White Paper 23-21

A Generalized Method for the Creation and Evaluation of Polygenic Scores

James R. Ashenhurst*, Jianan Zhan*, Michael L. Multhaup, Ryo Kita, Olga V. Sazonova, Becca Krock, Sarah B. Laskey, Suyash Shringarpure, Melanie Wallace, Peter Chisnell, Nicholas A. Furlotte, 23andMe Research Team, Bertram L. Koelsch

*These authors contributed equally to this work

Links to Appendices containing key information for models developed under these methods:

2020: https://permalinks.23andme.com/pdf/23_21-PRSMethodologyAppendix_2020.pdf

2021: https://permalinks.23andme.com/pdf/23_21-PRSMethodologyAppendix_2021.pdf

2022: https://permalinks.23andme.com/pdf/23_21-PRSMethodologyAppendix_2022.pdf

2023: https://permalinks.23andme.com/pdf/23_21-PRSMethodologyAppendix_2023.pdf

Alphabetic List with Links to Specific Report Information

Report Title	Year of Report Release and Link to Information
Anxiety	2022
Asthma	2022
Atrial Fibrillation	2020
Attention Deficit Hyperactivity Disorder (ADHD)	2023
Basal and Squamous Cell Carcinoma	2022
Cat Allergy	2021
Coronary Artery Disease	2020
Diverticulitis	2022
Dog Allergy	2021
Eczema (Atopic Dermatitis)	2021
Fibromyalgia	2022
Gallstones	2021
Gestational Diabetes	2021
Glaucoma	2022
Gout	2020
Hashimoto’s Disease	2022
HDL Cholesterol	2021
High Blood Pressure	2020
Insomnia	2023
Irritable Bowel Syndrome	2022
Kidney Stones	2021
LDL Cholesterol	2020
Lupus	2023
Melanoma	2022
Migraine	2020
Nearsightedness	2021

Nonalcoholic Fatty Liver Disease	2020
Obstructive Sleep Apnea	2020
Panic Attacks	2023
Polycystic Ovary Syndrome	2021
Preeclampsia	2023
Psoriasis	2022
Restless Legs Syndrome	2020
Rosacea	2022
Seasonal Allergies	2022
Severe Acne	2021
Triglycerides	2021
Uterine Fibroids	2020

Introduction

Polygenic scores (PGS) estimate the heritable portion of risk for many common chronic diseases and other traits. Genome-wide association studies (GWAS) frequently identify multiple genetic variants with small to moderate individual impact on risk for a condition. To quantify the cumulative impact of these variants on risk, machine learning methods are used to construct statistical models that generate polygenic scores. Recently, advances in modeling methodology have enabled massive increases in the number of genetic variants that can be included in polygenic models, leading to corresponding increases in the proportion of trait variance that these models explain (So & Sham, 2017; Yang et al., 2010). As a result, PGS are now being used to estimate heritable risk for a wide range of conditions and research is ongoing to evaluate their potential utility as part of clinical decision making (Khera et al., 2018).

The key factor that limits researchers' ability to create large polygenic models is the size of the training cohort. Very large sample sizes are necessary both to identify genetic variants associated with a disease and to estimate their joint contribution to risk (Dudbridge, 2013). Additionally, obtaining samples from diverse populations is necessary to create models that are calibrated to these populations, whether by assessing how well a model developed using data from a particular ancestry group (usually European) generalizes to other (usually non-European) groups, or by developing models using data from various populations themselves (Duncan et al.,

2019). With over thirteen million kits sold and approximately 80% of customers — including customers of many different ancestries — consenting to participate in research, 23andMe has a unique ability to develop large PGS that predict a wide range of health conditions and traits and to optimize and asses PGS performance across people with diverse ancestral backgrounds. . Analyses of the company's genetic and self-reported health data show that we can replicate GWAS on clinically collected health information (Tung et al., 2011). Over the last several years, 23andMe has used PGS as the basis of dozens of customer reports on topics ranging from the ability to match a musical pitch to the likelihood of developing type 2 diabetes (Furlotte et al., 2015; Multhaup et al., 2019).

Here we detail the modeling methodologies and evaluation procedures used to create the PGS behind 23andMe Health Predisposition and Wellness reports on common health conditions (Figure 1). As an example, we detail how this methodology was used to create and evaluate the PGS used in 23andMe's LDL Cholesterol report. The Appendices to this White Paper further summarize the performance and characteristics of each PGS used in recently released reports. We intend for this White Paper and the Appendices to be living documents that will be updated as methodologies change and new PGS-based genetic reports are released. A change log is provided at the bottom of this document to describe significant updates.

Methods

Phenotype validation

Previous analyses of 23andMe survey data have demonstrated the capacity of the research platform to replicate published results (Tung et al., 2011). Nevertheless, as all phenotypes are derived from self-reported survey data, we assess each phenotype used to create a PGS to determine whether it adequately captures the intended concept. First, we compare the prevalence of the phenotype across the dimensions of age, sex, and ancestry to prevalence values reported in published literature. While overall prevalence values may differ due to differences between the composition of 23andMe research participants and other large cohorts, demographic trends should be broadly consistent. In other words, a phenotype that is more prevalent among men than women or more common in older than younger individuals should show these trends in both the 23andMe research participant population and in other cohorts.

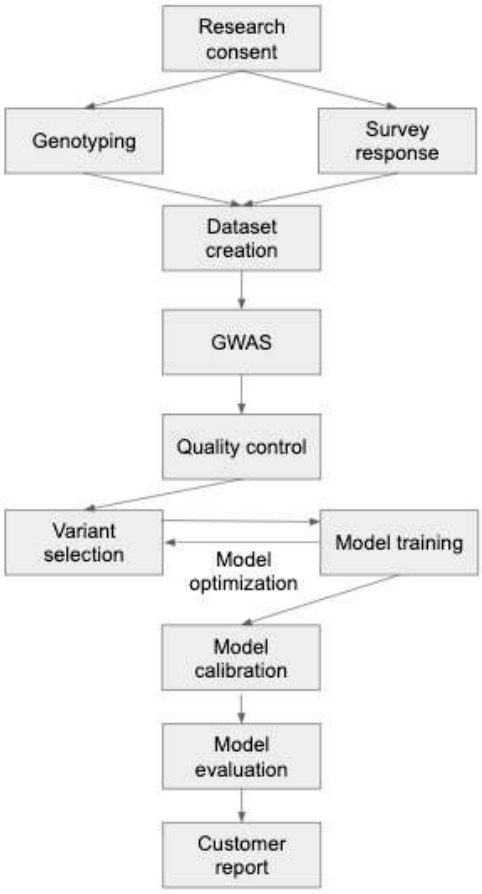


Figure 1: Outline of 23andMe's PGS creation procedure from self-report of survey data to generation of the polygenic model powering a health report.

Next, if there are well-established correlates or predictors of the phenotype and survey questions about these correlates are available in the 23andMe database, we attempt to replicate these associations using generalized linear models as an additional check of construct validity. For example, because body mass index (BMI), high LDL cholesterol, and type II diabetes are known risk factors for coronary artery disease (CAD; Arnett et al., 2019), we would expect associations between these characteristics at baseline and self-reported incident CAD to be comparable in direction and magnitude to clinically ascertained samples.

Lastly, we may assess whether the summary statistics from our GWAS replicate published GWAS results on the same or similar phenotype, if available. The primary metric for this comparison is the correlation between the effect sizes of independent genome-wide significant SNPs present in both summary statistic sets. If the 23andMe survey data is a good representation of the intended phenotype, we expect our GWAS of survey-based self-reported

phenotypes to substantially replicate published GWAS results for phenotypes obtained through clinical ascertainment or other methods. GWAS comparisons (if available) are provided for each phenotype in the Appendices to this White Paper.

