within the "C + T" and Lassosum construction methods (p value for "C + T" SNP sets; s and λ for Lassosum) and kept the PRS with the highest pseudo-R^2 for further analyses. For each PRS derived for each GWAS source/method combination, we assessed the following performance measures relative to observed disease status in MGI and UKB: (1) overall performance with Nagelkerke's pseudo-R^2 by using R package "rcompanion," (2) accuracy with Brier score by using R package "DescTools," and (3) ability to discriminate between cases and controls as measured by the area under the covariate-adjusted receiver operating characteristic (AROC; semiparametric frequentist inference) curve (denoted AAUC) by using R package "ROCnReg." Covariate-adjusted AUC (AAUC) is a weighted average of the areas under covariate-adjusted receiver operating characteristic (ROC) curves over the distribution of covariates in the study sample. In contrast to conventional AUC, the AAUC considers the covariates information to measure the classification accuracy. In addition, AAUC is context dependent because the calculation of the weights relies on the covariates' distributions. For example, even if the AUC (x) is the same in MGI and UKB cohorts, the AAUCs will be different because the distribution of x will be different as a result of data cohorts' particular covariate constitutions. We used Firth's bias reduction method to resolve the problem of separation in logistic regression (R package "brglm2").

Continuous traits. For the PRS evaluations of continuous traits, we used R^2 to select the tuning parameters within the "C + T" and Lassosum construction methods (p value for "C + T" SNP sets; s and λ for Lassosum) and kept the PRS with the highest R^2 for further analyses. For each PRS derived for each GWAS source/method combination, we assessed the prediction performance in terms of R^2 in MGI and UKB.

ExPRS primary association with the underlying exposure

Next, we assessed the strength of the relationship between these PRSs and the traits they were designed for. To do this, we fit the same model as Equation 1. Our primary interest is β_{PRS}, while the other factors (age, sex, and PCs) were included to address potential residual confounding. We used Firth's bias reduction method to resolve the problem of separation in logistic regression (Logistf in R package "EHR"). As an initial filtering step, we removed PRSs that were not significantly associated with their corresponding exposure in MGI or UKB cohorts ($p > 0.05$) for downstream analysis. The majority of these filtered PRSs were either based on discovery GWASs with small sample sizes that often did not identify any genome-wide significant hits or were evaluated for exposure with small sample sizes or both, indicating a potential lack of power in our analysis.

Illustrative examples showcasing the use of ExPRSs

Once we select the ExPRSs that were mostly and positively associated with the specific exposure, referred to as the best performing PRSs, we use these selected PRSs for various analyses to illustrate how a user may gainfully use these constructs in understanding disease etiology and mechanisms.

Phenome-wide exploration of ExPRS associations

We conducted PheWASs in MGI and UKB (if the GWAS source was not based on UKB) to identify phenotypes associated with the ExPRS. To evaluate ExPRS-phenotype associations, we conducted Firth bias-corrected logistic regression by fitting the following model for each ExPRS and each phenotype of the corresponding phenome.

$$\text{logit}(\text{P}(\text{Phecode is present}|\text{ExPRS},\ \text{Age}, \text{Sex}, \text{Array}, \text{PCs})) = \beta_0 + \beta_{ExPRS}\text{ExPRS} + \beta_{Age}\text{Age} + \beta_{Sex}\text{Sex} + \beta_{Array}\text{Array} + \beta_{PCs}\text{PCs}. \quad \text{(Equation 2)}$$

To adjust for multiple testing, we applied the conservative phenome-wide Bonferroni correction according to the total number of analyzed PheCodes (MGI: 1,685 phenotypes; UKB: 1,419 phenotypes, as described in Table S7). In Manhattan plots, we present –log10 (p value) corresponding to tests for association of the underlying phenotype with the ExPRS. Directional triangles on the PheWAS plot indicate whether a trait was positively (pointing up) or negatively (pointing down) associated with the ExPRS.

To investigate the possibility that the secondary trait associations with ExPRSs were completely driven by the exposure or extremes of the trait distribution, we performed a second set of PheWASs: for binary exposures, we excluded individuals with the binary exposures for which the ExPRS was constructed; for continuous exposures, we excluded individuals with measurements

outside of the normal range (Table S1). We referred to these PheWASs as "exclusion-PRS-PheWASs," as described previously.[41]

To evaluate whether the constructed ExPRS is a good proxy for the corresponding exposure, we also repeated the PheWAS by using the exposure or normal range exposure as the predictor instead. We referred to these PheWASs as "trait-PheWASs" and "exclusion-trait-PheWASs," respectively.

Utilities of ExPRSs on common chronic conditions

To investigate the utility of our constructed ExPRSs in predicting common chronic conditions in the MGI cohort (see MGI common chronic conditions above, Table S12), we first split the common chronic conditions into training (50% of the samples with gender ratio unchanged) and test set (50% of the samples with gender ratio unchanged). We conducted Firth bias-corrected logistic regression by fitting the following model for each of the best performing ExPRSs and each common chronic condition:

$$\text{logit}(P(\text{a common chronic condition is present}|\text{ExPRS}, \text{Age}, \text{Sex}, \text{Array}, \text{PCs})) = \beta_0 + \beta_{ExPRS}\text{ExPRS} + \beta_{Age}\text{Age} + \beta_{Sex}\text{Sex} + \beta_{Array}\text{Array} + \beta_{PCs}\text{PCs}. \quad \text{(Equation 3)}$$

Prediction performance was measured by Nagelkerke's pseudo-R^2, Brier score, and AAUC. Then we repeated the analysis by using the actual exposure as predictor to be trained and evaluated in the MGI cohort.

Next, we selected for each chronic condition the ExPRSs that reached nominal significance in the univariate model and performed clumping ($r < 0.5$). For each chronic condition, we combined the resulting sets of their associated ExPRSs by fitting a logistic regression in the training set to obtain the linear predictors that we defined as "multiExPRS" in the testing data.

To investigate whether such a multiExPRS can be helpful in predicting a common chronic condition "Y" beyond the condition-specific PRS "YPRS" (e.g., breast cancer PRS), we collected the YPRSs from public resources, except for type 2 diabetes and hypertension, for which we generated ExPRSs. More specifically, for type 2 diabetes (T2D) the T2D-PRS was used as the YPRS but never as the ExPRS, while for other conditions it was considered as the ExPRS. The same approach was applied for hypertension and the hypertension PRS. We downloaded PRS constructs/weights for lung cancer, prostate cancer, colorectal cancer, and breast cancer PRSs from Cancer PRSweb[51] and downloaded the following PRS weight from the PGS Catalog[24] (see web resources): stroke/transient ischemic attack, heart failure, glaucoma, chronic kidney disease, atrial fibrillation, and asthma PRSs. We harmonized the downloaded PRS weights to GRCh37/hg19 and determined overlap with the MGI genotype data. Non-ambiguous SNP alleles were flipped to the genomic plus strand. We fit three logistic models for each common chronic condition "Y" by using the following predictors adjusting for the set of covariates from above: (1) condition-specific PRS, "YPRS"; (2) the combined ExPRS, "multiExPRS"; and (3) "multiExPRS + YPRS." As before, we combined multiple predictors fitting a logistic regression in the training set to obtain the linear predictors that we used as combined score in the testing data. Our main interest is the comparison is between (2) and (3) because it tries to evaluate whether a multipleExPRS can improve prediction models beyond the condition-specific YPRS.

To study the ability of these three predictors to enrich patients for these chronic conditions, we binned the predictors according to their distribution in controls and compared the enrichment of cases in the three top bins "≤5%", "5%–10%", "10%–25%" (each coded as 1) versus the "40%–60%" (coded as 0) by using the multi-variate logistic model.

Poly-exposure score construction and comparison

To contrast the predictive power of a poly-exposure score (PXS) with combined ExPRSs (multiExPRSs, see above), we extracted the collected/measured exposure data from MGI. We removed three exposures (cystatin C, fasting plasma glucose, and estradiol levels) that because of their high missingness would have led to very small sample sizes in a complete case analysis across multiple exposures.

We retained the training/testing data split from the "ExPRS evaluation" (see above) and ran the following model for each of the remaining exposures and each of the selected common chronic conditions in the training data:

$$\text{logit}(P(\text{a common chronic condition is present}|\text{Exposure}, \text{Age}, \text{Sex}, \text{Array}, \text{PCs})) = \beta_0 + \beta_{Exposure}\text{Exposure} + \beta_{Age}\text{Age} + \beta_{Sex}\text{Sex} + \beta_{Array}\text{Array} + \beta_{PCs}\text{PCs}. \quad \text{(Equation 4)}$$

As with the multiExPRSs, we selected the significantly associated exposures and performed clumping to only retain the significantly associated exposures with a correlation < 0.5 with each other. We used the remaining set of exposures to create a complete case training dataset that we used to obtain effect sizes for each exposure that we used as weights to create weighted exposures in the complete case testing data. The weighted exposures were then combined into a single predictor that we refer to as poly-exposure score (PXS). Finally, we compared the AAUC of following four predictors adjusting for the set of covariates from above: the condition-specific PRS ("YPRS"), the combined ExPRS ("multiExPRS"), the "multiExPRS + YPRS," and the PXS.

Online visual catalog: ExPRSweb

The online open access visual catalog ExPRSweb (see web resources) was implemented with Grails as previously described.[20]

Unless otherwise stated, analyses were performed with R 4.1.1.

Results

Descriptive characteristics of study cohorts

For the generation and analysis of ExPRSs, we used two analytical datasets that were restricted to unrelated participants of broad European ancestry encompassing 46,782 individuals in MGI and 408,595 individuals in UKB (Table 1; subjects and methods).[34,39,41] The different prevalences of binary exposures and common chronic conditions in MGI and UKB most likely reflect the characteristics of a hospital-based study (MGI) and a healthier, population-based study (UKB), respectively (Table 1, Table S1). For example, there are marked differences between MGI and UKB regarding hypertension (49.8% versus 27.0%), diabetes (21.4% versus 7.2%), and lung cancer (2.2% versus 1.0%). Also, overweight individuals (74.7% versus 66.8%) and smokers (49.2% versus 39.4%) were more frequent in MGI (Figure S1).

Heritability estimates

In total, we identified 82 sets of GWAS summary statistics for 28 different exposures (21 quantitative, seven binary)

Table 1. Demographics and clinical characteristics of the analytic datasets

Characteristic	MGI	UKB
Demographics		
Study type	hospital-based	population-based
N	46,782	408,595
Females, n (%)	24,454 (52.3%)	220,896 (54.1%)
Mean age, years (SD)	56.7 (16.4)	56.9 (8.0)
Neighborhood deprivation index (SD)	0.9 (0.6)	not available
Townsend deprivation index (TDI)	not available	−1.3 (3.1)
Visits/measurements		
Median number of visits per participant	45	3[a]
Median time (years) between first and last visit	5.5	7.8[a]
Median lab orders per participant	59	34
Body mass index (BMI)	29.9 (7.1)	27.4 (4.8)
Underweight (BMI < 18.5), n (%)	498 (1.1%)	2,045 (0.5%)
Normal (BMI 18.5–24.9), n (%)	11,349 (24.3%)	132,264 (32.4%)
Overweight (BMI > 25.0), n (%)	34,916 (74.7%)	272,943 (66.8%)
Smoking status		
Yes	22,919 (49.2%)	160,954 (39.4%)
No	23,744 (50.8%)	247,641 (60.6%)
Selected common chronic conditions		
Hypertension, n (%)	23,314 (49.8%)	110,134 (27.0%)
Diabetes	10,012 (21.4%)	29,389 (7.2%)
Lung cancer	1,036 (2.2%)	3,885 (1.0%)

[a]Based on all available dates of first in-patient diagnoses.

that had matching exposure data in MGI and/or UKB; 52 solely based on UKB data and 30 on large GWASs (Table 2, Table S2). For each set, we estimated the narrow sense heritability[52] as PRSs are closely connected to it and because one PRS method (DBSLMM) relies on these estimates. After excluding three GWAS sets with negative h^2 estimates, we observed heritability estimates between 0.003 (sleep apnea) and 0.518 (height) that were in line with previous studies (Table S3).[4,53–57] Still, estimates from GWASs on the same exposure often varied (e.g., h^2[height]: 0.012–0.518 or h^2[vitamin D]: 0.009–0.100), implying different underlying frameworks (Figure S2).

ExPRS evaluation

Following the scheme in Figure 1, we generated 514 ExPRSs (379 for 25 exposures in MGI and 135 for 17 exposures in UKB; Table S4) and assessed association, overall performance, accuracy, and discrimination. A total of 336 ExPRSs for 27 exposures were nominally significant and positively associated with their corresponding exposures in MGI (262 ExPRSs; 24 exposures) and in UKB (74 ExPRSs; 14 exposures) and analyzed further (Table S4).

Performance comparison across methods

For the method comparison, we focus on MGI because it had a more comprehensive set of exposures covered by ExPRSs. PRS-CS produced the best performing ExPRSs for 18 of the 24 exposures, consistent with previous benchmarking (Table 3, Figure 2, and Figure S3).[58–60] Lassosum excelled for the alcohol and smoker exposure, DBSLMM for lipid levels, and both C + T implementations for exposures with low heritability, e.g., vitamins B12 and D (Figure 2). Further, we found that the C + T implementation that uses dosages for LD clumping had a slight edge over the one using best-guess genotypes, confirming previous findings.[61] Overall, these results suggested the methods' performances differed by trait, showcasing the benefit of screening multiple methods.

