Title

Uncertainty, portability and ancestry in polygenic scoring

Author

Ding, Yi

Publication Date

2024

UNIVERSITY OF CALIFORNIA

Los Angeles

Uncertainty, portability and ancestry in polygenic scoring

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioinformatics

by

Yi Ding

2024

ABSTRACT OF THE DISSERTATION

Uncertainty, portability and ancestry in polygenic scoring

by

Yi Ding
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2024
Professor Bogdan Pasaniuc, Chair

Polygenic score (PGS) is a tool for understanding an individual's predisposition to certain diseases or complex traits based on its genetic profile. In the burgeoning era of genomic medicine, PGS has emerged as a promising tool in advancing precision healthcare, demonstrating versatile utility such as patient risk stratification, disease risk prediction, and disease subtyping. However, its real application in clinical settings is limited by its uncertainty, bias, and low portability across diverse populations. For example, an individual may receive different genetic risk reports from different providers, and the score for a non-European individual may be less accurate than for a European individual. To fully understand and partially address these limitations, I first developed a Bayesian method to quantify the uncertainty in PGS at the individual level. I find trait-specific genetic architecture such as larger polygenicity and lower heritability combined with a small training sample size will lead to large uncertainty in PGS estimate, which in turn results in unreliable patient stratification in downstream analysis. Next, I expanded this approach to encompass individuals from varied genetic ancestry backgrounds. I find that the PGS performance varied from individual to individual with genetic distance playing a key role in impacting the performance of PGS; larger genetic distance from training data correlates with higher uncertainty and lower accuracy in testing individuals. These findings highlight the necessity of integrating individual-level PGS metrics in personalized medicine and the need for increasing genetic research diversity to ensure equitable and responsible use of PGS in clinical settings.

The dissertation of Yi Ding is approved.

Daniel H Geschwind

Paul Christopher Boutros

Sriram Sankararaman

Bogdan Pasaniuc, Committee Chair

University of California, Los Angeles

2024

To family.

Table of Contents

List of Figures

Acknowledgments

First and foremost, I want to express my deapest graditude to my advisor, Bogdan Pasaniuc. An old Chinese saying compares mentorship to the meticulous crafting of jade with chisel and stone. In a similar way, Bogdan's continuous guidance and constructive criticism in the past five years have been instrumental in shaping me into a more rigorous researcher and a more confident person. Working under the guidance of Bogdan, I learned the art of balancing overarching goals with instricate detials, and the ability of identifying and focusing the most important questions, which are invaluable lessons for my future career.

I would also like to express my sincere appreciation to the Bogdan Lab community. It has been a privilege to spend the past five years in such a friendly, collaborative, and supportive environment. I want to give my special thanks to Kangcheng Hou, Ruth Johnson, Kathy Burch and Rachel Mester, with whom I spent countless time talking about everything, from project ideas to career development and life advice. I would like to thank Arjun Bhattacharya and Harold Pimentel for giving me so many guidance when I looked for postdocs. I would also like to thank Ziqi Xu, Ella Petter, Sandra Lapinska, Kristin Boulier, Veronica Tozzo, Vidhya Venkateswaran, Helen Shang, Valerie Arboleda, Malika Kumar Freund, Tommer Schwarz, Jonatan Hervoso, Nicholas Mancuso, Yang Wu, Xinzhe Li, Alex Flynn-Carroll, Rob Brown, Claudia Giambartolomei, Sergey Knyazev, Igor Mandric, Arunabha Majumdar and many others for the intellectually-stimulating discussions during lab meetings, as well as the delightful and engaging conversations we've shared during lab lunches. Most importantly, the diverse culture and the respectful atmosphere among the lab members have taught me the value of being open-minded - both in embracing the diversity around me and in offering myself the same breadth of understanding and acceptance.

I owe special thanks to my commitee members Sriram Sankararaman, Paul Boutros and Dan Geschwind. Their profound knowledge and insightful critiques have significantly helped to cultivate my research. I want to thank collaborators within UCLA and other institutes: Timothy Chang, Bjarni Vilhjálmsson, Florian Privé, Tunc Morova, Alexander

Gusev, Matthew Freedman, Nathan Lack. I could not have done so much without their generous help.

I would also like to thank the entire Bioinformatics and Computational Medicine community for the tremendous inspiration and support. Scientific discussion on seminars and insightful conversations during donuts hour are invaluable memory of my graduate school. The experiences and friendship I gained here will be treasured forever.

In the end, I want to thank my friends and family for their unwavering support: Xinan Wang, Yumeng Ren, Eddie Li, Xinrui Cao, Jun Luo, Jiayi Tian, Shuang Gao, Yuan Meng, Hancong Wang and all my family members and many others. Their love and support are the bedrock of my motivation and resilience during the challenging phases of my PhD. Thank you all!

Permissions:

Chapter 2 is a version of Ding Y*, Hou K*, Burch KS, Lapinska S, Privé F, Vilhjálmsson B, Sankararaman S, Pasaniuc B. Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. Nat Genet. 2022 Jan;54(1):30-39. doi: 10.1038/s41588-021-00961-5. Epub 2021 Dec 20. PMID: 34931067 (* Denotes equal contributions)

Chapter 3 is a version of Ding Y, Hou K, Xu Z, Pimplaskar A, Petter E, Boulier K, Privé F, Vilhjálmsson BJ, Olde Loohuis LM, Pasaniuc B. Polygenic scoring accuracy varies across the genetic ancestry continuum. Nature. 2023 Jun;618(7966):774-781. doi: 10.1038/s41586-023-06079-4. Epub 2023 May 17. PMID: 37198491

Both reproductions are used here with permission from the authors and under the terms of their respective distribution licenses.

Curriculum Vitae

2010 – 2014	B.S. in Biotechnology
	Zhejiang University | Hangzhou, China
2016 – 2018	M.S. candidate in Biostatistics
	Harvard School of Public Health | Boston, MA
2018 – present	Ph.D. candidate in Bioinformatics
	Unversity of California, Los Angeles | Los Angeles, CA

Publications

Selected publications (of 21)

Ding Y, Hou K, Xu Z, Pimplaskar A, Petter E, Boulier K, Privé F, Vilhjálmsson BJ, Olde Loohuis LM, Pasaniuc B. Polygenic scoring accuracy varies across the genetic ancestry continuum. Nature. 2023 Jun;618(7966):774-781. doi: 10.1038/s41586-023-06079-4. Epub 2023 May 17. PMID: 37198491

Ding Y*, Hou K*, Burch KS, Lapinska S, Privé F, Vilhjálmsson B, Sankararaman S, Pasaniuc B. Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. Nat Genet. 2022 Jan;54(1):30-39. doi: 10.1038/s41588-021-00961-5. Epub 2021 Dec 20. PMID: 34931067

* Denotes equal contributions

CHAPTER 1

Introduction

Human traits are broadly classified into two major categories based on the number of genes influencing them: Mendelian (or monogenic) traits and complex traits [1]. Mendelian traits, like cystic fibrosis, typically result from mutations in a single gene. These mutations are highly penetrant and rare, leading to traits or diseases with direct and predictable manifestations. In contrast, complex traits exhibit a“polygenic” nature. They emerge from the cumulative impact of many genetic variants, each contributing a subtle effect to the overall phenotype. Unlike the more straightforward inheritance patterns of Mendelian traits, the polygenic basis of complex traits leads to less predictability and greater diversity in how these traits manifest in individuals. Most common diseases such as Type 2 diabetes (T2D) and cardiovascular disease follow such polygenic genetic patterns. Understanding the genetic component of these complex traits is crucial for predicting, preventing, and intervening of these diseases.

Polygenic score (PGS) is an estimate of an individual's genetic predisposition by aggregating the small effects of multiple variants across the genome [2]. The development of PGS begins with Genome-Wide Association Studies (GWAS), where researchers scan the genome of participants with diverse phenotype presentations (such as varying height) or different disease statuses (such as the diagnosis of T2D or not) to find genetic variants that are associated with specific traits or diseases[3]. Following GWAS, a set of genetic variants is selected with their effects inferred based on the GWAS results, which constitute a PGS model. To ensure accuracy and reliability, this model may be further validated in a dataset that is similar to the target dataset. The PGS for an individual is then calculated by summing the effects of selected genetic variants, which represents the individual's genetic predisposition to a certain trait or disease.

The ever-increasing GWAS sample sizes have significantly enhanced the predictive accuracy of PGS for a wide range of complex traits and diseases, establishing it as a promising tool in both genomic research and clinical decision-making[2, 4–12]. Numerous studies have demonstrated the versatile utility of PGS in patient stratification[13], personalized treatment[14], disease risk prediction[15], and prognosis assessment[4, 16]. This broad utility of PGS has drawn considerable attention from both academic and industrial sectors. In the healthcare industry, companies such as 23andMe, AncestryDNA, and Invitae offer direct-to-consumer genetic testing services for predicting disease predisposition to a variety of diseases, with certain conditions approved by FDA regulation[17]. Concurrently, in academia, consortia like the PRIMED Consortium[18], eMERGE Network[19], and INTERVENE project[20] are at the forefront of investigating the clinical applications of PGS, aiming to unlock their full potential in healthcare and to ensure responsible integration of PGS into clinical practice[21].

As enthusiasm for PGS surges, several concerns have emerged. First, PGS exhibits notable variability for a given individual, which has raised doubts about the reliability and reproducibility of PGS[17, 22, 23]. Many customers have expressed concerns about receiving different or even opposite results from different companies. The potential causes for the inconsistency include the use of different genetic variants and variations in the populations used for model training. However, the degree of variation and the underlying causes are not fully understood yet. It is important to quantify the individual-level PGS variation and investigate the factors contributing to this variation across a wide range of diseases/traits for the confident and responsible use of PGS in clinical application. Second, PGS exhibits significant accuracy gaps between ancestries, raising concerns over health disparities[24]. Despite the increasing PGS accuracy due to the growing GWAS sample size, this improvement disproportionately favors populations of European ancestry, which constitute over 85% of GWAS data. PGS models trained on datasets overrepresented by European ancestries are less predictive when applied to non-European populations. Such bias, if unaddressed, could potentially worsen existing health disparities if PGS becomes a standard tool in healthcare systems.

Driven by these two main challenges, my thesis focuses on evaluating the performance of PGS at the individual level among patients with diverse genetic backgrounds. First, I developed a general Bayesian framework to assess the uncertainty (variation) of PGS estimates at the individual level when PGS are trained and applied to individuals of European ancestries. By applying this framework to real data, I discovered a large uncertainty in PGS scores which leads to unreliable patient stratification and ranking in the downstream analysis. Second, I expanded the method to include scenarios where testing individuals come from a diverse genetic ancestry background which may differ from the original training population. In this enhanced method, I introduced a way to convert uncertainty to accuracy which is a more intuitive metric for PGS evaluation. I investigated the variation of PGS accuracy across a continuous genetic ancestries and found increased genetic distance corresponds to a lower PGS accuracy. This trend is significant even among so-called ”homogenous” populations like European Americans and more evident among admixed populations like Hispanic Latino Americans. These findings highlight the importance of incorporating individual-level PGS performance metrics in personalized medicine, emphasizing the need for tailored approaches in diverse populations.

The projects described above are organized into the following thesis chapters:

1. Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification.

2. Polygenic scoring accuracy varies across the genetic ancestry continuum.

CHAPTER 2

Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification

2.1 Introduction

PRSs have emerged as the main approach for predicting the genetic component of an individual's phenotype and/or common-disease risk (that is, genetic value (GV)) from large-scale genome-wide association studies (GWASs). Several studies have demonstrated the utility of PRSs as estimators of genetic values in genomic research and, when combined with nongenetic risk factors (for example, age, diet), in clinical decision-making [25–27]—for example, in stratifying patients [13], delivering personalized treatment [14], predicting disease risk [15], forecasting disease trajectories [4, 16] and studying shared etiology among traits [28, 29]. Increasingly large GWAS sample sizes have improved the predictive value of PRS for several complex traits and diseases [2, 4–12], thus paving the way for PRS-informed precision medicine.

