Within-family validation of polygenic risk scores in the UK Biobank: Investigating the role of principal component analysis and mixed model association

Ciaran Michael Kelly[1*], Onyedika Onuorah[1], Edmund Gilbert[1,2]

[1]School of Pharmacy and Biomolecular Sciences, RCSI University of Medicine and Health Sciences, Dublin, Ireland

[2]FutureNeuro SFI Research Centre, Royal College of Surgeons in Ireland, Dublin, Ireland

* Corresponding author

E-mail: ciarankelly@rcsi.com (CMK)

Abstract

A central challenge in the field of polygenic risk prediction has been measuring and controlling for confounding effects mediated through population stratification. Traditionally, control has been attempted through the inclusion of the top principal components (PCs) of variation and the use of linear mixed models in genome-wide association studies (GWAS). Calculating the genomic inflation factor (λ) and assessing the predictive ability of polygenic risk scores (PRS) in within-family settings have been considered important steps in validating that GWAS summary statistics are relatively unbiased estimates of causal effects.

In this study, we examined the relationship between the number of PCs included during GWAS/PRS model development and the observed attenuation in performance when moving from a population-level setting to a cohort composed of discordant sibling pairs. This was done for four complex diseases in the UK Biobank: coronary artery disease, type 2 diabetes, breast cancer, and prostate cancer. Comparisons were made

with educational attainment, a trait well-known to be highly prone to environmental confounding.

We find that, contrary to expectation, increasing the number of included PCs does not consistently reduce within-family attenuation. Furthermore, any reduction in predictive attenuation does not appear to closely correlate with reduction in the genomic inflation factor λ. Additionally, there appears to be little added benefit to prediction or reduced attenuation with the addition of mixed model-based approaches. Taken together, our results demonstrate that the effects of controlling for population stratification can be trait-specific and can vary substantially depending on the number of SNPs or PCs included in the PRS model. These results highlight the inherent complexity of controlling for population stratification in the UK Biobank and suggest that further research is needed to establish best practices for both detecting and effectively minimizing confounding during polygenic prediction tasks.

Introduction

Genome-wide association studies (GWAS) have been performed on many human complex traits and diseases to associate single-nucleotide polymorphisms (SNPs) with phenotypes of interest [37]. The summary statistics of a GWAS include an effect size and P-value for each examined SNP. At an individual level, these SNP effects can be summed by allele dosage and an overall polygenic risk score (PRS) calculated [8]. The potential ability of a PRS to explain a large part of population-level variation of human disorders offers an exciting opportunity for personalized medicine. For example, a polygenic score developed for coronary artery disease (CAD) identifies individuals in the tail end of the risk score distribution that are at a three-fold higher risk of CAD than the general population, where this level of increased risk is comparable to that observed in individuals with monogenic forms of CAD [19]. Thus, those individuals at high risk may be offered prophylactic treatment options, such as statins in the case of CAD, or preventative lifestyle interventions for other conditions such as Type 2 Diabetes (T2D) [23, 35]. Even in the cases where individual-level prediction is not accurate enough to justify prophylactic treatment, PRS could be an efficient and cost-effective population-level tool for choosing earlier ages of screening for disease, such as in the

case of breast and prostate cancer. Therefore, there is much interest in the use of polygenic scores as a central tool in personalized medicine.

One of the major challenges when associating a SNP with any complex phenotype is that of confounding, generally mediated through population structure in the data. Ideally, the SNP effect size estimates included in a PRS would be reflective of causal genetic effects, yet they may instead capture environmental or genetic background effects [2, 36]. Dealing with these unwanted effects of population structure has been a central task in the genomic prediction literature, with debate still remaining on how best to measure and minimize confounding [1, 26, 34, 43].

Perhaps the most common method that has been employed for dealing with population structure has been the inclusion of principal components (PCs) of variation as covariates as part of a GWAS. Principal component analysis (PCA) is a linear dimensionality reduction technique that generates vectors which capture the most variation in a dataset, with each sequential PC explaining less variation than the previous. It has been found that the first PCs of a genomic SNP dataset often capture broad ancestry differences in the population, which can be used to control for confounding [1, 24, 26]. In the context of the UK Biobank, it has previously been suggested that the first 16 PCs are generally sufficient to capture the population structure in this specific cohort [6, 29]. These PCs have been used for several GWAS and PRS conducted on this dataset in the wider literature [11, 13, 28].

The addition of mixed model approaches to GWAS has now become common [27, 41]. These methods employ a linear mixed model (LMM) and are commonly referred to as mixed-linear models of association (MLMAs). In this procedure, a genetic relatedness matrix (GRM) is constructed and is used to model the covariance structure of the random effect in the LMM. MLMA approaches are thought to better control for population stratification and cryptic relatedness than PC-adjustment alone. Extensions appropriate for binary phenotypes through the use of generalized linear mixed models (GLMMs) have also been developed [16].

Determining the success of methods that attempt to control for confounding in GWAS is challenging. The genomic inflation factor (λ) is a measure of global P-value inflation in the summary statistics outputted from a GWAS [9, 31, 40, 44]. When λ is significantly above one (i.e. P-values are systematically lower than the expectation

under the null) this suggests that population stratification may not have been successfully corrected for. Despite many such control techniques, some authors argue that the current methods may not yet be sufficient to truly correct for all the confounding present in GWAS effect size estimates [5, 42, 43]. As discussed above, much debate has surrounded the causal nature of the genetic effects estimated from GWAS and included in PRS models [2, 17, 36, 38]. This has been compounded by the fact that PRS models are often constructed with SNPs which fall well below the canonical genome-wide significance threshold, plausibly increasing the risk that genetic background and environmental effects will be exploited during prediction [8, 15].

One method that has been proposed as critical to validating the causal nature of SNP effects in PRS has been the within-family study design. Genetic background and environmental effects tagged by population structure are thought to be minimized in this framework [14, 22, 32]. The extent to which PRS can predict between family members such as siblings as to who will develop disease or not is likely reflective of the extent to which the PRS is picking up signal from truly biologically causal SNP effects.

The approach taken in this paper follows from the work of Lello et al. who have previously developed polygenic risk scores from the UK Biobank and validated them in a within-sibling cohort [22]. The authors found that polygenic predictive power mostly persists when moving from a population-level cohort to a within-family setting. However, the authors did not make use of principal components or mixed models in the development of their PRS on the basis that their top PCs do not explain a substantial amount of phenotypic variation in the context of human height [21].

Building on this previous work, our study systematically evaluates the extent to which including principal components and using mixed models during GWAS and PRS construction affects within-sibling predictive attenuation across multiple complex traits. In doing so, we aimed to assess whether these commonly used methods for controlling stratification appear to meaningfully impact the causal validity of PRS.

Materials and methods

Study approach

In this study, we aimed to examine the effects of controlling for population structure through PC-adjustment and mixed linear modeling on preventing attenuation when moving polygenic prediction from a population-level to within-family setting.

Using the 2018 UK Biobank genotype release, we performed several GWAS on four disease traits for which PRS holds promise in preventative medicine [6, 33]. As a positive control, we compared the effects of controlling for population stratification on a binarized version of the educational attainment phenotype which has been well-established as being highly susceptible to confounding [20, 22, 32].

Summary statistics from each GWAS were used to build PRS models used to classify between discordant pairs (consisting of one case and one control in each pair). Two test sets were evaluated, one consisting of sibling pairs and the other of unrelated pairs (non-siblings). To account for the effects of different levels of polygenicity, three SNP-set sizes were examined during PRS construction (1K, 10K, and 100K).

The attenuation in predictive performance between the population-level test set and the within-family test set was the central focus of this study.

Phenotype selection

Phenotype definitions followed those used in Lello et. al. (2020) with the exception of educational attainment which was converted to a pseudo-binary phenotype in order to facilitate comparisons in PRS performance and attenuation [22]. Dichotomizing educational attainment in this manner sacrifices information and interpretability and we employ it only to benchmark comparisons with the binary disease outcomes in terms of confounding. The precise encoding of each phenotype can be found in Supplementary Note 1. The sample size of each phenotype can be seen in Table 1.