Performance across exposures

Again, focusing on MGI, we selected for each exposure the ExPRS with the lowest association p value among its method/exposure GWAS combinations (Table S4). For quantitative exposures, the Pearson's correlation r with their corresponding ExPRS ranged from 0.049 (vitamin B12) to 0.373 (height). For binary exposures, the area under the covariate-adjusted area under the ROC curve (AAUC) ranged from

Table 2. Overview of the 28 included exposures traits

		Discovery GWAS		Heritability[a]		Evaluation cohort sample sizes (cases/controls or total)	
Exposure	**Category**	**Meta-analysis**	**UKB**	**h_g^2**	**SE**	**MGI**	**UKB**
Continuous traits							
HDL cholesterol	cardiovascular	n/a	2	0.228	0.021	18,639	n/a
LDL cholesterol	cardiovascular	n/a	2	0.113	0.016	18,576	n/a
Triglycerides	cardiovascular	n/a	2	0.200	0.022	19,184	n/a
Total cholesterol	cardiovascular	n/a	2	0.131	0.017	18,231	n/a
PUFAs	cardiovascular	1	n/a	0.148	0.081	n/a	174,277
CRP	cardiovascular	1	2	0.198	0.106	10,292	389,826
eGFR	renal biomarker	1	n/a	0.051	0.004	43,039	390,449
Creatinine	renal biomarker	3	4	0.260	0.038	40,792	390,449
Cystatin C	renal biomarker	3	2	0.230	0.025	213	390,609
Vitamin D	vitamin levels	1	4	0.100	0.017	13,854	373,768
Vitamin B12	vitamin levels	1	2	0.023	0.293	8,626	174,277
Fasting glucose plasma	blood sugar levels	2	n/a	0.071	0.011	570	n/a
Glucose	blood sugar levels	2	2	0.077	0.008	40,801	346,477
Estradiol	women's health	2	2	0.033	0.007	1,875	61,982
Age at menopause	women's health	1	n/a	0.109	0.010	n/a	139,773
Age at menarche	women's health	1	n/a	0.109	0.007	n/a	220,885
BMI	anthropometric	n/a	4	0.239	0.009	46,763	n/a
Height	anthropometric	2	2	0.518	0.034	46,699	407,750
DBP	Vitals	n/a	4	0.140	0.008	46,148	n/a
SBP	Vitals	n/a	4	0.148	0.008	46,144	n/a
Alcohol amount	health behavior	1	1	0.055	0.006	26,666	121,424
Binary traits							
Thinness	anthropometric	1	n/a	0.133	0.034	753/41,938	3,547/396,201
Drinker	health behavior	n/a	1	0.100	0.006	30,900/13,952	n/a
Smoker	health behavior	1	3	0.156	0.007	22,919/23,744	246,067/160,791
Type 2 diabetes	preexisting condition	3	1	0.307	0.027	9,843/32,794	19,780/386,988
Hypertension	preexisting condition	2	3	0.413	0.165	23,158 /23,465	77,740 / 329,912
Insomnia	preexisting condition	n/a	2	N/A[b]	N/A	5,524/31,654	n/a
Apnoea	preexisting condition	1	1	0.284	0.057	10,909/31,654	4,460/403,370

Details can be found in Tables S1–S3. Number of included discovery GWASs, estimated heritability (liability scale), and sample size of PRS evaluation cohorts are shown.
HDL, high-density lipoprotein; LDL, low-density lipoprotein; PUFAs, polyunsaturated fatty acids; CRP, C-reactive protein; eGFR, estimated glomerular filtration rate; BMI, body mass index; DBP, diastolic blood pressure; SBP, systolic blood pressure; n/a, not available; N/A, not applicable.
[a]Maximally observed heritability estimate if multiple discovery GWASs were available.
[b]Discovery GWAS on ordinal scale.

0.524 (insomnia) to 0.637 (T2D), confirming only modest discrimination by PRSs for complex traits.[62] The ExPRSs' performance generally agreed with the ranking of their heritability estimates (Figures S4 and S5).

Performance comparison across cohorts

As with MGI, we selected in UKB for each of the 14 exposures the ExPRS that reached the strongest association (Table S5): six were based on Lassosum, four on PRS-CS, three on C + T (DS), and one on C + T (GT). In contrast to MGI, Lassosum outperformed the other methods in UKB (Figures S6 and S7). The inconsistencies across cohorts might be the result of different underlying GWAS sets, i.e., for UKB ExPRSs, we only relied on non-UKB studies to avoid overfitting. Also, the methods' varying tuning

Table 3. Top ranked ExPRSs in MGI

Exposure	Discovery GWAS	Method	# SNPs in ExPRS	Association p	Brier score	Pearson's r	Adjusted AUC (95% CI)
Continuous traits							
HDL	UK Biobank	PRS-CS	1,113,830	1.3E−294	N/A	0.311	N/A
LDL	UK Biobank	DBSLMM	8,918,470	4.1E−174	N/A	0.274	N/A
TG	UK Biobank	DBSLMM	8,924,773	6.9E−304	N/A	0.348	N/A
TC	UK Biobank	PRS-CS	1,113,831	3.2E−168	N/A	0.265	N/A
CRP	UK Biobank	PRS-CS	1,113,831	6.2E−30	N/A	0.155	N/A
eGFR	Meta-analysis	PRS-CS	1,113,831	3.8E−150	N/A	0.13	N/A
Creatinine	UK Biobank	PRS-CS	1,113,831	4.6E−189	N/A	0.174	N/A
Vitamin D	UK Biobank	C + T(GT)	500	9.0E−46	N/A	0.166	N/A
Vitamin B_{12}	UK Biobank	C+T(DS)	9	0.0011	N/A	0.049	N/A
FPG	Meta-analysis	C + T(GT)	273	0.0095	N/A	0.099	N/A
Glucose	UK Biobank	PRS-CS	1,113,830	8.4E−112	N/A	0.156	N/A
Estradiol	UK Biobank	PRS-CS	1,113,823	0.046	N/A	0.0664	N/A
BMI	UK Biobank	PRS-CS	1,113,832	5.3E−607	N/A	0.319	N/A
Height	UK Biobank	PRS-CS	1,113,832	2.2E−1988	N/A	0.373	N/A
DBP	UK Biobank	PRS-CS	1,113,831	4.8E−170	N/A	0.166	N/A
SBP	UK Biobank	PRS-CS	1,113,831	1.7E−189	N/A	0.172	N/A
Alcohol amount	Meta-analysis	PRS-CS	1,116,497	1.4E−18	N/A	0.0746	N/A
Binary traits							
Thinness	Meta-analysis	C + T (DS)	256	0.044	0.018	N/A	0.532 (0.503, 0.561)
Drinker	UK Biobank	PRS-CS	1,113,832	5.1E−29	0.212	N/A	0.547 (0.539, 0.536)
Smoker	Meta-analysis	PRS-CS	1,109,786	1.10E−170	0.232	N/A	0.605 (0.598, 0.612)
T2D	Meta-analysis	PRS-CS	945,820	1.10E−159	0.139	N/A	0.637 (0.621, 0.653)
Hypertension	UK Biobank	PRS-CS	1,113,832	6.4E−213	0.182	N/A	0.630 (0.622, 0.639)
Insomnia	UK Biobank	PRS-CS	1,065,129	5.0E−06	0.126	N/A	0.524 (0.513, 0.536)
Sleep apnea	UK Biobank	PRS-CS	1,111,194	5.8E−10	0.182	N/A	0.527 (0.517, 0.536)

HDL, high-density lipoprotein; LDL, low-density lipoprotein; PUFAs, polyunsaturated fatty acids; CRP, C-reactive protein; eGFR, estimated glomerular filtration rate; BMI, body mass index; DBP, diastolic blood pressure; SBP, systolic blood pressure; N/A: not applicable. Details about the underlying discovery GWAS can be found in Table S2.

procedures, especially for Lassosum and C + T, might be affected by the larger sample sizes in UKB. For ExPRSs of quantitative exposures, the correlation with their corresponding exposures ranged from 0.015 (alcohol consumption) to 0.326 (height) (Table S5). For binary exposures, the AAUC ranged from 0.505 (hypertension) to 0.825 (T2D). When comparing ExPRSs on exposures that were present in both cohorts, we found generally consistent performances for quantitative traits such as C-reactive protein, creatine, vitamin D, and height, while for some binary traits such as T2D ($AAUC_{MGI}$: 0.64, $AAUC_{UKB}$: 0.83) and smoking ($AAUC_{MGI}$: 0.61, $AAUC_{UKB}$: 0.77), AAUC differed substantially (Table S6). Of note, the estimates in UKB might be heightened as a result of undetected, overlapping samples between their discovery GWAS and the UKB cohort[15,63] or caused by to the cohort's larger effective sample sizes.

Correlations of ExPRSs across exposures

Next, we assessed the relationships between ExPRSs and exposures in MGI. Figure 3 displays the pairwise correlation between 15 quantitative exposures, between their 15 corresponding ExPRSs, and between the ExPRSs and the quantitative exposures in MGI. The correlations between the quantitative exposures indicated positive and negative relationships (r between −0.1 and 0.92; Figure 3A), the strongest between closely related exposures: r[TC, LDL] = 0.92, r [eGFR, creatine] = −0.84, and r[SBP, DBP] = 0.53. The former two can be attributed to their underlying equations and related measurements, while the linear relationship between SBP and DBP is well established.[64,65] Several of the other observed correlations are also well documented, often reflecting related disease etiologies.[66–68] Similar but more attenuated patterns were seen for the ExPRSs whose correlations ranged from −0.78 to 0.72 (Figure 3B). The often lower

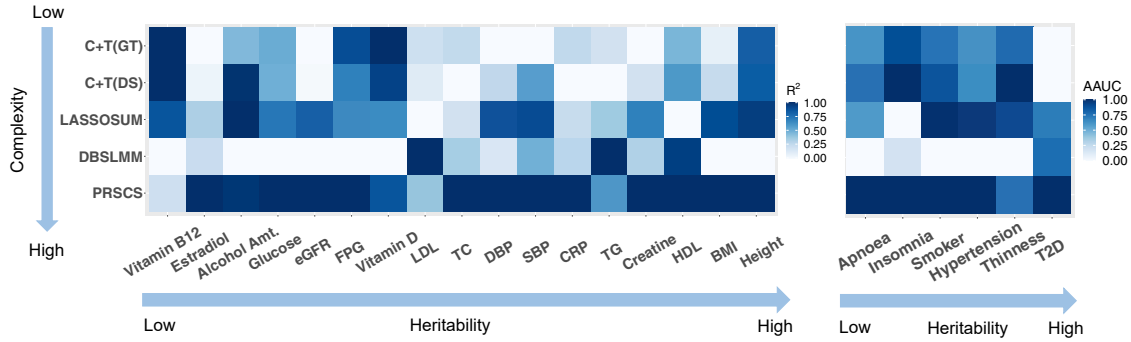

Figure 2. Prediction performance of the five applied PRS methods in MGI across continuous (left) and binary (right) traits
Here, the heatmap shows the relative prediction performance for each method across traits (values were scaled to 0–1 range) for better comparison. Specifically, the prediction performance is quantified with R^2 for continuous traits and covariate-adjusted AUC for binary traits. For a fair comparison, we selected the same summary statistic for each method (GWAS with the highest heritability estimate).

pairwise correlations (e.g., r[PRS_{TC}, PRS_{LDL}] = 0.72 and r[PRS_{eGFR}, $PRS_{creatine}$] = −0.78) were expected because ExPRSs capture only a fraction of the exposure's variance (see diagonal of Figure 3C). The consistent patterns suggested that several ExPRSs can replicate correlations of measured exposures relatively well and thus might be suitable surrogates for exposures, especially for studies where measurements might not be feasible or likely be biased.[66,69,70]

ExPRS applications

Phenome-wide association analyses

One application of ExPRSs is their use as predictors for phenome-wide association studies (PheWASs) to uncover phenotypes with a shared genetic component and thus disorders that might benefit from an early intervention. We showcase such ExPRS PheWASs by analyzing all 24 selected ExPRSs across up to 1,685 EHR-derived phenotypes (PheCodes) in MGI (Table S7). In total, we observed phenome-wide significant associations between 22 ExPRSs and 440 phenotypes (Bonferroni-corrected threshold at $p < 0.05/1{,}685$; Table S8). Overall, the number and the strength of observed associations seem to depend on the exposures' impact and heritability. For example, the PheWAS with the BMI ExPRS uncovered 329 associated phenotypes while the vitamin B_{12} ExPRS PheWAS only revealed two associations with closely related phenotypes. Besides the expected associations between BMI PRS and obesity-related phenotypes ($1.66 <$ odds ratio [OR] < 2.14, e.g., obesity, morbid obesity, and overweight), we also observed significant phenome-wide associations with hypertension (OR: 1.33 [1.30, 1.36]), T2D (OR: 1.41 [1.37, 1.45]), osteoarthrosis (OR: 1.15 [1.12, 1.17]), and sleep apnea (OR: 1.28 [1.25, 1.31]); all were previously reported for BMI[71–74] (Figure 4A; Table S8). The PheWAS with measured BMI revealed consistent associations (Figure 4C; Table S9), although with larger effects: hypertension (OR: 1.88 [1.84, 1.93]), T2D (OR: 2.00 [1.95, 2.06]), osteoarthrosis (OR: 1.29 [1.27, 1.32]), and sleep apnea (OR: 2.24 [2.18, 2.30]).

To assess whether these associations were driven by exposed individuals, i.e., individuals affected by a binary exposure or by low or high exposure values, we also performed "exclusion-PRS-PheWAS" analyses where we excluded such exposed individuals to remove direct and indirect associations of the exposure and potential treatment effects (see subjects and methods). While this exclusion of individuals markedly decreased sample sizes and thus power, we identified 198 phenotypes that remained significantly associated with 17 ExPRSs in the exclusion-PRS-PheWAS ($p < 0.05/1{,}685$; Table S8). For example, in the exclusion PheWAS with the BMI ExPRS, the associations with hypertension (OR: 1.17 [1.12, 1.23]) and T2D (OR: 1.18 [1.09, 1.27]) remained statistically significant (Figure 4B, Table S8). However, while the analysis of individuals with healthy BMI removed most of the obesity or overweight phenotypes, a strong association remained between BMI ExPRS and bariatric surgery (OR: 2.66 [2.08, 3.41]). A closer inspection revealed that 73 of 1,509 MGI participants who underwent bariatric surgery had recorded median BMI values that fell in the healthy BMI range ($18.5 <$ BMI < 25), indicating the BMI ExPRS's ability to capture pre-treatment exposures. Most interestingly, the corresponding exclusion PheWAS with measured BMI as predictor revealed many association signals that were reversed compared to the exclusion-PRS-PheWAS (Figures 4C and 4D, Tables S8 and S9). This finding might reflect biased measurements, e.g., due to treatment or interventions that result in normal BMI values, or the measured BMI's inability to capture central obesity.[75]

We performed similar sets of PheWASs in UKB. While based on a separate ExPRS generation restricted to UKB-independent GWAS summary statistics, most of the strong associations seen in the MGI were also seen in the UKB ExPRS PheWASs, e.g., obesity associated with T2D PRS (OR_{MGI}: 1.71 [1.15, 1.20] and OR_{UKB}: 1.63 [1.60, 1.66]; Figures S8 and S9). Because of the larger sample sizes in the UKB compared to MGI (Table 2), we often observed more and stronger secondary trait associations (Tables S10 and S11).