Under a linear additive genetic model, an individual's GV is the sum of the individual's dosage genotypes at causal variants (encoded as the number of copies of the effect alleles) weighted by the causal allelic effect sizes (expected change in phenotype per copy of the effect allele). In practice, the true causal variants and their effect sizes are unknown and must be inferred from GWAS data. Existing PRS methods generally fall into one of three categories based on their inference procedure: (1) pruning/clumping and thresholding (P + T) approaches, which account for linkage disequilibrium (LD) by pruning/clumping variants at a given LD and/or significance threshold and weight the remaining variants by their marginal association statistics [3, 30]; (2) methods that account for LD through regularization of effect sizes, including lassosum [31] and BLUP prediction [32, 33]; and (3) Bayesian approaches that explicitly model causal effects and LD to infer the posterior distribution of causal effect sizes [33–36].

Both the bias and the variability of a PRS estimator are critical to assessing its practical utility. Given that most PRS methods select variants and estimate their effect sizes, there are two main sources of uncertainty: (1) uncertainty about which variants are causal (that is, have nonzero effects) and (2) statistical noise in the causal effect estimates due to the finite sample size of GWAS training data and the presence of LD between variants. The impacts of sample size and LD on causal variant identification have been thoroughly investigated in the statistical fine-mapping literature, with uncertainty increasing as the strength of LD in a region increases and the sample size of the GWAS training data decreases [37, 38]. This uncertainty about which variant is causal propagates into uncertainty in the weights used for PRS, which can lead to different estimates of genetic value in a target individual. Evaluating how this uncertainty propagates to individual PRS estimation may improve subsequent analyses such as PRS-based risk stratification.

Unfortunately, studies that have applied PRS and/or examined PRS accuracy have largely ignored uncertainty in PRS estimates at the individual level [25], focusing instead on cohort-level metrics of accuracy such as R^2. Therefore, the degree to which uncertainty in causal variant identification impacts individual PRS estimation and subsequent analyses (for example, stratification) remains unclear. In contrast, in livestock breeding programs, prediction error variance (PEV) of estimated breeding values has been used for decades to evaluate the precision of individual estimates [39–41]. PEV can be directly computed by inverting the coefficient matrix of mixed-model equations [39, 42–48]. The uncertainty in other biomarkers and nongenetic risk factors has also been well studied [49]. For example, smoothing methods and error-correction methods are performed before biomarkers and nongenetic risk factors are included in predictive models [50, 51].

Motivated by potential clinical applications of PRS in personalized medicine, we focused on evaluating uncertainty in PRS estimates at the level of a single target individual. Our goal was to quantify the statistical uncertainty in individual PRS estimates ($\widehat{\mathrm{PRS}}_i$) conditional on the data used to train the PRS. First, we extended the Bayesian framework of LDpred2 (ref. [33]) to sample from the posterior distribution of an individual's GV (GV_i) to estimate (1) the posterior s.d.($\widehat{\mathrm{PRS}}_i$) and (2) ρ-level credible intervals for the genetic value (ρGV_i-CI) for different values of ρ. Second, we introduced an analytical form for the expectation across individuals of s.d.($\widehat{\mathrm{PRS}}_i$) as a function of heritability, number of causal variants and training data sample size, and showed that the analytical form is accurate in simulations and real data. Third, we show in simulations that ρGV_i-CI is well calibrated when the target sample matches the training data and that s.d.($\widehat{\mathrm{PRS}}_i$) increases as polygenicity (number of causal variants) increases and heritability and GWAS sample size decrease [52]. Analyzing 13 real traits in the UK Biobank, we observed large uncertainties in individual PRS estimates that greatly impacted the interpretability of PRS-based ranking of individuals. For example, on average across traits, only 0.2% (s.d. = 0.6%) of individuals with PRS point estimates in the top 1% also have corresponding 95% GV_i-CI fully contained in the top 1%. Individuals with PRS point estimates at the 90th percentile in a testing sample were ranked anywhere between the 34th and 99th percentiles in the same testing sample after their 95% credible intervals (CIs) were taken into account. Finally, we explored a probabilistic approach to incorporating PRS uncertainty in PRS-based stratification and demonstrated how such approaches can enable principled risk stratification under different cost scenarios.

2.2 Results

2.2.1 Sources of uncertainty in individual PRS estimation

Under a standard linear model relating genotype to phenotype (Methods), the estimand of interest for PRS is the genetic value of an individual i, defined as $\mathrm{GV}_i = \mathbf{x}_i^\top \beta$, where $\mathbf{x}_i$ is an $M \times 1$ vector of genotypes and β is the corresponding $M \times 1$ vector of unknown causal effect sizes [53] (Methods). Different PRS methods vary in how they estimate causal effects

$\hat{\beta}$ to construct the estimator $\widehat{\text{PRS}}_i = \mathbf{x}_i^\top \hat{\beta}$. Inferential variance in $\hat{\beta}$ propagates into the variance of $\widehat{\text{PRS}}_i$. In this work, we focus on quantifying the inferential uncertainty in $\widehat{\text{PRS}}_i$ and assessing its impact on PRS-based stratification.

To illustrate the impact of statistical noise in $\hat{\beta}$ on $\widehat{\text{PRS}}_i$, consider an example of a trait for which the observed marginal GWAS effects at three SNPs are equal (Fig. 2.1). The trait was simulated assuming that SNP1 and SNP2 are causal with the same effect, whereas SNP3 is not causal but tags SNP2 with high LD (0.9). The expected marginal effect is higher at SNP2 than at SNP3, thus implying that GWAS with infinite sample size would correctly identify the true causal variants and their effects. However, finite GWAS sample sizes induce statistical noise in the observed marginal effects. For example, the marginal effect at SNP3 (tag SNP) is higher than at SNP2 (true causal SNP) in 12-30% of GWASs simulated with sample size $n = 100,000$ under the LD structure of Fig. 2.1. Thus, the key challenge is that, given only GWAS marginal effects and LD, there is more than one plausible causal effect-size configuration. In Fig. 2.1, the observed marginal effects could be driven by SNPs (1 and 2) or (1 and 3) or (1, 2 and 3); in fact, (1 and 2) and (1 and 3) are equally probable in the absence of other information. In such situations, one can generate different PRS estimates for a given individual from the same training data. For example, P + T PRS methods and lassosum, which assume sparsity, would probably select either SNPs (1 and 2) or (1 and 3), whereas BLUP or Bayesian approaches would probably take an average over the possible causal configurations, splitting the causal effect of SNP2 between SNPs (2 and 3). Thus, in such cases, an individual with the genotype $\mathbf{x}_i = (0, 1, 0)^\top$ can be classified as being above or below a prespecified threshold, depending on the approach/assumptions used to estimate causal effects.

We explored inferential uncertainty in $\widehat{\text{PRS}}_i$ through two synergistic approaches. First, we provided a closed-form approximation for the expected s.d.$(\widehat{\text{PRS}}_i)$ under simplifying assumptions. Second, we sampled from the posterior distribution of the causal effects under the framework of LDpred2 to estimate s.d.$(\widehat{\text{PRS}}_i)$ and compute credible intervals (CIs) for GV_i at prespecified confidence levels (for example, $\rho = 95\%$) (Figure 2.2). As an example of the utility of such measures of uncertainty, we explored a probabilistic approach to PRS-based risk stratification that estimates the probability that GV_i is above a given threshold t (Fig. 2.2) and demonstrated how this probability can be used in conjunction with situation-specific cost functions to optimize risk stratification decisions.

2.2.2 Analytical derivation of individual PRS uncertainty

We focus on evaluating PRS uncertainty within a general Bayesian framework, where the posterior mean of the genetic effects conditional on a given GWAS, $\hat{\beta} \equiv \mathbb{E}(\beta|\mathbf{D})$, is used to estimate the genetic value of a given individual, $\mathbf{x}_i^\top \hat{\beta} \equiv \mathbb{E}(\mathbf{x}_i^\top \beta|\mathbf{D}, \mathbf{x}_i)$ ($\mathbf{D} = (\mathbf{X}, \mathbf{y})$ with access to individual-level genotypes $\mathbf{X}$ and phenotypes $\mathbf{y}$ or $\mathbf{D} = (\hat{\beta}_{\text{GWAS}}, \hat{\mathbf{R}})$ with access to marginal association statistics and LD). We define PRS uncertainty for individual i as the posterior variance of its genetic value, $\text{var}(\mathbf{x}_i^\top \beta|\mathbf{D}, \mathbf{x}_i)$. This quantity is an approximation to the PEV of estimated breeding values in livestock genetics[41, 43], which are analogous to genetic values in human genetics.

Assuming that every SNP has a nonzero causal effect drawn i.i.d. (independent and identically distributed) from $\beta_j \sim \mathcal{N}\left(0, \frac{h_g^2}{M}\right)$, one can derive a closed-form approximation to the expectation across individuals of the posterior variance of genetic value. Given a GWAS discovery dataset of N unrelated individuals drawn from a given population, the expected PRS uncertainty for a test individual i randomly drawn from the same population is:

$$\mathbb{E}_{\mathbf{x}_i}\left[\text{var}\left(\mathbf{x}_i^\top \beta|\mathbf{D}, h_g^2\right)\right] \approx \left(\frac{1}{h_g^2} + \frac{N}{M}\right)^{-1} \tag{2.1}$$

Under an infinitesimal model, the analytical form is an approximately unbiased estimator of the expected posterior variance, even in the presence of LD (Fig. 2.3). Under noninfinitesimal models, the analytical form underestimates the expected posterior variance, albeit by a relatively small amount. Notably, across 13 phenotypes in the UK Biobank, the analytical form provides relatively accurate estimates of the empirical average of the standard deviation $\text{s.d.}(\widehat{\text{PRS}}_i)$ computed from LDpred2 posterior sampling ($R^2 = 0.79$ across traits;

Fig. 2.3). Thus, the analytical form captures the interplay among SNP heritability, sample size, and number of causal variants, and provides a useful approximation to individual PRS uncertainty when posterior samples are unavailable.

2.2.3 Factors impacting individual PRS uncertainty in simulations

Next, we quantified the degree to which different parameters contribute to uncertainty in individual PRS estimates in simulations starting from real genotypes of unrelated ẃhite British' individuals in the UK Biobank (UKBB, $N = 291,273$ individuals ($N_{\text{train}} = 250,000$, $N_{\text{validation}} = 20,000$, $N_{\text{test}} = 21,273$) and $M = 459,792$ SNPs; Methods section).

First, we assessed the calibration of the ρ-level confidence intervals (CIs) for GV_i estimated by LDpred2. We compared the empirical coverage of the ρGV_i-CIs (proportion of individuals in a single simulation replicate whose ρGV_i-CI overlaps their true GV_i) with the expected coverage (ρ) across a range of values of ρ. We find that, overall, the ρGV_i-CIs are well calibrated, albeit slightly miscalibrated in high-heritability, low-polygenicity simulations (Fig. 2.4a). For example, across ten simulation replicates where $h_g^2 = 0.25$ and $p_{\text{causal}} = 1\%$, the 90% GV_i-CIs have an average empirical coverage of 0.92 (s.e.m. $= 0.005$) (Fig. 2.4a). The ρGV_i-CIs estimated by LDpred2 are also robust to training cohort sample size. As individuals with large PRS estimates might have a larger number of effect alleles and therefore accumulate more inferential variance, we investigate whether individual PRS uncertainty varies with respect to their true genetic value, and find no significant correlation between an individual's standard deviation $\text{s.d.}(\widehat{\text{PRS}}_i)$ and their true genetic value (Fig. 2.4b).