Table 1. Sample sizes in training GWAS for positive and negative phenotype classes in the UK Biobank.

Phenotype	Positive	Negative
Coronary Artery Disease	45,810	367,385
Type 2 Diabetes	30,936	385,931
Breast Cancer	9,548	217,739
Prostate Cancer	4,075	190,692
Educational Attainment	64,941	347,166

Discordant pair extraction

Siblings

Siblings were defined as those pairs which have an IBS0 >0.0012 and kinship coefficient >0.176. For each phenotype, all sibling pairs in which one individual was a case and the other a control were identified. Pairs were kept single-sex for prostate and breast cancer. Trios and higher-order relationships were not included in the analysis i.e. only the first pair were kept when more than one sibling pair was identified within a single family unit. The final sample sizes of the discordant sibling test sets can be seen in Table 2.

Table 2. Number of individuals in discordant sibling test sets from the UK Biobank. Additional test sets made up of equal number of unrelated pairs were generated for each phenotype.

Phenotype	Number of Siblings
Coronary Artery Disease	7,512
Type 2 Diabetes	5,024
Breast Cancer	1,196
Prostate Cancer	294
Educational Attainment	8,280

Non-siblings

After the creation of the discordant sibling test sets as described above, a KING relatedness threshold of ≥ 0.0884 was implemented to exclude closely related individuals from the post-QC dataset. Random case-control pairs were selected from the unrelated cohort, with sample sizes matched to the corresponding sibling test sets (see Table 2). For prostate and breast cancer, pairs were restricted to individuals of the same sex.

To match the age difference distribution observed in the sibling pairs, candidate non-sibling pairs were first stratified into equivalent age difference bins. Proportional sampling was then performed within these bins to replicate the sibling age difference

distribution. The resulting set was trimmed or supplemented to match the exact number of sibling pairs, thus ensuring balanced age distributions and sample sizes between groups.

Principal component analysis

Principal components were taken from UK Biobank data field 22009 and individuals with missing PC information were excluded (n=14,227). The first 16 principal components have previously been stated to be sufficient to capture population structure in the UK Biobank [29].

All UK-Biobank participants were plotted using their first 16 PCs. Individuals were coloured according to their country or region of origin. Country of birth information were taken from UK Biobank data fields 1647 and 20115. Individuals with missing or unassigned data were excluded. A maximum of 1,000 randomly sampled individuals per country were kept for ease of visualization purposes. For the global plot, all non-European countries were collapsed into four continental bins —Africa, Asia, Americas, Oceania, while the UK and the four core European UN Statistical Division (UNSD) subregions were kept as distinct categories. For the regional plot, only people born in the UK constituent countries and the Republic of Ireland were included.

Population selection and genotype quality control

As per Lello et al., all individuals with ethnic background codes 1, 1001, 1002, or 1003 (i.e. self-reported "White" individuals) were included in the GWAS and PRS analyses [22]. The following quality control measures were implemented using PLINK1.9: minor allele frequency (MAF) threshold of 0.1%, SNP missingness filter of 3%, and individual missing filter of 3%. The extended human leukocyte antigen (HLA) region and sex chromosomes were also excluded [30].

Association methods

Standard GWAS (GLM)

PLINK2's `--glm` function was used to perform basic GWAS across the phenotypes [7]. To investigate the effect of including varying numbers of principal components of

genetic variation - the top M number of PCs were included as covariates during association. The discordant test sets made up of sibling and non-sibling pairs were excluded from the training GWAS.

Mixed model GWAS

GCTA software was used to test for the added effect of performing a mixed model-based approach to controlling for population stratification [39]. The post-QC SNPs were pruned using PLINK1.9 for construction of the GRMs with the following parameters: window-size of 100kb, step-size of 5bp and squared correlation threshold of 0.5 [30]. The HLA region was excluded during pruning. GRMs were constructed using GCTA version 1.92.3 on the resultant set of 111,226 SNPs. A sparsity threshold of 0.05 was then applied to the GRM. A GCTA 1.94.4 generalized linear mixed model association analysis (fastGWA-GLMM) was then performed on each trait with the full set of 16 principal components of variation included as covariates [16].

Clumping

SNPs were clumped using PLINK1.9 [30]. The LD threshold for clumping was set to 0.1, the distance threshold to 500kb, and the index SNP significance threshold to 0.5.

Polygenic score construction

Polygenic scores were constructed for the discordant test set made up of sibling and non-sibling pairs using the top N clumped SNPs from the `--score` function of PLINK2 [7]. Raw scores were regressed on the top M principal components of genetic variation. The resulting residuals were used as the predictor variable for the phenotype of interest.

Classification

The outcome variable for classification of discordant pairs, denoted A, was defined as the proportion of pairs in which the positive-class individual had the higher polygenic score (with ties excluded) citeLello2020.

This metric is conceptually similar to the area under the ROC curve (AUC) which

represents the probability of correctly assigning a higher score to the case in a randomly selected discordant pair [12]. A major difference between the AUC and the classification metric used here is that the AUC represents an overall probability from all possible randomly selected discordant pairs in the population. The metric used here is based on a specific pre-defined and non-random set of pairs composed of the siblings and non-siblings.

Attenuation in performance was defined as the proportional reduction in above-chance classification accuracy when moving from unrelated pairs to siblings. An attenuation of 0% indicates no loss i.e. the PRS performed equally well in siblings and non-siblings. 100% attenuation corresponds to a complete loss of predictive power in the within-family setting. Negative attenuation values arose in the rarer instance where there was a greater predictive power in the sibling pairs as compared to the non-sibling pairs.

$$\text{Attenuation} = \left(1 - \frac{A_{\text{Siblings}} - 50}{A_{\text{Non-siblings}} - 50}\right) \times 100$$

Statistical analysis

A generalized linear mixed model (GLMM) was fit to investigate the overall effects of pair type, adjustment method, phenotype, and SNP-set size on classification accuracy. Three adjustment methods were compared: 0-PC, 16-PC, and GLMM+16-PC. An interaction term between method and pair type was included to assess whether the performance of the methods differed between sibling and non-sibling pairs. To account for repeated use of the same case-control pairs, a random intercept was included for each pair. The lme4 package was used to implement this GLMM using the BOBYQA optimizer [3, 25].

Results

Principal component analysis

An analysis of the first 16 principal components derived from UK Biobank SNP data revealed clear evidence of both global and intra-European population structure.

Specifically, PCs such as PC1, PC6, and PC11 captured broad ancestry differences, while PC4, PC7, and PC8 reflected finer-scale structure within Europe (Figure 1).

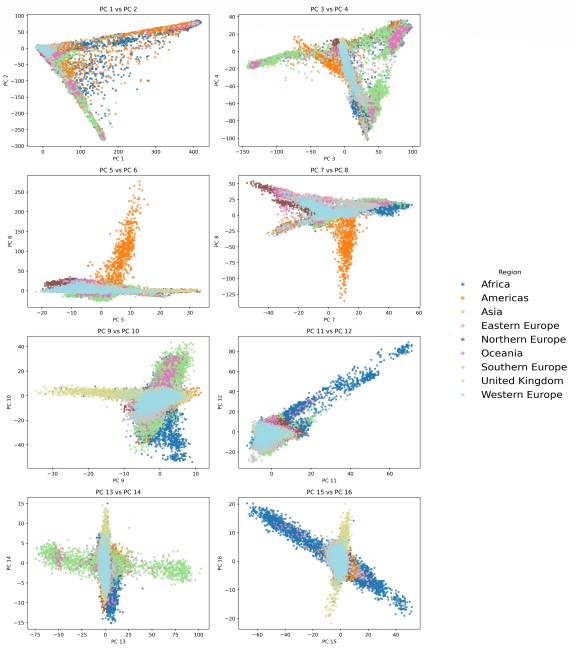


Fig 1. PC Scores 1–16 of UK Biobank Participants
PC scores were taken from Data Field 22009 of UK Biobank. A maximum of 1,000 individuals were included per country. Individuals not assigned to any country or population were removed.

