White Paper 23-25

A Generalized Method for the Creation and Evaluation of Polygenic Scores: Accounting for Genetic Ancestry as a Continuum

James R. Ashenhurst*, Sarah B. Laskey*, Jianan Zhan*, Aditya Ambati, Rafaela Bagur Quetglas, Anna Guan, Jingran Wen, Peter Wilton, Iris Yang, Bo Yoo, Sahil Hansalia, Yashashree Kokje, Ishana Raghuram, Akele Reed, Marcos Torres, 23andMe Research Team, Subarna Sinha, Theodore M. Wong, Bertram L. Koelsch

*These authors contributed equally to this work

Report-specific details and appendices

Breast Cancer (2024)

Colorectal Cancer (2024)

Prostate Cancer (2024)

Introduction

Polygenic scores (PGS) strive to estimate the heritable portion of risk for many common 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; many of these variants are commonly single nucleotide polymorphisms (SNPs). 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 (often non-European) groups, or by developing models using data from various populations themselves (Duncan et al., 2019). With over 14 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, Ashenhurst et al., 2020).

Here we detail the modeling methodologies and evaluation procedures used to create the PGS behind new 23andMe Health Predisposition and Wellness reports on common health conditions (Figure 1). The Appendices to this White Paper further summarize the performance and characteristics of each PGS used in recently released reports (starting in 2024). 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 males than females or more common in older than younger individuals should show these trends in both the 23andMe research participant population and in other cohorts.

Lastly, if there are well-established correlates or predictors of the phenotype and survey questions about these correlates are available in the 23andMe database, we may 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 2 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.

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 97.95% are discarded.

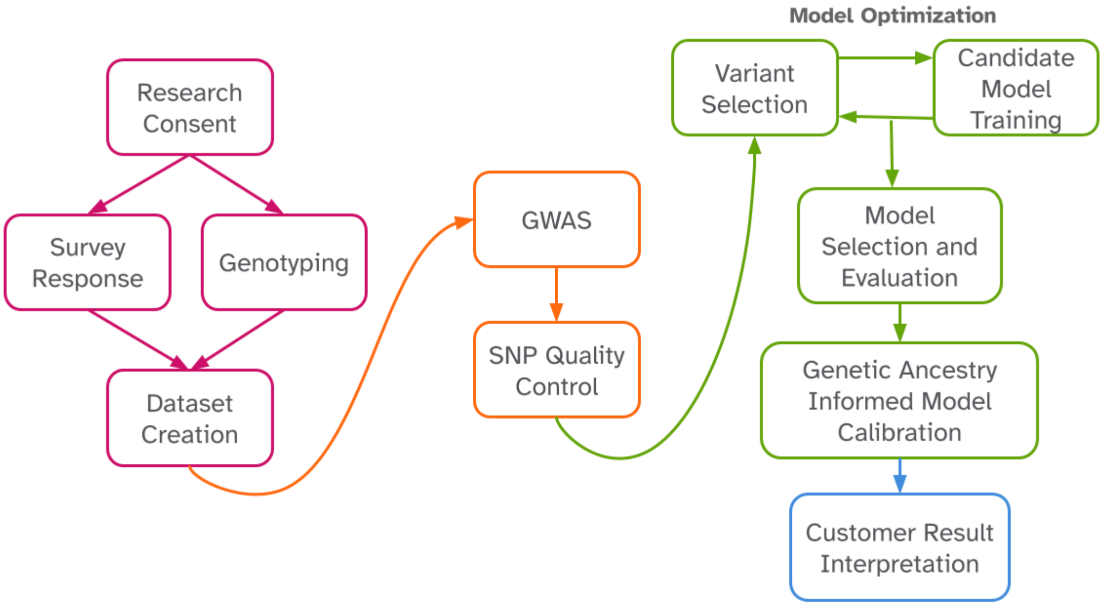

Figure 1. Outline of 23andMe's PGS creation procedure from self-report of survey data to generation of the polygenic model powering a customer's result.