We next assessed the impact of trait-specific genetic architecture parameters (heritability and polygenicity) on individual PRS uncertainty, defined as the posterior standard deviation (s.d.) of genetic value. First, we fixed heritability and varied polygenicity and found that the standard deviation $\text{s.d.}(\widehat{\text{PRS}}_i)$ increases from 0.10 to 0.50 when the proportion of causal variants increases from 0.1% to 100% (Fig. 2.4c). Second, we varied the heritability while keeping polygenicity constant. As different heritabilities lead to different variances explained by the PRS in the test sample, we scaled the individual s.d. ($\text{s.d.}(\widehat{\text{PRS}}_i)$) by the s.d. of

PRS point estimates across all tested individuals; we refer to this quantity as 'scaled s.d.' (Methods). We found that the scaled s.d. decreases with heritability and sample size (Fig. 2.4d). For example, when $h_g^2 = 0.05$ and $p_{\text{causal}} = 0.1\%$, a fivefold increase in training data sample size (from 50,000 to 250,000) reduces scaled s.d. by threefold (from 1.50 to 0.56); when $h_g^2 = 0.05$ and $p_{\text{causal}} = 1\%$, the same increase in training data sample size reduces the scaled s.d. by fourfold (from 1.10 to 0.39). Although the two simulation settings yield the same expected variance per causal variant under our simulation framework (that is, $h_g^2/(M \times p_{\text{causal}})$, Methods), we observe lower uncertainty across all sample sizes for $h_g^2 = 0.5$ and $p_{\text{causal}} = 1\%$, further emphasizing the impact of trait-specific genetic architecture on individual PRS uncertainty.

Next, we investigated the impact of different types of model misspecification on CI calibration and PRS uncertainty in simulations based on a set of 124,080 SNPs (the union of 36,987 UKBB array SNPs and 93,767 HapMap3 SNPs) on chromosome 2. First, we assessed the impact of imperfect tagging of causal variants by simulating phenotypes from the set of HapMap3 + UKBB SNPs ($h_g^2 = 0.02$, $p_{\text{causal}} = 0.01, 0.001$) and training the PRS on (1) 124,080 SNPs (HapMap3 + UKBB) and (2) 36,987 SNPs (UKBB only). The 'HapMap3 + UKBB' model contains all causal SNPs whereas the 'UKBB-only' model excludes approximately 70% of the causal SNPs, thus representing imperfect tagging of causal effects. As expected, the empirical coverage of the CIs is biased downward across a range of values of ρ when only the UKBB SNPs are used to train the model. This downward bias is less pronounced when polygenicity is higher (for example, $p_{\text{causal}} = 0.01$ versus 0.001) because the UKBB SNPs tag a larger proportion of heritability due to the increased causal SNP density. Second, to assess whether the coexistence of large and small causal effects impacts PRS uncertainty, we compared three simulation scenarios: (1) large effects only ($p_{\text{causal}} = 0.001$, $h_g^2 = 0.02$), (2) small effects only ($p_{\text{causal}} = 0.01$, $h_g^2 = 0.02$), and (3) a 'mixture of normal' model ($p_{\text{causal}} = 0.0055$, $h_g^2 = 0.02$ in total) composed of large effects ($p_{\text{causal}} = 0.0005$, $h_g^2 = 0.01$) and small effects ($p_{\text{causal}} = 0.005$, $h_g^2 = 0.01$). We found that the presence of a large number of small effects increases the uncertainty in individual PRS estimates. For example, the average standard deviation s.d.($\widehat{\mathrm{PRS}}_i$) among the 21,273 test individuals is 0.050, 0.087, and 0.11 for simulations (1), (2), and (3), respectively. In simulation (3), both PRS uncertainty and accuracy (squared Pearson's correlation between GV and PRS: $R^2_{\mathrm{GV}} = 0.90$, 0.51, 0.68 for (1), (2), and (3)) are approximate averages of simulations (1) and (2). Despite the LDpred2 model being misspecified in the 'mixture of normal' simulation, the GV-CIs remain well calibrated. Third, we compared PRSs obtained using external reference LD (a subsample of either 1,000 or 2,000 individuals held out from the UKBB training data) to those obtained using in-sample LD (all 250,000 individuals in the training data) and found similar degrees of PRS uncertainty and CI calibration.

2.2.4 Individual PRS Uncertainty in Real Data in the UK Biobank

We investigated individual PRS uncertainty across 13 traits in the UKBB: hair color, height, body mass index (BMI), bone mineral density in the heel (BMD), high-density lipoprotein (HDL), low-density lipoprotein (LDL), cholesterol, insulin growth factor 1 (IGF-1), creatinine, red blood cell count (RBC), white blood cell count (WBC), hypertension, and self-reported cardiovascular disease (CVD). First, we focused on PRS-based stratification. As most traits analyzed in the present study are not disease traits, we used 'above-threshold' and 'below-threshold' when referring to the results of stratification. We classified test individuals as above-threshold if their PRS point estimate (the posterior mean of their genetic value) exceeded a prespecified threshold t (that is, $\widehat{\mathrm{PRS}}_i > t$), where t is set to the 90th PRS percentile obtained from the test-group individuals. This threshold was chosen arbitrarily to provide an example of how one can compute and interpret PRS uncertainty; in practice, choosing a threshold requires careful consideration of various trait-specific factors such as prevalence and the intended clinical application. We then partitioned the above-threshold individuals into two categories: individuals whose $\rho\mathrm{GV}_i$-CIs are fully above the threshold t ('certain above-threshold') and individuals whose $\rho\mathrm{GV}_i$-CIs contain t ('uncertain above-threshold'). Similarly, we classified individuals with PRS estimates that lie below a prespecified threshold into 'certain below-threshold' and 'uncertain below-threshold' categories (Fig. 2.5a). At $t = 90th$ percentile and $\rho = 95\%$, only 0.8% (s.d. = 1.6%) of above-threshold individuals (averaged across traits) were certain above-threshold; the remaining above-threshold individuals had 95% GV_i-CIs that overlap t (Fig. 2.5b and Table 1). On the other hand, 21% (s.d. = 17.8%) of below-threshold individuals had 95% GV_i-CIs that do not overlap t. Consistent with simulations, we found that uncertainty is higher for traits that are more polygenic [54] (Table 2.1 and 2.2), with the average standard deviation s.d.$(\widehat{\text{PRS}}_i)$ ranging between 0.23 and 0.46 across the studied traits. We assessed the impact of quantile normalization of phenotypes and verified that, for mildly skewed distributions, the impact on uncertainty is small.

For completeness, we investigated the impact of the threshold t and credible level ρ on PRS-based stratification uncertainty, defined as the proportion of above-threshold individuals classified as 'certain above-threshold' for a given trait. As expected, the proportion of certain above-threshold classifications decreases as ρ increases (Fig. 2.6a). For traits with higher average uncertainty (scaled s.d.), we observed lower rates of certain classifications across all values of ρ. For example, at $t = 90\text{th}$ and $\rho = 95\%$, the proportion of above-threshold individuals classified with certainty is 0% for BMI (average scaled s.d. = 1.54) and 6.2% for hair color (average scaled s.d. = 0.62) (Fig. 2.6a). Height and HDL have similar average levels of uncertainty (average scaled s.d. of 0.95 for height and 0.96 for HDL) and similar proportions of above-threshold individuals classified with certainty (0.9% for height and 0.8% for HDL) (Fig. 2.6a and Table 2.1 and 2.2). Using a more stringent threshold t amplified the effect of uncertainty on PRS-based stratification (Fig. 2.6b). For example, for BMI and hair color, the proportion of certain classifications among above-threshold individuals dropped for all values of ρ when we increased the threshold from $t = 90\text{th}$ percentile to $t = 99\text{th}$ percentile (Fig. 2.6b).

We also quantified the impact of inferential variance in $\widehat{\text{PRS}}_i$ on PRS-based ranking of the test-group individuals. Using two random samples of genetic effects, we generated two independent rankings for all individuals in the test data and quantified the correlation in the rankings (Fig. 2.5c and Methods). We observed large variability in the rankings across the test data, with the correlation of rankings ranging from 0.25 to 0.78 across the 13 traits. We also estimated 95% CIs for the rankings of individuals at a given percentile (for example, 90th) (Table 2.3, Methods) and found high variability in the rankings. For example, in the case of HDL, an individual at the 90th (99th) percentile based on their PRS point estimate can lie within the 41st and 99th percentiles (72nd-99th) with 95% probability when the inferential variance in PRS estimation is taken into consideration (Table 2.3).

2.2.5 Integrating uncertainty into PRS-based stratification

In contrast to current PRS-based stratification practices, which compare an individual's PRS point estimate, $\widehat{\mathrm{PRS}}_i$, to a given threshold t, in the present study, we explored the use of the posterior probability that GV for individual i is above the threshold (that is, $Pr(GV_i > t)$). We estimated $Pr(GV_i > t)$ using Monte Carlo integration within the LDpred2 framework and showed in simulations that the probability is well calibrated for different causal effect-size distributions, despite slight miscalibration when polygenicity is high or when causal variants are not present in the training SNP panel(Methods).

As expected, for traits with higher PRS uncertainty, we observed a smaller proportion of testing individuals with deterministic classification ($Pr(GV_i > t) = 0$ or 1). We also found a tight correlation between $\widehat{\mathrm{PRS}}_i$ and $Pr(GV_i > t)$ across individuals in the test data. This is probably due to the relatively high polygenicity of the traits in the analysis; a lower correlation is expected for traits with lower polygenicity.

However, $Pr(GV_i > t)$ also contains information about individual-level false-positive (FP) and false-negative (FN) probabilities which, given a situation-specific cost function, can be used to calculate the expected cost of an above-threshold versus below-threshold classification (Methods). The cost functions for FP and FN should be carefully specified in the context of the clinical application; for example, in the case of bone density scans, the cost functions will depend on the actual cost of a low bone density versus risks associated with exposure to low-dose X-rays. Consider three cost functions that relate the relative costs of FP versus FN diagnoses: (1) equal cost for each FP and FN diagnosis ($C_{FP} =$

$C_{FN} = 1$); (2) 3× higher cost for FP diagnoses ($C_{FP} = 3, C_{FN} = 1$); and (3) 3× higher cost for FN diagnoses ($C_{FP} = 1, C_{FN} = 3$). For an individual with $Pr(GV_i > t) = 0.6$, the probability of a FP versus FN diagnosis is 0.4 versus 0.6, respectively. The expected costs of FP diagnoses ($Pr(FP) \times C_{FP}$) under each scenario are (1) 0.4, (2) 1.2, and (3) 0.4; the expected costs of FN diagnoses ($Pr(FN) \times C_{FN}$) are (1) 0.6, (2) 0.6, and (3) 1.8. Therefore, the classification for this individual that minimizes the expected cost under each scenario is (1) above-threshold, (2) below-threshold, and (3) above-threshold. More notably, as $Pr(GV_i > t)$ is well calibrated, we can estimate the expected cost for a population using the individual probabilities of being above-threshold. As a demonstration, in simulations, we generated the estimated cost curve on testing individuals (Methods) and found that it is very close to the true cost curve despite slight inflation (Fig. 2.6c). The estimated cost curves for the above-described cost functions achieve minimum cost at threshold = 0.5, 0.25, and 0.75, respectively, which is close to the minima of the true cost curves (0.5, 0.25, 0.7; Fig. 2.6c).