Genotyping

Genetic variants are assayed using Illumina BeadChip arrays as previously described in 23andMe White Paper 23-19 (Multhaup et al., 2019). In summary, DNA is extracted from saliva samples, and genotypes are determined by the National Genetics Institute (NGI), a subsidiary of the Laboratory Corporation of America and a Clinical Laboratory Improvement Amendments (CLIA)-certified clinical laboratory. To date, most samples were run on one of three Illumina BeadChip platforms: Illumina HumanHap550+ BeadChip platform augmented with a custom set of ~25,000 variants (V3); the Illumina HumanOmniExpress+ BeadChip with a baseline set of 730,000 variants and a custom set of ~30,000 variants (V4); and the Illumina Infinium Global Screening Array (GSA), consisting of 640,000 common variants supplemented with ~50,000 variants of custom content (V5). Samples with a call rate of less than 98.5% are discarded.

Dataset creation

Research participants included in datasets used for PGS creation are all 23andMe customers who have consented to participate in research and have answered survey questions required to define the phenotypes of interest. Both males and females and participants ages 20 to 80 are included unless otherwise specified in the Appendices. For any groups of related participants with identity-by-descent of more than 700 centimorgans, individuals are removed from the dataset until only one is left, preferentially retaining the less common phenotype class. Research participants are grouped as per Campbell et al. (2015) into Sub-Saharan African/African American, East/Southeast Asian, European, Hispanic/Latino, South Asian, and Northern African/Western Asian datasets. Any additional inclusion or exclusion criteria for each phenotype are described in their corresponding summaries in the Appendices. For each phenotype, training, validation, and testing cohorts are defined in groups with sufficient data. Details of how data representing different populations are split for each phenotype are found in the Appendices. Whereas the GWAS dataset includes individuals genotyped on multiple genotyping platforms, the training, validation, and testing datasets are restricted to individuals genotyped on the V5 array as these model results are delivered only to customers genotyped on this array.

Genome-wide association study (GWAS)

GWAS are performed as described previously (Tian et al., 2017), except that they are restricted to the union of variants genotyped on the V3, V4, and/or V5 arrays. Participants included in the GWAS may be in the model training set depending on their genotyping array version, but are not included in the validation or testing sets.

Variant and model selection

After running GWAS, variants are filtered to exclude those that do not pass GWAS quality control metrics: parent-offspring transmission, Hardy-Weinberg $p < 1e\text{-}20$, large sex effects, multiple reference sequence matches, significant genotyping date associations, genotype rate ≤ 0.95, imputed estimated R-squared ≤ 0.8, minor allele frequency ≤ 0.01, minor allele frequencies below 0.5% across several ethnicities, and other internal variant data quality filters.

To select variant sets, we perform pruning and thresholding with combinations of selection hyperparameters. For example: distance (kb) = [10, 100, 200, 1000, 2000], and GWAS *p*-value = [1e-2,1e-4, 1e-6,1e-8]. Variant sets up to a pre-specified maximum size are kept for hyperparameter evaluation. Variant selection hyperparameter evaluation is performed by fitting a model with each variant set in the training cohort and evaluating in the validation cohort. As described above, the validation cohort is distinct from the training and testing cohorts and no sample sets contain close relatives within or between sets.

Models typically include the first ten genomic principal components, age, and genetically determined sex (unless the phenotype is single sex only). The variant data are V5 platform genotype calls, and missing values are filled in with population mean dosages. The variant set with the highest area under the receiver operator curve (AUCROC) in the validation sets for a specific cohort is designated as the optimal feature set. Final fit statistics are obtained using the test set, which was held out of all upstream analyses. Variations in this approach are described in the phenotype-specific Appendices.

Model features

Features used in the model training typically include genomic principal components (PCs), demographic factors like age, sex, higher-order terms of age, interactions terms between demographic factors, and dosages for the variants. Variants on the X chromosome for males are modeled as a dominance effect (encoded 0 or 2). The purpose of including genomic PCs in the

regression is to account for any residual population substructure. While these genomic PCs and other non-genetic factors are used to create the PGS and related risk estimates, weights for features other than genetic variants are set to zero when computing the PGS so that the customer-facing qualitative results are based on only genetic variation identified in the GWAS. Absolute risk estimates associated with a PGS (the quantitative results) take into account self-reported birth sex and genetic ancestry as described previously (Campbell et al., 2015).

Model training

PGS are built using regression methods based on generalized linear models (GLM). Individual-level data, rather than GWAS summary statistics, are used to train these PGS. Features including genomic PCs, dosages for each variant, and demographic factors are treated as independent variables. For binary phenotypes, we use multivariate logistic regression under a general linear model framework. For quantitative phenotypes, we use linear regression. After a model is specified, weights for each feature are calculated through regression. We use Scikit-learn's LogisticRegression gradient descent algorithms (Pedregosa et al., 2011) to determine optimal parameter weights, typically with the liblinear solver and L2 regularization.

Methods for non-European ancestries

One of the biggest challenges for PGS today is transferability between ancestries (Martin et al., 2017). Individuals of European descent make up the overwhelming majority of genetics research participants even though they represent a minority of global genetic diversity (Popejoy & Fullerton, 2016). As a result, PGS trained with data from individuals of European descent typically perform worse among individuals of other ancestries.

We leverage our large research participant population with non-European ancestry to address this challenge using four possible approaches for each ancestry-phenotype pair, as sample sizes permit. Overall, no single method always works for every phenotype-ancestry combination. The specific method used for each ancestry group is considered a hyperparameter and optimized on a case-by-case basis as described in the Appendices for each phenotype. Note that all validation and testing are done in ancestry-specific datasets to avoid overestimation of performance metrics.

First, for phenotypes and ancestries with relatively large sample sizes, separate GWAS are run for each group, and ancestry-specific PGS are created from these ancestry-specific GWAS. However, for many phenotypes we do not have enough survey responses to run

sufficiently powered GWAS independently for all ancestry groups. Our second approach is to leverage information from the European GWAS to boost power for the non-European GWAS. To accomplish this, we perform a meta-analysis (Munafò & Flint, 2004) that combines information for each SNP across ancestries, and generate PGS leveraging training sets comprised of multiple ancestry groups (while controlling for population structure using genomic principal components). This method often yields effect size estimates that are more predictive for the specified non-European cohort. If this method does not improve performance for a given non-European ancestry group, the third approach we attempt is to run a GWAS and train a PGS using European-ancestry data, with model hyperparameters optimized based on performance in a validation dataset consisting of data from the non-European ancestry group. Finally, if none of these three methods is able to improve performance over simply applying the European-trained PGS to the non-European ancestry group, the European-trained PGS is used to deliver results to non-European customers.

Platt scaling

After model training, the PGS may be overfit to their training datasets. This can lead to miscalibration when applied to other datasets (especially those from individuals of non-European descent). To recalibrate the PGS model, the cumulative effect size of the PGS is re-estimated using a procedure known as Platt scaling, as described previously (Multhaup et al., 2019). Briefly, PGS values are calculated for each participant in all datasets. These original values are then standardized to fit the normal distribution. Then, separately in each test set, a secondary generalized linear model is fit to re-predict the outcome variable using the normalized PGS as a single predictor. These linear models are then used to adjust PGS scores for each individual. As these linear models are trained separately in each dataset, the coefficient of the PGS and the intercept in these models are specific to that dataset, accomplishing recalibration. The testing datasets are usually ancestry-specific or ancestry- and sex-specific.