In general, we found that agnostic ExPRS PheWASs can provide valuable insights into exposure-phenotype relationships, many of which were previously reported for measured exposures. However, thorough investigations are needed to distinguish between spurious and genuine signals.

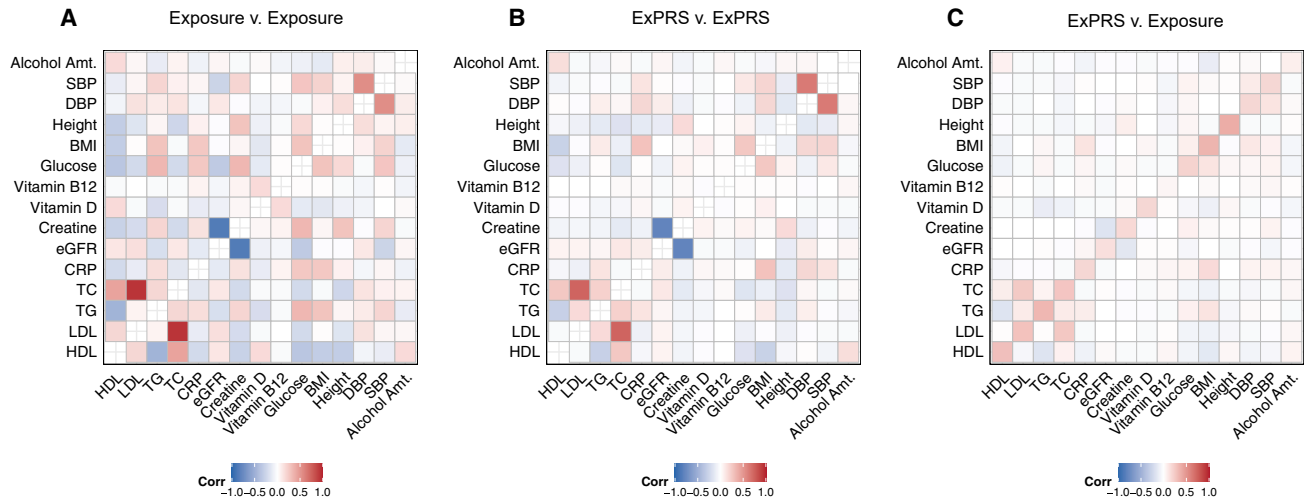

Figure 3. Comparison of the pairwise correlation of 15 ExPRSs and their corresponding continuous traits in MGI
Heatmap displays the pairwise correlation between (A) 15 continuous exposures in MGI; (B) ExPRSs; and (C) exposures (y axis) and ExPRSs (x axis). Here, pairwise Spearman correlation with nominally significant association p values (≤ 0.05) are shown. Fasting plasma glucose (exposure and ExPRS) was excluded because of the exposures low sample size in MGI.

Improving prediction models for common chronic conditions

As many exposures are important risk factors for common chronic conditions,[9,76,77] we performed analyses with a specific emphasis on 27 chronic conditions whose algorithms are used for Medicaid and Medicare claims and available from the Chronic Condition Data Warehouse (CCW, Table S12).[78] Because these were developed for the US health system and lack transferability to the UK, we limited our analysis to MGI. Related chronic conditions were already covered in the phenome-wide PheCode-based association analyses, therefore this targeted analysis of "real-world" phenotype algorithms aims to evaluate the ExPRSs' abilities to improve predictions. Basically, we are interested to see whether a prediction model that solely relies on a GWAS-based PRS for a chronic condition "Y" (YPRS) can be augmented with additional ExPRSs.

As a first step, we explored the association between the 27 conditions and the 24 ExPRSs. We found that even after excluding the directly related condition/exposure pairs (e.g., hypertension/SBP ExPRS, hyperlipidemia/TC ExPRS, etc.) all included 24 ExPRSs showed a nominally significant association with at least one condition at $p < 0.05$ (Table S13). Conversely, 26 of the 27 conditions were nominally significantly associated with at least one ExPRS substantiating the exposures' relevance. However, none of the ExPRSs were associated with Alzheimer disease, although many of the included exposures were reported risk factors.[79] The strongest risk-increasing effect was seen for BMI ExPRS and diabetes (OR: 1.393 [1.357, 1.430]), while the strongest protective effect was seen for HDL ExPRS and diabetes (OR: 0.823 [0.803, 0.844]) (Table S13).

Considering the relatively poor predictive performance of single ExPRSs for chronic conditions and that some of the chronic conditions were associated with several ExPRSs (Table S13), we next assessed whether the combination of ExPRSs ("multiExPRSs" see subjects and methods) can improve risk prediction of models that only include YPRSs (Table S14).

Because of the required cross-validation, limited sample sizes, and limited availability of YPRSs (Table S15), we restricted our comparisons to 12 conditions (Tables S12 and S15). We found that adding multiple ExPRSs enhanced models for several conditions (e.g., stroke/transient ischemic attack, heart failure, lung cancer, hypertension, chronic kidney disease, asthma; Table S16, Figure 5). For example, the AAUC for predicting hypertension increased from 0.627 to 0.637 when adding multiple ExPRSs (BMI, C-reactive protein, drinking status, fast plasma glucose, HDL, height, smoking status, T2D, triglycerides, apnea, and insomnia). In contrast, the addition of ExPRSs did not improve prediction accuracy for other conditions (e.g., glaucoma, prostate cancer, colorectal cancer, and atrial fibrillation). Nevertheless, the ability of specific ExPRSs to improve predictions indicates that some of the YPRSs often do not capture the entirety of an individual's genetic predisposition, most likely reflecting the lack of power of the condition's discovery GWAS compared to exposure GWASs, which as a result of larger sample sizes and continuous measurements, are often better powered.

Because these predictions yielded only moderate to poor discrimination (AAUC < 0.66), we also evaluated the ExPRSs' ability to augment risk stratification with YPRSs, i.e., to define subsets of individuals at high risk for the 12 conditions (Figure S10, Tables S17 and S18). Except for the heart failure PRS and the lung cancer PRS, ten of the 12 YPRSs were by themselves able to significantly enrich cases in at least one of the top bins ($\geq 5\%$, 5%–10%, or 10%–25%) compared to the center bin (40%–60%) of their distributions. For example, ten YPRSs could significantly enrich cases in the top $\leq 5\%$ bin at $p < 0.05$ with OR ranging from 1.26 (95% CI: 1.02, 1.54; chronic kidney disease) to 3.60 (95% CI: 2.83, 4.56; prostate cancer).

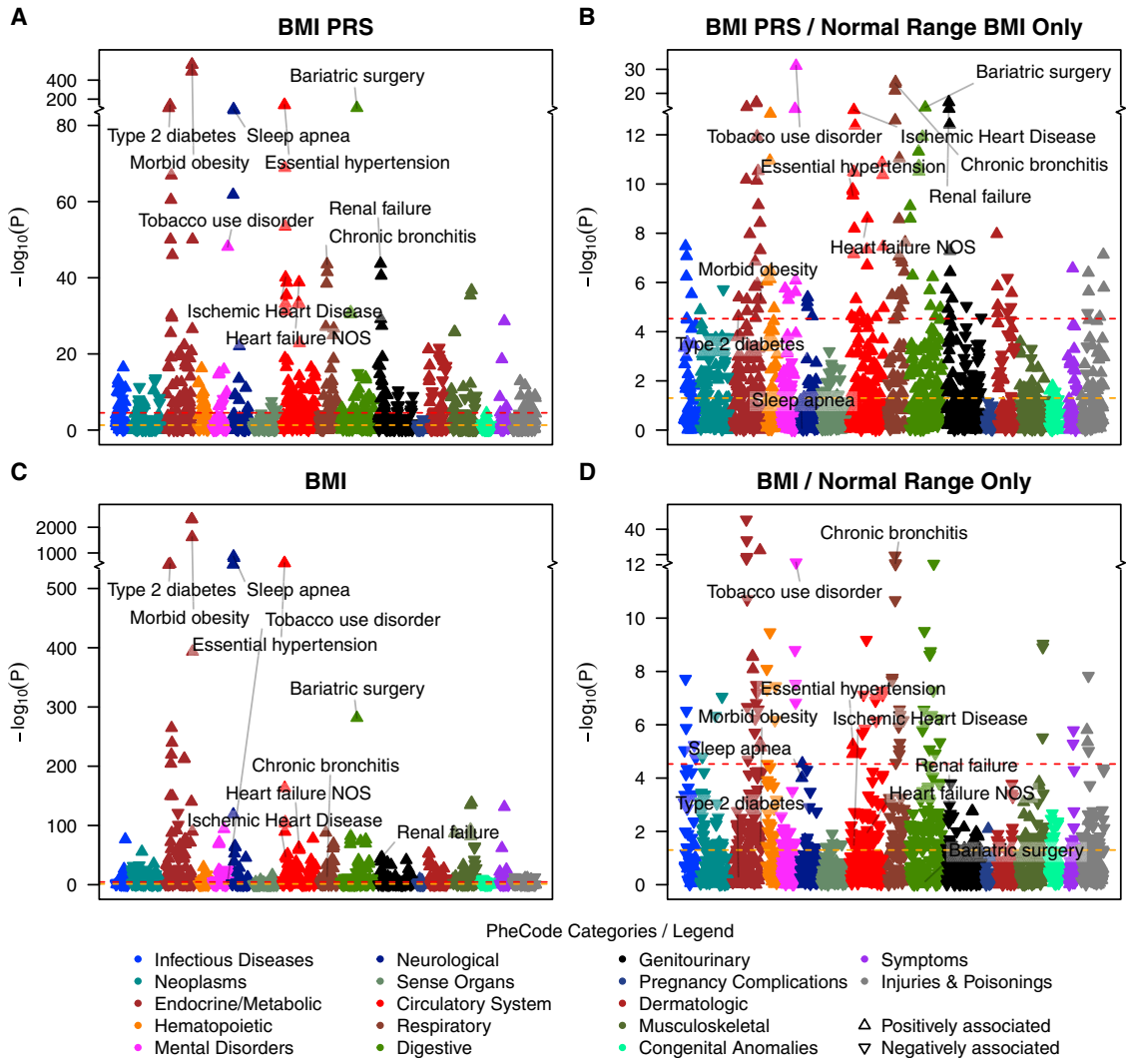

Figure 4. ExPRS PheWAS and exclusion-ExPRS-PheWAS as an example for continuous traits in MGI

(A) ExPRS PheWAS plot for BMI ExPRS.

(B) Exclusion-ExPRS-PheWAS plot is shown for using BMI ExPRS as predictor among the individuals with normal BMI value (18.5–24.9 kg/m^2).

(C) Trait PheWAS plot is shown for BMI trait.

(D) Exclusion-trait-PheWAS plot is shown for using BMI trait as predictor among the individuals with normal BMI value. The axis breaks were chosen so that the ten strongest signals fall in the top scale (y axis breaks for the four panels at $-\log_{10}(p)$ are 84, 13, 540, and 12, respectively). The red dashed line indicates genome-wide significance ($p < 0.05/1,685$), and the orange line indicates nominal significance ($p < 0.05$).

Adding the combined ExPRSs (multiExPRSs) to the "YPRS-only model" improved the enrichment of cases for nine of the 12 conditions when considering the top "≤ 5%" bin. The largest improvements were seen for the enrichment of cases in the top <5% with heart failure (YPRS: OR: 1.16 [0.94, 1.44] versus YPRS + multiExPRS: OR: 1.52 [1.23, 1.87]) and with T2D (YPRS: OR: 2.55 [2.19, 2.98] versus YPRS + multiExPRS: OR: 3.13 [2.69, 3.65]). However, adding multiple ExPRSs negatively affected the enrichment of cases with atrial fibrillation (YPRS: OR: 3.34 [2.80, 3.99] versus YPRS + multiExPRS: OR: 3.09 [2.60, 3.68]). Similar but less pronounced enrichments of cases were seen for the top 5%–10%%, and the top 10%–25% bins (Figure S10, Table S18).

Our explorations confirmed that individuals in the tails of PRS distributions are most informative for risks of chronic conditions.[80] Further, the consistent gain in risk stratification by adding multiple ExPRSs highlights their potential use.

Finally, we compared the application of the PRSs (YPRSs and/or the multiExPRSs) with poly-exposure scores (PXSs) that are based on measured/collected exposure data as previously described for type 2 diabetes.[81] Again, focusing on the 12 conditions (Table S12), we created a PXS for each condition in the MGI cohort by using up to 24 of 27 available exposures (subjects and methods). The number of incorporated exposures ranged from seven (glaucoma) to 19 (chronic kidney disease) (Tables S19 and S20). Although

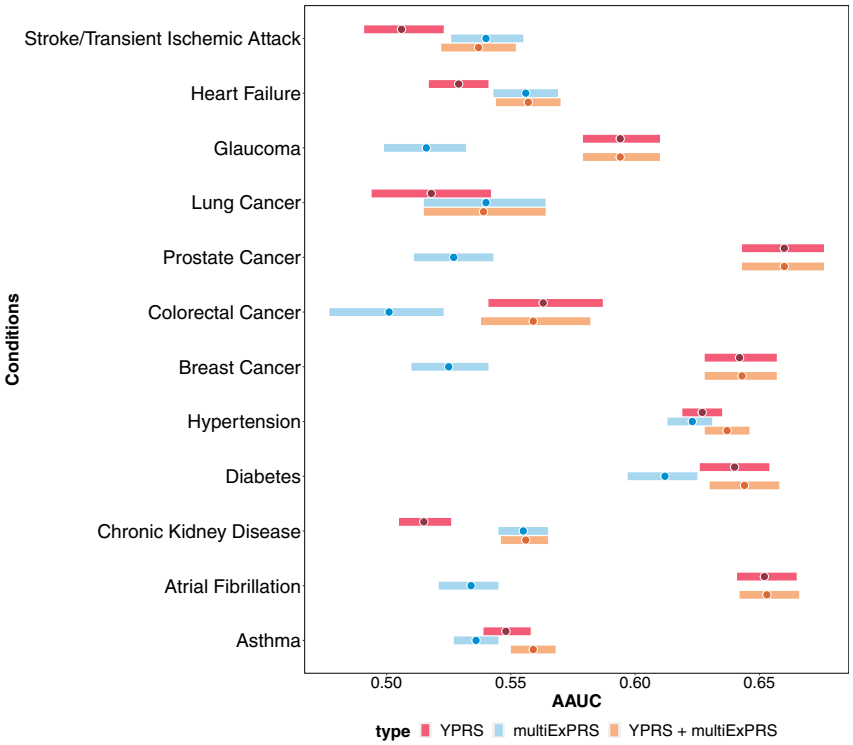

Figure 5. Comparisons of the prediction performance of different predictors for common chronic conditions in MGI cohort
AAUC paired with 95% confidence interval for condition specific PRS (YPRS, red), combined ExPRSs (multiExPRS, blue), and YPRS + multiExPRS (orange) were shown as forest plot. Each bar represents the 95% interval for the AAUC with the dot represents the AAUC estimate.

the evaluation cohorts were different in size, we observed that the PXSs mostly showed better discrimination than the models that only relied on PRSs (YPRS, multiExPRS, or YPRS + multiExPRS), except for colorectal, prostate, and female breast cancer, which underperformed (Figure S11, Table S18). Because PXSs were only obtainable for people who had complete data for each included exposure, they were only available for a small fraction of the genotyped MGI individuals for which YPRSs and ExPRSs were obtainable (2.5%–18.4%; glaucoma: 56.0%). Furthermore, the proportion of genotyped individuals with complete exposure data for their PXSs was significantly different between cases and controls for nine of the 12 analyzed conditions, indicating non-random missingness of exposures in the MGI EHR that most likely biased the analysis. The most extreme example was chronic kidney disease: cases were about four times more likely than controls to have complete exposure data for their PXS (OR 3.9 [3.5, 4.4], $p = 3.1 \times 10^{-107}$; Table S21).