When the analysis was restricted to participants born in the UK and Ireland, additional regional clustering became apparent. Whilst the first four principal components do not appear to pull out specific sub-regions, PCs 5, 9, 11, and 14 showed distinct separation among the constituent countries of the UK and Ireland (Fig. 2).

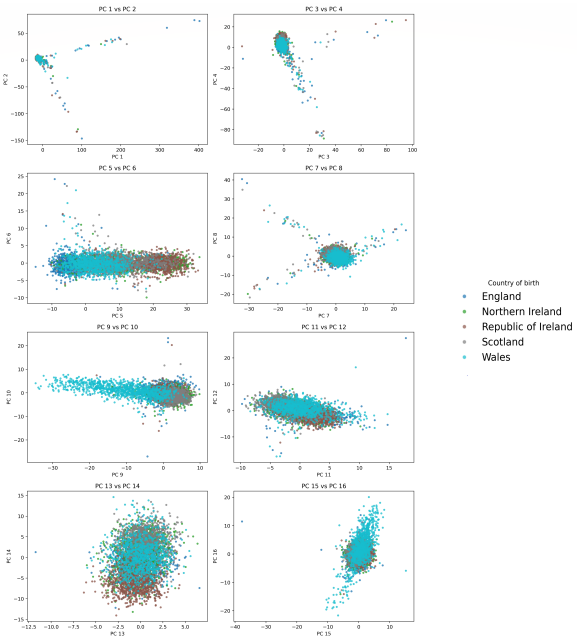


Fig 2. PC Scores 1–16 of UK Biobank Participants (UK and Ireland Only)
PC scores were taken from Data Field 22009 of UK Biobank. A maximum of 1,000 individuals were included per country.

PRS performance in population-level and within-family test sets

The performance of the standard GLM-based PRS models on the population-level and within-family test sets can be seen in Table 3. At a population-level, the greatest discrimination between cases and controls was achieved on type 2 diabetes and the constructed binary educational attainment phenotype. There was not a consistent increase in predictive power when increasing the number of SNPs in the GLM-based model, with the exception of educational attainment. In the case of prostate cancer, the most predictive PRS had only 1,000 top SNPs included.

As expected, classification accuracy generally decreased when moving from the non-sibling test set to the test set composed of discordant sibling pairs.This reduction in predictive performance between population-level and within-family pairs is also evident in Figures 3 and 7. Part of this decline can be attributed to the shared genetic background among siblings, which inherently limits the model's ability to discriminate between cases and controls. Additionally, the within-family design mitigates confounding due to population stratification, which may account for further attenuation. The most pronounced drop in classification accuracy was observed for educational

Table 3. Classification accuracy for standard GWAS-based PRS performance on sibling and non-sibling (population-level) discordant pairs in the UK Biobank. The full 16 PCs were included in both the GLM-based GWAS and PRS.

Phenotype	SNP-set size	**Classification accuracy (%)**	
		Non-sibling pairs	Sibling pairs
Coronary Artery Disease	1K	57.77	54.98
	10K	57.96	55.62
	100K	57.67	53.89
Type 2 Diabetes	1K	62.54	59.75
	10K	62.98	60.03
	100K	64.53	58.80
Breast Cancer	1K	56.35	54.35
	10K	56.35	55.02
	100K	53.18	53.34
Prostate Cancer	1K	61.22	61.90
	10K	56.46	59.18
	100K	58.50	53.74
Educational Attainment	1K	57.51	53.70
	10K	60.58	54.20
	100K	62.80	55.80

attainment, aligning with prior literature highlighting the high degree of environmental and stratification-related confounding associated with this trait [20, 22, 32].

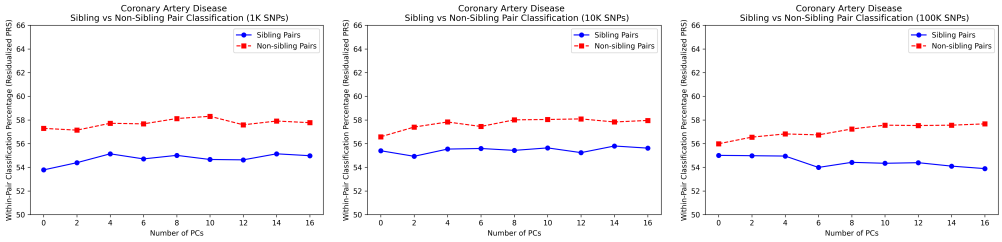
Influence of number of PCs on PRS accuracy

We next explored how the inclusion of increasing numbers of PCs during GWAS impacted classification accuracy in sibling and non-sibling test sets (Figures 3-7).

These results revealed heterogeneous and trait-specific effects. In the case of coronary artery disease (Fig. 3), there was a small but steady increase in population-level performance as more PCs were included in the GWAS. This is also reflected in the sibling set, however, when using 100,000 SNPs the trend appears to be reversed. Notably, the attenuation when moving from non-sibling to sibling pairs appeared to gradually get larger as more PCs are included in the 100,000 SNP CAD model (Fig. 3c). In contrast, in the case of type 2 diabetes (Fig. 4) there did not appear to be an obvious pattern to the effects of including varying numbers of PCs during training and prediction.

Breast cancer demonstrated an inverse pattern to coronary artery disease. In Figure 5, there appears to be a small trend towards decreasing population-level predictive

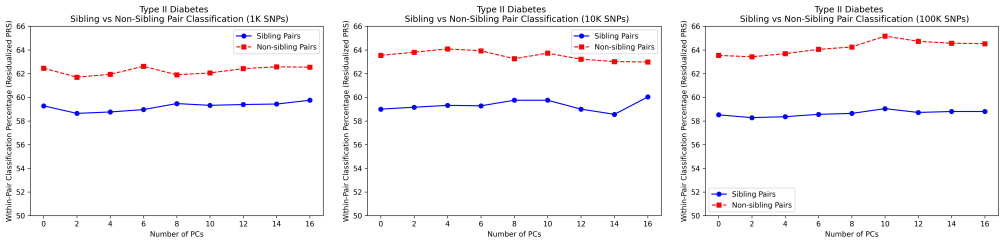
Coronary Artery Disease



(a) 1K SNPs **(b)** 10K SNPs **(c)** 100K SNPs

Fig 3. PRS Results for Coronary Artery Disease: PRS predictive performance among sibling and non-sibling pairs with increasing inclusion of PCs.

Type 2 Diabetes

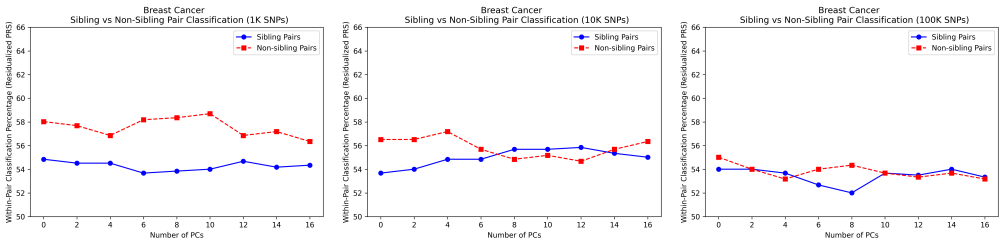


(a) 1K SNPs **(b)** 10K SNPs **(c)** 100K SNPs

Fig 4. PRS Results for Type 2 Diabetes: PRS predictive performance among sibling and non-sibling pairs with increasing inclusion of PCs.

power as more PCs are included. Notably, when only the first 8 PCs are included, performance was actually improved on sibling pairs relative to non-sibling pairs (see Figs. 5b and 5c). This highlights the variability possible when adjusting the specific number of PCs included during model development.

