# Risk factors affecting polygenic score performance across diverse cohorts

Daniel Hui[1], Scott Dudek[1], Krzysztof Kiryluk[2], Theresa L. Walunas[3], Iftikhar J. Kullo[4], Wei-Qi Wei[5], Hemant K. Tiwari[6], Josh F. Peterson[5], Wendy K. Chung[7], Brittney Davis[8], Atlas Khan[2], Leah Kottyan[9], Nita A. Limdi[8], Qiping Feng[10], Megan J. Puckelwartz[11], Chunhua Weng[12], Johanna L. Smith[4], Elizabeth W. Karlson[13], Regeneron Genetics Center[14], Gail P. Jarvik[15], Marylyn D. Ritchie[1]

[1]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.
[2]Division of Nephrology, Department of Medicine, Columbia University, NY, New York.
[3]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA.
[4]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN.
[5]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN.
[6]Department of Pediatrics, University of Alabama at Birmingham, Birmingham, AL.
[7]Departments of Pediatrics and Medicine, Columbia University Irving Medical Center, Columbia University, New York, NY.
[8]Department of Neurology, School of Medicine, University of Alabama at Birmingham, Birmingham, AL.
[9]The Center for Autoimmune Genomics and Etiology, Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH.
[10]Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA.
[11]Center for Genetic Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL.
[12]Department of Biomedical Informatics, Vagelos College of Physicians & Surgeons, Columbia University, New York, NY.
[13]Division of Rheumatology, Inflammation, and Immunity, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA.
[14]Regeneron Pharmaceuticals Inc., Tarrytown, NY.
[15]Departments of Medicine (Medical Genetics) and Genome Sciences, University of Washington Medical Center, Seattle, WA.

Correspondence to:
Marylyn D. Ritchie, PhD
marylyn@pennmedicine.upenn.edu

**Abstract**

Apart from ancestry, personal or environmental covariates may contribute to differences in polygenic score (PGS) performance. We analyzed effects of covariate stratification and interaction on body mass index (BMI) PGS ($PGS_{BMI}$) across four cohorts of European (N=491,111) and African (N=21,612) ancestry. Stratifying on binary covariates and quintiles for continuous covariates, 18/62 covariates had significant and replicable $R^2$ differences among strata. Covariates with the largest differences included age, sex, blood lipids, physical activity, and alcohol consumption, with $R^2$ being nearly double between best and worst performing quintiles for certain covariates. 28 covariates had significant $PGS_{BMI}$-covariate interaction effects, modifying $PGS_{BMI}$ effects by nearly 20% per standard deviation change. We observed overlap between covariates that had significant $R^2$ differences among strata and interaction effects – across all covariates, their main effects on BMI were correlated with their maximum $R^2$ differences and interaction effects (0.56 and 0.58, respectively), suggesting high-$PGS_{BMI}$ individuals have highest $R^2$ and increase in PGS effect. Using quantile regression, we show the effect of $PGS_{BMI}$ increases as BMI itself increases, and that these differences in effects are directly related to differences in $R^2$ when stratifying by different covariates. Given significant and replicable evidence for context-specific $PGS_{BMI}$ performance and effects, we investigated ways to increase model performance taking into account non-linear effects. Machine learning models (neural networks) increased relative model $R^2$ (mean 23%) across datasets. Finally, creating $PGS_{BMI}$ directly from GxAge GWAS effects increased relative $R^2$ by 7.8%. These results demonstrate that certain covariates, especially those most associated with BMI, significantly affect both $PGS_{BMI}$ performance and effects across diverse cohorts and ancestries, and we provide avenues to improve model performance that consider these effects.

## Introduction

Polygenic scores (PGS) provide individualized genetic predictors of a phenotype by aggregating genetic effects across hundreds or thousands of loci, typically estimated from genome-wide association studies (GWAS). In recent years it has become increasingly apparent that the transferability of PGS performance across different cohorts is poor (1). Most analyses to date have focused on ancestry differences as the main driver of this lack of portability (2–4). However, a growing body of evidence has demonstrated that PGS performance and effect estimates are influenced by differences in certain contexts i.e., environmental (classically termed "gene-environment" effects or interactions) or personal-level covariates – different phenotypes seem to be differently affected by these covariates, with adiposity traits such as body mass index (BMI) having substantial evidence for these effects (5–14). In one previous study, they showed that GWAS stratified by sample characteristics had better PGS performance in cohorts that matched the sample characteristics of the stratified GWAS, and that differences in heritability between the stratified cohorts partially explained this observation (13).

There are several gaps in current knowledge about these covariate-specific effects. Many analyses have assessed only a handful of these covariates due to the myriad of choices possible in typical large-scale biobanks. Little investigation has been done to systematically understand why certain covariates affect PGS performance, with such knowledge being useful to reduce the potential search for variables that impart context-specific effects. Furthermore, most studies investigating PGS-covariate interactions have been in European ancestry individuals; notably, comparing differences in PGS performance and prediction while controlling for differences in ancestry versus differences in context has not been assessed in previous studies. Moreover, covariate-specific effects are notorious for replicating poorly in human genetics studies, and previous studies of PGS-covariate interactions have been predominantly performed in the UK Biobank (UKBB) (15), where the majority of individuals are aged 40-69 (i.e., excluding young adults), are overall healthier than those from other e.g., hospital-based cohorts, and are predominantly European ancestry. Additionally, PGS performance is often assessed using linear models and in isolation of clinical covariates, which in practice would often be available. Machine learning models can have increased performance over linear models and are capable of modeling complex relationships and interactions between variables, which may serve to increase predictive performance, especially given evidence for PGS-covariate specific effects. Finally, given evidence for context-specific effects, it should be possible to directly incorporate SNP-covariate interaction effects from a GWAS directly to improve prediction performance, instead of relying on post-hoc interactions from a typical PGS calculated from main GWAS effects.

Using genetic data with linked-phenotypic information from electronic health records, we estimated the effects of covariate stratification and interaction on performance and effect estimates of PGS for BMI ($PGS_{BMI}$) – a flowchart summarizing our analyses is presented in Figure 1. These analyses were done across four datasets (Supplemental Table 1): UK Biobank (UKBB), Penn Medicine BioBank (PMBB) (15), Electronic Medical Records and Genomics (eMERGE) network dataset (16), and Genetic Epidemiology Research on Adult Health and Aging (GERA). These datasets include participants from two ancestry groups (N=491,111 European ancestry (EUR), N=21,612 African ancestry (AFR)), and 62 covariates (25 present in multiple datasets) representing laboratory, survey, and biometric data types typically associated with cardiometabolic health and adiposity. After constructing $PGS_{BMI}$ using out-of-sample multi-ancestry BMI GWAS, we assessed effects of covariate stratification on $PGS_{BMI}$ $R^2$, the significance of $PGS_{BMI}$-covariate interaction terms and their increases to model $R^2$ over models

only using main effects, as well as correlation of main effect, interaction effect, and $R^2$ differences. We then assessed ways to increase model performance through using machine learning models, and creating $PGS_{BMI}$ using GxAge GWAS effects. This study addresses a plethora of open issues considering performance and effects of PGS on individuals from diverse backgrounds.

## Results

Effect of covariate stratification on $PGS_{BMI}$ performance

We assessed 62 covariates for $PGS_{BMI}$ $R^2$ differences (25 present, or suitable proxies, in multiple datasets (Supplemental Table 2) after stratifying on binary covariates and quintiles for continuous covariates. With UKBB EUR as discovery (N=376,729), 18 covariates had significant differences (Bonferroni $p<.05/62$) in $R^2$ among groups (Figure 2a), including age, sex, alcohol consumption, different physical activity measurements, Townsend deprivation index, different dietary measurements, lipids, blood pressure, and HbA1c, with 40 covariates having suggestive ($p<.05$) evidence of $R^2$ differences. From an original $PGS_{BMI}$ $R^2$ of 0.076, $R^2$ increased to 0.094-0.088 for those in the bottom physical activity, alcohol intake, and high-density lipoprotein (HDL) cholesterol quintiles, and decreased to 0.067-0.049 for those in the top quintile, respectively, comparable to differences observed between ancestries (1). We note that the differences in $R^2$ due to alcohol intake and HDL were larger than those of any physical activity phenotype, despite physical activity having one of the oldest and most replicable evidence of interaction with genetic effects of BMI (17,18). Despite considerable published evidence suggesting covariate-specific genetic effects between BMI and smoking behaviors (6,8), we were only able to find suggestive evidence for $R^2$ differences when stratifying individuals across several smoking phenotypes (minimum $p=0.016$, for smoking pack years). $R^2$ differences due to educational attainment were also only suggestive ($p=0.015$), with published evidence on this association being conflicting (19–21).

We replicated these analyses in three additional large-scale cohorts of European and African ancestry individuals (Figure 2b, Supplemental Table 3), as well as in African ancestry UKBB individuals. Among covariates with significant performance differences in the discovery analysis, we were able to replicate significant ($p<.05$) $R^2$ differences for age, HDL cholesterol, alcohol intake frequency, physical activity, and HbA1c, despite much smaller sample sizes. We again observed mostly insignificant differences across cohorts and ancestries when stratifying due to different smoking phenotypes and educational attainment. For each covariate and ancestry combination, we combined data across cohorts and conducted a linear regression weighted by sample size, regressing $R^2$ values on covariate values across groupings. Slopes of the regressions across cohorts had different signs between ancestries for the same covariate (triglyceride levels, HbA1c, diastolic blood pressure, and sex), although larger sample sizes may be needed to confirm these differences are statistically significant.

Several observations related to age-specific effects on $PGS_{BMI}$ we considered noteworthy. First, in the weighted linear regression of all $R^2$ values across ancestries, expected $R^2$ for African ancestry individuals can become greater than that of European ancestry individuals among individuals within bottom and top age quintiles observed in these data. For instance, the predicted $R^2$ of 0.048 for 80 year-old European ancestry individuals would be lower than that of African ancestry individuals aged 24.7 and lower, indicating that differences in covariates can affect $PGS_{BMI}$ performance more than differences due to ancestry. Second, we obtained these results despite the average age of GWAS individuals being 57.8, which should increase $PGS_{BMI}$

$R^2$ for individuals closest to this age (13). This result suggests that PGS performance due to decreased heritability with age cannot be fully reconciled using GWAS from individuals of similar age being used to create $PGS_{BMI}$ (as heritability is an upper bound on PGS performance). Finally, we observed that $PGS_{BMI}$ $R^2$ increases as age decreases, consistent with published evidence suggesting that the heritability of BMI decreases with age (22,23).

PGS-covariate interaction effects

Next, we estimated difference in PGS effects due to interactions with covariates, by modeling interaction terms between $PGS_{BMI}$ and the covariate for each covariate in our list (described in Methods). We implemented a correction for shared heritability between covariates of interest and outcome (which can inflate test statistics (24)) to better measure the environmental component of each covariate, and show that this correction successfully reduces significance of interaction estimates (Supplemental Figure 1). Again, using UKBB EUR as the discovery cohort, we observed 28 covariates with significant (Bonferroni $p<.05/62$) PGS-covariate interactions (Table 1), with 38 having suggestive ($p<.05$) evidence (Supplemental Table 4). We observed the largest effect of PGS-covariate interaction with alcohol drinking frequency (20.0% decrease in PGS effect per 1 standard deviation (SD) increase, $p=2.62x10^{-55}$), with large effects for different physical activity measures (9.4%-12.5% decrease/SD, minimum $p=3.11x10^{-66}$), HDL cholesterol (15.3% decrease/SD, $p=1.71x10^{-96}$) and total cholesterol (12.7% decrease/SD, $p=1.64x10^{-71}$). We observed significant interactions with diastolic blood pressure (10.8% increase/SD, $p=6.06x10^{-60}$), but interactions with systolic blood pressure were much smaller (1.17% increase/SD, $p=4.41x10^{-3}$). Significant interactions with HbA1c (4.63% increase/SD, $p=5.37x10^{-14}$) and type 2 diabetes (27.2% PGS effect increase in cases, $p=1.83x10^{-7}$) were also observed. Other significant PGS-covariate interactions included lung function, age, sex, and LDL cholesterol – various dietary measurements also had significant interactions, albeit with smaller effects than other significant covariates. We were able to find significant interaction effects for smoking pack years (4.78% increase/SD, $p=3.68x10^{-7}$), but other smoking phenotypes had insignificant interaction effects after correcting for multiple tests (minimum $p=2.7x10^{-3}$); interactions with educational attainment were also insignificant ($p= 4.54x10^{-2}$).