Online visual catalog: ExPRSweb

In our current study, we generated and evaluated hundreds of ExPRSs in which predictive properties differed between GWAS source, exposure, method, and/or evaluation cohort (Table S4). To enable an exploration of the ExPRSs for 27 different exposures, we created a new PRSweb[20] instance called ExPRSweb (see web resources) that includes detailed metrics (association, performance, discrimination, and accuracy) and allows the selection of ExPRSs on the basis of properties for specific applications. The tables, such as Table 3 and Table S5, can be sorted, filtered, or downloaded. ExPRSweb also offers detailed information about each ExPRS, including GWAS source(s), LD reference panels and the included risk variants, effect/non-effect alleles, and weights. ExPRSweb also links to interactive ExPRS-PheWAS results for their evaluation cohort.

Discussion

In this study, we have constructed and evaluated a large set of ExPRSs by using 79 sets of GWAS summary statistics, applied various PRS methods, and while doing so, created over 514 ExPRSs, 336 of which showed promising performance for 27 different exposures in MGI and/or UKB.

We explored the performance of ExPRSs across methods, GWAS sources, and two cohorts and observed two key points that might be helpful to strategize future ExPRS generation projects. First, large exposure GWASs with higher SNP heritability estimates usually also resulted in the most predictive ExPRSs. Second, our results indicated that there might not be a one-size-fits-all approach for generating the most predictive ExPRSs but rather an array of choices one must make among many methods and available GWAS summary statistics. By comprehensively presenting our PRSs' underlying GWAS sources, the evaluations of an array of PRSs per trait in the same cohort, and some of their applications, we aimed to inform this choice.

What sets our work apart from other recent papers that systematically generated a broad set of polygenic scores, some of which for traits that overlap with our exposures,[82,83] is the comprehensive exploration of several methods across freely available GWAS summary statistics, the interactive presentation of their evaluation metrics in the same cohort, and their phenome-wide exploration.

While there is a wide range of health-related exposures,[84–86] we focused on 28 exposures for which we could find GWASs from external full summary statistics and for which we had sufficiently measured samples in MGI and/or UKB. The exposures can roughly be categorized into cardiovascular, renal biomarkers, vitamin levels, blood sugar levels, women's health, anthropometric measurements, vitals, health behaviors, and preexisting conditions. However, other relevant exposures were not explored in this study, e.g., dietary exposures (e.g., milk consumption,

coffee consumption[87,88]), telomere length,[89] and other biomarkers (e.g., transforming growth factor beta [TGF-beta],[90,91] circulating microRNA miR-34b[92]). While some exposures also have GWAS summary statistics available (e.g., coffee consumption, milk consumption summary statistics from UKB GWAS efforts), the exposures were not measured in MGI and thus could not be evaluated at this time.

We generated ExPRSs by using four methods (C + T, Lassosum, DBSLMM, and PRS-CS), all of which are computationally efficient, but skipped other new methods that have been proposed (SBayesR, LDPred, NPS, and SCT)[93–96] but often require massive computational resources, especially for large cohorts such as UKB and MGI. Additionally, several alternative methods were reported to improve predictive power by incorporating external information (e.g., functional annotations, pleiotropy across multiple traits), e.g., LDpred-funct,[97] AnnoPred,[98] and MTGBLUP.[99] Future implementations and systematic evaluations of these alternative choices are needed to further the availability of well-powered ExPRSs and their applications.

We focused our ExPRS generation and evaluation on samples of broadly European ancestry because of the limited diversity in MGI and UKB. However, the lacking transferability across ancestry groups increases the need to also construct ExPRSs for non-European ancestry groups.[29,95,100–102] When applying our top ranked ExPRSs to AFR, EAS, and CSA ancestry groups, we observed overall drop in predictive power in these non-EUR groups, i.e., weaker correlations for continuous or lower AAUC values for binary exposures, confirming previous studies that reported a transferability problem for EUR-based polygenic scores to other ancestries (Note S1, Figures S12 and S13, Tables S23–S25).[83,103,104] Nevertheless, our results indicated that some of the EUR-based ExPRSs can potentially be useful also for non-EUR individuals, although this only represents a compromise solution.[102] While efforts are underway to develop cross-ancestry PRS methods to increase transferability, ultimately an increased diversity in datasets is needed to counteract the European ancestry bias in GWASs that is passed on to PRS research.[102,105]

Our explorations of ExPRSs, mainly in the MGI cohort, revealed that some of the ExPRSs could be good surrogates for exposures and enable meaningful association analyses across medical phenomes or a collection of chronic conditions. Also, the combination of ExPRSs could to some degree improve predictions and risk stratification beyond the YPRSs, e.g., for asthma, heart failure, or hypertension. Yet, for some of the studied conditions, the additional of multiple ExPRSs did not improve models that already included YPRSs. This suggests that YPRSs, if based on very large sample sizes, might already have captured most of the genetic risk profiles reflecting direct and indirect (exposure-mediated) risk effects. Furthermore, it is important to bear in mind that the observed improvement in risk prediction by combining YPRSs with multiExPRSs was not validated outside the MGI cohort. Additional external studies are needed to explore the generalizability of the presented approach.

There are other applications of ExPRSs that gained attention in the recent years, e.g., mediation analyses to study polygenic pleiotropy[106] or their use as instrumental variables in Mendelian randomization analysis to uncover novel mechanisms that contribute toward disease susceptibility.[10,107–109] In our example applications, we showcased the use of ExPRSs for phenome-wide explorations to identify clinical phenotypes potentially associated with an exposure. In addition, we applied "exclusion-ExPRS-PheWAS" to assess whether the observed associations were mediated by the extremes of a quantitative exposure (outer quarters or non-normal range) and by "exposed" individuals of a binary exposure, respectively. Some of the observed associations may represent true causal relationships; however, additional follow-up analyses of such PRS PheWASs were recommended to substantiate any potential causal relationships, e.g., by determining the heterogeneity of the association across all variants of an ExPRS and to perform sensitivity analyses to uncover potential biases and pleiotropic effects.[110] The latter might be especially crucial for ExPRSs, which are based on thousands of variants and thus more likely to be affected by pleiotropy, biases, and context-dependent effects.[111] Of note, while an association between an ExPRS and disease may indicate an intermediate on the causal pathway to disease or simply a shared biological mechanism between the exposure and the disease, it does not necessarily mean that interventions targeting modifiable exposures will impact disease risk or onset.

A main application for ExPRSs might be their use as proxies for unmeasured exposures. Exposures relevant for many conditions are often only sparsely measured in the EHR datasets and their missingness can substantially reduce sample size when considering only complete case datasets (as seen here for PXSs). Furthermore, contrary to genotype data, the missingness can be non-random because testing generally is selective, diagnosis and symptom specific, as seen here for nine of the 12 analyzed conditions, and thus most likely would bias prediction models. Nevertheless, an ExPRS can even in the best scenario only capture the heritable fraction of the exposure's variance coming from variants assigned at birth but not the early, current, or lifelong exposure to environmental or consequences of behavioral factors.[112] Also, for a lowly heritable exposure, a derived ExPRS will only be weakly correlated with the exposure and consequently represent a poor proxy. Using ExPRSs for the imputation of incomplete exposure data could be worth further explorations but was not within the scope of the current study.

Being dependent on large GWASs and evaluation cohorts, we expect that future studies will provide more powerful YPRSs and ExPRSs. But even then, the interplay of genetic and non-genetic factors needs to be considered when assessing complex traits. Current large biobank efforts link genotype data with EHRs and often complement patient information on environmental, lifestyle, and

demographic variables via self-report.[113] The integration of these resources will most likely improve our models with the goal to prevent or treat conditions earlier.

Finally, we created an online repository called "ExPRSweb" that, like our cancer-specific PRS repository "Cancer PRSweb,"[20] provides an interactive platform to browse performance metrics of all generated ExPRSs in two independent biobanks. We also deposited all promising ExPRSs to the PGS catalog and linked it to ExPRSweb and our evaluations. We anticipate that ExPRSweb can serve as an example and a standardized platform to expedite ExPRS research and to facilitate easier access.

Data and code availability

Data cannot be shared publicly as a result of patient confidentiality. The data underlying the results presented in the study are available from the UK Biobank at http://www.ukbiobank.ac.uk/register-apply/ and from the MGI Study at https://precisionhealth.umich.edu/ourresearch/michigangenomics/ for researchers who meet the criteria for access to confidential data. The software and R packages supporting the current study are available online (see web resources). All generated ExPRS constructs, their evaluations, and PheWAS summary statistics are available online at https://exprsweb.sph.umich.edu:8443.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.09.001.

Acknowledgments

The authors acknowledge the Michigan Genomics Initiative participants, Precision Health at the University of Michigan, and the University of Michigan Medical School Data Office for Clinical and Translational Research; the University of Michigan Medical School Central Biorepository and the University of Michigan Advanced Genomics Core for providing data storage, management, processing, and distribution services; and the Center for Statistical Genetics in the Department of Biostatistics at the School of Public Health for genotype data curation, imputation, and management in support of the research reported in this publication. We would like to thank Alison Mondul and Brett Vanderwerff (University of Michigan School of Public Health) for careful reading of the manuscript.

Part of this research has been conducted with both the UK Biobank Resource under application number 24460 and with results and data generated by previous researchers who have used the UK Biobank Resource.

This material is based in part upon work supported by the National Institutes of Health/NIH (NCI P30CA046592 [L.G.F. and B.M.]), by the University of Michigan (UM-Precision Health Investigators Award U063790 [L.G.F., S.P., Y.M., and B.M.]), and by the National Science Foundation under grant number DMS-1712933. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Declaration of interests

The authors declare no competing interests.

Received: February 10, 2022
Accepted: August 31, 2022
Published: September 23, 2022

Web resources

Cancer PRSweb, https://prsweb.sph.umich.edu
CMS Chronic Condition Warehouse, https://www2.ccwdata.org/web/guest/home/
ExPRSweb, https://exprsweb.sph.umich.edu
FinnGen consortium, https://www.finngen.fi/en/access_results
Gemma and DBSLMM, https://xzlab.org/software.html
Lassosum, https://github.com/tshmak/lassosum
Locuszoom, https://github.com/statgen/locuszoom
NHGRI-EBI GWAS Catalog, https://www.ebi.ac.uk/gwas/summary-statistics
PGS Catalog, https://www.pgscatalog.org
PLINK, https://www.cog-genomics.org/plink2
PRS-CS, https://github.com/getian107/PRScs
Rprs, https://github.com/statgen/Rprs
The Comprehensive R Archive Network, https://cran.r-project.org
The Michigan Genomics Initiative (MGI), https://precisionhealth.umich.edu/our-research/michigangenomics/
UCSC Genome Browser Store, https://genome-store.ucsc.edu
UKB GWAS (Lee Lab), https://www.leelabsg.org/resources
UKB GWAS (Neale Lab), https://github.com/Nealelab/UK_Biobank_GWAS

References

1. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.
2. Génin, E. (2020). Missing heritability of complex diseases: case solved? Hum. Genet. *139*, 103–113.
3. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
4. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
5. Kamps, R., Brandão, R.D., Bosch, B.J.v.d., Paulussen, A.D.C., Xanthoulea, S., Blok, M.J., and Romano, A. (2017). Next-generation sequencing in oncology: genetic diagnosis, risk prediction and cancer classification. Int. J. Mol. Sci. *18*, E308.
6. Jostins, L., and Barrett, J.C. (2011). Genetic risk prediction in complex disease. Hum. Mol. Genet. *20*, R182–R188.
7. Ma, Y., and Zhou, X. (2021). Genetic prediction of complex traits with polygenic scores: a statistical review. Trends Genet. *37*, 995–1011.
8. Meigs, J.B., Wilson, P.W.F., Fox, C.S., Vasan, R.S., Nathan, D.M., Sullivan, L.M., and D'Agostino, R.B. (2006). Body mass index, metabolic syndrome, and risk of type 2 diabetes or cardiovascular disease. J. Clin. Endocrinol. Metab. *91*, 2906–2912.
9. Almirall, J., Serra-Prat, M., Bolíbar, I., and Balasso, V. (2017). Risk factors for community-acquired pneumonia in adults: a

