High trait variability in optimal polygenic prediction strategy within multiple-ancestry cohorts

B.C.L. Lehmann [*1,2], M. Mackintosh[3,4], G. McVean [†1], and C.C. Holmes [†1,2,3]

[1]Big Data Institute, University of Oxford, UK
[2]Department of Statistics, University of Oxford, UK
[3]The Alan Turing Institute, London, UK
[4]The Health Foundation, London, UK
[*]Corresponding author

Abstract

Polygenic scores (PGS) are individual-level measures that quantify the genetic contribution to a given trait. PGS have predominantly been developed using European ancestry samples and recent studies have shown that the predictive performance of European ancestry-derived PGS is lower in non-European ancestry samples, reflecting differences in linkage disequilibrium, variant frequencies, and variant effects across populations. However, the problem of how best to maximize performance within any one ancestry group given the data available, and the extent to which this varies between traits, are largely unexplored. Here, we investigate the effect of sample size and ancestry composition on the predictive performance of PGS for fifteen traits in UK Biobank and evaluate an importance reweighting approach that aims to counteract the under-representation of certain groups within training data. We find that, for a minority of the traits, PGS estimated using a relatively small number of Black/Black British individuals outperformed, on a Black/Black British test set, scores estimated using a much larger number of White individuals. For example, a PGS for mean corpuscular volume trained on only Black individuals achieved a 4-fold improvement on a corresponding PGS trained on only White individuals. For the remainder of the traits, the reverse was true; a PGS for height trained on only Black/Black British individuals explained less than 0.5% of the variance in height in a Black/Black British test set, compared to 3.9% for a PGS trained on a much larger training set consisting of only White individuals. We find that while importance weighting provides moderate benefit for some traits (for example, 40% improvement for mean corpuscular volume compared to no reweighting), the improvement is modest in most cases, arguing that only targeted collection of data from underrepresented groups can address differences in PGS performance.

[*]Corresponding author
[†]Joint supervision

Introduction

Polygenic scores (PGS) are quantitative measures that aim to predict complex traits from genetic data. As well as providing insights into the genetic architecture of complex traits, PGS have considerable clinical potential for screening and prevention strategies [9, 40]. Largely driven by significant increases in sample sizes, the predictive utility of PGS has improved substantially in recent years for a variety of traits [19], including cardiovascular disease [20], breast cancer [22], and Type I diabetes [37].

These improvements, however, have largely been limited to populations of European ancestry [26, 12, 25], reflecting the lack of ethnic diversity in genomic samples collected to date [25]. The lack of transferability of PGS across ancestries may be due to a number of factors, including population differences in allele frequencies and linkage disequilibrium [25]. There is also increasing evidence that the underlying variant effects differ across ancestries [5, 3], which may be due to gene-by-gene or gene-by-environment interactions [2]. This raises one of the most important technical and ethical challenges in the clinical utility and applications of PGS due to their potential negative impacts on health inequalities [25].

There are major, ongoing initiatives to collect genomic data from traditionally under-represented groups, such H3Africa [11], that aim to address the lack of global genetic diversity in research data. However, it may take many years to collect sufficient data to reduce the disparities in PGS performance. Statistical methods provide an alternative, short-term, cost-effective and complementary potential solution to mitigate against the negative effects of the lack of diversity in genomic datasets, by using modelling techniques to make use of all the existing data available, while allowing for some differences between groups. There has been a growing interest in statistical methods to improve the transferability of PGS, which have thus far focused on GWAS-derived PGS, i.e. PGS based on summary statistics from a genome-wide association study. Grinde et al. [15] suggest using European GWAS results to select variants, then using the non-European GWAS to estimate the variant weights. Marquez-Luna et al. [24] propose a multi-ethnic PGS by combining scores trained separately on different populations. In a related approach, Cavazos & Witte [6] use local ancestry weighting to construct PRS for admixed individuals. These efforts have yielded promising improvements in PGS performance, though there remains a significant gap in predictive accuracy between European and non-European target populations.

Here, we investigate the use of multiple-ancestry datasets, such as population-scale biobanks, to estimate PGS for a range of anthropometric and blood-sample traits, with the explicit objective of improving predictive accuracy for under-represented ancestries. Specifically, we ask whether there are consistent optimal strategies for borrowing information from across ancestry groups to maximize prediction accuracy in groups (which may be defined either on the basis of self-identified ethnicity or genetically-inferred ancestry) that have small sample sizes in available resources. For each trait, we construct training sets with varying numbers of individuals from each ancestry to assess the effect of sample size and composition on PGS accuracy, using both simulated and real data from UK Biobank [4]. Moreover, and to counteract the imbalance of ancestries in a multiple-ancestry training set, we investigate the use of an importance reweighting approach that places more weight on underrepresented ancestries during training. Given the availability of individual-level information in biobanks, we estimate PGS using regularised regression applied to full genotype and covariate data (as opposed to

genome-wide association summary statistics) to avoid introducing additional artefacts into our analysis via the reliance on assumptions about genotype and covariate correlation structure (including linkage disequilibrium) and GWAS methodology.

Our results show that the effect of sample size and composition on predictive performance is highly variable across traits. For some traits, polygenic scores estimated using a relatively small number of Black/Black British individuals outperformed on a Black/Black British test set scores estimated using a much larger number of White individuals. Moreover, adding White individuals to the training set did not always improvement performance and in some cases even led to poorer performance. Although importance weighting yields moderate improvement in performance for some traits, we find that sample size is a much more prominent factor, highlighting the limitations of statistical corrections and the importance of collecting more data from diverse sources.

Results

Overview of methods

To investigate the effect of sample size and ancestry composition on polygenic scores for different traits in UK Biobank [4], we constructed a range of training sets controlling the number of White and Black/Black British individuals. We considered three types of training sets: a single-ancestry set consisting only of White individuals, a single-ancestry set consisting of Black/Black British individuals, and a dual-ancestry set consisting of both White and Black/Black British individuals.

For each training set, we estimated PGS using L1-regularised regression, also known as the LASSO [39], which has previously been used in the context of genetic risk prediction (see for example [44, 31, 34, 35]). For the dual-ancestry training sets, we also estimated PGS using an importance reweighted LASSO, upweighting the Black/Black British individuals. Following Martin et al. [25], we assessed the predictive performance of a PGS using partial r^2 relative to a covariate-only model. See Materials and Methods for full details.

Adding individuals from one ancestry does not always improve PGS performance for a different ancestry

We first set out to examine whether prediction performance in an underrepresented ancestry group can be boosted by including individuals from a different ancestry among the training samples. First, we held the number of Black/Black British individuals fixed at approximately 5 000, corresponding to 80% of Black/Black British individuals with non-missing data for the trait in question. We then constructed six training sets, varying the number of White individuals from 0 to just under 50 000, so that the proportion of White individuals in the training set ranged from 0% to 90%. For each training set, we used the unweighted LASSO to generate polygenic scores.

Predictive performance on White individuals improved for each trait as the number of White individuals in the training set increased (Figure 1A). In contrast, predictive performance on Black/Black British individuals only increased markedly for height and platelet count, with the improvement for platelet count appearing to tail off between the two largest training sets. For three traits (mean

corpuscular haemoglobin, white blood cell count, mean corpuscular volume), predictive performance worsened as White individuals were added to the training set. For the remainder of the traits, predictive performance stayed mostly constant.

Next, we performed reciprocal analyses in which we held the number of White individuals fixed at approximately 50 000, and instead varied the number of Black/Black British individuals from 0 to approximately 5 000. The predictive performance on Black or British individuals increased for five traits: mean corpuscular haemoglogin (MCH), white blood cell count (WBC), platelet count, red blood cell count (RBC), and mean corpuscular volume (MCV), with the largest increase ($\Delta r^2 = 0.6$) observed for MCH (Figure 1B). For the remainder of the traits, predictive performance stayed largely stable. Predictive performance on White individuals also remained largely constant for each of the fifteen traits. These results demonstrate the potential for increasing prediction performance by including samples from a different population, but also high between-trait heterogeneity in response, suggestive of major differences in genetic architecture. This raises the possibility that more flexible schemes which combine all samples, but with reweighting, may enable optimal prediction performance.