Breast Cancer



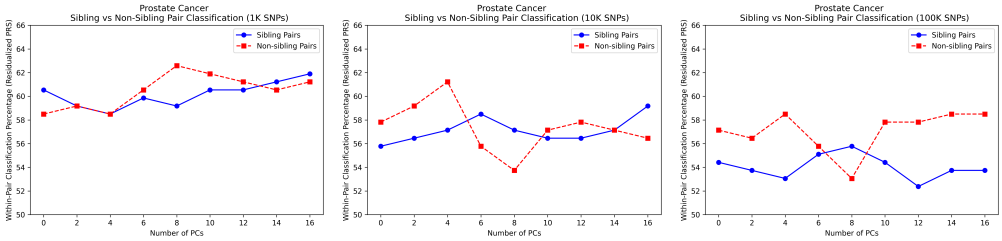
(a) 1K SNPs **(b)** 10K SNPs **(c)** 100K SNPs

Fig 5. PRS Results for Breast Cancer: PRS predictive performance among sibling and non-sibling pairs with increasing inclusion of PCs.

The same variability could be seen in the results for prostate cancer (Fig. 6). Performance on the population-level pairs decreased as the number of PCs grew to 8 with the 10,000 and 100,000 SNP models. The opposite was true, however, for the 1,000

SNP model, where performance is at its maximum when the same 8 PCs have been included. These inconsistencies might also reflect the relatively small number of discordant sibling pairs available for prostate cancer in the UK Biobank.

Prostate Cancer

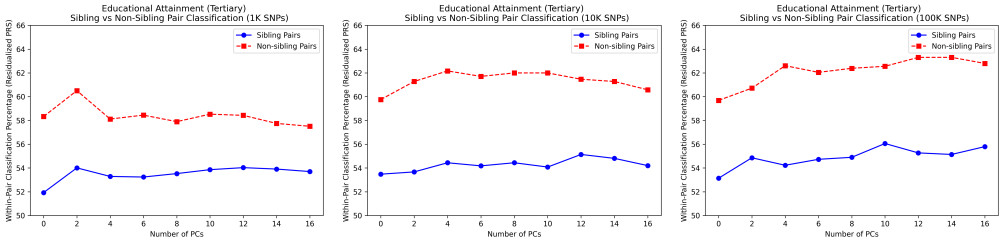


(a) 1K SNPs **(b)** 10K SNPs **(c)** 100K SNPs

Fig 6. PRS Results for Prostate Cancer: PRS predictive performance among sibling and non-sibling pairs with increasing inclusion of PCs.

As was expected, the largest attenuation in performance was seen for the educational attainment phenotype (Fig. 7). There was a significant gulf between the classification accuracy on non-sibling pairs vs. sibling pairs at all SNP-set levels. These results show that confounding may persist even after substantial PC-adjustment, although performance did improve along with PC number for both the non-sibling and sibling test sets in the 100,000 SNP model.

Educational Attainment



(a) 1K SNPs **(b)** 10K SNPs **(c)** 100K SNPs

Fig 7. PRS Results for Educational Attainment: PRS predictive performance among sibling and non-sibling pairs with increasing inclusion of PCs.

It is worth noting that the principal components used were computed across the entire UK Biobank cohort, including individuals of more diverse ancestry than were included in the GWAS conducted in this particular study. To address concerns that this might reduce the PCs' effectiveness for the ancestry-filtered GWAS cohort, we recomputed PCs using only individuals from the post-QC, ethnicity-specific dataset. However, when these recomputed PCs were used in GWAS and PRS construction, they

did not appear to greatly affect predictive accuracy across the diseases examined (see Supplementary Tables S2 and S3).

Effects of PC inclusion on attenuation

To better isolate the effect of PC inclusion on within-family attenuation, we directly examined attenuation percentages across traits and SNP-set sizes, which can be seen in Figure 8. The expectation that attenuation should, in general, decrease as more PCs are included in the GWAS and PRS did not appear to hold. For example, in Fig. 8a attenuation steadily increased as more PCs were included as covariates in the 100,000 SNP GWAS/PRS model. This stood in contrast to the 1,000 SNP educational attainment model in Figure 8e.

The aforementioned expectation was met up to around 8-12 PCS for the 10,000 SNP breast cancer model, whereupon the attenuation began to increase as more PCs are added. The opposite was seen in the 100,000 SNP model, however, where 8 PCs is when attenuation is at its peak. In prostate cancer, attenuation fluctuated widely depending on SNP-set size and PC count whilst no clear trend was apparent in the case of type 2 diabetes.

Taken together, these findings suggest that the relationship between PC inclusion and within-family attenuation is complex and trait-dependent. PC-adjustment in the UK Biobank does not necessarily reduce confounding as measured by within-family attenuation.

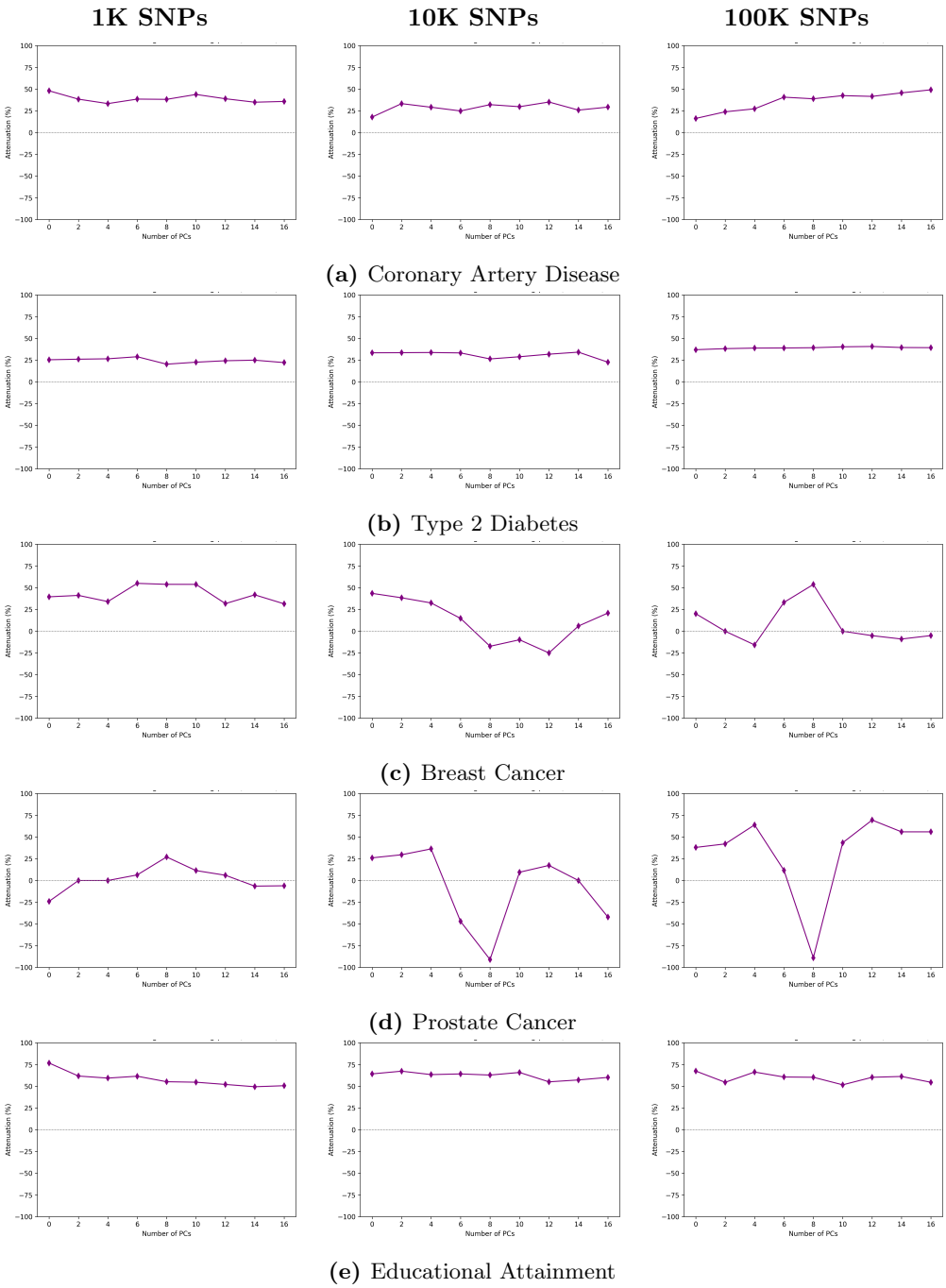


Fig 8. Attenuation in predictive performance with increasing PC inclusion between sibling and non-sibling (population-level) pairs across phenotypes.