2.3 Discussion

In the present study, we demonstrated that uncertainty in PRS estimates at the individual level can have a large impact on subsequent analyses such as PRS-based stratification and can be complementary to cohort-level metrics of PRS accuracy such as R^2. We proposed a general procedure for obtaining estimates of individual PRS uncertainty that can be applied to a wide range of existing PRS methods. Among 13 traits in the UKBB, we found that even with GWAS sample sizes on the order of hundreds of thousands of individuals, there is considerable uncertainty in individual PRS estimates that can impair the reliability of PRS-based stratification. We proposed a probabilistic approach to stratification that can be used in conjunction with situation-specific cost functions to help inform PRS-based decision-making, noting that such an approach is not necessarily useful for all downstream applications of PRS. As PRS must be combined with nongenetic risk factors (for example, age, lab values) to evaluate an individual's absolute risk for a given disease, the practical utility of PRS, including measures of uncertainty in PRS, is highly dependent on disease-specific factors such as heritability, age of onset and the costs/risks that would be incurred by initiating treatment, among many others [25, 27]. We note that this work focuses on estimating genetic value rather than predicting the phenotype; uncertainty in predictions of phenotype will be larger than the results reported here by $1 - h_g^2$ due to the addition of uncertainty in nongenetic factors [55], which can be further modeled and integrated [27, 50, 56–58]. We conjecture that measures of individual PRS uncertainty will be most useful for characterizing individuals whose combined risk scores (genetics + nongenetics factors) are at or close to the decision threshold for medical intervention; we leave an investigation of uncertainty in combined risk scores for future work.

We conclude with several caveats and future directions. First, we quantified individual PRS uncertainty by extending LDpred2 (ref. [33]), which is just one of many existing Bayesian methods that can be adapted for the same purpose [36, 59, 60]. Extensions of other methods, including analogous procedures for P + T [61] and regularization-based approaches [31, 32], could also be investigated. Overall, our methods produced well-calibrated CIs in realistic simulation parameter ranges, albeit with slight miscalibration when polygenicity is low and heritability is high. We hypothesized that this is due to several approximations employed in LDpred2 for computational efficiency. We leave investigation of the impact of approximation on calibration for future work.

Second, we proposed an analytical form to estimate the expected PRS uncertainty as a function of GWAS sample size, the number of causal SNPs, and SNP heritability. Although our analytical formula did provide a good approximation, systematic biases were observed, largely due to the omission of causal configuration uncertainty induced by LD. In practice, we recommend using samples from the posterior distribution, the properties of which are validated in our simulation studies.

Third, although we found broad evidence that both trait-specific genetic architecture parameters (for example, heritability, polygenicity) and individual-specific genomic features (for example, the cumulative number of effect alleles) can impact individual PRS uncertainty, both sources of uncertainty merit further exploration. For example, we performed simulations under a model in which each causal variant explained an equal portion of total SNP heritability but, in reality, genetic architecture can vary substantially among different traits. We did not find a correlation between an individual's cumulative number of effect alleles and their individual PRS uncertainty. This was primarily due to the high polygenicity of the traits being tested. Consequently, we observed a tight correlation between $\widehat{\text{PRS}}_i$ and $\Pr(GV_i > t)$ in most simulation scenarios except those with low polygenicity. Extending these analyses to traits with a wider range of genetic architectures, for example, traits with both monogenic and polygenic disease risk factors, will be of interest [62, 63]. It is also important to investigate the relative contributions of LD and small effect sizes to PRS uncertainty under various genetic architectures. We leave methods development for PRS uncertainty decomposition for future study.

Fourth, although we showed that our approach was robust to certain types of model misspecification (for example, mixture of normal effect-size distributions, imperfect tagging of causal effects), we do not exclude the possibility of nonlinear interaction effects such as GxE, GxG, and dominance effects [64–67]. An investigation of the impact of genotype imputation on uncertainty also merits further exploration. We leave a full investigation of these questions for future work.

Last, in the present study, we did not investigate individual PRS uncertainty in transethnic or admixed population settings. Causal variants, causal effect sizes, allele frequencies, and LD patterns can vary substantially across populations [68, 69]. Moreover, PRS prediction accuracy (measured via cohort-level metrics) is well known to depend heavily on the ancestry of the individuals in the GWAS training data [22, 70]. We leave a detailed exploration of individual PRS uncertainty with respect to ancestry for future work.

2.4 Methods

2.4.1 Individual PRS uncertainty

2.4.1.1 Definition of individual PRS uncertainty

Let y_i be a trait measured on the ith individual, $\mathbf{x}_i$ an $M \times 1$ vector of standardized genotypes and β an $M \times 1$ vector of corresponding standardized effects for each genetic variant. Under a standard linear model, the phenotype model is $y_i = \mathbf{x}_i^\top \beta + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_e^2)$. The goal of the PRS methods is to predict genetic value for individual i $\mathrm{GV}_i = \mathbf{x}_i^\top \beta$ of the phenotype. In practice, the genetic effects β are unknown and need to be inferred from GWAS data as $\hat{\beta}$. Therefore, the inferential variance in $\hat{\beta}$ propagates to the estimated genetic value of individual i $\widehat{\mathrm{PRS}}_i = \mathbf{x}_i^\top \hat{\beta}$. In the present study we studied the inferential variance in $\widehat{\mathrm{PRS}}_i = \mathbf{x}_i^\top \hat{\beta}$ as a noisy estimate of $\mathrm{GV}_i = \mathbf{x}_i^\top \beta$.

2.4.1.2 Connection between PEV and posterior variance

PEV, a widely used concept in the animal breeding literature, is defined as $\mathrm{var}_{\beta,\mathbf{y}}\left[\mathbf{x}_i^\top \hat{\beta} - \mathbf{x}_i^\top \beta\right]$, where $\mathbf{x}_i$ is the genotype of individual i and $\hat{\beta} = \mathbb{E}_{\beta|\mathbf{y}}[\beta]$ is the posterior mean of the causal effects. This variance is with respect to the randomness of both the prior β and phenotype y, conditional on a fixed genotype matrix X. Furthermore, assumptions can be made on X, to incorporate the randomness in X. PRS uncertainty with X fixed, which we derive here, will be a lower bound for PRS uncertainty with random X.

It follows from the law of total variance that $\mathrm{var}_{\beta,\mathbf{y}}[\beta] = \mathbb{E}_{\mathbf{y}}\left[\mathrm{var}_{\beta|\mathbf{y}}[\beta]\right] + \mathrm{var}_{\mathbf{y}}\left[\mathbb{E}_{\beta|\mathbf{y}}[\beta]\right]$. Using the fact that $\mathrm{var}_{\beta,\mathbf{y}}\left[\hat{\beta} - \beta\right] = \mathrm{var}_{\beta,\mathbf{y}}[\beta] - \mathrm{var}_{\beta,\mathbf{y}}\left[\hat{\beta}\right]$ (from ref. [40]), we have:

$$\begin{aligned}
\mathrm{var}_{\beta,\mathbf{y}}\left[\hat{\beta} - \beta\right] &= \mathrm{var}_{\beta,\mathbf{y}}[\beta] - \mathrm{var}_{\beta,\mathbf{y}}\left[\hat{\beta}\right] \\
&= \mathbb{E}_{\mathbf{y}}\left[\mathrm{var}_{\beta|\mathbf{y}}[\beta]\right] + \mathrm{var}_{\mathbf{y}}\left[\mathbb{E}_{\beta|\mathbf{y}}[\beta]\right] - \mathrm{var}_{\beta,\mathbf{y}}\left[\hat{\beta}\right] \\
&= \mathbb{E}_{\mathbf{y}}\left[\mathrm{var}_{\beta|\mathbf{y}}[\beta]\right].
\end{aligned}$$

Finally, by multiplying a fixed genotype vector $\mathbf{x}_i$ to both sides, we have:

$$\mathrm{var}_{\beta,\mathbf{y}}\left[\mathbf{x}_i^\top \hat{\beta} - \mathbf{x}_i^\top \beta\right] = \mathbb{E}_{\mathbf{y}}\left[\mathrm{var}_{\beta|\mathbf{y}}\left[\mathbf{x}_i^\top \beta\right]\right].$$

Therefore, the posterior variance is an unbiased estimator of prediction error variance. We also noted that under the infinitesimal model setting, the posterior variance of genetic value has the same matrix form as the inversion of coefficient matrix of the mixed-model equation for BLUP ([39, 40]).

2.4.2 PRS uncertainty analytical form under infinitesimal model

To facilitate understanding of PRS uncertainty, we derived an analytical estimator of PRS uncertainty under simplified assumptions: (1) all M SNPs are independent and causal and (2) effect sizes are i.i.d. and drawn from an infinitesimal model, $\beta_j \sim N\left(0, \frac{h_g^2}{M}\right)$ for $j = 1, \ldots, M$, where h_g^2 is the total heritability and M is the number of causal variants. Without loss of generality, we assume that genotypes are standardized to have mean zero and unit variance in the population, that is $\mathbb{E}\left(x_{ij}\right) = 0$ and $\mathrm{var}\left(x_{ij}\right) = 1$, where x_{ij} is the genotype at SNP j for individual i. Under this assumption, following Appendix A in ref. [35], the least squares estimate of the GWAS marginal effect $\hat{\beta}_{\mathrm{GWAS},j}$ was approximately distributed as:

$$\hat{\beta}_{\mathrm{GWAS},j}|\beta_j \sim N\left(\beta_j, \frac{1}{N}\left(1 - \frac{h_g^2}{M}\right)\right).$$

As the per-SNP heritability in this model, $\frac{h_g^2}{M}$, is small, the variance $\frac{1}{N}\left(1 - \frac{h_g^2}{M}\right)$ can be approximated as $\frac{1}{N}$. The posterior distribution of $\beta_j|\hat{\beta}_{\mathrm{GWAS},j}$ then becomes:

$$\beta_j|\hat{\beta}_{\mathrm{GWAS},j} \sim N\left(\left(1 + \frac{M}{h_g^2 N}\right)^{-1}\hat{\beta}_{\mathrm{GWAS},j}, \frac{1}{N}\left(1 + \frac{M}{h_g^2 N}\right)^{-1}\right).$$

Therefore, the posterior variance of genetic value for an individual with the genotype $\mathbf{x}_i$ can be approximated as:

$$\mathrm{var}\left(\mathbf{x}_i^\top \beta|\mathbf{x}_i, \mathbf{X}, \mathbf{y}, h_g^2\right) \approx \sum_{j=1}^{M} x_{ij}^2 \mathrm{var}\left(\beta_j|\hat{\beta}_{\mathrm{GWAS},j}\right) = \frac{\sum_{j=1}^{M} x_{ij}^2}{N}\left(1 + \frac{M}{h_g^2 N}\right)^{-1},$$

where the approximation is based on the fact that β_j and β_k are approximately independent in the posterior distribution.