Assessing model performance

Final performance statistics for European-trained models are determined in the European test set, which is not included in the GWAS, hyperparameter tuning, or any model fitting. Similarly, statistics for non-European groups are determined in test sets that were not included in any previous stage of analysis.

Ancestry-specific model performance is evaluated using the following metrics (and corresponding plots): 1) area under the receiver operator curve (AUROC), 2) risk stratification, estimated as odds ratios and relative risks for those in the upper segments of the distribution compared to those in the middle of the distribution (40th to 60th percentiles), 3) an estimation of AUROC within each decade of age — to assess age-related biases in model performance — and 4) calibration plots between PGS quantiles after Platt scaling and phenotype prevalences in each ancestry group.

PGS result categorization

For simplicity and clarity, we separate results from the PGS into three categories. The first represents individuals at increased likelihood of developing the condition and the second represents typical — i.e., not increased — likelihood of developing the condition. This is accomplished by determining a threshold (a specific level of risk defined by an odds ratio or relative risk) and then calculating the specific PGS value that corresponds to that threshold such that everyone with a higher PGS has at least that level of risk. A detailed explanation of this binarization into increased and typical likelihood is provided in 23andMe White Paper 23-19 (Multhaup et al., 2019). The third category represents individuals who have a very low likelihood of developing the condition, as defined by having an estimated likelihood of developing the condition (as defined in the next section) of less than 1 out of 1,000 . This category is intended to avoid giving overly precise results to customers, as the PGS and report do not factor in all genetic and lifestyle factors, and to help customers more easily interpret their overall level of absolute risk for developing the condition.

Estimated likelihoods

For each customer, the report result is presented as the likelihood of developing a condition by some target age (e.g., their 70's). This estimated likelihood is derived by multiplying an estimated genetic relative risk by an age- (and potentially sex- and ancestry-) specific baseline condition prevalence at the target age. Baseline prevalence values are derived from either external datasets, if available, or the 23andMe database. If there is not a clear match between a population in an externally derived baseline and a 23andMe ancestry group, the European baseline is provided instead because it is the largest available sample. The specific datasets used to calculate baseline prevalences for each phenotype are described in the Appendices.

The method for deriving estimated relative risks associated with a genetic result is as follows. First, PGS are standardized within each ancestry-specific test set, and PGS distributions are segmented into bins corresponding to percentiles. We use 92 bins, with the lowest and highest 5% of customers placed into single bins, and 90 intermediate bins each capturing 1% of the PGS distribution between these extremes. We chose to use larger bins at the extremes to avoid over- or under-estimating probabilities at the extreme tails of the PGS distribution. We bin participants rather than provide unique estimates for each individual because, in most cases, the customer-facing result is rounded to a whole percent. Customer-facing results are given to one decimal place only in the case of especially low estimates, for which this level of precision is necessary for customers to distinguish their unique likelihood estimates from what is considered typical (defined as the range between the minimum likelihood and the threshold between typical and increased likelihoods for the customer's specific genetic ancestry and birth sex).

Next, we calculate model-estimated prevalences for each genetic result bin at the target age of the report result. This is accomplished by re-estimating the prevalences for the test sets with the age parameter set as the target age (along with age-related covariates like any age-by-sex interaction terms) for the whole test set. In this way, we leverage the full (genetics + demographics) model to estimate prevalences for each ancestry group at the target age for both sexes. We generate these model-estimated prevalences because the sample size of every ancestry-specific test set is usually not sufficient to calculate reliable observed prevalences stratified by sex, age, and PGS percentile.

These estimated phenotype prevalences at the target age are Platt scaled to adjust for any miscalibration within each ancestry group. The parameters used for Platt scaling are based on the distribution of estimated probabilities given participants' actual ages (i.e., Platt scaling parameters are not re-estimated when age is fixed for the whole sample).

These scaled estimated phenotype prevalences are transformed into relative risks with reference to the median of each ancestry group's PGS distribution. In other words, the estimated prevalence for a particular genetic score percentile at the target age for a given sex is divided by the estimated prevalence at the median PGS for that group. The resulting values represent estimated relative risks based on the full model (including both genetic and demographic features) across the dimensions of genetic risk and demographics. We then multiply these relative risks by the baseline prevalence values to yield target age-linked estimated likelihoods.

These estimates should be interpreted in light of several limitations to this approach. First, for conditions linked to higher rates of mortality, baseline prevalence estimates at

advanced ages gathered from cross-sectional data sources likely undercount the true cumulative incidence of a condition. As such, these estimates represent the likelihood of having the condition assuming survival to a particular age. While other modeling strategies, like competing risks models (Gail et al., 1989), could be used to account for loss in participation due to mortality or other causes, they require detailed incidence data that are often unavailable. Furthermore, likelihood estimates as computed here only take into account risk stratification due to common variants. There are often rare variants that could be used to estimate a more comprehensive total lifetime risk. Additionally, many non-genetic factors, often including lifestyle, contribute to total risk for many conditions.

Quality control measures

Given that these polygenic models encompass thousands of variants, it is possible that an individual may not have genotype calls for a subset of markers included in a particular model. For those missing genotype calls, we impute to the population mean dosage to calculate an individual's score. Consequently, these missing values can introduce uncertainty as to whether or not a customer's score is above or below the binary qualitative result threshold.

In order to estimate this uncertainty, we use a metric similar to a Z score that includes information about a variant's effect size (β), its effect allele frequency (*p)*, and an individual's distance from the binary result threshold. For each missing genotype call *i* across *n* missing calls, we use the below equation to determine the ratio between the distance of an individual's score from the threshold and the uncertainty in the score due to missing values.

$$\frac{threshold - PGS}{\sqrt{\sum_{i=1}^{n} 2\beta_i^2 \cdot p_i(1-p_i)}}$$

As this metric approaches zero, the probability that a customer's score could be on the other side of the threshold increases to a maximum of 50%. If an individual's score has greater than a 1% chance of being on the other side of the binary threshold due to the specific missingness patterns in their data, the customer is alerted to the possibility that their qualitative result could differ if they were genotyped again and these missing values were called.

Irrespective of this metric, no result is provided to customers missing genotype calls at more than 10% of the markers in a particular model.

Validation in external datasets

In order to assess the generalizability of these models, we have assessed the performance of select PGS models in external datasets. Specifically, we are looking to understand how well these models can both stratify risk and provide accurate risk assessment outside of 23andMe research cohorts. To do this we conducted analyses on both White and Black cohorts from the Atherosclerosis Risk in Communities (ARIC) study (dbGaP accession phs000280.v7.p1). Both phenotype data (dbGaP accession pht000114.v3.p1) and genetic data (dbGaP accession phg000248.v1) for this analysis was from the ARIC Gene Environment Association Studies (dbGaP accession phs000090.v7.p1). Started in 1987, ARIC is a longitudinal study of nearly 16,000 participants that collected phenotype data on cardiometabolic conditions as well as genotype data on its participants. Within these cohorts we calculated PGS from our models, harmonized phenotypes to closely match those that were used to train our PGS models, and tested the performance of our PGS models on predicting these phenotypes. Our general methodology is described below, and the specific descriptions, data, and results for each phenotype are located in the White Paper appendices for each specific phenotype. Currently, phenotypes that have been externally validated are as follows: coronary artery disease, high blood pressure, LDL cholesterol, HDL cholesterol.

Phenotype Selection

The phenotypes available in the ARIC dataset are not precisely the same as those collected by 23andMe. We attempted to match phenotypes as closely as possible to those our PGS models were trained to assess. Cases and controls were defined using self-reported, clinical examination and biomarker data when available.