Importance reweighting can provide modest improvement in PGS performance

Next, we set out to evaluate the potential of reweighting through simulation. We generated 50 quantitative traits, with genetic effect correlations ranging from $\rho = 0.5, 0.6, \ldots, 0.9$, and for each trait we created five training sets with $n_{EUR} = 1\,800$ randomly selected individuals of European ancestry and $n_{AFR} = 200, 450, 771, 1\,200, 1\,800$ randomly selected individuals of African ancestry. We applied the importance reweighted LASSO for $\gamma = 0, 0.1, \ldots, 1$ to obtain estimates $\hat{\beta}$ for β. See Simulation Study for full details.

The predictive performance for African individuals improved substantially as the number of African individuals in the training set increased (Figure 2). For example, when $n_{AFR} = 200$ and the correlation between genetic effects ρ was 0.7, the mean proportion of variance explained by the unweighted LASSO ($\gamma = 0$) on was just under 0.2 for the African test sets, compared to 0.4 for the European test sets. While the performance on Europeans did not change markedly as the number of African training set individuals increased, performance on Africans increased to approximately 0.3 and 0.4 when $n_{AFR} = 771$ and $n_{AFR} = 1\,800$ respectively. As expected, the discrepancy between the two ancestry groups decreased as the correlation ρ between the European genetic effects and the African genetic effects increased.

Importance reweighting generally had a positive impact on predictive performance for African individuals for each set of simulations. The effect on predictive performance depended on the degree of reweighting, quantified by γ. As γ increased from 0 to 1, predictive performance typically increased up to an optimal value γ^* before decreasing. This reflects a crucial trade-off of importance reweighting in this context: while the bias of African ancestry genetic effect estimates may be lower with more reweighting, the increased variance of the weights results in a lower *effective* sample size.

The effect of reweighting was relatively small compared to the effect of increased sample size. For $n_{AFR} = 200$ and $\rho = 0.7$, the optimal amount of reweighting was $\gamma^* = 0.7$, with predictive performance of $r^2 = 0.21$, compared to $r^2 = 0.18$ for the unweighted LASSO ($\gamma = 0$). The amount of improvement

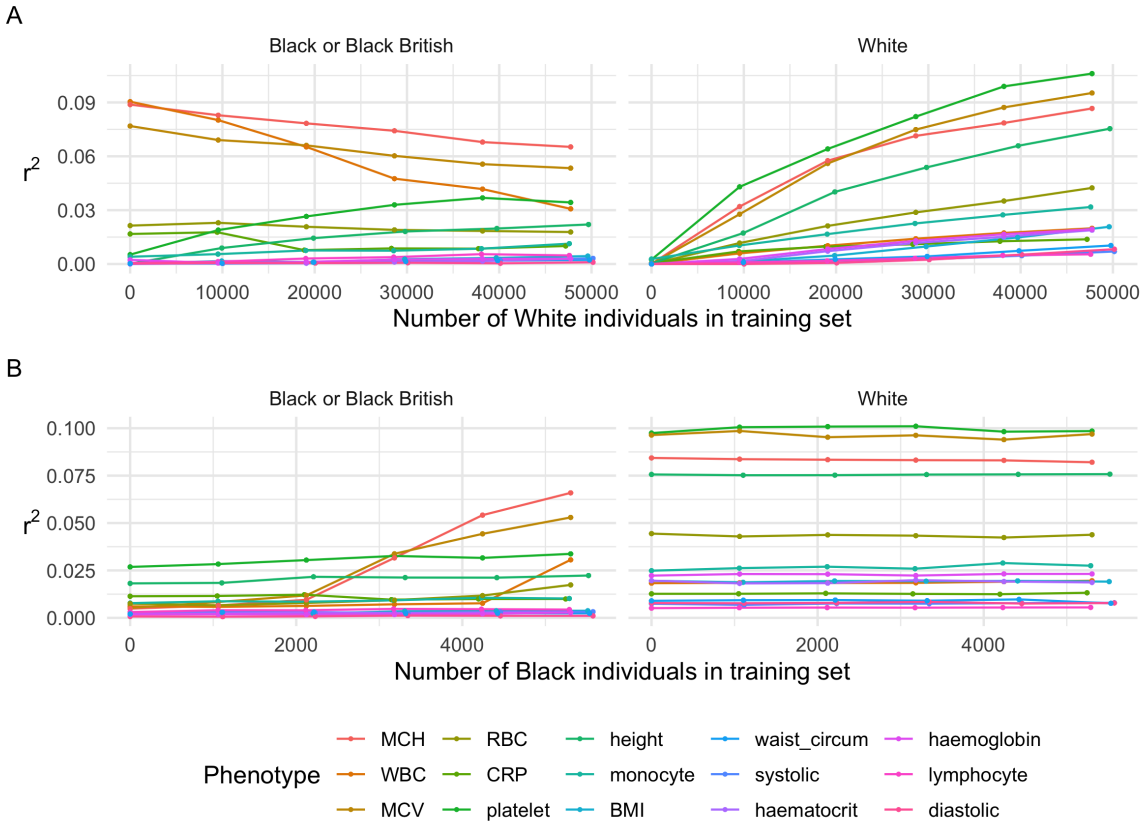

Figure 1: Partial r^2 against sample size for 15 quantitative traits in UK Biobank. A) Here, we fixed the number of Black/Black British individuals in the training set at approximately 5 000 for each trait and varied the number of White individuals so that the proportion of White in the training set ranged from 0% to 90%. While predictive performance on White individuals increased monotonically with sample size for each trait (right panel), the effect on performance on Black/Black British individuals varied by trait (left panel). B) Here, we instead fixed the number of White individuals in the training set at approximately 50 000 and varied the number of Black/Black British individuals from 0 to approximately 5 000. The predictive performance on Black or British individuals increased for five traits: mean corpuscular haemoglogin (MCH), white blood cell count (WBC), platelet count, red blood cell count (RBC), and mean corpuscular volume (MCV); and stayed largely stable for the remainder. Predictive performance on White individuals remained stable for each of the fifteen traits.

with reweighting decreased only slightly as n_{AFR} increased, and decreased more markedly as the correlation between genetic effects ρ increased.

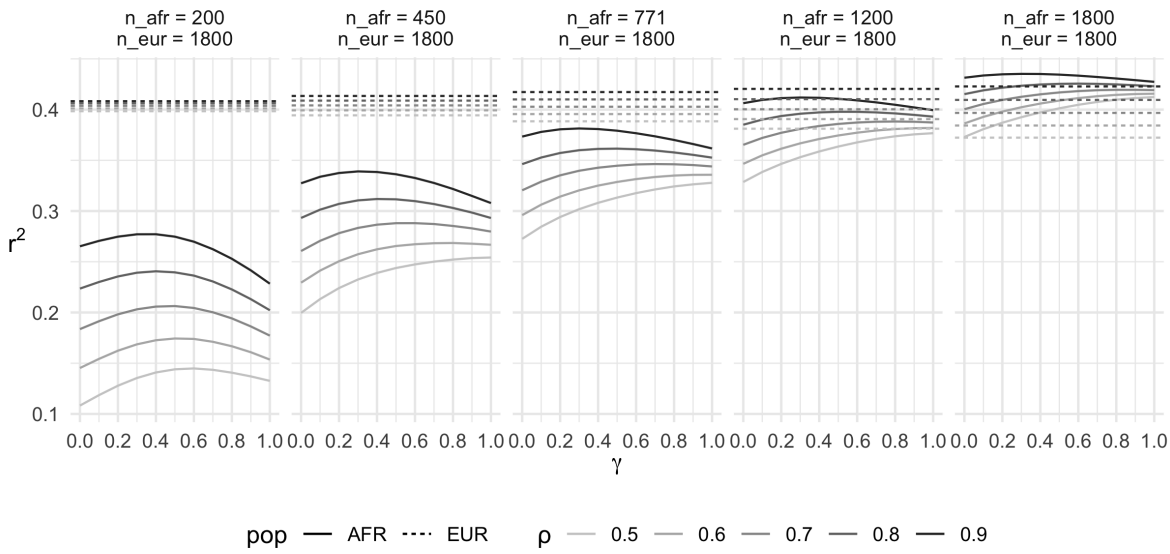

Figure 2: Predictive performance for African individuals improves as more Africans are added to the training set. Each panel corresponds to a different training set, with the number of Europeans fixed at $n_{EUR} = 1\,800$ and the number of Africans varying from $n_{AFR} = 200$ to $n_{AFR} = 1\,800$. Each line represents the mean proportion of variance explained r^2 across 50 simulated traits. The horizontal dashed lines correspond to r^2 for European test sets based on an unweighted LASSO, while the solid lines correspond to r^2 for African test sets. The correlation of genetic effects between ancestries ρ was varied from 0.5 (light grey lines) to 0.9 (black lines).