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 male and female 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 include only variants genotyped on the V5 array. 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 completing the GWAS, variants are filtered to exclude those that do not pass GWAS quality control metrics: parent-offspring transmission, large sex effects, multiple reference sequence matches, significant genotyping date associations, genotype rate ≤ 0.95, concordance between imputed and genotyped dosages, 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(s). 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 genetic 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 imputed to population mean dosages. The variant set with the highest area under the receiver operator curve (AUROC) in a given validation set is designated as the optimal feature set for the corresponding cohort(s). Final fit statistics are obtained using the test set, which was held out of model training and selection. Variations in this approach are described in the phenotype-specific Appendices.

Model features

Features used in the model training typically include genetic 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 genetic PCs in the regression is to account for any residual population substructure. Absolute and relative risk estimates associated with a PGS and reported as customer results take into account both self-reported birth sex and genetic ancestry.

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 genome-wide PCs, dosages for each variant, and demographic factors are treated as independent variables. For binary phenotypes, we use multivariate logistic regression under a GLM framework. For quantitative phenotypes, we use linear regression. After a model is specified, weights for each feature are calculated through regression. We use PyTorch to train models, typically with the L-BFGS solver (for logistic regression models) and L2 regularization with default penalization strength.

Transferability across populations

One of the most important challenges for PGS is transferability across groups with systematically different genetic ancestry (Martin et al., 2017). Individuals of European descent make up the overwhelming majority of genomics research participants even though they represent a minority of global genetic diversity (Popejoy & Fullerton, 2016). PGS trained with data exclusively from individuals of European descent often perform worse among individuals of other ancestries. We leverage our large, diverse research participant population to address this challenge using a multi-ancestry cohort for each step in the PGS development process as is possible. All validation and testing are done in ancestry-specific datasets to avoid overestimation of performance metrics.

As illustrated in Figure 1, we run separate GWAS in each of the six genetic ancestry-defined cohorts described above with sufficient sample size for a given phenotype. We perform a fixed-effect meta analysis (Munafò & Flint, 2004) to combine those separate GWAS and use the resulting *p*-values to select genetic features. We then train one model per hyperparameter set in a multi-ancestry training cohort, controlling for population structure by including genetic principal components as covariates. When sample sizes permit, we evaluate each of the models produced by different hyperparameter sets in ancestry-specific validation sets and select the one that performs best for each ancestry cohort. For ancestry groups that lack sufficient sample size for separate validation and test sets, we assign the model with the best performance in the largest (i.e., European-descent) validation set.

Assessing model performance

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 (ORs) 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 and phenotype prevalences in each ancestry group. These and other detailed results are presented in phenotype-specific Appendices to this White Paper.

Absolute lifetime risk estimates

Customer report results include an estimate of the absolute risk of developing each phenotype by a target age (e.g., their 70's). The target age is typically chosen to be the decade at which the highest proportion of participants report ever having been diagnosed or treated with a condition, so that this result can be interpreted as an approximate absolute lifetime risk estimate.

After model training, the raw PGS is often miscalibrated in some customer subsets, particularly when phenotype prevalence varies systematically with demographic factors. To derive well-calibrated absolute risk estimates for diverse customers from the potentially miscalibrated raw PGS, we recalibrate results empirically using sex and genetic ancestry information. To do this, first we define the PGS for customer i as:

$$P\hat{G}S_i = \sum_j \hat{\beta}_j x_{ij}$$

where $\widehat{\beta}_j$ is the estimated weight for SNP j, and x_{ij} is customer i's dosage for SNP j. A customer's more comprehensive polygenic risk score (PRS), which includes the impact of both genetics and demographic factors, is defined as:

$$P\hat{R}S_i = \hat{\beta}_{age} age_i + \hat{\beta}_{sex} sex_i + ... + P\hat{G}S_i$$

where ". . ." refers to any higher-order and/or interacting demographic terms included in the model. Note that the sex term is excluded for models trained only on individuals of a single sex. If we denote the target age as age_t, then a customer's estimated disease risk at the target age (which serves as an uncalibrated estimate of their total lifetime risk, in units of log odds) is:

$$P\hat{R}S_{t,i} = \hat{\beta}_{age} age_t + \hat{\beta}_{sex} sex_i + ... + P\hat{G}S_i$$

