Statistical and ethical considerations for the use of ancestry in polygenic risk scores

Alyna T. Khan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2024

Reading Committee:
Stephanie M. Fullerton, Chair
Alexander P. Reiner
Alison Fohner

Program Authorized to Offer Degree:
Public Health Genetics

University of Washington

Abstract

Statistical and ethical considerations for the use of ancestry in polygenic risk scores

Alyna T. Khan

Chair of the Supervisory Committee:
Stephanie M. Fullerton
Department of Bioethics and Humanities

The polygenic risk score (PRS) is an emerging tool in precision medicine that promises to provide individualized genetic risk estimates for developing a given trait or disease of interest. PRS have already shown improvement in risk assessments for certain conditions (e.g., cardiovascular disease) when used in conjunction with existing clinical risk assessment. However, many PRS have worse predictive accuracy for individuals of mixed or non-European genetic ancestry due to the overrepresentation of individuals of European descent in genomic research, leading to concerns of contributing to healthcare inequity. In this study, I explore the analytical and ethical challenges of addressing equity concerns in the context of using concepts of ancestry in PRS. First, I assessed the concordance of two PRS models for red blood cell traits in individuals with African ancestry to investigate whether both models identify the same individuals at high risk. The study revealed analytical challenges involved in working with underrepresented populations and small sample sizes with respect to PRS accuracy and reliability. Second, I conducted an ethical analysis focused on understanding the conceptualization of equity with respect to PRS and illustrated the practical constraints in achieving goals of equitable PRS development given the infrastructural constraints in genomic

research practice. Finally, I interviewed experts in the field who are actively developing PRS for diverse populations to capture their experiences, challenges, and outlook on the use of race, ethnicity, and ancestry in PRS. I conclude by summarizing results of this study and offering suggestions for moving forward in addressing the urgent issue of equity in PRS.

Table of Contents

List of figures

List of tables

Acknowledgements

The journey of working on this PhD has been one full of surprises, challenges, and growth, and I would not have been able to complete it without the support and guidance of many different people. I am grateful to my supervisory committee for helping me design and develop my projects, and for their attentiveness and critical feedback on my final product. I am especially thankful to my Committee Chair, Malia Fullerton, for showing me how to be a researcher with integrity, strength, and conviction, and for encouraging me to bring my whole self into this work. I am grateful for her unwavering support and unparalleled mentorship over the years, and for the wisdom and compassion she has shared with me.

I also thank the students, faculty, and staff of the Institute for Public Health Genetics. Annique Atwater has been a constant source of support as I navigated the logistics of the PhD, especially through the turbulent times of the global COVID pandemic. My PHG classmates have been wonderful peers and friends throughout this journey: Priyanka, Mckenna, Hanley, Diane, and many more. I am also deeply grateful to my colleagues at the GAC who have been enthusiastically supportive, encouraging, and inspiring throughout my time in the PhD. Working at the GAC has been a grounding experience where I had the pleasure and privilege to work alongside incredibly talented, thoughtful leaders in Genetics and Biostatistics. I want to especially thank Adrienne, Matt, Sarah, and Stephanie for their guidance, patience, and numerous insights they shared with me.

This research would not have been possible without the generosity of the research participants who volunteered their time and shared their experiences with me for my study via interviews. I

also thank the numerous participants who have donated their biosamples in support of this grand collective effort to understand ourselves in this world through science. Additionally, I am grateful to the many scholars before me whose work has taught me and guided me through my own research journey.

Along this road to obtaining the PhD, I was accompanied by the most supportive, loving, and dedicated community of family and friends I could possibly imagine. Completing this journey would not be possible without them. I thank all of my friends for their unyielding love, encouragement, and companionship through this time of my life: Abby, Anita, Anum, Becky, Binks, Danielle, Farty, Grace Ann, Sela, and many, many other wonderful people. I thank my Mom, Meher Khan, for her unconditional love, displayed in innumerable ways, including countless last minute cross country flights to visit me and be a source of strength in times of great uncertainty. I thank my Dad, my brothers, sisters-in-law, and my nieces and nephews for all the excitement, laughter, and fun you share with me. To all my family and friends, you make my life full and rich. Thank you for the joy and happiness you continually bring to my life.

March 27, 2024

Introduction

Polygenic risk scores (PRS) are an emerging tool in genomics and precision medicine that estimate an individual's genetic predisposition to developing complex traits (e.g., height) or diseases (e.g., type 2 diabetes). It is well-recognized that common, complex diseases are influenced by social determinants of health (e.g., diet, physical activity, built environment, structural harms) and environmental exposures (Braveman et al., 2014); but genomic studies have also shown that they can be influenced by many different genetic variants, each with small impacts on a given disease, but cumulatively can indicate whether an individual is at higher susceptibility to developing the disease. As such, PRS have received much attention in the genomics community as a potential tool that can aid in individualized clinical risk estimation of developing disease. The clinical use of PRS could aid in early detection of disease, predict response to pharmacological treatment, guide targeted therapies, and overall improve disease risk estimation. Widely accessible and predictively accurate PRS can be used to screen early in life for diseases and support preventative healthcare measures. Indeed, studies have shown that PRS for cardiovascular diseases provide improvement in risk estimation when used alongside existing clinical risk assessments (Khera et al., 2018). For example, PRS for coronary heart disease (CAD) has been shown to provide additive risk assessment benefits to traditional clinical risk factors, measured by a change in the C-statistic - or discriminatory power of a predictive model - which improved by 0.045 over 11 traditional risk factors alone (Hindy et al., 2020). In addition, PRS for CAD has aided in clinical decision-making by identifying those at high risk for experiencing a stroke event or myocardial infarction, and guiding treatment such as early statin use or surgical/procedural intervention (Klarin et al., 2022).

However, the variability in PRS accuracy across individuals of different (genetic and non-genetic) ancestry backgrounds has raised serious concerns of exacerbating existing healthcare inequities and presents a major hurdle in the clinical implementation of PRS. PRS make use of results from genome-wide association studies (GWAS) for risk estimation, and they perform better in individuals whose ancestry backgrounds are similar to those of the population in the GWAS used to inform the PRS. For example, Ding et al. (2023) demonstrate that the greater the difference in genetic ancestry between the training data and the scored individuals, the greater the decline in PRS predictive accuracy, with individuals most genetically distant from population of the training data having at least 14% lower accuracy relative to individuals with a shorter genetic distance to that of the training population. PRS accuracy has been shown to be as low as having 25% as predictive a score as those who are most genetically similar to the original study population used to derive the score (Martin et al., 2019; Ding et al., 2023). Historically, genomics research has predominantly focused on individuals of European ancestry, with over 80% of GWAS research conducted in European ancestry populations (Popejoy & Fullerton, 2016). Meanwhile, as of 2019, ~79% of all GWAS participants according to the GWAS Catalog were of European-descent (Martin et al., 2019), while only 2.8% of studies were done in populations of African ancestry (Morales et al., 2018). Similarly, reference panels (a frequently used tool in genomics research for operations such as imputation or genetic ancestry estimation) such as the Haplotype Reference Consortium (HRC), represent only a minority of diverse ancestry populations (Fatumo, 2022; O'Connell, 2021). Thus, given the overrepresentation of European-ancestry individuals in genomics research, PRS currently perform better for people with European ancestry than for those who have non-European or mixed ancestry.

In response to this notable disparity, the genomics community is working to correct the eurocentric bias by improving diverse representation in GWAS and PRS analyses (C. M. Lewis & Vassos, 2020; Polygenic Risk Score Task Force of the International Common Disease Alliance et al., 2021). There has been a greater focus on recruiting participants from underrepresented populations for genomic research (e.g., All Of Us), improvement in representation in reference panels (e.g., GenomeAsia), and encouragement of genomics consortia focused on diverse populations (e.g., PAGE, PRIMED). An increase in representation across the different components of genomic research (e.g., study cohorts, reference panels, study design) can enable variant discovery in diverse populations, which can then better inform what variants to include in a PRS. In doing so, the genomics community hopes to generate knowledge and develop PRS that are of equitable benefit to everyone.

This dissertation examines what equity might look like for an emerging tool that incorporates some measure of genetic ancestry in order to work equally accurately for everyone. Scholars have written about the inscription of biological race through medical devices like the spirometer, which has a “race correction” feature that allows health care providers to indicate the race of the patient on a device and adjust the results depending on the category selected (Braun, 2015). Currently, the clinical application of PRS would require healthcare providers to ascertain the genetic ancestry information of their patient in order to determine if and how PRS results may be applicable to them. In practice, genetic ancestry risks being conflated with social group membership (e.g., race, ethnicity, cultural background, religion, etc), especially since it is with social identity that people interact with the healthcare system and with one another (Bentz et al., 2024).

How are polygenic risk scores developed?

PRS use the results of genome-wide association studies (GWAS) to calculate one's genetic risk of developing a given trait or disease. GWAS results used in a PRS calculation consist of the set of single nucleotide polymorphisms (SNPs) that were observed to be associated with the trait of interest, the variant effect sizes - or measured impact on the trait - and the p-values of those variants. The score is then calculated by summing across the product of the effect sizes of the set of variants associated with the trait and the variant dosage (i.e. the number of copies of the variant) in a given individual's genome (Equation 1).

$$PRS_i = \sum_j^M \hat{\beta}_j \times dosage_{ij}$$

(Equation 1)

Reference: ASHG

where,
i = the individual
j = the variant
ß_*hat* = effect size of variant, *j*
dosage = number of copies of variant, *j* in the genotype of individual, *i*

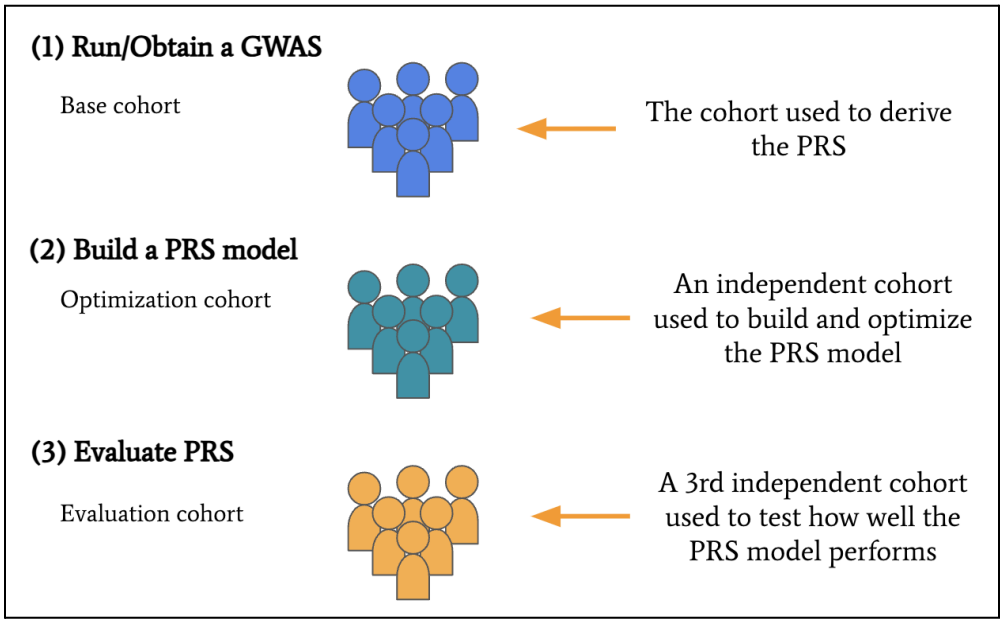
While the final computation of the risk score itself is fairly straightforward, the calculation of the PRS requires several intermediate steps that can be quite complex. These intermediate steps largely involve making adjustments to the set of SNPs and their effect size estimates to ensure that the variants identified from a GWAS conducted in a given population can accurately predict risk in another population. Variants associated with disease and measures of their effect sizes can vary due to differences in linkage disequilibrium (LD) and allele frequencies across populations, also known as population structure. LD refers to the tendency of certain variants being inherited

together; if it is not properly accounted for, a GWAS might identify many variants that appear to be associated with a given disease, but in fact are not. They may even obfuscate the variant that is truly associated with the disease (i.e. causal variant) because a GWAS may identify a lead variant that is in LD with the causal variant in some populations but not in others due to differing patterns of LD across populations. Accounting for population structure in PRS calculations may necessitate an external dataset (e.g., a reference panel) to approximate variant allele frequencies and LD patterns between SNPs in a given population. But defining what "a given population" is comes with many challenges, as will be discussed in more detail in later chapters. Finally, another complicating factor in polygenic risk scoring is the undoubted influence of non-genetic contributors to risk of disease, such as social determinants of health. Capturing, quantifying and incorporating social and environmental measures into polygenic risk models (as well as in other clinical risk prediction models) remains an active area of development, as the interaction of social and environmental factors with PRS is quite understudied.

The general process of generating a PRS model broadly includes two, sometimes three, phases: (1) Obtaining GWAS summary statistics from a given population (i.e. base cohort) to acquire a set of SNPs, effect sizes, and P-values; (2) Building and optimizing a PRS model in an independent population cohort (i.e. optimization cohort), which consists of the final set of SNPs, their (typically adjusted) effect size estimates, and thresholds for risk; (3) Evaluating the PRS model in a separate population (i.e. evaluation cohort) to assess PRS performance (Fig. A). While step 2 is not strictly required, it is oftentimes recommended in order to generate an optimized and least biased score, if the data is available. Each step in PRS development ideally uses separate, independent population cohorts to minimize risk of overfitting. However, due to

the different factors mentioned above that can impact PRS accuracy, it is important to select population cohorts with similar genetic (e.g., population structure) and non-genetic characteristics (e.g., environmental exposures).

Figure A. Overview of steps and respective cohorts involved in a PRS calculation



Deciding which population cohorts to use in a PRS analysis typically relies on the population descriptors available in study documentation. While describing study populations can be highly variable and can take many different forms (e.g., cultural, linguistic, tribal, national, regional, ethnic etc), standard practice for studies (conducted in the US) has often relied on the use of common race and ethnicity categories set by the Office of Management and Budget (OMB). There are a number of reasons for why scientists may want to standardize population descriptors to a particular system of classification, such as the desire to combine and harmonize multiple datasets to increase the statistical power of study (Stilp et al., 2021). But while standardization of data is typically seen as good scientific practice that supports replicability of studies,

standardization of population descriptors can result in loss of information and faulty scientific conclusions.

Race, ancestry, and genetics

"According to a story that is probably apocryphal but nonetheless telling, an American journalist once asked the late Papa Doc Duvalier of Haiti what percentage of the Haitian population was white. Duvalier's answer, astonishingly enough, was 'Ninety-eight percent.' The startled American journalist was sure he had either misheard or been misunderstood, and put his question again. Duvalier assured him that he had heard and understood the question perfectly well, and had given the correct answer. Struggling to make sense of this incredible piece of information, the American finally asked Duvalier: 'How do you define white?' Duvalier answered the question with a question: 'How do you define black in your country?' Receiving the explanation that in the United States anyone with any black blood was considered black, Duvalier nodded and said, 'Well, that's the way we define white in my country.'"

- Barbara J. Fields, *Ideology and Race in American History*

It is well-recognized that race is a social construct, and many scientists in biomedical fields like genetics and genomics have published statements condemning racial biology ("ASHG Denounces Attempts to Link Genetics and Racial Supremacy," 2018). Historically, the alignment of race and biology was intricately tied to imperial and colonial-era arguments for (unfree) labor (Fields & Fields, 2022; Loomba, 2015). The idea that certain groups of people were naturally suited for particular types of work was strengthened by a biological rationale, and for centuries, science was used to demonstrate the natural existence of racial typologies. Scientific methods were developed with this ideology in mind, and scientific evidence was used to justify social status, exploitative labor practices, and discriminatory access to resources[1] (Braun, 2015; Loomba, 2015a; Roberts, 2011b).

[1] Racism produced the concept of race, and science has been used as a tool to legitimize the idea of races. Historically, scientific support has strengthened racism as an ideology. Racism impacts health outcomes, leading to observed racial health disparities. These disparities are then identified as important and/or interesting to examine, and studies reemploy race to do so (Fields, 2022; Roberts, 2011). While examining health (and other) outcomes by race (and other social identifiers of group membership) is crucial for understanding how systems are disproportionately burdening some individuals more than others, scientists must carefully consider when race is a relevant variable to include in a study and when it is not.

Understanding how race and racism have historically functioned in biomedical science can help with identifying their remnants in scientific practice and translation today. As mentioned earlier, PRS currently rely on some estimation of genetic ancestry, or some definition of a population group, in order to work equally well across populations. Genetic ancestry generally refers to biological information inherited from one's ancestors. In GWAS and PRS analyses, genetic ancestry typically represents population structure observed through allele frequency differences and LD patterns. However, genetic ancestry is not wholly inseparable from non-genetic understandings of ancestry and social identity because social and environmental factors can influence the genetic information that is inherited (Mathieson & Scally, 2020).

This complex relationship between social identifiers and genetic ancestry can muddy efforts to eliminate problematic uses of race, ethnicity, and genetic ancestry in research practice. For example, observed correlations between race and genetic ancestry can sometimes be used to verify or assign "population group" membership, particularly in cases of missing race or ethnicity data for a given sample (i.e., imputing race or ethnicity data). One reason we might see a correlation between race and genetic ancestry is that the definition of race (especially in the US) has historically been tied to ancestry (or "blood" as it is described in the quote above) (Brubaker, 2016; Loomba, 2015a). Racial classification is a turbulent system, made evident in the erratic creation or dissolution of racial categories and inconsistent application of race to an individual (e.g., "Mexican" was removed as a racial category in 1940 and Mexican individuals were considered White, unless they were Indigenous (Ortiz & Telles, 2012)). But race is also typically attached to phenotype in addition to ancestry, which legitimizes and sustains a given

racial category (Brubaker, 2016; Loomba, 2015a). So while associations between race and genetics can occur, the relationship is a reflection of the socio-political context in which that data was collected and is specific to the data it is observed in; correlation between race and genetic ancestry is not an idea that can be generalized to other individuals or populations nor generalized across time or place. There is ample evidence showing that racial categories are completely arbitrary; anybody can be white, as Papa Doc implies in the quote above, depending on the definition they choose to employ. Thus, having an understanding that racial classification is fluid renders practices that associate race with genetic ancestry (e.g., imputing race or ethnicity from genetic data) questionable at best. Yet, alignment of race (and other social identifiers) with genetic ancestry occurs in various sectors of genomic research practice, which I discuss more in later chapters.

Statistical and ethical considerations of using race, ethnicity, and ancestry in polygenic risk score development

PRS development is in its nascent stages, which offers a unique opportunity to closely examine how race, ethnicity, and (genetic and non-genetic) ancestry (REA) are being used. As study data in genomic research diversify, whether and how population structure is accounted for in GWAS and PRS becomes a central analytical challenge. These challenges have implications for the reproducibility and generalizability of findings, as well as their responsible clinical translation, such as with polygenic risk score (PRS) use. For example, an analyst may want to stratify a study population into racially-, ethnically-, or genetically- defined subgroups as a means of accounting for population structure. Stratifying in this way operates on the assumption that the most powerful association signal comes from variants in linkage disequilibrium (LD) with the causal

variant, requiring stratification by ancestry due to different LD patterns across populations. This analytical assumption is ethically salient because stratification-based results are less generalizable, they risk loss of information due to exclusion of small populations or missing social or genetic ancestry information, and may be difficult to translate into clinical applications, especially for patients who identify with multiple racial, ethnic, and ancestral backgrounds. As a result, there has been warranted focus on understanding ways to develop PRS that offer equitable benefit to everyone.

This research takes a 3-pronged approach to gain a holistic overview of what it means for genomic tools like PRS to be equitable. I pursue three different research aims in service of better understanding one main question: *what does equity look like for a tool that uses ancestry-stratified data in its estimations of risk for developing a given trait or disease?* I investigated three research questions in pursuit of this goal:

1. Understand the statistical and analytical challenges entailed by conducting GWAS and developing PRS in an underrepresented population.
2. Examine the ethical concerns in the effort to build equitable PRS.
3. Explore researcher perspectives on the analytical and ethical trade-offs of using social identifiers like race, ethnicity, and ancestry in the development and translation of PRS.

Many studies have focused on the effort to build PRS that work equally well across populations with different (genetic and non-genetic) ancestry backgrounds, as PRS developed using data from one population does not transfer well when applied to individuals of a different population

(known as a “transferability problem”). It is not uncommon for populations to be defined using global identifiers (e.g., European ancestry, African ancestry, Asian ancestry, etc.), a practice that I examine in various ways in each chapter. I was curious to see how well PRS perform in populations that are considered to be of nominally the same ancestry. Chapter 1 examines the comparison of PRS performance for red blood cell traits within nominally African ancestry individuals. The main goal of this exercise was to see whether there would be consistency across PRS in discerning which individuals are at high or low risk, since it is these areas of the distribution that have been discussed as being most clinically relevant. Prior studies have looked into the transferability of PRS across different groups of individuals within the same ancestry background, as well as the “stability” of PRS generated from different discovery GWAS and applied to the same target populations (Kamiza et al., 2022; Schultz et al., 2022). While this analysis draws upon those findings, it is unique in several ways. First, the analysis focuses on underrepresented populations, despite the small cohort sample size, introducing challenges brought upon by compromising on statistical power. Second, the analysis focuses on individuals with African ancestry residing in two different countries - the US and the UK - with deep-rooted diaspora communities. Studies have shown that diasporic populations can have complex social identities and genetic ancestries (Belbin et al., 2021; Mathias et al., 2016), which can complicate clinical applications of PRS, especially when those populations have been historically marginalized and/or ignored in biomedical research and healthcare.

Insights from my analytically-focused aim inform the analysis of my second research question, which was centered around the idea of equity in PRS. Equity has commonly been characterized as building PRS that perform equally well for individuals of any ancestry background. Chapter 2

explores how this interpretation of equity impacts PRS development and implementation. This study is the first to date that uses the bioethical framework of Bounded Justice to examine the goal of achieving equity in PRS. Taking a bioethical approach, this analysis situates genomics research practice in the larger socio-historical context and illustrates how attempts to build equitable PRS are constrained by legacies of colonial ideology in genomic science.

Finally, I bring both analyses together in my third study, which explores researcher perspectives on the use of ancestry in genome-wide association studies and polygenic risk scores. Informed by my own experiences, I conducted a qualitative study in which I interviewed genomic scientists who are currently leading the effort to develop new methods and implement PRS in diverse populations. While many studies have been done to investigate how researchers conceptualize REA, this study is different in that it was focused on better understanding the analytical and methodological details for how REA is technically incorporated into analysis. Overall, discussions capture scientists' experiences, challenges, and outlook on using REA in PRS.

This dissertation offers an integrated analysis of PRS development and the concerns of equity in the context of research practice. While each project builds on a rich body of existing literature across many different disciplines, this work as a whole methodically brings together insights from three very different approaches to investigating the question of equity in PRS development.

Chapter 1
Comparing polygenic risk scores developed from different discovery GWAS of African ancestry populations

*Performed in collaboration with researchers at the University of North Carolina at Chapel Hill

Introduction

Ensuring that polygenic risk scores (PRS) are portable across diverse populations is an important area of focus in order to implement equitable access to genomic healthcare. Genomic data from major data repositories based in certain countries - like the US or UK - are major resources that are used to calculate and validate PRS on populations from all over the world.

The promise of PRS is to aid in early screening and detection of disease for those who are genetically at higher risk. Studies have shown that PRS can improve risk assessments for certain conditions, such as cardiovascular disease (Khera et al, 2018; Hindy et al., 2020; Khera et al., 2022; Klarin et al., 2022). They have shown to aid in the improvement of risk estimation for disease development, the prediction of response to treatment, and in early disease detection with general support of preventative medicine practices. As mentioned earlier, a PRS for CAD can help identify those at high risk for experiencing a myocardial infarction event (Klarin et al. 2022). However, the lack of representation of non-European populations in GWAS poses a challenge to the vision of PRS implementation in the clinic. PRS developed from primarily (or only) European-ancestry population data underperform in individuals of other population backgrounds; this is due to genetic contributions, such as differences in linkage disequilibrium (LD) and allele frequency patterns, as well as non-genetic influences such as environmental exposures and differences in lived experiences impacting health and disease (McArdle et al. 2021; Choi et al., 2020; Kaplan & Fullerton, 2022; A. C. F. Lewis & Green, 2021). Previous

studies have shown that PRS typically perform best in individuals with similar ancestry backgrounds (Ding et al., 2023b), and so the dearth of data from diverse populations is an obstacle for both the investigation of PRS transferability across individuals of different ancestries and the development of PRS in populations of mixed and non-European ancestry. For example, Martin et al. (2019) illustrated that ~79% of all GWAS participants according to the GWAS Catalog (to date of publication) were of European descent while Morales et al., (2018) showed that only 2.8% of studies in the GWAS catalog are done in individuals of African ancestry (Martin et al., 2019; Morales et al., 2018). Unsurprisingly, Martin et al., show that average prediction accuracy of PRS across 17 quantitative traits (including anthropometric traits, blood pressure, and blood cell traits) is far lower for individuals of African descent relative to those of European descent (prediction accuracy: EUR (ref) = 1; AFR = 0.25). Similarly, Ding et al. (2023) demonstrate a four-fold increase in predictive accuracy of PRS for height for individuals whose genetic ancestry most closely matches that of the population used to develop and train the PRS model; the least predictive scores were observed for individuals of African descent as compared to individuals of European descent, while predictive accuracy for individuals identifying as South Asian, East Asian, and Hispanic/Latino had intermediate performance between that of individuals of European and African descent (similar trends in PRS accuracy were observed across the 82 of the 84 traits tested in this study). The lack of diverse data also poses a challenge in approaches to study design. Especially for underrepresented populations, there is not enough data available for analysis. Thus, analysts may be interested in combining data from individuals who fall within a single presumed ancestry group, such as continental ancestry (e.g., “Asia” or “African” or “European”) yet are from different parts of the world (e.g., United States and Taiwan). However, the well-documented PRS transferability problem suggests

that combining heterogeneous populations into a single category may mask differences in PRS performance, perhaps leading to greater generalizability, but of poorer predictive performance for a given individual in that population (Schultz et al., 2020; Kamiza et al., 2022). Clinical implications of poorer predictive performance of PRS in some groups over others has been a concern highlighted in the literature (Martin et al., 2019; James et al., 2021; Lewis et al., 2022; Ding et al., 2023).

Many studies have looked into the genetic contributions to blood cell traits, and PRS for red blood cell traits in particular can be informative for better understanding risk of blood-related conditions such as anemia, which occurs at a higher prevalence in individuals of African descent. Given that previous studies demonstrate the importance of developing and optimizing PRS in cohorts of similar genetic ancestry to the target population, further investigation into PRS analysis and performance in individuals of African descent is necessary for improvement in predictive accuracy. Indeed, developing accurate and reliable PRS in populations of African ancestry can yield clinical insights and provide health care benefits in the treatment of blood-related conditions. Red blood cell traits serve as an informative model for investigating genetic contributions to phenotypes because of their polygenicity, influence of environmental factors, and evidence of a large number of population-enriched variants that are associated with them (Raffield et al., 2018).

Due to the increased genomic diversity and on average shorter haplotype blocks in populations of African ancestry, PRS performance tends to be particularly poor for individuals with significant African ancestry (Martin et al., 2019). The poor performance is likely due to the

decreased likelihood that, given short haplotype blocks, variants captured in the PRS are less likely to be in LD with the causal variant. Previous literature has shown poor transferability of PRS within African populations due to genetic and environmental differences (Kamiza et al., 2022), indicating that PRS developed in individuals with African ancestry do not perform equally well across all populations of nominally African ancestry. While prior research has compared the transferability of a single PRS model in multiple populations, in this study, we set out to assess the concordance of two PRS models derived from African ancestry population data in two major data repositories from two different countries: TOPMed (US) and UK Biobank (UK). By examining the performance of two PRS models in a single evaluation cohort that are derived from different summary statistics and optimized in the same cohort, we can get a better understanding of whether polygenic scoring substantially differs depending on summary statistics of geographically different populations with nominally similar continental-level ancestries and why that might be the case. Especially for underrepresented populations for whom there are few-to-no summary statistics available, analysts may look to larger datasets from individuals of nominally similar ancestry (oftentimes based on continental ancestry groupings) to develop and/or evaluate PRS performance; however, doing so may obscure important genetic and non-genetic contributions to disease, and therefore impact polygenic risk estimates.

We chose these two resources for several reasons. First, the US and the UK have complex histories of diaspora communities; diasporic populations can have unique population genetic structure as well as different lived experiences (Belbin et al., 2021; Mathias et al., 2016), which can impact PRS analyses and application of results to individuals of similar ancestry backgrounds but of different diasporic communities. However, the use of continental, or

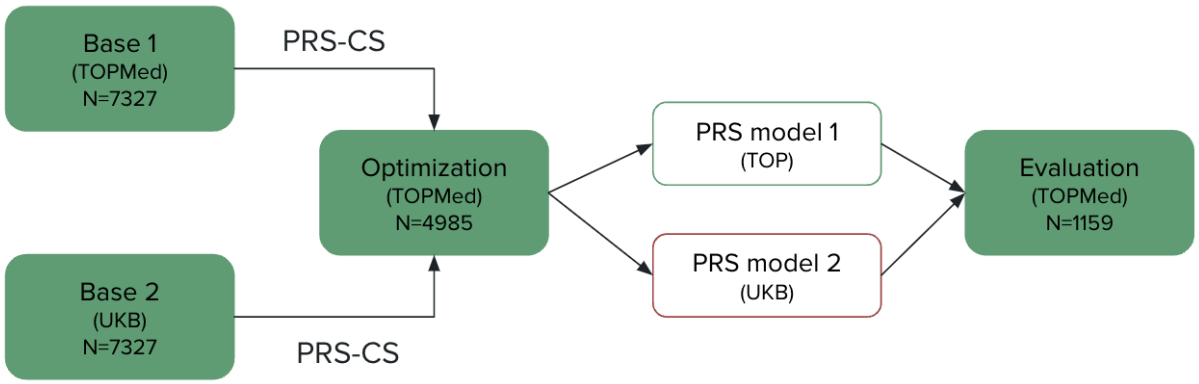
otherwise geographic-based descriptors currently remains common in PRS development; such descriptors can obscure both genetic and non-genetic heterogeneity in diaspora populations that may be relevant to health outcomes. Second, PRS development and validation require multiple independent data sets of large sample sizes. With the current underrepresentation of non-European ancestry populations in genomics research, analysts may look to develop PRS using data from study populations recruited in one country and optimizing or evaluating the PRS in an independent cohort of study data from another country (Ge et al., 2022). Both TOPMed and UK Biobank are major resources of data, collected in two different countries with different environmental and social contexts, including different African diaspora communities. Specifically, many African or African-descent individuals represented in the UK Biobank are from the Caribbean or from many different countries all over Africa, while many individuals of African-descent represented in TOPMed are African Americans, with West African and European ancestry (Hu et al., 2021; Sun et al., 2021).

In addition to the lack of transferability of PRS between populations of different ancestries, previous studies have also shown modest to poor stability (i.e., concordance in score) of PRS when generated from different discovery GWAS (Schultz et al., 2022), though more studies have looked into PRS performance across nominally different ancestry populations. It is also clinically problematic if PRS developed from different summary statistics identify different sets of individuals at high or low risk, as it can lead to misrepresentation of risk and faulty clinical decision-making, contributing to health care inequity, particularly for individuals of underrepresented communities and populations with high genetic heterogeneity, such as those with African ancestry. Here, we leveraged the TOPMed and UK Biobank data resources to

examine the performance of PRS models in populations from different countries but of nominally the same ancestry (Fig. 1.1). We focused on African ancestry population data because we were interested in working with data from underrepresented populations, while also trying to achieve enough power in our analyses. We chose red blood cell traits as the phenotype of interest because of the accessibility of the data, the known genetic contributions to phenotypic variance, and the clinical relevance of red blood cell traits to health complications such as anemia, a condition reported to be more prevalent in individuals of African descent (Chen et al., 2013). Additionally, the varying reported heritability across traits (e.g., RBCC=56%; MCV=40-52%; HGB=34-90%) provide an opportunity to examine how PRS performance may vary with heritability (Kullo et al., 2010; Barrera-Reyes, 2019; Patel, 2008). We were primarily interested in investigating the degree of sample overlap between the PRS models generated from each set of summary statistics in both the optimization and evaluation cohorts. Consistent with previous studies, we focused on the upper and lower quantiles (Choi et al., 2020; Kamiza et al., 2022) and observed that PRS models generated from different discovery GWAS do not reliably identify the same individuals (Schultz et al., 2022). We hypothesized that both PRS models would perform similarly, as measured by comparing score quantiles from both models for each individual in the target cohort. We used PRS-CS to generate PRS models, as we were interested in using a method that is commonly used for single-ancestry or single-population PRS computation and that applies an evidence-based weighting algorithm to the effect size estimates of variants associated with the trait, relieving the need for arbitrary p-value thresholding. While performance metrics, such as proportion of variance explained (PVE), were similar for both PRS models, we observed different risk scores between the two models for the majority of individuals in the target cohort.

As mentioned earlier, the differential predictive performance of PRS has led to concerns of worsening existing health and healthcare inequities. Different PRS models that provide varying estimates of risk for a given individual can lead to confusing or incorrect recommendations for care (James et al., 2021). It is crucial to develop equally predictive PRS, but evaluating PRS performance by both summary metrics and risk estimation reliability at the individual level (especially for underrepresented populations) across PRS models developed from different datasets is important for equitable clinical implementation. A closer examination of individual-level risk estimation by different PRS models, particularly for underrepresented populations where a lack of data introduces greater analytical challenges, can elucidate what factors impact PRS accuracy and create greater opportunities to investigate how to better translate summary-level measures to individual-level risk assessment. This analysis gives us further insight into concordances of multiple PRS models being applied to the same target population, and highlights some of the complexities of PRS model development for underrepresented populations.

Figure 1.1 Diagram of study design



Methods

Study population

The study population for the reported analyses included participants with African ancestry from the Trans Omics for Precision Medicine (TOPMed) Program and the UK Biobank (UKBB). Consistent with multiple prior projects (e.g., in TOPMed and MVP) (Chen et al., 2020; Hu et al., 2021; Little et al., 2022; Mikhaylova et al., 2021; Sun et al., 2021), we used a combination of reported race/ethnicity and genetic ancestry clustering to identify individuals for analysis.