This simulation study illustrates the relative influence of sample size and importance weighting on the transferability of PGS as a function of the correlation of genetic effects between ancestries. Reducing the disparity in the numbers of samples in a multiple-ancestry dataset reduces the gap in predictive performance between the majority ancestry and the minority ancestry. On the other hand, we find that importance weighting has a relatively limited effect on performance, particularly when the imbalance of ancestries is high.

Optimal ancestry composition of training sets varies among traits

Next, we sought to explore the value of the reweighting approach in empirical data by considering a range of traits in UK Biobank. We considered three training sets: a White training set of approximately 300 000 White individuals, a Black-only training set of approximately 5 000 Black/Black British individuals, and a dual-ancestry training set made up of the Black training set combined with White individuals so that the proportion of Black individuals was 10%. For the dual-ancestry training set, we used the importance weighted LASSO with $\gamma = 0, 0.2, \ldots, 1$ to construct the PGS, while for the single-ancestry training sets we just considered the standard, unweighted LASSO.

Figure 3A illustrates the predictive performance of the PGS for mean corpuscular volume. For

the Black/Black British test set, the PGS estimated using only Black or British individuals outperformed the PGS estimated using only White individuals, despite the former sample size being much smaller ($n = 5\,304$ versus $n = 295\,774$). With no reweighting, the PGS estimated using both White and Black/Black British individuals, performed slightly worse than the PGS estimated using only Black/Black British individuals. Importance reweighting yielded a moderate improvement in partial r^2 with the optimal amount of reweighting at $\gamma^* = 0.6$. For the White test set, the PGS estimated using only White individuals performed best, yielding a partial r^2 of 0.18. The PGS estimated using only Black/Black British individuals offered no predictive power on top of a covariate-only model. Importance reweighting on the dual-ancestry training set reduced predictive performance for White individuals, reflecting the reduced effective sample size. The number of variants selected in the White PGS (i.e. those with non-zero effect estimates) was approximately 20 000 compared to approximately 2 100 for the unweighted dual-ancestry PGS and 83 from the Black PGS (Figure S1).

The predictive performance of the polygenic scores for height displayed somewhat different characteristics for the Black/Black British test set (Figure 3B). In this case, the score estimated using only White individuals performed best, followed by the score estimated on both White and Black/Black British individual. The score estimated using only Black/Black British individuals performed worst, offering very little predictive power over a covariate-only model. Importance reweighting had very little effect, with partial r^2 staying largely constant over the range $\gamma = 0, 0.2, 0.4, 0.6$ and decreasing slightly for $\gamma = 0.8, 1$. The number of variants selected in the White PGS was approximately 25 000 compared to approximately 9 000 for the unweighted dual-ancestry PGS and 250 for the Black PGS (Figure S2).

The fifteen traits we considered could be split into two groups according to whether the score estimated using only a relatively small number of Black/Black British individuals outperformed the score estimated using a much larger number of White individuals (Figure 4). The first group consisted of mean corpuscular haemoglobin, white blood cell count, mean corpuscular volume, red blood cell count, and C-reactive protein for which the Black score outperformed the White score. For these traits, importance reweighting had a noticeably positive effect for the dual-ancestry scores but did not significantly outperform the Black scores. For the remaining ten traits in the second group, the White score outperformed the Black score. For these traits, importance reweighting had a much reduced effect on predictive performance.

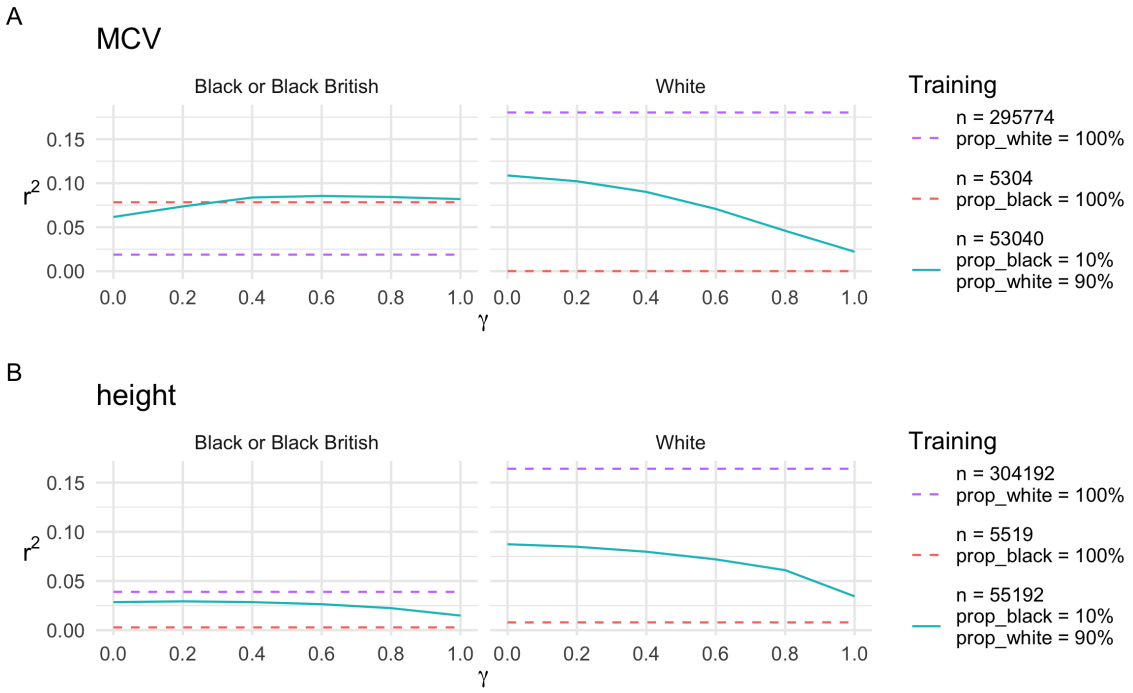

Figure 3: Partial r^2 on of polygenic scores for A) mean corpuscular volume (MCV) and B) height using Black-only (red dashed line), White (purple dashed line) and dual-ancestry (blue solid line) training sets. The single-ancestry scores were estimated using a standard, unweighted LASSO. The dual-ancestry scores were constructed using an importance weighted LASSO with various degrees of reweighting γ.