Recalling that genotype is standardized so that $\mathbb{E}\left(x_{ij}^2\right) = 1$, the expected posterior variance of genetic value in the population can be approximated by:

$$\mathbb{E}_{\mathbf{x}_i}\left(\mathrm{var}\left(\mathbf{x}_i^\top \beta | \mathbf{x}_i, \mathbf{X}, \mathbf{y}, h_g^2\right)\right) \approx \frac{M\mathbb{E}\left(x_{ij}^2\right)}{N}\left(1 + \frac{M}{h_g^2 N}\right)^{-1} = \left(\frac{1}{h_g^2} + \frac{N}{M}\right)^{-1}.$$

2.4.3 Estimating individual uncertainty in Bayesian PRS models

2.4.3.1 Overview of the framework for estimating individual uncertainty

Next, we showed how Bayesian models for estimating $\widehat{\mathrm{PRS}}_i$ can be extended to evaluate the variance of its estimate. We focused on LDpred2, a widely used method, although a similar approach could be incorporated in most Bayesian approaches. LDpred2 assumes that causal effects at SNP j are drawn from a mixture distribution with spike at 0 as follows:

$$\beta_j \sim \begin{cases} \mathcal{N}\left(0, \frac{h_g^2}{M p_{\mathrm{causal}}}\right), & \text{with probability } p_{\mathrm{causal}} \\ 0, & \text{with probability } 1 - p_{\mathrm{causal}} \end{cases}$$

where M is the total number of SNPs in the model, h_g^2 is the heritability of the trait, and p_{causal} is the proportion of causal variants in the model (that is, polygenicity). Let $\hat{\beta}_{\mathrm{GWAS}}$ and $\hat{\mathbf{R}}$ represent GWAS marginal effects and LD matrix computed from GWAS samples. By combining the prior probability $p(\beta | h_g^2, p_{\mathrm{causal}})$ and the likelihood of observed data $p(\hat{\beta}_{\mathrm{GWAS}} | \beta, \hat{\mathbf{R}})$, we can compute a posterior distribution as $p(\beta | \hat{\beta}_{\mathrm{GWAS}}, \hat{\mathbf{R}}, h_g^2, p_{\mathrm{causal}})$. The posterior distribution is intractable and therefore LDpred2 uses Markov Chain Monte Carlo (MCMC) to obtain posterior samples from $p(\beta | \hat{\beta}_{\mathrm{GWAS}}, \hat{\mathbf{R}}, h_g^2, p_{\mathrm{causal}})$. For simplicity, we used $\tilde{\beta} \sim p(\beta | \hat{\beta}_{\mathrm{GWAS}}, \hat{\mathbf{R}}, h_g^2, p_{\mathrm{causal}})$ to refer to the samples from the posterior distribution, and $p(\tilde{\beta})$ to refer to $p(\beta | \hat{\beta}_{\mathrm{GWAS}}, \hat{\mathbf{R}}, h_g^2, p_{\mathrm{causal}})$ whenever the context was clear. The posterior samples of the causal effects are summarized using the expectation $\mathbb{E}\left[\tilde{\beta}\right] = \int \tilde{\beta} p\left(\tilde{\beta}\right) d\tilde{\beta}$, leading to $\widehat{\mathrm{PRS}}_i = \mathbf{x}_i^\top \mathbb{E}\left[\tilde{\beta}\right]$.

Unlike existing methods that summarize the posterior samples of causal effects into the expectation and then estimate $\widehat{\mathrm{PRS}}_i$, we sampled from the posterior of PRS_i to construct a $\rho\mathrm{GV}_i$-CI for each individual. The Bernstein-von Mises theorem provides the basis that, under certain conditions, such constructed Bayesian CI will asymptotically be of coverage probability ρ[71]. This property of the Bayesian CI provides an intuitive explanation of the uncertainty. Concretely, we obtain B MCMC samples from the posterior distribution of causal effects $p(\tilde{\beta})$: $\tilde{\beta}^{(1)}, \tilde{\beta}^{(2)}, \ldots, \tilde{\beta}^{(B)}$. Then we computed a PRS estimate for individual i from each sample of $p(\tilde{\beta})$: $\mathbf{x}_i^\top \tilde{\beta}^{(1)}, \mathbf{x}_i^\top \tilde{\beta}^{(2)}, \ldots, \mathbf{x}_i^\top \tilde{\beta}^{(B)}$ to approximate the posterior distribution of PRS_i. From the B samples of posterior, we obtained empirical $\frac{1-\rho}{2}$ and $\frac{1+\rho}{2}$ quantiles as lower and upper bound estimates of $\rho\mathrm{GV}_i$-CI. As B goes to infinity, such MCMC estimates converge to the $\left[Q_{(1-\rho)/2}\left(\mathbf{x}_i^\top \tilde{\beta}\right), Q_{(1+\rho)/2}\left(\mathbf{x}_i^\top \tilde{\beta}\right)\right]$, where $Q_\alpha\left(\mathbf{x}_i^\top \tilde{\beta}\right)$ represents the α-quantile (here, $\alpha = (1-\rho)/2, (1+\rho)/2$) for the distribution of $p(\mathbf{x}_i^\top \tilde{\beta})$. Similarly, we summarized the posterior samples using the second moment to estimate $\mathrm{s.d.}(\widehat{\mathrm{PRS}}_i) = \mathrm{s.d.}(\mathbf{x}_i^\top \tilde{\beta})$. In practice, we used $B = 500$ because this leads to stable results. We investigated the autocorrelation statistics and found no evidence of autocorrelation at various lags in our experiment. We recommend checking autocorrelation in practice. The MCMC samplings should be thinned when there is strong evidence of autocorrelation, which will otherwise lead to underestimation of variance.

Although in the present study we focused on LDpred2, the above-described procedure is generalizable to a wide range of Bayesian methods (for example, SBayesR[36], PRS-CS[59], and AnnoPred[60]). Methods that are not based on Bayesian principle could potentially use Bootstrap to obtain individual uncertainty intervals[72].

2.4.3.2 PRS analysis using LDpred2

We ran LDpred2 for both simulation and real data analysis with the following settings. We calculated the in-sample LD with functions provided by the LDpred2 package, using the window size parameter of 3 cM. We estimated the heritability $h^2_{\mathrm{chr}_i}, i = 1, \ldots, 22$ for each chromosome with built-in constrained LD score regression[73] function. We ran LDpred2-

grid per chromosome with a grid of 17 polygenicity parameters p_{causal} from 10^{-4} to 1 equally spaced in log(space), three heritability parameters $\{0.7h^2_{\text{chr}_i}, 1.0h^2_{\text{chr}_i}, 1.4h^2_{\text{chr}_i}\}$, and with the sparsity option both enabled and disabled, as recommended by LDpred2. We chose the model with the highest R^2 between the predicted posterior mean and the (adjusted) phenotype on the validation set as the best model to apply to testing data. We extracted 500 posterior samples of causal effects $\tilde{\beta}^{(1)}, \tilde{\beta}^{(2)}, \dots, \tilde{\beta}^{(500)}$ after 100 burn-in iterations from the MCMC sampler of the model to approximate the posterior distribution of causal effects. For each individual with genotype $\mathbf{x}_i$, we calculated $\mathbf{x}_i^\top \tilde{\beta}^{(1)}, \mathbf{x}_i^\top \tilde{\beta}^{(2)}, \dots, \mathbf{x}_i^\top \tilde{\beta}^{(500)}$ to approximate the GV posterior distribution for the individual i. We then calculated summary statistics of GV posterior distribution, including the posterior mean $\widehat{\text{PRS}}_i$, ρGV_i-CI, and the probability of above-threshold t (Pr(GV$_i$ ¿ t)).

2.4.3.3 Software implementation

Our method was implemented in the LDpred2 package. In the function ‘snp_ldpred2_grid‘’, setting the option ‘return_sampling_betas = TRUE‘ will output B posterior samples of the causal genetic effects. Posterior samples of an individual's GV were obtained by multiplying the individual's genotype by the $M \times B$ weight matrix. One could subsequently obtain the posterior mean, posterior variance, and other quantities of interest from the posterior of the GV. We noted that the time required to estimate the causal effects remains the same; the only additional computational costs came from storing the $M \times B$ weight matrix and from multiplying the genotype vector by an $M \times B$ matrix rather than an $M \times B$ vector. The memory required to store 500 samples of causal effects for 459,792 SNPs is approximately 2 GB. Given the B posterior samples of causal effects, the runtime for computing the posterior distribution of genetic value for 10,000 testing individuals was less than 5 minutes.

2.4.4 Simulation experiments on PRS uncertainty under various genetic architecture

2.4.4.1 Simulation setup

We designed simulation experiments in various settings and different sample sizes to understand the properties of uncertainty in PRS estimates. We used simulation starting from genotypes in UKBB[74]. We excluded SNPs with minor allele frequency < 0.01 and genotype missingness > 0.01, and those SNPs that fail the Hardy-Weinberg test at significance threshold 10^{-7}, which left us 459,792 SNPs. We preserved 'white British individual', with self-reported British white ancestry, and filtered pairs of individuals with kinship coefficient $< \frac{1}{2^{(9/2)}}$[74]. We further filtered individuals who were outliers for genotype heterozygosity and/or missingness, and obtained 291,273 individuals for all analyses.

Given the genotype matrix X, heritability h_g^2, proportion of causal variants p_{causal}, standardized effects and phenotypes are generated as follows:

$$\beta_j \sim \begin{cases} N\left(0, \frac{h_g^2}{Mp_{\text{causal}}}\right), & c_j = 1, \text{ with probability } p_{\text{causal}} \\ 0, & c_j = 0, \text{ with probability } 1 - p_{\text{causal}} \end{cases}$$

$$(y_1, \ldots, y_N)^\top \sim N(\mathbf{X}\beta, (1 - h_g^2)\mathbf{I}_N)$$

Finally, given the phenotypes $\mathbf{y} = (y_1, \ldots, y_N)^\top$ and genotypes X, we simulated the GWAS marginal association statistics with $\hat{\beta}_{\text{GWAS}} = \frac{1}{N}X^\top\mathbf{y}$. We simulated the data using a wide range of parameters, $h_g^2 \in \{0.05, 0.1, 0.25, 0.5, 0.8\}$, $p_{\text{causal}} \in \{0.001, 0.01, 0.1, 1\}$, a total of 20 simulation settings, with each repeated 10 times. The total population of individuals is randomly assigned to 250,000 individuals as the training population, 20,000 individuals as the validating population, and the remaining 21,273 individuals as the testing population, following the usual practice for the PRS model-building process. When investigating how sample sizes in the training cohort change PRS uncertainty, we varied the sample sizes in the training population to 20,000, 50,000, 100,000, 150,000, and 250,000, while holding the validation population and testing population intact, to enable a fair comparison between sample sizes.

2.4.4.2 Calculating and evaluating the coverage

We evaluated the coverage properties of $\rho\mathrm{GV}_i$-CI in simulation: we checked whether $\mathbb{P}\left(\mathbf{x}_i^\top\beta \in \left[Q_{(1-\rho)/2}\left(\mathbf{x}_i^\top\tilde{\beta}\right), Q_{(1+\rho)/2}\left(\mathbf{x}_i^\top\tilde{\beta}\right)\right]\right) = \rho$. To evaluate this property, for each simulated dataset, we calculated the frequency of the true genetic risk lying in the predicted interval, that is, the frequency of $\mathbf{x}_i^\top\beta \in \left[Q_{(1-\rho)/2}\left(\mathbf{x}_i^\top\tilde{\beta}\right), Q_{(1+\rho)/2}\left(\mathbf{x}_i^\top\tilde{\beta}\right)\right]$ for every individual in the testing population, for $\rho \in \{0.1, 0.2, \ldots, 1.0\}$. This property provided us an intuitive understanding of the predicted interval: for an individual with a predicted interval $\left[Q_{(1-\rho)/2}\left(\mathbf{x}_i^\top\tilde{\beta}\right), Q_{(1+\rho)/2}\left(\mathbf{x}_i^\top\tilde{\beta}\right)\right]$, its true genetic risk was expected to be in this interval with a probability ρ.