For TOPMed study participants, we used harmonized race and ethnicity variables to select study samples for inclusion in analysis, for which details have been described elsewhere (Fang et al., 2019; Stilp et al., 2021b). Briefly, race and ethnicity data were harmonized according to US OMB (Office of Management and Budget) categories (see Supplementary Information for more details); notably, part of the harmonization process employs a method called HARE (Harmonized Ancestry, Race, and Ethnicity) that applies a supervised learning model to impute the race/ethnicity of individuals (who have missing race/ethnicity data) based on where their samples fall in PC space relative to the genetic PCs of samples from individuals with known race/ethnicity values (race and ethnicity values identified using HARE were known as the "HARE group") (Fang et al., 2019). Harmonized race and ethnicity variables, including those that employed the HARE method, were used to select individuals with a reported race of Black or African American. The TOPMed cohort included participants from nine participating studies (n=14208): Jackson Heart Study (JHS, n=2903); Genetic Epidemiology of COPD Study (COPDGene, n=1670); Atherosclerosis Risk in Communities Study (ARIC, n=1902); Women's Health Initiative (WHI, n=1447); Mount Sinai BioMe Biobank (BioME, n=3112); Coronary

Artery Risk Development in Young Adults (CARDIA, n=1359); Cardiovascular Health Study (CHS, n=691); Multi-Ethnic Study of Atherosclerosis (MESA, n=641); and Genetic Studies of Atherosclerosis Risk (GeneSTAR, n=729).

For UKBB study participants (n=7327), to maintain similarity with ancestry ascertainment approaches for the TOPMed cohort , individuals included for analysis were those who both (1) reported identifying as as either “Black or Black British”, “Mixed”, “Other ethnic group”, “Do not know”, “Prefer not to answer” during the initial Assessment Centre visit (variable “ethnic_background”, Data-Field 21000 of the UKBB data); and (2) whose samples cluster genetically with 1000 Genomes AFR samples in principal component (PC) space using K-means clustering consistent with prior studies (Q. Sun et al., 2021).

Given the sample size discrepancy between the TOPMed cohort and the UKBB cohort, which would impact the statistical power of the analyses, we downsampled the TOPMed cohort to match the sample size of the UKBB cohort by performing a random selection of 7327 individuals out of the total cohort size, resulting in the same sample size for both TOPMed and UKBB GWAS (referred to as the “base cohort”). The remaining samples in TOPMed were then used as the independent optimization sample set for PRS model generation (N = 4958 (MCV); N=7127 (HGB)) referred to as the “optimization cohort”). Relatedness was computed using PC-Relate and a sparse kinship matrix was examined for related individuals, defined as having a kinship coefficient of less than the 3rd degree ($KC > 2^{(-9/2)}$); less than 1% of individuals were related between the optimization and base cohorts. Finally, 1159 TOPMed participants from the HCHS/SOL study with a HARE_group assignment of “Black” or “Dominican” were randomly

selected for the evaluation cohort; we sought to include diverse participants who identified as Caribbean and/or with African ancestry backgrounds to test the two optimized PRS models, given that the base cohorts of the summary statistics used to develop the PRS models contained participants of African and Afro-Caribbean descent.

Red blood cell trait measurement

Six RBC traits (hemoglobin, hematocrit, mean corpuscular volume, mean corpuscular hemoglobin concentration, red cell distribution width, and red blood cell count) were initially chosen based on previous studies indicating their clinical relevance (such as anemia) and demonstrating substantial genetic components (Kullo et al., 2010); however, this analysis describes investigations into two traits: mean corpuscular volume (MCV) and hemoglobin (HGB). We narrowed this initial analysis to two traits due to time constraints and we specifically focused on MCV and HGB to enable assessment of PRS concordance in traits with varying heritability estimates (MCV=0.52; HGB=0.34-0.9).

Red blood cell traits (RBC) in TOPMed were measured from freshly collected whole blood samples at local clinical laboratories using automated hematology analyzers calibrated to manufacturer recommendations according to clinical laboratory standards. RBC data was harmonized following harmonization algorithms developed by domain experts and data harmonizers at the TOPMed DCC (Hu, Stilp et al., 2021; Stilp et al., 2021). This analysis focuses on two RBCs to start: mean corpuscular volume (MCV) and hemoglobin (HGB). MCV is the average volume of red blood cells, measured in femtoliters. HGB was defined as the mass per volume (grams per deciliter) of hemoglobin in the blood. The distribution of MCV and HGB were analyzed and extreme values likely due to measurement of recording errors were removed

from the analysis (see supplemental methods). Participants were excluded if they had extreme values for MCV (MCV > 150 fl) and HGB (HGB > 30 g/dL and HGB < 5 g/dL). UKBB has extensive data collected from questionnaires, physical measures, and stored blood samples. Measurement of red blood cell traits in the UKBB has been previously described elsewhere; briefly, blood samples were collected at baseline visits and specific trait measurements were either directly measured (HGB) or derived (MCV) (Bycroft et al., 2018; Vuckovic et al., 2020; Q. Sun et al., 2021).

WGS, genotyping, and imputation

WGS was performed as part of the NHLBI TOPMed program and has been described elsewhere (Taliun et al., 2021b). Genomic locations of variants are based on GrCh38. Genotyping methods for the UK Biobank are described elsewhere; briefly, the The UK Biobank Affymetrix Axiom array was used to perform genome-wide exome sequencing (Astle et al., 2016). Imputation was performed using the TOPMed imputation server and all genotypes were imputed to the TOPMed freeze 8 reference panel (Das et al., 2016; Taliun et al., 2021b) in order to better capture ancestry-specific variation. SNPs that passed quality control (INFO > 0.4 and MAF > 0.001) were used in model-fitting. Genomic locations of all variants are based on GrCh38. Further details are described in previous literature (Q. Sun et al., 2021).

Single-variant association analyses

A single-variant association test (Conomos, 2017) on the TOPMed base cohort data was performed for hemoglobin and MCV using linear mixed models (LMMs) and a fully adjusted two-stage procedure for rank-normalization following the same method used in Hu et al(Hu et al., 2021). The covariates included were sex, age at trait measurement, study, study phase,

indicators for stroke, COPD, and VTE events, and the first 11 principal components. We accounted for heterogeneous residual variances by study group to aid in genomic control.

Collaborators at the University of North Carolina Chapel Hill performed a single-variant association test on 6 red blood cell traits in the UKBB base cohort following a two-stage adjustment model to improve power and control of false positives. First, a regression was performed on raw phenotypes, age, age-squared, sex, and study center. Population structure and relatedness was adjusted for using EMMAX (see Supplemental Information for more detail). Residuals from the model were obtained and then inverse normalized for association analysis. Genome-wide association testing was performed using the EMMAX test in EPACTS across all SNPs.

Polygenic risk score calculation

Summary results from the single-variant association tests were used to calculate polygenic risk scores for each red blood cell trait (Figure 1.2). For each trait, we calculated a PRS using the TOPMed summary statistics and a second PRS using the UK Biobank summary statistics. We then calculated the score of each of the two PRS for each individual in the optimization cohort. Subsequently, we tested both models in an independent cohort (evaluation cohort) of 1159 participants from the TOPMed HCHS/SOL study. For MCV, we also developed PRS models using only genome-wide significant SNPs in the optimization cohort.Variants from both TOPMed and UKB summary results use TOPMed variants from build hg38.

Following approaches in prior studies, we used PRS-CS to generate scores. PRS-CS (Ge et al., 2019) is a scoring method that employs a Bayesian approach to infer posterior SNP effect sizes

under continuous shrinkage priors using GWAS summary statistics and an external linkage disequilibrium (LD) reference panel, relieving the need to use arbitrary cutoffs for clumping to account for LD and P-value thresholds. Bi-allelic single nucleotide polymorphisms were included in polygenic scoring across all chromosomes. The 1000 Genomes reference panels were made available for download from the PRS-CS github page. Reference samples of Yoruba individuals from Ibadan were used to adjust for LD for all PRS because it provided the best match to the ancestries used in both base GWAS cohorts. Due to the small sample size of the GWAS summary statistics and the varying polygenicity of red blood cell traits, multiple models were generated using four different values for the global search parameter, phi: $1x10^{-6}$, $1x10^{-4}$, $1x10^{-2}$, and 1. All other optional parameters used default values: parameters a and b in the gamma-gamma prior were set to 1 and 0.5, respectively. The number of MCMC iterations and burnin iterations were 1000 and 500, respectively. The thinning factor of the Markov chain was set to 5. Python version 3.2 was used to run PRS-CS v1.0.0 for all models.

PRS-CS operates on autosomes. However, we were interested in including the X chromosome because there is a strong signal of association between G6PD on X and red blood cell traits, which we did not want to ignore. As a result, a combination of pruning and thresholding of the X chromosome was done using a sliding window of 1 kb and a p-values threshold of $p < 0.01$. Full dosage compensation was assumed for X genotypes. No adjustments were made to the effect sizes of X chromosome variants used in the PRS.

Association of PRS and red blood cell traits

Linear mixed models were used to test the association of the PRS with each red blood cell trait for all four PRS models generated using different phi parameters. The same covariates used in

the GWAS (noted above) were also included in the PRS association test. A score test, which used a null model and vector of polygenic risk scores for all individuals, was used to assess the significance of the PRS association; the estimated PRS effect size and proportion of variance explained (PVE) were approximated using the score statistics (Hu et al., 2021). The best PRS model was selected after comparing the PVE Analyses were conducted using the GENESIS package in R (Conomos, 2017).

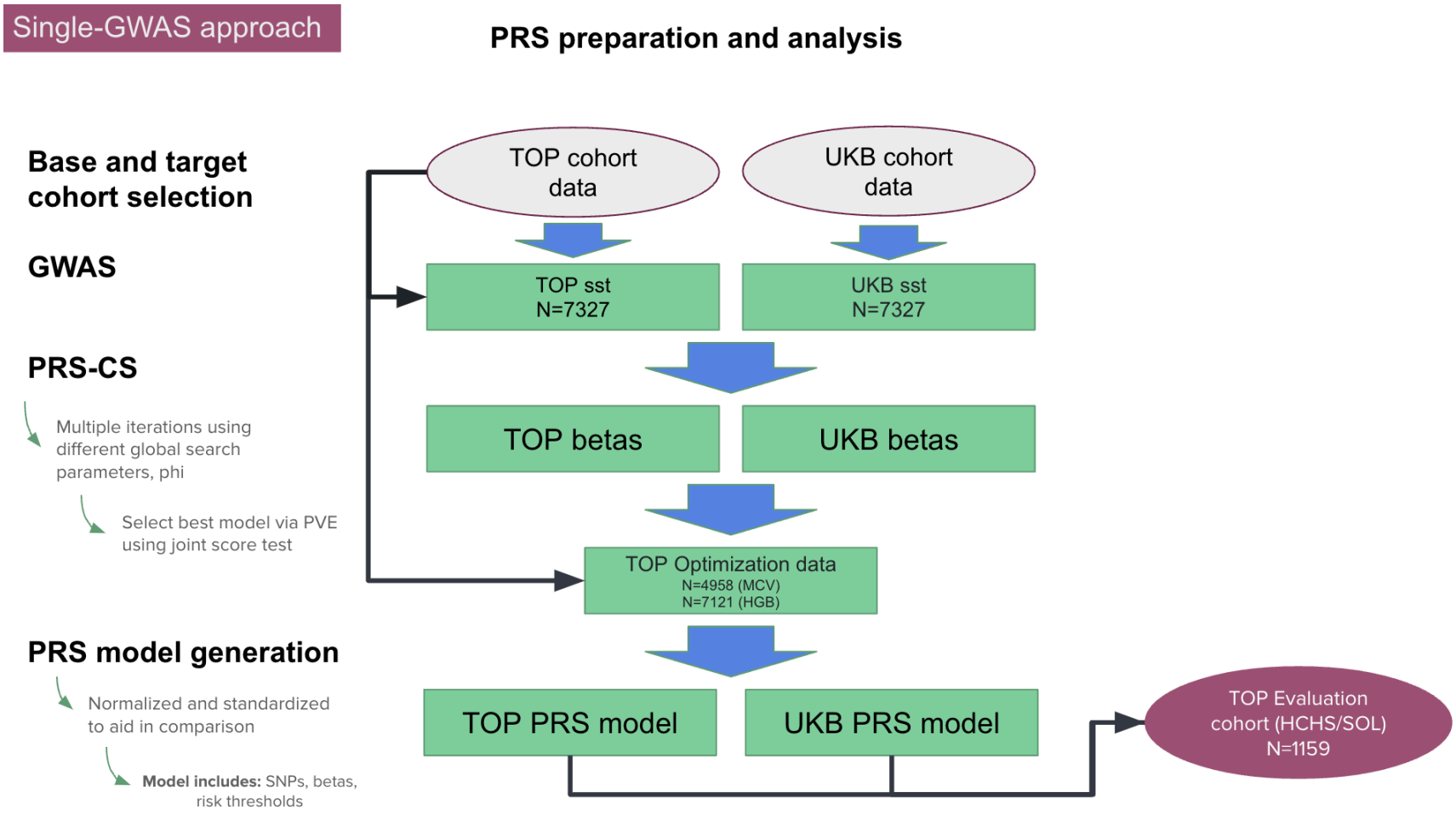
Resampling TOPMed data and performing iterative PRS analyses

In order to assess whether the results of our PRS model comparison was due to chance, we resampled the TOPMed cohort without replacement five times to generate five cohort pairs (i.e., 10 total cohorts). Each pair consisted of a base cohort of the same sample size as all other iterations (N=7327) to match the sample size of the UK Biobank GWAS and an optimization cohort (N=4958 (MCV); N=7121 (HGB)). A GWAS was performed on each of the five base cohorts, and a PRS model was generated from each of the five resulting summary statistics.

Assessing PRS concordance by sample overlap

We identified participants in the highest and lowest 5th percentiles for each PRS model. We counted the number of participants in the top and bottom 5th percentiles for each PRS model and assessed how many participants overlapped between the two models. As an example, for the MCV trait PRS, 247 individuals were assigned to the high score quantile (i.e., top 5th percentile) based on their standardized, scaled PRS by each PRS-model. We then counted the number of overlapping individuals who were assigned to the high score stratum (N=67) and divided it by the total number of individuals assigned to that stratum (67/247) to get the percentage overlap within that group (27%).

Figure 1.2 PRS Preparation and analysis



Results

Results from genome-wide association analyses

Single-variant association was performed on TOPMed data for two blood cell traits: mean corpuscular volume (MCV) and hemoglobin (HGB). The genomic inflation factor suggested sufficient genomic control ($lambda_{MCV}$ = 1.007 $lambda_{HGB}$ = 1.008). Given that detailed PRS analysis was performed on MCV, we briefly report summary results for MCV. A total of 1014 variants reached genome-wide significance ($P < 5 \times 10^{-8}$). The strongest signals were found across seven chromosomes (number of statistically significant variants provided in parentheses): chromosomes 3 (4), 9 (49), 10 (9), 11 (70), 15 (1), 16 (867), 17 (1), and X (271). Variants reaching genome-wide significance on chromosome 16 are likely markers tagging a structural variant (16p13.3 (alpha-globin) locus) known to be associated with red blood cell traits. Results from MCV GWAS performed on UKB data revealed a total of 838 variants reaching genome-wide significance. The strongest signals were found across four chromosomes (number of significant variants in brackets): chromosomes 6 [5], 11 [118], 16 [715], and X [265]. Similar to what was observed in GWAS performed on the TOPMed cohort, variants reaching genome-wide significance on chromosome 16 are likely markers tagging a structural variant known to be associated with red blood cell traits (16p13.3 (alpha-globin) locus) (Wheeler et al., 2022). Significant variants observed in TOPMed and UKB GWAS for chromosomes 3 (rs112097551), 19 (rs73494666), and 22 (rs228914) are consistent with prior findings in analyses of individuals of African-descent reported in Hu et al. (2021). Similarly, several variants from genome-wide significant genes from aggregated association (HBB, OR52H1, HGB1, HBA1/2))

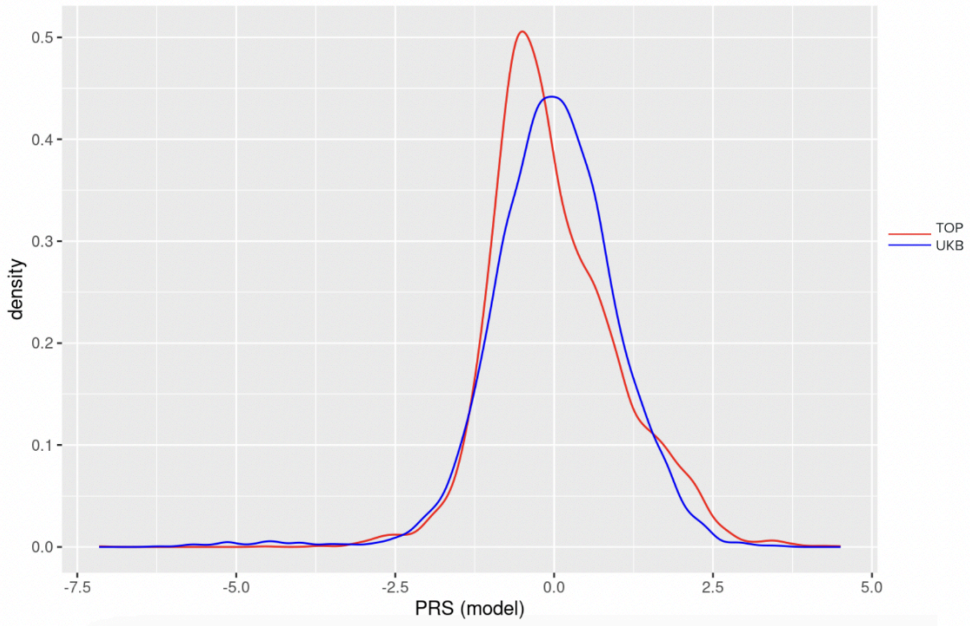
analyses done by Hu et al. (2021) were found to meet genome-wide significance thresholds and were included in the PRS models, as well (Table S4).

Comparing PRS model performance

Two PRS models were generated from the summary statistics obtained by performing a GWAS on a TOPMed base cohort and a UKB base cohort. Demographic information for both base cohorts can be found in the Supplementary Information (Table S2). TOPMed base cohorts for MCV and HGB analysis differed slightly due to differences in participants with complete blood count data. For the TOPMed base cohort for MCV analysis, the mean age was 51.3 years and the mean MCV was 87.5 fl (range=[53,140]). Of the total cohort, 61.2% were female. For the TOPMed base cohort for the HGB analysis, the mean age was 52.6 years with a cohort mean hemoglobin value of 13.12 (range=[5.3,19.2]). About 63.6% of the cohort was female. The UKB base cohort for MCV and HGB analyses had a mean age of 52.0 years with mean trait values of 87.7fl (MCV) and 13.6 g/dL. Of the total cohort, 57.2% were female (N=4190). The optimization cohort (n=4958), drawn from TOPMed data, was also different due to differences in complete blood count data. The TOPMed optimization cohort for the MCV analysis had a mean age of 51.7 years and a mean cohort MCV value of 87.4fl. Of the total number of participants, 62.6% were female. The TOPMed optimization cohort for HGB analysis (N=7127) had a mean age of 53.0 years with a cohort mean for HGB of 13.3 g/dL (range=[5.5,19.8]). Of the total cohort, 65.4% were female. We computed all PRS models across four different phi parameter values as a small-scale grid search as was recommended by PRS-CS documentation when computing PRS for smaller sample sizes (N < 10,000). The best-performing PRS models were selected by inspecting the proportion of variance (PVE) explained by the PRS and the effect size estimate of the PRS after performing a score test using the null model. The PVE of PRS-TOP

and PRS-UKB were similar to each other for both MCV (PVE_{top}=0.028; PVE_{ukb}=0.023) and HGB (PVE_{top}=0.008; PVE_{ukb}=0.007), though the PVE for MCV was overall higher than that of HGB. This is consistent with previous studies which found similar values for trait variance explained by PRS in both African ancestry and non-African ancestry populations which have demonstrated R^2 values ranging from 0.0035-0.0197 for MCV and 0.0125-0.06 for HGB (Chen et al. 2020; Vuckovic et al., 2020). Each model included a similar total number of variants for both MCV (PRS-TOP = 1375284; PRS-UKB = 1368046) and HGB (PRS-TOP=1374708; PRS-UKB=1367901) of which the majority were shared between the two models (MCV PRS=1236574; HGB PRS=1218994). However, only a small portion of the total number of variants included in the models were found to be statistically significant in each GWAS for both MCV (PRS-TOP=114; PRS-UKB=107) and HGB (PRS-TOP=7; PRS-UKB=13). PRS scores were normalized and standardized to aid in comparability. Because the upper and lower quantiles of the PRS distribution are the areas with most clinical relevance, we were especially interested in the top and bottom 5th percentile scores, which differed between the two models for both MCV (PRS-TOP=[-1.34,1.86]; PRS-UKB=[-1.47,1.56]) and HGB (PRS-TOP=[-1.63,1.67]; PRS-UKB=[-1.48, 1.82]). We calculated the correlation between the two PRS models for each trait, which was stronger for MCV than for HGB (MCV PRS r=0.46; HGB PRS r=0.28). Model metrics for the best performing PRS for MCV and HGB can be found in the Supplementary Information.

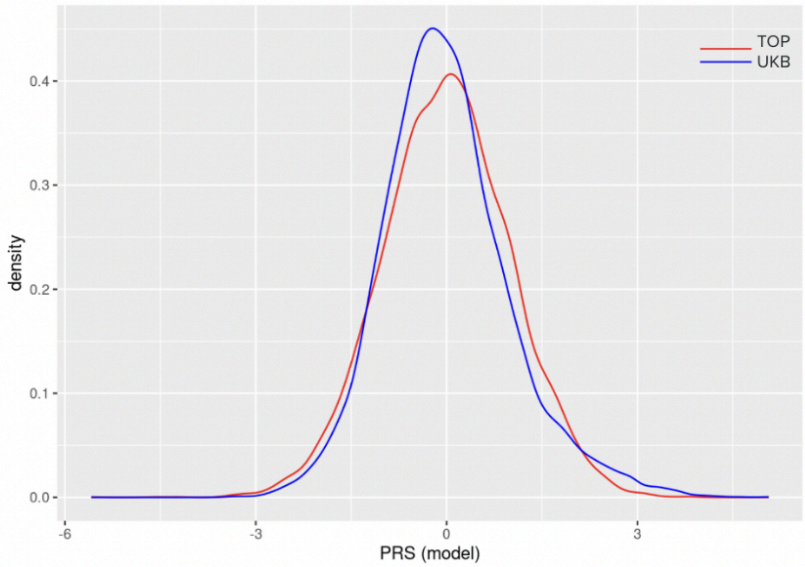
Figure 1.3. Distribution of MCV PRS optimized in the TOPMed cohort



Performance metrics for optimized MCV models, PRS-TOP and PRS-UKB

Original	phi	PVE	Beta	P-value	Bot. 5th %ile	Top 5th %ile
PRS-TOP-Orig	1e+00	0.0277	1.1438	1.367×10^{-31}	-1.34	1.86
PRS-UKB-Orig	1e-02	0.0234	1.1778	7.221×10^{-27}	-1.47	1.56

Figure 1.4. Distribution of HGB PRS optimized in the TOPMed cohort



Performance metrics for optimized HGB models, PRS-TOP and PRS-UKB

Original	phi	PVE	Beta	P-value	Bot. 5th %ile	Top 5th %ile
PRS-TOP-Orig	1e-06	0.008	0.1196	3.178×10^{-14}	-1.63	1.67
PRS-UKB-Orig	1e-06	0.007	0.1514	8.803×10^{-13}	-1.48	1.82

To examine how trait values vary with increasing PRS, we then compared the stratification of the PRS results for the red blood cell traits across 20 deciles based on a recommendation from Choi et al (2020). Current standard practice is to use the top 5th percentile of the PRS distribution as a clinically-relevant cut off (Kamiza et al., 2022; J. Sun et al., 2021). In the optimization cohorts, for MCV, study results show that individuals in the top 5th percentile of the PRS had on average higher trait values than those in the bottom 5th percentile (Fig. 1.5). For HGB, while the average trait measurement per quantile is much lower in the bottom 5th percentile than in the top 5th percentile, there seems to be no pattern in the mean trait values in the middle quantiles (Fig. 1.6).

Next, we compared scores of the two PRS models for each individual, after phi parameter tuning and selection of the best PRS model based on PVE (described in 'Methods'). As mentioned earlier, it could be clinically problematic if PRS models identified different sets of individuals as "high" risk. With quantitative traits, having extremely low or extremely high measurements can indicate risk of disease outcomes. Therefore, we examined trait values across 20 deciles. Based on prior work, we used a cut off of the top 5th percentile to designate the "high" trait-value category (i.e., the 20th decile), the bottom 5th percentile as "low" trait-value (i.e., the 1st decile) and the middle deciles as "average" (i.e., 2nd-19th deciles) (Choi et al., 2020; Kamiza et al., 2022). Notably sample overlap within each decile was poor, especially in the PRS for HGB. We especially focused on the participants whose risk assignment switched between "high PRS stratum" and "average PRS stratum" or "low PRS stratum" and "average PRS stratum" between the two PRS models, as this would be highly relevant to clinical decision making. Less than half of individuals were assigned to the top 5th percentile by both PRS models for both traits (MCV=27.1% agreement; HGB=15.7% agreement).

Figure 1.5a. Mean PRS by quantile (MCV)

Figure 1.5b. Mean trait value by quantile (MCV)

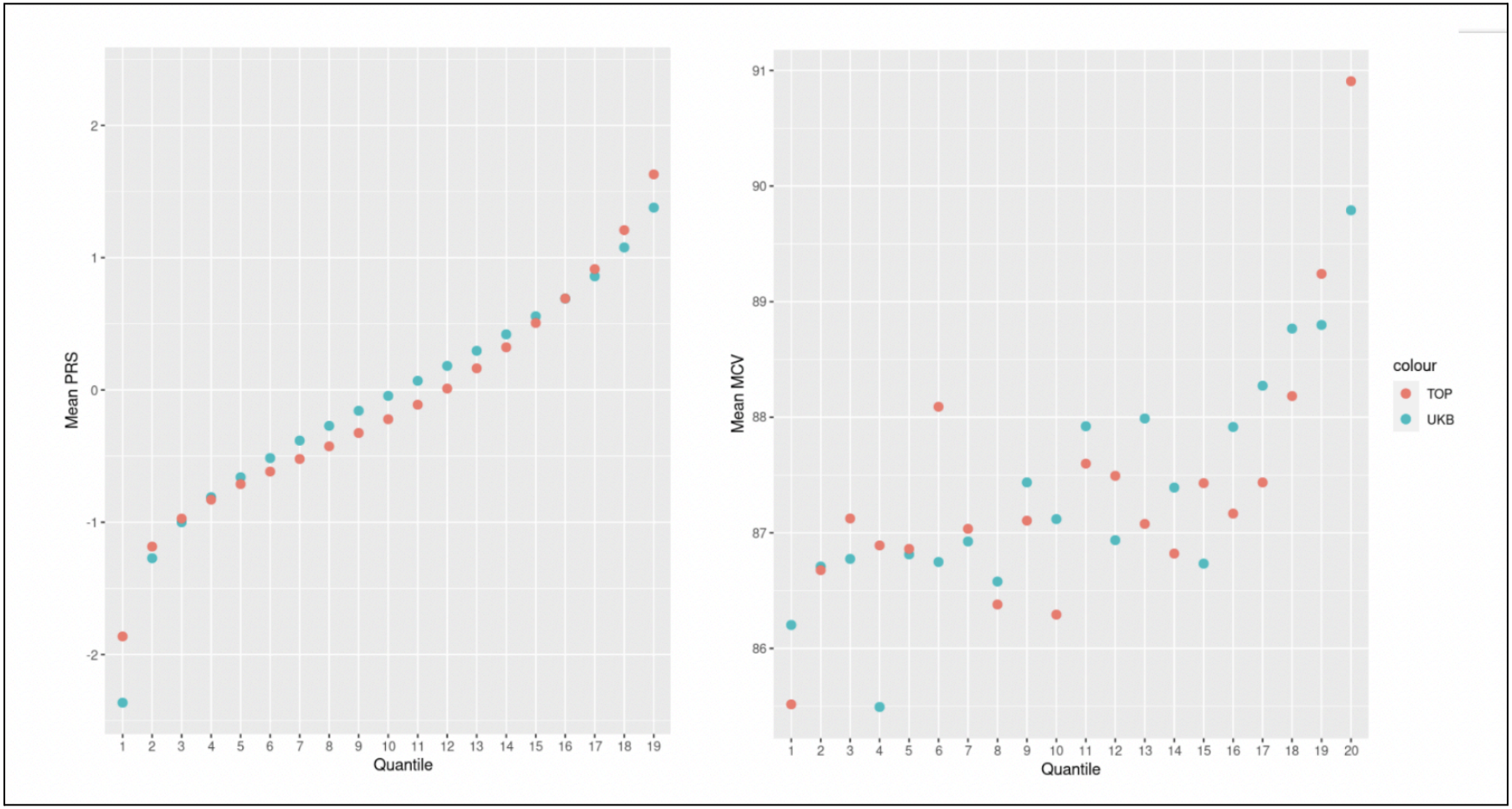


Figure 1.6a. Mean PRS by quantile (HGB)

Figure 1.6b. Mean trait value by quantile (HGB)

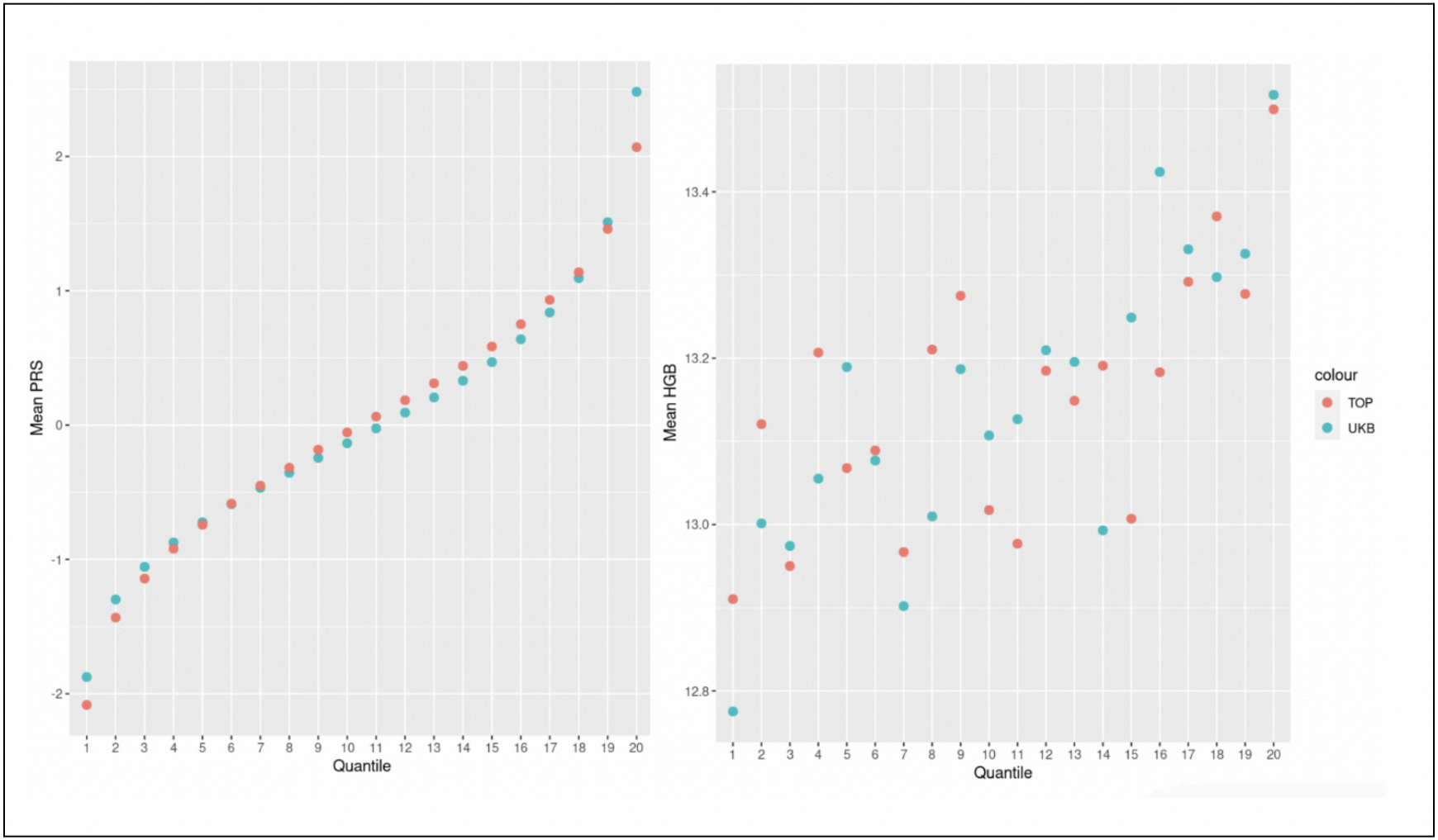


Figure 1.7. Standardized and scaled PRS across models (Optimization cohort, MCV)