We replicated these analyses across ancestries and the other non-UKBB EUR cohorts (Figure 3, Supplemental Table 4). For age and sex, which were available for all cohorts, interactions were significant ($p<.05$) and directionally consistent across cohorts and ancestries (except for GERA AFR which had small sample size (N=1,789)). We were able to test interactions with alcohol intake frequency and physical activity in GERA, and replicated significant and directionally consistent associations. We observed poor replication for LDL cholesterol, HbA1c, and smoking pack years, with insignificant and directionally inconsistent interaction effects across cohorts. Educational attainment was available in GERA, and interactions were once again insignificant. We observed significant and directionally consistent interaction effects for TG in eMERGE EUR and PMBB EUR, while the effect was inconsistent in UKBB EUR despite much larger sample size.

However, despite significance of interaction terms, increases in model $R^2$ when including PGS-covariate interaction terms were small. For instance, the maximum increase among all covariates in UKBB EUR was only 0.0024 from a base $R^2$ of 0.1049 (2.1% relative increase), for alcoholic drinks per week. Across all cohorts and ancestries, the maximum increase in $R^2$ was only 0.0058 from a base $R^2$ of 0.09454 (6.1% relative increase), when adding a PGS-age interaction term for eMERGE EUR ($p=5.40x10^{-46}$) – this was also the largest relative increase

among models with significant interaction terms. This result suggests that, while interaction effects can significantly modify $PGS_{BMI}$ effect, their overall impact on model performance is relatively small, despite large differences in $R^2$ when stratifying by covariates.

Correlations between $R^2$ differences, interaction effects, and main effects

We next investigated the relationship between interaction effects, maximum $R^2$ differences across quintiles, and main effects of covariates on BMI. We first estimated main effects of each covariate on BMI (Methods, Supplemental Table 5), and then calculated the correlation weighted by sample size between main effects, maximum $PGS_{BMI}$ $R^2$ across quintiles, and PGS-covariate interaction effects (Figure 4) across all cohorts and ancestries – GERA data were excluded from these analyses due to slightly different phenotype definition (Supplemental Table 6), as were binary variables. Interaction effects and maximum $R^2$ differences had a 0.80 correlation ($p=2.1x10^{-27}$), indicating that variables with larger interaction effects also had larger effects on $PGS_{BMI}$ performance across quintiles, and that covariates that increase $PGS_{BMI}$ effect also have the largest effect on $PGS_{BMI}$ performance i.e., individuals most at risk for obesity will have both disproportionately larger $PGS_{BMI}$ effect and $R^2$. Main effects and maximum $R^2$ differences had a 0.56 correlation ($p=1.3x10^{-11}$), while main effects and interaction effects had a 0.58 correlation ($p=7.6x10^{-12}$) again suggesting that $PGS_{BMI}$ are more predictive in individuals with higher values of BMI-associated covariates, although less predictive than estimating the interaction effects themselves directly. However, this result demonstrates that covariates that influence both $PGS_{BMI}$ effect and performance can be predicted just using main effects of covariates, which are often known for certain phenotypes and easier to calculate, as genetic data and PGS construction would not be required.

Increase in PGS effect for increasing percentiles of BMI itself, and its relation to $R^2$ differences when stratifying by covariates

Given large and replicable correlations between main effects, interaction effects, and maximum $R^2$ differences for individual covariates, it seemed these differences may be due to the differences in BMI itself, rather than any individual or combination of covariates. To assess this, we used quantile regression to evaluate the effect of $PGS_{BMI}$ on BMI at different deciles of BMI itself. We observed that the effect of $PGS_{BMI}$ consistently increases from lower deciles to higher deciles across all cohorts and ancestries (Figure 5) – for instance, in European ancestry UKBB individuals, the effect of $PGS_{BMI}$ (in units of log(BMI)) when predicting the bottom decile of log(BMI) was 0.716 (95% CI: 0.701-.732), and increased to 1.31 (95% CI: 1.29-1.33) in the top decile. Across all cohorts and ancestries, the effect of $PGS_{BMI}$ between lowest and highest effect decile ranged from 1.43-2.06 times larger, with all cohorts and ancestries having non-overlapping 95% confidence intervals between their effects (except for African ancestry eMERGE individuals, which had much smaller sample size).

While this analysis showed that the effect of $PGS_{BMI}$ increases as BMI itself increases, which may help explain significant interaction effects between $PGS_{BMI}$ and different covariates, it does not directly explain differences in $R^2$ when stratifying by different covariates – we describe several points that help explain this result and suggest they may actually be closely related. Essentially, as the magnitude of the slope of a regression line increases while the mean squared residual does not increase, model $R^2$ will increase – we demonstrate this using simulated data (Supplemental Figure 2). As the magnitude of the regression line's slope decreases, the regression line becomes a comparatively worse predictor compared to just using the mean, which

decreases $R^2$ despite the mean error being the same across models. To demonstrate this in real data, we compared simple univariable models of log(BMI) ~ $PGS_{BMI}$ (in units of log(BMI)) between the bottom and top age quintiles in the European ancestry UKBB (Supplemental Figure 3). As shown in previous sections, $R^2$ and $PGS_{BMI}$ beta are higher in younger individuals ($R^2$ = 0.088 versus $R^2$ = 0.066, and beta=1.12 and 0.87, respectively), which seem to be a direct consequence of one another, as the mean squared error in younger individuals is actually higher (0.027 versus 0.022, respectively). This description suggests that the use of $R^2$ as the sole performance metric for evaluation of PGS may not always be appropriate, despite its overwhelming usage. Furthermore, this explanation helps explain the seemingly paradoxical results of significant interaction terms yet small increases in overall model $R^2$, and comparably much larger differences in $R^2$ in the stratified analyses.

Effects of machine learning approaches on predictive performance

Given evidence of PGS-covariate dependence, we aimed to assess increases in $R^2$ when using machine learning models (neural networks), which can inherently model interactions and other nonlinearities, over linear models even with interaction terms. We first included age and sex as the only covariates (along with genotype PCs and $PGS_{BMI}$), as age and sex were present in all datasets and had significant and replicable evidence for PGS-dependence across our analyses. Three models were assessed – L1-regularized (i.e., LASSO) linear regression without any interaction terms, LASSO including a PGS-age and PGS-sex interaction term, and neural networks (without interaction terms). When comparing neural networks to LASSO with interaction terms, the relative 10-fold cross-validated $R^2$ increased up to 67% (mean 23%) across cohorts and ancestries (Figure 6, Supplemental Table 7). The inclusion of interaction terms increased cross-validated $R^2$ up to 12% (mean 3.9%) when comparing LASSO including interaction terms to LASSO with main effects only.

We then modeled all available covariates and their interactions with PGS for each cohort and did similar comparisons. Cross-validated $R^2$ increased by up to 17.6% (mean 9.5%) when using neural networks versus LASSO with interaction terms, and up to 7.0% (mean 2.0%) when comparing LASSO with interaction terms to LASSO with main effects only. Increases in model performance using neural networks were smaller in UKBB, perhaps due to the age range being smaller than other cohorts (all participants aged 39-73). This result suggests that additional variance explained through non-linear effects with age and sex are explained by other variables present in the remainder of the datasets. Our findings show machine learning methods can improve model $R^2$ that include $PGS_{BMI}$ as variables beyond including interaction terms in linear models, even when variable selection is performed using LASSO, demonstrating that model performance can be increased beyond modeling nonlinearities through linear interaction terms and a feature selection procedure.

Calculating PGS directly from GxAge GWAS effects

Previous studies (13) have created PGS using GWAS stratified by different personal-level covariates, but for practical purposes this leads to a large loss of power as the full size of the GWAS is not utilized for each strata and continuous variables are forced into bins. We developed a novel strategy where PGS are instead created from a full-size GWAS that includes SNP-covariate interaction terms (Methods). We focused on age interactions, given their large and replicable effects based on our results – similar to a previous study (13), we conducted these analyses in the European UKBB. We used a 60% random split of study individuals to conduct

three sets of PGS using GWAS of the following designs: main effects only, main effects also with a SNP-age interaction term, and main effects but stratified into quartiles by age. 20% of the remaining data were used for training and the final 20% were held-out as a test set. The best performing PGS created from SNP-age interaction terms ($PGS_{GxAge}$) increased test $R^2$ to 0.0771 (95% bootstrap CI: 0.0770-0.0772) from 0.0715 (95% bootstrap CI: 0.0714-0.0716), a 7.8% relative increase compared to the best performing main effect PGS (Figure 7, Supplemental Table 8 – age-stratified PGS had much lower performance than both other strategies (unsurprising given reasons previously mentioned). Including a $PGS_{GxAge}$-age interaction term only marginally increased $R^2$ (0.0001 increase), with similar increases for the other two sets of PGS, further demonstrating that post-hoc modeling of interactions cannot reconcile performance gained through directly incorporating interaction effects from the original GWAS. The strategy of creating PGS directly from full-sized SNP-covariate interactions is potentially quite useful as it increases PGS performance without the need for additional data – there are almost certainly a variety of points of improvement (described more in Discussion), but we consider their investigation outside the scope of this study.

## Discussion

We uncovered replicable effects of covariates across four large-scale cohorts of diverse ancestry, on both performance and effects of $PGS_{BMI}$. When stratifying by quintiles of different covariates, certain covariates had significant and replicable evidence for differences in $PGS_{BMI}$ $R^2$, with $R^2$ being nearly double between top and bottom performing quintiles for covariates with the largest differences. When testing PGS-covariate interaction effects, we also found covariates with significant interaction effects, where, for the largest effect covariates, each standard deviation change affected $PGS_{BMI}$ effect by nearly 20%. Across analyses, we found age and sex had the most replicable interaction effects, with levels of serum cholesterol, physical activity, and alcohol consumption having the largest effects across cohorts. Interaction effects and $R^2$ differences were strongly correlated, with main effects also being correlated with interaction effects and $R^2$ differences, suggesting that covariates with the largest interaction effects also contribute to the largest $R^2$ differences, with simple main effects also being predictive of expected differences in $R^2$ and interaction effects. Relatedly, we observed the effect of $PGS_{BMI}$ increases as BMI itself increases, and reason that differences in $R^2$ when stratifying by covariates are largely a consequence of difference in $PGS_{BMI}$ effects. Next, we employed machine learning methods for prediction of BMI with models that include $PGS_{BMI}$ and demonstrate that these methods outperform regularized linear regression models that include interaction effects. Finally, we employed a novel strategy that directly incorporates SNP interaction effects into PGS construction, and demonstrate that this strategy improves PGS performance when modeling SNP-age interactions compared to PGS created only from main effects.

These observations are relevant to current research and clinical use of PGS, as individuals above a percentile cutoff are designated high-risk (40), implying that individuals most at-risk for obesity have both disproportionately higher predicted BMI and increased BMI prediction performance compared to the general population. More broadly, these results may likely extend to single variant effects instead of those aggregated into a PGS, which may inform the cause of previous GxE discoveries – for instance, variants near *FTO* that interact with physical activity discovered through GWAS of BMI are robust and well-documented. However, individuals engaging in physical activity will generally have lower BMI than those that are sedentary, and these results suggest it may not be the difference in physical activity that's driving the

interaction, rather the difference in BMI itself. This concept may also apply to other traits – for instance, sex-specific analyses are commonly performed, and variants with differing effects between males and females GWAS may largely be explained by phenotypic differences, rather than any combination of biological or lifestyle differences.