2.4.4.3 Scaled s.d. in individual PRS estimates

To compare the relative order of s.d. across different genetic architecture, especially across genetic architecture with different heritability, we defined the quantity, scaled s.d., in individual PRS estimates (scaled $\mathrm{s.d.}(\widehat{\mathrm{PRS}}_i)$), to enable fair comparison. The quantity is defined for every individual i, as $\mathrm{s.d.}_{\tilde{\beta}}\left[\mathbf{x}_i^\top\tilde{\beta}\right]/\mathrm{s.d.}_{\mathbf{x}_i}\left[\mathbf{x}_i^\top\hat{\beta}\right]$, where the numerator term $\mathrm{s.d.}_{\tilde{\beta}}\left[\mathbf{x}_i^\top\tilde{\beta}\right]$ refers to the s.d. due to the posterior sampling of $\tilde{\beta}$ of the ith individual. Recalling that $\mathbf{x}_i^\top\hat{\beta} = \mathbb{E}\left[\mathbf{x}_i^\top\tilde{\beta}\right]$, the denominator term $\mathrm{s.d.}_{\mathbf{x}_i}\left[\mathbf{x}_i^\top\hat{\beta}\right]$ refers to the variation of the point estimate across individuals in the population.

2.4.5 Real data analysis

We performed real data analysis with 13 real traits from UKBB, including hair color, height, BMI, BMD, HDL, LDL, cholesterol, IGF-1, creatinine, RBC and WBC, hypertension, and CVD. The genotype was processed in the same way as the simulation study, where we have 459,792 SNPs and 291,273 individuals. We randomly partitioned the total of 291,273 individuals into 250,000 training, 20,000 validation, and 21,273 testing groups. Training samples were used to estimate PRS weights; validation samples were used to estimate hyperparameters (for example, heritability and polygenicity) for LDpred2, and testing samples were used to evaluate accuracy and uncertainty. The random partition was repeated five times to average the randomness of results due to sample partition. For each round of random partition of the individuals, we calculated marginal association statistics between genotype and quantile-normalized phenotype in the training group with PLINK, using age, sex, and the first 20 genetic principal components as the covariates. Then we applied LDpred2 to obtain the individual posterior distribution of the genetic value, as described above. We regressed out covariates from the phenotypes to obtain adjusted phenotypes, where the regressing coefficients were first estimated from the training population, and then applied to the phenotype from training, validation, and testing populations, respectively. We evaluated the accuracy of PRS estimates in validation and testing groups using Pearson's correlation between PRS estimates and adjusted phenotypes.

2.4.6 Posterior individual ranking interval

The relative rank of individual PRS $\mathbf{x}_i^\top \tilde{\beta}^{(\mathrm{b})}$ in the population $\mathbf{x}_j^\top \tilde{\beta}^{(\mathrm{b})}, j = 1, \ldots, N$ varied across different MCMC samplings of posterior causal effects. To evaluate the uncertainty of ranking for individual i, we computed $r_i^{(b)}$ as the ranking of $\mathbf{x}_i^\top \tilde{\beta}^{(b)}$ in the population $\mathbf{x}_j^\top \tilde{\beta}^{(b)}, j = 1, \ldots, N$ for each of the $b = 1, \ldots, B$ posterior samples to approximate posterior distribution of the relative rank. We could obtain ρ-level CIs of ranking as $\left[Q_{(1-\rho)/2}\left(r_i\right), Q_{(1+\rho)/2}\left(r_i\right)\right]$ for each individual i. To assess the uncertainty of ranking for individuals at the 90th (99th) percentile threshold based on PRS estimates, we selected individuals within 1 percentile of thresholds (89.5-90.5%, 98.5-99.5%) and computed mean and s.d. for lower and upper bounds of $\rho = 95\%$ posterior ranking interval, across the selected individuals.

With the B posterior causal effect samples $\tilde{\beta}^{(1)}, \tilde{\beta}^{(2)}, \ldots, \tilde{\beta}^{(B)}$ after burn-in and N individuals in the testing population $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, we computed PRS for each individual, $\mathbf{x}_1^\top \tilde{\beta}^{(b)}, \ldots, \mathbf{x}_N^\top \tilde{\beta}^{(b)}$ and its relative rank in the population $r_1^{(b)}, \ldots, r_N^{(b)}$ for each posterior sample $\tilde{\beta}^{(b)}$. Then, for each pair of different b_1th and b_2th posterior samples, $\tilde{\beta}^{(b_1)}, \tilde{\beta}^{(b_2)}$, we calculated Spearman's correlation between $r_1^{(b_1)}, \ldots, r_N^{(b_1)}$ and $r_1^{(b_2)}, \ldots, r_N^{(b_2)}$, representing the variability of the ranks across MCMC samplings. We computed the rank correlation for 1,000 pairs of different MCMC samplings, and got the distribution of the rank correlation.

2.4.7 Probabilistic risk stratification

We defined the notion of probabilistic framework for risk stratification based on the posterior distribution of GV_i. Given a prespecified threshold t, for every individual, we could calculate the posterior probability of the genetic risk larger than the given threshold t, $\Pr(\mathrm{GV}_i > t)$, with MCMC integration as:

$$\Pr(\mathrm{GV}_i > t) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\mathbf{x}_i^\top \tilde{\beta}^{(b)} > \mathrm{t}).$$

We used the previous simulation settings to show that this probability is well calibrated. For each simulation, we divided the individuals based on their posterior probability of being at above-threshold into ten bins with $\{0, 0.1, \ldots, 1.0\}$ as breaks. For each bin, we calculated the proportion of individuals with true genetic risk higher than the threshold as the empirical probability, and the average posterior probability as theoretical probability. The empirical probability was expected to be the same as theoretical probability.

The individualized posterior distribution of genetic value provided extra information for patient stratification. We considered a scenario where there is a cost associated with the decision that classifies (1) an individual with low genetic risk into a high genetic risk category, C_{FP}, and (2) an individual with high genetic risk into a low genetic risk category, C_{FN}. For an individual with posterior probability $\Pr(\mathrm{GV}_i > t)$, we wanted to decide an action, whether to classify this individual to be at high genetic risk, and perform further screening. If we classified this individual as above-threshold, we would have probability $1 - \Pr(\mathrm{GV}_i > t)$ that this individual was in fact below-threshold, inducing an expected cost $C_{\mathrm{FP}}(1 - \Pr(\mathrm{GV}_i > t))$. Conversely, if we classified this individual as below-threshold, we would have probability $\Pr(\mathrm{GV}_i > t)$ that this individual will be in the high genetic risk, inducing an expected cost $C_{\mathrm{FN}} \Pr(\mathrm{GV}_i > t)$. To minimize the expected cost, we would decide according to which action leads to the least cost. The critical value in this scenario was $\frac{C_{\mathrm{FN}}}{C_{\mathrm{FP}}+C_{\mathrm{FN}}}$: if $\Pr(\mathrm{GV}_i > t) > \frac{C_{\mathrm{FN}}}{C_{\mathrm{FP}}+C_{\mathrm{FN}}}$, we would choose to classify this individual as above-threshold, otherwise below-threshold. For Fig. 6c, given the cost parameters C_{FP} and C_{FN}, and a threshold t, for every decision threshold, we calculated the estimated cost by summing up $C_{\mathrm{FP}}(1 - \Pr(\mathrm{GV}_i > t))$ for those individuals classified as high genetic risk category, and $C_{\mathrm{FN}} \Pr(\mathrm{GV}_i > t)$ for those individuals classified as low genetic risk category in the testing data. Correspondingly, for every decision threshold, we also calculated the true cost based on the ground truth of genetic values in the simulation.

2.5 Figures

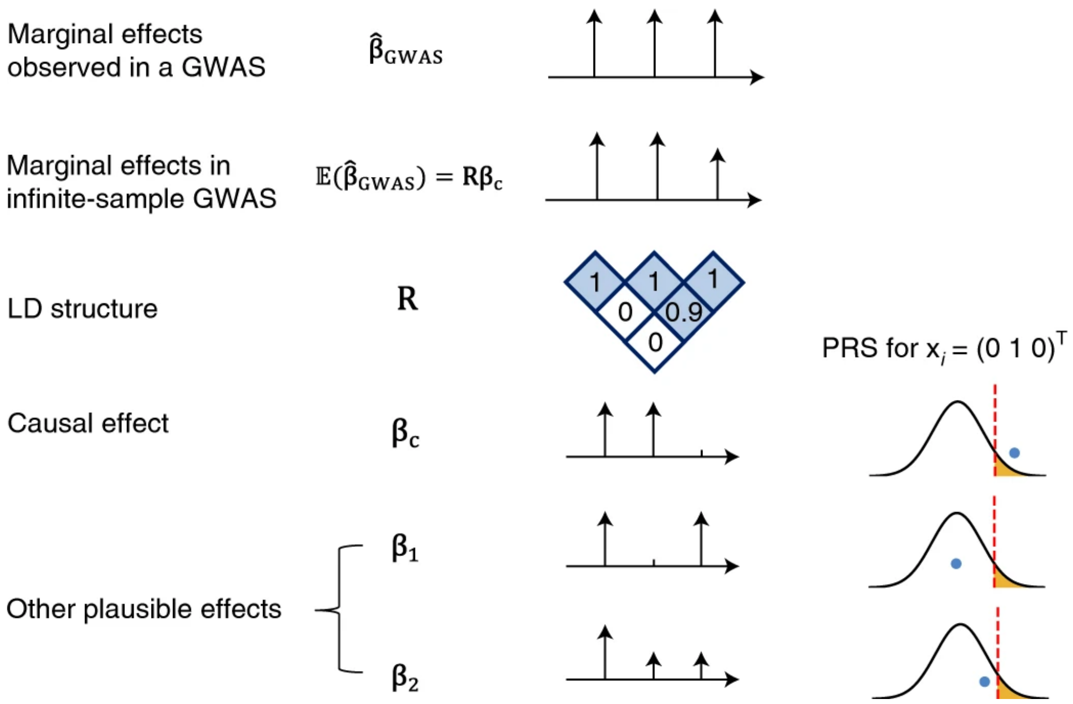
Marginal effects observed in a GWAS
$\hat{\beta}_{GWAS}$
Marginal effects in infinite-sample GWAS
$\mathbb{E}(\hat{\beta}_{GWAS}) = R\beta_c$
LD structure
R
1
1
1
0
0.9
0
PRS for $x_i = (0\ 1\ 0)^T$
Causal effect
β_c
Other plausible effects
β_1
β_2

Figure 2.1: **LD and finite GWAS sample size introduce uncertainty into PRS estimation.** We simulated a GWAS of N individuals across 3 SNPs with LD structure $\mathbf{R}$ (SNP2 and SNP3 are in LD of 0.9 whereas SNP1 is uncorrelated with other SNPs), where SNP1 and SNP2 are causal with the same effect size $\beta_c = (0.016, 0.016, 0)$ such that the variance explained by this region is $\mathrm{var}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_c) = 0.5/1000$, corresponding to a trait with total heritability of 0.5 uniformly distributed across 1,000 causal regions. The marginal effects observed in the GWAS, $\hat{\beta}_{\mathrm{GWAS}}$, have an expectation of $\mathbf{R}\beta_c$ and variance-covariance $(\sigma_{\mathrm{e}}^2/N)\mathbf{R}$, thus showcasing the statistical noise introduced by finite sample size of GWAS (N). For example, the probability of the marginal GWAS effect at tag SNP3 exceeding the marginal effect of true causal SNP2, although it decreases with N, remains considerably high for realistic sample and effect sizes (12% at $N = 100,000$ for a trait with $h_g^2 = 0.5$ split across 1,000 causal regions). Given such an observation, in addition to the true causal effects β_c, other causal configurations are probable: $\beta_1 = (0.016, 0, 0.016)$ or $\beta_2 = (0.016, 0.008, 0.008)$. An individual with genotype $\mathbf{x}_i = (0, 1, 0)^\top$ will attain different PRS estimates under these different causal configurations. Most importantly, in the absence of other prior information, β_1 and β_c are equally probable given the data, thus leading to different PRS estimates for individual $\mathbf{x}_i = (0, 1, 0)^\top$.