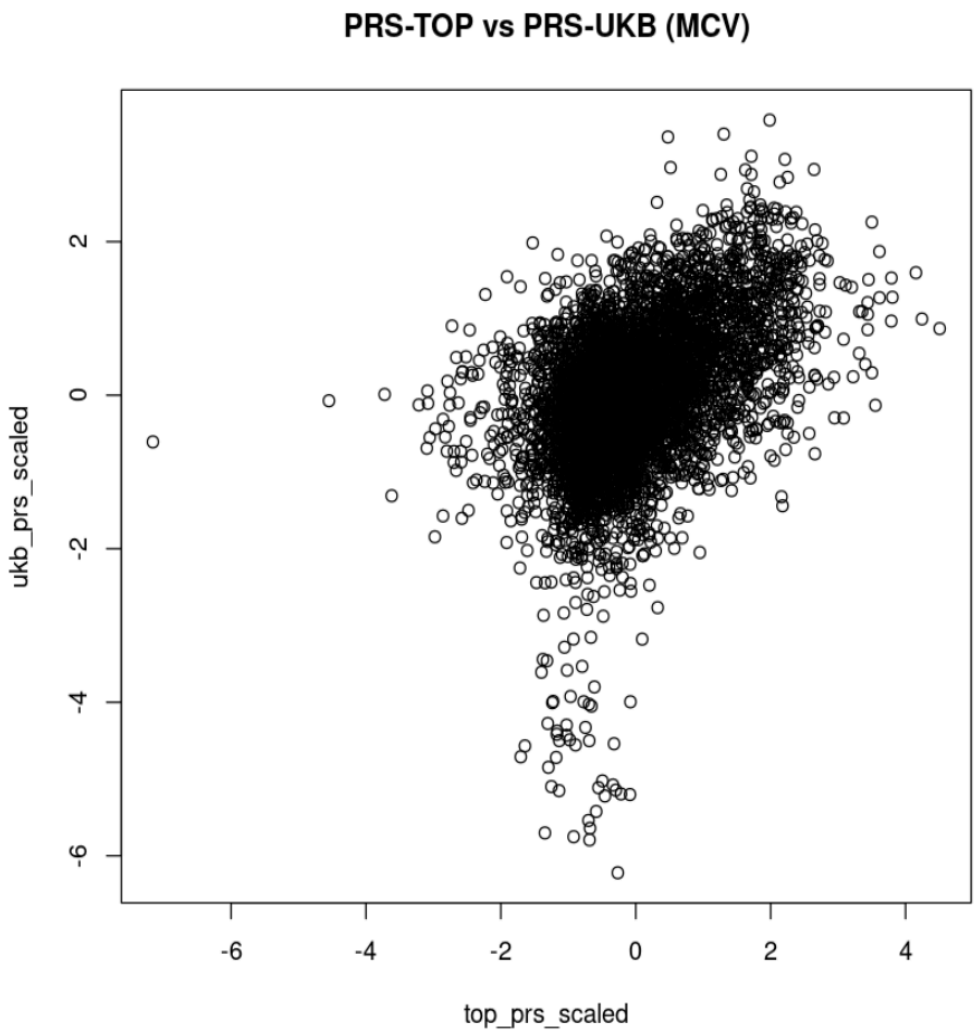


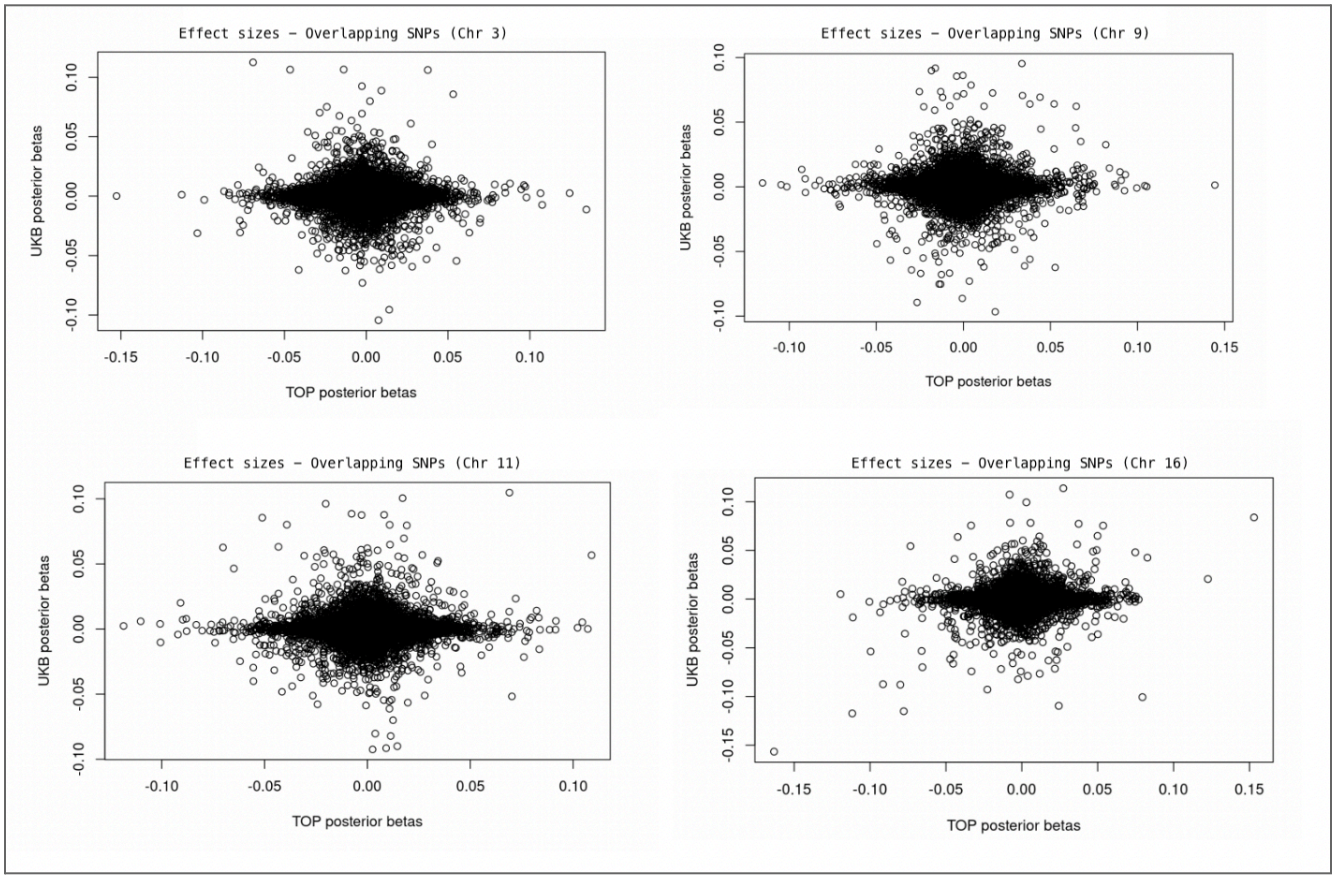
Table 1.1. Counts of individuals assigned to PRS strata by each PRS model

PRS-TOP	**Avg**	**Avg**	**Avg**	**High**	**High**	**High**	**Low**	**Low**	**Low**
PRS-UKB	**Avg**	**High**	**Low**	**Avg**	**High**	**Low**	**Avg**	**High**	**Low**
N	4059	179	225	180	67	0	224	1	23

High = top 5th percentile (qtl.20); Low = bottom 5th percentile (qtl.1); Avg=other quantiles (qtl. 2-19)

To investigate whether differences in SNP selection or in the beta values of the selected SNPs were the main contributors to the difference in score assignment, we identified overlapping SNPs used in both optimal PRSs and compared the posterior beta values. Optimal PRS models for PRS-TOP and PRS-UKB used the same phi parameter value for HGB but that was not the case for MCV. We further examined beta shrinkage patterns for MCV PRS models between both models across all values of phi, and observed similar shrinkage trends for both PRS-TOP and PRS-UKB (see Supplemental Information). Due to a greater association of the PRS with MCV than with HGB as measured by PVE (MCV: PVE_{top}=0.028; PVE_{ukb}=0.023; HGB: PVE_{top}=0.008; PVE_{ukb}=0.007) and the beta (MCV: $Beta_{top}$ = 1.1438; $Beta_{ukb}$ = 1.1778; HGB: $Beta_{top}$ = 0.1196; $Beta_{ukb}$ = 0.1514) of the PRS from the regression analysis, we focused additional follow up analysis on the PRS for MCV (preliminary analysis of HGB was performed and more detail is in the Supplementary Information). In general, scatter plots comparing beta values of SNPs that were included in the PRS model show that shrinkage may have been applied to a set of SNPs in one PRS but was not applied to the same set of SNPs in the other (Fig. 1.8).

Figure 1.8 Beta comparisons of overlapping SNPs across chr 3, 9, 11, 16 (MCV)



We examined the proportion of genome-wide significant variants from each set of summary statistics that were included in their respective PRS models and whether they were selected for shrinkage (proportion_topmed = 0.09; proportion_ukb = 0.11). Not all significant variants identified in a GWAS were included in the PRS model due to being absent in either the target population or the reference panel. Of the 1014 autosomal variants reaching genome-wide significance in the TOPMed GWAS, only 9.47% of them were included in PRS-TOP, while 11.1% of the 838 total autosomal variants reaching genome-wide significance in the UKB GWAS were included in PRS-UKB, indicating that they were the significant variants selected after PRS-CS that serve as a proxy for all genome-wide signals (Fig. 1.9). We compared the

posterior betas for all overlapping significant variants ($P < 5x10^{-8}$) that were included in the PRS models (PRS-TOP N=128; PRS-UKB N=117, overlap N=61). Our analysis was focused on variants on chromosomes 11, chromosome 16 (Fig. 1.10), and chromosome X because we observed the most signal contributing to the PRS from variants on those chromosomes. Prior literature has identified that the strongest association with RBCs on chromosome 16 is a structural variant (SV) (Hu et al., 2021), so tagging SNPs may not necessarily tag the SV well and could be different across summary statistics. In general, we observed that significant variants included in both PRS models had similar beta values. In some cases, significant variants were observed in a chromosome in one set of summary statistics, but not the other, possibly leading to their beta values being reduced in one model, while the beta values of those same variants may have been left unshrunk in the other model.

Figure 1.9. Proportion of SNPs reaching genome-wide significance (MCV) included in the PRS model across autosomes (Note: only displays chromosomes for which there was at least one variant reaching genome-wide significance)

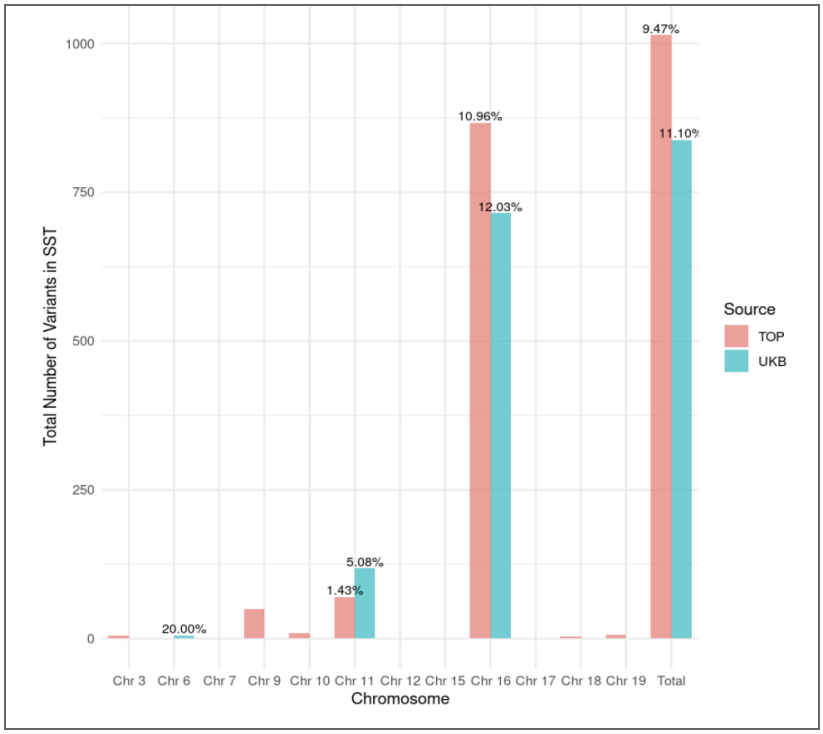
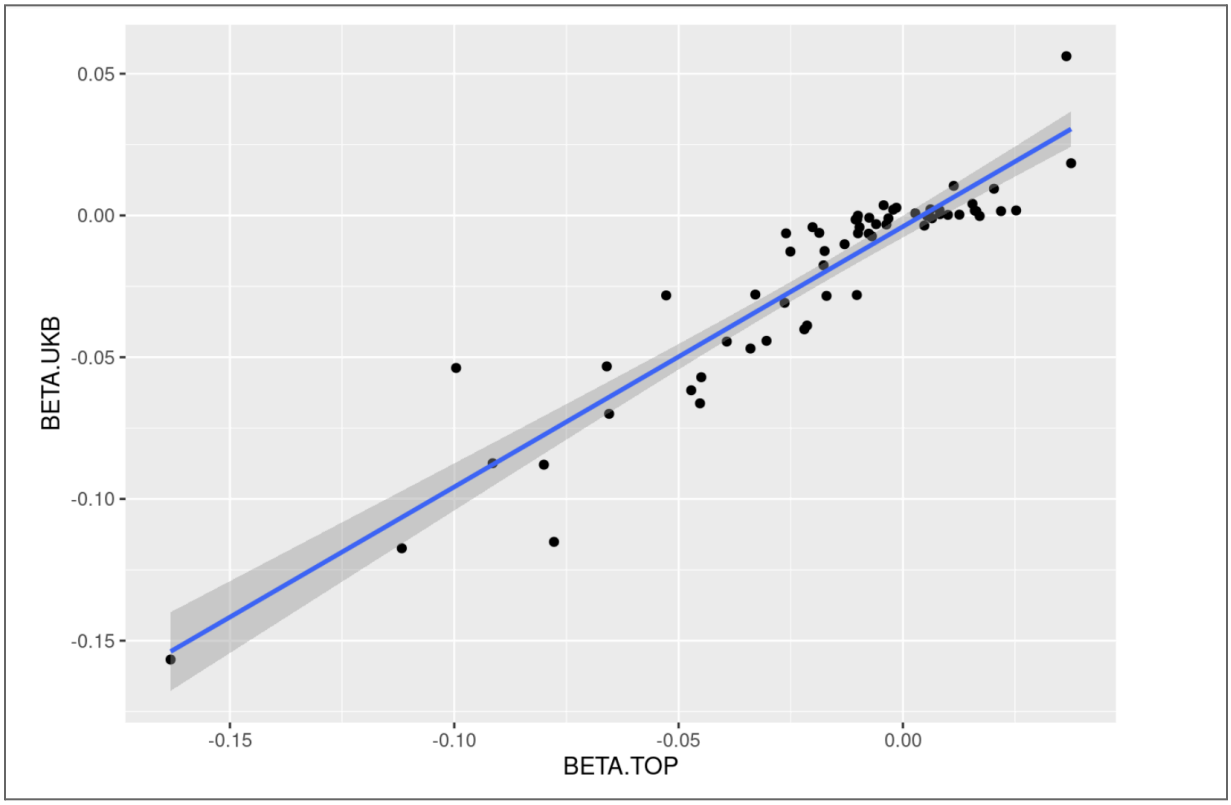


Figure 1.10. Betas of overlapping SNPs reaching genome-wide significance (chr 16, MCV)



We then constructed PRS models using only SNPs reaching genome-wide significance in each set of summary statistics to see if we might observe more agreement in score quantile assignments between models. The PRS-TOP model included 128 genome-wide significant SNPs, while PRS-UKB used 117 genome-wide significant SNPs. In the top 5th percentile (the high score quantile), we observed 85.4% agreement in score quantile assignment between the two PRS models. In the bottom 5th percentile (the low score quantile), we observed 58% concordance in score quantile agreement. Thus, 14.6% of individuals were nevertheless classified differently by the two PRS models in the highest-score decile, while 42.3% of individuals were classified differently in the lowest-score decile, an observation which warrants further examination.

PRS-CS uses a randomized component to its algorithm and this may impact variability in PRS results. To assess whether we observed similar trends in beta transformations and polygenic risk scoring, we resampled the TOPMed base and target cohorts and generated five iterations of PRS models. A second iteration showed patterns of beta transformations similar to what was observed in the primary computation, though there was some variation in the beta values (Supplementary Information). This may be a function of the global search parameter, phi, a tuning parameter that is impacted by assumptions about the sparseness of the genetic architecture of the trait (Ge et al., 2019). More work is needed to evaluate the impact of phi on beta transformations for summary statistics data of small sample sizes.

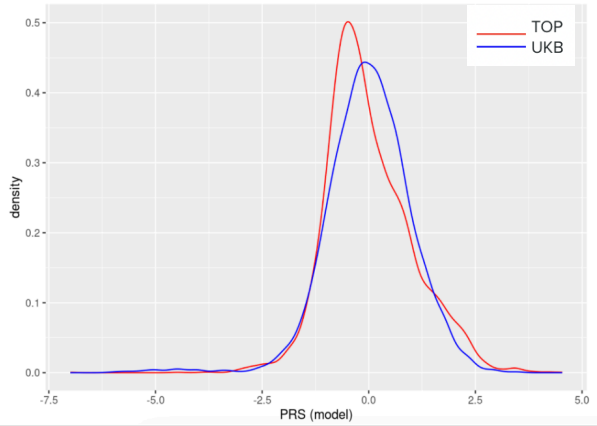
An advantage of PRS-CS is that it does not require an analyst to arbitrarily select SNPs for inclusion in the PRS model because it uses continuous shrinkage to adaptively transform genetic markers in relation to the strength of their association with the trait (Ge et al., 2019). However, given the small sample sizes of the summary statistics (N=7327) used in our analyses to derive the PRS models, we wanted to know whether the inclusion of rare variants impacted the shrinkage applied across all markers, thereby influencing the overall PRS results. We therefore regenerated PRS models after filtering out rare variants (MAF > 0.01) and compared performance metrics to the non-filtered PRS models (Fig. 1.11). Both PRS models with rare variants filtered out before applying PRS-CS performed similarly to the models where prior rare variant filtering was not performed. This seems consistent with what we would expect given that PRS-CS applies continuous shrinkage to the betas of SNPs based on strong evidence of

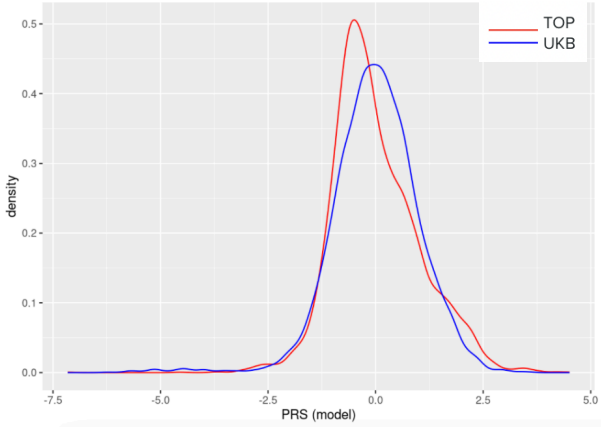
association between the variant and the trait, which may not have been the case with many variants with low allele frequencies.

Figure 1.11. Comparison of PRS models (Optimization) with and without rare variants (MCV)

Original	phi	PVE	Beta	Bot. 5th %ile	Top 5th %ile
PRS-TOP-Orig	1e+00	0.0277	1.1438	-1.34	1.86
PRS-UKB-Orig	1e-02	0.0234	1.1778	-1.47	1.56

No-Rare-Vars	phi	PVE	Beta	Bot. 5th %ile	Top 5th %ile
PRS-TOP-FIL	1e+00	0.0279	1.1515	-1.36	1.86
PRS-UKB-FIL	1e-02	0.0230	1.1685	-1.48	1.57



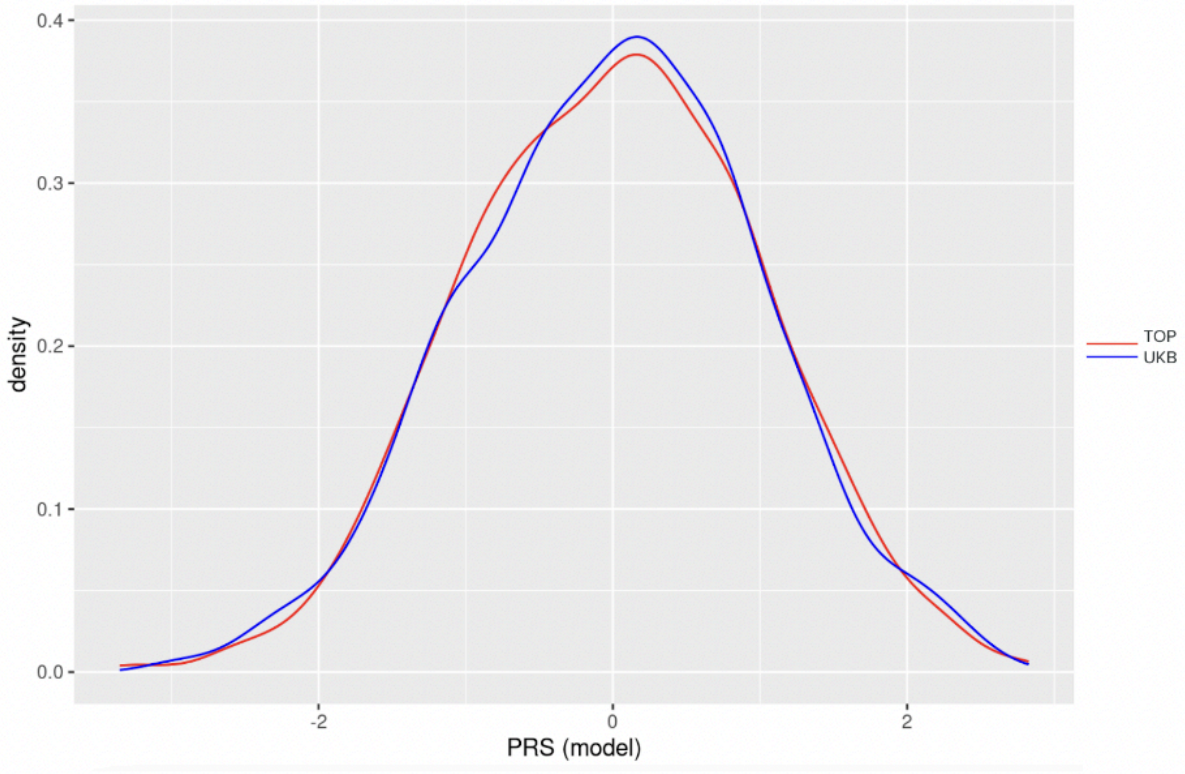


Testing the PRS models on an evaluation cohort

We tested each model on a third, independent dataset (N=1159) with samples from individuals from the TOPMed study, HCHS/SOL (referred to as the "evaluation cohort"). To try to best match the demographic background of individuals in the TOPMed and UKB base cohorts (African American - West African and African-Caribbean, respectively), samples that were assigned a HARE group of "Black" (N=23) or "Dominican" (N=1136) were selected for the evaluation cohort. Overall, both PRS models for MCV poorly explained variance in the evaluation cohort (PRS-TOP PVE=0.0134; PRS-UKB PVE=0.0016) as compared to the optimization cohort (PRS-TOP PVE = 0.0277; PRS-UKB PVE=0.0234), which may be due to differences in genetic admixture between African American individuals in US TOPMed studies and African British individuals in the UK Biobank, as well as other factors like technical

differences (e.g. study-specific differences) or non-genetic contributions. While the scores delineating the top and bottom 5th percentiles were similar across both models (PRS-TOP=[-1.62, 1.63]; PRS-UKB=[-1.64, 1.62]), model metrics showed PRS-TOP perform better than PRS-UKB (Fig. 1.12, 1.13).

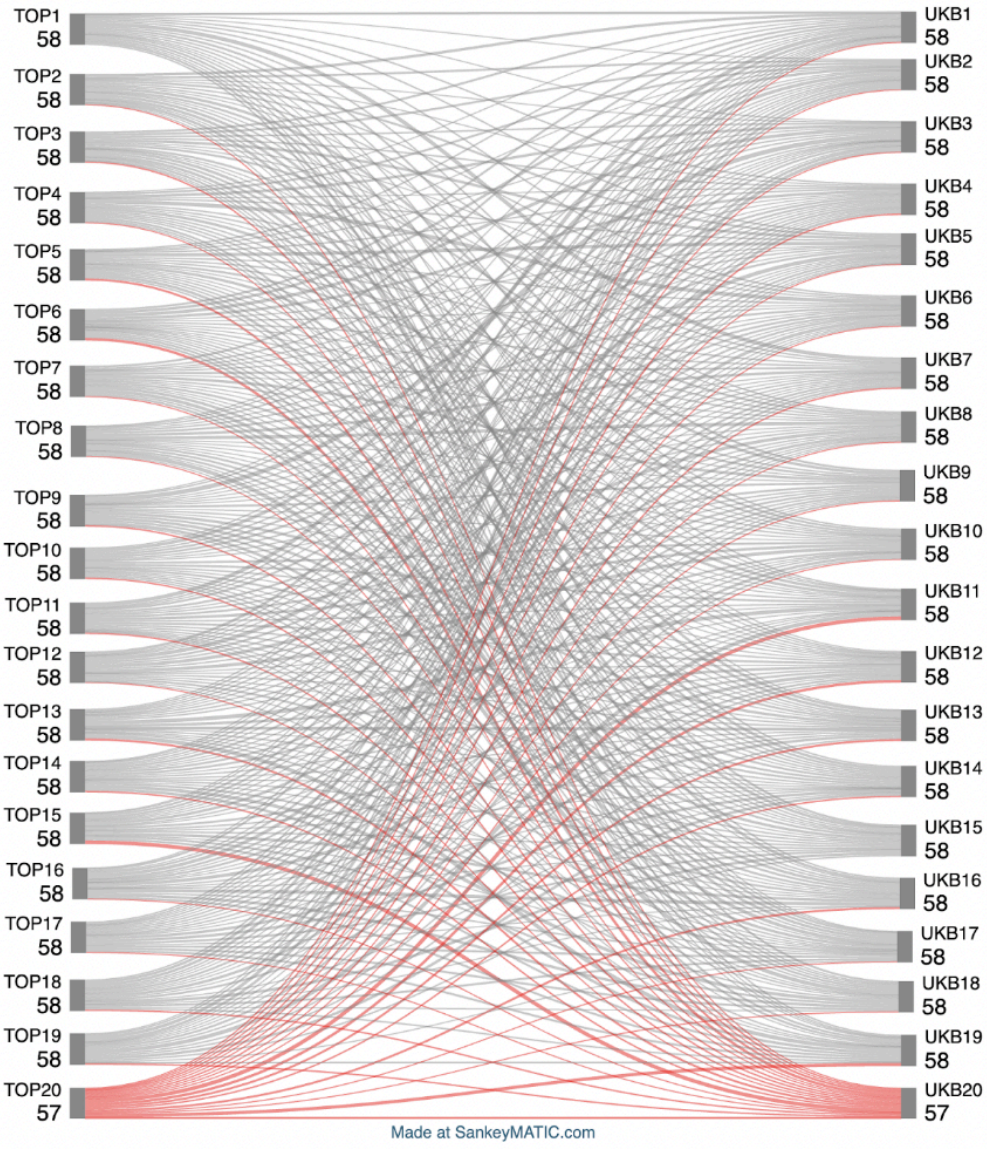
Figure 1.12. Distribution of PRS in the TOPMed evaluation cohort (MCV)



Performance metrics for PRS-TOP and PRS-UKB in Evaluation cohort

Evaluation	phi	PVE	Beta	Bot. 5th %ile	Top 5th %ile
PRS-TOP-Eval	1e+00	0.0134	0.7079	-1.62	1.62
PRS-UKB-Eval	1e-02	0.0016	0.2515	-1.64	1.62

Figure 1.13. Comparison of PRS decile assignment across PRS models (Evaluation, MCV). Individuals assigned a score in the top 5th percentile are highlighted in red.



Frequency of score distribution between two PRS models (N=1159)

PRS-TOP	Average	Average	Average	High	High	High	Low	Low	Low
PRS-UKB	Average	High	Low	Average	High	Low	Average	High	Low
N	938	52	54	53	3	1	53	2	3

Discussion

We examined the performance of two PRS models generated from different summary statistics of African ancestry study participants in TOPMed and UK Biobank cohorts and optimized the models in the same optimization cohort of African American individuals from TOPMed. We compared the association and the effect size estimates of the PRS with RBC phenotypes and assessed the degree of sample overlap within high- and low-score strata. For MCV, there was 27.1% agreement in standardized and scaled scores in the top 5th percentile between the two PRS models. The level of agreement resembles a similar trend to what has been found in a previous study by Schultz et al. 2022, who examined the percent agreement of individuals identified in the top 5th percentile between two PRS models derived from different GWAS summary statistics for multiple traits (e.g., post traumatic stress disorder (PTSD), type 2 diabetes (T2D), and height) within African ancestry populations as well as within European ancestry populations. For example, they found 36% agreement in the top 5th percentile between two PRS models for PTSD in African ancestry individuals. Similarly, they found low agreement between PRS across traits with European ancestry populations (PTSD=19.5%; T2D=34.4%; Height=41.6%). In comparison to these findings, the findings in this study show greater concordance between PRS quantile assignments for MCV in African ancestry populations than was found in Schultz et al. for PTSD in African ancestry individuals and for PTSD, T2D, and Height for European ancestry individuals. Concordance in PRS for MCV was further improved in the top and bottom 5th percentile (85.4% and 58%, respectively) observed when using only genome-wide significant SNPs in the PRS model, suggesting that PRS for MCV can be improved by using a subset of SNPs with the greatest evidence of association with the trait. These results also suggest that levels of concordance across different PRS models can vary

across traits, underscoring the need to better assess what characteristics about a trait are important to consider when developing reliable PRS models. For example, with RBC traits, having higher heritability, determining the optimal SNP set (e.g., SNPs most strongly associated with the trait), and identifying population-enriched SNPs for a given population of interest seem to be important factors to consider when assessing PRS performance. Our analysis differs from previous analyses in that it investigates PRS generation and concordance specifically in smaller population sizes and focuses on two red blood cell traits known to have population-specific variant associations for populations in malaria-endemic regions of the world. For example, the HBA gene (alpha globin region of chromosome 16 (16p13.3)) contains a copy number variation structural variant that is enriched in individuals of African ancestry and is strongly associated with red blood cell traits (Hu et al., 2021; Raffield et al., 2018). Variants on G6PD on chromosome X have been identified as having an allele frequency of 20-30% in African American individuals. This analysis offers an initial examination of PRS performance as defined by score concordance between two models in African ancestry populations with known genetic associations with red blood cell traits not easily captured currently by PRS. In order to effectively integrate PRS into clinical practice, PRS must reliably and consistently identify individuals with high genetic liability.

Observed discordance in score assignment between the two PRS models may be due to several factors, including genetic factors (e.g., differences in LD patterns and allele frequencies), non-genetic factors (e.g., differences in environmental exposures), and technical/analytical reasons (e.g., low-powered summary statistics, choice of imputation reference panel and genotyping array). Low-powered GWAS are less likely to identify variants causally associated

with a trait, particularly for polygenic traits with varying heritability like MCV and HGB, and can be more susceptible to variability in identifying associated SNPs tagging the causal variant. Therefore, it may be possible that, in a given LD block, a different SNP is selected for shrinkage by PRS-CS in one model than in the other model. The differences in selection of which SNPs are left unshrunk may occur because updates to the effect sizes are performed in multivariate block updates, where the shrinkage parameter reduces the effect sizes of noisy estimates while keeping SNPs with evidence of large signals left unshrunk. The possibility that different SNPs are effectively being used to inform the PRS may help explain why some individuals are assigned to different score deciles between the two PRS. The use of a reference panel that does not fully or accurately reflect the underlying LD patterns of the target sample may compound the effect, as ineffectually identifying SNPs in LD can lead to inflation or overfitting of PRS results. It might also be impacted by the use of imputed data from the UKB dataset versus typed data from TOPMed. The concordance of samples within score deciles between multiple iterations of PRS-CS was moderate (Supplementary Information), suggesting that some discordance in sample overlap may also be due to randomness in the PRS-CS algorithm.

This analysis expands upon previous work on the comparison of PRS performance in underrepresented populations by focusing on individual-level differences between two PRS models in study cohorts of participants with African ancestry. We chose to use data from the TOPMed consortium and the UK Biobank because they are two large and commonly used data resources for genomic analyses. They include individuals who were recruited from either within the US (TOPMed) or within the UK (UK Biobank). Picking data sources from two different countries was of interest due to contextual differences based on country, such as differing

diasporic communities (e.g., of individuals with African ancestry) and differences in the reporting and collection of race and/or ethnicity data (Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research et al., 2023b). We chose to conduct separate genome-wide association analyses for each data source and used similar approaches as well as identical sample sizes to reduce differences in statistical power and in methodology. We attempted to restrict major differences in cohorts to differences in national contexts because of the different African diaspora communities living in the US and UK (and therefore, likely different patterns of genetic admixture) by selecting cohorts with other similar characteristics. There were similarities between the TOPMed and UK Biobank base cohorts with regards to age (TOPMed= 51.3; UKB=51.6) , mean trait values (TOPMed=87.5 (MCV), 13.12 (HGB); UKB=87.7 (MCV); 13.6 (HGB)), and proportion of females vs. males in the cohort (TOPMed=61.8% (MCV), 64.8% (HGB); UKB=57.2%). In general, the majority of participants in both TOPMed and UKB datasets had trait values for both MCV and HGB within the normal range (TOPMed st.dev=6.92 (MCV), 1.54 (HGB); UKB st.dev = 6.23 (MCV), 1.41 (HGB)), which may have limited PRS predictive accuracy as the majority of study participants would not have extreme RBC values. We found the proportion of variance explained by the PRS was greater for MCV than it was for HGB, consistent with results from prior studies. This is likely due to the higher heritability of MCV as compared to HGB (Kullo et al., 2010; Viel et al., 2007), enabling the PRS to potentially explain a greater proportion of variability in MCV than HGB (Fig. 1.3, 1.4). With environmental and genetic admixture being two major differences between cohorts, one possible explanation for the observed discordance and poor proportion of variance explained by the PRS may be due to differences in genetic admixture between African American individuals in US-based TOPMed studies and African British individuals represented in the UK

Biobank. It is well known that there is a high amount of genetic heterogeneity among individuals of African ancestry, and genetic admixture varies across different diaspora communities (Mathias et al., 2016; Constantinescu et al., 2022; Kachuri et al., 2023). For example, G6PD on chromosome X and the alpha globin gene region on chromosome 16 (16p13.3) containing the large structural variant enriched in populations of African ancestry are known to be significantly associated with red blood cell traits (Hu et al. 2021). However, individuals with different admixture patterns, such as those who have greater proportions of genetic similarity to those of European ancestry, may not carry these variants. Base cohorts that include individuals of varying genetic admixture patterns may lead to weaker association signals of variants known to be more prevalent in populations of West African ancestry, or may not detect those signals at all, thereby impacting GWAS results. Thus, PRS models derived from GWAS performed in individuals of differing genetic admixture patterns may contain different variants, or may contain variants with different effect sizes. Either case would affect the PRS model and would impact who the PRS model would be applicable to. Prior studies have shown that differences in African ancestries between base and target cohorts can contribute to PRS variability in addition to other cohort differences, such as sex or age (Majara et al., 2023), indicating the importance of genetic admixture in PRS associations and accuracy of risk estimation. Previous studies have also shown that the inclusion of population-enriched variants that are clinically meaningful can enhance PRS estimates for clinical support, particularly for conditions like anemia or chronic kidney disease (Vuckovic et al., 2020). These complicated patterns of genetic admixture and population structure provide unique opportunities to further examine applicability of PRS models across diaspora communities.

We chose PRS-CS for PRS computation because it can accommodate varying effect size distributions in complex traits, incorporates local LD structure in its posterior beta transformations, and it uses all the information available in a set of summary statistics, relieving the need to set arbitrary pruning and thresholding cutoffs (Ge et al., 2019). It uses global and local tuning parameters to apply shrinkage to betas of SNPs that have demonstrated a weak signal association with the trait, while leaving data-supported SNPs with greater association to the trait unshrunk. We employed a small-scale grid search which tested four phi values and selected the best model from the set. We noticed that the amount of shrinkage changed across the different phi values, further indicating that the choice of phi may influence PRS output. However, given the highly polygenic nature of RBC traits as well as the small sample size ($N < 10,000$) of the base cohorts and optimization cohorts used to develop the PRS, other scoring methods may be more appropriate for this analysis, such as Pruning and Thresholding (P+T) or LDPred, though the latter requires an additional dataset of individual-level data for the LD matrix component of the algorithm, which would be difficult to obtain because of challenges in data access, as well as due to the underrepresentation of African ancestry individuals in genomics research.