Future work may include replicating these analyses across additional traits, and trying to understand why these differences occur, as well as further exploring machine learning and deep learning methods on other phenotypes to determine if this trend of inclusion of PGS along with covariate interaction effects outperforms linear models for risk prediction. Additionally, inclusion of a PGS for the covariate to better measure its environmental effect is potentially worth exploring further, and should improve in the future as PGS performance continues to increase. A slight limitation of this method in our study is that for the UKBB analyses the GWAS used for PGS construction were also from UKBB thus not out-of-sample, although many of the covariates only have GWAS available through UKBB individuals. Furthermore, a variety of improvements are likely possible when creating PGS directly from SNP-covariate interaction terms. First, we used the same SNPs that were selected by pruning and thresholding based on their main effect p-values, but selection of SNPs based on their interaction p-values should also be possible and would likely improve performance. Additionally, performance of pruning and thresholding-based strategies have largely been overtaken by methods that first adjust all SNP effects for LD and don't require exclusion of SNPs, and a method that could do a similar adjustment for interaction effects would likely outperform most current methods for traits with significant context-specific effects. Next, incorporating additional SNP-covariate interactions (e.g., SNP-sex) would also likely further improve prediction performance, although any SNP selection/adjustment procedures may be further complicated by additional interaction terms. Finally, if SNP effects do truly differ according to differences in phenotype, then SNP effects would differ depending on the alleles one has, implying epistatic interactions are occurring at these SNPs.

While difference in phenotype itself may be able to explain difference in genetic effects, it still may be that specific environmental or lifestyle characteristics are driving the differences. We propose several ideas about why BMI-associated covariates have larger interaction effects and impact on $R^2$ for $PGS_{BMI}$. First, age may be a proxy for accumulated gene-environment interactions as younger individuals have less exposure to environmental influences on weight compared to older individuals; therefore, one may expect that in younger individuals their phenotype could be better explained by genetics compared to environment. Second, PGS may more readily explain high phenotype values especially for positively skewed phenotypes, as large effect variants (e.g., associated with very high weight or height (25)) may be more responsible for extremely high phenotypic values. For example, the distribution of BMI is often positively skewed, and effects in trait-increasing alleles may have a larger potential to explain trait variation compared to trait-reducing variants. This explanation would likely be better suited to positively skewed traits and is not fully satisfactory as first log-transforming or rank-normal transforming the phenotype, as was done in this study, may invalidate this explanation.

PGS is a promising technique to stratify individuals for their risk of common, complex disease. To achieve more accurate predictions as well as promote equity, further research is required regarding PGS methods and assessments. This research provides firm evidence supporting the context-specific nature of PGS and the impact of nonlinear covariate effects for improving polygenic prediction of BMI, promoting equitable use of PGS across ancestries and cohorts.

## Methods

### Study datasets

Individual inclusion criteria and sample sizes per cohort are described below – additional information is available in Supplemental Table 1.

*UKBB*

Individual-level quality-control and filtering have been described elsewhere (26) for European ancestry individuals. Briefly, individuals were split by ancestry according to both genetically inferred ancestry and self-reported ethnicity (15). Individuals with low genotyping quality and sex mismatch were removed, only unrelated individuals (pi-hat < 0.250) were retained, and variants were filtered at INFO > 0.30 and minor allele frequency (MAF) > 0.01. For African ancestry, individuals were first selected based on self-reported ethnicity "Black or Black British", "Caribbean", "African", or "Any other Black background". Individuals that were low quality i.e., "Outliers for heterozygosity or missing rate", and that were Caucasian from "Genetic ethnic grouping" were removed. Of these individuals, those that were +/- 6 standard deviations from the mean of the first 5 genetic principal components provided by UKBB were excluded. Finally, only unrelated individuals were retained up to the second degree using plink2 (27) "--king-cutoff 0.125". After QC and consideration of phenotype, a total of 7,046 individuals in the UKBB AFR data who also had BMI available were used for downstream analyses. In total, 383,775 individuals were used for analysis ($N_{EUR}$=376,729, $N_{AFR}$=7,046).

*eMERGE*

Ancestry and relatedness inference have been described elsewhere (15). Individuals were split into European and African ancestry cohorts, and related individuals were removed (pi-hat > 0.250) such that all were unrelated. 35,064 individuals ($N_{EUR}$=31,961, $N_{AFR}$=3,103) were used for analysis.

*GERA*

Ancestry inference has been described elsewhere (28), and study individuals were divided into European and African ancestry cohorts. Related individuals were removed using plink2 "--king-cutoff 0.125". 57,838 individuals ($N_{EUR}$=56,049, $N_{AFR}$=1,789) were used for analysis.

*PMBB*

Ancestry inference and relatedness inference have been described elsewhere (29). Individuals were split into European and African ancestry cohorts, and related individuals were removed at pi-hat > 0.250. 36,046 individuals ($N_{EUR}$=26,372 , $N_{AFR}$=9,674) were used for analysis.

### Choice of covariates

A total of 62 covariates were included in the analyses, 25 of which were present (or similar proxies) in multiple datasets. These covariates were selected based on relevance to cardiometabolic health and obesity, and previous evidence of context-specific effects with BMI (5,6,8,30–32). For UKBB, phenotype values were used from the collection that was closest to recruitment, and for PMBB the median values were used – for GERA and eMERGE only one value was available. Additional details on covariate constructions, transformations, filtering, and cohort availability are in Supplemental Table 2.

461
462 PGS construction
463 PGS for BMI ($PGS_{BMI}$) were constructed using PRS-CSx (33), using GWAS summary statistics
464 for individuals of European (34), African (35), and East Asian (36) ancestry that were out-of-
465 sample of study participants. A set of 1.29 million HapMap3 SNPs provided by PRS-CSx was
466 used for PGS calculation, which are generally well-imputed and variable frequency across global
467 populations. Default settings for PRS-CSx (downloaded November 22, 2021) were used, which
468 have been demonstrated to perform well for highly polygenic traits such as BMI (list of
469 parameters in Supplemental Table 9). The final $PGS_{BMI}$ per ancestry and cohort was calculated
470 by regressing log(BMI) on the $PGS_{BMI}$ per ancestry without covariates – the combined, predicted
471 value was used as a single $PGS_{BMI}$ in downstream analyses.
472
473 For GERA, BMI was not transformed, as it was already binned into a categorical variable with 5
474 levels (18≤, 19-25, 26-29, 30-39, >40). Additionally, for GERA the uncombined ancestry-
475 specific $PGS_{BMI}$ were used in the final models, as it had higher $R^2$ than using the combined
476 $PGS_{BMI}$ (data not shown).
477
478 $PGS_{BMI}$ performance after covariate stratification
479 Analyses were performed separately for each cohort and ancestry. For each covariate, individuals
480 were binned by binary covariates or quintiles for continuous covariates. Incremental $PGS_{BMI}$ $R^2$
481 was calculated by taking the difference in $R^2$ between:
482
483 $\log(BMI) \sim PGS_{BMI} + Age + Sex + PCs_{1\text{-}5}$
484 $\log(BMI) \sim Age + Sex + PCs_{1\text{-}5}$
485
486 We performed 5,000 bootstrap replications to obtain a bootstrapped distribution of $R^2$. P-values
487 for differences in $R^2$ were calculated between groups by calculating the proportion of overlap
488 between two normal distributions of the $R^2$ value using the standard deviations of the bootstrap
489 distributions. Again for GERA, BMI was not transformed.
490
491 $PGS_{BMI}$ interaction modelling
492 Evidence for interaction with each covariate with the $PGS_{BMI}$ was evaluated using linear
493 regression. It has been reported that the inclusion of covariates that are genetically correlated
494 with the outcome can inflate test statistic estimates (24,37,38). To assuage these concerns, we
495 introduced a novel correction, where we first calculated a PGS for the covariate ($PGS_{Covariate}$) and
496 included it in the model, as well as an interaction term between $PGS_{BMI}$ and $PGS_{Covariate}$. The
497 $PGS_{Covariate}$ terms were calculated using the European ancestry Neale Lab summary statistics
498 (URLs) and PRS-CS (39). To standardize effect sizes across analyses, $PGS_{BMI}$ and Covariate
499 were first converted to mean zero and standard deviation of 1 (binary covariates were not
500 standardized). We demonstrate inclusion of $PGS_{Covariate}$ terms successfully reduced significance
501 of the $PGS_{BMI}$*Covariate term (Supplemental Figure 1). The final model used to evaluate
502 $PGS_{BMI}$ and Covariate interactions was:
503
504 $\log(BMI) \sim PGS_{BMI}*Covariate + PGS_{BMI} + Covariate + PGS_{Covariate} + PGS_{BMI}*PGS_{Covariate} + Age + Sex + PCs_{1\text{-}5}$
505
506 We report the effect estimates of the $PGS_{BMI}$*Covariate term, and differences in model $R^2$ with
507 and without the $PGS_{BMI}$*Covariate term. Again for GERA, BMI was not transformed.

Correlation between $R^2$ differences, interaction effects, and main effects

We estimated main effects of each covariate on BMI with the following model:

$$\log(BMI) \sim Covariate + Age + Sex + PCs_{1\text{-}5}$$

Note that we ran new models with main effects only, instead of using the main effect from the interaction models (as the main effects in the interaction models depend on the interaction terms, and main effects used to create interaction terms are sensitive to centering of variables despite the scale invariance of linear regression itself (40)). We then estimated the correlation between main effects, interaction effects, and maximum $R^2$ differences across all cohorts and ancestries weighting by sample size, analyzing quantitative and binary variables separately.

Quantile regression to measure PGS effect across percentiles of BMI

The effect of $PGS_{BMI}$ on BMI at different deciles of BMI was assessed using quantile regression. Tau – the parameter that sets which percentile to be predicted – was set to .10, .20, …, .90. Models included age, sex, and the top 5 genetic PCs as covariates. Analyses were stratified by ancestry and cohort, and BMI was first log transformed. GERA was excluded from these analyses, as a portion of the models failed to run (as BMI values from GERA were already binned, some deciles all had the same BMI value – additionally, difference in effects between bins would be harder to evaluate as BMI within each decile would be more homogeneous).

Machine learning models

UKBB EUR and GERA EUR models were restricted to 30,000 random individuals, for computational reasons – BMI distributions did not differ from the full-sized datasets (Kolmogorov-Smirnov p-value of 0.29 and 0.57, respectively). $PGS_{BMI}$ and top five genetic principal components were included as features in all models. Two sets of models were evaluated for each cohort and ancestry: including age and sex as features, and including all available covariates in each cohort as features. Interactions terms between PGS and each covariate were included for models using interaction terms. 'Ever Smoker' status was used in favor of 'Never' vs 'Current smoking' status (if present), as individuals with 'Never' vs 'Current' status are a subset of those with 'Ever Smoker' status. UKBB AFR with all covariates was excluded due to small sample size (N=53).

Neural networks were used as the model of choice, given their inherent ability model interactions and other nonlinear dependencies. Prior to modeling, all features were scaled to be between 0 and 1. We used average 10-fold cross validation $R^2$ to evaluate model performance. Separate models were trained using untransformed and log(BMI). L1-regularized linear regression was used with 18 values of lambda ($1.0, 5.0\times10^{-1}, 2.0\times10^{-1}, 1.0\times10^{-1}, 5.0\times10^{-2}, 2.0\times10^{-2}, \ldots, 1.0\times10^{-5}, 5.0\times10^{-6}, 2.0\times10^{-6}$). Models were trained without inclusion of interaction terms (which neural networks can implicitly model), using 1,000 iterations of random search with the following hyperparameter ranges: size of hidden layers [10, 200], learning rate [0.01, 0.0001], type of learning rate [constant, inverse scaling], power t [0.4, 0.6], momentum [0.80, 1.0], batch size [32, 256], and number of hidden layers [1, 2].

GxAge $PGS_{BMI}$ creation and assessment

Analyses were conducted in the European UKBB (N=376,629), as was done in a study on a similar topic (13). Three sets of analyses were performed, using GWAS conducted in a 60%

random split of individuals using the following models (BMI was rank-normal transformed before GWAS):

1) $BMI \sim SNP + Age + Sex + PCs_{1-5}$

2) $BMI \sim SNP + Age*SNP + Age + Sex + PCs_{1-5}$

3) Using the model in 1) but stratified into quartiles by age. BMI was rank-normal transformed within each quartile.