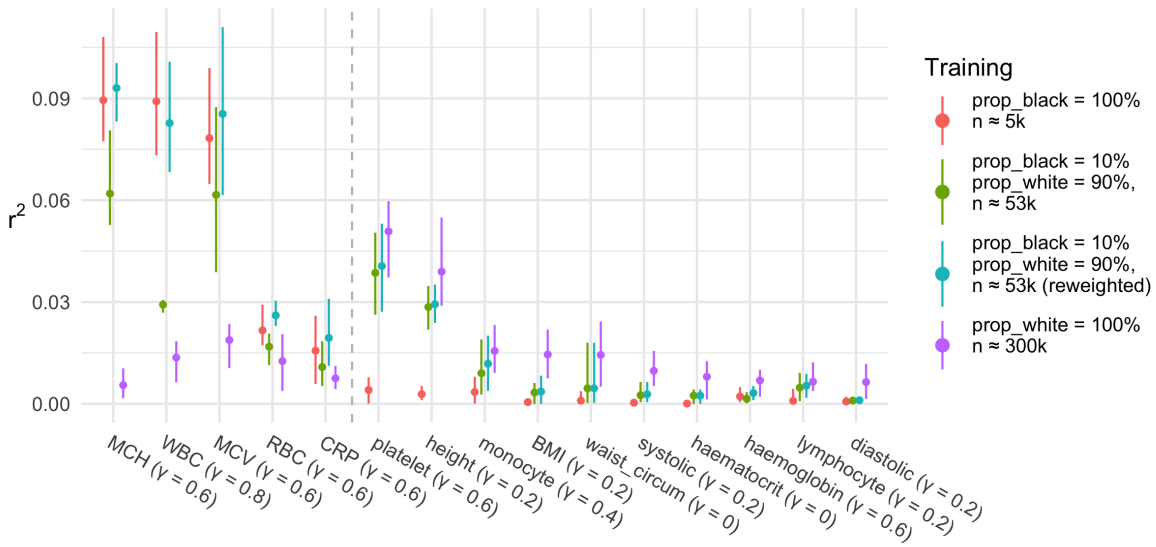

Figure 4: Partial r^2 for Black/Black British individuals for fifteen quantitative traits, using different training sets. The traits can be split into two groups indicated by the grey dashed line: those for which a Black PGS outperforms the White PGS (MCH, WBC, MCV, RBC, CRP) and those for which the reverse is true. The performance of the unweighted dual-ancestry PGS sits between the White and Black PGS for each trait. Reweighting has most effect on the traits for which the Black PGS performs best.

Differences in trait architecture explain variable performance by ancestry

Finally, to investigate the reasons why optimal training approaches vary across traits, we investigated the contribution of variants at different allele frequencies to prediction accuracy. Specifically, we measured the partial r^2 for different subsets of variants grouped by minor allele frequency (MAF) in a given ethnic group. We grouped variants into four sets: 'rare' (MAF $\leq 1\%$), 'uncommon' ($1\% <$ MAF $\leq 5\%$), 'intermediate' ($5\% <$ MAF $\leq 20\%$), and 'common' (MAF $> 20\%$). Given a set of variants $\mathcal{K}$ and the genotype matrix of the test set G, let $G_{\mathcal{K}}$ denote the submatrix given by the columns of G that are in $\mathcal{K}$. For a set of variants $\mathcal{K}$, we define $\hat{Y}_{\mathcal{K}} = X\hat{\theta} + G_{\mathcal{K}}\hat{\beta}_{\mathcal{K}}$ to be the score associated with the set $\mathcal{K}$, where X is the test covariate matrix, and $(\hat{\theta}, \hat{\beta})$ are the LASSO estimates. We defined the partial r^2 to be the difference in r^2 between models

$$Y = X\theta_1 + 1\hat{Y}_{\mathcal{K}}\eta + \epsilon_1 \tag{1}$$

$$Y = X\theta_2 + \epsilon_2. \tag{2}$$

We note that because we are considering the same set of genotyped variants for each trait, observed differences among traits in composition likely reflect differences in trait architecture, rather than differences in how the studies were performed.

Figure 5 illustrates the contribution to prediction accuracy explained by different segments of the allele frequency spectrum for the PGS for mean corpuscular volume and height. We observe striking

differences between these traits in the contribution of different allele frequency segments. For MCV in Black/Black British individuals, the majority of the predictive performance in each of the polygenic scores was due to variants that were intermediate or common in Black/Black British individuals (see top row, LHS). Notably, over half the contribution to prediction accuracy of the Black and dual-ancestry PGS could be attributed to variants that have MAF $< 5\%$ in the White subgroup (second row, LHS). Conversely, prediction accuracy in the White subgroup was driven by variants that are intermediate or common in White individuals (fourth row, LHS). And approximately one quarter of the accuracy could be attributed to variants with MAF $< 5\%$ in Black/Black British individuals.

In contrast, for height we find that predictive performance in White individuals has an allele frequency decomposition with a slightly greater contribution of variants that are common in White individuals (bottom row, RHS). But, predictive performance in Black/Black British individuals is dominated by variants that are common (MAF $> 5\%$) in both populations (top two rows, RHS). The effect of reweighting follows a similar pattern; for MCV shifting predictive power towards variants that are higher-frequency in the Black/Black British individuals (and considerably rarer in White individuals) and, for height, tending to favour variants that are common in both groups.

These results suggest differences in genetic architecture between the two traits. For mean corpuscular volume, the variance explained by the dual-ancestry PGS and the Black PGS could largely be attributed to variants that were relatively common in Black/Black British individuals but rare in White individuals. Since the number of Black/Black British individuals in these training sets was relatively small, this suggests that the corresponding effect sizes are comparatively large. On the other hand, the variance explained by each of the height PGS could be attributed to variants that were common in both White and Black/Black British individuals. Such differences in trait architecture are indicative of variation in the evolutionary and selective forces that have shaped trait variation [38, 45, 36].

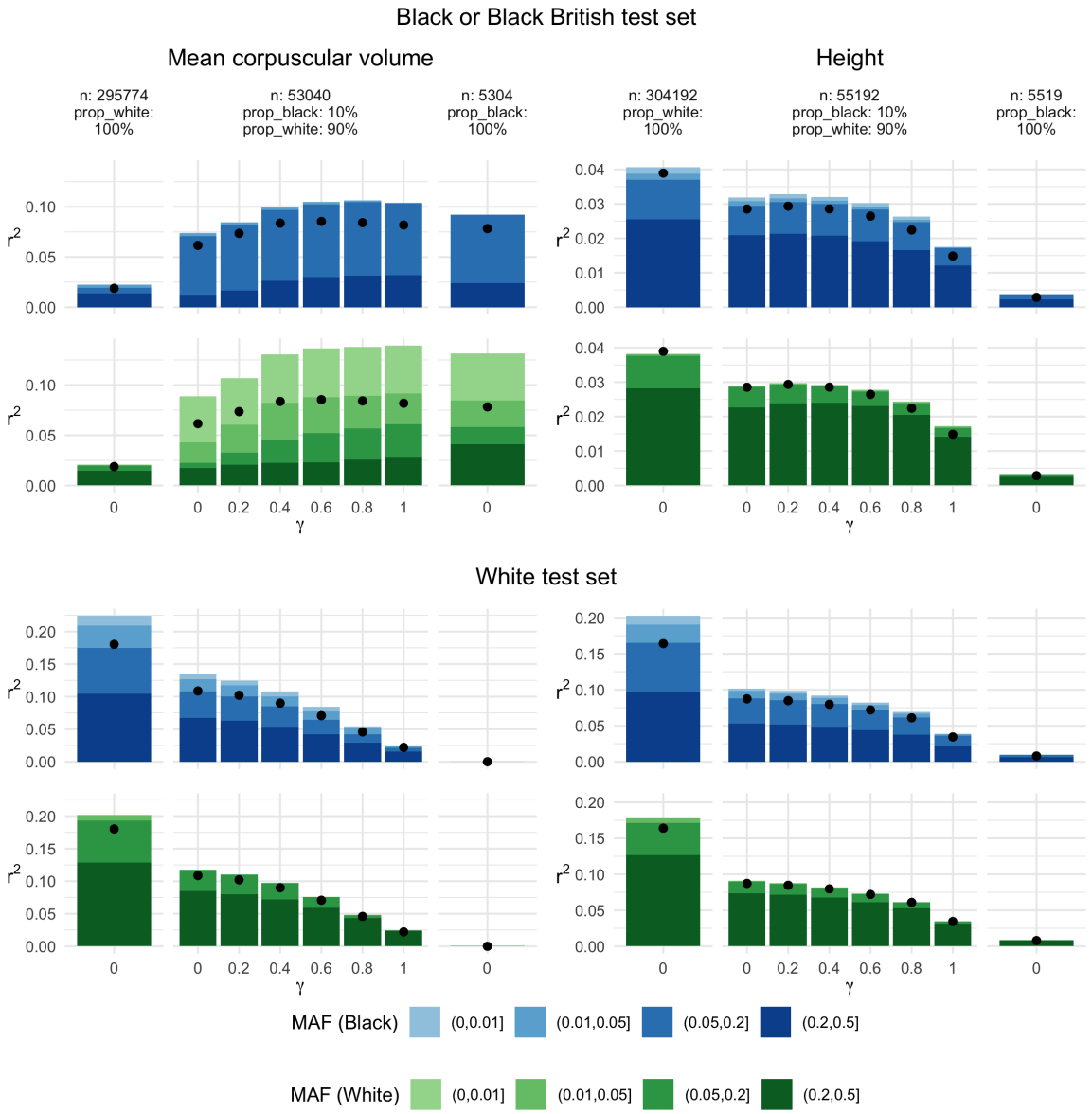

Figure 5: Allele frequency composition of variance explained by PGS for mean corpuscular volume (left) and height (right) in a Black/Black British test set (top) and a White test set (bottom). The black dots represent partial r^2 for all the variants, i.e. the entire polygenic score. Variants were grouped according to their minor allele frequency in Black/Black British individuals (blue palette) or in White individuals (green palette). Each bar represents the sum of the partial r^2 values for each subset of variants in a given polygenic score. Note that the height of the bar is generally higher than corresponding dot due to linkage disequilibrium between variants.