There are several limitations and analytical choices to take note of. Differences in genotyping methods and analysis software could contribute to slight differences in beta estimates, thereby affecting the PRS values. TOPMed genomic data is whole genome sequence data, while UK Biobank genotype data was imputed to the TOPMed reference panel, which may introduce differences in beta values or may make the data susceptible to oversight in strand flip or allele switches. We also converted variant IDs during the use of PRS-CS. Information might have been

lost when converting TOPMed variant IDs to rsIDs, reducing the set of SNPs considered for use in the PRS. However, a strong impact might result from the choice of LD panel used in PRS-CS. In this analysis, 1000Genomes AFR reference panel was used for LD adjustment in PRS-CS, and PRS performance may be improved if another panel more representative of the data used in the GWAS and target cohorts were used, such as the TOPMed reference panel. Additionally, PRS-CS operates on autosomes. Therefore, we manually included the X chromosome in the PRS score, selecting statistically significant variants on G6PD (a gene known to be associated with red blood cell traits) and applying unadjusted effect size estimates in the PRS calculation. Given the strong association of variants on the X chromosome with red blood cell traits, incorporation of unadjusted beta values in polygenic scoring may have significant impact on the results. Another major consideration is the relatively small sample size of the GWAS cohorts used in the derivation of the PRS models. The combination of small sample size, leading to decreased statistical power, and few causal variants may contribute to the overfitting of the PRS model, and therefore reduce PRS prediction accuracy. While we observed the greatest degree of overlap in score assignment in the top 5th percentile (85.4%) after computing a coarse PRS using only SNPs reaching genome-wide significance in each set of summary statistics (PRS-TOP=128; PRS-UKB=117), the low score quantile showed modest concordance (58%); however, 14.6% of individuals were nevertheless classified differently by the two PRS models in the highest-score decile, while 42.3% of individuals were classified different in the lowest-score decile, and observation which warrants further examination. Known associations with RBC traits of G6PD on the X chromosome and the structural variant in the alpha-globin gene region on chromosome 16 in individuals of African ancestry suggest that PRS computation for MCV and related traits may benefit from more specific SNP selection enabled by simpler methods like Pruning and

Thresholding (P+T). Prior studies have shown better predictive performance of PRS for blood cell traits when based on smaller SNPs sets vs. larger ones (Vuckovic et al., 2020). Other PRS methods, such as LD-Pred, which might upweight conserved SNPs or functional SNPs are also important to consider, particularly when working with summary statistics of smaller sample sizes, comparatively (Prive et al., 2020). However, such methods may also require an additional independent dataset for use as an LD reference. Curating SNP sets for PRS calculation using methods that are not limited to the use of autosomes and can reliably incorporate associated SNPs on the X chromosome, and identifying ways to incorporate structural variants, or SNPs accurately tagging strongly associated structural variants, may improve PRS predictive accuracy for MCV.

The underperformance of PRS in cohorts of small sample sizes has serious implications for clinical implementation of PRS. The highly polygenic characteristics and varying heritability of RBC traits demands larger sample sizes for PRS analyses, for both base cohorts (i.e., sample size of summary statistics) as well as optimization cohorts (i.e., the cohort used to tune parameters and optimize the PRS). While increasing sample sizes by improving representation in genomics research study populations can help mitigate the problem of poorer predictive accuracy of PRS, it might not be feasible to do due to recruitment challenges or other barriers. Furthermore, known associations of X chromosome variants on G6PD and the alpha-globin structural variant on chromosome 16 with RBC traits, both found in higher frequencies in individuals of African ancestry (Wheeler et al., 2022), underscores the need to improve PRS computation methods to better include sex chromosome variants and account for structural variants in polygenic risk estimation in order to generate accurate predictive scores, particularly for a population already

known to have higher prevalence of anemia and other blood-related conditions (Chen et al., 2020). Additionally, numerous minor, but impactful, analytical decisions are made when using a single PRS algorithm, and currently, many PRS approaches need large datasets or multiple datasets in order to generate accurate scores. Fluctuations in posterior beta transformations as a function of phi in this analysis suggest that analyses with small sample sizes are sensitive to analytical decisions; a better understanding of the impact of parameter tuning and algorithm choice on PRS results – especially for cohorts of smaller sample sizes or with heterogeneous genetic data – will be important for establishing reliability in routine applications of PRS in a clinical setting.

Further investigation of individual-level concordance between different PRS models for a given trait or disease is needed. A greater focus on the inclusion of underrepresented populations, even if small in population size, will inform future PRS development efforts. To realize the promise of PRS, we must not only ensure that it performs equally well across diverse populations, but we must also make sure that it performs reliably and consistently across different methodological approaches to development.

Chapter 2
Applying a bounded justice framework to examine equity in polygenic risk score use

Introduction

Polygenic risk scores (PRS) are an emerging class of genomic tools used to estimate an individual's genetic contribution to developing complex diseases that are influenced by both genetics and the environment. They are derived from the results of genome-wide association studies (GWAS). Currently, PRS developed from GWAS performed in European ancestry populations poorly predict the risk of disease for individuals of non-European or admixed ancestry, resulting in unreliable test results with limited actionability (Martin et al., 2019a). Due to the underrepresentation of populations of non-European-descent in GWAS, only a small fraction of human genomic variation is captured by current PRS models. The differential predictive performance of PRS risks exacerbating existing healthcare inequities. Specifically, as genetic ancestry can sometimes correlate with socially-defined identities, poor PRS performance for non-European ancestry patients can worsen disease burden and access to healthcare for marginalized and underrepresented groups (Martin et al., 2019; Kurniansyah et al., 2023; Rebbeck et al., 2022). To address this concern, scientists are focusing recruitment on individuals from underrepresented populations to improve the diversity of genetic effects captured by GWAS and generate equitable genomic research (Hindorff et al., 2018; Fatumo et al., 2022).

Achieving equity in PRS development and use is usually focused on the effort to develop PRS which will predict disease risk equivalently across all population groups (Martin et al., 2019a). But framing the solution in this way ignores other inequities, baked into the process of defining populations in genomic analyses (A. C. F. Lewis & Green, 2021a). As described in Chapter 1,

how populations are defined in PRS analysis can impact PRS applicability and generalizability and influence the accuracy of a PRS from certain individuals. Ongoing efforts have highlighted the need to improve uses of population descriptors and definitions in genomics research, but these are oftentimes focused on more visible acts of defining populations, such as in recruitment efforts or in reporting and interpreting results in publications and presentations. However, populations can be defined in more subtle and less noticeable ways in the genomics research cycle, such as in the use of reference panels or in the process of harmonizing raw data in preparation for analysis (Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research et al., 2023b). These practices are typically a result of legacy methods that were incorporated into routine genetic analysis methodology decades ago in parallel with, and oftentimes in response to, cultural and political changes in the way science is executed (Roberts, 2011). Thus, well-meaning efforts to develop equally predictive PRS may fall short of realizing their equitable potential because they are bound by greater socio-historical constraints.

The framework of "bounded justice" (Creary, 2021) can be used as a theoretical framework to better understand these constraints. This framework, which was first developed in the context of contributing to discussions of equity and justice policymaking , suggests that efforts to reduce health and healthcare disparities must be contextualized within the underlying social and physical infrastructures that drive them. In genomic analyses, for example, it is well-acknowledged that the ways in which populations are characterized have significant impacts on the results of a study (Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research et al., 2023b) and, subsequently, the performance

of PRS across different groups of individuals (Ding et al., 2023). Critically evaluating how genomic research infrastructure influences analytical decisions about populations can help to elucidate some of the barriers to achieving equity in PRS.

This paper applies the bounded justice framework to two different components of genomic research infrastructure: reference populations and data models. I focus on these two aspects because they can be easily overlooked yet have large downstream effects on how populations are analyzed in genomic research. Both reference panels and data models are also two components that are crucial to the genomics research methodology, but they are developed outside of the genomic research pipeline. In other words, projects to develop reference panels are different from projects that use those reference panels in analysis. Similarly, data models are oftentimes developed by data model experts who are not necessarily trained in genomics research specifically. Thus, the goals for building both of these research components may not always align with the goals of the analyses that use them. For both components, this paper reviews the social and historical context for the use of population groupings, provides an example of a technical use of these groups in analysis, and highlights the implications of those practices for equity in PRS use. It illustrates how these two overlooked aspects of genomics research infrastructure inadvertently ascribe genetic meaning to socially-defined populations, with real-world consequences for PRS development and implementation in the clinic. So long as associations between genetic information and social identity persist in research practice, increasing diversity in genomic research by including underrepresented populations will be ineffective, and potentially even counterproductive, in the effort to achieve equity in PRS.

PRS, genetic ancestry, and increasing diversity

A polygenic risk score model consists of a set of risk variants, their effect sizes, and a likelihood measure of developing the disease of interest (e.g., an odds ratio) (Choi et al., 2020). The generation of a polygenic risk score model requires use of at least two different population cohorts: the "base" cohort (i.e., the population used to generate the GWAS summary statistics) and the "target" cohort (population used to build the PRS model using the results from the GWAS summary statistics). Sometimes a third, independent reference cohort is needed in order to estimate the allele frequency distribution of the variants in the target population. The predictive accuracy of PRS largely depends on matching the target population to the base population(s), with regard to both genetic similarity and non-genetic similarity; the greater the difference between the two groups, the less accurately the PRS will be able to predict disease risk for the target population (Ding et al., 2023b).

In practice, analysts employ one of several different approaches to generating PRS models that will perform equivalently across groups. In one approach, scores can be computed for separate population groups, oftentimes stratified with respect to genetic ancestry, race/ethnicity, or a combination. This might be done to ensure that variants that are enriched in certain populations (i.e., variants that are more common in one group than others) with stronger associations to the disease of interest are included in the PRS model. Another approach is to develop a PRS using multiple, ancestry-stratified datasets at the same time and present it as a single PRS that can be applied to anyone (Ge et al., 2019; Ruan et al., 2022). A third approach is to develop PRS using pooled summary statistics with no ancestry stratification, which might typically use genome-wide significant variants that are common across populations, and can be implemented

with methods like P+T (pruning and thresholding) (Choi et al., 2020). Scientists are also working on developing methods to make PRS translatable for individuals of any background by applying *post hoc* ancestry calibration methods that readjust the score for a given individual based on a within-test estimate of their genomewide (sometimes referred to as "global") genetic ancestry (Ge et al., 2022a; Rosenthal et al., 2023).

Regardless of method, most approaches to PRS development require some definition of a population and this has implications for generalizability and applicability. For example, how a population is defined determines who belongs to that population group and who does not; oftentimes this is a decision that is usually made by analysts when designing a study, recruiting participants, and/or gathering data. Who does and does not belong to a defined group can impact the sample size of a study cohort; small sample sizes in genomic studies can yield low-confidence results, especially for traits influenced by many genomic variants each with small effects, and requires high power to identify variants with statistically significant associations to the disease of interest. For diseases with fewer variants of stronger associations to the disease of interest, small sample sizes can work, but this limits what health concerns to focus on. Additionally, for traits with strong environmental influences, that are influenced by many different gene variants, and/or are not very heritable, predictive PRS may need to be highly population-specific, and therefore may not be generalizable. In general, as long as a population must be defined in order to create PRS, there will be concerns of applicability, transferability, access, and belonging.

The genomics community has actively pursued efforts to diversify genomics research with the hope of improving genomic discovery and a better understanding of all genetic contributions to disease risk; to the extent that these efforts are successful, PRS performance and transferability is also expected to benefit. In recent years, multiple NIH-funded consortia have prioritized enhancing diversity in genomic analyses, including Population Architecture through Genetics and Environment (PAGE), the Trans-Omic Precision Medicine program (TOPMed), the All Of Us Research Program, the Electronic Medical Records and Genomics (eMERGE) Network, and the Polygenic Risk Methods in Diverse Populations (PRIMED) consortium.

While such efforts to diversify genomics research are underway, a logical next question is whether and how such diversity can be leveraged to promote equitable applications. Capturing a better catalog of genetic variation and disease association across different population genetic backgrounds can indeed improve the performance of PRS for individuals whose genetic background is represented among the populations studied. But an equally important consideration is translating observed diversity into a PRS model in a way that scores will be universally applicable and fairly deployed.

Using bounded justice to examine equity and PRS

Equity and justice in healthcare can be interpreted in many different ways. In this paper, I consider health equity as "the principle underlying a commitment to reducing disparities in health and its determinants" as defined by Braveman et al. Therefore, health equity is intimately tied to addressing health disparities, which Braveman et al. describe as "a specific subset of health differences of particular relevance to social justice because they may arise from

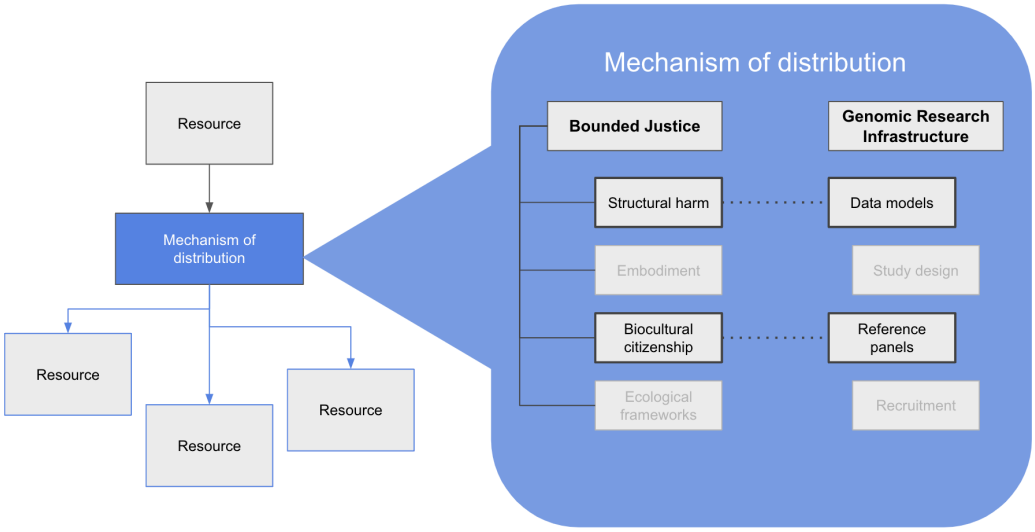
intentional or unintentional discrimination or marginalization and, in any case, are likely to reinforce social disadvantage and vulnerability" (Braveman et al., 2011). These definitions of health equity and health disparities are grounded in the ethical concept of distributive justice and fair resource allocation. However, the just distribution of healthcare resources is not automatically guaranteed. Instead, Bounded Justice contends that fair resource allocation may be thwarted when the institutions and systems responsible for the distribution of goods are themselves unfair due to historical and structural injustices (Creary, 2021).

In this sense, Bounded Justice is a fitting framework to examine what equity means in PRS development, because it integrates the influence of lived experiences into distributive justice practices (Fig. 2.1). It combines multiple ideas and frameworks to help highlight how structural harm and systematic disempowerment of specific groups of people create inequality. Therefore, Bounded Justice asserts that attempts to address inequity are bounded by larger historical and socio-political constraints and in order to design effective policy, approaches must incorporate individual and collective experiences of factors such as colonialism and racism (among other -isms).

Legacies of colonialism and racism persist in genomics research and influence different aspects of contemporary research practice, from study design and methods development to knowledge production and application (Roberts, 2011a). In this analysis, I focus on two elements highlighted by the Bounded Justice framework, namely the potential for structural harm and the influence on notions of biocultural citizenship. Structural harm refers to features of the research process that preserve social hierarchies and power imbalances, even in the production and

interpretation of knowledge. Biocultural citizenship examines the ways that research influences ideas of "belonging" (e.g., social status, disease status) (Creary, 2018; Creary, 2021). These two concepts were chosen because, together, they illustrate how legacy practices can lead to exclusionary applications of genomic research. By focusing on reference populations and data models, this paper uses Bounded Justice to critically assess equity-driven approaches to building PRS intended to benefit everyone equally.

Figure 2.1. Situating bounded justice framework within genomic research infrastructure



Bounded justice and genomics research

Reference panels and PRS

Colonial and racial ideologies on population difference have persisted in the creation of reference panels commonly used in genomics research. Reference panels are comprised of a set of individuals (commonly referred to as reference populations) whose genetic data are used to estimate various types of genomic information in comparison to study samples (Sengupta et al.,

2023a; Taliun et al., 2021b; The 1000 Genomes Project Consortium et al., 2015). The current framework of genomic analysis relies on reference sequences at different stages of the genomic analysis pipeline. They can be used to impute (i.e. infer) unknown/missing genotypes during genomic data preparation; estimate genetic ancestry to account for confounding due to population structure in a GWAS; and approximate linkage disequilibrium (LD) of target populations and allele frequencies of single nucleotide polymorphisms (SNPs) in PRS development. While analysts may also use in-sample estimates of LD and allele frequencies, this requires high computational power and large samples of individual-level data, which may not be feasible to obtain in many cases. Even if feasible to do, analysts may want to compare estimates to a reference such as 1000Genomes in order to contextualize results in the literature.

Early efforts to catalog the diversity of human genetic variation created several of the reference panels commonly used today by analysts. In the late 1990s and early 2000s, international projects such as the CEPH Human Genome Diversity Panel (HGDP), the International Haplotype Map Consortium (HapMap), and the 1000 Genomes Project (1KG) collected data from groups of individuals sampled from distant geographic locations or from individuals deemed representative of "isolated" populations (Leng, 1995; Bliss, 2009).

The motivation for this approach to sampling was driven largely by an interest in novel variant discovery or in making inferences about human demographic history using genetic information. Scholars have critiqued this sampling approach as exhibiting forms of racism and biocolonialism (Greely, 2001). For example, in seeking a globally representative set of samples, HGDP prioritized collecting DNA from isolated populations that were feared to be "threatened with

extinction" (Bliss, 2009); Indigenous groups around the world charged the HGDP project with biopiracy and colonialism (Bliss, 2009; Greely, 2001; Roberts, 2011). Meanwhile, the "super-populations" in 1KG are organized in alignment with typological notions of human populations in part because sampling strategy and data annotation were largely guided by US racial categories or continental groupings that recapitulate racial categories (Roberts, 2011b).

As mentioned earlier, reference panels play a prominent role in GWAS research and PRS development. Analysts using 1KG as a reference panel in their analyses might typically organize their study data in alignment with the reference panel for purposes of genetic ancestry estimation (e.g., by projecting principal components of study data onto principal components of 1KG data to identify which genetic ancestry clusters study data points most closely resemble) or by estimating linkage disequilibrium for PRS development (Koenig et al., 2023). Analysts might also compare self-reported race and ethnicity identifiers with genetic ancestry estimates to see whether those who identify with a particular social identity also share genetic ancestry as a measure of quality control (Khan et al., 2022) (discussed further in Chapter 3). Assumptions of racial or ethnic alignment with genetic ancestry have also been embedded in methods developed to retain the use of study participant data with missing demographic information by "imputing" racial and ethnic identity using genetic ancestry (Fang et al., 2019). Imputing racial and ethnic identity was seen as preferable to misassigning participants to reference panel derived groupings that would be inconsistent with their self-identified race or ethnicity, especially if available those panels did not have data representing the populations that the participants identify with. While this example shows how the unavailability of diverse data can drive the development of controversial analysis methods to compensate for the lack of diversity in genomic research, it

also reveals how ideologies of biological racism can be preserved in genomic research practice. These practices reinforce the idea of genetic concordance with social identity and reflect a racialized and colonialized approach to data collection built into the genomic research infrastructure. Scholars have pointed out that classifying genomic data using typological nomenclature wrongly attributes genetic meaning to socially-defined identifiers (Lewis et al., 2022), but despite these concerns, such resources are commonly employed in GWAS research and PRS development.

There are several reasons why these reference panels continue to be used. HGDP and 1KG are open access resources with high quality individual-level data from samples that represent subsets of global diversity that may not otherwise be accessible to researchers (Koenig et al., 2023). With improvements in sequencing technologies and updates to data curation techniques, new releases of the HGDP and 1KG data reveal new genomic variants, contribute to updated phasing and imputation efforts, and have been adapted for use in cloud-computing environments (Byrska-Bishop et al., 2022). The ease of access, the development of accompanying analytical support tools, and the extensive representation of human genetic diversity available in such collections incentivize researchers to make use of these resources in pipelines for emerging technologies, like PRS.

While the main aim of building reference panels in the early days of genomic research was to support novel variant discovery, the current use of reference panels in PRS development is largely to account for confounding due to genomic population structure and thus to optimize SNP selection for risk calculations (Bliss, 2009; Greely, 2001; Roberts, 2011). Recall that PRS

use reference populations in model development to account for LD and allele frequency differences across populations, and that a PRS developed in one population is less predictive in other populations. Previous work has shown that decreased genetic similarity between base and target cohorts contributes to decreased PRS accuracy, suggesting that matching on genetic similarity is important for developing accurate PRS (Ding et al., 2023b). As genomics research becomes a more global enterprise, scientists are recognizing the limitations of extant global reference panels (Fatumo et al., 2022), promoting the expansion of reference panels to populations from different regional locations.

As mentioned earlier, recruitment for early reference panels was influenced largely by notions of race or continental groupings (e.g., African, European, Asian, Oceania). People who were considered to have mixed ancestry, as well as recent immigrants to the geographic region being sampled, were excluded (Greely, 2001; Roberts, 2011), implying that such individuals do not fit the criteria for membership in a reference population, and prompting the question of what makes someone eligible to be considered representative of a specific population. This can be interpreted as a form of biocultural citizenship; in the case of reference populations, having mixed ancestry or a specific immigrant status prevents individuals from being included in a genomic reference panel, and promotes a link between social identifiers and a particular, largely biological, understanding of a reference population.

This approach to building reference panels for genetic association studies shapes population characterizations and genetic ancestry estimations in GWAS and PRS analyses used today (Bonham et al., 2018), which in turn impacts the predictive accuracy of PRS across populations,

as noted above. Some reflection on the purpose of a reference panel and how to build it can encourage a line of inquiry promoted by a bounded justice framework: What makes a group of individuals, or the categories they belong to, useful as a reference panel? Who decides what those categories are? Who gets membership to what groups? Who controls and monitors the boundaries of those categories? Reexamining the way genome scientists conceptualize the idea of a reference population may be an important first step in the direction of dismantling the legacy of colonialism and racism that persists in genomic research infrastructure. For example, given that the purpose of a reference panel for use in PRS is not about novel variant discovery, but rather to account for genetic ancestry in score estimation, perhaps reference panels could prioritize data that are representative of the populations in the healthcare system it is meant to serve. Or, perhaps a reference "panel" is too limiting, and instead, an interconnected network of databases would be more suitable. Such a network could be structured in a way that links information across different domains of knowledge about a particular disease or trait of interest (e.g., epidemiologic, sociological, environmental, genetic, and clinical) and integrates different types of knowledge systems, such as Indigenous and pre-colonial knowledge systems (Barnhardt, 2016; Menon, 2022). That way, when estimating an individual's (polygenic) risk, comparisons are not made to a single population of presumed similarity, but rather a vast network of populations with different combinations of genetic and non-genetic characteristics informing risk of developing disease. There would, of course, be ethical concerns to take note of, such as issues of privacy, matters of data security, control, management, and access, and possibilities of misuse (Arshad, 2021; Saylor, 2023). With acknowledgement of the obstacles, a collective reimagining of a reference resource for PRS estimation would nevertheless provide an opportunity to rebuild a crucial component of genomic research infrastructure to better serve our

purposes today. But in order to do so, genome scientists must also reconsider the data models used to organize, standardize, and structure demographic data used in genomics research.

Data models and population descriptors

A data model is a conceptualization used for the organization and structure of data elements. One main goal of a data model is to standardize data in a way that supports reproducibility of analyses. It is a fundamental component of research infrastructure that supports the storage, access, and management of the data that researchers use in their analyses. In genomic research, it is typical for research organizations to develop their own schema for organizing and accessing genomic and biomedical data for research purposes (Danese et al., 2019a). In some cases, the transformation of raw data into a data model can be fairly straightforward. It can be as simple as standardizing the variable name (e.g., naming a variable measuring height as "height" as opposed to naming it "HT" or "V1" or any number of other ways a variable might be referred to in a data file), standardizing variable units (e.g., converting height measured in inches to centimeters) or standardizing categorical enumerations (e.g., cancer staging). While it may be simple to transform some types of data into an agreed upon standardized format, it is not so simple to standardize data representing demographic information, such as race, ethnicity, or (non-genetic and genetic) ancestry.

Data models will often consist of data tables, a way of storing data in a tabular form that facilitates convenient storage and manipulation of the data for analysis. In genetics research, it is common for a data model to have a "participant table" which stores information about each research participant. This can include administrative information relating to the study such as the

participant's anonymous study ID number or specifics of the consent agreement they signed, as well as personal information like age, gender identity, or race and ethnicity.

In the United States, the race and ethnicity of a research participant is almost always collected in government-funded biomedical research following the passage of the NIH Revitalization Act in 1993 (*Women and Health Research*, 1994). Therefore, data models are typically built to include field entries for the race and ethnicity of the participant in the participant table. Oftentimes the standardized enumerated values for these two fields are modeled off the US OMB (Office of Management and Budget) census categories for race and ethnicity (Danese et al., 2019a). This means that when an analyst prepares their data in accordance with the data model, they need to transform population descriptors of their research participants to accord with the US OMB categories for race and ethnicity. In practice, this can lead to analysts making ad hoc judgment calls on the identity of an individual (e.g., deciding who should be considered White, Asian, Black, etc.). This subsequently impacts who the results will be applicable to, making this practice yet another subtle, but consequential, example of employing biocultural citizenship in genomics. Social identifiers are not the only descriptors to be pigeonholed by OMB categories. It is not uncommon to use continental labels to describe the genetic ancestry of study participants (Lewis et al., 2022). As was noted above regarding 1KG reference population labels, continental ancestry mirrors OMB racial categories and reflects typological thinking (Bliss, 2009; Roberts, 2011). It creates a direct link between racialized categories and genetic information, and similarly impacts who a PRS would be useful to, further entangling race, biology, and belonging. The genomics community is currently confronting challenges regarding the applicability of PRS to individuals whose identities fall outside the widely-accepted framework of continental

ancestry (Martschenko et al., 2023). While many may diagnose this as a methodological problem of addressing genetic admixture, bounded justice can help us see that the issue of applicability will not be solved by methodological advancements, but rather by a shift in the way researchers conceptualize human populations.

Having fields for only race and ethnicity in a data model, as the sole means of capturing a participant's social identity, is analytically limiting in several ways. Recall that the way in which populations are defined in PRS impacts its accuracy and applicability. If raw population descriptor data must conform to a defined set of categories for race and ethnicity, information that is important to a PRS model may be lost (Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research et al., 2023b). For example, until 2020, there were no descriptors for individuals who identify as Middle Eastern or North African (MENA) in the US Census, which means data models that employ US census categories would have no infrastructure to support population descriptors for MENA individuals prior to 2020 (Office of Management and Budget, "Initial Proposals For Updating OMB's Race and Ethnicity Statistical Standards" 2023). Accordingly, individuals who do not identify with the allowable categories in the data model are either misrepresented (e.g., if incorrectly assigned), obfuscated (e.g., labeled as "other") or not represented at all (e.g., excluded) in the data. This is especially the case for individuals of mixed racial, ethnic, and ancestry backgrounds and for immigrants (Belbin et al., 2021; Martschenko et al., 2023). Following research reports from a mid 2010s investigation on representation of MENA populations in the US (*2015 National Content Test: Race and Ethnicity Analysis Report*, 2015), the OMB has been considering amending the census to include MENA as a separate category. Currently the MENA category is included as a

subcategory under “White” an issue MENA individuals have argued erases their presence, as they consider themselves different from “White” (Maghbouleh et al., 2022). This illustrates both the fluid and political nature of racial categories and the ways that data models that use such categories are subject to political contingencies and limited by the forms of population representation entailed in such frameworks. Relevant cultural and demographic information can be lost early in the research cycle when more specific social identities are prematurely collapsed into broader categories in order to conform to a data model during the preparation stages of data analysis. Previous research has shown that PRS do not transfer well across populations due to differences in cultural and environmental factors, even with similar, broadly-defined genetic ancestry backgrounds (e.g., “African ancestry”), further indicating that the sociocultural information associated with social identity is an important element of PRS accuracy and applicability (McArdle et al., 2021).

Furthermore, race and ethnicity are context-dependent social identifiers that change across social environments, so identifying an individual according to the US OMB race categories may not translate to their social identity in another country (Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research et al., 2023b). Similarly, the population descriptor most relevant to them may not be a descriptor commonly used in the US. In either circumstance, there is the risk of either misrepresenting those individuals’ social identities, or not representing them at all. Lack of representation of admixed individuals and immigrants poses a major problem for the clinical application of PRS and risks exacerbating existing health inequities, an issue that is well-recognized in the genomics community (Sun et al., 2024; Belbin et al., 2021; Martschenko et al., 2023).

The ways in which data are organized influences the ways in which we conceptualize the knowledge we draw from those data (Loomba, 2015a). Data models are responsible for organizing raw data into a form that can be used for analysis, and designing them to rely on US OMB racial categories (or similar systems) to systematically represent social identity is a structural shortcoming that can have major downstream consequences. Using race and ethnicity as the sole fields in a data model that represent something as complex, dynamic, and diverse as social identity, impacts how we perceive the relevance and applicability of study results and risks misattributing genetic meaning to social identities.

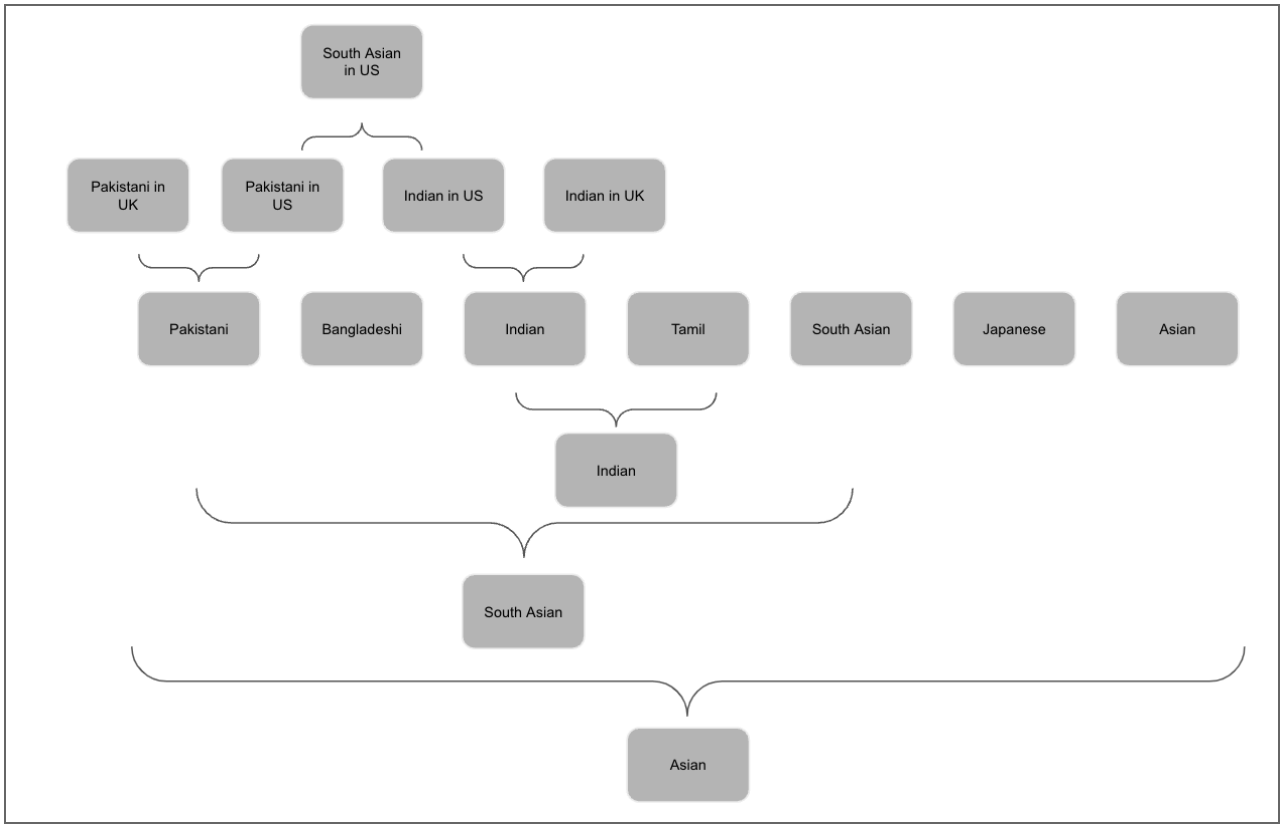
Recent recommendations have encouraged analysts to retain as much granularity in population descriptors as possible during data preparation, analysis, and interpretation (Khan, 2022; Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research, 2023). But as long as the models used for analysis are bounded by a limited set of population descriptors, rich and nuanced information on social identity will be reduced to coarse categories, regardless of how detailed the initial population descriptor variables are. In order to support inclusivity in genomics research and retain detailed data on population descriptors, data models will need to be flexible and versatile. Analysts can then preserve important details about study populations and can make hypothesis-driven analytical decisions on how to group participants in an analysis. A model that maintains population descriptor data in as close to its raw (i.e. original) form as possible also enables traceability and replication, since an analyst would be able to document the choices made with regard to how populations were defined for specific analyses. Given the importance of population definitions on the accuracy of PRS, having

transparency on how populations were defined for an analysis can be beneficial for purposes of reproducibility and applicability of results.

The Polygenic Risk Methods in Diverse Populations (PRIMED) Consortium has recently developed a versatile data model for population descriptors. As a consortium tasked with developing methods to improve PRS for diverse populations, the standard data model structure using OMB race and ethnicity categories as fields in a participant table for capturing population descriptors was regarded insufficient. This was in part because of the international nature of study cohorts involved in PRIMED. Data from global study populations are heterogeneous. How research participants are described can vary greatly across different regional and political contexts; using a limited set of descriptors (from any one or several systems) for describing populations would not adequately account for the different populations and communities included in PRIMED. Instead, the PRIMED data model includes a data table specifically designed to capture and organize population descriptors of any type (e.g., race, ethnicity, nationality, tribe, etc.), and it allows any value for each descriptor type (Figure 3; for more detail, see Supplementary Table 1). This enables analysts to retain in the data model all details pertaining to population descriptors of the study population. Then, when an analyst prepares data for a specific analysis, they can organize study cohorts in whatever way makes the most sense for their analysis. With such a data model, analysts can implement recommendations on how to incorporate population descriptors into their analyses responsibly. On top of that, a model like this enables traceability and transparency, and it facilitates making documentation that delineates how groups were defined and how decisions were made. The flexibility of the model enables an analyst to make analytical choices about how to group study participants in a way that makes the

most sense for the research question. While the responsible use of population descriptors in genomics research ultimately lies with the analyst, a data model that accepts any type of population label will not inherently systematically discriminate against any one kind of social identifier.