Using each set of GWAS, PGS was first calculated in a 20% randomly selected training set of the dataset using pruning and thresholding using 10 p-value thresholds (0.50, 0.10, …, $5.0x10^{-5}$, $1.0x10^{-5}$) and remaining settings as default in Plink 1.9. For 2), GxAge $PGS_{BMI}$ were calculated using SNPs clumped by their main effect p-values from 1), and additionally incorporating the GxAge interaction effects per SNP. In other words, instead of typical PGS construction as:

$$PGS_i = \beta_1 k_1 + \beta_2 k_2 + \ldots + \beta_n k_n$$

For an individual *i*'s PGS calculation, with main SNP effect β, and *n* SNPs, PGS incorporating GxAge effects ($PGS_{GxAge}$) was calculated as:

$$PGS_{GxAge,i} = \beta_1 k_1 + \beta_{GxAge,1} k_1 Age_i + \beta_2 k_2 + \beta_{GxAge,2} k_2 Age_i \ldots + \beta_n k_n + \beta_{GxAge,n} k_n Age_i$$

Where $\beta_{GxAge}$ is the GxAge effect for each SNP *n* and $Age_i$ is the age for individual *i*.

For each of the three analyses, the parameters and models resulting in the best performing PGS (highest incremental $R^2$, using same main effect covariates as in the three GWAS) from the training set were evaluated in the remaining 20% of the study individuals. For 3), models were first trained within each quartile separately. To maintain sense of scale across quartiles (after rank-normal transformation), $R^2$ between all predicted values and true values was calculated together. For $R^2$ confidence intervals, the training set was bootstrapped and evaluated on the test set 5,000 times.

**URLs**

Neale Lab UKBB summary statistics: http://www.nealelab.is/uk-biobank

**Data Availability statement**

UK Biobank data was accessed under project #32133. eMERGE data is available at dbGaP in phs001584.v2.p2. GERA data is available at dbGaP in phs000674.v3.p3. Summary statistics for PMBB are available from authors upon request.

**Regeneron Genetics Center Banner Author List and Contribution Statements**

RGC Management and Leadership Team

Goncalo Abecasis, PhD , Aris Baras, M.D. , Michael Cantor, M.D. , Giovanni Coppola, M.D. , Andrew Deubler , Aris Economides, Ph.D. , Luca A. Lotta, M.D., Ph.D. , John D. Overton, Ph.D. , Jeffrey G. Reid, Ph.D. , Katherine Siminovitch, M.D. , Alan Shuldiner, M.D.

Sequencing and Lab Operations
Christina Beechert , Caitlin Forsythe, M.S. , Erin D. Fuller , Zhenhua Gu, M.S. , Michael Lattari , Alexander Lopez, M.S. , John D. Overton, Ph.D. , Maria Sotiropoulos Padilla, M.S. , Manasi Pradhan, M.S. , Kia Manoochehri, B.S. , Thomas D. Schleicher, M.S. , Louis Widom , Sarah E. Wolf, M.S. , Ricardo H. Ulloa, B.S.

Clinical Informatics
Amelia Averitt, Ph.D. , Nilanjana Banerjee, Ph.D. , Michael Cantor, M.D. , Dadong Li, Ph.D. , Sameer Malhotra, M.D. , Deepika Sharma, MHI , Jeffrey Staples , Ph.D.

Genome Informatics
Xiaodong Bai, Ph.D. , Suganthi Balasubramanian, Ph.D. , Suying Bao, Ph.D. , Boris Boutkov, Ph.D. , Siying Chen, Ph.D. , Gisu Eom, B.S. , Lukas Habegger, Ph.D. , Alicia Hawes, B.S. , Shareef Khalid , Olga Krasheninina, M.S. , Rouel Lanche, B.S. , Adam J. Mansfield, B.A. , Evan K. Maxwell, Ph.D. , George Mitra, B.A. , Mona Nafde, M.S. , Sean O'Keeffe, Ph.D. , Max Orelus, B.B.A. , Razvan Panea, Ph.D. , Tommy Polanco, B.A. , Ayesha Rasool, M.S. , Jeffrey G. Reid, Ph.D. , William Salerno, Ph.D. , Jeffrey C. Staples, Ph.D. , Kathie Sun, Ph.D.

Analytical Genomics and Data Science
Goncalo Abecasis, D.Phil. , Joshua Backman, Ph.D. , Amy Damask, Ph.D. , Lee Dobbyn, Ph.D. , Manuel Allen Revez Ferreira, Ph.D. , Arkopravo Ghosh, M.S. , Christopher Gillies, Ph.D. , Lauren Gurski, B.S. , Eric Jorgenson, Ph.D. , Hyun Min Kang, Ph.D. , Michael Kessler, Ph.D. , Jack Kosmicki, Ph.D. , Alexander Li , Ph.D. , Nan Lin, Ph.D. , Daren Liu, M.S. , Adam Locke, Ph.D. , Jonathan Marchini, Ph.D. , Anthony Marcketta, M.S. , Joelle Mbatchou, Ph.D. , Arden Moscati, Ph.D. , Charles Paulding, Ph.D. , Carlo Sidore, Ph.D. , Eli Stahl, Ph.D. , Kyoko Watanabe, Ph.D. , Bin Ye, Ph.D. , Blair Zhang, Ph.D. , Andrey Ziyatdinov, Ph.D.

Therapeutic Area Genetics
Ariane Ayer, B.S. , Aysegul Guvenek, Ph.D. , George Hindy, Ph.D. , Giovanni Coppola, M.D. , Jan Freudenberg, M.D. , Jonas Bovijn M.D. , Katherine Siminovitch, M.D. , Kavita Praveen, Ph.D. , Luca A. Lotta, M.D. , Manav Kapoor, Ph.D. , Mary Haas, Ph.D. , Moeen Riaz , Ph.D. , Niek Verweij, Ph.D. , Olukayode Sosina, Ph.D. , Parsa Akbari, Ph.D. , Priyanka Nakka, Ph.D. , Sahar Gelfman, Ph.D. , Sujit Gokhale, B.E. , Tanima De, Ph.D. , Veera Rajagopal, Ph.D. , Alan Shuldiner, M.D. , Bin Ye, Ph.D. , Gannie Tzoneva, Ph.D. , Juan Rodriguez-Flores, Ph.D.

Research Program Management & Strategic Initiatives
Esteban Chen, M.S. , Marcus B. Jones, Ph.D. , Michelle G. LeBlanc, Ph.D. , Jason Mighty, Ph.D., Lyndon J. Mitnaul, Ph.D. , Nirupama Nishtala, Ph.D. , Nadia Rana, Ph.D. , Jaimee Hernandez

**Acknowledgements**
We acknowledge David Crosslin for helping clean the eMERGE data.

**Funding**

M.D.R., S.D., and D.H. are funded by AI077505 and HL169458. W.K.C is funded by grant NIDDK DK52431. J.F.P. is funded by grant U01 HG011166. C.W. is funded by U01 HG008680. This phase of the eMERGE Network was initiated and funded by the NHGRI through the following grants: U01HG008657 (Group Health Cooperative/University of Washington); U01HG008685 (Brigham and Women's Hospital); U01HG008672 (Vanderbilt University Medical Center); U01HG008666 (Cincinnati Children's Hospital Medical Center); U01HG006379 (Mayo Clinic); U01HG008679 (Geisinger Clinic); U01HG008680 (Columbia University Health Sciences); U01HG008684 (Children's Hospital of Philadelphia); U01HG008673 (Northwestern University); U01HG008701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG008676 (Partners Healthcare/Broad Institute); U01HG008664 (Baylor College of Medicine); and U54MD007593 (Meharry Medical College).

## References

1. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019 Apr;51(4):584–91.

2. Wang Y, Guo J, Ni G, Yang J, Visscher PM, Yengo L. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat Commun. 2020 Jul 31;11(1):3865.

3. Galinsky KJ, Reshef YA, Finucane HK, Loh PR, Zaitlen N, Patterson NJ, et al. Estimating cross-population genetic correlations of causal effect sizes. Genet Epidemiol. 2019 Mar;43(2):180–8.

4. Shi H, Gazal S, Kanai M, Koch EM, Schoech AP, Siewert KM, et al. Population-specific causal disease effect sizes in functionally important regions impacted by selection. Nat Commun. 2021 Feb 17;12(1):1098.

5. Rask-Andersen M, Karlsson T, Ek WE, Johansson Å. Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. PLoS Genet. 2017 Sep;13(9):e1006977.

6. Robinson MR, English G, Moser G, Lloyd-Jones LR, Triplett MA, Zhu Z, et al. Genotype-covariate interaction effects and the heritability of adult body mass index. Nat Genet. 2017 Aug;49(8):1174–81.

7. Sulc J, Mounier N, Günther F, Winkler T, Wood AR, Frayling TM, et al. Quantification of the overall contribution of gene-environment interaction for obesity-related traits. Nat Commun. 2020 Mar 13;11(1):1385.

8. Justice AE, Winkler TW, Feitosa MF, Graff M, Fisher VA, Young K, et al. Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. Nat Commun. 2017 Apr 26;8:14977.

9. Helgeland Ø, Vaudel M, Juliusson PB, Lingaas Holmen O, Juodakis J, Bacelis J, et al. Genome-wide association study reveals dynamic role of genetic variation in infant and early childhood growth. Nat Commun. 2019 Oct 1;10(1):4448.

10. Vogelezang S, Bradfield JP, Ahluwalia TS, Curtin JA, Lakka TA, Grarup N, et al. Novel loci for childhood body mass index and shared heritability with adult cardiometabolic traits. PLoS Genet. 2020 Oct;16(10):e1008718.

11. Couto Alves A, De Silva NMG, Karhunen V, Sovio U, Das S, Taal HR, et al. GWAS on longitudinal growth traits reveals different genetic factors influencing infant, child, and adult BMI. Sci Adv. 2019 Sep;5(9):eaaw3095.

12. Choh AC, Lee M, Kent JW, Diego VP, Johnson W, Curran JE, et al. Gene-by-age effects on BMI from birth to adulthood: the Fels Longitudinal Study. Obes Silver Spring Md. 2014 Mar;22(3):875–81.

13. Mostafavi H, Harpak A, Agarwal I, Conley D, Pritchard JK, Przeworski M. Variable prediction accuracy of polygenic scores within an ancestry group. eLife. 2020 Jan 30;9:e48376.

14. Elks CE, den Hoed M, Zhao JH, Sharp SJ, Wareham NJ, Loos RJF, et al. Variability in the heritability of body mass index: a systematic review and meta-regression. Front Endocrinol. 2012;3:29.

15. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203–9.

16. Stanaway IB, Hall TO, Rosenthal EA, Palmer M, Naranbhai V, Knevel R, et al. The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. Genet Epidemiol. 2019 Feb;43(1):63–81.

17. Kilpeläinen TO, Qi L, Brage S, Sharp SJ, Sonestedt E, Demerath E, et al. Physical activity attenuates the influence of FTO variants on obesity risk: a meta-analysis of 218,166 adults and 19,268 children. PLoS Med. 2011 Nov;8(11):e1001116.

18. Rampersaud E, Mitchell BD, Pollin TI, Fu M, Shen H, O'Connell JR, et al. Physical activity and the association of common FTO gene variants with body mass index and obesity. Arch Intern Med. 2008 Sep 8;168(16):1791–7.

19. Amin V, Dunn P, Spector T. Does education attenuate the genetic risk of obesity? Evidence from U.K. Twins. Econ Hum Biol. 2018 Sep;31:200–8.

20. Li Y, Cai T, Wang H, Guo G. Achieved educational attainment, inherited genetic endowment for education, and obesity. Biodemography Soc Biol. 2021 Jun;66(2):132–44.

21. Frank M, Dragano N, Arendt M, Forstner AJ, Nöthen MM, Moebus S, et al. A genetic sum score of risk alleles associated with body mass index interacts with socioeconomic position in the Heinz Nixdorf Recall Study. PloS One. 2019;14(8):e0221252.

22. Ge T, Chen CY, Neale BM, Sabuncu MR, Smoller JW. Phenome-wide heritability analysis of the UK Biobank. PLoS Genet. 2017 Apr;13(4):e1006711.

23. Min J, Chiu DT, Wang Y. Variation in the heritability of body mass index based on diverse twin studies: a systematic review. Obes Rev Off J Int Assoc Study Obes. 2013 Nov;14(11):871–82.

24. Aschard H, Vilhjálmsson BJ, Joshi AD, Price AL, Kraft P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. Am J Hum Genet. 2015 Feb 5;96(2):329–39.

25. Robinson PN, Arteaga-Solis E, Baldock C, Collod-Béroud G, Booms P, De Paepe A, et al. The molecular genetics of Marfan syndrome and related disorders. J Med Genet. 2006 Oct;43(10):769–87.