systematic review of observational studies. Respiration *94*, 299–311.
10. Pierce, B.L., Kraft, P., and Zhang, C. (2018). Mendelian randomization studies of cancer risk: a literature review. Curr. Epidemiol. Rep. *5*, 184–196.
11. Kachuri, L., Graff, R.E., Smith-Byrne, K., Meyers, T.J., Rashkin, S.R., Ziv, E., Witte, J.S., and Johansson, M. (2020). Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. Nat. Commun. *11*, 6084.
12. Haneuse, S. (2016). Distinguishing Selection Bias and Confounding Bias in Comparative Effectiveness Research. Med. Care *54*. e23–*29*.
13. Beesley, L.J., and Mukherjee, B. (2020). Statistical Inference for Association Studies Using Electronic Health Records: Handling Both Selection Bias and Outcome Misclassification (Biometrics).
14. Loos, R.J.F. (2020). 15 years of genome-wide association studies and no signs of slowing down. Nat. Commun. *11*, 5900.
15. Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D.M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., et al. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat. Genet. *51*, 237–244.
16. 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.
17. Lambert, S.A., Abraham, G., and Inouye, M. (2019). Towards clinical utility of polygenic risk scores. Hum. Mol. Genet. *28*, R133–R142.
18. Tam, C.H.T., Lim, C.K.P., Luk, A.O.Y., Ng, A.C.W., Lee, H.M., Jiang, G., Lau, E.S.H., Fan, B., Wan, R., Kong, A.P.S., et al. (2021). Development of genome-wide polygenic risk scores for lipid traits and clinical applications for dyslipidemia, subclinical atherosclerosis, and diabetes cardiovascular complications among East Asians. Genome Med. *13*, *29*.
19. Ma, Y., and Zhou, X. (2021). Genetic prediction of complex traits with polygenic scores: a statistical review. Trends Genet. *37*, 995–1011.
20. Fritsche, L.G., Patil, S., Beesley, L.J., VandeHaar, P., Salvatore, M., Ma, Y., Peng, R.B., Taliun, D., Zhou, X., and Mukherjee, B. (2020). Cancer PRSweb: an online repository with polygenic risk scores for major cancer traits and their evaluation in two independent biobanks. Am. J. Hum. Genet. *107*, 815–836.
21. Andrews, S.J., Fulton-Howard, B., O'Reilly, P., Marcora, E., Goate, A.M.; and collaborators of the Alzheimer's Disease Genetics Consortium (2021). Causal associations between modifiable risk factors and the alzheimer's phenome. Ann. Neurol. *89*, 54–65.
22. Li, S., and Schooling, C.M. (2021). A phenome-wide association study of genetically mimicked statins. BMC Med. *19*, 151.
23. Richardson, T.G., Harrison, S., Hemani, G., and Davey Smith, G. (2019). An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. Elife *8*, e43657.
24. 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.
25. Wray, N.R., Goddard, M.E., and Visscher, P.M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. *17*, 1520–1528.
26. International Schizophrenia Consortium, Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
27. Euesden, J., Lewis, C.M., and O'Reilly, P.F. (2015). PRSice: Polygenic Risk Score software. Bioinformatics *31*, 1466–1468.
28. 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.
29. 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.
30. 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.
31. Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H.M., Stambolian, D., Chew, E.Y., Branham, K.E., Heckenlively, J., FUSION Study, and Fulton, R., et al. (2014). Ancestry estimation and control of population stratification for sequence-based association studies. Nat. Genet. *46*, 409–415.
32. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., and Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.
33. Alexander, D., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109.
34. 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.
35. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.
36. Abraham, K.J., and Diaz, C. (2014). Identifying large sets of unrelated individuals and unrelated markers. Source Code Biol. Med. *9*, 6.
37. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64, 976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.
38. Ho, D.E., Imai, K., King, G., and Stuart, E.A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. J. Stat. Softw. *42*, 1–28.
39. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779.

40. 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.
41. Fritsche, L.G., Gruber, S.B., Wu, Z., Schmidt, E.M., Zawistowski, M., Moser, S.E., Blanc, V.M., Brummett, C.M., Kheterpal, S., Abecasis, G.R., and Mukherjee, B. (2018). Association of polygenic risk scores for multiple cancers in a phenome-wide study: results from the michigan genomics initiative. Am. J. Hum. Genet. *102*, 1048–1061.
42. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. *50*, 1335–1341.
43. 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.
44. Levey, A.S., Stevens, L.A., Schmid, C.H., Zhang, Y.L., Castro, A.F., 3rd, Feldman, H.I., Kusek, J.W., Eggers, P., Van Lente, F., Greene, T., et al. (2009). A new equation to estimate glomerular filtration rate. Ann. Intern. Med. *150*, 604–612.
45. Zhou, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. Ann. Appl. Stat. *11*, 2027–2051.
46. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.
47. Rao, C. (1971). Estimation of variance and covariance components—MINQUE theory. J. Multivar. Anal. *1*, 257–275.
48. Rao, C.R. (1970). Estimation of heteroscedastic variances in linear models. J. Am. Stat. Assoc. *65*, 161–172.
49. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.
50. Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination. Biometrika *78*, 691–692.
51. Fritsche, L.G., Patil, S., Beesley, L.J., VandeHaar, P., Salvatore, M., Ma, Y., Peng, R.B., Taliun, D., Zhou, X., and Mukherjee, B. (2020). Cancer PRSweb: an online repository with polygenic risk scores for major cancer traits and their evaluation in two independent biobanks. Am. J. Hum. Genet. *107*, 815–836.
52. Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era—concepts and misconceptions. Nat. Rev. Genet. *9*, 255–266.
53. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., Robinson, M.R., Perry, J.R.B., Nolte, I.M., van Vliet-Ostaptchouk, J.V., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. *47*, 1114–1120.
54. Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D.I., Cornes, B.K., Davis, C., Dunkel, L., De Lange, M., Harris, J.R., Hjelmborg, J.V.B., et al. (2003). Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res. *6*, 399–408.
55. Johnson, D.A., Billings, M.E., and Hale, L. (2018). Environmental determinants of insufficient sleep and sleep disorders: implications for population health. Curr. Epidemiol. Rep. *5*, 61–69.
56. 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*, 1776–1810.
57. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. *97*, 576–592.
58. Kulm, S., Marderstein, A., Mezey, J., and Elemento, O. (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.
59. 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.
60. 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.
61. Choi, S.W., and O'Reilly, P.F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. GigaScience *8*, giz082.
62. Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat. Rev. Genet. *17*, 392–406.
63. Xue, A., Wu, Y., Zhu, Z., Zhang, F., Kemper, K.E., Zheng, Z., Yengo, L., Lloyd-Jones, L.R., Sidorenko, J., Wu, Y., et al. (2018). Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. Nat. Commun. *9*, 2941–3014.
64. Schillaci, G., and Pucci, G. (2010). The dynamic relationship between systolic and diastolic blood pressure: yet another marker of vascular aging? Hypertens. Res. *33*, 659–661.
65. Gavish, B., Ben-Dov, I.Z., and Bursztyn, M. (2008). Linear relationship between systolic and diastolic blood pressure monitored over 24 h: assessment and correlates. J. Hypertens. *26*, 199–209.
66. Tam, C.H.T., Lim, C.K.P., Luk, A.O.Y., Ng, A.C.W., Lee, H.-m., Jiang, G., Lau, E.S.H., Fan, B., Wan, R., Kong, A.P.S., et al. (2021). Development of genome-wide polygenic risk scores for lipid traits and clinical applications for dyslipidemia, subclinical atherosclerosis, and diabetes cardiovascular complications among East Asians. Genome Med. *13*, 1–18.
67. Timpson, N.J., Nordestgaard, B.G., Harbord, R.M., Zacho, J., Frayling, T.M., Tybjærg-Hansen, A., and Smith, G.D. (2011). C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. Int. J. Obes. *35*, 300–308.
68. Unger, G., Benozzi, S.F., Perruzza, F., and Pennacchiotti, G.L. (2014). Triglycerides and glucose index: a useful indicator of insulin resistance. Endocrinol. Nutr. *61*, 533–540.
69. Beesley, L.J., Fritsche, L.G., and Mukherjee, B. (2020). An analytic framework for exploring sampling and observation process biases in genome and phenome-wide association studies using electronic health records. Stat. Med. *39*, 1965–1979.
70. Farmer, R., Mathur, R., Bhaskaran, K., Eastwood, S.V., Chaturvedi, N., and Smeeth, L. (2018). Promises and pitfalls of electronic health record analysis. Diabetologia *61*, 1241–1248.

71. Gray, N., Picone, G., Sloan, F., and Yashkin, A. (2015). The relationship between BMI and onset of diabetes mellitus and its complications. South. Med. J. *108*, 29–36.
72. Wolk, R., Shamsuzzaman, A.S.M., and Somers, V.K. (2003). Obesity, sleep apnea, and hypertension. Hypertension *42*, 1067–1074.
73. Wolfe, B.M., Kvach, E., and Eckel, R.H. (2016). Treatment of obesity: weight loss and bariatric surgery. Circ. Res. *118*, 1844–1855.
74. Shivakumar, S., Srivastava, A., and C Shivakumar, G. (2018). Body mass index and dental caries: a systematic review. Int. J. Clin. Pediatr. Dent. *11*, 228–232.
75. Coutinho, T., Goel, K., Corrêa de Sá, D., Kragelund, C., Kanaya, A.M., Zeller, M., Park, J.S., Kober, L., Torp-Pedersen, C., Cottin, Y., et al. (2011). Central obesity and survival in subjects with coronary artery disease: a systematic review of the literature and collaborative analysis with individual subject data. J. Am. Coll. Cardiol. *57*, 1877–1886.
76. Ng, R., Sutradhar, R., Yao, Z., Wodchis, W.P., and Rosella, L.C. (2020). Smoking, drinking, diet and physical activity—modifiable lifestyle risk factors and their associations with age to first chronic disease. Int. J. Epidemiol. *49*, 113–130.
77. Wynder, E.L., Williams, C.L., Laakso, K., Levenstein, M., Lippert, P., Hoffmeister, H., Puska, P., Vartiainen, E., Choay, P., and Morla, S. (1981). Screening for risk factors for chronic disease in children from fifteen countries. Prev. Med. *10*, 121–132.
78. Chronic Conditions Data Warehouse. CCW chronic condition categories.
79. Xu, W., Tan, L., Wang, H.F., Jiang, T., Tan, M.S., Tan, L., Zhao, Q.F., Li, J.Q., Wang, J., and Yu, J.T. (2015). Meta-analysis of modifiable risk factors for Alzheimer's disease. J. Neurol. Neurosurg. Psychiatry *86*, 1299–1306.
80. Choi, S.W., and O'Reilly, P.F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. GigaScience *8*, giz082.
81. He, Y., Lakhani, C.M., Rasooly, D., Manrai, A.K., Tzoulaki, I., and Patel, C.J. (2021). Comparisons of polyexposure, polygenic, and clinical risk scores in risk prediction of Type 2 diabetes. Diabetes Care *44*, 935–943.
82. Tanigawa, Y., Qian, J., Venkataraman, G., Justesen, J.M., Li, R., Tibshirani, R., Hastie, T., and Rivas, M.A. (2022). Significant sparse polygenic risk scores across 813 traits in UK Biobank. PLoS Genet. *18*, e1010105.
83. 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*, 373–423.
84. Caldwell, M., Martinez, L., Foster, J.G., Sherling, D., and Hennekens, C.H. (2019). Prospects for the primary prevention of myocardial infarction and stroke. J. Cardiovasc. Pharmacol. Ther. *24*, 207–214.
85. Reis, J.P., Loria, C.M., Sorlie, P.D., Park, Y., Hollenbeck, A., and Schatzkin, A. (2011). Lifestyle factors and risk for new-onset diabetes: a population-based cohort study. Ann. Intern. Med. *155*, 292–299.
86. Guilbert, J.J. (2003). The world health report 2002 - reducing risks, promoting healthy life. Educ. Health *16*, 230.
87. Ellingjord-Dale, M., Papadimitriou, N., Katsoulis, M., Yee, C., Dimou, N., Gill, D., Aune, D., Ong, J.-S., MacGregor, S., Elsworth, B., et al. (2021). Coffee consumption and risk of breast cancer: A Mendelian randomization study. PLoS One *16*, e0236904.
88. Grosso, G., Micek, A., Godos, J., Sciacca, S., Pajak, A., Martínez-González, M.A., Giovannucci, E.L., and Galvano, F. (2016). Coffee Consumption and Risk of All-Cause, Cardiovascular, and Cancer Mortality in Smokers and Non-smokers: A Dose-Response Meta-Analysis (Springer).
89. Xu, J., Chang, W.-S., Tsai, C.-W., Bau, D.-T., Xu, Y., Davis, J.W., Thompson, T.C., Logothetis, C.J., and Gu, J. (2020). Leukocyte telomere length is associated with aggressive prostate cancer in localized prostate cancer patients. EBioMedicine *52*, 102616.
90. Soleimani, A., Pashirzad, M., Avan, A., Ferns, G.A., Khazaei, M., and Hassanian, S.M. (2019). Role of the transforming growth factor-β signaling pathway in the pathogenesis of colorectal cancer. J. Cell. Biochem. *120*, 8899–8907.
91. Kubiczkova, L., Sedlarikova, L., Hajek, R., and Sevcikova, S. (2012). TGF-β–an excellent servant but a bad master. J. Transl. Med. *10*, 183–224.
92. Wang, H., Peng, R., Wang, J., Qin, Z., and Xue, L. (2018). Circulating microRNAs as potential cancer biomarkers: the advantage and disadvantage. Clin. Epigenetics *10*, 1–10.
93. 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.
94. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2020). LDpred2: better, faster, stronger. Bioinformatics *36*, 5424–5431.
95. Chun, S., Imakaev, M., Hui, D., Patsopoulos, N.A., Neale, B.M., Kathiresan, S., Stitziel, N.O., and Sunyaev, S.R. (2020). Non-parametric polygenic risk prediction via partitioned GWAS summary statistics. Am. J. Hum. Genet. *107*, 46–59.
96. Privé, F., Vilhjálmsson, B.J., Aschard, H., and Blum, M.G.B. (2019). Making the most of clumping and thresholding for polygenic scores. Am. J. Hum. Genet. *105*, 1213–1221.
97. Márquez-Luna, C., Loh, P.R., South Asian Type 2 Diabetes SAT2D Consortium; and SIGMA Type 2 Diabetes Consortium, and Price, A.L. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet. Epidemiol. *41*, 811–823.
98. Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X., and Zhao, H. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput. Biol. *13*, e1005589.
99. Maier, R., Moser, G., Chen, G.B., Ripke, S., Cross-Disorder Working Group of the Psychiatric Genomics, C., Coryell, W., Potash, J.B., Scheftner, W.A., Shi, J., Weissman, M.M., et al. (2015). Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am. J. Hum. Genet. *96*, 283–294.
100. 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.
101. 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.