To adjust for miscalibration in this estimate due to demographic confounding, we train a *recalibration logistic regression model* with the form:

$$y \sim sex + \sum_{k=1}^{5} PC_k * P\hat{R}S_t$$

where $y \in \{0, 1\}$ is the binary disease outcome, PC_k are the first five genetic principal components, and * denotes the interaction terms. Because we do not have sufficient samples across diverse genetic ancestry backgrounds to perform this recalibration step in the training or validation sets, the recalibration model is trained in the multi-ancestry test set. If we denote the fitted coefficients of this recalibration model as $\widehat{\alpha}$, then the log-odds of a customer's recalibrated PRS is:

$$\hat{y}_i = \sum_{i=1}^{5} \hat{\alpha}_{PC_k} PC_{k,i} + \sum_{i=1}^{5} \hat{\alpha}_{PC_k P\hat{R}S_t} PC_{k,i} \times P\hat{R}S_{t,i}$$
$$+ \hat{\alpha}_{sex} sex_i + \hat{\alpha}_{P\hat{R}S_t} P\hat{R}S_{t,i}$$

The absolute lifetime risk estimate result shown in the customer's report is computed by taking the inverse logit transformation of this value:

$$\hat{AR_i} = expit(\hat{y}_i) = \frac{e^{\hat{y}_i}}{1 + e^{\hat{y}_i}}$$

In most cases, the customer-facing lifetime absolute risk estimate is rounded to a whole percent; customer-facing results are given to one decimal place only in the case of especially low estimates.

These estimates should be interpreted in light of several limitations to this approach. First, for conditions linked to higher rates of mortality, prevalence at the chosen target age likely undercounts the true cumulative incidence of a condition. Rather, these estimates represent the likelihood of having the condition assuming survival to a particular age. While other modeling strategies, like competing risk 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 many examples of 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.

Relative risk estimates

Customers also receive a relative risk estimate indicating whether their result is associated with an increased or a typical — i.e., not increased — likelihood of developing the condition. Relative risks are by nature comparative, and we determined that the relevant comparison group for an individual's PGS-based risk is other people who share their basic demographic characteristics; otherwise, demographic factors often predominate in relative risk estimates. In other words, we determined that this result should answer the question, "is the customer's absolute lifetime risk estimate increased or not, relative to what it would have been without knowing their PGS?" We train the following *reference linear regression model*:

$$\hat{y} \sim sex + \sum_{k=1}^{5} PC_k$$

where $\hat{y}$ is the recalibrated PRS derived above, and PC_k are the first five genetic principal components. The cohort used to train this model is downsampled for over-represented and un-admixed populations, so that it emphasizes greater diversity and genetic admixture. As with the model used to generate absolute risk estimates, this model is trained in a cohort of test set individuals. Denoting the fitted coefficients of this reference model as $\hat{\gamma}$, we use the model to compute the average PGS of a theoretical reference group with the same sex and genetic ancestry as customer *i*:

$$\tilde{y}_i = \hat{\gamma}_{sex} sex_i + \sum_{k=1}^{5} \hat{\gamma}_{PC_k} PC_k$$

We then calculate the odds ratio between the customer's calibrated absolute risk estimate and this reference PGS value:

$$\hat{OR}_i = e^{\hat{y}_i - \tilde{y}_i}$$

If this odds ratio is greater than some threshold δ (typically ≥ 1.5), the customer receives an "increased likelihood" relative risk result.

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 relative risk result threshold.

In order to estimate this uncertainty, we use a metric similar to a z-score that includes information about missing SNPs' effect sizes ($\hat{\beta}$), effect allele frequencies (*p)*, and the threshold for an individual customer between a "typical likelihood" and an "increased likelihood" relative risk result, defined as:

$$\Delta_i = log(\delta) + \tilde{y}_i$$

where δ is the odds ratio threshold described in the previous section.