Effects of number of included PCs on the genomic inflation factor λ

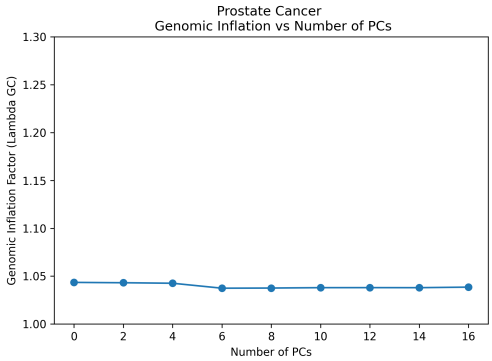
The effects of increasing the number of included PCs on the genomic inflation factor (λ), which is also used as a proxy for confounding, can be seen in Fig. 9 and the first two columns of Table 4.

Table 4. Genomic inflation factor (λ) of GLM- and GLMM Based GWAS.

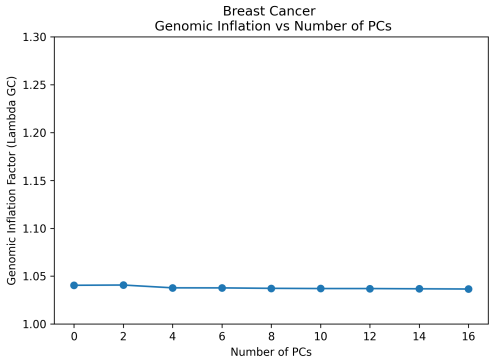
Phenotype	GLM	GLM+16PC	GLMM+16PC
Coronary Artery Disease	1.191	1.153	1.152
Type 2 Diabetes	1.232	1.214	1.211
Breast Cancer	1.040	1.037	1.037
Prostate Cancer	1.043	1.039	1.036
Educational Attainment	1.737	1.324	1.317

The greatest reduction in λ can be seen for the educational attainment phenotype which matches with prior expectation of confounding in the trait. λ is above 1.7 when no PCs are included in the GWAS and this is decreased to around 1.3 with 10 PCs included, where the effect of including more PCs tapers out.

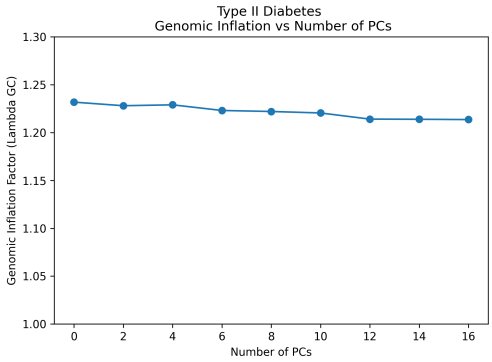
The same effect is visible for coronary artery disease, but to a lesser degree. For the remaining three traits, the effects of PC-adjustment on λ appears to be small in magnitude. These findings suggest that while PCs do reduce global P-value inflation, particularly for highly confounded traits, the relationship between λ reduction and improved within-family performance is not straightforward.



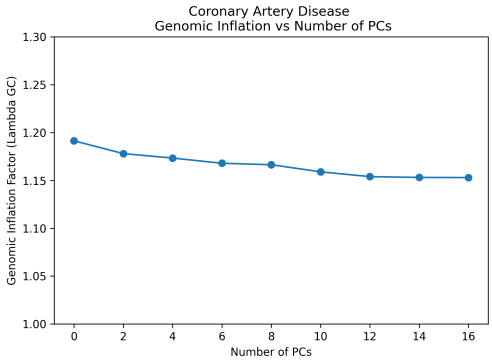
(a) Prostate Cancer



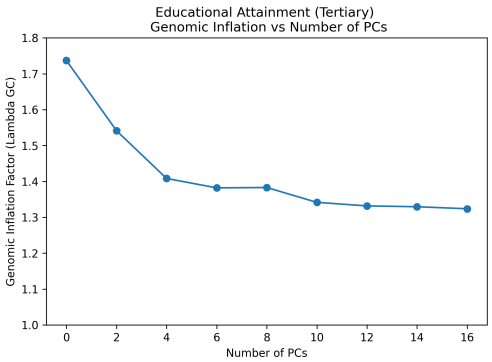
(b) Breast Cancer



(c) Type 2 Diabetes



(d) Coronary Artery Disease



(e) Educational Attainment

Fig 9. Effects of PCs on the genomic inflation factor λ. Summary statistics to calculate λ were generated from GLM-based GWAS. Note that the y-axis limit is higher for the educational attainment phenotype due to additional confounding.

Mixed model GWAS

To investigate whether GRM-based mixed models offer additional control over confounding, we performed GWAS using fastGWA-GLMM with the full set of 16 PCs included as covariates. The PRS results from this GWAS can be seen in Table 5.

With respect to predictive performance on a population-level test set, comparing Tables 3 and 5 demonstrates very little gain in performance when including a GRM during association.

Table 5. Classification accuracy for GLMM-based GWAS PRS performance on sibling and non-sibling (population-level) discordant pairs in the UK Biobank. The full 16 PCs were included in both the GLMM-GWAS and PRS.

Phenotype	SNP-set size	**Classification accuracy (%)**	
		Non-sibling pairs	Sibling pairs
Coronary Artery Disease	1K	57.72	54.98
	10K	57.77	55.62
	100K	57.69	53.94
Type 2 Diabetes	1K	62.70	59.67
	10K	63.02	59.83
	100K	64.49	59.00
Breast Cancer	1K	55.69	54.85
	10K	55.02	53.68
	100K	55.85	54.52
Prostate Cancer	1K	59.18	61.22
	10K	53.74	61.22
	100K	58.50	56.46
Educational Attainment	1K	57.51	53.62
	10K	60.72	54.18
	100K	62.87	55.58

Likewise, in the within-family setting, a similar result is also apparent. Ultimately, classification accuracy using mixed models, evaluated in both sibling and non-sibling test sets, was nearly identical to that of standard GLM models.

Similarly, genomic inflation factors were reduced minimally compared to GLMs with 16 PCs (Table 4) and attenuation remained largely unchanged (Table 6).

These results suggest that in large, relatively homogeneous cohorts such as the UK Biobank, the combination of principal component adjustment and standard association methods may already be sufficient to address population stratification. The added complexity of GRM-based modeling did not meaningfully improve prediction or reduce confounding in this context.

These same results are presented in terms of attenuation in Table 6. For the three traits with the largest sample sizes (coronary artery disease, type 2 diabetes, and educational attainment) there would appear to be little positive effect of a GRM on reducing attenuation when moving from population-level to a within-family setting. The results are more variable for the two cancer phenotypes - although the pattern is not consistent and was dependent on SNP-set size.

Table 6. Attenuation in PRS performance between non-sibling (population-level) and sibling discordant pairs, comparing GLM-based GWAS and GLMM-based GWAS (both with 16 PCs) in the UK Biobank.

Phenotype	SNP-set size	**Attenuation (%)**	
		GLM	GLMM
Coronary Artery Disease	1K	35.91	35.49
	10K	29.40	27.67
	100K	49.28	48.76
Type 2 Diabetes	1K	22.25	23.86
	10K	22.73	24.50
	100K	39.44	37.89
Breast Cancer	1K	31.50	14.76
	10K	20.94	26.69
	100K	-5.03	6.80
Prostate Cancer	1K	-6.06	-22.22
	10K	-42.11	-200.00
	100K	56.00	24.00
Educational Attainment	1K	50.73	51.80
	10K	60.30	61.01
	100K	54.69	56.64

Statistical analysis

To assess the factors influencing classification accuracy, we fit a generalized linear mixed model (GLMM) with phenotype, SNP-set size, pair type (sibling vs. non-sibling), and adjustment method as fixed effects.