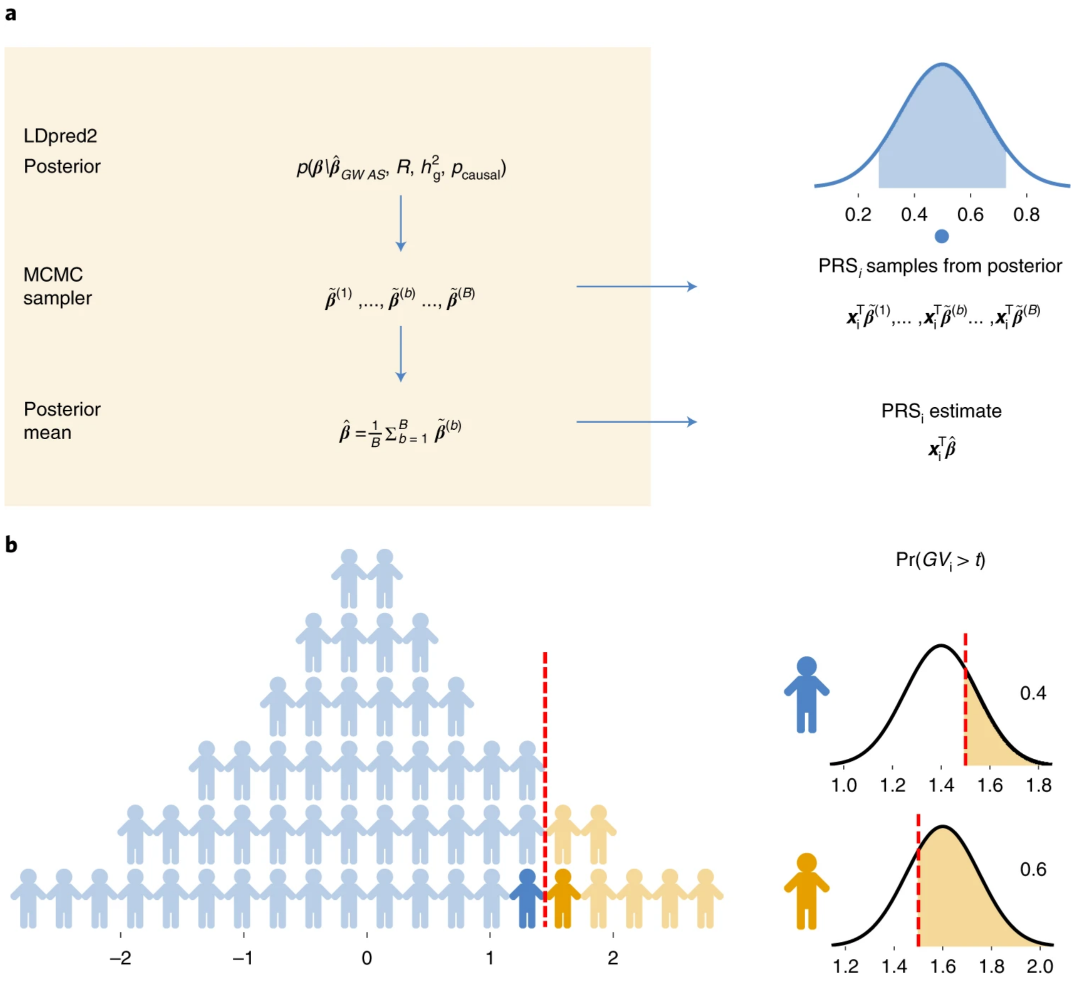
a
LDpred2
Posterior
$p(\beta \backslash \hat{\beta}_{GWAS}, R, h_g^2, p_{causal})$
MCMC
sampler
$\tilde{\beta}^{(1)}, \ldots, \tilde{\beta}^{(b)} \ldots, \tilde{\beta}^{(B)}$
Posterior
mean
$\hat{\beta} = \frac{1}{B}\Sigma_{b=1}^{B} \tilde{\beta}^{(b)}$
0.2
0.4
0.6
0.8
PRS_i samples from posterior
$x_i^T\tilde{\beta}^{(1)}, \ldots, x_i^T\tilde{\beta}^{(b)} \ldots, x_i^T\tilde{\beta}^{(B)}$
PRS_i estimate
$x_i^T\hat{\beta}$
b
−2
−1
0
1
2
$\Pr(GV_i > t)$
0.4
1.0
1.2
1.4
1.6
1.8
0.6
1.2
1.4
1.6
1.8
2.0

Figure 2.2: **Framework for Extracting Uncertainty from Bayesian Methods for Probabilistic PRS-Based Stratification. (a)** Procedure to obtain uncertainty from LDpred2. LDpred2 uses MCMC to sample from the posterior causal effect distribution given GWAS marginal effects and LD. It outputs the posterior mean of the causal effects for estimating the posterior mean genetic value (the PRS point estimate). The density plot represents the posterior distribution of GV for an individual. The shaded area represents a ρ-level CI. The dot represents the posterior mean. **(b)** Probabilistic PRS-based stratification framework. Given a threshold t, probabilistic PRS-based stratification assigns each individual a probability of being above-threshold $\Pr(\text{GV}_i > t)$.

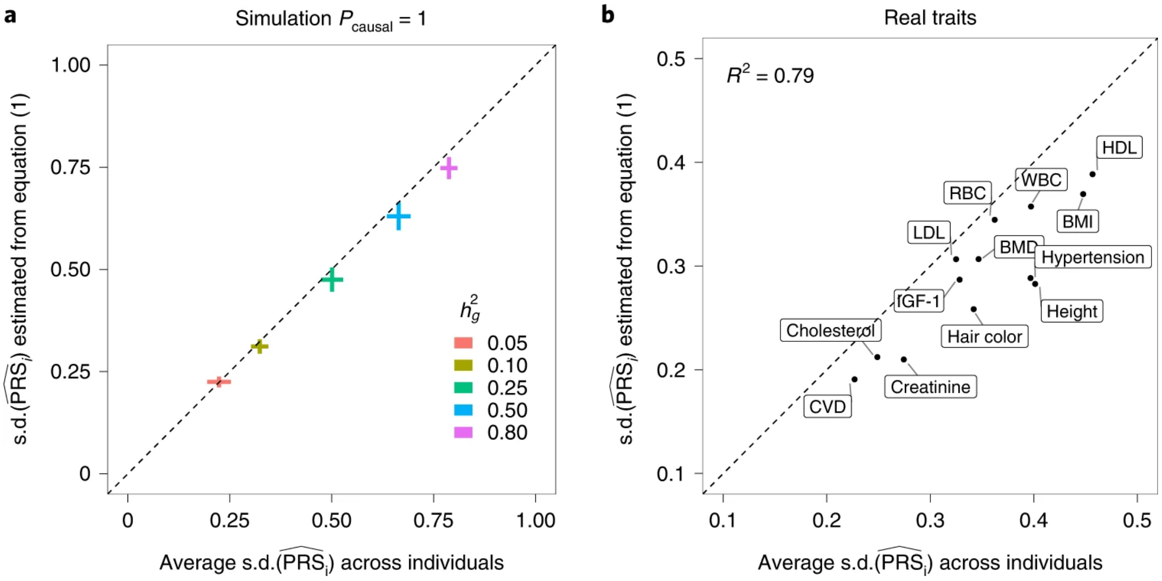

Figure 2.3: **Expected standard deviation (s.d.($\widehat{\textbf{PRS}}_i$)) estimated as a function of heritability, polygenicity, and training GWAS sample size is highly correlated with average standard deviation (s.d.($\widehat{\textbf{PRS}}_i$)) across testing individuals.** **(a)** The analytical form provides approximately unbiased estimates of expected s.d.($\widehat{\text{PRS}}_i$) in simulations when $p_{\text{causal}} = 1$. The x axis is the average s.d.($\widehat{\text{PRS}}_i$) in testing individuals. The y axis is the expected s.d.($\widehat{\text{PRS}}_i$) computed from equation (1). Each dot is an average of ten simulation replicates for each $h_g^2 \in \{0.05, 0.1, 0.25, 0.5, 0.8\}$. The horizontal whiskers represent ±1.96 standard deviation of average s.d.($\widehat{\text{PRS}}_i$) across ten simulation replicates. The vertical whiskers represent ±1.96 standard deviation of expected s.d.($\widehat{\text{PRS}}_i$) across ten simulation replicates. **(b)** The analytical estimator of expected s.d.($\widehat{\text{PRS}}_i$) is highly correlated with estimates obtained via posterior sampling for real traits. The x axis is the average s.d.($\widehat{\text{PRS}}_i$) in testing individuals. The y axis is the expected s.d.($\widehat{\text{PRS}}_i$) computed from equation (1), where M is replaced with the estimated number of causal variants and heritability is replaced with estimated SNP heritability.

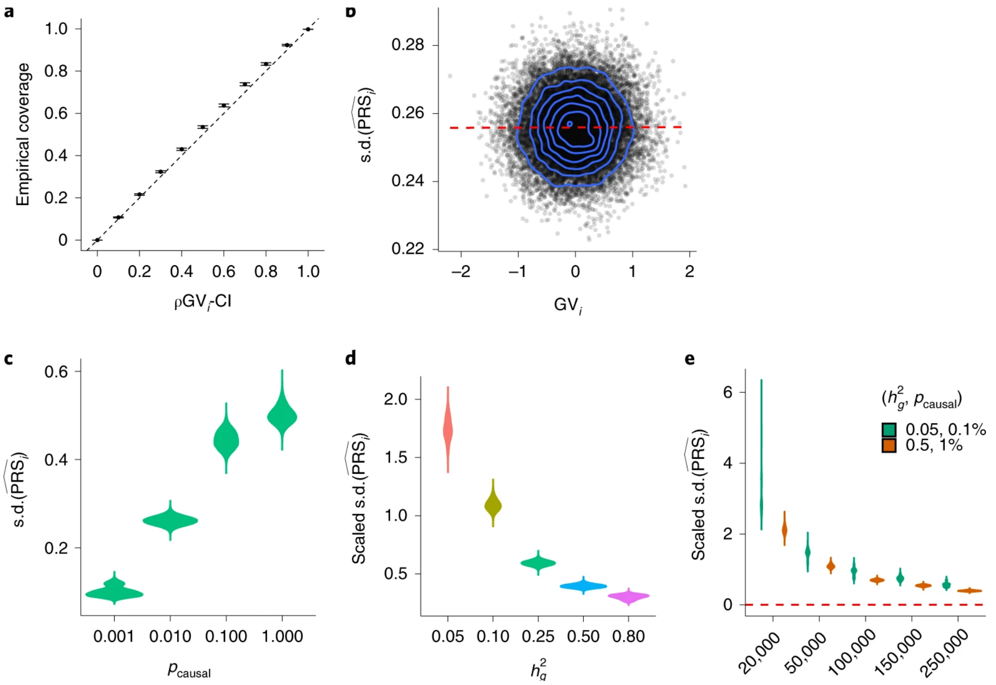

Figure 2.4: **Genetic architecture (polygenicity, p_{causal}; SNP heritability, h_g^2) and GWAS sample size impact uncertainty in PRS estimates in simulations. (a)** Individual confidence intervals (CIs) are well calibrated ($h_g^2 = 0.25$, $p_{\text{causal}} = 1\%$). Empirical coverage is calculated as the proportion of individuals in a single simulation whose ρ-level CIs contain their true genetic risk. The dots and error bars represent the mean ±1.96 standard error of the mean (s.e.m.) of the empirical coverage calculated from ten simulations. **(b)** Correlation between uncertainty and true genetic value ($h_g^2 = 0.25$, $p_{\text{causal}} = 1\%$). Each dot represents an individual. The x axis is the true genetic value; the y axis is the standard deviation of the individual PRS estimate. **(c)** Distribution of individual PRS uncertainty estimates with respect to polygenicity. **(d)** Distribution of individual PRS uncertainty estimates with respect to heritability. Each violin plot represents scaled standard deviation for 21,273 testing individuals across ten simulation replicates. **(e)** Distribution of individual uncertainty estimates with respect to training GWAS sample size. Each violin plot represents scaled standard deviation of individual PRS for 21,273 testing individuals across ten simulation replicates.