For each missing genotype call *m* across *M* missing calls, we compute the ratio between the distance of an individual's score from the threshold Δ_i and the uncertainty in the score due to missing values.

$$\frac{\Delta_i - \hat{y}_i}{\hat{\alpha}_{P\hat{R}S_t \sqrt{\sum_{m=1}^{M} 2\hat{\beta}_m^2 * p_m (1-p_m)}}}$$

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 relative risk 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 with available data in external datasets. Specifically, we sought to understand how well these models can both stratify risk and provide accurate risk assessment outside of 23andMe research cohorts.

Validation in the UK Biobank

We conducted validation analyses on the Caucasian cohort ("Caucasian" for the 22006 Data Field, Genetic Ethnic Grouping) within the UK Biobank study. The UK Biobank is a vast biomedical database and research resource, housing comprehensive genetic and health data from a population of half a million UK participants. These participants were all between the ages of 40 and 69 at the time of recruitment, which occurred between 2006 and 2010. The database encompasses a wide array of information, including blood samples, cardiac and cerebral scans, genetic profiles, and lifestyle details. For our validation work (under Application Number 95801), we computed PGS using our models, aligned UK Biobank phenotypes as closely as possible to those used in training our PGS models, and assessed the predictive performance of our PGS models and the calibration of our absolute lifetime risk estimates for these phenotypes. To ensure data quality, we applied filters that excluded samples where the reported sex did not match genetic sex and samples with sex chromosome aneuploidy.

Phenotype Selection

The phenotypes available in the UK Biobank dataset are not exactly the same as those collected by 23andMe. For these assessments, we attempted to harmonize UK Biobank and 23andMe phenotypes as closely as possible. Cases and controls were defined using self-reported, clinical examination, and biomarker data when available.

PGS Calculation

PGS were calculated by summing the weighted SNP dosages from the UK Biobank, using the PGS weights from 23andMe. Overlapping SNPs were used between the 23andMe PGSs and the UK Biobank's WTCHG imputation panel. The alignment of SNPs between the PGS and the UK Biobank data was performed based on chromosome number and position, for the accurate identification and matching of SNPs between the different datasets. Original UK Biobank SNP information was based on Genome Reference Consortium human genome build 37 (GRCh37) and was converted to genome build 38 (GRCh38) before matching. To ensure consistent interpretation of 23andMe PGS across the two datasets, we inverted the beta estimates for any SNPs with differing coding alleles between the UK Biobank and 23andMe PGS.

PGS Performance Validation

For each PGS we assessed model performance in the UK Biobank cohort using AUROC and odds ratios, and we assessed absolute risk estimates using calibration plots showing actual prevalence versus predicted prevalence for each 20% percentile of PGS score. To compute absolute lifetime risk estimates for UK Biobank participants from the 23andMe PGS, we set the y-intercept based on the measured prevalence of the condition in the UK Biobank cohort, and then scaled zero-centered PGS using the PGS recalibration model coefficients. For the interaction terms between genetic principle components and the PGSs in the recalibration model, we set the genetic principle component values to the mean values of a cohort of 23andMe samples whose “British & Irish” Ancestry Composition is greater than 90%. Relatedly and as a reminder, this analysis only uses "Caucasian" individuals from UK Biobank, as described above.

1: Breast cancer PRS validation in UK Biobank

Phenotype definition

The process of identifying breast cancer cases and controls relied on two UKBB concepts: 'Diagnoses - ICD10' (Data-Field 41270) and 'Cancer code, self-reported' (Data-Field 20001). For the ICD10 phenotype, any sample with a code of C50 is defined as a case, and as a control otherwise. Regarding the self-reported cancer code, anyone who explicitly specified breast cancer in the data field (cancer code 1002, breast cancer) was categorized as a case, with all four self-reported instances combined, meaning that an individual was classified as a case if at least one of the instances indicated a case. These two concepts were subsequently integrated, so that individuals identified as cases in either ICD10 or self-reported data were categorized as breast cancer cases, while the remaining participants were classified as controls. Individuals not genetically identified as female (Data-Field 22001) were excluded from the analysis.