Discussion

The lack of diversity in human genetic studies has been brought into focus by a number of articles (e.g. [29, 33, 25]), revealing that around 80% of all GWAS participants are of European descent. In the context of polygenic scores, this bias has been reflected in the lack of transferability of scores across ancestries: PGS developed using European-ancestry samples tend to perform poorly in non-European test sets [26, 25, 12]. This raises serious technical, clinical and ethical issues, with likely substantial impacts on health disparities and inequalities if left unaddressed [25]. Correspondingly, our results found that, for a range of quantitative traits, PGS estimated using individuals self-reporting as White performed relatively poorly on individuals self-reporting as Black/Black British.

Here, we have considered the extent to which PGS performance in underrepresented ethnic groups can be boosted by training in datasets of multiple ancestries, but where statistical methods are used to focus weight on the ethnic group in question. The ability to ask such a question has been made possible by the advent of large-scale biobanks, which combine population-scale individual-level genetic data from multiple ancestries, accompanied by a wide range of phenotypic measurements. Previous studies on the lack of transferability of PGS have generally estimated scores using summary statistics from genome-wide association studies of single-ancestry populations [26, 25, 12]. These summary statistic approaches are often highly efficient computationally and typically achieve highly competitive predictive performance relative to full genotype approaches [21, 23, 43, 35]. However, combining such data such raises complexities, including assumptions that have to be made about the correlation structure of untyped variants and the comparability of phenotype definitions. By utilising a common set of genotyped variants on individual-level data across multiple ancestries and where phenotypic data is (at least approximately) comparable, enables us to focus on the cause of differences in performance and to consider strategies for their mitigation.

Our primary finding is that traits vary substantially in their optimal strategy for combining data across ancestries. Through simulation, we have shown that there are plausible regions of parameter space, notably where effect sizes are correlated across ethnic groups but not identical, where reweighting strategies can boost prediction performance using a mixed-ancestry training set. However, when applied to empirical data from the UK Biobank, relatively weak benefit from reweighting was observed. Rather, traits fell into two broad categories. For those in the first category (such as mean corpuscular volume), the most valuable training data was the small dataset from the minority group - here Black/Black British - alone (and where performance decreases by the inclusion of any individuals of the majority - here White - ethnicity). For those in the second category (such as height), the best performance in the underrepresented group was achieved by training in the much larger majority group. Although the first category was smaller, it included important biomarkers (for example, CRP) and even when the best strategy was to use the PGS trained in White individuals, the performance of the PGS in Black/Black British individuals was considerably worse than the performance of the same PGS in White individuals.

Our findings highlight the limitations of statistical corrections alone in reducing the disparities of PGS performance between ancestries. There are numerous sources of potential bias that adversely affect the application of an analytical pipeline to individuals of minority and/or underrepresented groups, including problem selection, data collection, outcome definition, model development and lack

of real-world impact assessments and considerations [10]. Therefore, statistical solutions must be considered in combination with efforts to address the global health research funding gap [42], diversity of the bioinformatics workforce [17] and assessing the impact and translation of analyses to real-world data [30] in addition to efforts to increase the number of non-European participants in genetic research. Each of these partial solutions, in combination, provide essential contributions to reduce and remove opportunities for negative effects on health inequalities, particular amongst those from different ethnic backgrounds [32].

It is becoming increasingly clear that the vast disparities in PGS performance can only be bridged by collecting more samples from minority and under-represented groups. Perhaps counter-intuitively, with more diverse data, statistical tools such as importance reweighting may eventually play a more important role as we seek to boost predictive utility by using all the available data. Reweighting strategies have the benefit of overcoming the lack of universal definitions of race, ethnicity and ancestry, which causes considerable confusion and imprecision [27, 28]. The categorisation of individuals into discrete ethnic groups to explain differences between behaviours and exposures may be unhelpful [13], whereas continuous representations of genetic ancestry, such as genetic principal components (Figure 6) enable identification of areas of genetic ancestry where performance is stronger or weaker as well as how the trait itself varies across ancestry. Ultimately, approaches to genetic prediction must acknowledge both the many similarities of human biology, but also the differences in history, cultural heritage, exposure, and behaviour that can lead to certain factors being of greater relevance for particular groups of individuals.