26. Zhang X, Lucas AM, Veturi Y, Drivas TG, Bone WP, Verma A, et al. Large-scale genomic analyses reveal insights into pleiotropy across circulatory system diseases and nervous system disorders. Nat Commun. 2022 Jun 14;13(1):3428.

27. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7.

28. Banda Y, Kvale MN, Hoffmann TJ, Hesselson SE, Ranatunga D, Tang H, et al. Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics. 2015 Aug;200(4):1285–95.

29. Penn Medicine BioBank [Internet]. Available from: https://pmbb.med.upenn.edu/

30. Tyrrell J, Wood AR, Ames RM, Yaghootkar H, Beaumont RN, Jones SE, et al. Gene-obesogenic environment interactions in the UK Biobank study. Int J Epidemiol. 2017 Apr 1;46(2):559–75.

31. Young AI, Wauthier F, Donnelly P. Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index. Nat Commun. 2016 Sep 6;7:12724.

32. Winkler TW, Justice AE, Graff M, Barata L, Feitosa MF, Chu S, et al. The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. PLoS Genet. 2015 Oct;11(10):e1005378.

33. Ruan Y, Lin YF, Feng YCA, Chen CY, Lam M, Guo Z, et al. Improving polygenic prediction in ancestrally diverse populations. Nat Genet. 2022 May;54(5):573–80.

34. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015 Feb 12;518(7538):197–206.

35. Ng MCY, Graff M, Lu Y, Justice AE, Mudgal P, Liu CT, et al. Discovery and fine-mapping of adiposity loci using high density imputation of genome-wide association studies in individuals of African ancestry: African Ancestry Anthropometry Genetics Consortium. PLoS Genet. 2017 Apr;13(4):e1006719.

36. Sakaue S, Kanai M, Tanigawa Y, Karjalainen J, Kurki M, Koshiba S, et al. A cross-population atlas of genetic associations for 220 human phenotypes. Nat Genet. 2021 Oct;53(10):1415–24.

37. Kerin M, Marchini J. Inferring Gene-by-Environment Interactions with a Bayesian Whole-Genome Regression Model. Am J Hum Genet. 2020 Oct 1;107(4):698–713.

38. Vanderweele TJ, Ko YA, Mukherjee B. Environmental confounding in gene-environment interaction studies. Am J Epidemiol. 2013 Jul 1;178(1):144–52.

39. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019 Apr 16;10(1):1776.

40. Afshartous D, Preston RA. Key Results of Interaction Models with Centering. J Stat Educ. 2011 Nov;19(3):1.

**Figures**

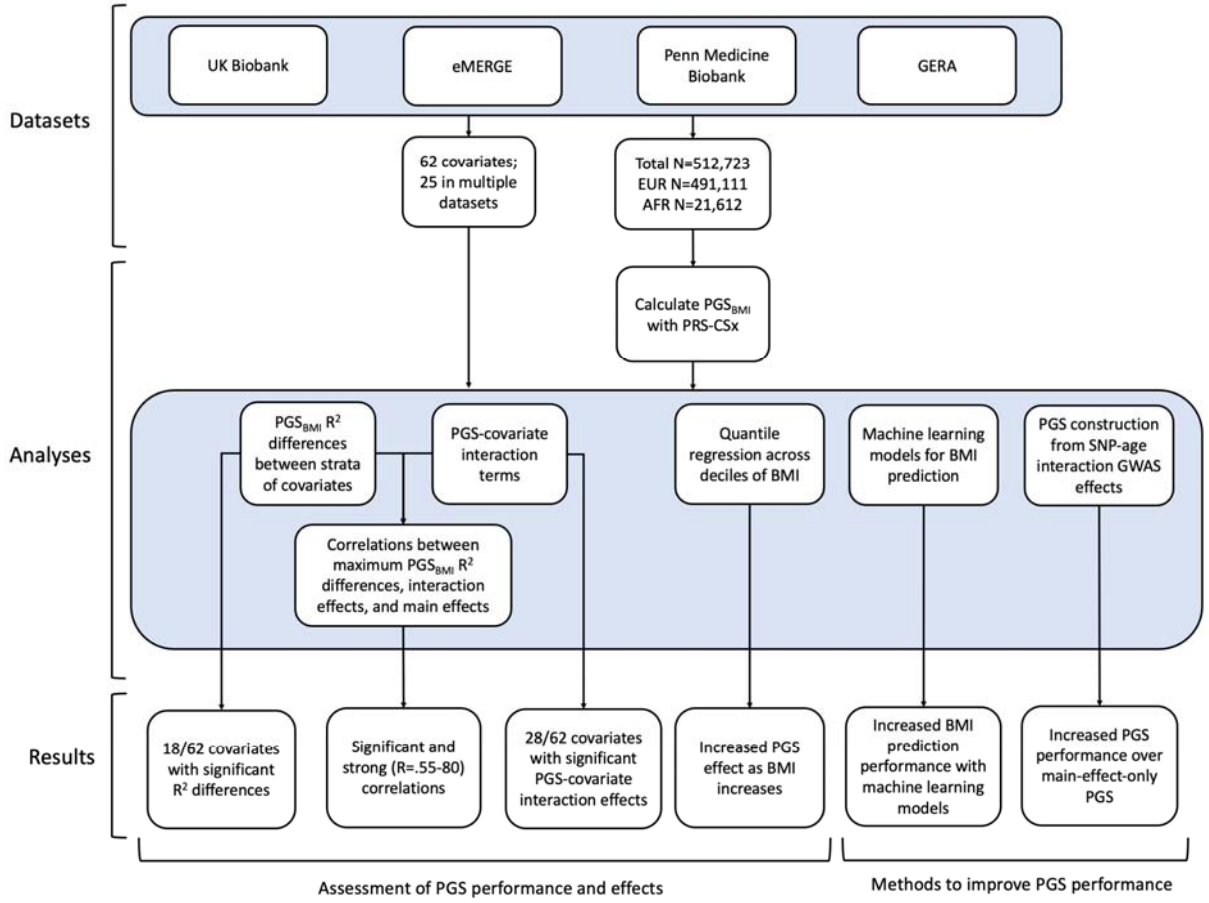
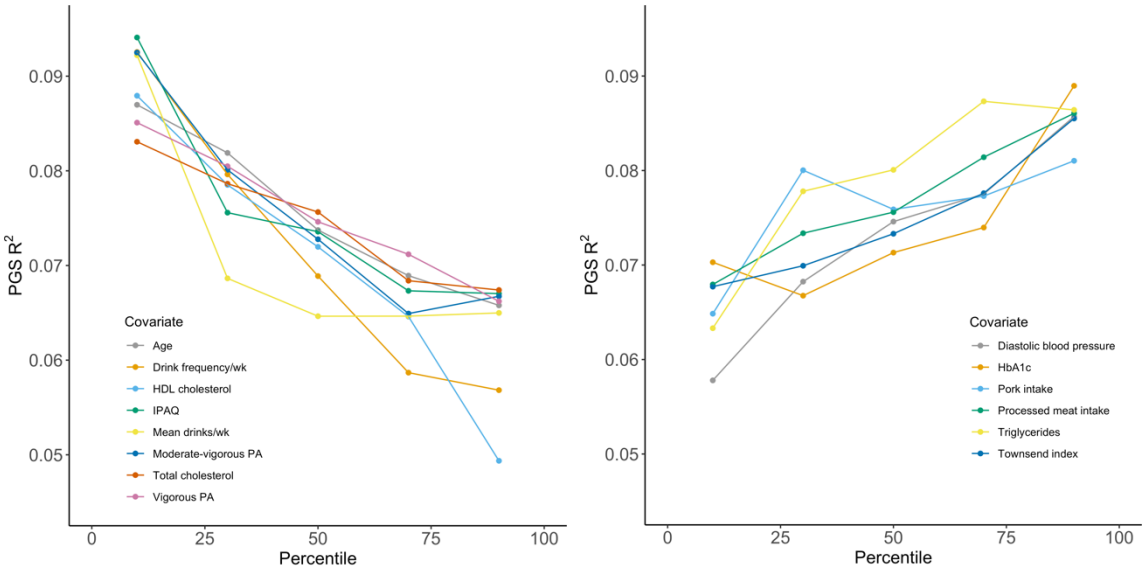



Figure 1. A flowchart of the project.

a)




b)

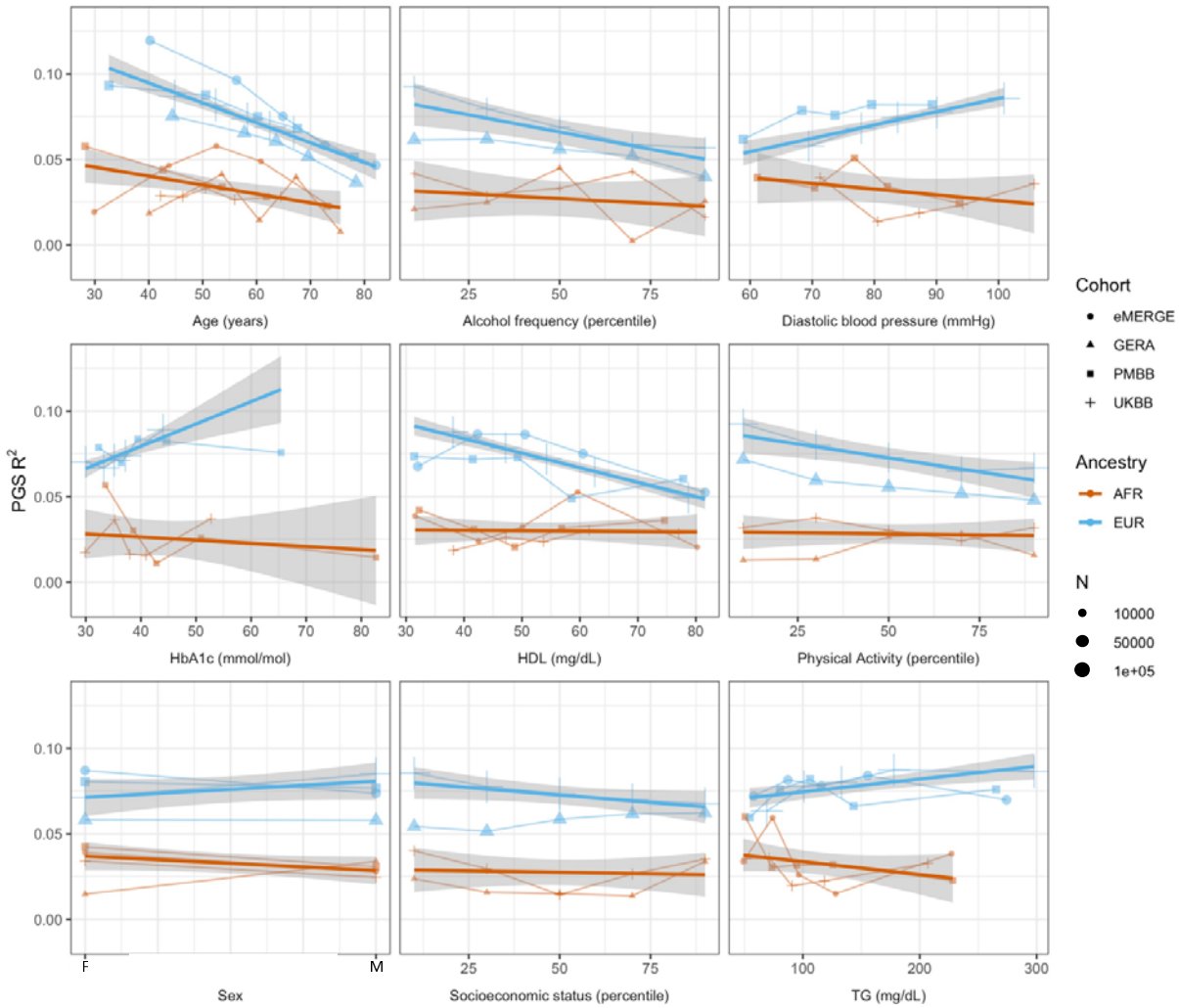


Figure 2. PGS $R^2$ stratified by quintiles for quantitative variables and by binary variables. a) Continuous covariates with significant ($p < 8.1x10^{-4}$) $R^2$ differences across quintiles in UKBB EUR. Pork and processed meat consumption per week were excluded from this plot in favor of pork and processed meat intake. b) Covariates with significant differences that were available in multiple cohorts. When traits had the same or directly comparable units between cohorts we show the actual trait values (and show percentiles for physical activity, alcohol intake frequency, and socioeconomic status, which had slightly differing phenotype definitions across cohorts) plotted on x-axis. Townsend index and income were used as variables for socioeconomic status UKBB and GERA, respectively. Note that the sign for Townsend index was reversed, since increasing Townsend index is lower socioeconomic status, while increasing income is higher socioeconomic status. Abbreviations: physical activity (PA), International Physical Activity Questionnaire (IPAQ).