Table 1-1. UK Biobank cohort description for breast cancer

Cohort	N	Mean age (SD)	Prevalence (%)
UK Biobank Caucasian	198,765	72.1 (8.0)	9.26%
23andMe European (test set)	185,129	49.0 (16.5)	3.51%

Table 1-2. Breast cancer PGS performance in UK Biobank

		Odds Ratios (OR)	
Cohort	Genetics Only AUC (95%CIs)	Top 5% versus average	Top 5% versus bottom 5%
UK Biobank Caucasian	0.6376 (0.6338 to 0.6414)	3.04 (2.79 to 3.31)	7.87 (6.98 to 8.86)
23andMe European (test set)	0.6284 (0.6221 to 0.6347)	2.60 (2.37 to 2.86)	6.60 (5.43 to 8.05)

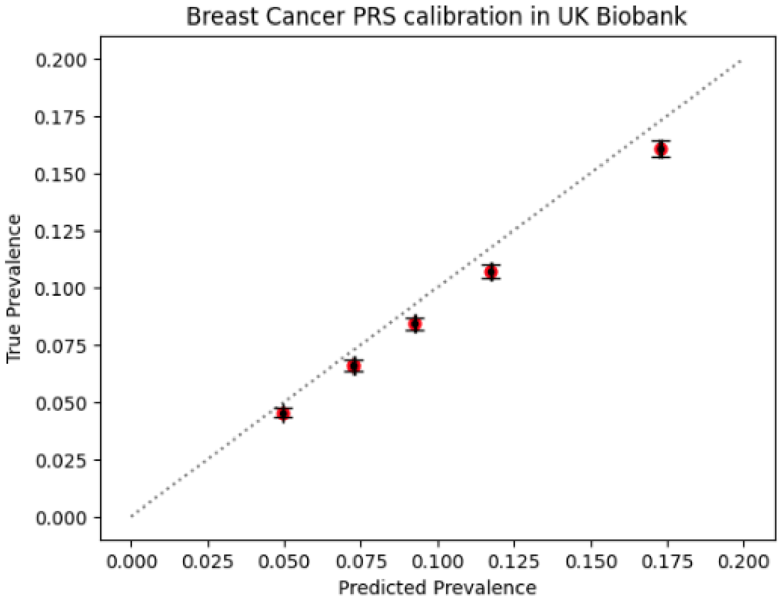

Figure 1-1. Calibration of breast cancer PGS in UK Biobank cohort across quintiles of the PGS distribution. Error bars represent 95% CIs.

2: Prostate cancer PRS validation in UK Biobank

Phenotype definition

The process of identifying prostate cancer cases and controls relied on two UKBB concepts: 'Diagnoses - ICD10' (Data-Field 41270) and 'Cancer code, self-reported' (Data-Field 20001). For the ICD10 phenotype, any sample with a code of C61 is defined as a case, and as a control otherwise. Regarding the self-reported cancer code, anyone who explicitly specified prostate cancer in the data field (cancer code 1044, prostate cancer) was categorized as a case, with for all four self-reported instances combined, meaning that an individual was classified as a case if at least one of the instances indicated a case. These two concepts were subsequently integrated, so that individuals identified as cases in either ICD10 or self-reported data were categorized as prostate cancer cases, while the remaining participants were classified as controls. Individuals not genetically identified as male (Data-Field 22001) were excluded from the analysis.