PGS Calculation

PGSs were calculated as weighted summations of ARIC SNP dosages with 23andMe PGS weights, using the overlapping SNPs between the PGSs and ARIC's 1000G imputation panel. The alignment of SNPs between the PGSs and the ARIC data was performed based on chromosome number and position, using the Genome Reference Consortium human genome

build 37 (GRCh37) as the reference genome, allowing for the accurate identification and matching of SNPs between the two datasets. To ensure consistency in the interpretation of the 23andMe PGs across the two datasets, we reversed the direction of the beta estimate of any SNPs that had a different coding allele in the ARIC study versus the 23andMe PGS.

PGS Performance Validation

For each PGS we assessed model performance in the ARIC cohorts using AUROC, odds ratios, and calibration plots showing actual prevalence versus predicted prevalence for each ventile of PGS score. For the calibration plots, we set the baseline log-odds intercept based on the measured prevalence of the condition in the cohort, and then scaled z-scores from the ARIC cohorts using the PGS model platt coefficient.

Case study: High LDL Cholesterol

Here we demonstrate the above methods applied to a model predicting the genetic risk of ever having high LDL cholesterol (LDL-C) levels. This section is intended to walk the reader through the generation of the consumer-facing results for this model. Details of the remaining PGS-based health reports are described in the separate Appendices to this White Paper.

Dataset

Individuals eligible for the development of the LDL cholesterol model were 23andMe customers who provided informed consent and answered survey questions pertaining to LDL-C and a history of cholesterol-lowering medication (Table 1).

High LDL cholesterol phenotype definition

The phenotype used to develop polygenic models represented self-report of ever having had high LDL-C or ever having been prescribed medication to lower cholesterol, an indication that a physician likely determined that the respondent had high LDL-C. This phenotype combined responses from three questions pertaining to the most recent LDL-C, highest ever LDL-C, and medication history. Prevalence increased with advancing age (Figure 2). Additional detail about survey questions is provided in the 2020 Appendix to this White Paper.

Comparison to previously published GWAS

The largest external GWAS to date on measured LDL-C levels (as published by the Global Lipid Genetics Consortium, GLGC) yielded many genome-wide significant loci (Willer et al., 2013). In order to validate the results of the 23andMe GWAS, we compared the summary statistics of these two GWAS. A Manhattan plot below (Figure 3) shows an overlay of the *p*-values obtained by the two GWAS, rescaled to have similar peak heights. This plot demonstrates the substantial overlap between the two GWAS, confirming that the self-reported phenotype used in the 23andMe data, while not as granular as quantitative laboratory measures of lipid levels, does indeed capture genetic signals highly similar to those captured by GWAS of LDL-C levels measured from blood. Among the 471 genome-wide significant hits in the GLGC data (obtained after pruning and clumping the summary statistics), 407 (86.4%) were also genome-wide significant in the 23andMe GWAS.

Table 1: High LDL-C participant cohort descriptives

Platform	Ancestry Group	Sample Use	N	Age mean (SD)	Sex (% female)	High LDL-C Prevalence (%)
V1 to V5	European	GWAS	617,165	56.2 (13.8)	54.60%	41.99%
V5	European	Training the European Model	511,469	55.0 (13.9)	55.50%	40.60%
		Testing	56,749	55.1 (14.0)	55.24%	40.94%
	Sub-Saharan African/African American	Testing	18,710	50.1 (13.5)	59.02%	40.94%
	East/Southeast Asian	Testing	18,357	44.7 (14.2)	57.51%	27.07%
	Hispanic/Latino	Testing	72,806	47.8 (14.0)	56.46%	33.86%
	South Asian	Testing	6,128	44.3 (13.0)	37.73%	34.48%
	Northern African/ Western Asian	Testing	5,267	49.4 (14.7)	40.38%	38.47%

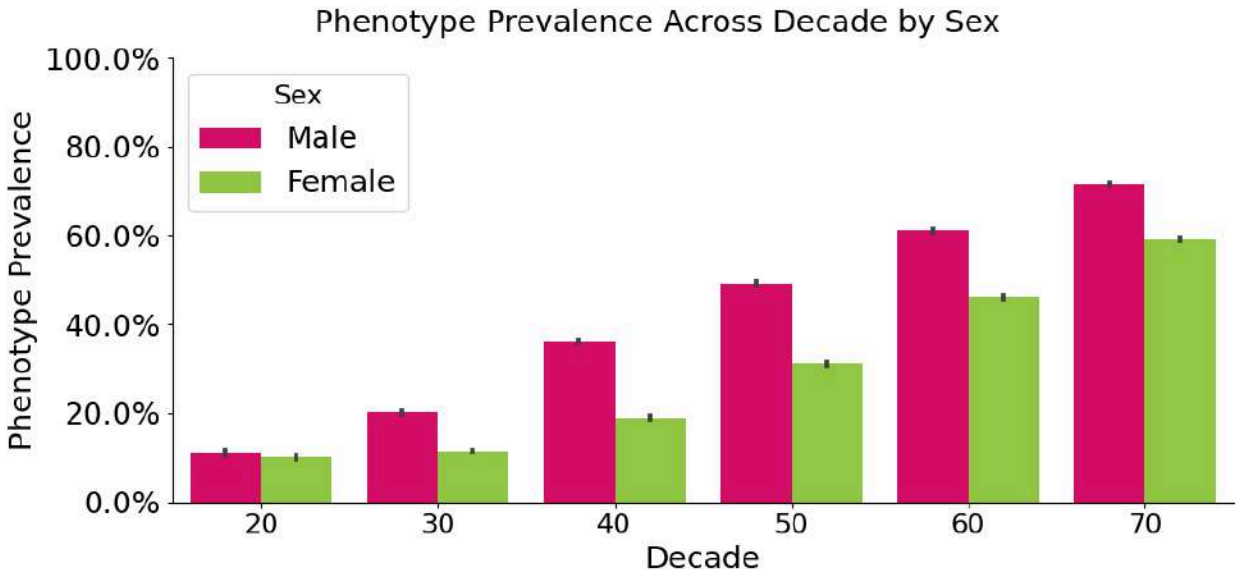


Figure 2: High LDL-C phenotype distribution in the European ancestry training set

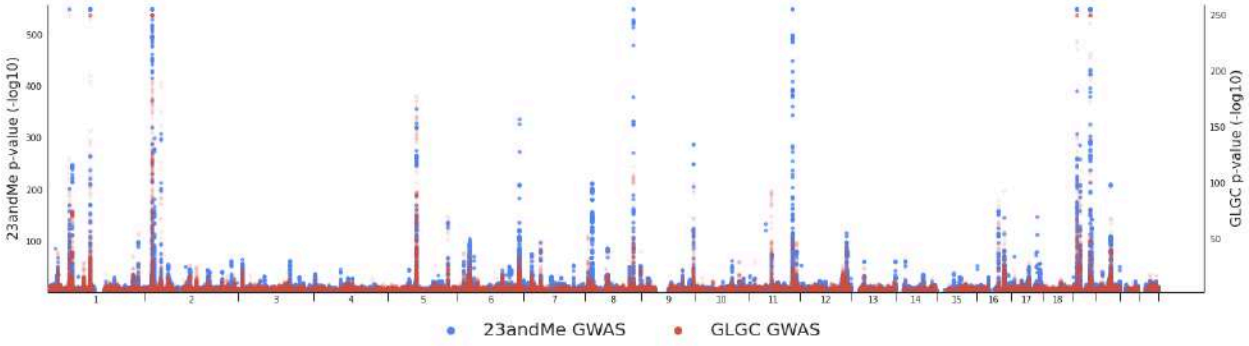


Figure 3: Manhattan plot of 23andMe and Willer et al., (2013) GWAS summary statistics for LDL-C. The *p*-values for each variant shared between the 23andMe (blue) and Global Lipids Genetics Consortium (GLGC; red) genome-wide association studies (GWASs) are depicted. Chromosomal location is represented on the horizontal axis, and the negative log of the *p*-value is represented on the vertical axis (scaled separately for each analysis).