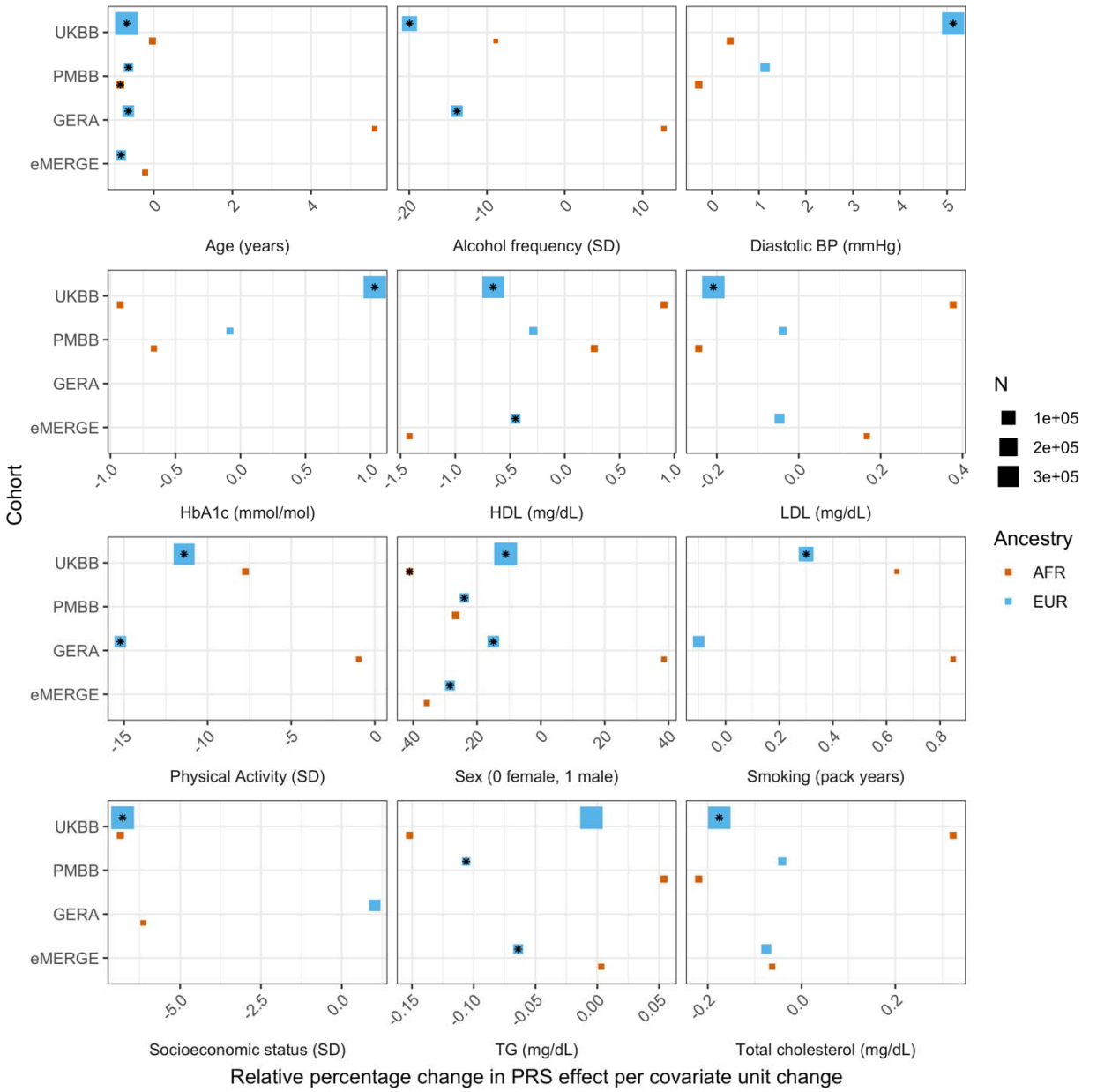



Figure 3. Relative percentage changes in PGS effect per unit change in covariate, for covariates that significantly changed PGS effect (i.e., significant interaction beta at Bonferroni $p < 8.1 \times 10^{-4}$ – denoted by asterisks) and were present in multiple cohorts and ancestries. Same covariate groupings and transformations were performed as in Figure 1. Similarly, actual values were used when variables had comparable units across cohorts, and standard deviations (SD) used otherwise.

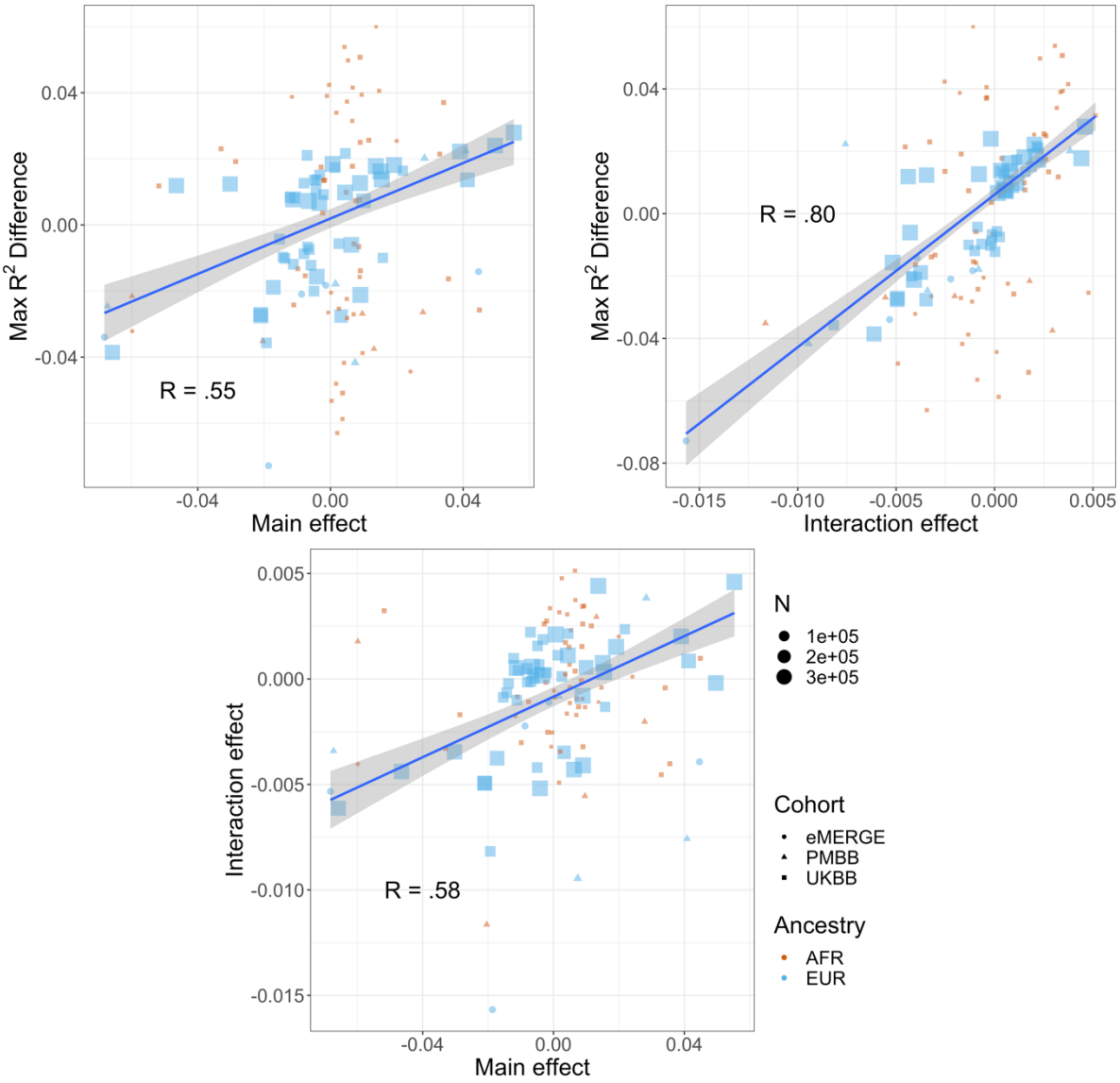



Figure 4. Relationships (Pearson correlations weighted by sample size) between maximum $R^2$ differences across strata, main effects of covariate on log(BMI), and PGS-covariate interaction effects on log(BMI). Main effect units are in standard deviations, interaction effect units are in PGS standard deviations multiplied by covariate standard deviations. Only continuous variables are plotted and modeled. GERA was excluded due to slightly different phenotype definitions.

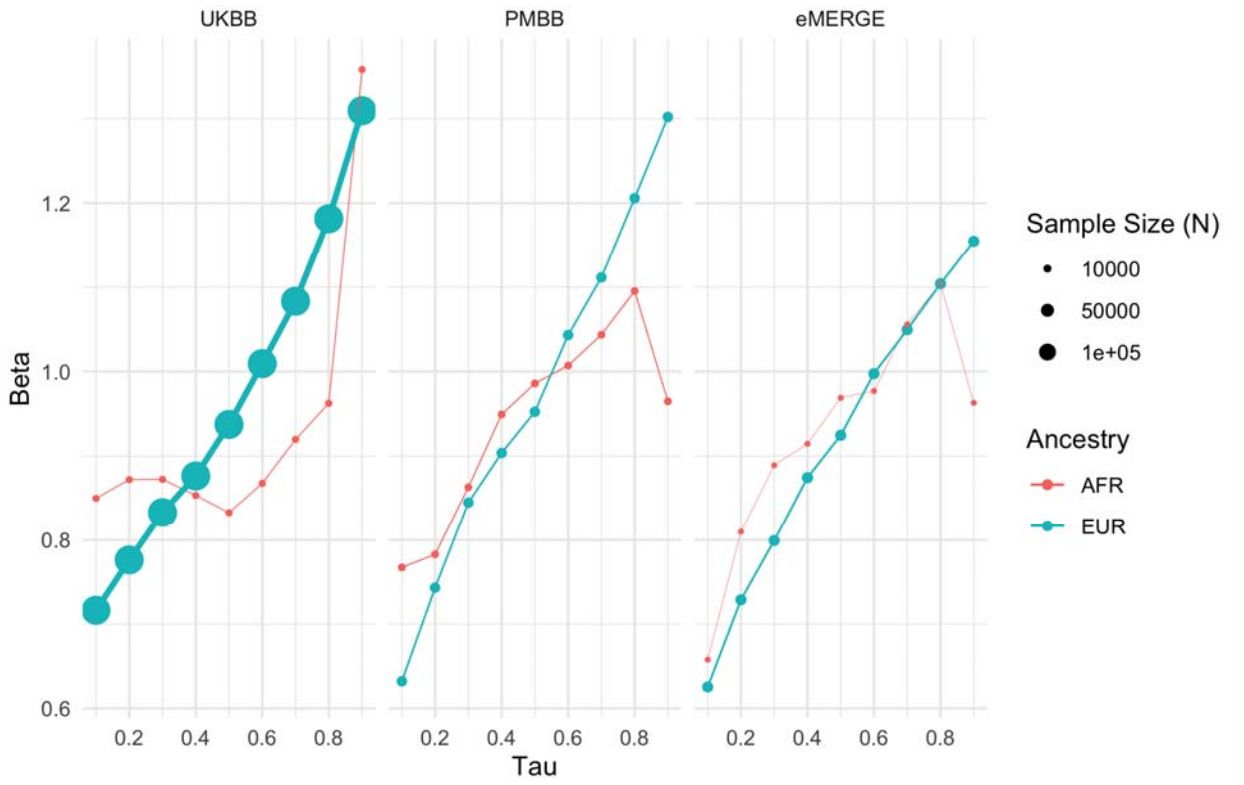



Figure 5. Quantile regression effects of $PGS_{BMI}$ (in units of log(BMI)) on log(BMI) at each decile of BMI in each cohort and ancestry. The effect of $PGS_{BMI}$ increases as BMI itself increases, suggesting that no individual covariate-PGS interaction is responsible for the nonlinear effect of $PGS_{BMI}$.