Table 2-1. UK Biobank cohort description for prostate cancer

Cohort	N	Mean age (SD)	Prevalence (%)
UK Biobank Caucasian	166,700	72.8 (8.0)	8.20%
23andMe European (test set)	133,363	50.4 (16.7)	2.49%

Table 2-2. Prostate cancer PGS performance in UK Biobank cohort

		Odds Ratios (OR)	
Cohort	**Genetics Only AUC (95%CIs)**	**Top 5% versus average**	**Top 5% versus bottom 5%**
UK Biobank Caucasian	0.6575 (0.6532 to 0.6618)	3.50 (3.16 to 3.88)	11.51 (9.83 to 13.49)
23andMe European (test set)	0.6783 (0.6703 to 0.6863)	3.7 (3.25 to 4.18)	18.3 (12.46 to 26.79

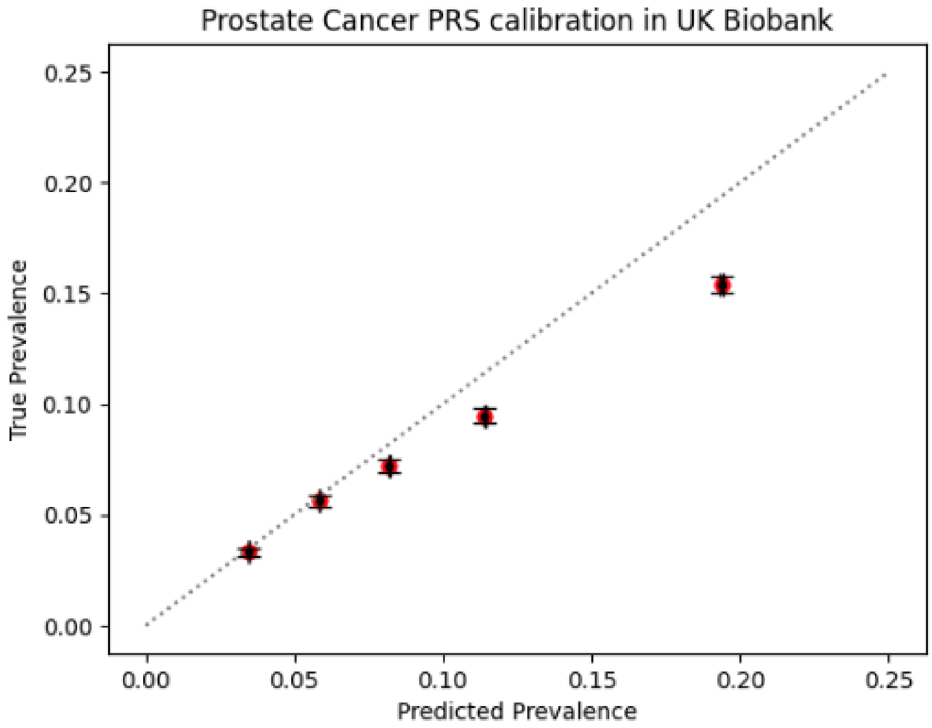

Figure 2-1. Calibration of prostate cancer PGS in UK Biobank cohort across quintiles of the PGS distribution. Error bars represent 95% CIs.

3. Colorectal cancer PRS validation in UK Biobank

Phenotype definition

The process of identifying colorectal cancer cases and controls relied on two UKBB concepts: 'Diagnoses - ICD10' (Data-Field 41270) and 'Cancer code, self-reported' (Data-Field 20001). For the ICD10 phenotype, any sample with a code of C18 or C19 or C20 is defined as a case, and as a control otherwise. Regarding the self-reported cancer code, anyone who explicitly specified colorectal cancer in the data field (cancer code 1020, large bowel cancer/colorectal cancer; cancer code 1022 colon cancer/sigmoid cancer; cancer code 1023, rectal cancer) was categorized as a case, with for all four self-reported instances combined, meaning that an individual was classified as a case if at least one of the instances indicated a case. These two concepts were subsequently integrated, so that individuals identified as cases in either ICD10 or self-reported data were categorized as colorectal cancer cases, while the remaining participants were classified as controls.