The analysis confirmed that phenotype was a strong predictor of classification accuracy ($\chi^2 = 114$, $p < 2 \times 10^{-16}$). Both larger SNP-set sizes ($\chi^2 = 21.6$, $p = 2 \times 10^{-5}$) and non-sibling status ($\chi^2 = 80.1$, $p < 2 \times 10^{-16}$) also had significantly higher predictive accuracy when considering all the data. The adjustment methods differed only modestly overall ($\chi^2 = 9.6$, $p = 0.008$), with the 16-PC and GLMM+16-PC each increasing accuracy only slightly relative to the 0-PC model. The

`method × pair-type` interaction was not significant ($\chi^2 = 0.008$, p = 0.996), indicating that all three methods exhibit a similar attenuation when moving from unrelated to sibling pairs. This result challenges the assumption that more sophisticated correction strategies necessarily yield better within-family transferability of PRS.

Discussion

Many commonly held assumptions about population stratification control in PRS development did not consistently hold across traits in this analysis. In particular, increasing the number of principal components included during GWAS and PRS construction did not reliably reduce within-family attenuation of predictive performance. Whilst a general decrease was observed in the genomic inflation factor with the inclusion of additional PCs, this did not consistently translate into improvements in the within-family prediction task, raising questions about the practical utility of λ as a proxy for confounding.

The use of linear mixed models for GWAS summary statistics also did not significantly reduce either λ or attenuation beyond that of PC-control alone. This is especially notable given the widespread adoption of LMM approaches in the field. For several traits, including breast and prostate cancer, the effects of attempting stratification control appeared minimal regardless of the number of PCs included or the use of LMMs. This potentially reflects low baseline levels of confounding in these traits, or limited statistical power to detect and minimize its effects. Even for educational attainment, which is known to be highly confounded, the success of PC or LMM-based correction in fully accounting for stratification was questionable. While λ decreased with additional PCs as expected, the corresponding attenuation in sibling prediction remained substantial, suggesting residual confounding or inherent limitations in the correction methods themselves.

Two important observations were the lack of correlation between both higher genomic inflation values and increasing SNP-set size with a higher level of within-family attenuation. This stands in contrast to the expectation that λ and low polygenicity may be used as proxies for PRS causality and low risk of confounding.

The predictive gap between the population-level test sets and the within-family

cohort was sometimes minimal, which might also raise questions about the suitability of using within-family comparisons as a gold standard validation step in PRS development. Notably, we still lack a complete understanding of the factors contributing to predictive attenuation when moving from population-level to within-family analyses [43]. While part of the attenuation can be attributed to the shared genetic and environmental backgrounds between siblings, other contributing factors may remain unidentified or unresolved. Although within-family validation should still be considered a vital benchmark for assessing the causal validity of polygenic risk scores, it is unclear whether it fully meets the expectations of this standard, at least in this setting.

Although this study focused only on linear polygenic risk scores, the question of establishing causality may increase in importance if the field of genomic prediction shifts toward reliance on "black-box" machine-learning models such as neural networks [4, 18]. Non-linear models could exploit complex confounding patterns that cannot be adequately addressed through linear approaches such as PCA. For this reason, focus on the development of robust validation frameworks for assessing the causal interpretation of PRS may increase in the years ahead.

Taken together, these findings indicate that the impact of population stratification control is highly trait-dependent, and current best practices such as adjusting for a pre-defined fixed number of PCs may not generalize well across phenotypes. Using the UK Biobank, Lello at al. have previously shown in the case of height that population stratification may not contribute a significant amount to predictive accuracy, and these results validate this finding in the context of human disease using a different framework [21]. Lello at al. regressed height against the principal components and found low explanatory power for the PCs. Interestingly, while we did find that some PCs were significantly associated with each phenotype investigated here, there was no clear association between significance and its effects on performance or attenuation (Supplementary Table S1).

Limitations of this analysis include the number of traits studied and the sample sizes of the test sets, especially for the cancer phenotypes. The size of current within-family cohorts is limited, however there do exist external independent datasets that could be used to potentially validate some of these findings [14]. Furthermore, the results shown here only apply to those of self-declared White ancestry and present in the UK Biobank.

The nature of the recruitment processes and relatively small geographic catchment area might limit the extent to which environmental confounding is a problem in the GWAS being conducted on this data. It is entirely possible that analyses on other populations or more diverse cohorts could potentially see much greater effects of stratification control, especially in those of admixed continental ancestry. Nonetheless, it is currently standard practice to employ these techniques on the UK Biobank and these findings, along with the work of Lello et al., may indicate that current stratification control methods may not always be warranted on this dataset [21].

In conclusion, caution should be exercised when attempting to control for population stratification on polygenic risk scores derived from UK Biobank data. The relationship between genomic inflation, principal component or mixed model adjustment, and the within-family attenuation is complex and not predictable in advance. No one heuristic will likely be sufficient to maximize predictive performance and causality, and serious consideration of trait-specific prior investigation may be warranted during any future PRS development from this dataset.

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 103770. We thank all participants of the UK Biobank for generously sharing their data. We also wish to thank Gianpiero Cavalleri for his support and guidance.

References

1. Fentaw Abegaz, Kridsadakorn Chaichoompu, Emmanuelle Génin, David W Fardo, Inke R König, Jestinah M Mahachie John, and Kristel Van Steen. Principals about principal components in statistical genetics. *Briefings in Bioinformatics*, 20(6):2200–2216, nov 2019.

2. Nick Barton, Joachim Hermisson, and Magnus Nordborg. Population Genetics: Why Structure Matters. *eLife*, 8, mar 2019.

3. Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

4. Pau Bellot, Gustavo de Los Campos, and Miguel Pérez-Enciso. Can Deep Learning Improve Genomic Prediction of Complex Human Traits? *Genetics*, 210(3):809–819, 2018.

5. Jeremy J Berg, Arbel Harpak, Nasa Sinnott-Armstrong, Anja Moltke Joergensen, Hakhamanesh Mostafavi, Yair Field, Evan August Boyle, Xinjun Zhang, Fernando Racimo, Jonathan K Pritchard, and Graham Coop. Reduced signal for polygenic adaptation of height in uk biobank. *eLife*, 8, 3 2019.

6. Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 2018.

7. Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7, dec 2015.

8. Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F. O'Reilly. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, 15:2759–2772, 9 2020.

9. B. Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55:997–1004, 12 1999.

10. Kevin J. Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J. Patterson, and Alkes L. Price. Fast principal-component analysis reveals convergent evolution of adh1b in europe and east asia. *The American Journal of Human Genetics*, 98:456–472, 3 2016.

11. Lachlan Gilchrist, Thomas P. Spargo, Rebecca E. Green, Jonathan R. I. Coleman, David M. Howard, Jackson G. Thorp, Brett N. Adey, Jodie Lord, Helena L. Davies, Jessica Mundy, Abigail R. ter Kuile, Molly R. Davies, Christopher Hübel, Shannon Bristow, Sang Hyuck Lee, Henry Rogers, Charles Curtis, Saakshi Kakar, Chelsea M. Malouf, Gursharan Kalsi, Ryan Arathimos, Anne Corbett, Clive Ballard, Helen Brooker, Byron Creese, Dag Aarsland, Adam Hampshire, Latha Velayudhan, Thalia C. Eley, Gerome Breen, Alfredo Iacoangeli, Sulev Kõks, Cathryn M. Lewis, and Petroula Proitsi. Depression symptom-specific genetic associations in clinically diagnosed and proxy case alzheimer's disease. *Nature Mental Health*, 3:212–228, 1 2025.

12. J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 4 1982.

13. Kangcheng Hou, Ziqi Xu, Yi Ding, Ravi Mandla, Zhuozheng Shi, Kristin Boulier, Arbel Harpak, and Bogdan Pasaniuc. Calibrated prediction intervals for polygenic scores across diverse contexts. *Nature Genetics*, 56:1386–1396, 7 2024.