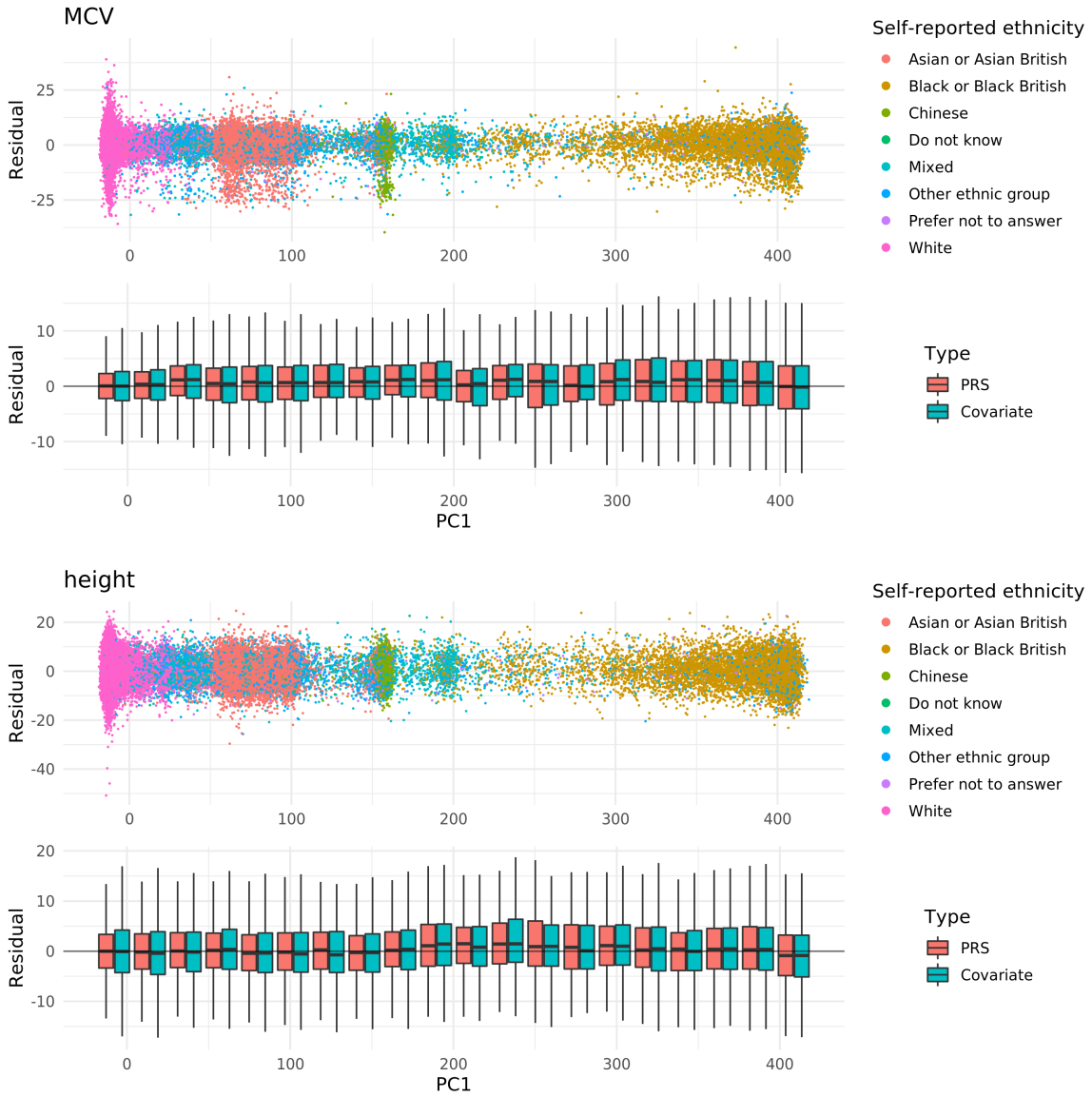

Figure 6: Residuals of a White PGS for mean corpuscular volume (top) and height (bottom) against the first genetic principal component. In the top panel, each dot corresponds to a single individual in the test set, coloured by self-reported ethnicity. The bottom panel shows box plots of the residuals for model 11 (PGS) and model 12 (Covariate), grouped into twenty bins across the range of the first genetic principal component. The boxes represent interquartile ranges, and the whiskers represent 1.5 times these ranges. Points outside the range of the whiskers are not shown. The changing explanatory power of PGS can be seen from the difference between the distribution of residuals in the covariate-only and PGS models. The distribution of the covariate-only residuals demonstrates how the trait variance varies by genetic ancestry and the location of the median residuals indicate how trait means vary by ancestry.

Materials and Methods

PGS Estimation using the LASSO

To construct PGS from full genotype data, we use L1-penalised regression, also known as the LASSO [39]. The LASSO has previously been used in the context of genetic prediction (see for example [44, 31, 34, 35] and references therein) and is suitable for high-dimensional problems where one expects the number of non-zero predictors to be small relative to the number of total predictors. Although there exist other full genotype PGS methods such as linear mixed models (see e.g. [46]), we focus on the LASSO largely for its computational efficiency. We provide an analysis of predictive performance versus sample size, focusing on differences in predictive performance between European ancestry individuals and African and Afro-Caribbean ancestry.

We first briefly recap the LASSO algorithm for constructing polygenic scores. Let n be the number of individuals in the training set. Denote $X \in \mathbb{R}^{n \times q}$ to be the matrix of q covariates, G to be the $n \times p$ genotype matrix, and y to be the n-vector of observed phenotype values. We assume a linear relationship between the phenotype and predictors,

$$y = X\theta + G\beta + \epsilon, \tag{3}$$

where θ is a q-vector of covariate effects, β is a p-vector of variant effects, and ϵ is an environmental noise term with mean 0.

The LASSO aims to minimise the objective function,

$$L(\theta, \beta) = ||y - X\theta - G\beta||_2^2 + \lambda||\beta||_1, \tag{4}$$

with respect to (θ, β), where λ is a regularisation parameter. The purpose of the $||\beta||_1$ term is to penalise large values of $|\beta|$ and thus encourage *sparse* solutions, that is, solutions with a relatively low number of non-zero coefficients. Note that the parameters θ are not penalised. The choice of λ controls the degree of penalisation with higher values of λ encouraging smaller values of β. To select λ, we use 5-fold (on real data) or 10-fold (on simulated data) cross-validation. We optimise (4) using the R packages `glmnet` [14] and `snpnet` [35] on simulated and real data respectively. The `snpnet` package is an extension of `glmnet` designed to interface directly with the PLINK software [8, 7] to handle large-scale single nucleotide polymorphism datasets such as the genotype data in UK Biobank.

Multiple-ancestry Datasets

The above description assumes that the genetic effects β are the same for all individuals. While this assumption may be reasonable within a single ancestry or ethnic group, it is unlikely to hold more generally. We consider a hypothetical model that allows genetic effects to vary among individuals. To this end, we introduce a low-dimensional latent variable z representing the ancestry of a given individual. For an individual i,

$$y_i = x_i^T\theta(z_i) + g_i^T\beta(z_i) + \epsilon_i, \tag{5}$$

where $(\theta(\cdot), \beta(\cdot))$ are now *functions* parameterised by z. The only assumption we make about these functions is one of *smoothness*, so that individuals with similar ancestry have similar covariate and genetic effects. Otherwise, we do not make any explicit statements regarding the form of the functions.

In this paper, we consider the special case with $J = 2$ ancestry groups, such that $z_i \in \{1, 2\}$. We denote $(\theta(z_i), \beta(z_i)) = (\theta^{(j)}, \beta^{(j)})$ for an individual i in ancestry group j. Using superscripts to denote the ancestry group (for example, $y^{(j)}$ is the vector of observed phenotypes for group j),

$$y^{(j)} = X^{(j)}\theta^{(j)} + G^{(j)}\beta^{(j)} + \epsilon^{(j)}, \quad j = 1, 2. \tag{6}$$

Denoting n_j to be the number of individuals of ancestry j in the training set, we assume $n_1 > n_2$ to reflect a European-ancestry dominated training set. Applying the standard LASSO to such a training set, we would expect the estimate $(\hat{\beta}, \hat{\theta})$ to be closer to $(\theta^{(1)}, \beta^{(1)})$ than to $(\theta^{(2)}, \beta^{(2)})$. All else being equal, this would result in better predictive performance for test individuals in group 1.

Importance Reweighting

We return for a moment to the more general model in Equation (5). Suppose we wish to construct a polygenic score for a test set with ancestry distribution $\pi^{TE}(z)$ given a training set of observations $(y_i, x_i, g_i)_{i=1}^n$ from individuals with ancestry $z_1, \ldots, z_n$, how should one obtain a single estimate for (θ, β)? Our proposed approach is to use *importance reweighting*, assigning each of the training individuals a weight w_i that quantifies the 'similarity' of the ancestry variable z_i to the test set ancestries, with individuals who have a similar ancestry to the test set assigned a higher weight. More concretely, assuming an ancestry distribution $\pi^{TR}(\cdot)$ on the training set, we set

$$w_i = \frac{\pi^{TE}(z_i)}{\pi^{TR}(z_i)}. \tag{7}$$

Note that we do not have access to the distributions π^{TR}, π^{TE} so in practice we must approximate (7). Given weights w_i, to estimate (θ, β), we minimise the following *weighted* LASSO objective function:

$$L_\lambda(\theta, \beta) = \sum_{i=1}^{n} w_i \left(y_i - x_i^T\theta + g_i^T\beta\right)^2 + \lambda||\beta||_1. \tag{8}$$

For the special case with two ancestry groups (Eqn. 6), we consider importance weights w_1, w_2 such that $w_2 \geq w_1$ with the aim of improving estimates for group 2, i.e. the underrepresented group. Specifically, we use weights of the form:

$$w_i = \left(\frac{n_1 + n_2}{n_i}\right)^\gamma, \tag{9}$$

where $\gamma \in [0, 1]$ is a hyperparameter that controls the degree of reweighting. We normalised the weights so that $n_1 w_1 + n_2 w_2 = n_1 + n_2$. With these weights, we thus minimise the objective function,

$$L_\gamma(\theta, \beta) = w_1||y^{(1)} - X^{(1)}\theta - G^{(1)}\beta||_2^2 + w_2||y^{(2)} - X^{(2)}\theta - G^{(2)}\beta||_2^2 + \lambda||\beta||_1, \tag{10}$$

with respect to (θ, β).

UK Biobank

The UK Biobank is a large-scale cohort study with genomic and phenotypic data collected on approximately 500 000 individuals aged 40-69 at the time of recruitment [4]. Participants reported their ethnic backgrounds, with the majority (94%) self-identifying as a subgroup within the broad-level group 'White'. There are approximately 7 000 (1.7%) individuals with a self-reported ethnic background of Black/Black British. To harmonise terms across analyses on both UK Biobank data and the simulated data used in this study, the terms self-reported ethnicity and ancestry are used interchangeably.