102. Fritsche, L.G., Ma, Y., Zhang, D., Salvatore, M., Lee, S., Zhou, X., and Mukherjee, B. (2021). On cross-ancestry cancer polygenic risk scores. PLoS Genet. *17*, e1009670.
103. Fahed, A.C., Aragam, K.G., Hindy, G., Chen, Y.D.I., Chaudhary, K., Dobbyn, A., Krumholz, H.M., Sheu, W.H.H., Rich, S.S., Rotter, J.I., et al. (2021). Transethnic transferability of a genome-wide polygenic score for coronary artery disease. Circ. Genom. Precis. Med. *14*, e003092.
104. 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.
105. Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019). The missing diversity in human genetic studies. Cell *177*, 1080–1131.
106. Zeng, P., Shao, Z., and Zhou, X. (2021). Statistical methods for mediation analysis in the era of high-throughput genomics: current successes and future challenges. Comput. Struct. Biotechnol. J. *19*, 3209–3224.
107. Guo, Y., Warren Andersen, S., Shu, X.-O., Michailidou, K., Bolla, M.K., Wang, Q., Garcia-Closas, M., Milne, R.L., Schmidt, M.K., Chang-Claude, J., et al. (2016). Genetically predicted body mass index and breast cancer risk: Mendelian randomization analyses of data from 145, 000 women of European descent. PLoS Med. *13*, e1002105.
108. Shen, X., Howard, D.M., Adams, M.J., Hill, W.D., Clarke, T.K., Major Depressive Disorder Working Group of the Psychiatric Genomics, C., Deary, I.J., Whalley, H.C., and McIntosh, A.M. (2020). A phenome-wide association and Mendelian Randomisation study of polygenic risk for depression in UK Biobank. Nat. Commun. *11*, 2301.
109. Richardson, T.G., Harrison, S., Hemani, G., and Davey Smith, G. (2019). An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. Elife *8*, e43657.
110. Burgess, S., Butterworth, A., and Thompson, S.G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. Genet. Epidemiol. *37*, 658–665.
111. Morrison, J., Knoblauch, N., Marcus, J.H., Stephens, M., and He, X. (2020). Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nat. Genet. *52*, 740–747.
112. Shi, J., Swanson, S.A., Kraft, P., Rosner, B., De Vivo, I., and Hernán, M.A. (2021). Instrumental variable estimation for a time-varying treatment and a time-to-event outcome via structural nested cumulative failure time models. BMC Med. Res. Methodol. *21*, 258.
113. Beesley, L.J., Salvatore, M., Fritsche, L.G., Pandit, A., Rao, A., Brummett, C., Willer, C.J., Lisabeth, L.D., and Mukherjee, B. (2020). The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. Stat. Med. *39*, 773–800.

The American Journal of Human Genetics, Volume *109*

Supplemental information

ExPRSweb: An online repository with polygenic risk scores for common health-related exposures

Ying Ma, Snehal Patil, Xiang Zhou, Bhramar Mukherjee, and Lars G. Fritsche

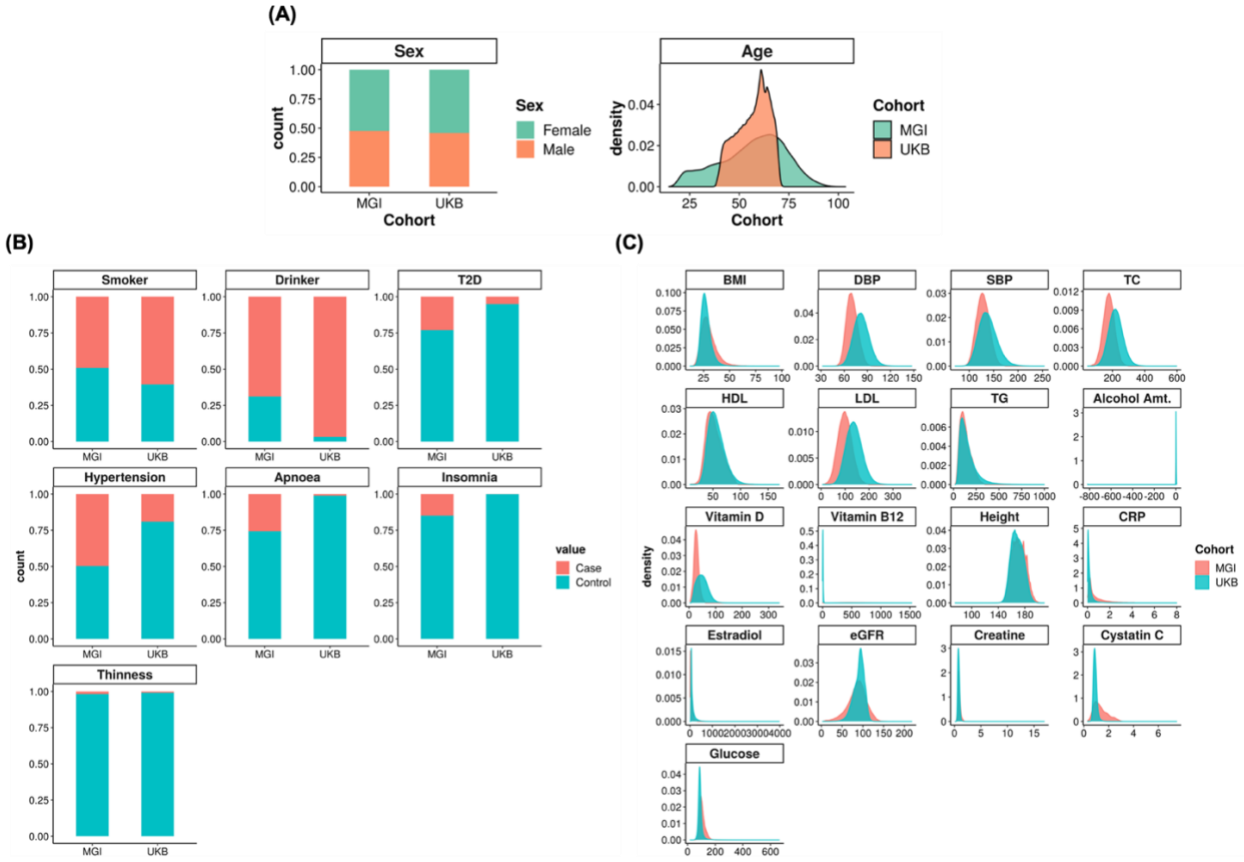

Figure S1. Descriptive statistics comparing traits in MGI and UKB. For continuous traits, the plot of distribution density was shown. For binary traits, box plots were shown.

Figure S2. Heritability estimates of 79 collected summary statistics with positive heritability estimates on the liability scale. Here, three summary statistics were excluded due to the negative estimate on the heritability. Detailed statistics see **Table S3**.

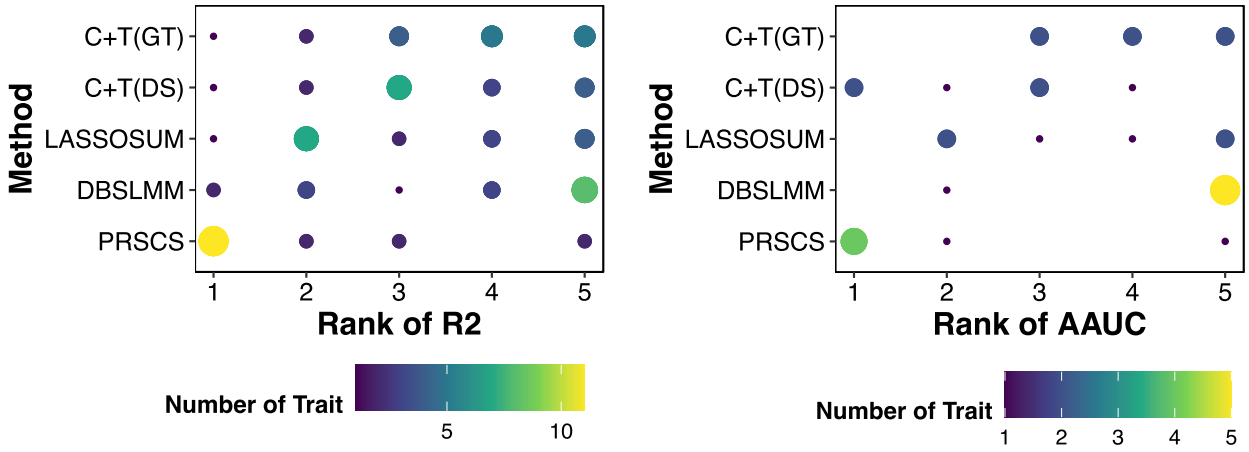

Figure S3. Rank plot of prediction performance of different PRS methods in MGI across ten continuous (left) and six binary traits (right). For each trait, all methods were ranked according to their prediction performance (R^2 for continuous traits and covariates adjusted AUC for binary traits) in MGI traits. The summed number of achieved ranks is depicted by color and circle size. Method that generated non-significant PRS were ranked as five. For a fair comparison we selected the same summary statistic for each method (GWAS with the highest heritability estimate).

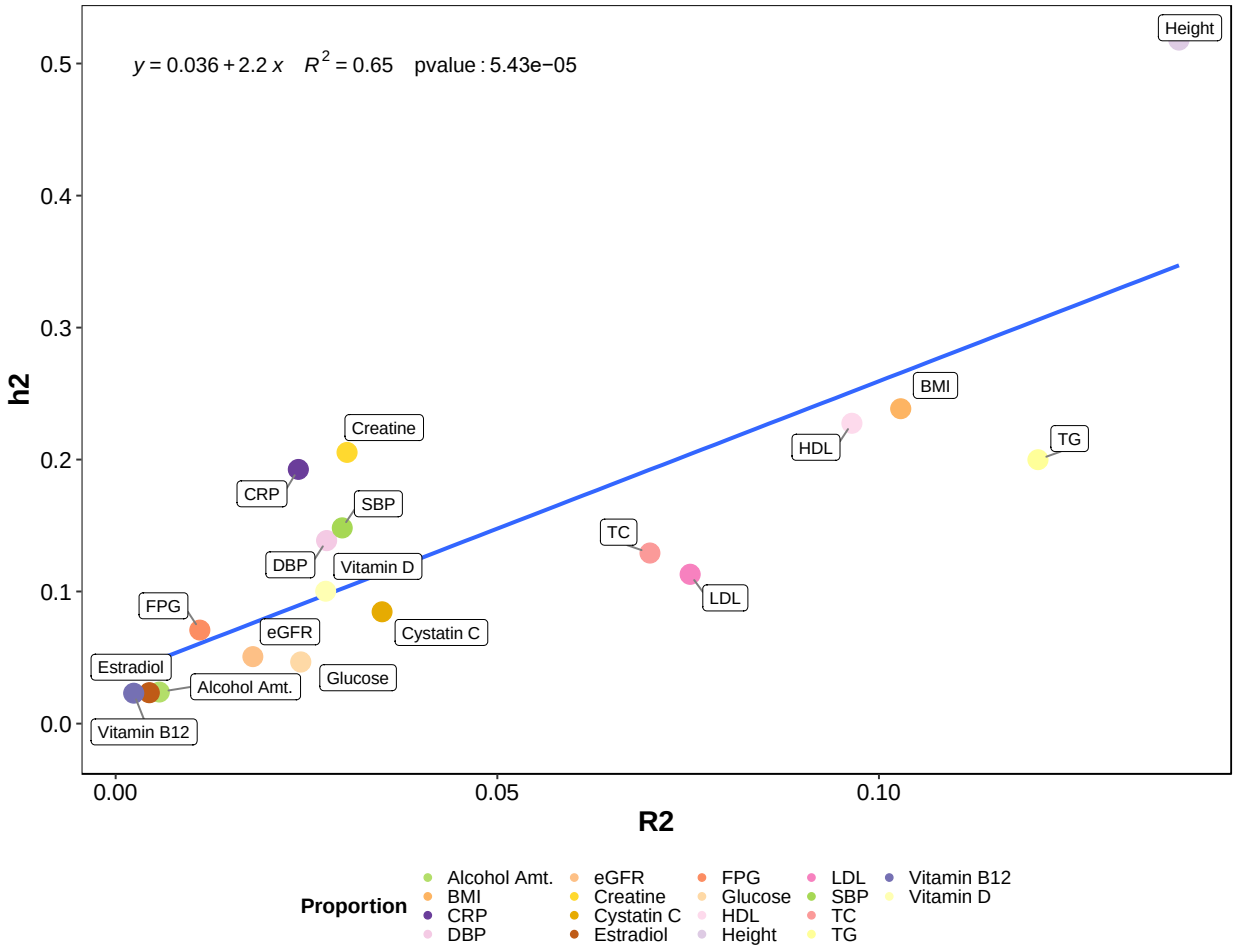

Figure S4. The relationship between heritability estimates and prediction R^2 for the summary statistics that generate the best PRS for a specific quantitative exposure trait in MGI cohort.

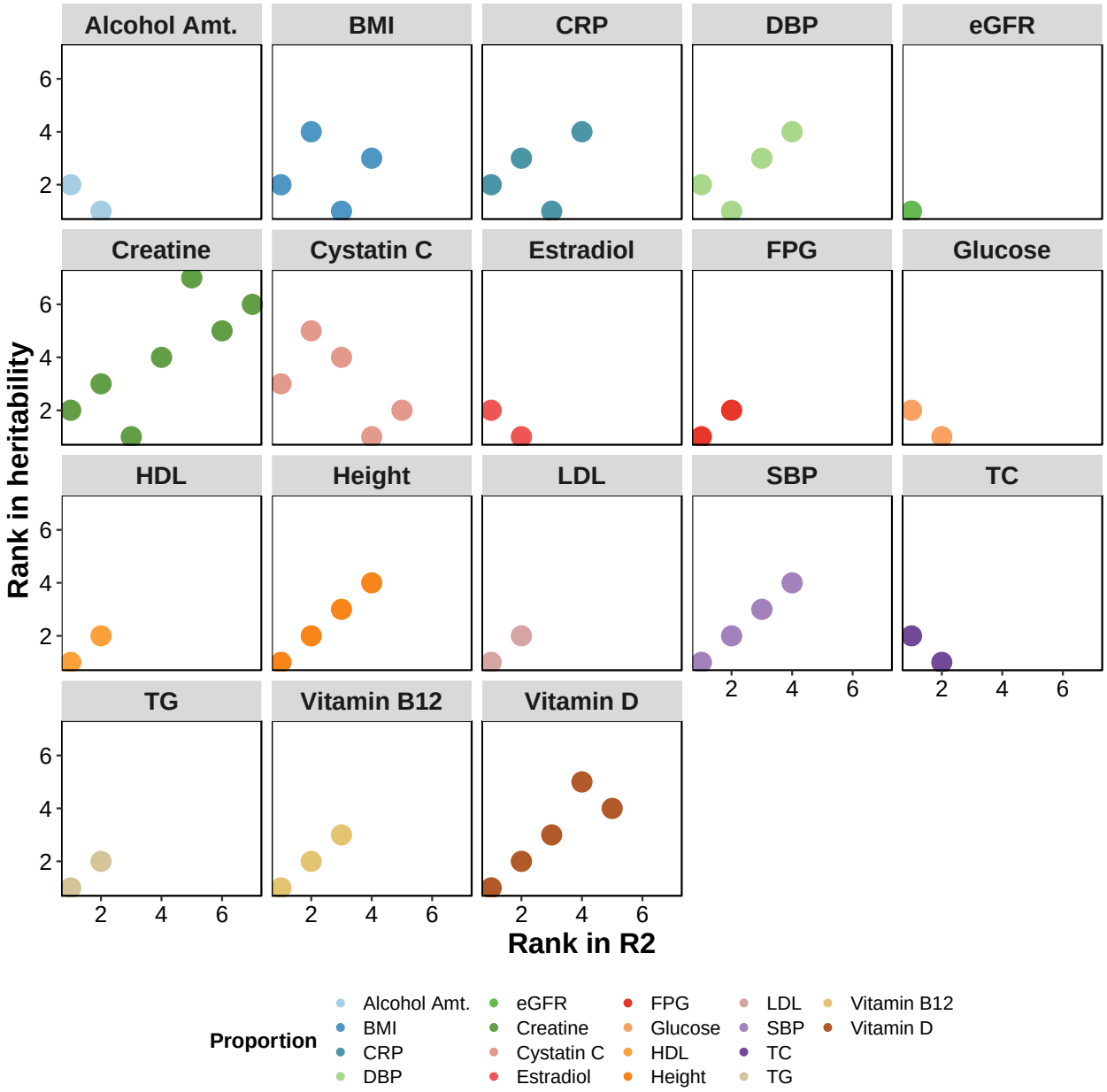

Figure S5. The relationship between the rank of heritability estimates and the rank for prediction R^2 of the constructed PRS from each summary statistic across 18 quantitative traits in MGI.