14. Laurence J. Howe, Michel G. Nivard, Tim T. Morris, Ailin F. Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Rafael Ahlskog, Penelope A. Lind, Teemu Palviainen, Matthijs D. van der Zee, Rosa Cheesman, Massimo Mangino, Yunzhang Wang, Shuai Li, Lucija Klaric, Scott M. Ratliff, Lawrence F. Bielak, Marianne Nygaard, Alexandros Giannelis, Emily A. Willoughby, Chandra A. Reynolds, Jared V. Balbona, Ole A. Andreassen, Helga Ask, Aris Baras,

Christopher R. Bauer, Dorret I. Boomsma, Archie Campbell, Harry Campbell, Zhengming Chen, Paraskevi Christofidou, Elizabeth Corfield, Christina C. Dahm, Deepika R. Dokuru, Luke M. Evans, Eco J. C. de Geus, Sudheer Giddaluru, Scott D. Gordon, K. Paige Harden, W. David Hill, Amanda Hughes, Shona M. Kerr, Yongkang Kim, Hyeokmoon Kweon, Antti Latvala, Deborah A. Lawlor, Liming Li, Kuang Lin, Per Magnus, Patrik K. E. Magnusson, Travis T. Mallard, Pekka Martikainen, Melinda C. Mills, Pål Rasmus Njølstad, John D. Overton, Nancy L. Pedersen, David J. Porteous, Jeffrey Reid, Karri Silventoinen, Melissa C. Southey, Camilla Stoltenberg, Elliot M. Tucker-Drob, Margaret J. Wright, Hyeokmoon Kweon, Philipp D. Koellinger, Daniel J. Benjamin, Patrick Turley, Laurence J. Howe, Michel G. Nivard, Tim T. Morris, Ailin F. Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Rafael Ahlskog, Penelope A. Lind, Teemu Palviainen, Matthijs D. van der Zee, Rosa Cheesman, Massimo Mangino, Yunzhang Wang, Shuai Li, Lucija Klaric, Scott M. Ratliff, Lawrence F. Bielak, Marianne Nygaard, Alexandros Giannelis, Emily A. Willoughby, Chandra A. Reynolds, Jared V. Balbona, Ole A. Andreassen, Helga Ask, Dorret I. Boomsma, Archie Campbell, Harry Campbell, Zhengming Chen, Paraskevi Christofidou, Elizabeth Corfield, Christina C. Dahm, Deepika R. Dokuru, Luke M. Evans, Eco J. C. de Geus, Sudheer Giddaluru, Scott D. Gordon, K. Paige Harden, W. David Hill, Amanda Hughes, Shona M. Kerr, Yongkang Kim, Antti Latvala, Deborah A. Lawlor, Liming Li, Kuang Lin, Per Magnus, Patrik K. E. Magnusson, Travis T. Mallard, Pekka Martikainen, Melinda C. Mills, Pål Rasmus Njølstad, Nancy L. Pedersen, David J. Porteous, Karri Silventoinen, Melissa C. Southey, Camilla Stoltenberg, Elliot M. Tucker-Drob, Margaret J. Wright, John K. Hewitt, Matthew C. Keller, Michael C. Stallings, James J. Lee, Kaare Christensen, Sharon L. R. Kardia, Patricia A. Peyser, Jennifer A. Smith, James F. Wilson, John L. Hopper, Sara Hägg, Tim D. Spector, Jean-Baptiste Pingault, Robert Plomin, Alexandra Havdahl, Meike Bartels, Nicholas G. Martin, Sven Oskarsson, Anne E. Justice, Iona Y. Millwood, Kristian Hveem, Øyvind Naess, Cristen J. Willer, Bjørn Olav Åsvold, Jaakko Kaprio, Sarah E. Medland, Robin G. Walters, David M. Evans, George Davey Smith, Caroline Hayward, Ben Brumpton, Gibran Hemani, Neil M. Davies, John K. Hewitt, Matthew C. Keller,

Michael C. Stallings, James J. Lee, Kaare Christensen, Sharon L. R. Kardia, Patricia A. Peyser, Jennifer A. Smith, James F. Wilson, John L. Hopper, Sara Hägg, Tim D. Spector, Jean-Baptiste Pingault, Robert Plomin, Alexandra Havdahl, Meike Bartels, Nicholas G. Martin, Sven Oskarsson, Anne E. Justice, Iona Y. Millwood, Kristian Hveem, Øyvind Naess, Cristen J. Willer, Bjørn Olav Åsvold, Philipp D. Koellinger, Jaakko Kaprio, Sarah E. Medland, Robin G. Walters, Daniel J. Benjamin, Patrick Turley, David M. Evans, George Davey Smith, Caroline Hayward, Ben Brumpton, Gibran Hemani, and Neil M. Davies. Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nature Genetics*, 54:581–592, 5 2022.

15. A Cecile J W Janssens. Validity of polygenic risk scores: are we measuring what we think we are? *Human Molecular Genetics*, 28:R143–R150, 11 2019.

16. Longda Jiang, Zhili Zheng, Hailing Fang, and Jian Yang. A generalized linear mixed model association tool for biobank-scale data. *Nature Genetics*, 53:1616–1621, 11 2021.

17. Jonathan Michael Kaplan and Stephanie M. Fullerton. Polygenic risk, population structure and ongoing difficulties with race in human genetics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1852), jun 2022.

18. Ciaran Michael Kelly and Russell Lewis McLaughlin. Comparison of machine learning methods for genomic prediction of selected arabidopsis thaliana traits. *PLOS ONE*, 19:e0308962, 8 2024.

19. Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, and Sekar Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224, 2018.

20. James J. Lee, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, Richard Karlsson Linnér, Mark Alan Fontana, Tushar Kundu, Chanwook Lee, Hui Li, Ruoxi Li, Rebecca Royer, Pascal N. Timshel, Raymond K. Walters,

Emily A. Willoughby, Loïc Yengo, Maris Alver, Yanchun Bao, David W. Clark, Felix R. Day, Nicholas A. Furlotte, Peter K. Joshi, Kathryn E. Kemper, Aaron Kleinman, Claudia Langenberg, Reedik Mägi, Joey W. Trampush, Shefali Setia Verma, Yang Wu, Max Lam, Jing Hua Zhao, Zhili Zheng, Jason D. Boardman, Harry Campbell, Jeremy Freese, Kathleen Mullan Harris, Caroline Hayward, Pamela Herd, Meena Kumari, Todd Lencz, Jian'an Luan, Anil K. Malhotra, Andres Metspalu, Lili Milani, Ken K. Ong, John R. B. Perry, David J. Porteous, Marylyn D. Ritchie, Melissa C. Smart, Blair H. Smith, Joyce Y. Tung, Nicholas J. Wareham, James F. Wilson, Jonathan P. Beauchamp, Dalton C. Conley, Tõnu Esko, Steven F. Lehrer, Patrik K. E. Magnusson, Sven Oskarsson, Tune H. Pers, Matthew R. Robinson, Kevin Thom, Chelsea Watson, Christopher F. Chabris, Michelle N. Meyer, David I. Laibson, Jian Yang, Magnus Johannesson, Philipp D. Koellinger, Patrick Turley, Peter M. Visscher, Daniel J. Benjamin, and David Cesarini. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, 50:1112–1121, 8 2018.

21. Louis Lello, Steven G Avery, Laurent Tellier, Ana I Vazquez, Gustavo de los Campos, and Stephen D H Hsu. Accurate Genomic Prediction of Human Height. *Genetics*, 210(2):477–497, oct 2018.

22. Louis Lello, Timothy G. Raben, and Stephen D. H. Hsu. Sibling validation of polygenic risk scores and complex trait prediction. *Scientific Reports*, 10(1):13190, dec 2020.

23. Anna C. F. Lewis and Robert C. Green. Polygenic risk scores in the clinic: new perspectives needed on familiar ethical issues. *Genome Medicine*, 13(1):14, dec 2021.

24. John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens, and Carlos D. Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, nov 2008.