We investigated a range of quantitative traits that are likely to be influenced by or correlated with a number of polygenic diseases. These included three common anthropometric traits (height, body mass index, waist circumference) and twelve blood-related traits (systolic blood pressure, diastolic blood pressure, C-reactive protein, platelet count, white blood cell count, red blood cell count, haemoglobin concentration, haematocrit percentage, mean corpuscular volume, mean corpuscular haemoglobin, lymphocyte count, and monocyte count). We used genotypes from individuals self-identifying as "White" or "Black/Black British" and who satisfied certain quality-control metrics, namely those individuals used in PCA calculations of [4] and who were not identified as displaying sex chromosome aneuploidy. To control for population structure, we included the following covariates in our model: age, sex, the first ten genetic principal components (PCs), and interactions between sex and the ten genetic PCs.

Construction of Training Sets

To assess the effect of sample size and composition on PGS performance, we used subsets of the above data as training sets, controlling the number of White and Black/Black British individuals. The remaining individuals were then used as a held-out test set. For each trait, we constructed three types of training sets using individuals with non-missing data for that particular trait. The first two types of training sets were single-ancestry datasets, one consisting solely of White individuals and the other consisting solely of Black/Black British individuals.

1. *White single-ancestry*
   This set consisted of a randomly-sampled 80% of the quality-controlled White individuals (approximately 300 000 with 3/4 of these used for training and the remaining 1/4 used as a validation set to select the regularisation parameter λ).

2. *Black single-ancestry*
   This set consisted of a randomly-sampled 80% of the quality-controlled Black/Black British individuals, corresponding to approximately 5 500 individuals. Given the relatively small number of individuals, we used 5-fold cross-validation to select λ.

3. *Dual-ancestry*
   This set consisted of both Black/Black British individuals and White individuals. The basic form of this training set was made up of the Black training set described above, combined with White individuals so that the proportion of Black individuals was 10%. The White individuals were matched to the Black/Black British individuals on age and sex. We also considered variants

of this training set by removing a proportion of either the White individuals or the Black/Black British individuals (see Results for more details).

To each dataset, we applied a minor allele frequency threshold of 0.1% and a missing genotype call rate filter of 5%. Note that, as a result, the sets of variants generally differed by training set. For the single-ancestry datasets, we used the standard, unweighted LASSO to construct the PGS. For the dual-ancestry datasets, we used the weighted LASSO with $\gamma = 0, 0.2, 0.4, 0.6, 0.8, 1$. Note that $\gamma = 0$ corresponds to the unweighted LASSO.

Predictive Accuracy

We evaluated predictive accuracy of each PGS using individuals that were not included in the corresponding training sets. To assess the genetic predictive accuracy of a PGS, we calculated the partial r^2 attributable to the PGS, relative to a covariate-only model, following Martin et al. [25]. Specifically, we fit the following nested linear models,

$$Y = X\theta_1 + \hat{Y}\eta + \epsilon_1 \tag{11}$$

$$Y = X\theta_2 + \epsilon_2, \tag{12}$$

where X is the covariate matrix and $\hat{Y}$ is the vector consisting of polygenic scores for each individual in the test set. We defined the partial r^2 to be the difference in r^2 between models 11 and 12. To obtain a more reliable estimate of predictive accuracy, we performed 5-fold cross-validation. That is, we repeated five times the process of training set construction and PGS estimation to obtain five estimates of partial r^2. Finally, we took the mean across these five folds as our measure of predictive accuracy.

Simulation Study

To evaluate the effect of importance reweighting on PGS performance under various settings, we undertook a simulation study. We used the simulation engine `msprime` [18] in the standard library of population genetic simulation models `stdpopsim` v0.1.2 [1] to generate African and European genotypes, following a similar simulation framework to that used by Martin et al. [26]. For each ancestry group, we generated 200 000 genotypes for chromosome 20, based on a three-population 'out-of-Africa' demographic model [16], using a mean mutation rate of 1.29×10^{-8} and a recombination map of GRCh37. Note that this model also generates genotypes for East Asian individuals but we do not use these in our analysis. To reduce the computational burden, we used only the first 10% of the chromosome, applying a minor allele frequency threshold of 0.1% to each ancestry group separately, resulting in a total of 7 563 variants.

We simulated phenotypes from these genotypes assuming a normal linear model,

$$y^{(j)} = G^{(j)}\beta^{(j)} + \epsilon^{(j)}, \quad j = 1, 2, \tag{13}$$

where $\epsilon^{(j)} \sim \mathcal{N}(0, \sigma^2)$. The noise variance parameter σ^2 controls the SNP heritability h_j^2 for each ancestry group and was chosen to yield an average SNP heritability of 0.5 across the groups. To reflect

the sparsity of genetic effects, we randomly selected $p_0 = 100$ of the $p = 7,563$ variants to be causal. Motivated by Trochet et al. [41], we drew the causal effect sizes from a bivariate normal distribution to model the similarity of genetic effects between ancestries. Denoting $\mathcal{C} = \{k_1, \ldots, k_{100}\}$ to be the indices of the causal SNPs, and $\beta_C^{(j)} = (\beta_{k_1}^{(j)}, \ldots, \beta_{k_{100}}^{(j)})$ we have the equation

$$\begin{pmatrix} \beta_C^{(1)} \\ \beta_C^{(2)} \end{pmatrix} \sim \mathcal{N}\left(0\ , \quad \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \tag{14}$$

where ρ is the correlation between the ancestry-specific genetic effects. We set $\beta_i^{(1)} = \beta_i^{(2)} = 0$ for $i \notin C$. For each value of $\rho = 0.5, 0.6, \ldots, 0.9$, we repeated the above procedure 50 times to generate 50 quantitative traits.

To investigate the effect of sample size, for each trait we created five training sets with $n_{EUR} = 1\,800$ randomly selected individuals of European ancestry and $n_{AFR} = 200, 450, 771, 1\,200, 1\,800$ randomly selected individuals of African ancestry. For each of these training sets, we computed weights $w_{EUR}^{(\gamma)} = (10/9)^\gamma$, $w_{AFR}^{(\gamma)} = 10^\gamma$ and where γ is a hyperparameter controlling the degree of reweighting. Note that $\gamma = 1$ corresponds to inverse proportion reweighting when $n_{AFR} = 200$ and $n_{EUR} = 1\,800$. We normalised the weights to ensure that their sum equalled n, in line with the unweighted case.

To assess predictive performance for each ancestry, we constructed an African test set and a European test set by randomly selecting 2 000 individuals from each ancestry out of those not included in the training set. We used the proportion of variance explained, denoted r^2, as the measure of predictive performance,

$$r^2 = \frac{\widehat{\mathrm{Var}}(G\hat{\beta})}{\widehat{\mathrm{Var}}(y)}, \tag{15}$$

where $\widehat{\mathrm{Var}}$ denotes the sample variance. Note that this is equivalent to the partial r^2 relative to an intercept-only covariate model. For $\rho = 0.5, 0.6, \ldots, 0.9$, we repeated the above process 20 times with different randomly selected individuals in the training and test sets to calculate a mean r^2 for each trait.

Acknowledgements

BCLL and CM were supported by the UK Engineering and Physical Sciences Research Council through the Bayes4Health programme [Grant number EP/R018561/1]. BCLL also gratefully acknowledges funding from Jesus College, Oxford. CCH was also supported by The Alan Turing Institute, Health Data Research UK, the Medical Research Council UK, and AI for Science and Government UK Research and Innovation (UKRI). GM was supported by the Wellcome Trust [Grant Number 100956/Z/13/Z] and the Li Ka Shing Foundation. This research has been conducted using data from UK Biobank, a major biomedical database, under application 12788. The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

References

[1] Jeffrey R Adrion et al. "A community-maintained standard library of population genetic models". In: *eLife* 9 (June 2020). Ed. by Graham Coop et al., e54967. ISSN: 2050-084X. DOI: 10.7554/eLife.54967. URL: https://doi.org/10.7554/eLife.54967.

[2] Amy R Bentley et al. "Multi-ancestry genome-wide gene–smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids". In: *Nature genetics* 51.4 (2019), pp. 636–648.