Figure 2.2. Example schematic for label-mapping process



Conclusion

Diversifying genomics research by prioritizing the inclusion of underrepresented populations in studies is commonly endorsed as a necessary first step in order to develop equitable PRS. The hope is that the analysis of heterogeneous data will provide novel insight into biological mechanisms of disease and enable the development of precision medicine approaches to better serve individuals of those communities. This frames the issue as a technical problem that can be

addressed with a technical solution (i.e. better recruitment and sampling approaches, improved methods). But decisions about ancestry and population groups in PRS are not value-neutral and apolitical, as demonstrated by a closer examination of genomic research infrastructure. While including historically excluded and marginalized communities is necessary for even the possibility of developing equitable PRS, it is not sufficient as a standalone response. The framework of bounded justice can help scientists and policymakers identify and understand the web of factors that contribute to inequity in genomics, disproportionate burden of disease, and disparities in healthcare.

Reference panels and data models are two components of genomic research infrastructure in which racialized ideologies of population difference are preserved. Once genetic association with social group membership and disease is embedded in tools like PRS, it affects attitudes and practices in ordinary ways to become part of a legacy of explaining health disparities and disease outcomes using a biological rationale (Liao and Carbonell, 2023). Reconstructing reference panels to reflect the goal of developing equitable PRS and implementing more flexible data models provides opportunities for more inclusive research practices. Reforming research infrastructure can help minimize risk of typological thinking in genomics by making it easy and expedient for researchers to organize and apply data in ways that do not ascribe genetic meaning to social identifiers.

Understanding the socio-historical context of the current genomic research enterprise can guide scientists and policymakers in their well-intentioned efforts by identifying key issues that are inextricably linked to the structural disenfranchisement of the people they want to support.

Research infrastructure is one piece of a larger network of genomic research practice. By recognizing the broad array of constraints, the genomics community can take a more holistic approach to developing policies and programs, and devise coordinated efforts across different domains for intervention and reform.

Chapter 3
"My science is not separate from who I am and the values I have"
Genomic scientists' perspectives on the use of ancestry in genome-wide association studies and polygenic risk scores

Introduction

A polygenic risk score (PRS) is a new type of precision medicine tool that estimates an individual's genetic liability for developing a disease or trait (Choi et al., 2020). PRSs use single nucleotide polymorphisms (SNPs) identified by genome-wide association studies (GWAS) found to be associated with a particular trait or disease of interest. PRSs are anticipated to support existing clinical risk estimation tools and ultimately improve screening methods, health care delivery, and health outcomes (Dikilitas, 2020; Kullo, 2022). Currently, the use of PRS in a clinical setting requires some determination of the patient's population group membership. In other words, for a PRS to accurately predict disease risk in a patient, that individual must closely resemble the study population that was used in the construction of the PRS, both in genetic similarity and in a socio-environmental context (Choi, 2020; Ding, 2023). Thus, the ways in which populations are conceptualized and operationalized in GWAS and PRS construction have implications for the applicability and generalizability of a PRS.

Due to the overrepresentation of European descent individuals in genomic research, PRS have largely been developed in populations with European ancestry and predict disease risk for European descent individuals better than for individuals of non-European descent of mixed ancestry. This has led to serious concerns that PRS may exacerbate existing health disparities (Martin et al., 2019), and there has been a concerted effort to develop PRS methods that are equally predictive for everyone, regardless of background. As a result, researchers have directed

their efforts towards exploring (1) the ethical considerations for using population descriptors in genomics research and (2) the statistical considerations and novel analytical methods needed for the analysis of heterogeneous data afforded by inclusion of diverse populations (Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research et al., 2023b; Peterson et al., 2019; Zhang et al. 2023).

However, how these two domains relate to one another has only partially been explored, and the lack of integration limits the impact of ethical and analytical investigations. If recommendations grounded in ethical concerns do not take into account study goals and analytical constraints, they may not be feasible to implement. Similarly, if analytical methods are implemented without taking into account ethical considerations, then study results can inadvertently perpetuate harmful conclusions. By considering ethical and analytical concerns together, scientists will be better positioned to develop recommendations and methodology that achieve the intended goals.

To gain a more in-depth understanding of how ethical considerations are (or, at least, could be) incorporated into analytical practices of performing GWAS, we conducted a qualitative study involving interviews with genomic scientists (in this case, referring to those who conduct research using genomic data) who have familiarity with GWAS and PRS analyses. Using my own experiences from the quantitative analysis performed in Chapter 1 and the insights I gained from the ethical analysis conducted in Chapter 2, I set out to better understand how professional genomic researchers view the use and the challenges of using race, ethnicity, and ancestry in their analyses. Capturing the experiences, challenges, and goals of genomic researchers can expose where efforts to implement ethical considerations into analytical practice fall short,

highlight areas where integration of ethical and analytical practice succeed, and identify opportunities for further intervention.

Materials and methods

Study design

I conducted semi-structured confidential interviews with genomic scientists who had direct experience with conducting GWAS and/or PRS analyses in diverse populations. Interviews were conducted virtually, audio-recorded, and transcribed for subsequent qualitative analysis. The motivation for this work was to better understand the decision-making process that researchers experience when using race, ethnicity, and/or genetic ancestry in their genetic analyses. All study activities were reviewed and approved by the University of Washington Human Subjects Division and Institutional Review Board.

Recruitment

As the focus of this study was on understanding the methodology for constructing PRS, including discussions on conducting genomic analyses in diverse populations, we sought to recruit investigators with familiarity of different GWAS and/or PRS methods. We focused our recruitment, in the first instance, on investigators involved in either the Polygenic RIsk MEthods in Diverse populations (PRIMED) or the Trans-Omics for Precision Medicine (TOPMed) consortia, beginning with contacting researchers in leadership positions (e.g., a study/site principal investigator or a working group co-chair). We then employed a snowball sampling method where we asked each respondent for recommendations of researchers who would be interested in participating in the study.

The goal was to recruit individuals with different training backgrounds, including but not limited to genetics and genetic counseling, population genetics, biology, medicine, and epidemiology. Individuals were eligible to participate if they had some experience conducting GWAS or PRS analysis. To capture various experience levels, we included participants from post-doctoral researchers to principal investigators; we excluded graduate students because we wanted to focus on perspectives of researchers who had completed their training and had experience with conducting independent research studies. We also did not offer incentives for participation. Using a shared calendar system set up by the first author, participants were invited to select a 1-hour time block that was convenient for them. A Study Information Sheet containing details about the purpose of the study, study procedures, and the risks and benefits was provided along with the initial contact email. Interview candidates responded via email either confirming or declining participation in the study. For each participant who accepted the invitation, verbal consent was obtained at the start of the interview session, prior to recording. Of the 35 invited, a total of 18 investigators accepted and successfully completed the interview, 4 declined and 13 did not respond. Of the 4 that declined, 2 suggested alternative investigators to contact. One participant was based in Europe, while all other participants were based in the United States. Of the 17 participants residing in the US, six identified as "being from" outside of the U.S (e.g., being born outside the US and/or having grown up outside the US), including Europe, Asia, and Africa.

Interview guide

The interview guide was developed based on author experience of study design and genomic analysis, and was informed by conducting an ethical analysis of health equity in genomics and medicine (see Chapter 2). Specifically, interview questions spanned three overarching domains

centered on the use of race, ethnicity, and genetic ancestry in GWAS and PRS research: (1) Analytical considerations; (2) Ethical considerations; and (3) Researcher positionality. Questions from domain (1) focused on understanding the methods that the participant uses when incorporating or adjusting for population background and the reasoning surrounding those choices. Questions from domain (2) addressed issues of equity in PRS implementation and perceptions with respect to how equity can best be achieved in this area. Questions from domain 3 invited participants to reflect on their positionality relative to the research they do and sought to explore whether their personal social and/or professional identities influence how they approach their work. Demographic information (Table 3.1) was collected at the end of the interview by verbally asking participants about their social identities (providing standard OMB options) and their field of profession, or through follow up communications. The interview guide was refined through four pilot interviews before proceeding with official interviews. Data from the pilot interviews are not reported here.

Data collection

Interviews were conducted by the first author over the course of five months (from December 2022-April 2023) over Zoom. We reconfirmed verbal consent from each participant at the beginning of the interview and sought permission to record. Completed recordings were subsequently transcribed and redacted for identifiability. Each transcript was assigned a unique identifier to ensure confidentiality, and a participant ID link was safely stored on a secure, HIPAA-compliant cloud platform. All recordings were deleted after transcripts were generated and checked for accuracy. All interviews were conducted in English.

Data analysis

Qualitative analysis software Atlas.ti v.9 was used to support coding, analysis, and data management. Transcripts were analyzed using a blend of directed and conventional content analysis (Hsieh & Shannon, 2005a). A mixed approach of top-down and open coding was used until code saturation was reached (Hennink et al., 2017; Maguire & Delahunt, 2017). We prioritized coding of expressed and implied perspectives on (1) the value of various methods employed to incorporate population descriptors in analysis; (2) concerns about the consequences of methodological choices for research and clinical interpretation of study results; and (3) personal reflections on the positionality and responsibility of the researchers. One of the authors acted as a second coder and independently coded and reviewed a 4 out of 18 transcripts. Coding differences were resolved through discussion. We amalgamated the final codes into overall themes and translated them into concepts as they emerged from the data. During the late-stage analysis, we used the Atlas.ti metrics of "groundedness" (i.e. total occurrence of a given code) and "pervasiveness" (i.e. occurrence of coded text over unique transcripts) to maximize our ability to identify all relevant themes.

Results

In total, we invited 35 investigators via email to participate in a 1-hour confidential interview (51.4% response rate). The participants in our study represented four general disciplinary fields, broadly categorized based on the professional titles they provided (Table 3.1). Of the 35 invited, a total of 18 investigators accepted and successfully completed the interview, 4 declined and 13 did not respond. Of the 4 that declined, 2 suggested alternative investigators to contact. One participant was based in Europe, while all other participants were based in the United States. Of the 17 participants residing in the US, six identified as "being from" outside of the U.S (e.g.,

being born outside the US and/or having grown up outside the US), including Europe, Asia, and Africa. There was a broad range in age (29-67 [49]) of the participants, and the number of years of experience working in Genetics ranged from several years to several decades. All participants reported experience with conducting genome-wide association studies and most participants had generated PRS themselves or had team members who did so; those who had not performed PRS analyses directly were familiar with the computation of such scores and their intended applications. The majority of participants identified as non-Hispanic White according to the US OMB (Office of Management and Budget) categories.

Table 3.1. Participant characteristics

N		18
Age		
	Range (mean)	29-67 (46)
Race (OMB, self-report)		
	White	14 (77%)
	Black	1 (5%)
	Asian	2 (11%)
	Other	1 (5%)
Ethnicity (OMB, self-report)		
	Hispanic	1 (5%)
	Non-Hispanic	17 (94%)
Field (profession)		
	Epidemiology	9 (50%)
	Genetics/Population Genetics/Statistical Genetics	5 (28%)
	Medicine	2 (11%)
	Informatics/Computation	2 (11%)

Three major themes emerged from participant perspectives on the use of ancestry in PRS development and its impact on creating equitable scores, summarized in Table 3.2. In general,

participants acknowledged the need for more diverse data in order to produce findings and risk estimates that would be applicable to everyone. There was overall agreement that some use of ancestry is needed in order to develop accurate PRS, and that in addition to having greater representation in study data, more specificity with regard to the types of data collected (e.g., more granular information such as country of birth, geographic residence, social and environmental data) is necessary to more precisely characterize populations. Participants also commented on their personal views on the use of ancestry in GWAS and PRS, with many respondents advocating for improvements in methodological approaches for diverse study populations.

Table 3.2 Domains and themes

Domain	Theme 1	Theme 2	Theme 3
Analytical	Diversity in data is necessary, but questions about integration into existing infrastructure and methodology remain.	Convenience drives the use of ancestry, yet definitions of ancestry are contested.	Research methodology can be at odds with researchers' ideals.
Ethical			
Personal Reflection			

Theme 1: Diversity in data is necessary, but questions about integration into existing infrastructure and methodology remain

Need to diversify study populations

Participants widely acknowledged that improving diversity and representation in genomics research is imperative for the production of science that is applicable and beneficial to everyone. All participants acknowledged the need to improve diversity across study populations as well as among the genomics research workforce. However, many respondents raised practical questions, such as "what do we [i.e., the broader members of the genomics research community] mean by

diversity" (A312), how to access it, and how to use it in an advantageous and responsible manner.

When asked about what is needed to achieve equity in PRS development and validation, all respondents agreed that, for genomics research, improving representation from underrepresented populations is paramount. Some respondents believed that the genomics community cannot "effectively move forward without having data from those [i.e., underrepresented] populations" (A304), and that "you can develop all the complicated analytical methods you want, but if you don't have the data there from the people you want to study…you can only do so much with what you have" (A312). Others, however, indicated that having access to diverse data can only help so much, and that a focus on methods development will also be critical in order to generate equitable PRS. As one participant explained:

"...my stance is data can't be the only solution…I love the fact that we're focusing on increasing diversity in data sets, but I feel like we're never truly going to catch up to representing diverse genomes in proportion to their actual frequency in the world's population. Like, we're not going to stop sequencing white people. So in that sense, what excites me is that I think methods have to be part of the solution here." (A303)

Implicit here is the concern for sample size. This participant was not alone in conveying apprehension about ensuring adequate representation of all people in genomics research. Other respondents commented on challenges of building PRS for small communities, where a GWAS might not be able to be powered enough to identify variants relevant to disease, and therefore a PRS would not be as predictive for that population. Overall, participants acknowledged that the practical effort to improve diversity in genomics research is a complex, multidimensional

endeavor. It spans the full spectrum of the research cycle, from study design and recruitment, to methods in data preparation and analysis, to interpretation and translation of results.

In addition to improving representation, determining the correct sampling frame for a research study is another methodological choice that an investigator needs to make when increasing diversity is a priority. As one participant noted, the ultimate aim of the study matters to such decision-making:

"So if your aim is to find new biological insights, you should not focus on a few big global populations. We should stop studying Europeans, that's for sure. And we should go and start looking at every isolated population we can. If the interest is in using PRS for clinical translatability, I guess that's a different purpose. And then you really got to build up the populations that you want to serve." (A311)

Many participants mentioned downstream clinical relevance as an important guide to sampling design. For a PRS to be clinically relevant within a health system, for example, "all ancestries represented in that health system must be included in the generation of the model…so the polygenic score must be generated with data that adequately represents that population" (A310). Participants reflected on how this would practically be achieved, and most acknowledged that recruitment of diverse populations in genomics research would require a great amount of time and money, among other resources. Some participants were hopeful that consortia dedicated to PRS methods development and implementation, such as PRIMED and eMERGE, would help to address these concerns.

Need for new methods

Respondents also expressed optimism with regard to generating equitable science through the use of novel methods. They recognized that the requirement of many large datasets to generate

reliable PRS is a challenge that has yet to be addressed. Several respondents noted that for some communities that are small in number, a GWAS may yield less reliable results due to a higher possibility of false positive findings. Yet combining populations of small sample sizes with other populations in a pooled analysis may risk obscuring signals of variant associations due to overrepresentation of other populations in the study.. One participant proposed adding a penalizing factor to PRS models that have overrepresentation of one population as a way to incorporate an equity measure and level the playing field. However, more respondents seemed hesitant when asked about the reliability of performing GWAS in populations with small sample sizes and developing PRS models for individuals from those communities. One participant suggested that framing association studies to focus on aggregate level effects (e.g., on gene regions or pathways) to "move beyond the SNP" (A303) could help get traction for performing association testing with small sample sizes, since fewer elements would need to be tested. Participants also noted challenges with data sharing policies that make access to individual-level data difficult, leaving analysts to forego studying data from those populations or to work with summary statistics and meta-analysis methods to perform PRS analyses. As one participant described:

"You know, obviously…if you need individual level data, it's much harder than if you can get by with summary statistics, which is particularly important if you want to grow your sample size and you cannot get from somebody the individual level data which, is like, 'ohh it takes many years to get all these agreements, whatever, in place.' So methods that can use summary statistics definitely are often the winner just because of sample size access" (A313)

Other aspects of study design were highlighted as important factors in the development of equitable PRS, and that employing greater variability in approaches to study design might help address remaining equity concerns (e.g., systematically examining PRS performance across a variety of environmental contexts). One example given was the pursuit of more sensitivity

analyses following a GWAS since GWAS results are "essentially correlations in very arbitrarily-defined populations" (A314). However, the same respondent noted that different study types "are not currently valued and rewarded and rarely followed up on in any meaningful way" (A314). In order for change to occur, taking different approaches to study design must be incentivized in the field, they argued.

Need for new infrastructure

In addition to diversifying study populations and developing new methods, respondents also pointed to a need for changing research infrastructure to be able to accommodate the influx of heterogeneous data. Many respondents discussed the importance of incorporating non-genetic information into analyses, such as variables representing social determinants of health. It was often noted that enhanced representativeness must expand beyond the consideration of race, ethnicity, and ancestry, not only to improve the accuracy of PRS, but also because "there are other ways we could be partitioning people that might be equally interesting" (A303). Some examples participants offered included diversifying sampling across age, gender identities, socioeconomic status, geographic location, and health status, among other factors. However, while participants mentioned the *importance* of different types of information to include in a PRS, they generally agreed that some use of ancestry is *needed* to develop and use a PRS model. While different types of non-genetic information can provide more clarity on the multiple factors influencing disease, patterns of genetic variation, represented by genetic ancestry, remains necessary to account for in risk estimation in order to accurately assess the variants contributing to disease onset.

Theme 2: Convenience drives the use of ancestry, yet definitions of ancestry are contested

Participants were asked why genetic ancestry is important to, and what role it plays in, the development of a PRS. The most commonly reported reason was to address confounding due to population structure as a result of differences in linkage disequilibrium (LD) and allele frequency across populations. Studies have shown that the greater the genetic distance between the source population and the target population in a PRS model, the less accurate the PRS will be in estimating risk for the target group (Ding et al., 2023b). Most respondents described a common procedure that they follow to account for genetic ancestry is to perform principal components analysis (PCA), a mathematical method that summarizes patterns of variation in data (in this case, single nucleotide polymorphisms (SNPs)), with the top principal components (PCs) explaining the greatest amount of variation. There were generally two applications of PCs described: (1) PCs were used as a covariate in analysis to represent potential association effects due to genetic ancestry, and (2) as a tool to identify groups of genetically similar individuals that represent variation in genetic ancestry. Using PCs to identify genetic similarity is oftentimes used in conjunction with stratification by race/ethnicity to identify population groups. Several participants described performing a comparison with reported race/ethnicity as an internal consistency check (to guard against sample mix ups) and to identify outliers (i.e. individuals for which shared racial/ethnic identity and genetic ancestry [as inferred from PC relationships] do not align). One participant described the process in this way:

"I think the most common current approach is to stratify populations by ancestry ideally, but often that kind of starts off as looking at by race/ethnicity, and then doing a principal component analysis to see if people really fall into the given race/ethnicity that they self reported, and if they don't, then we exclude people who are outliers, or who don't fall neatly into one category" (A304)

This reflects a change in perspective on the utility of PCs in genomics research from over a decade ago when PCs were being introduced into research practice as an alternative to using race to account for population structure. Prior research suggested researcher ambivalence towards the use of PCs to adjust for population structure, in part due to its lack of interpretability and the potential for misinterpretation of study results that use PCs as an adjustment tool (Yu et al., 2012). In our interviews, all participants described their routine use of PCs to account for population structure; most respondents acknowledged that the PC values themselves are not directly interpretable, but that they are a representation of the variation in the study sample, capturing differences in allele frequencies and LD patterns.

While all participants mentioned a similar procedure to control for confounding, there was nevertheless variability in how ancestry estimation was approached analytically. As one participant described:

"The genetic ancestry component to it is also super complicated because it depends on, you know, how do you actually infer the genetic ancestry? And there's a term called genetically inferred ancestry, GIAs. Essentially, you can infer genetic ancestry at the level of a continent, at the level of a country, at the level of different groupings. And essentially there's no unique GIA. Because it all depends on the reference panels. It depends on the method that you're looking at, and it also depends on how do you define these boundaries, right? How do you define this clustering?" (A318)

As this participant mentioned, there are many factors that influence how genetic ancestry estimation is approached in practice, and one analytical decision (e.g., choice of reference panel) can impact downstream analytical choices (e.g., inclusion/exclusion of data points). For example, analysts typically select the number of PCs to use as a covariate in a GWAS. This choice currently seems to be done *ad hoc* and for different reasons. For example, some investigators

decide on the number of PCs that explains the most variation in the data; others use whatever is standard practice in their labs. In other cases, analysts use what previous studies have reported using, particularly when working with summary statistics in meta-analyses. Others prefer performing PCA in an already-stratified set of samples to identify finer population structure due to concerns regarding the influence of fine-scale population structure on rare variants that contribute to disease. However, analysis of smaller groups leads to complications in how populations should be defined, which variants to keep in a PRS analysis and who that would impact, and overall, it can make PRS less generalizable, and therefore less applicable to other populations.

Some participants suggested that analysts could eliminate the need to use ancestry groups by performing in-sample relatedness calculations or building a universal reference panel instead that would include unstratified genetic information along with non-genetic information, such as social determinants of health and detailed information on social identifiers. Nevertheless, they acknowledged that this would be possible only with access to large amounts of individual level data and high-efficiency computational resources. Currently, using summary statistics to develop PRS models typically requires some matching on population group due to the use of a LD matrix in PRS calculations, which is typically categorized by population, as different populations are known to have different patterns of LD. The LD matrix is typically used to help identify the variants that are in LD with one another and therefore, impacts the effect of that variant in the PRS calculation. Hence some definition of genetic ancestry to define populations. While interviewees generally believed that changing this approach to sample stratification was

methodologically feasible, they seemed dubious of its probability due to a general reluctance to modify established approaches. As one participant stated:

"A lot of the polygenic score methods right now require a linkage disequilibrium matrix… even the ones that use multi-group, multi-ancestry, it requires you to split and then have these arbitrary cut offs. So that's an area of development that would require active efforts to get away from that. And I don't see a lot of effort on that right now, just because it's really hard to sort of break out of that way of thinking" (A314)

It seems that, in addition to race, ethnicity, and principal components, technical constraints (e.g., sample size, access to reference panels, availability of individual data vs. summary statistics) influence how ancestry is inferred for an analysis. The convergence on an overall approach (e.g., combination of race/ethnicity and principal components) seems to imply the routinization of defining ancestry groups for PRS development, but the variety of methods to do so suggests that "ancestry" or "population group" remains an elusive concept that is hard to subject to the same rules of precision and rigor that other variable definitions are subject to in genomic analyses.

Theme 3: Research practice and methodology can be at odds with scientists' ideals

Many participants shared their personal perspectives on the practice of genomics research and the state of the field in general. The majority of participants had a positive outlook on the future of genomics and the clinical utility of PRS, in particular. At the same, the majority of participants also expressed that they (1) had conflicting feelings about current standard research practices or (2) had general concerns over the state of genomics research with regards to diversity, inclusion, anti-racism, and the pace of change in the field. As one participant explained:

"*It is hard, so hard to change procedure…once you get focused on how you're getting things done, you're gonna get things done in the same way that you have always gotten things done.*" (A310)

Several participants conveyed dismay at the approach of identifying genetic clusters using PCA, particularly noting concerns about the exclusion of participants who are found to not fall within genetic clusters identified by PCA.

"*So we'll use that [PCs] as a QC [quality control] check and look at each self-reported race/ethnic group for outliers. You may do it and you may see some extreme outliers. This is actually an area that bothers me, because I think what we usually do with those people is we exclude them. You know, I hate the idea of excluding people, because they work just as hard as anybody else to contribute*." (A308)

Multiple participants predicted confidently that, if their own genetic data were collected for an association study, they would be identified as one of a small sample, or as an outlier, and likely excluded. This is because individuals who share their genetic ancestry are not currently adequately represented among those for whom GWAS have been previously conducted. Therefore, and wholly ironically, these respondents felt that PRS would be minimally applicable to them, if at all. Many interviewees included some mention of their own identity in their responses, including researchers who self-identified as White, as they felt themselves underrepresented in other aspects of their identity. Participants were also invited to reflect on their own positionality in their work as genetic researchers, with some respondents having a neutral attitude or remarking on their position of privilege within the genomics research community. Others examined their sense of belonging, and commented on their contributions to a field that repeatedly excludes from investigation the communities with which they are affiliated. As noted by one participant:

"I want something to mean something, right? And right now, nothing that currently exists in the translational space for genetics applies to really anyone in my family…And so like, what's the point? And I think part of it's like, I should be allowed to be angry about that, right? I should be allowed to be like, 'hey, my science is not separate from who I am and the values I have.'" (A314)

Over the last decade, genomic scientists have acknowledged their responsibility to foster inclusive research practices and generate equitable science. Similar sentiments were shared by several participants in this study, who took pains to explicitly state the problematic history of genetics as a field and the need for researchers to deliberately take action to avoid reinscribing past wrongs. As one participant decisively put it:

"I think a phrase we often put a lot of weight in as scientists is 'do no harm'. And I think currently, our research has the potential to do harm, and it is our responsibility to mitigate that potential…I think as a field, especially as a field that has its roots in racism, racial biology, and eugenics, it is our duty, it is our utmost ethical duty, to not do harm, and to try to acknowledge, undo, and interact with the harm that we've already caused." (A301)

Generally, participants felt that many members of the genomics community are aware of the potential harm of taking a dispassionate approach to using race, ethnicity, ancestry, and other population descriptors in genomics research; however, they soberly remarked that it would require cross-disciplinary, coordinated efforts to replace and rebuild research infrastructure and instill new research practice.

Discussion

While prior literature has shown an evolving and heterogeneous understanding of ancestry among genomic scientists and across the biomedical field (Byeon et al., 2021; Dauda et al., 2023a), this study specifically examined the ways in which genetic ancestry is understood and used analytically in the development of PRS and the challenges analysts face when doing so. When asked what the role of ancestry is, and whether it is needed, in PRS development, participants generally insisted that the use of ancestry was context-dependent: it depends on the research question or it depends on the methods being used. This is consistent with prior research

investigating implicit and explicit understandings of ancestry based on how it is used in genetic analysis (Byeon et al., 2021; Yu et al., 2012).

This study reveals a growing tension between the ways that ancestry is understood and how it is used in PRS. Participants described ancestry as a complex, dynamic concept with social and genetic aspects to it, both of which influence health outcomes, yet isolating these two factors into separate components for use as scientific variables in a genomic analysis remains a major challenge. Researchers in this study frequently noted the need for including more granular population descriptor data and more social determinants of health data in GWAS and PRS analyses in order to improve accuracy, suggesting that genomic scientists are aware of the deeply entangled relationship between social and genetic causes of disease. Another challenge in using ancestry in PRS is its inconsistent (and unpredictable) relationship to social identity, especially for individuals of mixed backgrounds, making it difficult to assess the absolute risk of disease for a given population. This study revealed that it is common practice for researchers to compare inferred genetic ancestry with reported racial/ethnic identity to identify outliers to exclude prior to analysis. The use of PCs (seen as a representation capturing allele frequency differences and patterns of population structure) have become preferable to the use of ancestry informative markers (AIMs) over time. However, it appears as though the rationale driving its use is similar - namely that human population groups can be identified using genetics, even if no longer through AIMs, but instead through PCs. Though previous studies have illustrated a correlation between genetic ancestry and social identifiers like race and ethnicity, such a correlation is both a poor and risky model for their relationship. As noted earlier, social identity is fluid and therefore does not align with specific patterns of genetic variation. However, it is sometimes a difficult

association to avoid, particularly when using data that was collected and curated decades ago (i.e., legacy data) when alignment of race and genetics were cemented in data processing techniques, such as using race-specific imputation panels. When using such legacy data today, it can be onerous and sometimes infeasible to reprocess the data in a way that does not tie race and genetics together. As a result, researchers may repeatedly find themselves in debate about whether to use that data anyway, find alternate uses for it that would not promote racial associations with genetics, or to discard that data altogether. Because genetic ancestry cannot reliably be mapped to social identity - which is generally the kind of information available in clinical practice - there remains a large discrepancy between how ancestry is being used in PRS development and how it is being interpreted in clinical practice. If clinical implementation of PRS is important for guiding PRS development, as many participants in this study noted, this disconnect between genetic ancestry and social identity will continue to be a problem.

There are several limitations to this study to take note of. The majority of participants in this study are members of consortia dedicated to improving PRS performance for diverse populations, which may contribute to a response bias regarding the use of ancestry in PRS and the challenges discussed. Previous research and participants in this study have indicated that the specific goals of a research study can influence how ancestry is perceived and used in analysis. Hence, the overrepresentation of scientists with similar research goals may have led to disproportionate portrayal of challenges in using ancestry in PRS development that may not be perceived by other scientists who are not involved in PRS development for diverse populations. Second, because this was a self-funded research project all study participants were volunteers who were not compensated for their involvement in the study. The sample may, therefore, have

been skewed towards scientists who are intrinsically motivated to tackle questions about ancestry in PRS and may have more experience experimenting with novel approaches to use, or not use, ancestry, as opposed to researchers who may be less aware or less interested in topics of ancestry application in genomic analyses. Finally, given the focus of this study, it is possible that participants may have felt obligated to share socially or ethically acceptable responses, or were aware of the interviewer's views on the subject matter and so shared what they presumed the interviewer would have wanted to hear. However, this seems unlikely because there were multiple occasions where participants verified their anonymity with the interviewer when sharing their thoughts, suggesting that they were willing to share views that may be seen as unpopular.

Insights from this study build on a legacy of empirical research examining the use and interpretation of ancestry in genomic research and medicine. Ancestry continues to be an elusive concept in scientific practice, and scientists are still figuring out whether and how it fits into different types of genetic analyses. Especially when researchers perceive the role of ancestry to be context-dependent, more efforts are needed to elucidate when genetic ancestry is relevant and when it is not. To examine these questions in more depth, genomics research infrastructure will need to expand its boundaries to better handle the complex relationship between ancestry and disease, but it will also need to recognize its own limitations.

Conclusion

Polygenic risk scores, and their development and implementation in diverse populations, present an interesting opportunity for us in the genomics community to interrogate our own understandings of social ideology and scientific research pursuits. This research took a multidisciplinary approach to examining the ways in which race, ethnicity, and ancestry are considered analytically and ethically in PRS development, and the impact their application has on building equitable PRS. Here, I summarize the unique scientific contributions of this work, highlight areas for further inquiry, and offer ideas for moving forward.

Summary

This research described here contributes to the literature in three important ways: First, it adds to a growing body of work exploring PRS development and application to diverse populations by centering study goals, design, and analysis in quantitative, ethical, and qualitative projects on PRS in underrepresented populations. Second, it offers new understandings of the role of genomic research infrastructure in striving for equity in PRS performance. Third, it reveals novel and interesting insights into the perspectives of genomic scientists on the use of ancestry and other population descriptors in GWAS and PRS.

There are many variables to consider when developing a PRS model that can impact the results, such as what PRS algorithm was used, how parameters within the algorithm were optimized, as well as what populations were chosen for testing and validation and why. While developing methods to make PRS "transferable" across populations is necessary, it is also important to assess how concordant PRS models are within populations that are - conceptually at least -

regarded as the same, however they may be defined. The study described in Chapter 1 found discordance between two PRS models developed from different summary statistics. Both models were optimized in the same population cohort, using only data from individuals of African descent. The disagreement between the two models was likely due to a combination of technical differences (e.g., analytical software used, randomness introduced by algorithm), analytical choices (e.g., selection of reference panel, small population cohort size), and differences in genetic and non-genetic cohort characteristics (e.g., environmental exposures and different demographic histories, hence different patterns of genetic admixture). Further work is needed to better manage genomic association studies of underrepresented populations, especially while access to diverse data remains limited. Phenotypes like red blood cell traits are highly polygenic, influenced by population-enriched variants, and impacted by environmental differences. Yet developing highly population-specific and/or environmentally-specific PRS would not only limit who can benefit from a PRS, but it would also be untenable to develop. Additional research to improve the PRS concordance and reliability within and across populations will be important for the equitable clinical applications of PRS.

The practical challenges I confronted in Chapter 1 gave rise to a series of questions, several of which I explored in the ethical analysis described in Chapter 2. How are populations used in GWAS and PRS development defined and who defines them? What is the purpose of a reference population and who can be considered a part of one? Who controls and monitors the boundaries of population definitions and categories? How do new categories get created and by whom? These questions, and their consideration through the lens of the ‘bounded justice’ ethical framework, put a spotlight on the hidden and embedded forces that keep legacies of colonialism

and racism alive in genomic research. Scholars have documented how the persistence of colonial and racial ideologies in science exacerbate health inequities (Fields & Fields, 2022; Loomba, 2015a; Roberts, 2011b). Thus, regardless of whether or how much diverse data is obtained and novel methods developed, PRS will not be equitable so long as remnants of colonialism and racialized thinking remain built into genomics research infrastructure, which then allows these outmoded ideologies to persist. To my knowledge, this ethical analysis is the first of its kind to apply the framework of Bounded Justice to critically examine approaches to building equitable PRS.

Experiences from both of these research endeavors informed the discussions I had with genomic scientists performing GWAS and PRS analyses. Through my interviews with investigators who had experience using race, ethnicity, and/or (genetic and non-genetic) ancestry in GWAS and PRS, I learned that, although there was unanimous agreement that more diverse data was needed in order to generate more equitable PRS, there remained many questions about how to use that diversity responsibly in genomic analyses. I learned that it was common practice to use race, ethnicity, and genetic ancestry to identify population groups for data analysis - a practice I myself employed in my own GWAS and PRS analyses, though with hesitation and unease. I also learned from these interviews that I was not alone in that sense of uneasiness. Many participants in this study felt that their views were at odds with what standard practices in the field expected or entailed. I concluded that a major impediment to developing equitable PRS are the constraints on research practice imposed by limitations in the research infrastructure.

Moving forward

Much of this dissertation focused on the challenges of using REA in PRS within the research domain. However, there is a growing body of literature that discusses specific implications and challenges experienced with the use of ancestry-informed PRS in a clinical setting (Aragam & Natarajan, 2020; James et al., 2021; A. C. F. Lewis, Perez, et al., 2022; C. M. Lewis & Vassos, 2020). One area to explore further is better understanding the social and relational burdens that healthcare providers (HCP) and patients experience as a result of using ancestry-informed PRS. For example, some healthcare providers (HCP) have expressed confusion regarding interpretation of PRS for individuals who do not fit into any one category as well as communication of PRS results to patients (A. C. F. Lewis, Perez, et al., 2022). This exposes a disconnect between genomic research and clinical medicine where information relevant to genetic ancestry is difficult to translate into a clinical context, where the relevant and accessible information in a clinical setting is typically race, ethnicity or other social identifiers, and there is no direct link between the two types of information.

As alluded to in earlier chapters, ancestry-informed PRS also place an undue burden on individuals whose ancestry backgrounds are not represented by categories included in the PRS. This is especially true for individuals of mixed backgrounds and for whom PRS development remains a highly active area of research. One can imagine a scenario where a patient A has heard of the benefits of polygenic risk scores in predicting risk of developing a disease, but later learns that they do not yet accurately predict risk for someone of their background. The option to proceed with the PRS is another decision to weigh: should patient A use the PRS and then assess whether the results are relevant to them? Or should they forego the test altogether? Compare this

experience to patient B whose population background is represented in available PRS models and can proceed with receiving a genetic risk estimate with little concern for inaccuracy. The experiences of these two patients is imbalanced; not only would patient B receive additional beneficial care in the wake of a high-risk test result, but they would also be protected from the extra burden of decision-making and personal risk assessment that patient A must confront.