25. Michael JD Powell et al. The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26:26–46, 2009.

26. Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 2006.

27. Alkes L. Price, Noah A. Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11:459–463, 7 2010.

28. Florian Privé, Hugues Aschard, Shai Carmi, Lasse Folkersen, Clive Hoggart, Paul F. O'Reilly, and Bjarni J. Vilhjálmsson. Portability of 245 polygenic scores when derived from the uk biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics*, 109:12–23, 1 2022.

29. Florian Privé, Keurcien Luu, Michael G B Blum, John J McGrath, and Bjarni J Vilhjálmsson. Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics*, 36:4449–4457, 8 2020.

30. Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575, sep 2007.

31. D E Reich and D B Goldstein. Detecting association in a case-control study while correcting for population stratification. *Genetic epidemiology*, 20:4–16, 1 2001.

32. Saskia Selzam, Stuart J. Ritchie, Jean-Baptiste Pingault, Chandra A. Reynolds, Paul F. O'Reilly, and Robert Plomin. Comparing Within- and Between-Family Polygenic Score Prediction. *The American Journal of Human Genetics*, 105(2):351–363, aug 2019.

33. Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779, March 2015.

34. Jae Hoon Sul, Lana S. Martin, and Eleazar Eskin. Population structure in genetic studies: Confounding factors and mixed models. *PLOS Genetics*, 14(12):e1007309, dec 2018.

35. Ali Torkamani, Nathan E. Wineinger, and Eric J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590, sep 2018.

36. Bjarni J. Vilhjálmsson and Magnus Nordborg. The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, 14(1):1–2, jan 2013.

37. Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101(1):5–22, jul 2017.

38. Naomi R. Wray, Jian Yang, Ben J. Hayes, Alkes L. Price, Michael E. Goddard, and Peter M. Visscher. Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, 14(7):507–515, jul 2013.

39. Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88(1):76–82, jan 2011.

40. Jian Yang, Michael N Weedon, Shaun Purcell, Guillaume Lettre, Karol Estrada, Cristen J Willer, Albert V Smith, Erik Ingelsson, Jeffrey R O'Connell, Massimo Mangino, Reedik Mägi, Pamela A Madden, Andrew C Heath, Dale R Nyholt, Nicholas G Martin, Grant W Montgomery, Timothy M Frayling, Joel N Hirschhorn, Mark I McCarthy, Michael E Goddard, and Peter M Visscher.

Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19:807–812, 7 2011.

41. Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100–6, feb 2014.

42. Alexander I. Young, Stefania Benonisdottir, Molly Przeworski, and Augustine Kong. Deconstructing the sources of genotype-phenotype associations in humans. *Science*, 365:1396–1400, 9 2019.

43. Alexander Strudwick Young. Genome-wide association studies have problems due to confounding: Are family-based designs the answer? *PLOS Biology*, 22:e3002568, 4 2024.

44. Gang Zheng, Boris Freidlin, and Joseph L. Gastwirth. Robust genomic control for association studies. *The American Journal of Human Genetics*, 78:350–356, 2 2006.

Supporting information

Supplementary Note 1 - Phenotype Definitions

Type 2 diabetes

Individuals assigned into the positive class were those with any of the following codes in "Diagnoses - main ICD10", "Diagnoses - secondary ICD10", or "Diagnoses - ICD10": "E110", "E111", "E112", "E113", "E114", "E115", "E116", "E117", "E118", "E119".

Coronary artery disease

Individuals assigned into the positive class were those with any of the following codes in "Diagnoses - main ICD10", "Diagnoses - secondary ICD10", or "Diagnoses - ICD10": "I200", "I201", "I208", "I209", "I210", "I211", "I212", "I213", "I214", "I219", "I220", "I221", "I228", "I229", "I235", "I236", "I238", "I240", "I250", "I251", "I252", "I253", "I254", "I255", "I256", "I258", "I259" and "I21.X"

Breast cancer

Individuals assigned into the positive class were those with any of the those with self-reported cancer code 1002.

Prostate cancer

Individuals assigned into the positive class were those with any of the those with self-reported cancer code 1044.

Educational attainment

Codes in UKBB Qualifications Field (6138) were converted to years of education via the following mapping: (1,2,3,4,5,6,-7,-3) $\rightarrow$ (20,13,10,10,19,15,7,NA). From this, a constructed binary phenotype was generated consisting of those with $\leq$7 years of education (positive class) and those with >7 years of education (negative class). Dichotomizing educational attainment in this manner sacrifices information and interpretability. We employ it only to allow for confounding comparisons across binary

disease outcomes. Results for this proxy should not be interpreted as estimates for a true binary educational phenotype.

Association between PCs and phenotypes

Table S1. Linear regression results between PCs and each phenotype. P-values are shown with significant PCs in bold.

	Coronary Artery Disease	Type 2 Diabetes	Prostate Cancer	Breast Cancer	Educational Attainment
PC1	7.7e-01	**2.0e-03**	9.6e-01	2.1e-01	**9.7e-05**
PC2	9.2e-01	7.1e-01	3.2e-01	8.7e-01	**5.1e-04**
PC3	1.2e-01	**5.6e-03**	1.5e-01	3.3e-01	8.3e-01
PC4	**1.4e-04**	**4.4e-02**	1.6e-01	2.7e-01	**8.6e-50**
PC5	**2.3e-05**	**2.0e-05**	**2.7e-06**	4.8e-01	**6.5e-26**
PC6	1.0e-01	**4.2e-04**	6.9e-01	1.5e-01	**3.4e-03**
PC7	2.8e-01	**1.7e-02**	4.4e-01	4.9e-01	**3.0e-48**
PC8	**2.6e-05**	**7.5e-03**	**4.9e-03**	**2.2e-02**	**2.0e-02**
PC9	**4.6e-16**	5.4e-01	9.7e-02	9.0e-01	**3.0e-117**
PC10	**5.8e-03**	7.3e-02	7.5e-01	2.3e-01	**3.3e-03**
PC11	**2.2e-16**	**1.4e-14**	7.0e-01	9.4e-01	**1.4e-38**
PC12	1.4e-01	**3.5e-03**	9.7e-01	5.9e-02	7.4e-02
PC13	8.2e-02	3.1e-01	9.9e-01	2.0e-01	5.4e-01
PC14	**8.1e-03**	5.2e-02	5.5e-01	**3.9e-02**	**1.6e-03**
PC15	9.0e-01	3.4e-01	2.3e-01	3.1e-01	3.4e-01
PC16	**1.9e-02**	**5.5e-04**	2.1e-01	1.3e-01	**2.4e-31**

GWAS/PRS results with recomputed PCs

The first 16 PCs were re-calculated on post-QC self-reported white dataset using PLINK2's approximate PCA method [7, 10]. 126,575 pruned SNPs were used as input for this PCA using a window-size of 50kb, step-size of 5bp and squared correlation threshold of 0.05 [30]. The HLA region was excluded during pruning. A standard GLM-based GWAS and PRS were performed as in the main text using these re-computed 16 PCs as well as age and sex as covariates.

Table S2. PRS with recomputed PCs: Classification accuracy (%) in standard GLM-based GWAS-based PRS performance on discordant sibling pairs in the UK Biobank. Age, sex and the first 16 recomputed PCs were included in the GWAS and PRS.

Phenotype	1K SNPs	10K SNPs	100K SNPs
Coronary Artery Disease	56.07	55.11	56.07
Type 2 Diabetes	59.24	59.75	58.28
Breast Cancer	53.51	54.68	54.18
Prostate Cancer	56.46	61.90	53.74

Table S3. PRS with recomputed PCs: Classification accuracy (%) in standard GLM-based GWAS-based PRS performance on non-sibling discordant pairs in the UK Biobank. Age, sex and the first 16 recomputed PCs were included in the GWAS and PRS.

Phenotype	1K SNPs	10K SNPs	100K SNPs
Coronary Artery Disease	57.45	58.39	57.29
Type 2 Diabetes	62.30	63.54	64.25
Breast Cancer	58.36	55.69	54.35
Prostate Cancer	61.22	56.46	61.22