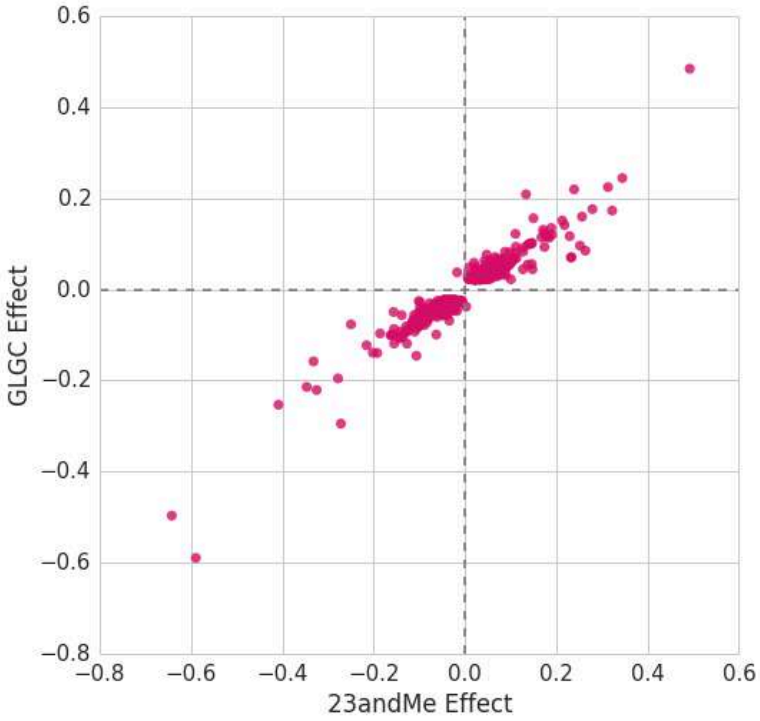


Figure 4: Scatter plot showing the estimated effect sizes for 23andMe (change in log-odds per unit predictor change) and Global Lipids Genetics Consortium (GLGC; linear betas; Willer et al., 2013) genome-wide significant hits shared between the two GWAS for LDL cholesterol.

Next, as an additional validation of the 23andMe GWAS, the effect sizes of all independent genome-wide significant loci found in both sets of summary statistics were compared. These effect sizes should be similar in scale and with the same positive or negative valence. We determined the correlation between these two sets of effect sizes after reformatting the data to align all strand and reference alleles and selecting independent variants using clumping and pruning procedures in PLINK (Chang et al., 2015; Purcell et al., 2007; parameters *p*-value = 5e-8, r^2 = 0.5, distance = 250kb). As shown in Figure 4, all but two genome-wide significant loci showed the same positive or negative valence in the GWAS, and the effect sizes were strongly correlated. The replication of the majority of previously identified loci in addition to the correlated effect sizes demonstrates that the 23andMe GWAS based on self-reported data adequately captured the results of the external GLGC GWAS, which was based on clinically ascertained laboratory values.

Model performance

Demographic covariates included in polygenic modeling for LDL-C were age, sex, age^2, as well as sex-by-age and sex-by-age^2 interaction terms. Model training and hyperparameter tuning was performed in samples of European descent, as described in Methods. The final selected model contained 2,950 genetic variants.

Performance and calibration statistics were assessed as described (see Methods). As expected, the PGS performed best in individuals of European ancestry, followed by individuals of Hispanic/Latino, South Asian, and Northern African/Western Asian ancestry, and finally in Sub-Saharan African/African American and East/Southeast Asian ancestries (Table 2, Figures 5-7). In all these populations, however, the odds ratio for high LDL-C for individuals in the top 5% of the (genetics-only) PGS versus individuals with average PGS was close to or higher than two, indicating that the PGS was able to stratify a substantial amount of risk for those at the right tail of the distribution. Additionally, Platt-scaled calibration plots illustrate a high correlation of predicted versus real prevalences in all ancestries (Figure 8).

Qualitative result thresholding

We used standardized (within each population) polygenic scores to determine the population-specific threshold corresponding to an odds ratio of 1.5 relative to the 40th to 60th percentile of each population's distribution. Table 3 shows the proportion of customers above this threshold, who would thus receive the "increased likelihood" result. Likelihood ratios associated with the "increased likelihood" result are also provided in Table 3.

Table 2: High LDL-C PGS performance characteristics

Ancestry Group (test sets)	**Full Model AUROC**	**Genetics Only AUROC**	**Odds Ratio top 5% versus average (95%CIs)**	**Odds Ratio top 5% versus bottom 5% (95%CIs)**
European	0.7770	0.6456	2.81 (2.58 to 3.07)	10.24 (9.02 to 11.63)
Sub-Saharan African/African American	0.7312	0.5985	1.91 (1.67 to 2.23)	4.10 (3.34 to 5.05)
East/Southeast Asian	0.7635	0.5888	1.91 (1.64 to 2.22)	4.30 (3.43 to 5.39)
Hispanic/Latino	0.7561	0.6179	2.31 (2.15 to 2.49)	5.87 (5.27 to 6.55)

South Asian	0.7828	0.6222	2.69 (2.08 to 3.47)	7.75 (5.29 to 11.37)
Northern African/Western Asian	0.7776	0.6188	2.81 (2.13 to 3.72)	7.49 (5.04 to 11.14)

Table 3: High LDL-C qualitative result characteristics

Ancestry Group (test sets)	**Odds Ratio for Result Threshold**	**Percent Above Threshold**	**Likelihood Ratio of "Increased" Result (95% CIs)**
European	1.5	22.79%	1.97 (1.91 to 2.03)
Sub-Saharan African/African American	1.5	12.32%	1.69 (1.56 to 1.82)
East/Southeast Asian	1.5	10.37%	1.63 (1.50 to 1.78)
Hispanic/Latino	1.5	17.19%	1.80 (1.74 to 1.86)
South Asian	1.5	18.29%	1.82 (1.64 to 2.02)
Northern African/Western Asian	1.5	17.47%	1.85 (1.64 to 2.08)

Baseline prevalences used for likelihood estimate

Ancestry- and sex-specific baseline prevalences of ever having had high LDL cholesterol were derived from the 2017 data release of the Behavioral Risk Factor Surveillance System (BRFSS; Centers for Disease Control and Prevention [CDC], 2017). The specific calculated variable (coded _RFCHOL1) represents the concept: adults who have had their cholesterol checked and have been told by a doctor, nurse, or other health professional that it was high. The ancestry variable used (coded _RACE) included the categories White non-Hispanic, Black non-Hispanic, Asian non-Hispanic, and Hispanic. Analysis was restricted only to those between the ages of 70 and 79, to capture this decade of age (coded _AGEG5YR). The baselines used for each sex and ancestry combination and how they map to each 23andMe ancestry group are shown in Table 4.