Implementing a version of PRS that has the potential to be beneficial to everyone but is accessible only to a small subset of the population - as is occurring in clinical trials (James et al., 2021) or as DTC (direct to consumer) products (e.g., MyHeritage) - with the hope that it will, at some point in the future, be equitable still exacerbates health inequities *now* and does not guarantee they will eventually be mitigated. As illustrated by the imaginary scenario above, some forms of inequity exist even within the interactions between people (e.g., patient and provider) or between people and products (e.g., patients and PRS), and may not be immediately obvious or measurable, yet nevertheless have real consequences on health outcomes (Fourie et al., 2015). As the genomics community strives to improve representation in genetics research, we need to continually reexamine our approaches to research so that we can find ways to better foster inclusiveness and cultivate a sense of belonging in scientific and medical practice - particularly for individuals of historically marginalized or overlooked communities - and lay the groundwork for building more equitable tools in genomic precision medicine.

References

ATLAS.ti Scientific Software Development GmbH. (2023). ATLAS.ti Mac (version 23.2.1) [Qualitative data analysis software]. https://atlasti.com

2015 National Content Test: Race and Ethnicity Analysis Report. (n.d.).

Aragam, K. G., & Natarajan, P. (2020). Polygenic Scores to Assess Atherosclerotic Cardiovascular Disease Risk: Clinical Perspectives and Basic Implications. *Circulation Research*, *126*(9), 1159–1177. https://doi.org/10.1161/CIRCRESAHA.120.315928

Arshad, S, Arshad, J, Khan, Muhammad M, Parkinson, S. (2021) Analysis of security and privacy challenges for DNA-genomics applications and databases. *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2021.103815.

ASHG Denounces Attempts to Link Genetics and Racial Supremacy. (2018). *The American Journal of Human Genetics*, *103*(5), 636. https://doi.org/10.1016/j.ajhg.2018.10.011

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M. A., Lambourne, J. J., Sivapalaratnam, S., Downes, K., Kundu, K., Bomba, L., Berentsen, K., Bradley, J. R., Daugherty, L. C., Delaneau, O., … Soranzo, N. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, *167*(5), 1415-1429.e19. https://doi.org/10.1016/j.cell.2016.10.042

Barnhardt, R. (n.d.). INDIGENOUS KNOWLEDGE SYSTEMS AND CROSS-CULTURAL RESEARCH.

Barrera-Reyes, P. K., & Tejero, M. E. (2019). Genetic variation influencing hemoglobin levels and risk for anemia across populations. Annals of the New York Academy of Sciences, 1450(1), 32–46. https://doi.org/10.1111/nyas.14200

Belbin, G. M., Cullina, S., Wenric, S., Soper, E. R., Glicksberg, B. S., Torre, D., Moscati, A., Wojcik, G. L., Shemirani, R., Beckmann, N. D., Cohain, A., Sorokin, E. P., Park, D. S., Ambite, J.-L., Ellis, S., Auton, A., Bottinger, E. P., Cho, J. H., Loos, R. J. F., … Kenny, E. E. (2021). Toward a fine-scale population health monitoring system. *Cell*, *184*(8), 2068-2083.e11. https://doi.org/10.1016/j.cell.2021.03.034

Bentz, M., Saperstein, A., Fullerton, S. M., Shim, J. K., & Lee, S. S.-J. (2024). Conflating race and ancestry: Tracing decision points about population descriptors over the precision medicine research life course. *Human Genetics and Genomics Advances*, *5*(1), 100243. https://doi.org/10.1016/j.xhgg.2023.100243

Bird KA and Carlson J (2024) Typological thinking in human genomics research contributes to the production and prominence of scientific racism. Front. Genet. 15:1345631. doi: 10.3389/fgene.2024.1345631

BLISS, C. (2009). *GENOME SAMPLING AND THE BIOPOLITICS OF RACE*.

Bliss, C. (2020). Conceptualizing Race in the Genomic Age. *Hastings Center Report*, *50*, S15–S22. https://doi.org/10.1002/hast.1151

Bliss. Genome Sampling and the Biopolitics of Race.pdf. (n.d.).

Bonham, V. L., Green, E. D., & Pérez-Stable, E. J. (2018). Examining How Race, Ethnicity, and Ancestry Data Are Used in Biomedical Research. *JAMA*, *320*(15), 1533. https://doi.org/10.1001/jama.2018.13609

Braun, L. (2015). Race, ethnicity and lung function: A brief history. *Canadian Journal of Respiratory Therapy: CJRT = Revue Canadienne de La Therapie Respiratoire: RCTR*, *51*(4), 99–101.

Braun, L., & Hammonds, E. (2008). Race, populations, and genomics: Africa as laboratory. *Social Science & Medicine*, *67*(10), 1580–1588. https://doi.org/10.1016/j.socscimed.2008.07.018

Braveman, P., & Gottlieb, L. (2014). The social determinants of health: it's time to consider the causes of the causes. Public health reports (Washington, D.C. : 1974), 129 Suppl 2(Suppl 2), 19–31. https://doi.org/10.1177/00333549141291S206

Braveman, P. A., Kumanyika, S., Fielding, J., LaVeist, T., Borrell, L. N., Manderscheid, R., & Troutman, A. (2011). Health Disparities and Health Equity: The Issue Is Justice. *American Journal of Public Health*, *101*(S1), S149–S155. https://doi.org/10.2105/AJPH.2010.300062

Brubaker, R. (2016). *Trans: Gender and race in an age of unsettled identities*. Princeton University Press.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Byeon, Y. J. J., Islamaj, R., Yeganova, L., Wilbur, W. J., Lu, Z., Brody, L. C., & Bonham, V. L. (2021). Evolving use of ancestry, ethnicity, and race in genetics research-A survey spanning seven decades. *American Journal of Human Genetics*, *108*(12), 2215–2223. https://doi.org/10.1016/j.ajhg.2021.10.008

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Paul Flicek, Germer, S., Brand, H., Hall, I. M., Talkowski, M. E., … Xiao, C. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, *185*(18), 3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004

Caliebe, A., Tekola‐Ayele, F., Darst, B. F., Wang, X., Song, Y. E., Gui, J., Sebro, R. A., Balding, D. J., Saad, M., Dubé, M., & IGES ELSI Committee. (2022). Including diverse and admixed populations in genetic epidemiology research. *Genetic Epidemiology*, *46*(7), 347–371. https://doi.org/10.1002/gepi.22492

Census.gov. (n.d.).

https://www.census.gov/quickfacts/fact/note/US/RHI625222#:~:text=OMB%20requires%20five%20minimum%20categories,report%20more%20than%20one%20race.

Chen, M.-H., Raffield, L. M., Mousas, A., Sakaue, S., Huffman, J. E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., Bao, E. L., Zhong, X., Manansala, R., Laplante, V., Chen, M., Lo, K. S., Qian, H., Lareau, C. A., Beaudoin, M., … Lettre, G. (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell*, *182*(5), 1198-1213.e14. https://doi.org/10.1016/j.cell.2020.06.045

Constantinescu, AE., Mitchell, R.E., Zheng, J. et al. A framework for research into continental ancestry groups of the UK Biobank. Hum Genomics 16, 3 (2022). https://doi.org/10.1186/s40246-022-00380-5

Chen, Z., Tang, H., Qayyum, R., Schick, U. M., Nalls, M. A., Handsaker, R., Li, J., Lu, Y., Yanek, L. R., Keating, B., Meng, Y., Van Rooij, F. J. A., Okada, Y., Kubo, M., Rasmussen-Torvik, L., Keller, M. F., Lange, L., Evans, M., Bottinger, E. P., … Reiner, A. P. (2013). Genome-wide association analysis of red blood cell traits in African Americans: The COGENT Network. Human Molecular Genetics, 22(12), 2529–2538. https://doi.org/10.1093/hmg/ddt087

Choi, S. W., Mak, T. S.-H., & O'Reilly, P. F. (2020). Tutorial: A guide to performing polygenic risk score analyses. *Nature Protocols*, *15*(9), 2759–2772. https://doi.org/10.1038/s41596-020-0353-1

Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research, Board on Health Sciences Policy, Committee on Population, Health and Medicine Division, Division of Behavioral and Social Sciences and Education, & National Academies of Sciences, Engineering, and Medicine. (2023b). *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field* (p. 26902). National Academies Press. https://doi.org/10.17226/26902

Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research, Board on Health Sciences Policy, Committee on Population, Health and Medicine Division, Division of Behavioral and Social Sciences and Education, & National Academies of Sciences, Engineering, and Medicine. (2023a). *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field* (p. 26902). National Academies Press. https://doi.org/10.17226/26902

Considerations for clinical implementation of polygenic risk scores in diverse US populations. (2024). *Nature Medicine*, *30*(2), 354–355. https://doi.org/10.1038/s41591-024-02801-5

Creary, M. S. (2018). Biocultural citizenship and embodying exceptionalism: Biopolitics for sickle cell disease in Brazil. *Social Science & Medicine*, *199*, 123–131. https://doi.org/10.1016/j.socscimed.2017.04.035

Creary, M. S. (2021). Bounded Justice and the Limits of Health Equity. *Journal of Law, Medicine & Ethics*, *49*(2), 241–256. https://doi.org/10.1017/jme.2021.34

Dai, C. L., Vazifeh, M. M., Yeang, C.-H., Tachet, R., Wells, R. S., Vilar, M. G., Daly, M. J., Ratti,

C., & Martin, A. R. (2020). Population Histories of the United States Revealed through Fine-Scale Migration and Haplotype Analysis. *The American Journal of Human Genetics*, *106*(3), 371–388. https://doi.org/10.1016/j.ajhg.2020.02.002

Danese, M. D., Halperin, M., Duryea, J., & Duryea, R. (2019a). The Generalized Data Model for clinical research. *BMC Medical Informatics and Decision Making*, *19*(1), 117. https://doi.org/10.1186/s12911-019-0837-5

Danese, M. D., Halperin, M., Duryea, J., & Duryea, R. (2019b). The Generalized Data Model for clinical research. *BMC Medical Informatics and Decision Making*, *19*(1), 117. https://doi.org/10.1186/s12911-019-0837-5

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

Dauda, B., Molina, S. J., Allen, D. S., Fuentes, A., Ghosh, N., Mauro, M., Neale, B. M., Panofsky, A., Sohail, M., Zhang, S. R., & Lewis, A. C. F. (2023a). Ancestry: How researchers use it and what they mean by it. *Frontiers in Genetics*, *14*, 1044555. https://doi.org/10.3389/fgene.2023.1044555

Dauda, B., Molina, S. J., Allen, D. S., Fuentes, A., Ghosh, N., Mauro, M., Neale, B. M., Panofsky, A., Sohail, M., Zhang, S. R., & Lewis, A. C. F. (2023b). Ancestry: How researchers use it and what they mean by it. *Frontiers in Genetics*, *14*, 1044555. https://doi.org/10.3389/fgene.2023.1044555

Dikilitas, O., Schaid, D. J., Kosel, M. L., Carroll, R. J., Chute, C. G., Denny, J. C., Fedotov, A., Feng, Q., Hakonarson, H., Jarvik, G. P., Lee, M. T. M., Pacheco, J. A., Rowley, R., Sleiman, P. M., Stein, C. M., Sturm, A. C., Wei, W.-Q., Wiesner, G. L., Williams, M. S., … Kullo, I. J. (2020). Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups. *The American Journal of Human Genetics*, *106*(5), 707–716. https://doi.org/10.1016/j.ajhg.2020.04.002

Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., Privé, F., Vilhjálmsson, B. J., Olde Loohuis, L. M., & Pasaniuc, B. (2023b). Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, *618*(7966), 774–781. https://doi.org/10.1038/s41586-023-06079-4

Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., Privé, F., Vilhjálmsson, B. J., Olde Loohuis, L. M., & Pasaniuc, B. (2023a). Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, *618*(7966), 774–781. https://doi.org/10.1038/s41586-023-06079-4

Fang, H., Hui, Q., Lynch, J., Honerlaw, J., Assimes, T. L., Huang, J., Vujkovic, M., Damrauer, S. M., Pyarajan, S., Gaziano, J. M., DuVall, S. L., O'Donnell, C. J., Cho, K., Chang, K.-M., Wilson, P. W. F., Tsao, P. S., Sun, Y. V., Tang, H., Gaziano, J. M., … Striker, R. (2019). Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *The*

American Journal of Human Genetics, *105*(4), 763–772. https://doi.org/10.1016/j.ajhg.2019.08.012

Fatumo, S., Chikowore, T., Choudhury, A., Ayub, M., Martin, A. R., & Kuchenbaecker, K. (2022). A roadmap to increase diversity in genomic studies. *Nature Medicine*, *28*(2), 243–250. https://doi.org/10.1038/s41591-021-01672-4

Ferryman, K. (2023). Bounded Justice, Inclusion, and the Hyper/Invisibility of Race in Precision Medicine. *The American Journal of Bioethics*, *23*(7), 27–33. https://doi.org/10.1080/15265161.2023.2207515

Fields, K. E., & Fields, B. J. (2022). *Racecraft: The soul of inequality in American life* (Paperback edition). Verso.

Fourie, C., Schuppert, F., & Wallimann-Helmer, I. (Eds.). (2015). *Social Equality: On What It Means to be Equals*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199331109.001.0001

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., & Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, *10*(1), 1776. https://doi.org/10.1038/s41467-019-09718-5

Ge, T., Irvin, M. R., Patki, A., Srinivasasainagendra, V., Lin, Y.-F., Tiwari, H. K., Armstrong, N. D., Benoit, B., Chen, C.-Y., Choi, K. W., Cimino, J. J., Davis, B. H., Dikilitas, O., Etheridge, B., Feng, Y.-C. A., Gainer, V., Huang, H., Jarvik, G. P., Kachulis, C., … Karlson, E. W. (2022a). Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. *Genome Medicine*, *14*(1), 70. https://doi.org/10.1186/s13073-022-01074-2

Ge, T., Irvin, M. R., Patki, A., Srinivasasainagendra, V., Lin, Y.-F., Tiwari, H. K., Armstrong, N. D., Benoit, B., Chen, C.-Y., Choi, K. W., Cimino, J. J., Davis, B. H., Dikilitas, O., Etheridge, B., Feng, Y.-C. A., Gainer, V., Huang, H., Jarvik, G. P., Kachulis, C., … Karlson, E. W. (2022b). Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. *Genome Medicine*, *14*(1), 70. https://doi.org/10.1186/s13073-022-01074-2

Greely, H. T. (2001). Human genome diversity: What about the other human genome project? *Nature Reviews Genetics*, *2*(3), 222–227. https://doi.org/10.1038/35056071

Hennink, M. M., Kaiser, B. N., & Marconi, V. C. (2017). Code Saturation Versus Meaning Saturation: How Many Interviews Are Enough? *Qualitative Health Research*, *27*(4), 591–608. https://doi.org/10.1177/1049732316665344

Herd, P., Mills, M. C., & Dowd, J. B. (2021). Reconstructing Sociogenomics Research: Dismantling Biological Race and Genetic Essentialism Narratives. *Journal of Health and Social Behavior*, *62*(3), 419–435. https://doi.org/10.1177/00221465211018682

Hindorff, L. A., Bonham, V. L., Brody, L. C., Ginoza, M. E. C., Hutter, C. M., Manolio, T. A., & Green, E. D. (2018a). Prioritizing diversity in human genomics research. *Nature Reviews. Genetics*, *19*(3), 175–185. https://doi.org/10.1038/nrg.2017.89

Hindorff, L. A., Bonham, V. L., Brody, L. C., Ginoza, M. E. C., Hutter, C. M., Manolio, T. A., &

Green, E. D. (2018b). Prioritizing diversity in human genomics research. *Nature Reviews Genetics*, *19*(3), 175–185. https://doi.org/10.1038/nrg.2017.89

Hindy, G., Aragam, K. G., Ng, K., Chaffin, M., Lotta, L. A., Baras, A., Regeneron Genetics Center, Drake, I., Orho-Melander, M., Melander, O., Kathiresan, S., & Khera, A. V. (2020). Genome-Wide Polygenic Score, Clinical Risk Factors, and Long-Term Trajectories of Coronary Artery Disease. Arteriosclerosis, thrombosis, and vascular biology, 40(11), 2738–2746. https://doi.org/10.1161/ATVBAHA.120.314856

Hsieh, H.-F., & Shannon, S. E. (2005a). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Hsieh, H.-F., & Shannon, S. E. (2005b). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Hu, Y., Stilp, A. M., McHugh, C. P., Rao, S., Jain, D., Zheng, X., Lane, J., Méric de Bellefon, S., Raffield, L. M., Chen, M.-H., Yanek, L. R., Wheeler, M., Yao, Y., Ren, C., Broome, J., Moon, J.-Y., de Vries, P. S., Hobbs, B. D., Sun, Q., … NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. (2021). Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: The NHLBI TOPMed program. *American Journal of Human Genetics*, *108*(5), 874–893. https://doi.org/10.1016/j.ajhg.2021.04.003

Hu, Y., Stilp, A. M., McHugh, C. P., Rao, S., Jain, D., Zheng, X., Lane, J., Méric De Bellefon, S., Raffield, L. M., Chen, M.-H., Yanek, L. R., Wheeler, M., Yao, Y., Ren, C., Broome, J., Moon, J.-Y., De Vries, P. S., Hobbs, B. D., Sun, Q., … Reiner, A. P. (2021). Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: The NHLBI TOPMed program. *The American Journal of Human Genetics*, *108*(5), 874–893. https://doi.org/10.1016/j.ajhg.2021.04.003

Initial Proposals For Updating OMB's Race and Ethnicity Statistical Standards.pdf. (n.d.).

James, J. E., Riddle, L., Koenig, B. A., & Joseph, G. (2021). The limits of personalization in precision medicine: Polygenic risk scores and racial categorization in a precision breast cancer screening trial. *PLOS ONE*, *16*(10), e0258571. https://doi.org/10.1371/journal.pone.0258571

Kachuri, L., Chatterjee, N., Hirbo, J., Schaid, D. J., Martin, I., Kullo, I. J., Kenny, E. E., Pasaniuc, B., Polygenic Risk Methods in Diverse Populations (PRIMED) Consortium Methods Working Group, Auer, P. L., Conomos, M. P., Conti, D. V., Ding, Y., Wang, Y., Zhang, H., Zhang, Y., Witte, J. S., & Ge, T. (2024). Principles and methods for transferring polygenic risk scores across global populations. *Nature Reviews Genetics*, *25*(1), 8–25. https://doi.org/10.1038/s41576-023-00637-2

Kamiza, A. B., Toure, S. M., Vujkovic, M., Machipisa, T., Soremekun, O. S., Kintu, C., Corpas, M., Pirie, F., Young, E., Gill, D., Sandhu, M. S., Kaleebu, P., Nyirenda, M., Motala, A. A., Chikowore, T., & Fatumo, S. (2022). Transferability of genetic risk scores in African populations. *Nature Medicine*, *28*(6), 1163–1166. https://doi.org/10.1038/s41591-022-01835-x

Kaplan, J. M., & Fullerton, S. M. (2022). Polygenic risk, population structure and ongoing difficulties with race in human genetics. *Philosophical Transactions of the Royal Society B:*

Biological Sciences, *377*(1852), 20200427. https://doi.org/10.1098/rstb.2020.0427

Khan, A. T., Gogarten, S. M., McHugh, C. P., Stilp, A. M., Sofer, T., Bowers, M. L., Wong, Q., Cupples, L. A., Hidalgo, B., Johnson, A. D., McDonald, M.-L. N., McGarvey, S. T., Taylor, M. R. G., Fullerton, S. M., Conomos, M. P., & Nelson, S. C. (2022). Recommendations on the use and reporting of race, ethnicity, and ancestry in genetic research: Experiences from the NHLBI TOPMed program. *Cell Genomics*, *2*(8), 100155. https://doi.org/10.1016/j.xgen.2022.100155

Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. 50, 1219–1224 (2018).

Klarin, D., & Natarajan, P. (2022). Clinical utility of polygenic risk scores for coronary artery disease. Nature reviews. Cardiology, 19(5), 291–301. https://doi.org/10.1038/s41569-021-00638-w

Koenig, Z., Yohannes, M. T., Nkambule, L. L., Goodrich, J. K., Kim, H. A., Zhao, X., Wilson, M. W., Tiao, G., Hao, S. P., Sahakian, N., Chao, K. R., gnomAD Project Consortium, Rehm, H. L., Neale, B. M., Talkowski, M. E., Daly, M. J., Brand, H., Karczewski, K. J., Atkinson, E. G., & Martin, A. R. (2023). *A harmonized public resource of deeply sequenced diverse human genomes* [Preprint]. Genomics. https://doi.org/10.1101/2023.01.23.525248

Kullo, I. J., Ding, K., Jouni, H., Smith, C. Y., & Chute, C. G. (2010). A Genome-Wide Association Study of Red Blood Cell Traits Using the Electronic Medical Record. *PLoS ONE*, *5*(9), e13011. https://doi.org/10.1371/journal.pone.0013011

Kullo, I. J., Lewis, C. M., Inouye, M., Martin, A. R., Ripatti, S., & Chatterjee, N. (2022). Polygenic scores in biomedical research. *Nature Reviews Genetics*, *23*(9), 524–532. https://doi.org/10.1038/s41576-022-00470-z

Kullo IJ, Ding K, Jouni H, Smith CY, Chute CG (2010) A Genome-Wide Association Study of Red Blood Cell Traits Using the Electronic Medical Record. PLoS ONE 5(9): e13011. https://doi.org/10.1371/journal.pone.0013011

Kurniansyah, N., Goodman, M. O., Khan, A. T., Wang, J., Feofanova, E., Bis, J. C., Wiggins, K. L., Huffman, J. E., Kelly, T., Elfassy, T., Guo, X., Palmas, W., Lin, H. J., Hwang, S.-J., Gao, Y., Young, K., Kinney, G. L., Smith, J. A., Yu, B., … Sofer, T. (2023). Evaluating the use of blood pressure polygenic risk scores across race/ethnic background groups. *Nature Communications*, *14*(1), 3202. https://doi.org/10.1038/s41467-023-38990-9

Liao, S., & Carbonell, V. (2023). Materialized Oppression in Medical Tools and Technologies. The American Journal of Bioethics, 23(4), 9–23. https://doi.org/10.1080/15265161.2022.2044543

Leng, C. H. (n.d.). *Bioethics and human population genetics research; 1995*.

Lewis, A. C. F., & Green, R. C. (2021a). Polygenic risk scores in the clinic: New perspectives needed on familiar ethical issues. *Genome Medicine*, *13*(1), 14. https://doi.org/10.1186/s13073-021-00829-7

Lewis, A. C. F., & Green, R. C. (2021b). Polygenic risk scores in the clinic: New perspectives needed on familiar ethical issues. *Genome Medicine*, *13*(1), 14. https://doi.org/10.1186/s13073-021-00829-7

Lewis, A. C. F., Green, R. C., & Vassy, J. L. (2021). Polygenic risk scores in the clinic: Translating risk into action. *Human Genetics and Genomics Advances*, *2*(4), 100047. https://doi.org/10.1016/j.xhgg.2021.100047

Lewis, A. C. F., Molina, S. J., Appelbaum, P. S., Dauda, B., Di Rienzo, A., Fuentes, A., Fullerton, S. M., Garrison, N. A., Ghosh, N., Hammonds, E. M., Jones, D. S., Kenny, E. E., Kraft, P., Lee, S. S.-J., Mauro, M., Novembre, J., Panofsky, A., Sohail, M., Neale, B. M., & Allen, D. S. (2022). Getting genetic ancestry right for science and society. *Science*, *376*(6590), 250–252. https://doi.org/10.1126/science.abm7530

Lewis, A. C. F., Perez, E. F., Prince, A. E. R., Flaxman, H. R., Gomez, L., Brockman, D. G., Chandler, P. D., Kerman, B. J., Lebo, M. S., Smoller, J. W., Weiss, S. T., Blout Zawatksy, C. L., Meigs, J. B., Green, R. C., Vassy, J. L., & Karlson, E. W. (2022). Patient and provider perspectives on polygenic risk scores: Implications for clinical reporting and utilization. *Genome Medicine*, *14*(1), 114. https://doi.org/10.1186/s13073-022-01117-8

Lewis, C. M., & Vassos, E. (2020). Polygenic risk scores: From research tools to clinical instruments. *Genome Medicine*, *12*(1), 44. https://doi.org/10.1186/s13073-020-00742-5

Liao, S., & Carbonell, V. (2023). Materialized Oppression in Medical Tools and Technologies. *The American Journal of Bioethics*, *23*(4), 9–23. https://doi.org/10.1080/15265161.2022.2044543

Little, A., Hu, Y., Sun, Q., Jain, D., Broome, J., Chen, M.-H., Thibord, F., McHugh, C., Surendran, P., Blackwell, T. W., Brody, J. A., Bhan, A., Chami, N., De Vries, P. S., Ekunwe, L., Heard-Costa, N., Hobbs, B. D., Manichaikul, A., Moon, J.-Y., … Raffield, L. M. (2022). Whole genome sequence analysis of platelet traits in the NHLBI Trans-Omics for Precision Medicine (TOPMed) initiative. *Human Molecular Genetics*, *31*(3), 347–361. https://doi.org/10.1093/hmg/ddab252

Loomba, A. (2015a). *Colonialism/postcolonialism* (Third edition). Routledge, Taylor & Francis Group.

Loomba, A. (2015b). *Colonialism/postcolonialism* (Third edition). Routledge, Taylor & Francis Group.

Maghbouleh, N., Schachter, A., & Flores, R. D. (2022). Middle Eastern and North African Americans may not be perceived, nor perceive themselves, to be White. *Proceedings of the National Academy of Sciences*, *119*(7), e2117940119. https://doi.org/10.1073/pnas.2117940119

Mathias, R., Taub, M., Gignoux, C. et al. A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. Nat Commun 7, 12522 (2016). https://doi.org/10.1038/ncomms12522

Maguire, M., & Delahunt, B. (2017). *Doing a Thematic Analysis: A Practical, Step-by-Step Guide for Learning and Teaching Scholars. 8*(3).

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019a). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, *51*(4), 584–591. https://doi.org/10.1038/s41588-019-0379-x

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019b). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, *51*(4), 584–591. https://doi.org/10.1038/s41588-019-0379-x

Martschenko, D. O., Wand, H., Young, J. L., & Wojcik, G. L. (2023). Including multiracial individuals is crucial for race, ethnicity and ancestry frameworks in genetics and genomics. *Nature Genetics*, *55*(6), 895–900. https://doi.org/10.1038/s41588-023-01394-y

Mathias, R. A., Taub, M. A., Gignoux, C. R., Fu, W., Musharoff, S., O'Connor, T. D., Vergara, C., Torgerson, D. G., Pino-Yanes, M., Shringarpure, S. S., Huang, L., Rafaels, N., Boorgula, M. P., Johnston, H. R., Ortega, V. E., Levin, A. M., Song, W., Torres, R., Padhukasahasram, B., … Barnes, K. C. (2016). A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. *Nature Communications*, *7*(1), 12522. https://doi.org/10.1038/ncomms12522

Mathieson, I., & Scally, A. (2020). What is ancestry? *PLoS Genetics*, *16*(3), e1008624. https://doi.org/10.1371/journal.pgen.1008624

Matthew P. Conomos, T. T. (2017). *GENESIS* [Computer software]. Bioconductor. https://doi.org/10.18129/B9.BIOC.GENESIS

McArdle, C. E., Bokhari, H., Rodell, C. C., Buchanan, V., Preudhomme, L. K., Isasi, C. R., Graff, M., North, K., Gallo, L. C., Pirzada, A., Daviglus, M. L., Wojcik, G., Cai, J., Perreira, K., & Fernandez-Rhodes, L. (2021). Findings from the Hispanic Community Health Study/Study of Latinos on the Importance of Sociocultural Environmental Interactors: Polygenic Risk Score-by-Immigration and Dietary Interactions. *Frontiers in Genetics*, *12*, 720750. https://doi.org/10.3389/fgene.2021.720750

Menon, M. (2022). Indigenous knowledges and colonial sciences in South Asia. South Asian History and Culture, 13(1), 1–18. https://doi.org/10.1080/19472498.2021.2001198

Mikhaylova, A. V., McHugh, C. P., Polfus, L. M., Raffield, L. M., Boorgula, M. P., Blackwell, T. W., Brody, J. A., Broome, J., Chami, N., Chen, M.-H., Conomos, M. P., Cox, C., Curran, J. E., Daya, M., Ekunwe, L., Glahn, D. C., Heard-Costa, N., Highland, H. M., Hobbs, B. D., … Auer, P. L. (2021). Whole-genome sequencing in diverse subjects identifies genetic correlates of leukocyte traits: The NHLBI TOPMed program. *The American Journal of Human Genetics*, *108*(10), 1836–1851. https://doi.org/10.1016/j.ajhg.2021.08.007

Ortiz, V., & Telles, E. (2012). Racial Identity and Racial Treatment of Mexican Americans. *Race and Social Problems*, *4*(1), 41–56. https://doi.org/10.1007/s12552-012-9064-8

Office of Management and Budget, "Initial Proposals for Updating OMB's Race and Ethnicity

Statistical Standards," 88 Fed. Reg. 5375 (January 27, 2023).

Patel, KV. Variability and heritability of hemoglobin concentration: an opportunity to improve understanding of anemia in older adults. Haematologica 2008;93(9):1281-1283; https://doi.org/10.3324/haematol.13692.

Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C. Y., Popejoy, A. B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R. J., Brick, L., Carey, C. E., Martin, A. R., Meyers, J. L., Su, J., Chen, J., Edwards, A. C., Kalungi, A., Koen, N., Majara, L., Schwarz, E., … Duncan, L. E. (2019). Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. Cell, 179(3), 589–603. https://doi.org/10.1016/j.cell.2019.08.051

Polygenic Risk Score Task Force of the International Common Disease Alliance, Adeyemo, A., Balaconis, M. K., Darnes, D. R., Fatumo, S., Granados Moreno, P., Hodonsky, C. J., Inouye, M., Kanai, M., Kato, K., Knoppers, B. M., Lewis, A. C. F., Martin, A. R., McCarthy, M. I., Meyer, M. N., Okada, Y., Richards, J. B., Richter, L., Ripatti, S., … Zhou, A. (2021). Responsible use of polygenic risk scores in the clinic: Potential benefits, risks and gaps. *Nature Medicine*, *27*(11), 1876–1884. https://doi.org/10.1038/s41591-021-01549-6

Privé F., Arbel J., Vilhjálmsson, BJ., LDpred2: better, faster, stronger, Bioinformatics, Volume 36, Issue 22-23, December 2020, Pages 5424–5431, https://doi.org/10.1093/bioinformatics/btaa1029

Reardon, J. (2005). *Race to the finish: Identity and governance in an age of genomics*. Princeton University Press.

Rebbeck, T. R., Mahal, B., Maxwell, K. N., Garraway, I. P., & Yamoah, K. (2022). The distinct impacts of race and genetic ancestry on health. *Nature Medicine*, *28*(5), 890–893. https://doi.org/10.1038/s41591-022-01796-1

Roberts, D. E. (2011a). Chapter 3: Redefining race in genetic terms. In *Fatal invention: How science, politics, and big business re-create race in the twenty-first century*. New Press.

Roberts, D. E. (2011b). *Fatal invention: How science, politics, and big business re-create race in the twenty-first century*. New Press.

Rosenthal, E. A., Hsu, L., Thomas, M., Peters, U., Kachulis, C., Patterson, K., & Jarvik, G. P. (2023). *Comparing ancestry calibration approaches for a trans-ancestry colorectal cancer polygenic risk score* [Preprint]. Genetic and Genomic Medicine. https://doi.org/10.1101/2023.10.23.23296753

Ruan, Y., Feng, Y.-C. A., Chen, C.-Y., Lam, M., Sawa, A., Martin, A. R., Qin, S., Huang, H., & Ge, T. (n.d.). *Improving Polygenic Prediction in Ancestrally Diverse Populations*. 17.