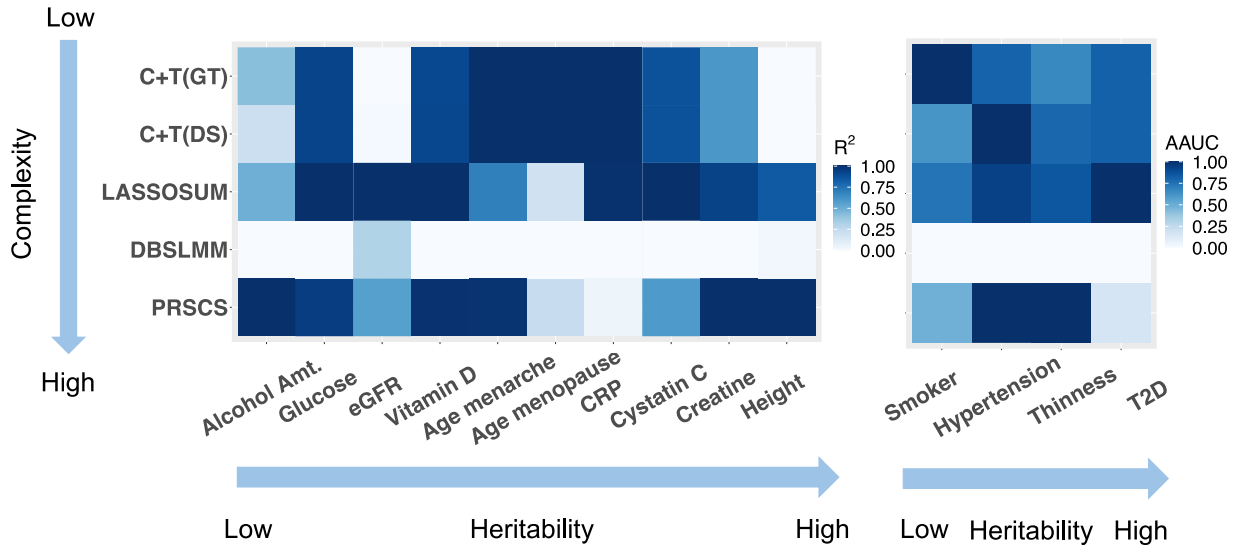

Figure S6. Prediction performance of different PRS methods in UKB across traits. (A) The color represents the absolute prediction performance for each method across traits. (B) The color is scaled to 0-1 range and represents the relative prediction performance for each method across traits. For both (A) and (B), the prediction performance is quantified as R^2 for continuous traits AUC for binary traits. We keep the order of the traits corresponding to the same order of traits in MGI evaluation. For the traits that are not available or the traits that have no best performing PRS in MGI (e.g., age menarche, age menopause, and Cystatin C), we order them by their maximum heritability estimates. Here the summary statistic that generates the best performing exposure PRS was selected for comparison.

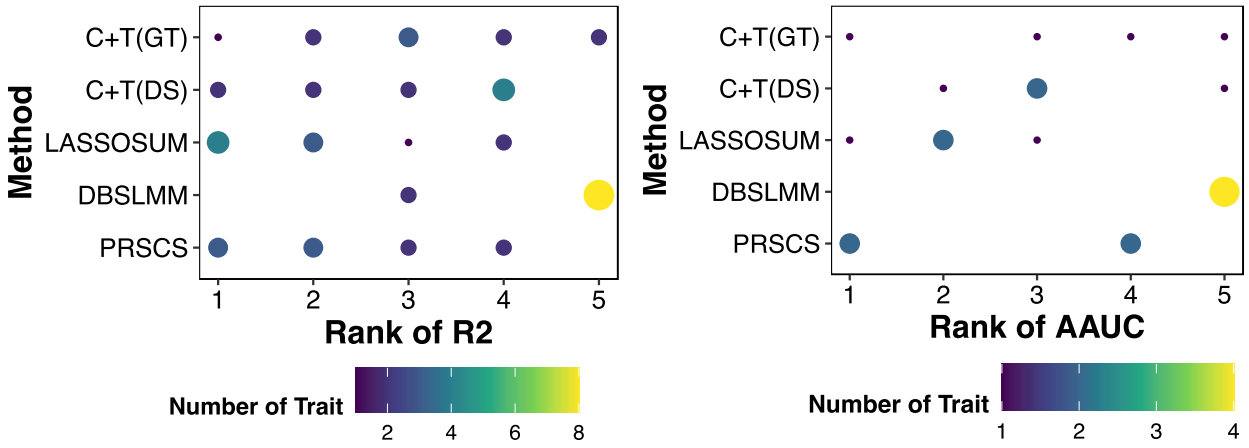

Figure S7. Rank plot of prediction performance of different PRS methods in UKB across traits. For each trait, all methods were ranked according to their prediction performance (R^2 for continuous traits and AUC for binary traits) in UKB traits. The unique ranks each method achieved is shown, colored according to the number of diseases corresponding to that rank. For the method that generated non-significant PRS, we rank it as five. Here the summary statistic that generates the best performing exposure PRS was selected for comparison.

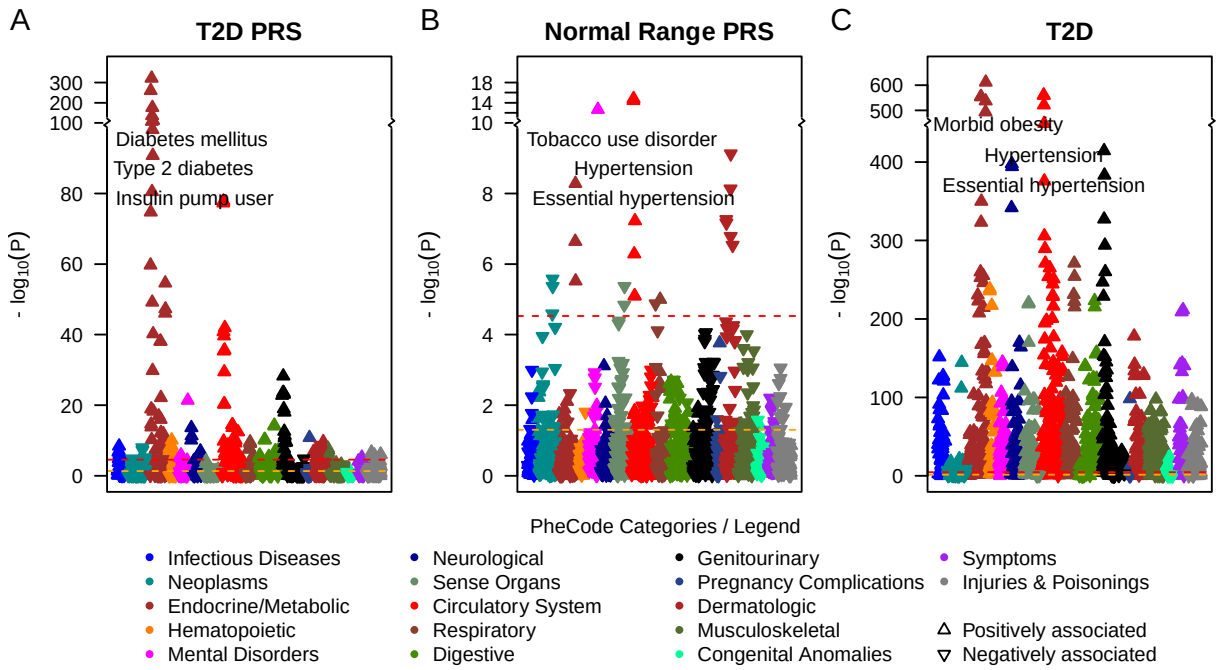

Figure S8. PRS PheWAS and exclusion PRS PheWAS for an example binary trait in MGI. (A) PRS PheWAS plot is shown for type 2 diabetes PRS predictor (B) Exclusion PRS PheWAS plot is shown for non-type 2 diabetes PRS predictor (C) PRS PheWAS plot is shown for type 2 diabetes predictor (D) Exclusion PRS PheWAS plot is shown for non-type 2 diabetes.

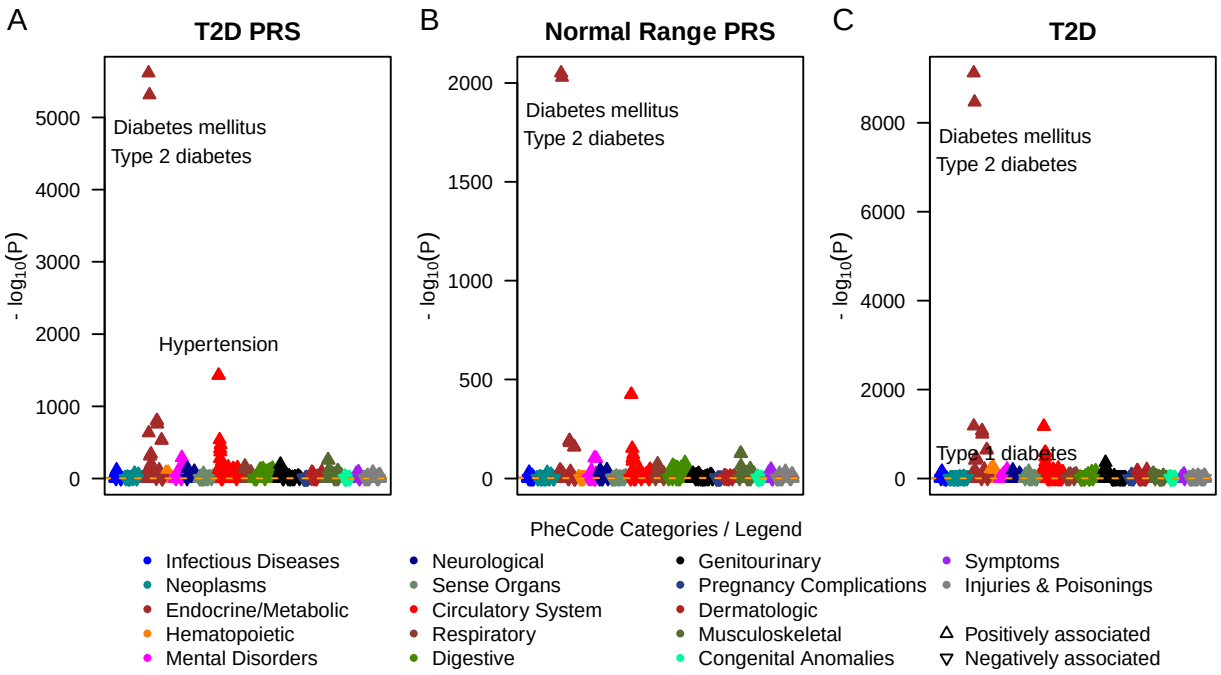

Figure S9. PRS PheWas and exclusion PRS PheWas for example binary trait in UKB. (A) PRS PheWAS plot is shown for T2D PRS predictor (B) Exclusion PRS PheWAS plot is shown for non- T2D PRS predictor (C) Trait PheWAS plot is shown for T2D predictor

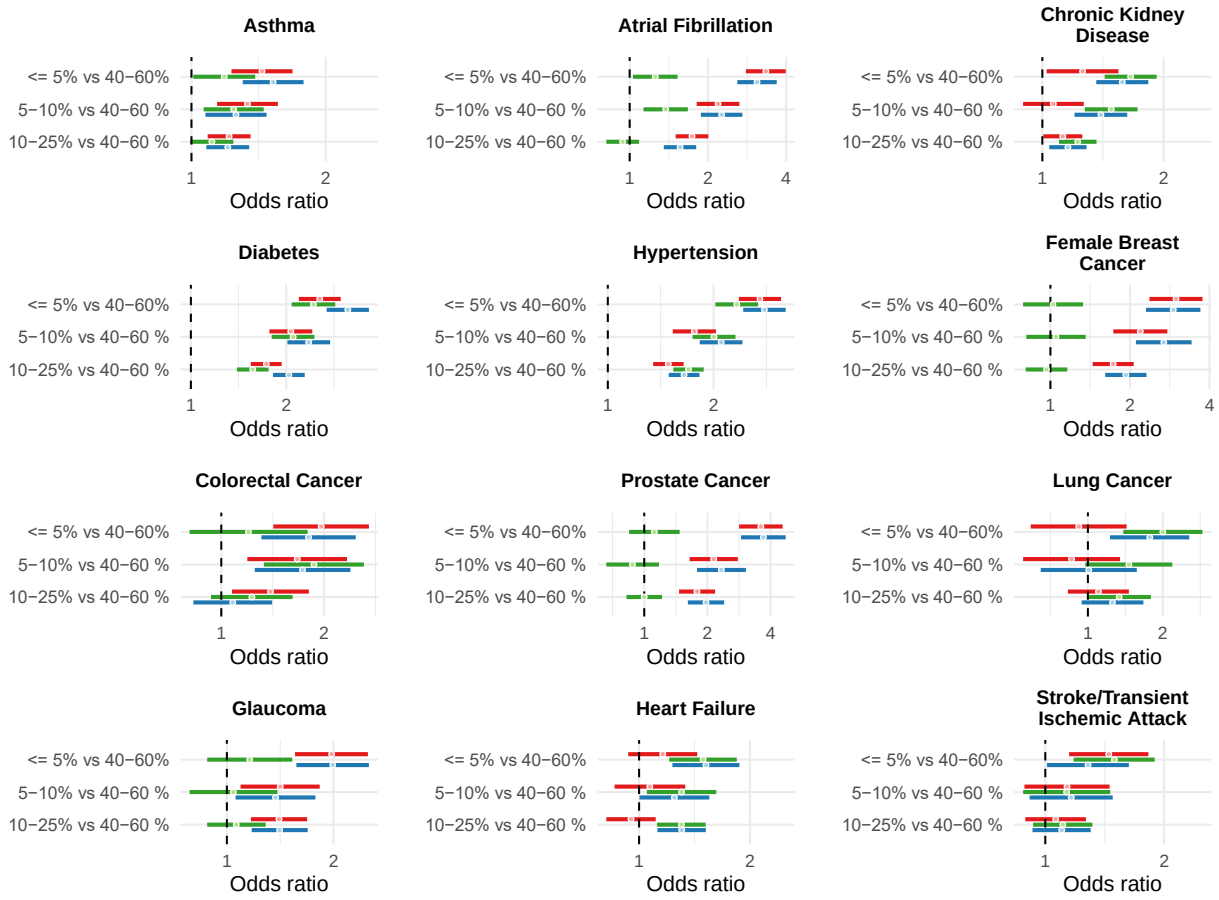

Figure S10. Utility of exposure PRS in risk stratification for 12 chronic conditions. Odds ratio of observing cases in selected top bins of the PRS distributions (Top 5%, Top 5 – 10%, and Top 10-25%) versus the center 40-60% bin for condition specific PRS (YPRS, red), multipleExPRS (green), and YPRS + multipleExPRS (blue) are shown for each condition. Details can be found in **Table S18**.