Table 4: High LDL-C baseline prevalences in BRFSS data between ages 70 to 79

Group	Matched 23andMe Population(s)	Sex	N	Prevalence	95% CI
White Non-Hispanic	European, Northern African/Western Asian, Other	Male	23,256	55.02%	54.38% to 55.66%
		Female	33,369	54.22%	53.69% to 54.76%
Black Non-Hispanic	Sub-Saharan African/African American	Male	1,335	52.58%	49.91% to 55.26%
		Female	2,795	53.42%	51.57% to 55.27%
Asian Non-Hispanic	East/Southeast Asian, South Asian	Male	327	46.18%	40.77% to 51.58%
		Female	386	51.30%	46.31% to 56.28%
Hispanic	Hispanic/Latino	Male	951	46.58%	43.41% to 49.75%
		Female	1,619	51.95%	49.51% to 54.38%

Figure 5: High LDL-C AUROC across ancestry-specific test sets

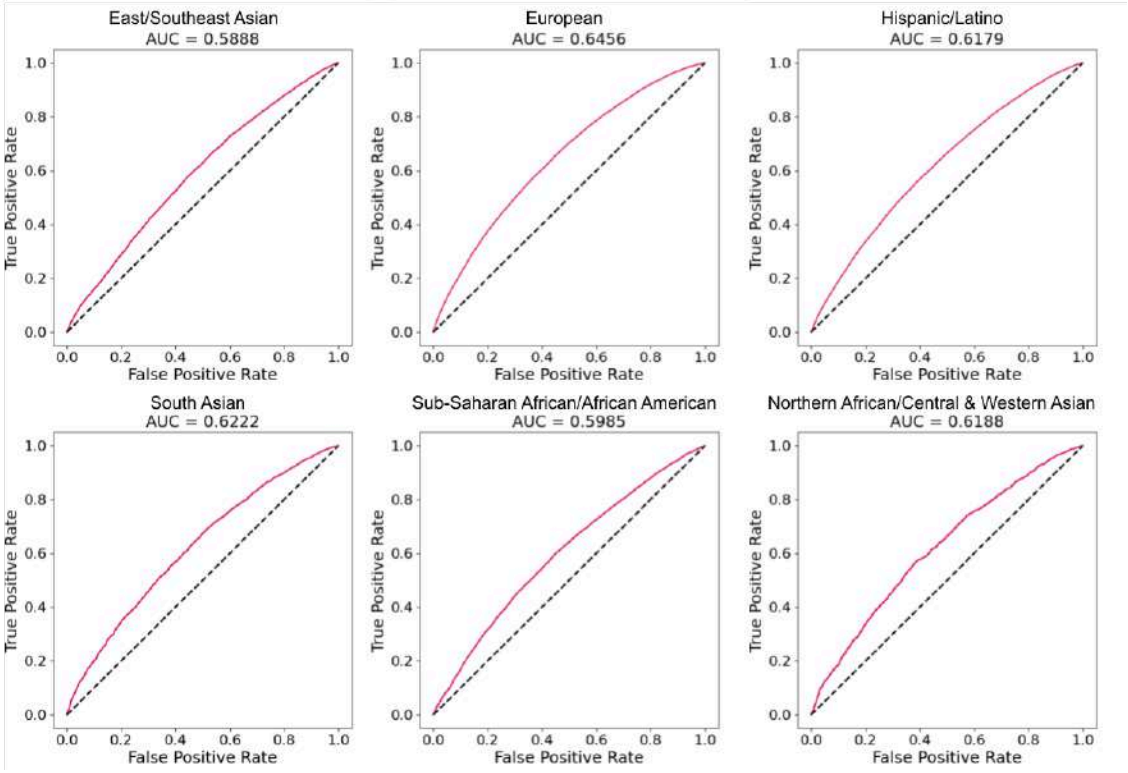


Figure 6: High LDL-C AUROC within each decade of age across ancestry-specific test sets

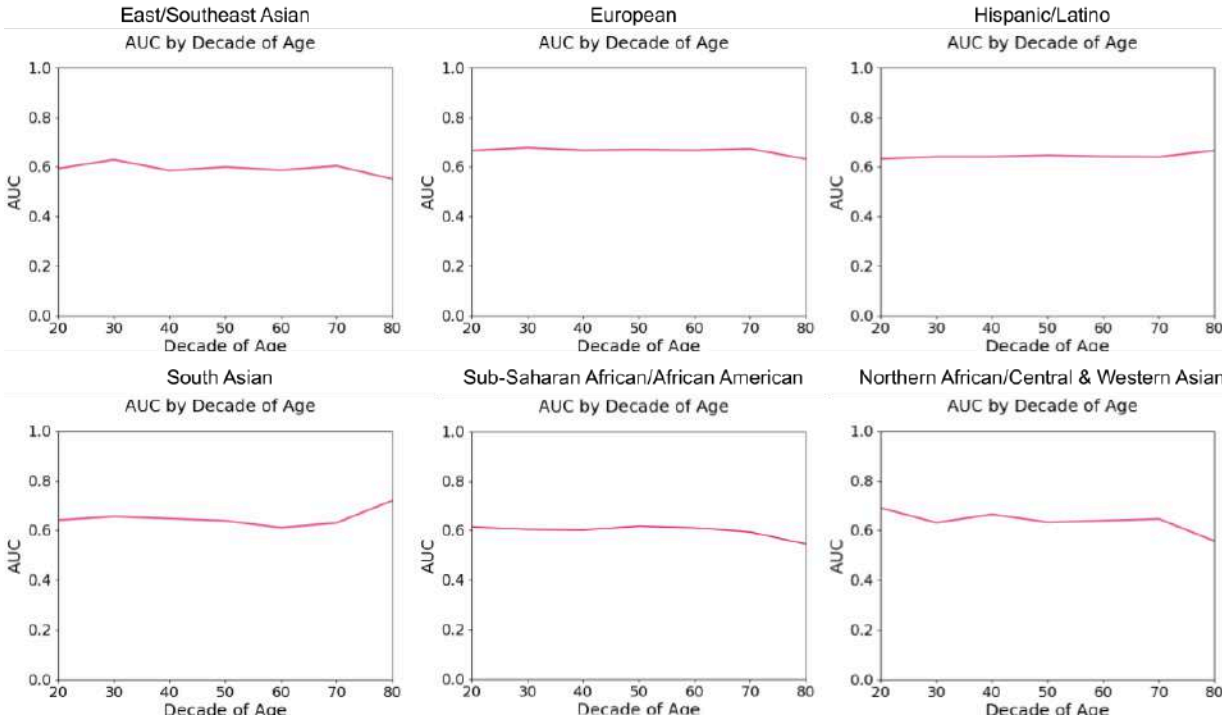


Figure 7: High LDL-C case/control standardized PGS distributions across ancestry-specific test sets