[3] Brielin C Brown et al. "Transethnic genetic-correlation estimates from summary statistics". In: *The American Journal of Human Genetics* 99.1 (2016), pp. 76–88.

[4] Clare Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726 (2018), pp. 203–209.

[5] Christopher S. Carlson et al. "Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study". In: *PLOS Biology* 11.9 (Sept. 2013), pp. 1–11. DOI: 10.1371/journal.pbio.1001661. URL: https://doi.org/10.1371/journal.pbio.1001661.

[6] Taylor B. Cavazos and John S. Witte. "Inclusion of Variants Discovered from Diverse Populations Improves Polygenic Risk Score Transferability". In: *bioRxiv* (2020). DOI: 10.1101/2020.05.21.108845. eprint: https://www.biorxiv.org/content/early/2020/10/05/2020.05.21.108845.full.pdf. URL: https://www.biorxiv.org/content/early/2020/10/05/2020.05.21.108845.

[7] Christopher C Chang and Shaun M Purcell. *PLINK 1.9*. www.cog-genomics.org/plink/1.9/.

[8] Christopher C Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *Gigascience* 4.1 (2015), s13742–015.

[9] Nilanjan Chatterjee, Jianxin Shi, and Montserrat García-Closas. "Developing and evaluating polygenic risk prediction models for stratified disease prevention". In: *Nature Reviews Genetics* 17.7 (2016), p. 392.

[10] Irene Y Chen et al. "Ethical machine learning in health". In: *arXiv preprint arXiv:2009.10576* (2020).

[11] H3Africa Consortium et al. "Enabling the genomic revolution in Africa". In: *Science* 344.6190 (2014), pp. 1346–1348.

[12] L Duncan et al. "Analysis of polygenic risk score usage and performance in diverse human populations". In: *Nature communications* 10.1 (2019), pp. 1–9.

[13] Morris W Foster and Richard R Sharp. "Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity". In: *Genome research* 12.6 (2002), pp. 844–850.

[14] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software, Articles* 33.1 (2010), pp. 1–22. ISSN: 1548-7660. DOI: 10.18637/jss.v033.i01. URL: https://www.jstatsoft.org/v033/i01.

[15] Kelsey E Grinde et al. "Generalizing polygenic risk scores from Europeans to Hispanics/Latinos". In: *Genetic epidemiology* 43.1 (2019), pp. 50–62.

[16] Ryan N. Gutenkunst et al. "Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data". In: *PLOS Genetics* 5.10 (Oct. 2009), pp. 1–11. DOI: 10.1371/journal.pgen.1000695. URL: https://doi.org/10.1371/journal.pgen.1000695.

[17] Bas Hofstra et al. "The Diversity–Innovation Paradox in Science". In: *Proceedings of the National Academy of Sciences* 117.17 (2020), pp. 9284–9291.

[18] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. "Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes". In: *PLoS Comput Biol* 12.5 (May 2016), pp. 1–22. DOI: 10.1371/journal.pcbi.1004842. URL: http://dx.doi.org/10.1371%2Fjournal.pcbi.1004842.

[19] Amit V Khera et al. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations". In: *Nature genetics* 50.9 (2018), pp. 1219–1224.

[20] Joshua W. Knowles and Euan A. Ashley. "Cardiovascular disease: The rise of the genetic risk score". In: *PLOS Medicine* 15.3 (Mar. 2018), pp. 1–7. DOI: 10.1371/journal.pmed.1002546. URL: https://doi.org/10.1371/journal.pmed.1002546.

[21] Luke R Lloyd-Jones et al. "Improved polygenic prediction by Bayesian multiple regression on summary statistics". In: *Nature communications* 10.1 (2019), pp. 1–11.

[22] Paige Maas et al. "Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States". In: *JAMA oncology* 2.10 (2016), pp. 1295–1302.

[23] Timothy Shin Heng Mak et al. "Polygenic scores via penalized regression on summary statistics". In: *Genetic epidemiology* 41.6 (2017), pp. 469–480.

[24] Carla Márquez-Luna et al. "Multiethnic polygenic risk scores improve risk prediction in diverse populations". In: *Genetic epidemiology* 41.8 (2017), pp. 811–823.

[25] Alicia R Martin et al. "Clinical use of current polygenic risk scores may exacerbate health disparities". In: *Nature genetics* 51.4 (2019), pp. 584–591.

[26] Alicia R Martin et al. "Human demographic history impacts genetic risk prediction across diverse populations". In: *The American Journal of Human Genetics* 100.4 (2017), pp. 635–649.

[27] Iain Mathieson and Aylwyn Scally. "What is ancestry?" In: *PLoS Genetics* 16.3 (2020), e1008624.

[28] Tesfaye B Mersha and Tilahun Abebe. "Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities". In: *Human genomics* 9.1 (2015), p. 1.

[29] Anna C Need and David B Goldstein. "Next generation disparities in human genomics: concerns and remedies". In: *Trends in Genetics* 25.11 (2009), pp. 489–494.

[30] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453.

[31] Sebastian Okser et al. "Regularized Machine Learning in the Genetic Prediction of Complex Traits". In: *PLOS Genetics* 10.11 (Nov. 2014), pp. 1–9. DOI: 10.1371/journal.pgen.1004754. URL: https://doi.org/10.1371/journal.pgen.1004754.

[32] Roseann E Peterson et al. "Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations". In: *Cell* 179.3 (2019), pp. 589–603.

[33] Alice B Popejoy and Stephanie M Fullerton. "Genomics is failing on diversity". In: *Nature News* 538.7624 (2016), p. 161.

[34] Florian Privé, Hugues Aschard, and Michael G. B. Blum. "Efficient Implementation of Penalized Regression for Genetic Risk Prediction". In: *Genetics* 212.1 (2019), pp. 65–74. ISSN: 0016-6731. DOI: 10.1534/genetics.119.302019. eprint: https://www.genetics.org/content/212/1/65.full.pdf. URL: https://www.genetics.org/content/212/1/65.

[35] Junyang Qian et al. "A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank". In: *PLOS Genetics* 16.10 (Oct. 2020), pp. 1–31. DOI: 10.1371/journal.pgen.1009141. URL: https://doi.org/10.1371/journal.pgen.1009141.

[36] Armin P Schoech et al. "Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection". In: *Nature communications* 10.1 (2019), pp. 1–10.

[37] Seth A Sharp et al. "Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis". In: *Diabetes Care* 42.2 (2019), pp. 200–207.

[38] Yuval B. Simons et al. "A population genetic interpretation of GWAS findings for human quantitative traits". In: *PLOS Biology* 16.3 (Mar. 2018), pp. 1–20. DOI: 10.1371/journal.pbio.2002985. URL: https://doi.org/10.1371/journal.pbio.2002985.

[39] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[40] Ali Torkamani, Nathan E Wineinger, and Eric J Topol. "The personal and clinical utility of polygenic risk scores". In: *Nature Reviews Genetics* 19.9 (2018), pp. 581–590.

[41] Holly Trochet et al. "Bayesian meta-analysis across genome-wide association studies of diverse phenotypes". In: *Genetic epidemiology* 43.5 (2019), pp. 532–547.

[42] D Vidyasagar. "Global notes: the 10/90 gap disparities in global health research". In: *Journal of Perinatology* 26.1 (2006), pp. 55–56.

[43] Bjarni J Vilhjálmsson et al. "Modeling linkage disequilibrium increases accuracy of polygenic risk scores". In: *The american journal of human genetics* 97.4 (2015), pp. 576–592.

[44] Patrik Waldmann et al. "AUTALASSO: an automatic adaptive LASSO for genome-wide prediction". In: *BMC bioinformatics* 20.1 (2019), pp. 1–10.

[45] Jian Zeng et al. "Signatures of negative selection in the genetic architecture of human complex traits". In: *Nature genetics* 50.5 (2018), pp. 746–753.

[46] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. "Polygenic Modeling with Bayesian Sparse Linear Mixed Models". In: *PLOS Genetics* 9.2 (Feb. 2013), pp. 1–14. DOI: 10.1371/journal.pgen.1003264. URL: https://doi.org/10.1371/journal.pgen.1003264.