Ruan, Y., Lin, Y.-F., Feng, Y.-C. A., Chen, C.-Y., Lam, M., Guo, Z., Stanley Global Asia Initiatives, Ahn, Y. M., Akiyama, K., Arai, M., Baek, J. H., Chen, W. J., Chung, Y.-C., Feng, G., Fujii, K., Glatt, S. J., Ha, K., Hattori, K., Higuchi, T., … Ge, T. (2022). Improving polygenic

prediction in ancestrally diverse populations. *Nature Genetics*, *54*(5), 573–580. https://doi.org/10.1038/s41588-022-01054-7

Saylor, K. M. (2023). Biometric Technologies and Global Security (IF11783 Version 4; Congressional Research Service). https://crsreports.congress.gov/

Schultz, L. M., Merikangas, A. K., Ruparel, K., Jacquemont, S., Glahn, D. C., Gur, R. E., Barzilay, R., & Almasy, L. (2022). Stability of polygenic scores across discovery genome-wide association studies. *Human Genetics and Genomics Advances*, *3*(2), 100091. https://doi.org/10.1016/j.xhgg.2022.100091

Sengupta, D., Botha, G., Meintjes, A., Mbiyavanga, M., Hazelhurst, S., Mulder, N., Ramsay, M., & Choudhury, A. (2023a). Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations. *Cell Genomics*, *3*(6), 100332. https://doi.org/10.1016/j.xgen.2023.100332

Sengupta, D., Botha, G., Meintjes, A., Mbiyavanga, M., Hazelhurst, S., Mulder, N., Ramsay, M., & Choudhury, A. (2023b). Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations. *Cell Genomics*, *3*(6), 100332. https://doi.org/10.1016/j.xgen.2023.100332

Stilp, A. M., Emery, L. S., Broome, J. G., Buth, E. J., Khan, A. T., Laurie, C. A., Wang, F. F., Wong, Q., Chen, D., D'Augustine, C. M., Heard-Costa, N. L., Hohensee, C. R., Johnson, W. C., Juarez, L. D., Liu, J., Mutalik, K. M., Raffield, L. M., Wiggins, K. L., De Vries, P. S., … Laurie, C. C. (2021b). A System for Phenotype Harmonization in the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Program. *American Journal of Epidemiology*, *190*(10), 1977–1992. https://doi.org/10.1093/aje/kwab115

Stilp, A. M., Emery, L. S., Broome, J. G., Buth, E. J., Khan, A. T., Laurie, C. A., Wang, F. F., Wong, Q., Chen, D., D'Augustine, C. M., Heard-Costa, N. L., Hohensee, C. R., Johnson, W. C., Juarez, L. D., Liu, J., Mutalik, K. M., Raffield, L. M., Wiggins, K. L., De Vries, P. S., … Laurie, C. C. (2021a). A System for Phenotype Harmonization in the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Program. *American Journal of Epidemiology*, *190*(10), 1977–1992. https://doi.org/10.1093/aje/kwab115

Sun, J., Wang, Y., Folkersen, L., Borné, Y., Amlien, I., Buil, A., Orho-Melander, M., Børglum, A. D., Hougaard, D. M., Regeneron Genetics Center, Lotta, L. A., Jones, M., Baras, A., Melander, O., Engström, G., Werge, T., & Lage, K. (2021). Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction. *Nature Communications*, *12*(1), 5276. https://doi.org/10.1038/s41467-021-25014-7

Sun, Q., Graff, M., Rowland, B., Wen, J., Huang, L., Miller-Fleming, T. W., Haessler, J., Preuss, M. H., Chai, J.-F., Lee, M. P., Avery, C. L., Cheng, C.-Y., Franceschini, N., Sim, X., Cox, N. J., Kooperberg, C., North, K. E., Li, Y., & Raffield, L. M. (2021). Analyses of biomarker traits in diverse UK biobank participants identify associations missed by European-centric analysis strategies. *Journal of Human Genetics*. https://doi.org/10.1038/s10038-021-00968-0

Sun, Q., Graff, M., Rowland, B., Wen, J., Huang, L., Miller-Fleming, T. W., Haessler, J., Preuss, M. H., Chai, J.-F., Lee, M. P., Avery, C. L., Cheng, C.-Y., Franceschini, N., Sim, X., Cox, N. J.,

Kooperberg, C., North, K. E., Li, Y., & Raffield, L. M. (2022). Analyses of biomarker traits in diverse UK biobank participants identify associations missed by European-centric analysis strategies. *Journal of Human Genetics*, *67*(2), 87–93. https://doi.org/10.1038/s10038-021-00968-0

Sun, Q., Rowland, B. T., Chen, J., Mikhaylova, A. V., Avery, C., Peters, U., Lundin, J., Matise, T., Buyske, S., Tao, R., Mathias, R. A., Reiner, A. P., Auer, P. L., Cox, N. J., Kooperberg, C., Thornton, T. A., Raffield, L. M., & Li, Y. (2024). Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-differential effects via GAUDI. *Nature Communications*, *15*(1), 1016. https://doi.org/10.1038/s41467-024-45135-z

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., … Abecasis, G. R. (2021a). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290–299. https://doi.org/10.1038/s41586-021-03205-y

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S.-B., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., … Abecasis, G. R. (2021b). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290–299. https://doi.org/10.1038/s41586-021-03205-y

The 1000 Genomes Project Consortium, Corresponding authors, Auton, A., Abecasis, G. R., Steering committee, Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

The All of Us Research Program Genomics Investigators, Manuscript Writing Group, Bick, A. G., Metcalf, G. A., Mayo, K. R., Lichtenstein, L., Rura, S., Carroll, R. J., Musick, A., Linder, J. E., Jordan, I. K., Nagar, S. D., Sharma, S., Meller, R., All of Us Research Program Genomics Principal Investigators, Basford, M., Boerwinkle, E., Cicek, M. S., Doheny, K. F., … Denny, J. C. (2024). Genomic data in the All of Us Research Program. *Nature*. https://doi.org/10.1038/s41586-023-06957-x

Viel, K. R., Howard, T., Curran, J. E., Almasy, L., Moses, E. K., Glahn, D. C., Ameri, A., & John, B. (2007). Heritability of Commonly Measured Blood Cell Phenotypes in the San Antonio Family Heart Study. *Blood*, *110*(11), 3687–3687. https://doi.org/10.1182/blood.V110.11.3687.3687

Vuckovic, D., Bao, E. L., Akbari, P., Lareau, C. A., Mousas, A., Jiang, T., Chen, M. H., Raffield, L. M., Tardaguila, M., Huffman, J. E., Ritchie, S. C., Megy, K., Ponstingl, H., Penkett, C. J., Albers, P. K., Wigdor, E. M., Sakaue, S., Moscati, A., Manansala, R., Lo, K. S., … Soranzo, N. (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell, 182(5), 1214–1231.e11. https://doi.org/10.1016/j.cell.2020.08.008

Wheeler, M. M., Stilp, A. M., Rao, S., Halldórsson, B. V., Beyter, D., Wen, J., Mihkaylova, A. V., McHugh, C. P., Lane, J., Jiang, M. Z., Raffield, L. M., Jun, G., Sedlazeck, F. J., Metcalf, G.,

Yao, Y., Bis, J. B., Chami, N., de Vries, P. S., Desai, P., Floyd, J. S., … Reiner, A. P. (2022). Whole genome sequencing identifies structural variants contributing to hematologic traits in the NHLBI TOPMed program. Nature communications, 13(1), 7592. https://doi.org/10.1038/s41467-022-35354-7

Women and Health Research: Ethical and Legal Issues of Including Women in Clinical Studies, Volume 1 (p. 2304). (1994). National Academies Press. https://doi.org/10.17226/2304

Yu, J.-H., Taylor, J. S., Edwards, K. L., & Fullerton, S. M. (2012). What Are Our AIMs? Interdisciplinary Perspectives on the Use of Ancestry Estimation in Disease Research. *AJOB Primary Research*, *3*(4), 87–97. https://doi.org/10.1080/21507716.2012.717339

Yudell, M., Roberts, D., DeSalle, R., & Tishkoff, S. (2020). NIH must confront the use of race in science. *Science*, *369*(6509), 1313. https://doi.org/10.1126/science.abd4842

Zhang, H., Zhan, J., Jin, J. et al. A new method for multiancestry polygenic prediction improves performance across diverse populations. Nat Genet 55, 1757–1768 (2023). https://doi.org/10.1038/s41588-023-01501-z

Appendix

The Appendix contains Supplemental Information, links to resources such as statistical code, qualitative codings, and documentation referenced throughout the dissertation.

Supplemental figures and tables

Figure S1. Boxplots of MCV trait values and age distribution of base cohort in TOPMed GWAS

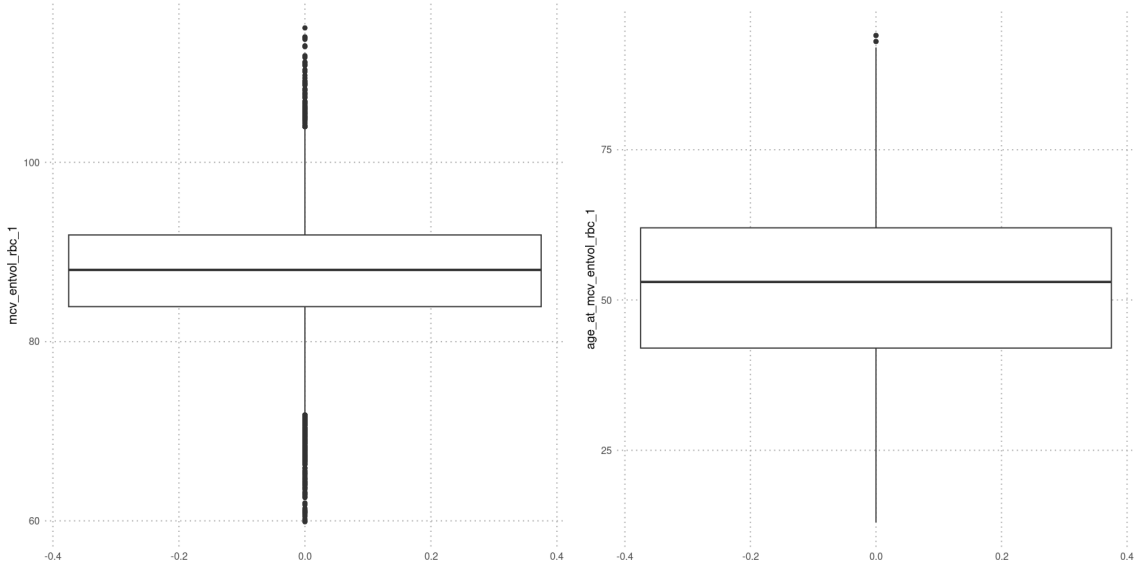


Figure S2. Rank inverse normal distribution of MCV (TOPMed)

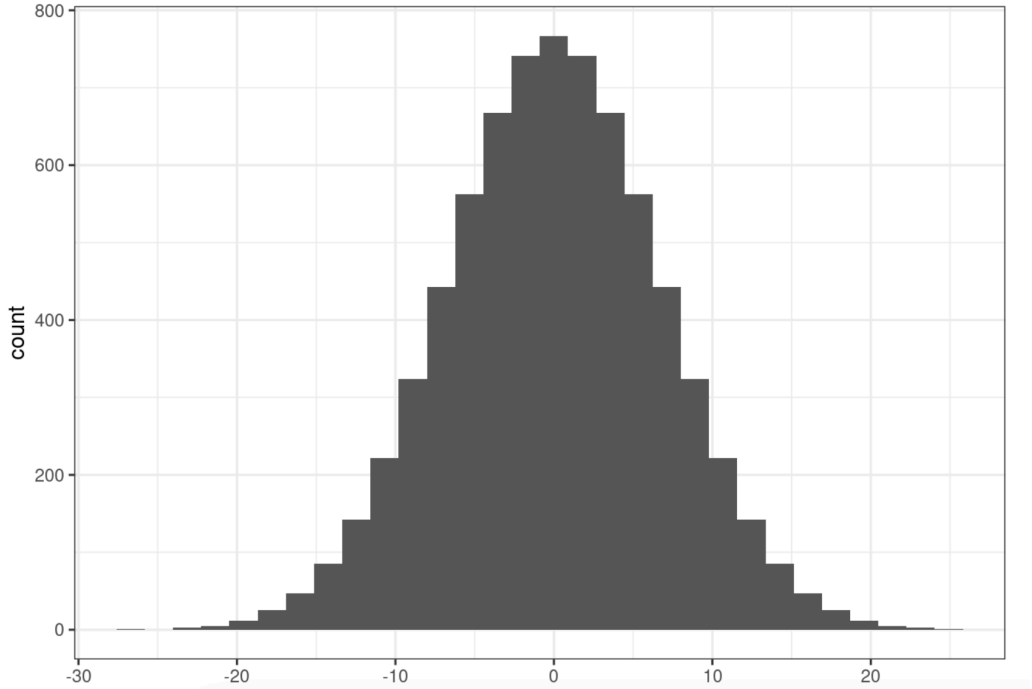


Figure S3. Boxplots of HGB trait values and age distribution of base cohort in TOPMed GWAS

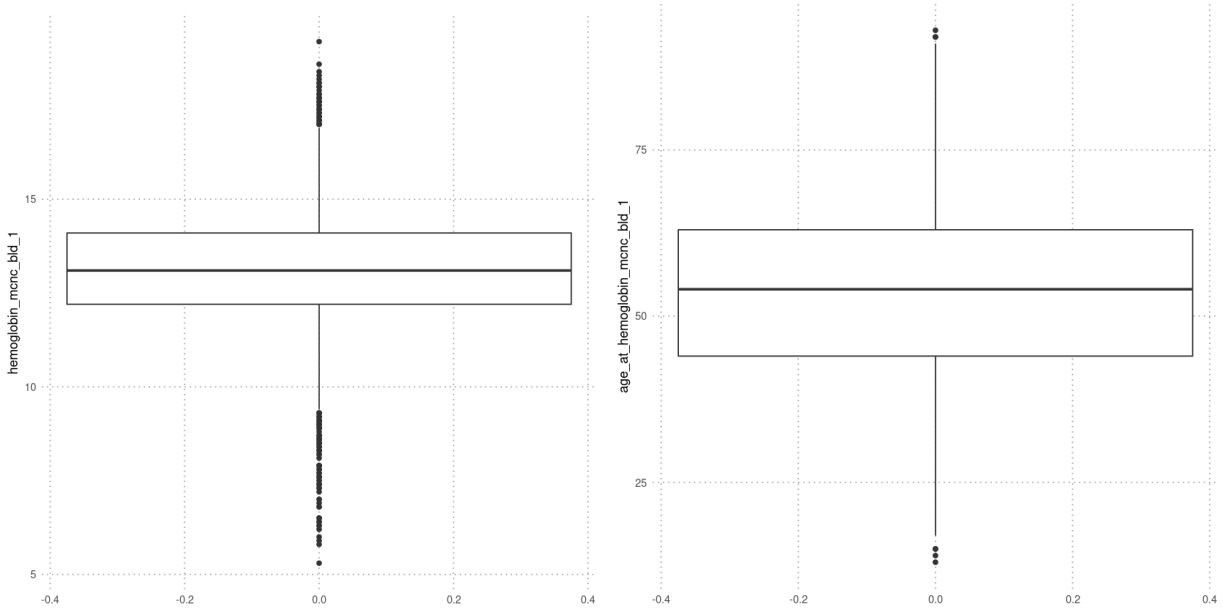


Figure S4. Rank inverse normal distribution of HGB (TOPMed)

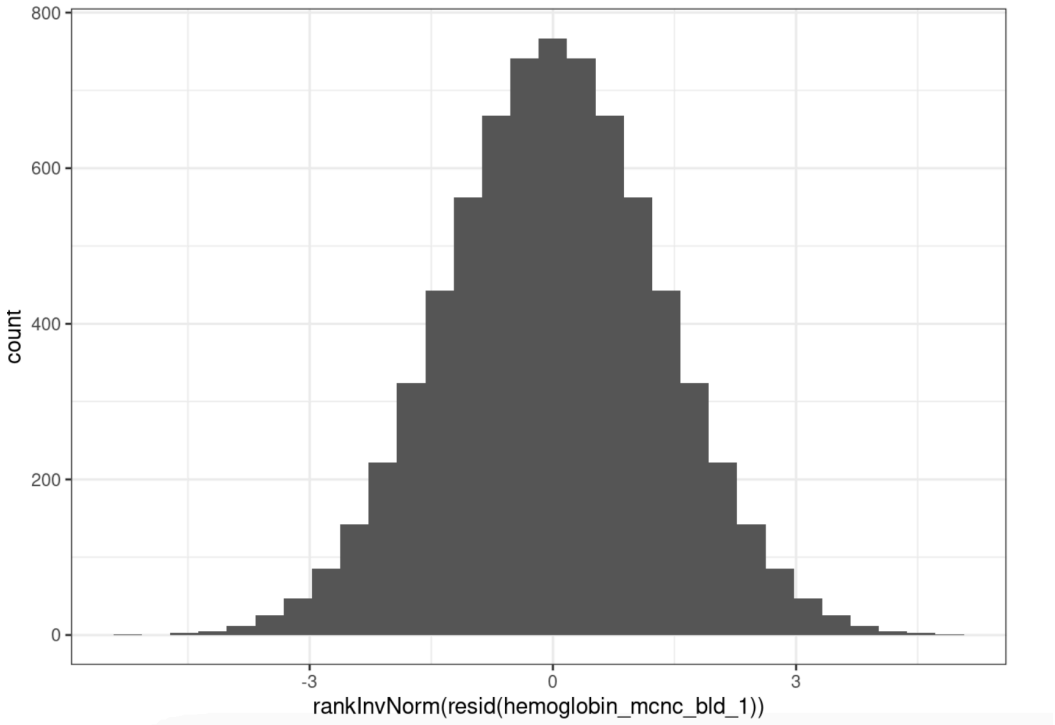


Figure S5. Box plot of MCV trait values and age of TOPMed optimization cohort

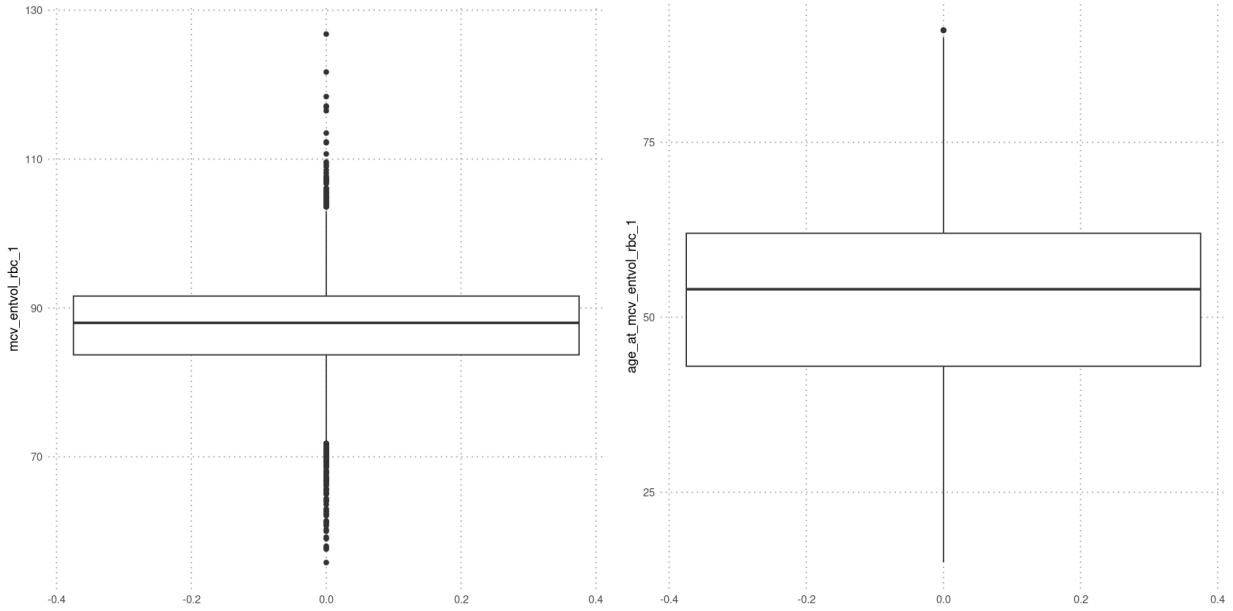


Figure S6. Box plot of MCV trait values and age of TOPMed evaluation cohort

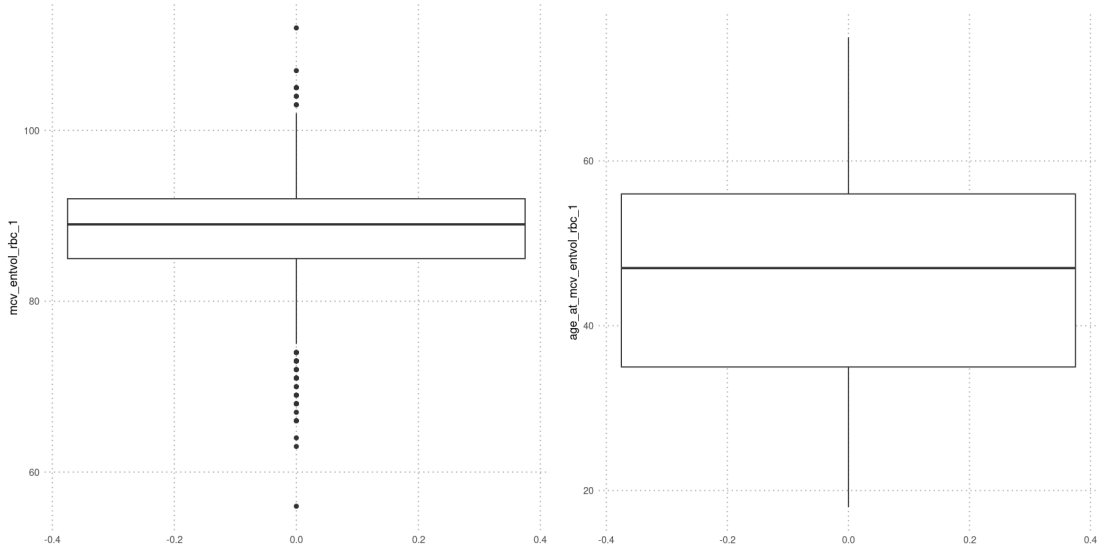


Table S1. Aim 1 Cohort demographics (TOPMed)

Base cohort: MCV			
	N		7327
	mean age		51.3
	Sex		
		Male	2800
		Female	4527
	MCV range (mean)		87.5 [53,140]
Target cohort: MCV			
	N		4958
	mean age		51.7
	Sex		
		Male	1837
		Female	3121
	MCV mean (range)		87.36 [55.8,126.8]
Base cohort: HGB			
	N		7327
	mean age		52.62
	Sex		

		Male	2582
		Female	4745
	HGB range (mean)		13.12 [5.3, 19.2]
Target cohort: HGB			
	N		7127
	mean age		52.96
	Sex		
		Male	2468
		Female	4659
	HGB mean (range)		13.25 [5.50, 19.8]

Table S2. Aim 1 Cohort demographics (UKB)

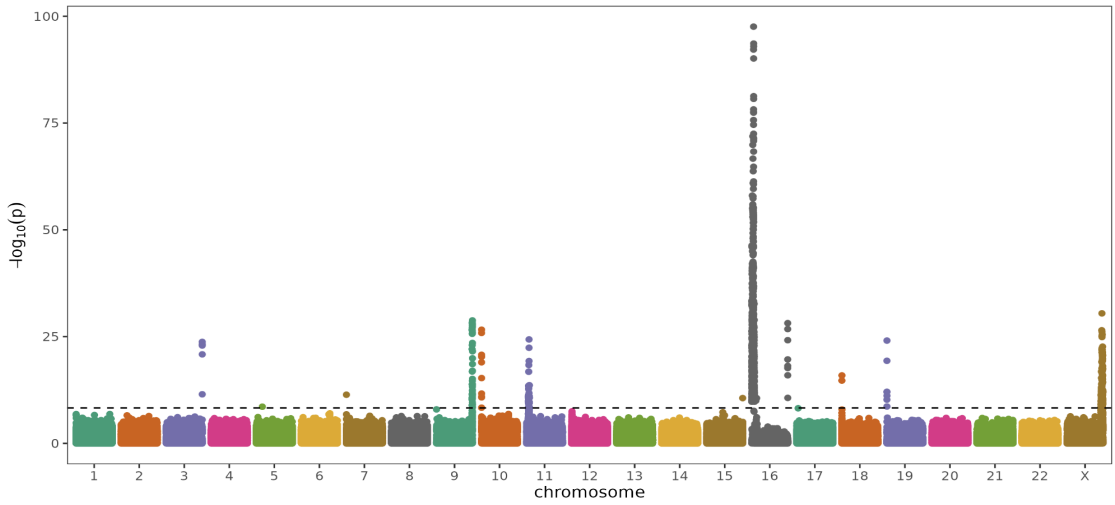
	Female (N=4190)	**Male (N=3137)**	**Overall (N=7327)**
age			
Mean (SD)	52.0 (7.89)	51.0 (8.04)	51.6 (7.97)
Median [Min, Max]	51.0 [40.0, 70.0]	49.0 [39.0, 70.0]	50.0 [39.0, 70.0]
MCV			
Mean (SD)	87.5 (6.43)	88.0 (5.93)	87.7 (6.23)
Median [Min, Max]	88.0 [56.9, 114]	88.4 [65.0, 108]	88.2 [56.9, 114]
Hgb			
Mean (SD)	12.8 (1.12)	14.6 (1.06)	13.6 (1.41)
Median [Min, Max]	12.8 [4.78, 16.7]	14.6 [8.40, 18.4]	13.6 [4.78, 18.4]

Table S3. Findings for genome-wide significant loci and genes as compared to those identified by Hu et al (2021)

						TOPMed		UKB	
		chr	**pos-start**	**pos-end**	**Gene**	**Genome-wide significance?**	**Included in PRS?**	**Genome-wide significance?**	**Included in PRS?**
MCV	Genome-wide significant genes	11	5224309	5225461	AC104389	no	no	no	yes
		11	5225464	5229395	HBB	yes	yes	yes	no
		11	5224448	5224639	RF00621	no	no	no	yes
		11	5544489	5548533	OR52H1	yes	yes	no	yes
		11	5248079	5249859	HBG1	yes	yes	no	yes
		16	176680	177522	HBA1	yes	yes	yes	yes
		22	37065436	37109713	TMPRSS6	no	no	yes	yes
		X			G6PD	yes	yes	yes	yes
		chr	**pos**	**rsid**	**Gene**	**Genome-wide significance?**	**Included in PRS?**	**Genome-wide significance?**	**Included in PRS?**
	Genome-wide significant loci	3	128,603,774	rs112097551	RPN1	yes	no	yes	no
		11	5,226,774	rs11549407	HBB	no	no	no	no
		16	55,649	rs868351380	HBA1	no	no	no	no
		19	1,253,643	rs73494666	MIDN	yes	no	yes	no
		22	37,108,472	rs228914	TMPRSS6	yes	no	yes	no

Figure S7. Manhattan plots of TOPMed and UK Biobank GWAS results for MCV and HGB

A. Manhattan plot for MCV TOPMed GWAS results (lambda = 1.007)



B. Manhattan plot for HGB TOPMed GWAS results (lambda = 1.008)

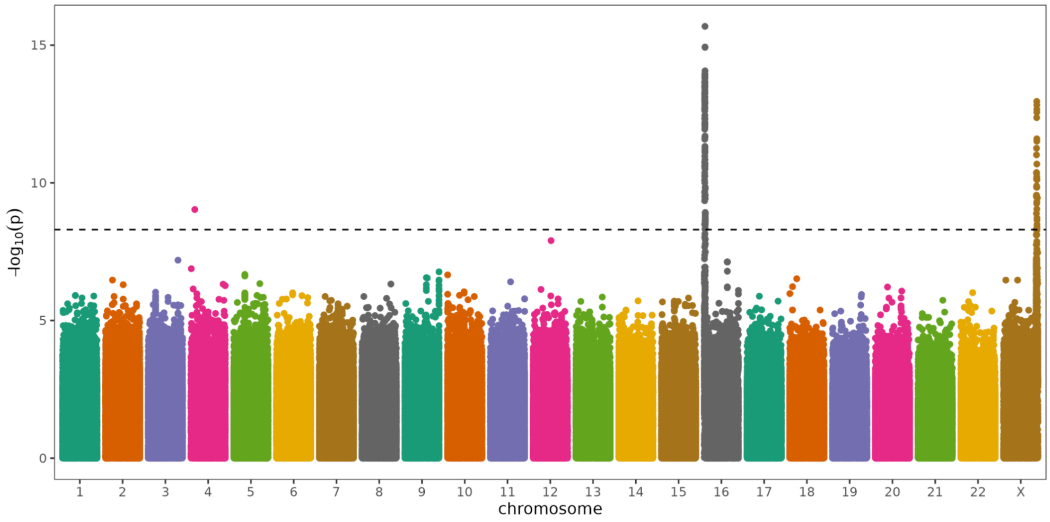
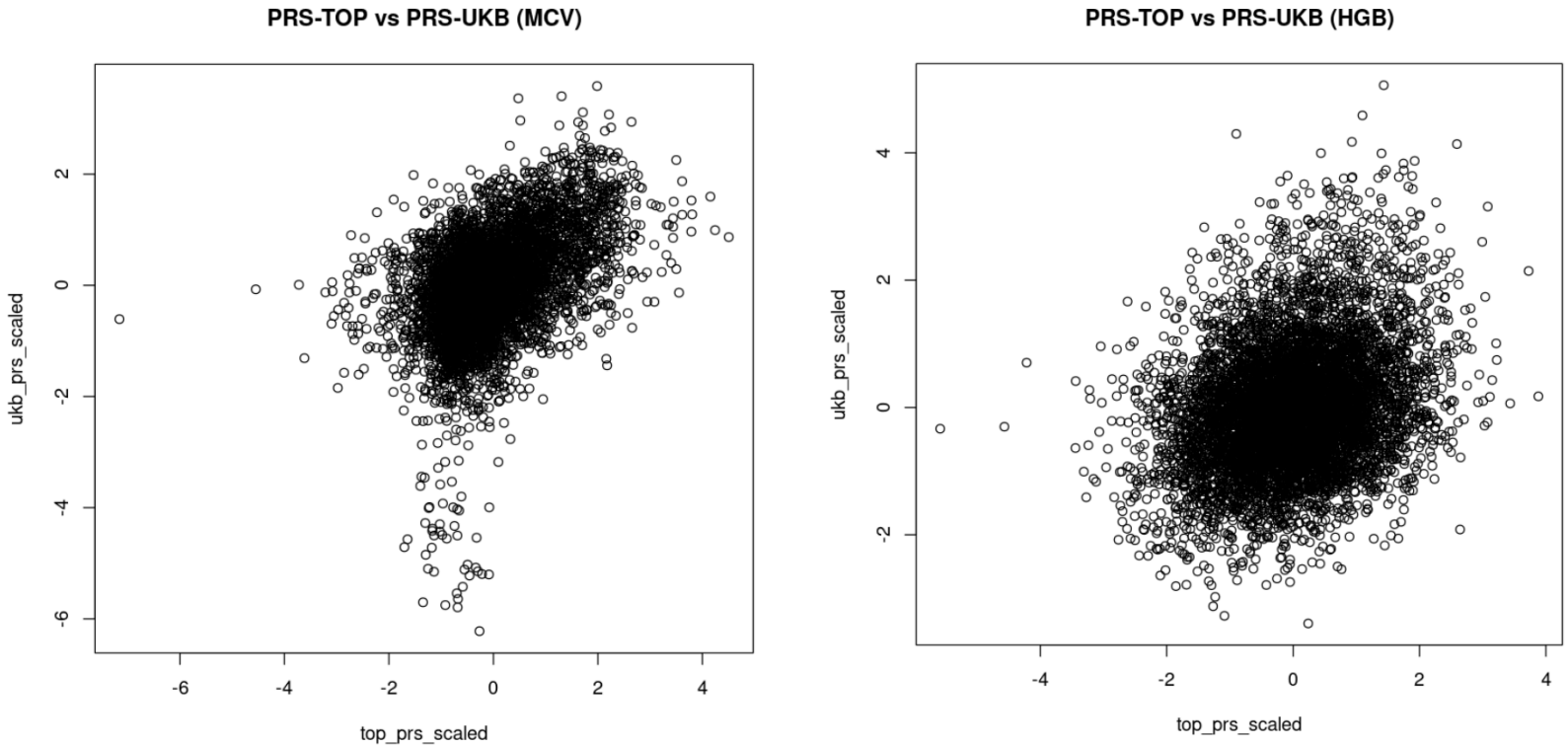


Table S4. Metrics for best performing PRS models for MCV and HGB

Trait	Model	Phi	PVE	Beta	P-value	Number of variants	Number of significant variants (autosomes)	Number of significant variants (X)	Number of overlapping variants	Top 5th %ile	Bottom 5th %ile
MCV	PRS-TOP	1.00E+00	0.0277	1.1438	1.37E-31	1375284	96	18	1236574	1.86	-1.34
	PRS-UKB	1.00E-02	0.0234	1.1778	7.22E-27	1368046	93	14		1.56	-1.47
HGB	PRS-TOP	1.00E-06	0.0081	0.1196	3.18E-14	1374708	3	4	1218994	1.67	-1.63
	PRS-UKB	1.00E-06	0.0072	0.1514	8.80E-13	1367901	13	0		1.82	-1.48

Figure S8. Comparison of PRS scores (Optimization) for MCV and HGB.