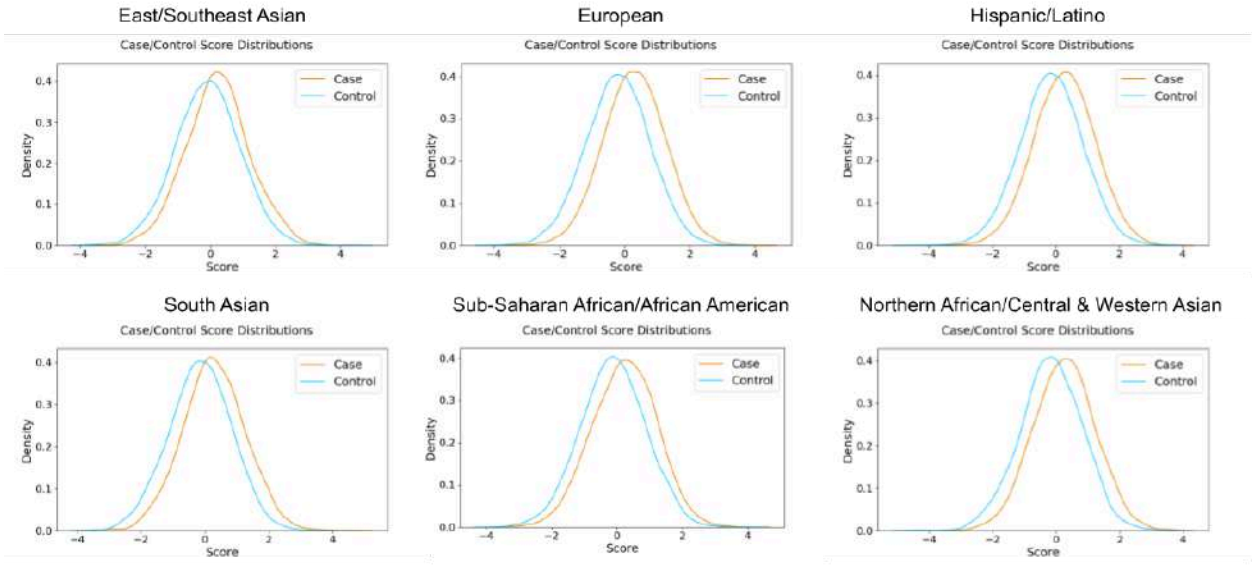
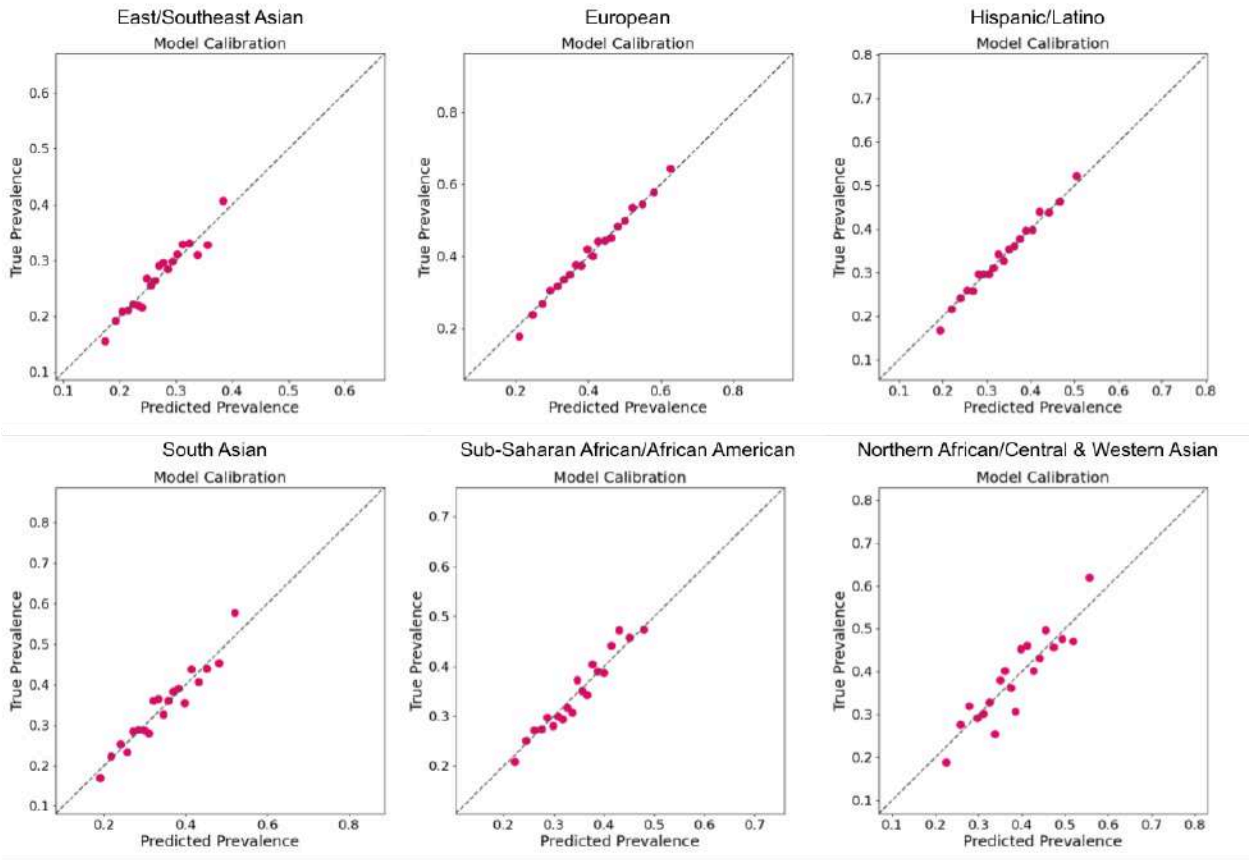


Figure 8: High LDL-C Platt-scaled calibration plots across ancestry-specific test sets



External Validation

We conducted an external analysis on our LDL PGS model in the Atherosclerosis Risk in Communities (ARIC) study cohort. To determine LDL status, we combined coded phenotypes ldlsiu02, cholmdcode01, and cholmdcode02. These phenotypes report measured LDL levels, whether the participant is taking medications to treat high cholesterol, and whether the participant is taking secondary medications that might affect cholesterol levels. Any participant who reported taking cholesterol lowering medication or had LDL > 160mg/dL was considered a case. All other participants were considered controls. Any participant taking secondary medication that might affect cholesterol levels was removed from the analysis.

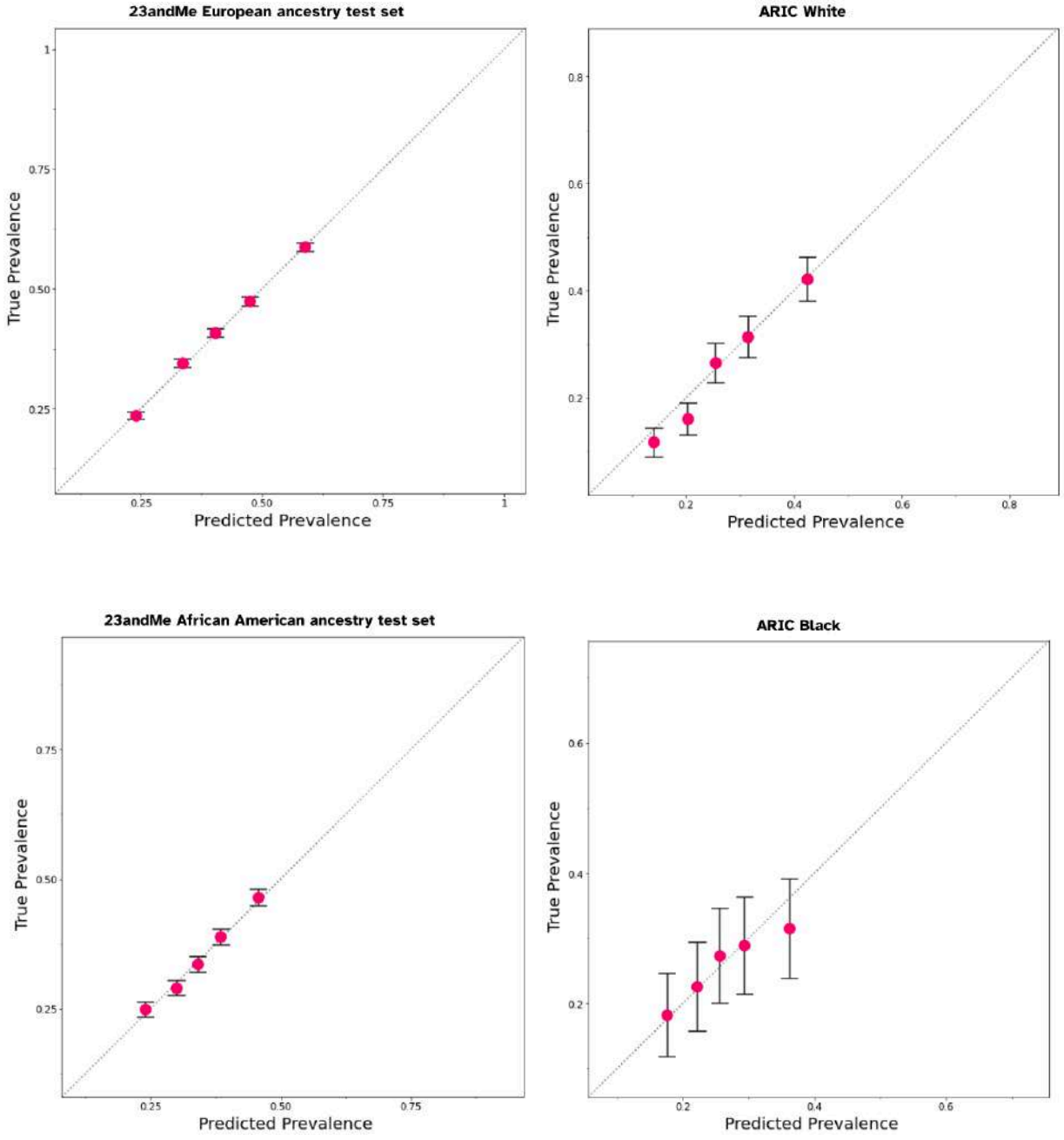
Table 5: ARIC cohort statistics for High LDL Cholesterol