Table 3-1. UK Biobank cohort description for colorectal cancer

Cohort	N	Mean age (SD)	Sex (% female)	Prevalence (%)
UK Biobank Caucasian	365,465	72.4 (8.0)	54.39%	2.52%
23andMe European (test set	281,599	49.5 (16.5)	59.5%	0.44%

Table 3-2. Colorectal Cancer PGS performance in UK Biobank cohort

		Odds Ratios (OR) and 95% CIs	
Cohort	**Genetics Only AUC (95%CIs)**	**Top 5% versus average**	**Top 5% versus bottom 5%**
UK Biobank Caucasian	0.5592 (0.5535 to 0.5648)	1.56 (1.38 to 1.76)	2.66 (2.31 to 3.07)
23andMe European (test set	0.5711 (0.556 to 0.5862)	1.5 (1.2 to 1.95)	2.2 (1.51 to 3.14)

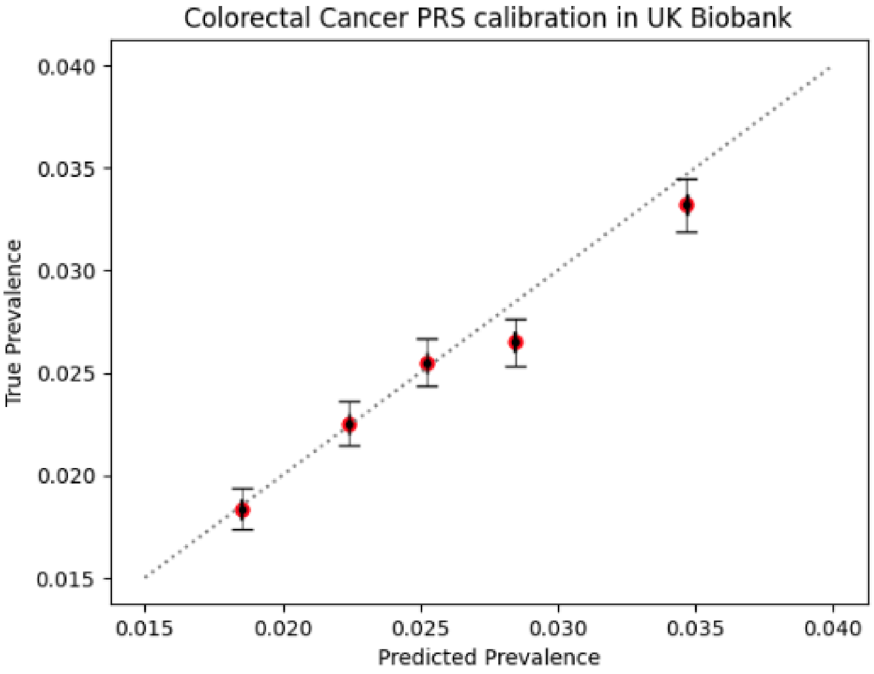

Figure 3-1. Calibration of colorectal cancer PGS in UK Biobank cohort across quintiles of the PGS distribution. Error bars represent 95% CIs.

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 the 23andMe Product Science and Genomic Health teams, along with all current and past employees of 23andMe who have made this research and this data-driven product possible. The PGS validation part of the work has been conducted using the UK Biobank Resource under Application Number 95801.

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

Ashenhurst, J. R., Zhan, J., Multhaup, M. L., Kita, R., Sazonova, O. V., Krock, B., et al. (2020). A Generalized Method for the Creation and Evaluation of Polygenic Scores. 23andMe, Inc. Available at: https://permalinks.23andme.com/pdf/23_21-PRSMethodology_May2020.pdf

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

Evangelou, E., Warren, H. R., Mosen-Ansorena, D., Mifsud, B., Pazoki, R., Gao, H., Ntritsos, G., Dimou, N., Cabrera, C. P., Karaman, I., Ng, F. L., Evangelou, M., Witkowska, K., Tzanis, E., Hellwege, J. N., Giri, A., Velez Edwards, D. R., Sun, Y. V., Cho, K., … Caulfield, M. J. (2018). Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nature Genetics, 50(12), 1755. https://doi.org/10.1038/s41588-018-0297-3

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

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

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., Bastarache, L. A., Wei, W.-Q., Denny, J. C., Lin, M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., & Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature Genetics, 50(9), 1335–1341. https://doi.org/10.1038/s41588-018-0184-y