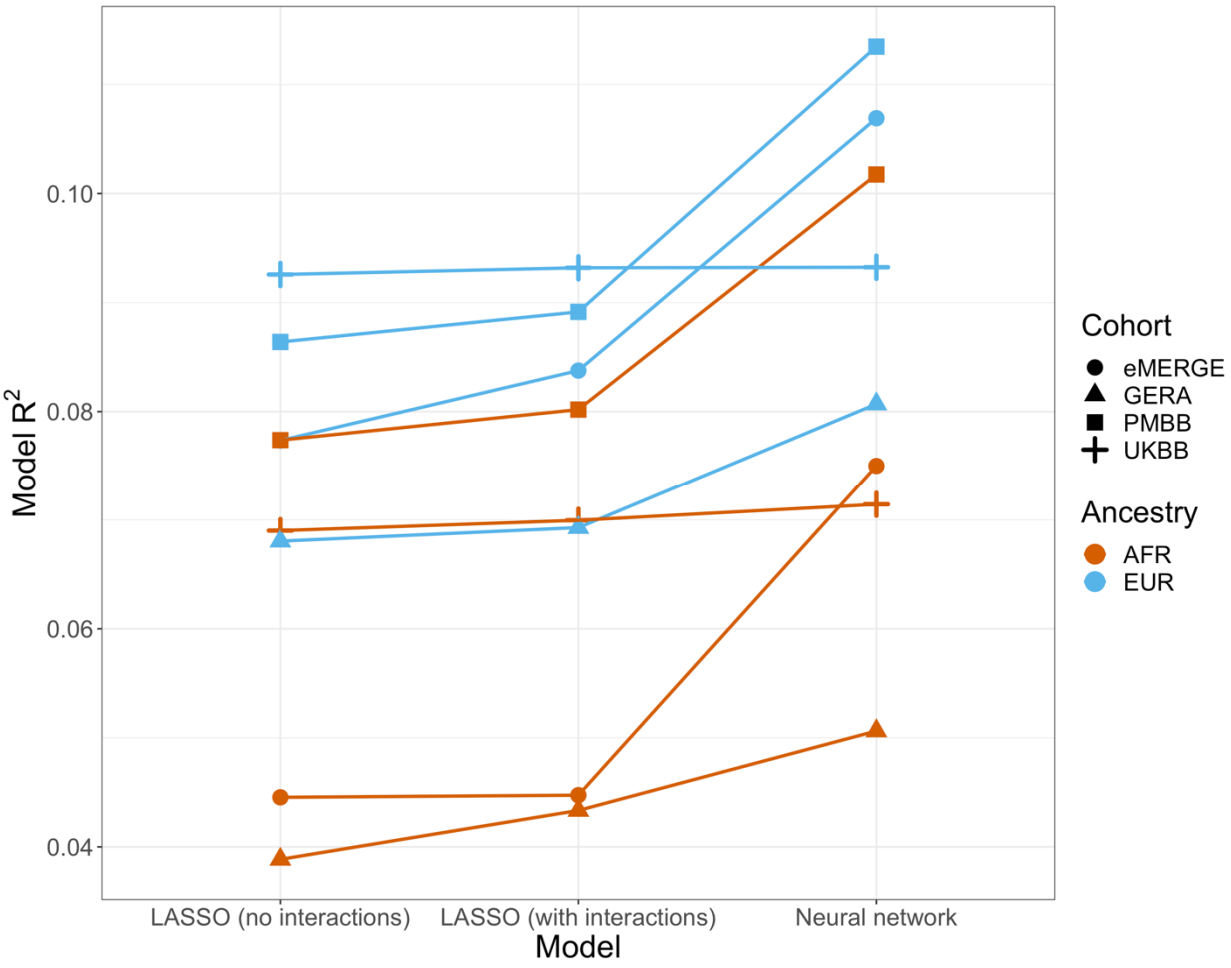



Figure 6. Model $R^2$ from different machine learning models across cohorts and ancestries using age and gender as covariates (along with $PGS_{BMI}$ and PCs 1-5). Across all cohorts and ancestries, LASSO with PGS-age and PGS-gender interaction terms had better average 10-fold cross-validation $R^2$ than LASSO without interaction terms, while neural networks outperformed LASSO models.

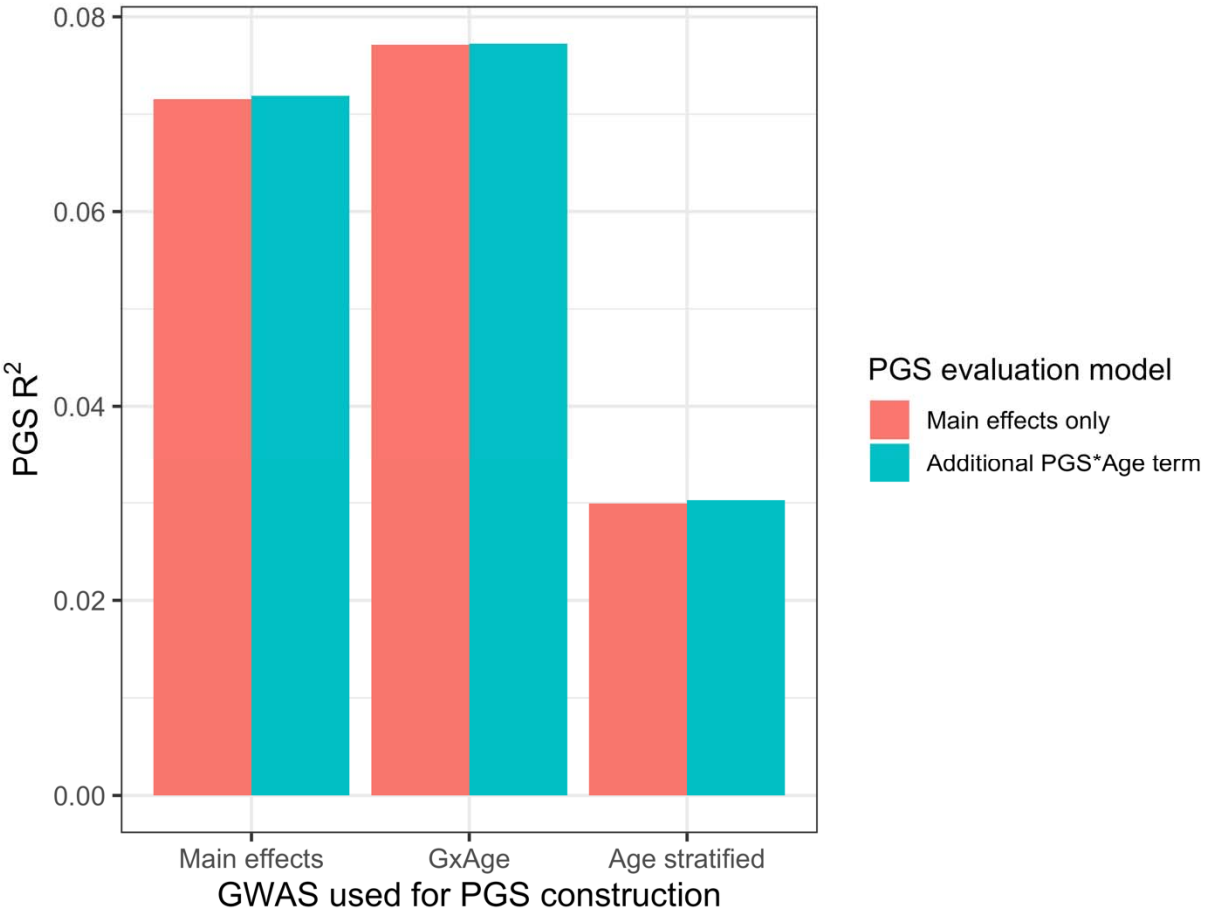
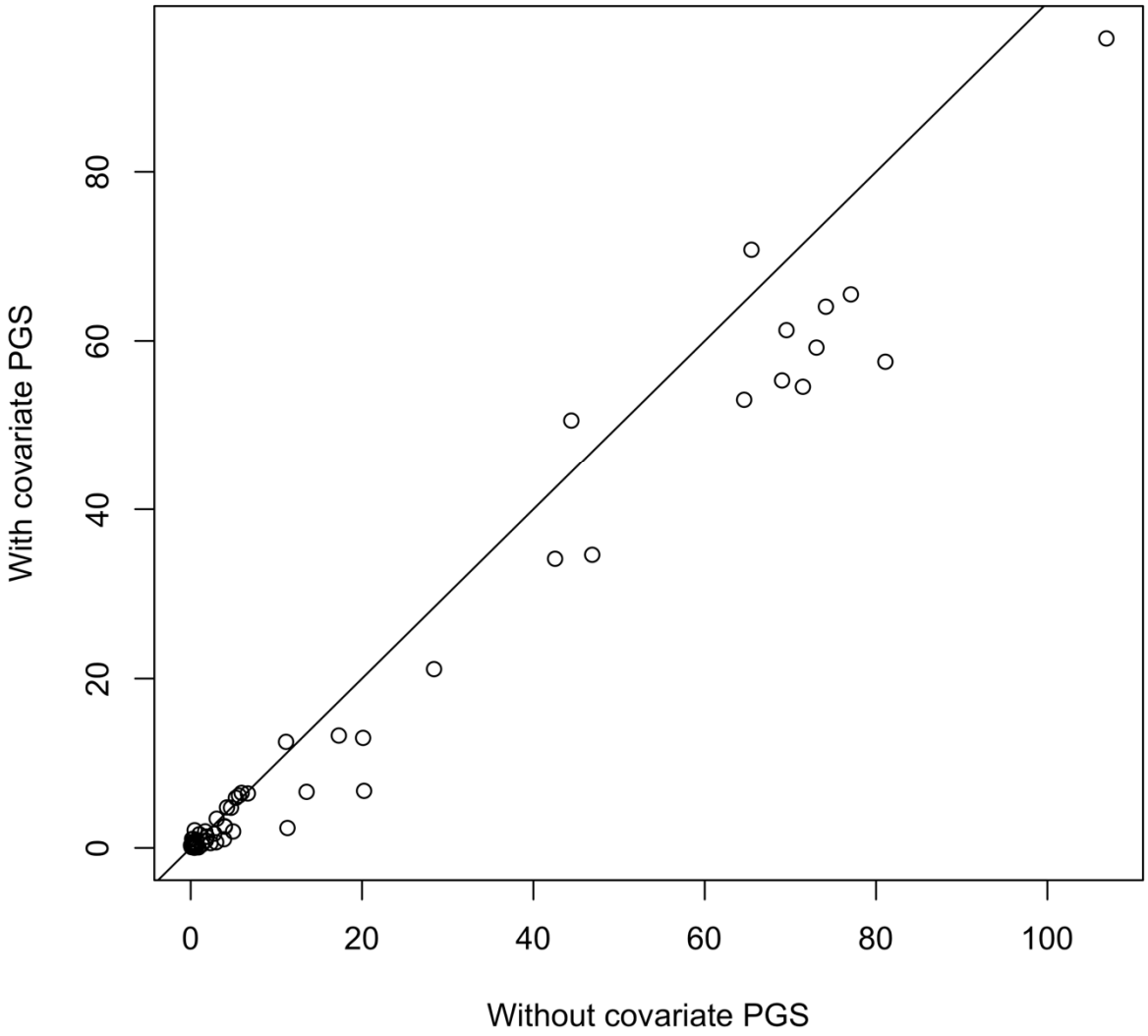
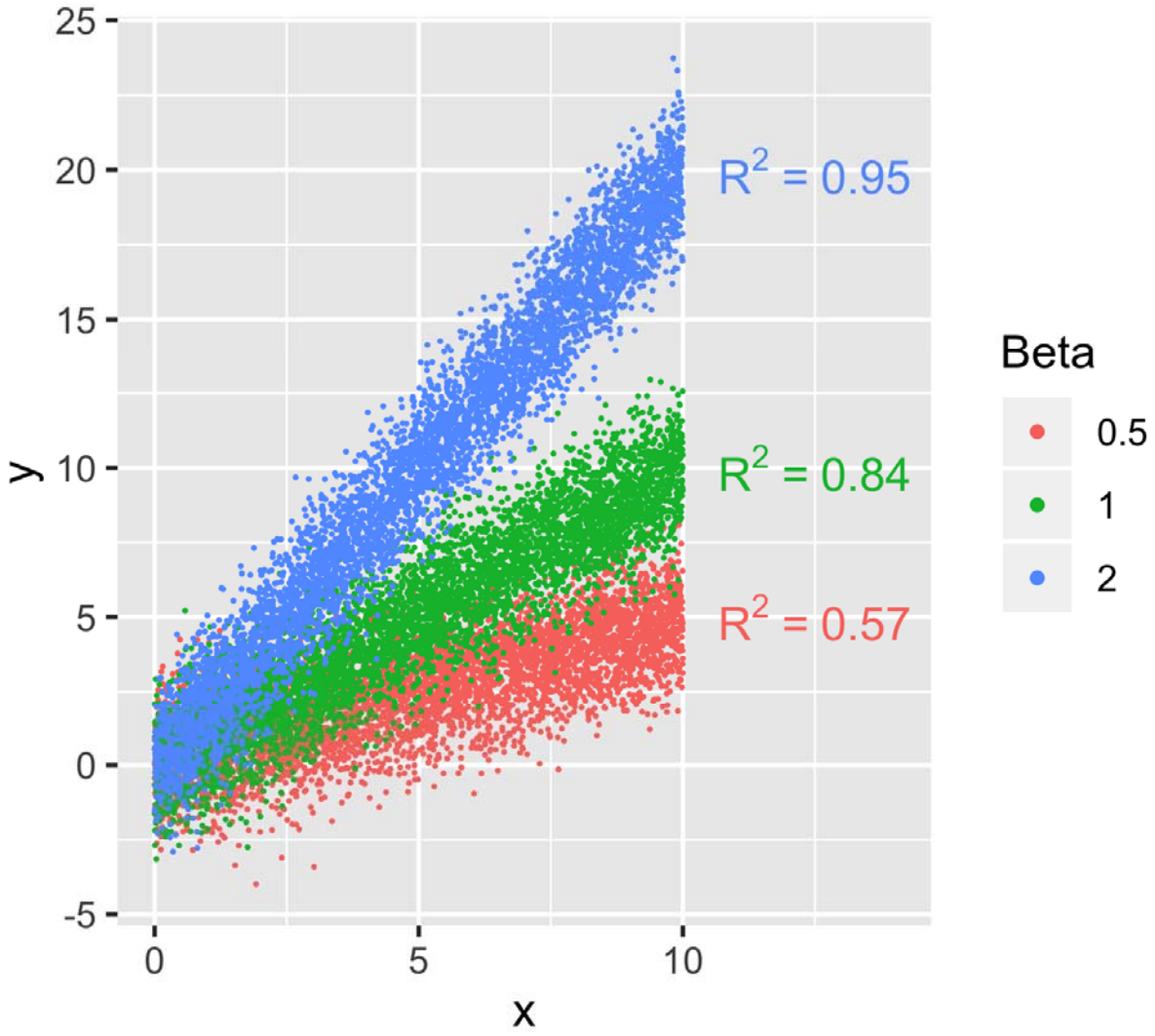