Cohort	N	Mean max age (SD)	Sex (% female)	High LDL Prevalence (%)
ARIC White	2,779	53.8 (5.6)	51.85%	25.51%
ARIC Black	713	52.9 (5.7)	56.52%	25.67%

Table 6: High LDL Cholesterol PGS performance in 23andMe and ARIC cohorts: PGS stratifies risk similarly in 23andMe and external cohorts.

Cohort	Genetics Only AUC (95%CIs)	Odds Ratio top 20% versus average (95%CIs)	Odds Ratio top 20% versus bottom 20% (95%CIs)
23andMe European ancestry test set	0.6456 (0.6409 to 0.6503)	2.06 (1.96 to 2.18)	4.62 (4.37 to 4.90)
ARIC White Cohort	0.6716 (0.6475 to 0.6957)	2.02 (1.57 to 2.60)	5.49 (4.03 to 7.47)
23andMe African American ancestry test set	0.5985 (0.5898 to 0.6072)	1.72 (1.56 to 1.88)	2.62 (2.38 to 2.89)
ARIC Black Cohort	0.5735 (0.5245 to 0.6226)	1.22 (0.74 to 2.04)	2.07 (1.19 to 3.59)

Figure 9: Calibration of High LDL PGS in ARIC cohorts: PGS scores remain well calibrated in both ARIC cohorts.



Acknowledgements

We thank the customers of 23andMe for answering surveys and participating in this research. Our mission is to help people access, understand and benefit from the human genome. Trust in and engagement with our research platform is critical to that mission. We also thank all current and past employees of 23andMe, who together have made this research and this data-driven product possible.

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). The authors thank the staff and participants of the ARIC study for their important contributions. Funding for GENEVA was provided by National Human Genome Research Institute grant U01HG004402 (E. Boerwinkle).

References

Arnett, D. K., Blumenthal, R. S., Albert, M. A., Buroker, A. B., Goldberger, Z. D., Hahn, E. J., Himmelfarb, C. D., Khera, A., Lloyd-Jones, D., McEvoy, J. W., Michos, E. D., Miedema, M. D., Muñoz, D., Smith, S. C., Virani, S. S., Williams, K. A., Yeboah, J., & Ziaeian, B. (2019). 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*, *140*(11), e596–e646. https://doi.org/10.1161/CIR.0000000000000678https://doi.org/10.1038/ng.608

Campbell, C. L., Furlotte, N. A., Eriksson, N., Hinds, D., & Auton, A. (2015). Escape from crossover interference increases with maternal age. *Nature Communications*, *6*, 6260. https://doi.org/10.1038/ncomms7260

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015).

Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7. https://doi.org/10.1186/s13742-015-0047-8

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, *9*(3), e1003348. https://doi.org/10.1371/journal.pgen.1003348

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., & Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, *10*(1), 3328. https://doi.org/10.1038/s41467-019-11112-0

Furlotte, N., Kleinman, A., Smith, R., & Hinds, D. A. (2015). *23andMe White Paper 23-12: Estimating Complex Phenotype Prevalence Using Predictive Models*. 23andMe. https://permalinks.23andme.com/pdf/23-12_predictivemodel_methodology_02oct2015.pdf

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., & Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, *81*(24), 1879–1886. https://doi.org/10.1093/jnci/81.24.1879

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., & Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, *50*(9), 1219–1224. https://doi.org/10.1038/s41588-018-0183-z

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., Daly, M. J., Bustamante, C. D., & Kenny, E. E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *American Journal of Human Genetics*, *100*(4), 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004

Multhaup, M., Kita, R., Krock, B., Eriksson, N., Fontanillas, P., Asilbekyan, S., Del Gobbo, L., Shelton,

J., Tennen, R., Lehman, A., Furlotte, N., & Koelsch, B. (2019). *23andMe White paper 23-19: The science behind 23andMe's Type 2 Diabetes report*. 23andMe. https://permalinks.23andme.com/pdf/23_19-Type2Diabetes_March2019.pdf

Munafò, M. R., & Flint, J. (2004). Meta-analysis of genetic association studies. *Trends in Genetics*, *20*(9), 439–444. https://doi.org/10.1016/j.tig.2004.06.014

Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, & Dubourg. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164. https://doi.org/10.1038/538161a

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

So, H.-C., & Sham, P. C. (2017). Exploring the predictive power of polygenic scores derived from genome-wide association studies: A study of 10 complex traits. *Bioinformatics (Oxford, England)*, *33*(6), 886–892. https://doi.org/10.1093/bioinformatics/btw745

Tian, C., Hromatka, B. S., Kiefer, A. K., Eriksson, N., Noble, S. M., Tung, J. Y., & Hinds, D. A. (2017). Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nature Communications*, *8*(1), 599. https://doi.org/10.1038/s41467-017-00257-5

Tung, J. Y., Do, C. B., Hinds, D. A., Kiefer, A. K., Macpherson, J. M., Chowdry, A. B., Francke, U., Naughton, B. T., Mountain, J. L., Wojcicki, A., & Eriksson, N. (2011). Efficient replication of over 180 genetic associations with self-reported medical data. *PloS One*, *6*(8), e23473.

https://doi.org/10.1371/journal.pone.0023473

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., Beckmann, J. S., Bragg-Gresham, J. L., Chang, H.-Y., Demirkan, A., Den Hertog, H. M., Do, R., Donnelly, L. A., Ehret, G. B., Esko, T., … Global Lipids Genetics Consortium. (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, *45*(11), 1274–1283. https://doi.org/10.1038/ng.2797

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., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. https://doi.org/10.1038/ng.608

Change Log

- This document was updated in November, 2022 to reflect updates in the methods used to present low likelihood estimate results, as well as minor updates to the general methods to be consistent with parameters used for the majority of reports developed between 2020 and 2022.
- This document was updated in August, 2023 to reflect updated methods whereby genetic ancestry is used to interpret polygenic score results instead of self-reported ancestry.
- This document was updated in September, 2023 to include validation of select PGS using data from the ARIC study.