Larger correlation between the two PRS models for MCV (r=0.46) than for HGB (r=0.28)

Degree of overlap increases at higher percentile strata. For example, if defining “high score stratum” by the 10th percentile (instead of 5th percentile), then the degree of sample overlap between the two PRS models for both MCV and HGB in the high score stratum increases (MCV: 48.6%; HGB=27.4%)

Figure S9. Beta comparisons of overlapping SNPs across phi values (MCV, chr 16, i=1)

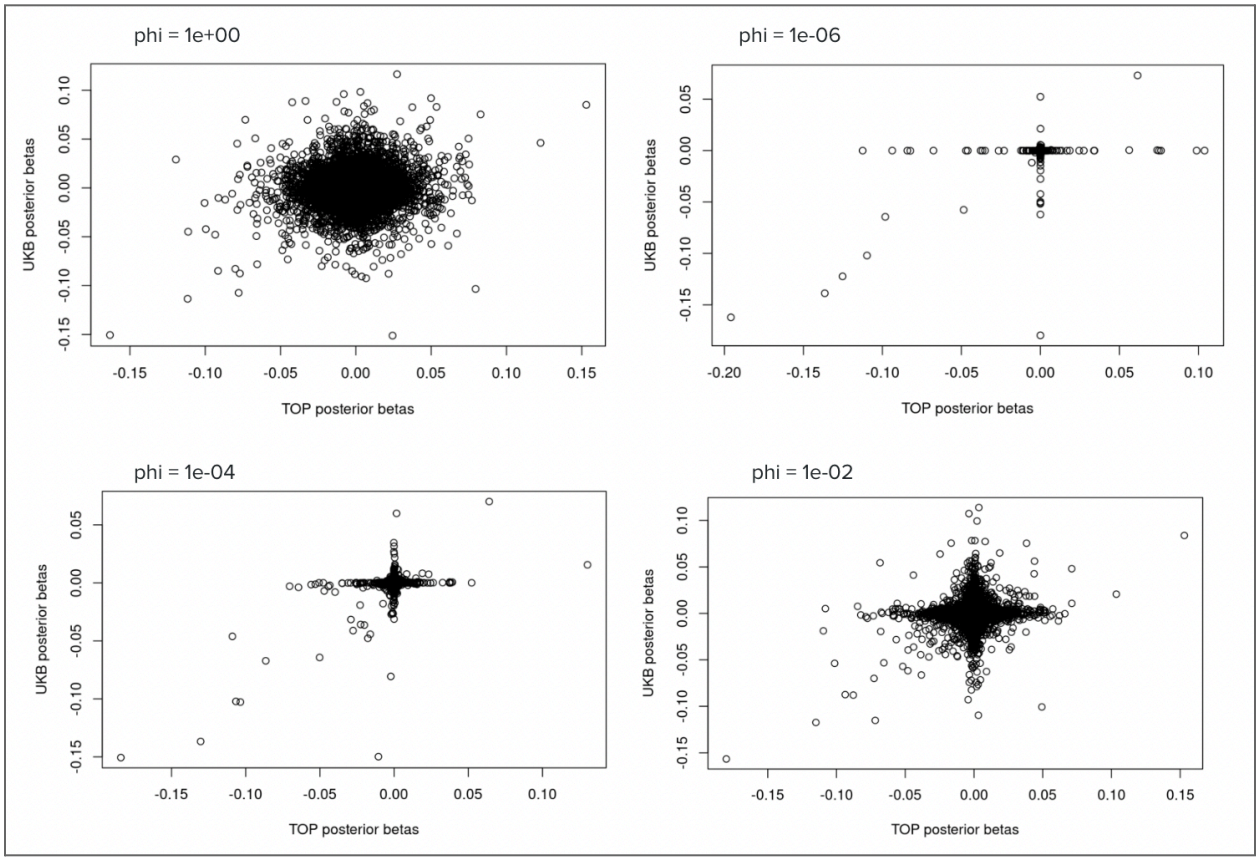


Figure S10. Beta Comparisons of overlapping SNPs across phi values (MCV, chr 16, i=2)

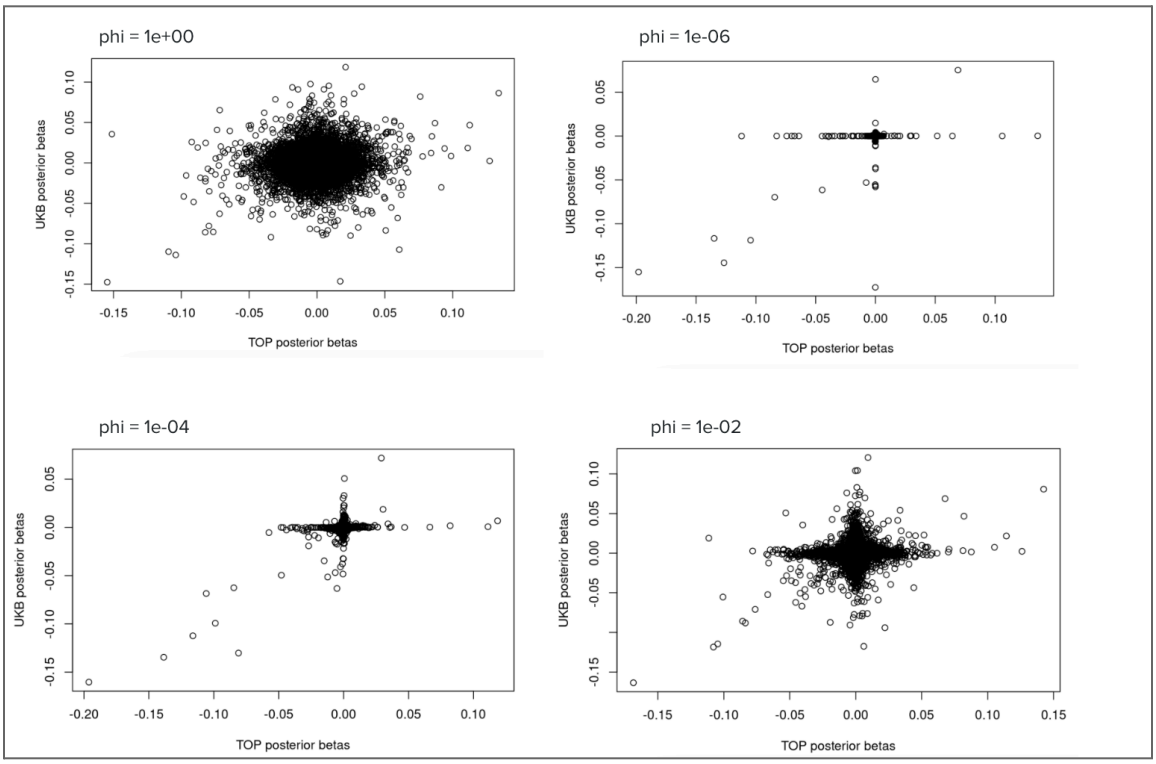


Figure S11. Comparing betas of overlapping SNPs across values of phi (MCV, chromosome X, i=1)

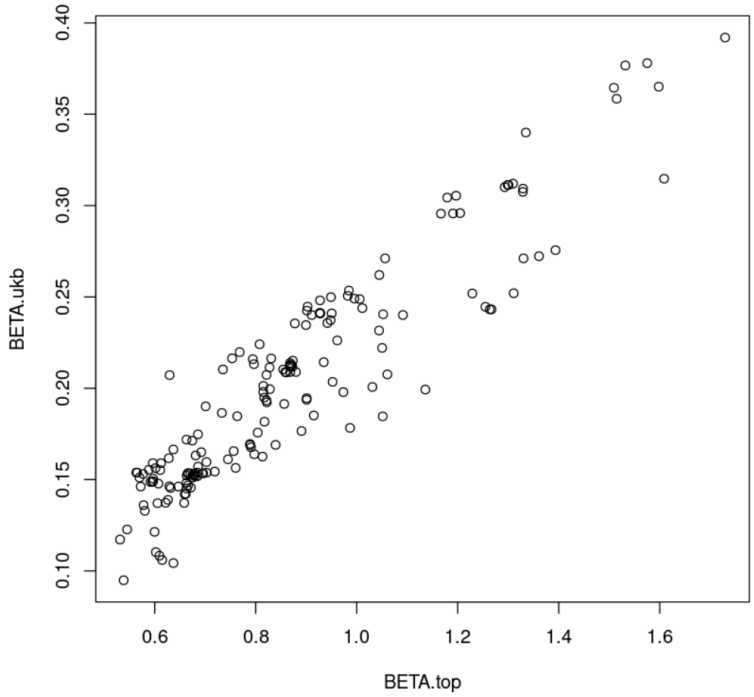
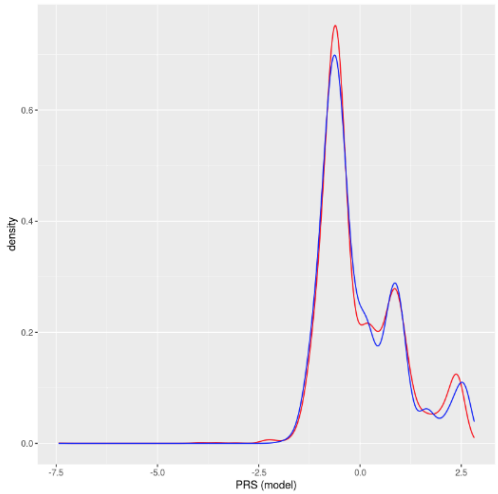


Figure S12. Estimated PRS using genome-wide significant SNPs only (i=1). Peaks in the PRS distribution are likely driven by a subset of individuals with few variants of high effect size, driving the peak in PRS score. There is greater concordance in the "high score" category between the two PRS models (i.e., the majority of individuals categorized in the top 5th percentile by one PRS model were also categorized in the top 5th percentile by the other PRS model). However, this was not observed for the bottom 5th percentile. There was no overlap of samples in the bottom 5th percentile between the two PRS models. While we observed the greatest degree of overlap in score assignment in the top 5th percentile (85.4%) after computing a coarse PRS using only SNPs reaching genome-wide significance in each set of summary statistics (PRS-TOP=128; PRS-UKB=117), the low score quantile showed modest concordance (58%); however, 14.6% of individuals were nevertheless classified differently by the two PRS models in the highest-score decile, while 42.3% of individuals were classified different in the lowest-score decile, and observation which warrants further examination.



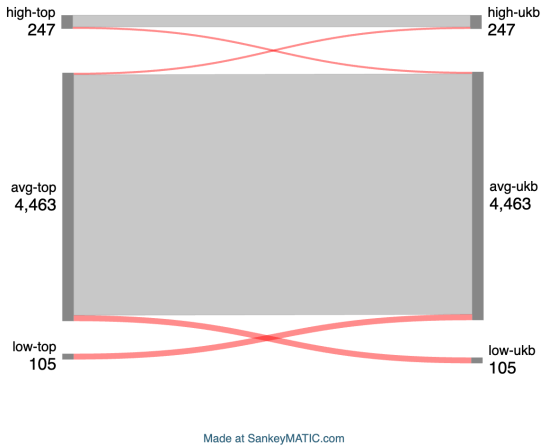


Table S5. Frequencies across score deciles

PRS-TOP-Sig	**Average**	**Average**	**Average**	**High**	**High**	**Low**
PRS-UKB-Sig	**Average**	**High**	**Low**	**Average**	**High**	**Average**
N	4322	36	105	36	211	105

Figure S13. PRS distributions for HGB

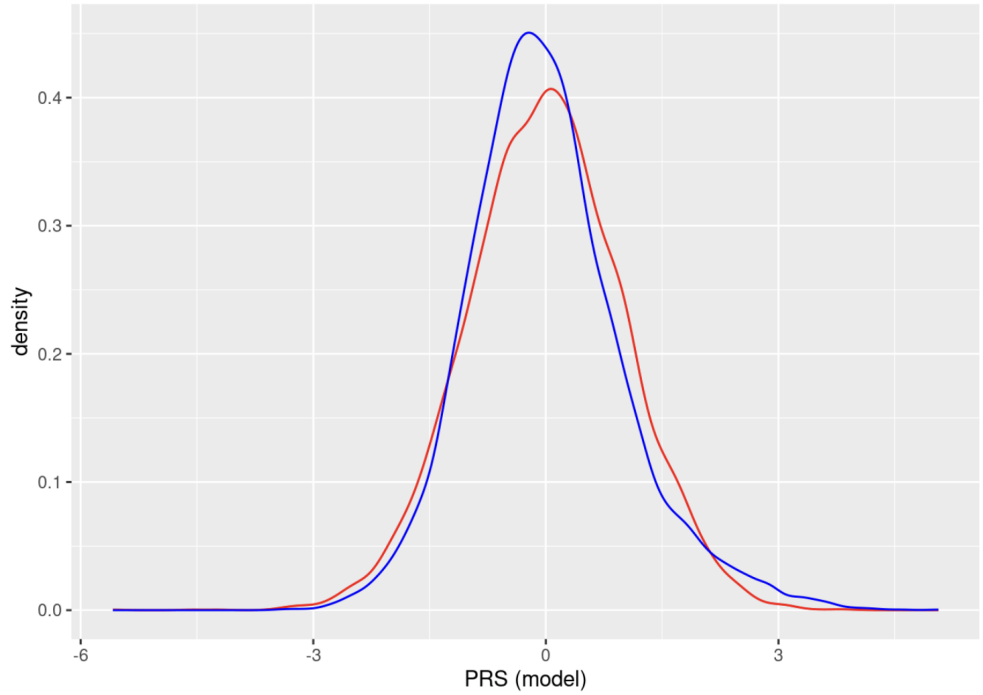


Figure S14. Number of participants assigned to each PRS stratum

prs_top	prs_ukb	n
avg	avg	5821
avg	high	299
avg	low	294
high	avg	297
high	high	56
high	low	3
low	avg	296
low	high	1
low	low	60

High = top 5th percentile; avg = middle quantiles (2-19); Low=bottom 5th percentile.

Figure S15. Overlapping SNPs across all autosomes in the HGB PRS models followed a cross shape as in the figure below.

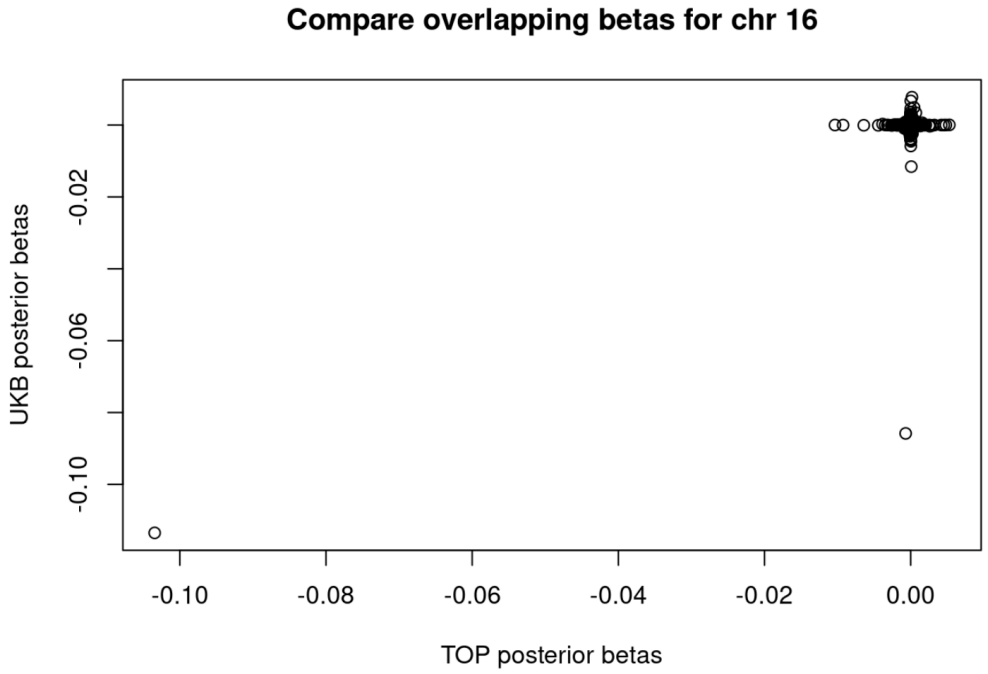


Table S6a. Example original study data

subject_id	Race	Ethnicity	Study documentation
US_01	Asian-Pakistani, Native American	Hispanic	Place of birth - US
US_02	Asian-South Asian, White-Arab	Not Hispanic	Place of birth - US
US_03	Asian	Not Hispanic	Place of birth - US
US_04	Black	Not Hispanic	Place of birth - US
UK_01		Pakistani, Dutch	Place of birth - NED
UK_02		Pakistani	Place of birth - PAK
UK_03		Indian	Place of birth - UK
UK_04		Bangladeshi	Place of birth - UK

Table S6b. Example original study data formatted to a Population descriptors data table

subject_id	population_descriptor	population_label	country_of_recruitment	country_of_birth
US_01	Race | Race | Ethnicity	Pakistani | Native American | Hispanic	US	US
US_02	Race | Race | Ethnicity	South Asian | Arab | Not Hispanic	US	US
US_03	Race | Ethnicity	Asian | Not Hispanic	US	US
US_04	Race | Ethnicity	Black | Not Hispanic	US	US
UK_01	Race | Race | Background	White | Pakistani | Western	UK	NED
UK_02	Ethnicity	Pakistani	UK	PAK
UK_03	Ethnicity	Indian	UK	UK
UK_04	Ethnicity	Bangladeshi	UK	UK

Table S6c. Population label mapping

population_label	Mapping-1	Mapping-2
Pakistani | Native American |	Admixed South Asian and Native American	Admixed American

Hispanic	in US	
South Asian | Arab | Not Hispanic	Admixed Asian in US	Admixed Asian ancestry
Asian	Asian in US	Asian ancestry
Black	Black in US	African ancestry
Pakistani | Dutch	Admixed South Asian in UK	South Asian ancestry
Pakistani	South Asian in UK	South Asian ancestry
Indian	South Asian in UK	South Asian ancestry
Bangladeshi	South Asian in UK	South Asian ancestry

Links to resources and documentation

1. TOPMed DCC harmonization of race and ethnicity: https://github.com/UW-GAC/topmed-dcc-harmonized-phenotypes/blob/master/harmonized-variable-documentation/demographic/race_us_1.json
2. EPACTS documentation: https://github.com/statgen/EPACTS
3. PRS-CS documentation: https://github.com/getian107/PRScs

Codebook

Category	*Code*	*Definition*	*Inclusion/exclusion*	*Examples*
Ethical	Access	Comments about access to PRS as a clinical tool.	Includes patient access, provider access, or discussing access more generally, like whether an entire population may or may not have access to PRS	"So I think, you know, access to these tests, it's a much bigger problem than whether this test will be slightly more beneficial in one population versus the other."
Analytical	Admixture	Comments on admixture.	Includes comments on mixed ancestry	"I think that is a really non-inclusive approach, because it doesn't allow for individuals who have high admixture to be included in in the analysis, and there are some populations that have really high admixture where we may exclude a lot of participants" "Like so many people have mixed ancestry. What do you do?"
Analytical	Allele frequency	Comments on allele frequency		
	Ancestry	Comments about ancestry. It can be referring to genetic ancestry or non-genetic ancestry. Phrases or sentences that include the word ancestry in them.	Includes general comments about different or multiple ancestries. Excludes comments on population- or ancestry-specific variants or effects or observations	"for each person, they're calculating a polygenic risk score, but you've got people of different ancestry backgrounds, so you know this doesn't predict quite as well."

	Ancestry-specific or population-specific	Comments about or using terms like ancestry-specific or population-specific.	Includes comments discussion population-specific scenarios or events (population or ancestry specific variants, effects, observations). Excludes comments that mention "different ancestries" because that is not a comment about a specific population.	" So, if you look at the prevalence of diabetes across those 8 bins, there is a gradient from lowest at the no European ancestry to highest, or lowest, being no native of ancestry to the highest, being full native ancestry, in terms of the diabetes risk. And if you look at the frequency of this haplotype, it's in the other direction. Highest among folks who have no native ancestry, lowest among folks who have native ancestry"
General	Background experience	Captures comments on the participant's background education, work history, experience with GWAS or PRS.		
Secondary	Bad/con/challenges	Comments that are negative or discuss challenges		
Analytical	Causal variant	Comments about causal variants, causality	Includes negative comments about causality (ie, comments referring to something being not causal, not having a causal relationship, not finding the causal variant, etc)	" Polygenic risk scores are more correlative, or are sometimes used to get at biology when they don't really. Or they may not really"

Secondary	Change	Comments that indicate a change in thought, change in behavior, change in a practice. It can refer to the individual's personal change or a change that has occurred on a broader, bigger scale (a change in the field, a change in practice, etc).		"you know, it reminds her about the time of when we did a lot of imputation and thinking we needed to do the imputation for different racial/ethnic groups separately and eventually we learned, oh, you know, just throw them all together. And that is better."
Ethical	Clinical or healthcare	Comments referring to clinical or healthcare considerations, applications, motivations, or other types and references to clinical or healthcare.	Includes comments about healthcare systems (e.g., their readiness to handle PRS or data required for PRS)	"and then the second one being that they're not immediately interpretable, and I wonder how well clinicians are able to really deal with these... to deal with the score and to know what it means and how to interpret it and what.... How they should work with the patient with whatever given score they have."
Ethical	Communication	Comments about communication.	Includes communication of scientific results, communication of risk to a patient, reporting, or communication more broadly.	" we really need to test it in clinical trials and see how people respond to knowing this information. How do they like it? Can they understand it? We need to do a lot of, you know, in terms of communication of risk."
Analytical	Confounding or correlation or association	Comments about confounding, correlation, or		" Polygenic risk scores are more correlative, or are sometimes used to get at biology when they don't really.

		association.		Or they may not really"
	Consequences or results or impact	Comments about the consequences of something, the results of something (like a study or a practice), or the impact of doing something (like impact, consequence, or result of using a specific method).		"So they've shown that, you know, communicating high genetic risk to providers can result in ordering imaging tests, ordering statin treatment when that might not otherwise have occurred" "So we'll use that as like a QC check, and look at each self-reported race/ethnic group for outliers. You may do it and you may see some extreme outliers. This is actually an area that bothers me, because I think what we usually do with those people we exclude them. You know, I hate the idea of excluding people, because you know, they work just as hard as anybody else to contribute. And yeah, the idea of not using data."
	Continental ancestry	Comments about or using the term "continental ancestry."		
Technical	Data infrastructure/access/type	Comments on data infrastructure, data access, or data type.	Includes comments on summary statistics, individual level data, SDOH, etc.	" there's always the limitations that you have to have individual level data to test your PRS. So the derivation sample can be summary statistics, which are available for most of the phenotypes by now. But to test a PRS efficiently in an independent testing example, you need individual level data and this can be, of course,

				with some smaller sample size than the derivation sample or just individual level data is not readily available. So that's one of the big limitations in the PRS analysis."
Technical	Data we have	Comments about the data we currently have.	Includes comments that reference what data we do or do not have currently.	"I mean there's really not... I mean you can develop all the complicated analytical methods you want. But if you don't have the data there from the people that you want to study, it's really not going to be.... I mean, you can only do so much with what you have, I guess is the challenge, right?"
Ethical	Decision-making/rationale	Comments describing the participant's decision-making or rationale for making particular analytical choices.	Includes any explanation of how a particular method is used and why; how particular decisions within a method were made (e.g., number of PCs). Includes how components within a method were defined in an analytical sense (e.g., how an ancestry group was defined using a particular method). NOTE: this does not refer to definitions of concepts or definitions of a practice or methodology itself.	" I think that was one of our tactics, we approached, you know, looking mostly at generic ancestry. And then, but. how do you define the ancestry group? So what we did was you take the center, and then you take like 4 or 5 standard deviations, you know, within, around that, and then that will be the ancestral group." "And I think the weaker odds ratio in Africans in particular, is because they have lower LD. And there are other reasons as well - allele frequencies, LD blocks, which are very different from other groups. So that's why, if you want to increase the ratio and bring parity, you really need to get very, very large samples of African individuals so that you can create a more specific ancestry-directed PGS

				for those individuals."
Secondary	Definition	Definition of a term, concept, or practice. Also refers to comments about the definition of something.		"I guess I would say a GWAS is broadly, a study of the genetic influences or genetic risk factors for a disease or phenotype" "I think there are a lot of issues in terms of defining base and target populations and a lot of issues in terms of defining and harmonizing phenotypes."
Ethical	Different ORs-is not okay	Answer is that having different odds ratios is not okay.		
Ethical	Different ORs-is okay	Answer is that having different odds ratios is okay.		
Ethical	Different ORs-maybe-I don't know	Answer is that participant isn't sure whether having different odds ratios in PRS results is okay or is not okay. Or indicates that sometimes it may be, other times it may not be. Or answer is unclear.		
Secondary	Disagree	A code to disagree with the assigned code.		

Ethical	Diversity or representation	Comments about diversity or representation.		
	DTC tests	Comments about DTC tests		
Ethical	Ethical or equity	Comments about ethical considerations or thoughts on equity.		
Analytical	Generalizable	Comments on generalizability of results. Can refer to PRS or GWAS.		"So I think in this case, Okay, they're saying they've identified which variants are the risk variants in these different populations, but I don't know how generalizable that is"
Analytical	Genetic factors	Comments about genetic factors involved in an observation or phenomenon (e.g., GWAS results), specifically referred to opposite to non-genetic factors.		
Secondary	Good/pro/benefits	Comments that are positive or discuss benefits		
	Grants/Funding	Comments referring to grants and funding as impacting a decision.		
General	GWAS general	General comments		

		on GWAS.		
Analytical	GxE	Comments about gene by environment interactions.		
Analytical	Imputation	Comments about imputation.		
	It depends / Depends on the trait	Refers to when a participant says "it depends" in response to a question. Specific attention to when an explanation involves being dependent on a trait		
Secondary	Juicy quote	An especially interesting, controversial, or otherwise compelling quotation.		
	Legacy/Back in the day	Comments referring to doing what's been done in the past		"as the majority of methods that existed at the time were designed for homogeneous [panoptic] populations, like the large European data sets that were out there at the time. European ancestry"
Analytical	Linkage disequilibrium	Comments on linkage disequilibrium.		

	Maybe-I don't know-no thought yet	Refers to when a participant is unsure, ambivalent, or doesn't know due to not having given thought to the topic, being unfamiliar with the topic, or because they just don't know (e.g., despite presence of evidence).		" Honestly, I'm a little bit unsure. I guess you know, missing heritability is a concept. It's a problem." "And that's where I don't know enough about the history of the field, and that's where these questions about benchmarking like, What does it I mean if adding in PRS helps you explain 2% more phenotypic variation. I don't know."
	Maybe-jury is still out-need more evidence	Refers to when a participant is unsure, ambivalent, or doesn't know due to the lack of evidence or lack of consensus in the scientific community.	Includes comments where the evidence is present, but it isn't sufficiently convincing to participant.	"So I guess that's my feeling about it. I haven't gotten enough... I think as a population genetics study, I haven't gotten enough evidence."
	Measuring risk	Comments on measuring risk. This includes models that measure risk, metrics for measuring risk, general comments about measuring risk, such as the challenges in doing so, or the motivations or	Includes comments on absolute and relative risk.	"But when we're building these risk models and thinking about implementing them clinically, and we want to return absolute risk numbers to people, we're not just returning a relative risk, we're also returning the absolute risk, and the absolute risks vary a lot from population to population, for reasons that have nothing to do with polygenic risk score."

		benefits of doing so.		
Analytical	Method or adjustment	Comments on methods or adjustments used in GWAS or PRS analyses. This includes general comments about methods as a topic itself.		"Methods are the solution"
Analytical	Model-noun	Comments relating to a model.		"So you know what I learned about population stratification bias, it was very much in the sort of balls and urns model of population genetics."
Analytical	Model-verb	Comments relating to modeling something.	Includes comments on developing a model	"You model the thing that you can't measure"
Secondary	No	A no answer		
Secondary	No code / Needs a code	There is no existing code for this quotation.		
Analytical	Non-genetic factors	Comments about non-genetic factors. Esp with respect to how GWAS and PRS are conducted, results interpreted, and applied		"There just haven't been very many non-genetic risk factors identified for prostate cancer" "But I think for a different disease where there is a stronger environmental component, then it really muddies the whole picture of what you would expect to see"

Secondary	Not sure which code	Not sure which code applies to this quotation.		
	One size fits all	Refers to when a participant uses this phrase or something similar, or applying this idea.		
	Odds ratios			
Analytical	Odds ratios from PRS	Comments about odds ratios from PRS, interpreted as the odds based on PRS of developing a disease		
Analytical	Odds ratios and effect sizes from GWAS	Comments about odds ratios from GWAS, interpreted as effect sizes or odds ratios		
	Outlook	Refers to the participant's outlook on the future or something. Particularly on the future of PRS, the future of the field.		
Analytical	Performance metric	Comments about performance metrics for PRS or for GWAS		"what measures we should be using to adjudicate the utility of a PRS score, I think, is very much open for debate"

	Personal / Personal beliefs	Comments sharing personal beliefs, what a participant thinks personally, or comments referring to something personal about the participant (e.g., family, family history, personal anecdotes, personal opinions on a practice, etc)	Excludes comments on personal approaches to methods in genetic analysis or background experience	(Include example) "I think it's not yet prime time. I doubt it. For me, that's my view. I think there's a lot of conflict. And I think that's the common view among the scientists. But there are people who have a certain hubristic view, a hubristic attitude that, you know, it's ready to go. Not really. It's not actually even ready within the context of European ancestry population. Because there are a lot of things which need to be refined." (exclude example) "So to me it's kind of a best practice that if you have, again like CHS, if we've got the whites and blacks, and we're going to do GWAS or PRS separately and then say, meta-analyze, really in my lab, we like to do the PCs separately, and then you adjust the PCs separately."
Analytical	Pooled	Comments on pooling in GWAS or PRS. This includes pooled analysis, or discussing the characteristics of pooling samples.		

Analytical	Population structure/substructure/stratification	Comments about population structure, substructure, or population stratification.	Includes general comments about structure in populations, or specific comments that use the words "strucutre" or "substructure" in terms of popuations.	"Because I know population in Africa is structured and most of the ancestry that is represented in GWAS in Western world is population which were predominantly sourced from West Africa."
Analytical	Prediction/Precision/Accuracy	Comments on the predictive ability of PRS, on PRS precision, or on PRS accuracy.		(Prediction) "You know, as genetic distance increases between base and target populations, you find a pretty linear decrease in predictive ability" (Accuracy) "Yeah, and what does that mean if they're accurate?" (Precision) "And then when you think about precision or accuracy, you think about the standard errors that you attached to it. And obviously the lower the standard error, the more accurate that predictor, the more useful that predictor could be, right?"
Analytical	Principal components	Comments on principal components.		
General	PRS general	General comments on PRS.		
	Publication	Comments referring to publication as impacting a deicision.		

	Race and/or ethnicity	Comments on race or ethnicity.		
Analytical	Reference populations	Comments on reference populations.		
Analytical	Relatedness	Comments about relatedness or relatedness matrix or other uses and implications of relatedness in genetic analysis.		
Analytical	Replicable	Comments on replicability of results. Can refer to PRS or GWAS.		
Analytical	Sample size	Comments about sample size.		"And the biggest limitation is we do not have the same sample sizes"
	Solution or what needs to be done	Comments on proposed solutions to a challenge, or what needs to be done to overcome or address a challenge.		"I'm going to lean heavy on the sort of design side, and say, that's why we need more diverse studies from more diverse contexts" "I just think that we should have like a really high burden of proof for PRS deployment"
Analytical	Stats: power	Comments about statistical power.		" And now more methods are coming out in which you don't actually lose any power by putting everybody together and including different factors"

				"I think small underpowered GWASs have muddied the field an awful lot"
Analytical	Stratified	Comments on stratification in GWAS or PRS. This includes stratified analyses or discussing the characteristics of stratifying samples. It is not limited to stratification by race, ethnicity, or ancestry. It can include comments on stratification by other variables, as well.		
Analytical	Stratified vs. pooled	Comments specifically comparing stratified analysis to pooled analysis, or taking a stratified approach vs. a pooled approach.		
Analytical	Study design/study question	Comments about study design.		
General	The field	Comments about the field of genetics. This includes comments on the history of the field,		

		the direction the field is going in, challenges in the field, changes in the field, comparing the field to other disciplines.		
Analytical	Transferability	Comments about transferability or transportability of PRS.		" And that, unfortunately, also affects portability to other ancestry groups, because the LD blocks are different, and a tag that is tagging the causal variant may not do it in the other ancestry groups, and that's why it's weaker" "The context behind how questions were asked to patients, and how data was collected can very deeply, I think, affect performance and transferability or portability of these scores"
	We're all admixed	Refers to this phrase or similar to it.		
Analytical	Stats: Who to sample	Comments about who to sample for analyses.		"We need to stop doing GWAS in Europeans" "All of them. If we're talking about the United States, I think we're really missing indigenous American populations in these studies."
Secondary	Yes	A yes answer		

Interview guide

General curiosity: *What are your perspectives on and experience with using race, ethnicity, and ancestry in genetic analyses and what you think the value and role of using these variables are? What are the challenges and benefits of using it (ie what, how, why, so what?)*

Before the call:

1. Send reminder 1-2 days before with Zoom info.
2. Set Zoom to record (but do *not* set up the Zoom meeting with the setting "Automatically record meeting" - interviewer will manually start recording after confirming with interviewee). Record to the Cloud.
3. Have Study ID in hand.
4. Sign into your UW Zoom account before joining

Before you start recording:

1. Confirm availability for the scheduled duration.
2. Verbal consent:

> ***The interview is designed to take no more than one hour. I will ask you a series of open-ended questions and you are completely free to decline answering any question. Your participation in this study is voluntary, you can withdraw from this interview at any time without consequences. I will audio record this interview with your permission and transcribe our conversation afterwards in order to review and analyze it. And you may see me taking notes occasionally, and that is simply me jotting down notes to help me keep in mind the topics we cover. Do I have your permission to begin audio recording?***

3. Confirm ok to audio record.

[Start recording here]

Introduction

Thank you for taking the time to speak with me today. To briefly introduce myself, my name is Alyna and I am a PhD student at the University of Washington at the Institute for Public Health

Genetics, and this interview is a part of a research study within my dissertation project, under the guidance of my chair, Malia Fullerton.

As you'll recall from the study information sheet I sent, the purpose of this study is to better understand the analytical motivations alongside the ethical considerations of conducting GWAS and PRS analyses in diverse populations. The interview questions are designed to help me understand the views of genetic scientists on the roles that race, ethnicity, and ancestry play in analyses and serve to structure a broad and open ended conversation.

I am interested in your thoughts, from your perspective as a researcher conducting GWAS and PRS in diverse populations, about the ways that race, ethnicity, and ancestry are used, interpreted, and translated into clinical practice. I'll start with a brief warm-up, and then cover three broad areas:

1. Analytical considerations when using race, ethnicity, and ancestry in analysis
2. Ethical considerations on the roles of race, ethnicity, and ancestry in genomics
3. Reflections (a reflective discussion)

Do you have questions for me before we begin?

Warm Up

1. To start off, I would love to know: Can you tell me a little bit about your background knowledge and experience working with GWAS and/or PRS in diverse populations?
    a. In your own words in plain language, how would you define a GWAS or a PRS; **what do you see is the purpose of a GWAS or PRS is?**

2. PRS are directly derived from GWAS and are intended to be used in the clinical space to inform people of their genetic risk of developing a particular trait or disease. **What in your mind are the main advantages to PRS?**
    a. How about some **potential limitations to using PRS?** This can include challenges in PRS methods and development or clinical implementation.

Analytical considerations

3. There are many methods and adjustments that an analyst can use to account for race, ethnicity, and ancestry in genetic analyses. **What kinds of methods or adjustments have you used in your analyses (and what was its purpose / what aspect of the analysis was it addressing)?(e.g., using PCs, performing stratified analyses,**

imputation strategies, using algorithms like HARE or performing k-means clustering, software like ADMIXTURE).

a. Can you tell me a little bit more about why/how you chose that (method)? Or if possible, walk me through an example of arriving at that decision point? (E.g., did the method involve X?)

   i. ***"Can you tell me more" examples:***
      1. *Can you tell me about how you got to 11 PCs? How you decided which samples belonged to which genetic ancestry groups (threshold, combined R/E info)? How you ended up with 5 groups in your ADMIXTURE analysis?*

   ii. ***(summarize) What I'm hearing is:*** *Type 1 error is a big concern that can be reduced if the data adhere to assumptions of normality. To adhere to assumptions of normality, various statistical techniques might be used.* ***Why is avoiding type 1 error important? Can it be done without using race/ethnicity/ancestry information in this way?***

   iii. Have your analyses included admixed individuals, and if so, how did you handle those populations?

Ethical considerations

[intro to ethical considerations section] We discussed how many analytical decisions can impact how people are grouped or categorized in GWAS. We also know that the less similar the PRS test population is to the population used in the GWAS, the less accurate a PRS will be. It has also been widely discussed that PRS are not equally predictive or accurate for people with non-European ancestry or have admixed ancestry. Specifically, ORs (for indicating likelihood of developing a disease) are different for different populations. Some people have vocalized that this could lead to healthcare disparities. **Could you share your thoughts on this concern?** (Do you think having different ORs in different populations is a problem? For whom might something like this be a problem? Under what circumstances might this be a problem?)

iv. [*if no, not really a concern that PRS are not equitable]* **Can you tell me a little bit more about that? Why do you think that is?**

v. *[if yes it's a concern]* **What do you think needs to happen to make PRS more equitable?**

   1. [*if "increase diversity so that we can make PRS equitable across all ancestries" response*

a. *Can you tell me a little more about how increasing diversity in GWAS would help make PRS more equitable? What would need to happen next?*

vi. *[prompt: there have been calls for equity by making PRS equally accurate for all/every ancestry(ies)"*] **What does "all/every ancestry/(ies) encompass?**

1. **What ancestry categories are absolutely necessary to include in this PRS equity work? Are there any others?**

a. **What makes you prioritize those ancestries over others?** What makes these categories useful as genetic ancestry categories?

Reflection

We have X minutes left, and so I want to bring us into the last main section, which is meant to be more of an opportunity for personal reflection.

4. In our discussion we've talked about various methods and adjustments in analyses that are made to meet various statistical requirements of normality or other things, and we've talked about some of the challenges in the implementation of PRS in the clinic. **If you were a participant in a study, or a patient in a clinic, where do you think your data point would fall in a genetic analysis or a polygenic risk score assessment?**
    a. **Do you think this has influenced your work at all?**

Wrap-up

Thank you very much for your time today. Now that we've been through this whole discussion, I'd like to ask:

5. Is there anything you think we should have asked you that we didn't?

    a. Is there something people often misunderstand/"get wrong" that you would want to clarify?

6. Would you be willing to speak with us again in the future (e.g. f/up questions)?

7. Is there anyone else who it may be helpful for us to speak with?

Demographics - NIH categories

Your age: _____

Your gender:
□ Male
□ Female
□ Other

Your race:
□ American Indian/Alaska Native
□ Asian
□ Native Hawaiian or Other Pacific Islander
□ Black or African American
□ White
□ More than one race

Your ethnicity:
□ Hispanic
□ Not Hispanic

Your disciplinary training (field): _____
Your disciplinary profession (field): _____

[Stop recording here]