Figure 7. PGS $R^2$ based on three sets of GWAS setups. "Main effects" were from a typical main effect GWAS, "GxAge" effects were from a GWAS with a SNP-age interaction term, and "Age stratified" GWAS had main effects only but were conducted in four age quartiles. PGS $R^2$ was evaluated using two models: one with main effects only, and one with an additional PGS*Age interaction term.

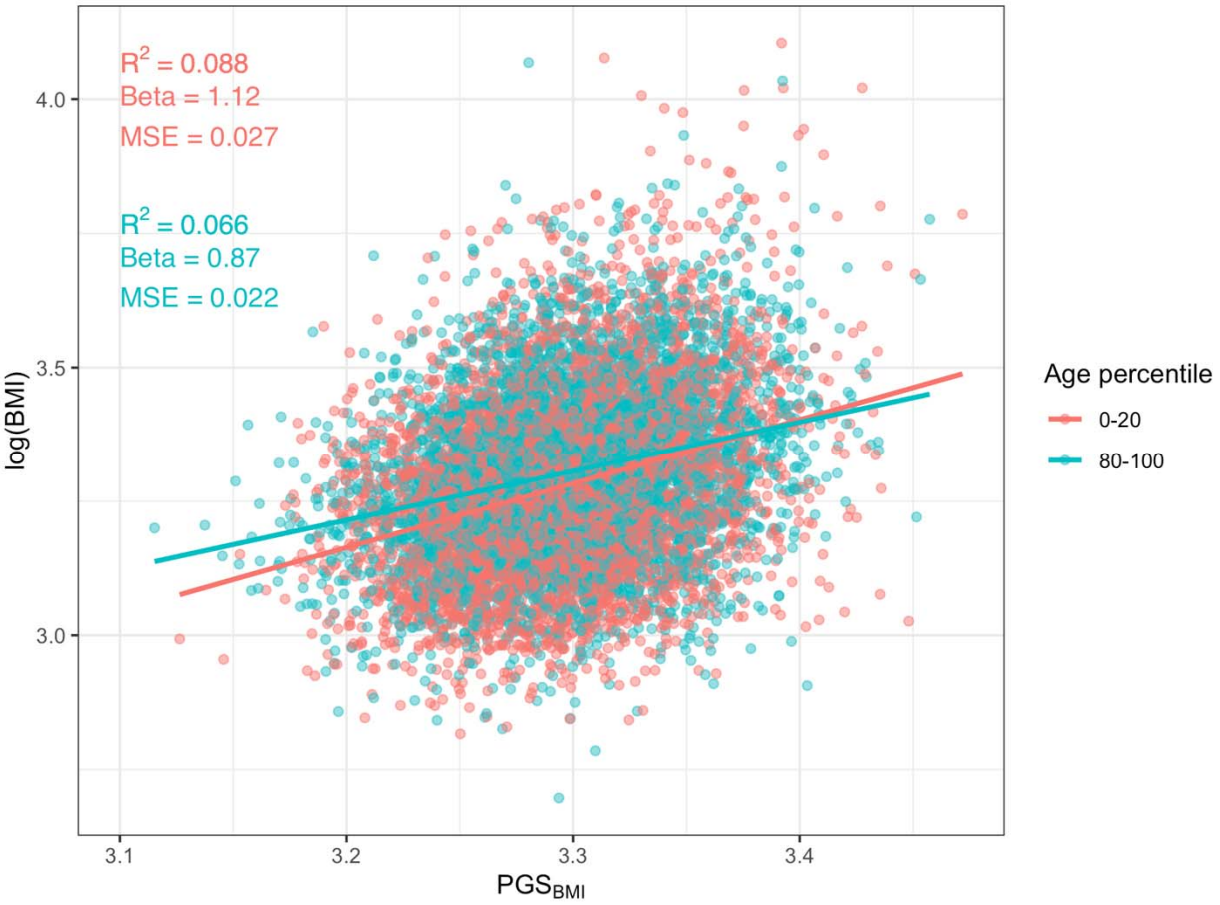
Supplemental Figures:




S Figure 1. PGS-covariate interaction term -$\log_{10}$(p-values) in UKBB EUR, with and without including the covariate PGS in the model – the mean -$\log_{10}$(p) is reduced from 18.0899 to 14.97072 with their inclusions. Note age and sex PGS were not calculated, and their interaction p-values are excluded from this figure.




S Figure 2. Three sets of simulated data with varying regression line slopes, showing how model $R^2$ changes when regression line slope changes, all else being equal. Residuals were sampled from a normal distribution (mean=0, sigma=sqrt(π/2)) to give mean squared error=1. 5,000 x-values were sampled for each line, uniformly distributed from 0-10. Despite having the same mean squared error, model $R^2$ increases as beta increases.




S Figure 3. Univariable association of $PGS_{BMI}$ and log(BMI) in European UKBB, separately for the bottom and top quintiles of age. $R^2$ is higher in younger individuals, which is partially a consequence of the larger effect (as shown in S Figure 2), despite the mean squared error actually being higher.

**Table**

| Variable type | Covariate | % change in $\beta_{PGS}$ per covariate unit change | Interaction P | $R^2$ increase with interaction term | N |
|---|---|---|---|---|---|
| Continuous | HDL cholesterol | -15.29 | $1.71 \times 10^{-96}$ | 0.0012 | 328719 |
| | Total cholesterol | -12.70 | $1.64 \times 10^{-71}$ | 0.00082 | 359221 |
| | IPAQ | -12.50 | $3.11 \times 10^{-66}$ | 0.001 | 304951 |
| | Moderate-vigorous PA | -11.41 | $8.92 \times 10^{-65}$ | 0.001 | 304951 |
| | Diastolic BP | 10.84 | $6.06 \times 10^{-60}$ | 0.0007 | 352804 |
| | Townsend Index | 6.78 | $2.86 \times 10^{-58}$ | 0.00089 | 376283 |
| | Age | -9.02 | $3.60 \times 10^{-57}$ | 0.00061 | 376729 |
| | FVC | -9.66 | $4.69 \times 10^{-56}$ | 0.0008 | 343467 |
| | Drink frequency/wk | -19.96 | $2.62 \times 10^{-55}$ | 0.0024 | 122281 |
| | LDL cholesterol | -9.86 | $2.63 \times 10^{-51}$ | 0.00058 | 358556 |
| | N days vigorous PA/wk | -9.37 | $2.42 \times 10^{-35}$ | 0.0007 | 299963 |

| | | | | | |
|---|---|---|---|---|---|
| | FEV1 | -7.38 | $7.15x10^{-35}$ | 0.0005 | 343544 |
| | Mean alcohol consumption | -7.38 | $7.65x10^{-22}$ | 0.00113 | 126756 |
| | HbA1c | 4.63 | $5.37x10^{-14}$ | 0.0002 | 358798 |
| | Mean drinks/wk | -7.66 | $1.01x10^{-13}$ | 0.0008 | 112204 |
| | Water intake | 4.60 | $2.97x10^{-13}$ | 0.00014 | 347472 |
| | Processed meat intake | 3.70 | $2.38x10^{-7}$ | 0.0002 | 376205 |
| | Starch mean | 5.51 | $3.15x10^{-7}$ | 0.00018 | 128346 |
| | Smoking pack years | 4.78 | $3.68x10^{-7}$ | 0.0002 | 114135 |
| | Protein mean | 4.82 | $6.52x10^{-7}$ | 0.00018 | 128181 |
| | Saturated fat mean | 4.92 | $1.23x10^{-6}$ | 0.00017 | 127899 |
| | Fat mean | 4.40 | $1.64x10^{-5}$ | 0.00013 | 128092 |
| | Saturated fat grams/wk | 2.46 | $1.79x10^{-5}$ | 4.00E-05 | 364629 |
| | Retinol mean | 3.77 | $3.54x10^{-4}$ | 9.00E-05 | 126029 |
| Binary | IPAQ | -12.68 | $5.30x10^{-62}$ | 0.0009 | 304951 |
| | Vigorous PA/wk | -20.55 | $9.07x10^{-54}$ | 0.0009 | 304951 |
| | Sex | -11.02 | $1.41x10^{-24}$ | 0.00025 | 376729 |
| | Diabetes | 27.19 | $1.83x10^{-7}$ | 0.0004 | 375903 |

Table 1. Model descriptive statistics on 28 of 62 covariates, which have significant ($p<.05/62$) PGS-covariate interaction terms, in UKBB EUR. The third column is the percentage change in PGS effect per unit change (standard deviations for continuous variables, binary variables encoded as 0 or 1) in covariate. The fifth column is the increase in model $R^2$ with a PGS-covariate interaction term versus a main effects only model. Abbreviations: blood pressure (BP), physical activity (PA), forced vital capacity (FVC), forced expiratory volume in 1-second (FEV1), International Physical Activity Questionnaire (IPAQ).