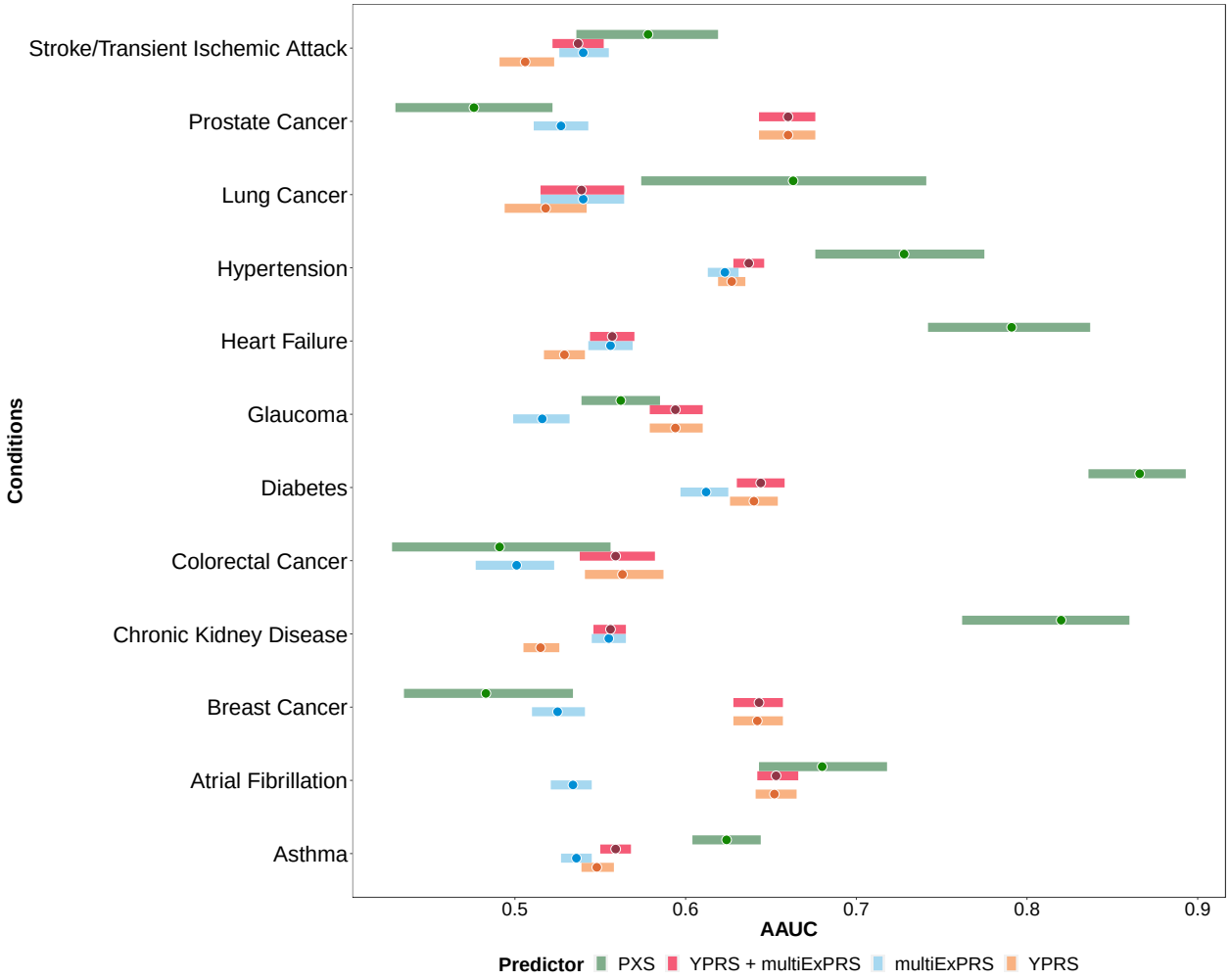

Figure S11. Comparisons of the prediction performance of genetic predictors and the exposure trait score (PXS) for common chronic conditions in MGI cohort. AAUC paired with 95% confidence interval for condition PXS (green), specific PRS (red), exposure PRS (blue) and trait + exposure PRS (orange) were shown in the format of forest plot. Each bar represents the 95% interval for the AAUC with the dot represents the AAUC estimate.

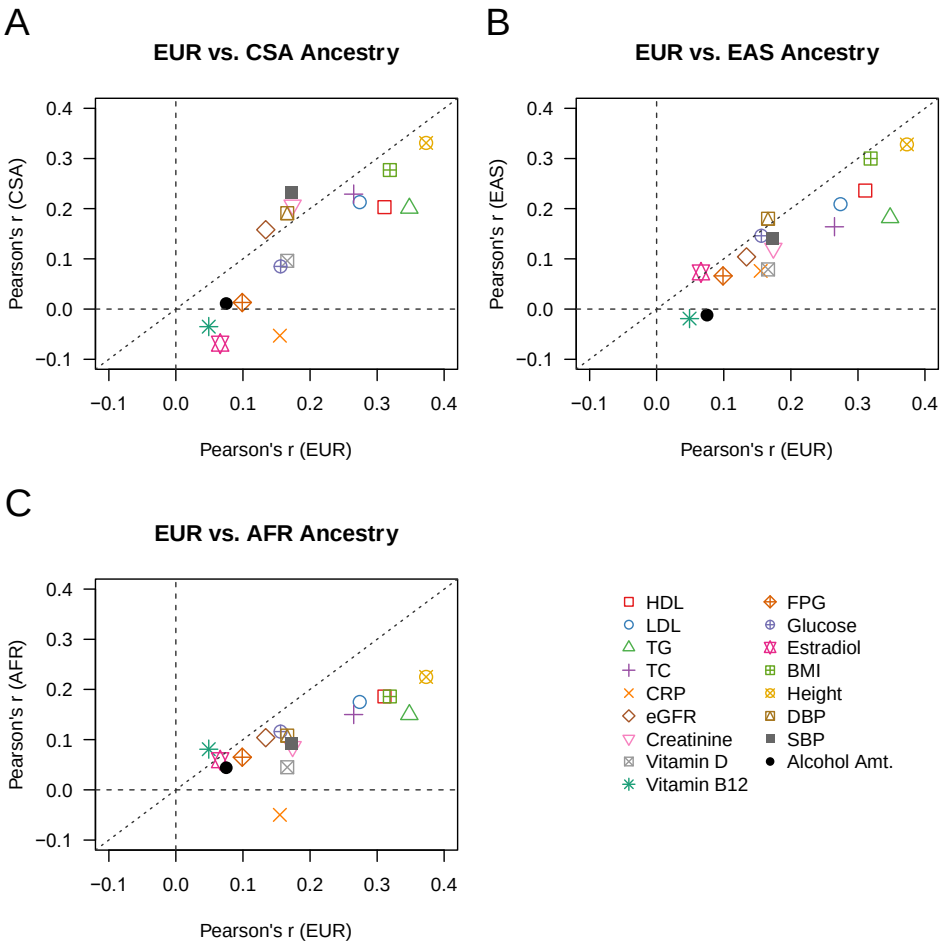

Figure S12. Comparison of the pairwise Pearson's correlation of 17 top ranked ExPRS and their corresponding continuous traits between European (EUR) and non-EUR ancestry groups (A: Central South Asian [CSA], B: East Asian [EAS], and C: African [AFR]) in MGI. Details can also be found in **Table S23**.

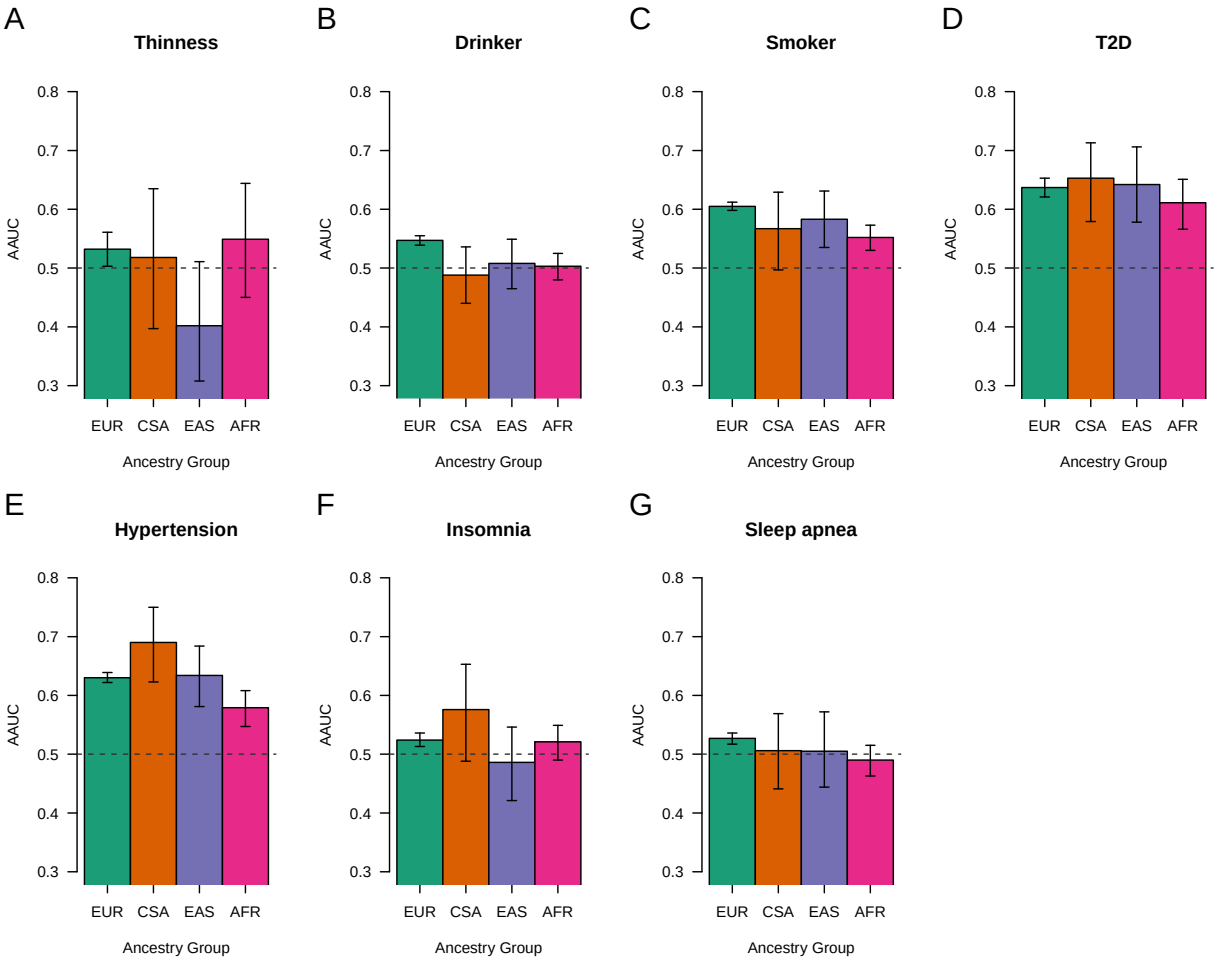

Figure S13. Comparison of the adjusted AUC (AAUC) of 7 top ranked ExPRS and their corresponding binary traits (A: thinness, B: alcohol drinker status, C: smoking status, D: type 2 diabetes [T2D], E: hypertension, F: insomnia, and G: sleep apnea) between European (EUR) and non-EUR ancestry groups (Central South Asian [CSA], East Asian [EAS], and African [AFR]) in MGI. Details can also be found in **Table S24.**

Note S1. Transferability of ExPRS

Considering the lack of discovery exposure GWAS and small sample sizes for parameter tuning (further reduced by the missingness of measured exposure data, Table R1), we explored a compromise solution in MGI, namely the use of the derived EUR-based ExPRS for non-EUR ancestry groups after scaling the ExPRS within each ancestry group [77]. We found that the majority (CSA: 14 of 17; EAS: 15 of 17, and AFR: 16 of 17) of the ExPRS for continuous exposures showed a positive correlation with their corresponding exposures (**Table S23 and S24**; **Figure S12**). Among the seven ExPRS for binary exposures, only the ExPRS for smoking status, type 2 diabetes, and hypertension, showed AAUC values in non-EUR ancestry groups that were comparable to the EUR group, though the confidence intervals were due to the small sample sizes substantially wider (**Table S23** and **S25**; **Figure S13**).