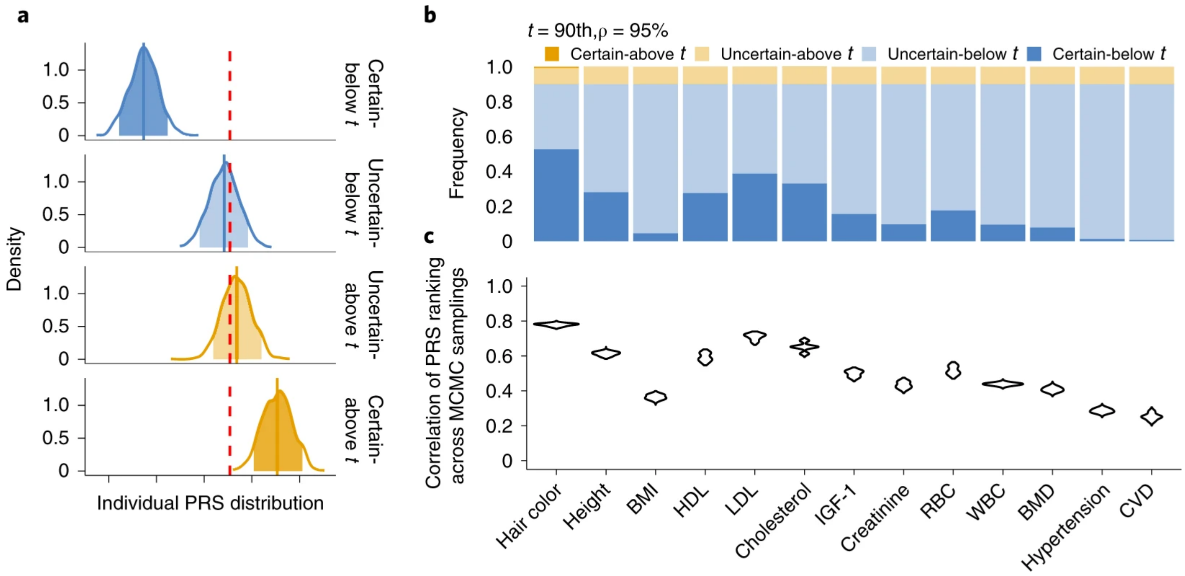

Figure 2.5: **Uncertainty in real data and its influence on PRS-based stratification.** **(a)** Example of posterior PRS distributions for individuals with certain below-threshold (dark blue), uncertain below-threshold (light blue), uncertain above-threshold (light yellow) and certain above-threshold (dark yellow) classifications for HDL. Each density plot is a smoothed posterior PRS distribution of an individual randomly chosen from that category. The solid vertical lines are posterior means. The shaded areas are 95% CIs. The red dotted line is the classification threshold. **(b)** Distribution of classification categories across 13 traits (t = 90%, $\rho = 95\%$). Each bar plot represents the frequency of testing individuals who fall into each of the four classification categories for one trait. The frequency is averaged across five random partitions of the whole dataset. **(c)** Correlation of PRS rankings of test individuals obtained from two MCMC samplings from the posterior of the causal effects. For each trait, we drew two samples from the posterior of the causal effects, ranked all individuals in the test data twice based on their PRS from each sample, and computed the correlation between the two rankings across individuals. Each violin plot contains 5,000 points (1,000 pairs of MCMC samples and five random partitions).

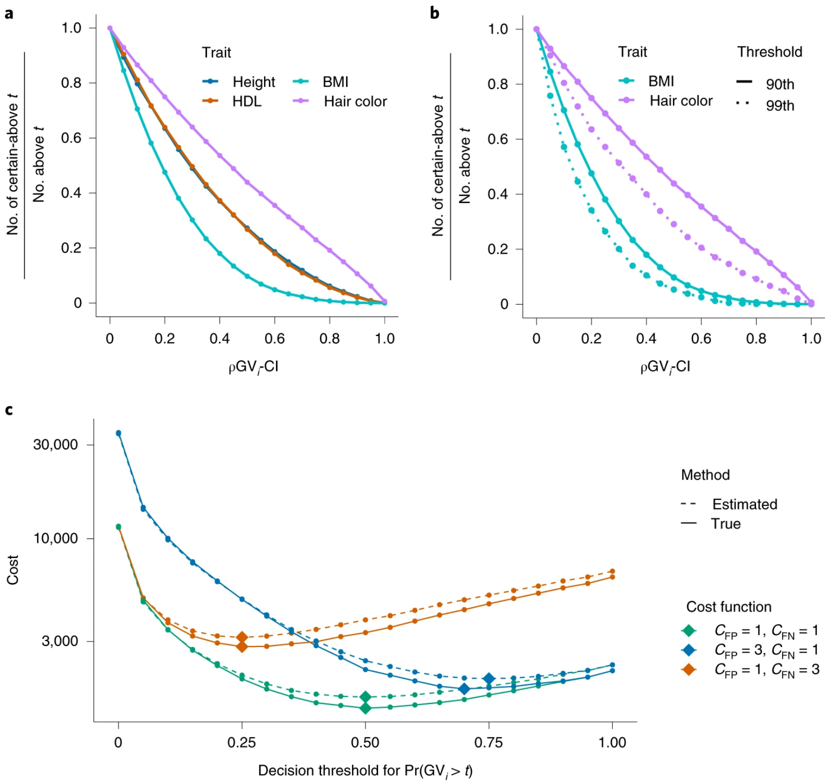

Figure 2.6: **Stratification uncertainty at different thresholds t and credible set levels ρ.** **(a)** Proportion of above-threshold classifications that are 'certain' for four representative traits. The x axis shows ρ varying from 0 to 1 in increments of 0.05. The stratification threshold t is fixed at 90%. **(b)**Proportion of above-threshold classifications that are 'certain' for two representative traits and two stratification thresholds (t = 90th and t = 99th percentiles). **(c)**Flexible cost optimization with probabilistic individual stratification under various cost functions. Each color corresponds to one cost function: equal cost for each FP and FN diagnosis ($C_{FP} = C_{FN} = 1$, green); 3 × higher cost for FP diagnoses ($C_{FP} = 3$, $C_{FN} = 1$, blue); and 3 × higher cost for FN diagnoses ($C_{FP} = 1$, $C_{FN} = 3$, orange). The probability threshold for classification varies along the x axis. Solid lines represent cost calculated using true genetic risk, and dotted lines represent cost estimated from the probability of an individual being above-threshold. Diamond symbols represent the optimal classification threshold for each curve (the minima). Simulation parameters are fixed to $h_g^2 = 0.25$, $p_{\text{causal}} = 1\%$.

2.6 Tables

Table 2.1: PRS-based individual stratification uncertainty across 13 complex traits in UKBB at 90th percentile stratification threshold

Trait	**PRS $<t$ ('below-threshold')**		**PRS $>t$ ('above-threshold')**	
	No. of certain (s.d.)	**% of certain (s.d.)**	**No. of certain (s.d.)**	**% of certain (s.d.)**
Hair color	18,398.6 (208.4)	87.4 (1.0)	4.4 (1.5)	2.1 (0.7)
Height	14,442.6 (147.6)	68.6 (0.7)	0.6 (0.9)	0.3 (0.4)
BMI	5,254.4 (739.1)	24.9 (3.5)	0.2 (0.4)	0.1 (0.2)
HDL	14,167.6 (691.4)	67.3 (3.3)	0.2 (0.4)	0.1 (0.2)
LDL	15,615.8 (448.1)	74.1 (2.1)	0.6 (0.5)	0.3 (0.3)
Cholesterol	14,793.2 (668.3)	70.2 (3.2)	0.2 (0.4)	0.1 (0.2)
IGF-1	11,049.2 (597.9)	52.5 (2.8)	0.2 (0.4)	0.1 (0.2)
Creatinine	8,337.2 (702.7)	39.6 (3.3)	0 (0)	0 (0)
RBC	1,1532.8 (1,056.9)	54.8 (5.0)	0 (0)	0 (0)
WBC	8,496.6 (370.7)	40.3 (1.8)	0 (0)	0 (0)
BMD	7,816.0 (511.1)	37.1 (2.4)	0.0 (0)	0.0 (0)
Hypertension	2,378.8 (390.7)	11.3 (1.9)	0 (0)	0 (0)
CVD	1,506.6 (512.3)	7.2 (2.4)	0 (0)	0 (0)
Average	**1,0291.5 (5,220.4)**	**48.9 (24.8)**	**0.49 (1.2)**	**0.2 (0.6)**

Table 2.2: PRS-based individual stratification uncertainty across 13 complex traits in UKBB at 99th percentile stratification threshold

Trait	**PRS $< t$ ('below-threshold')**		**PRS $> t$ ('above-threshold')**	
	No. of certain (s.d.)	**% of certain (s.d.)**	**No. of certain (s.d.)**	**% of certain (s.d.)**
Hair color	18,398.6 (208.4)	87.4 (1.0)	4.4 (1.5)	2.1 (0.7)
Height	14,442.6 (147.6)	68.6 (0.7)	0.6 (0.9)	0.3 (0.4)
BMI	5,254.4 (739.1)	24.9 (3.5)	0.2 (0.4)	0.1 (0.2)
HDL	14,167.6 (691.4)	67.3 (3.3)	0.2 (0.4)	0.1 (0.2)
LDL	15,615.8 (448.1)	74.1 (2.1)	0.6 (0.5)	0.3 (0.3)
Cholesterol	14,793.2 (668.3)	70.2 (3.2)	0.2 (0.4)	0.1 (0.2)
IGF-1	11,049.2 (597.9)	52.5 (2.8)	0.2 (0.4)	0.1 (0.2)
Creatinine	8,337.2 (702.7)	39.6 (3.3)	0 (0)	0 (0)
RBC	1,1532.8 (1,056.9)	54.8 (5.0)	0 (0)	0 (0)
WBC	8,496.6 (370.7)	40.3 (1.8)	0 (0)	0 (0)
BMD	7,816.0 (511.1)	37.1 (2.4)	0.0 (0)	0.0 (0)
Hypertension	2,378.8 (390.7)	11.3 (1.9)	0 (0)	0 (0)
CVD	1,506.6 (512.3)	7.2 (2.4)	0 (0)	0 (0)
Average	**1,0291.5 (5,220.4)**	**48.9 (24.8)**	**0.49 (1.2)**	**0.2 (0.6)**

Table 2.3: Average 95% posterior ranking CIs for individuals at two stratification thresholds for 13 traits

Trait	t = 90th percentile		t = 99th percentile	
	Lower bound (s.d.)	Upper bound (s.d.)	Lower bound (s.d.)	Upper bound (s.d.)
Hair color	57.9 (1.8)	97.9 (0.22)	88.0 (2.2)	99.8 (0.05)
Height	43.4 (2.1)	98.6 (0.18)	74.9 (3.4)	99.9 (0.04)
BMI	22.9 (2.1)	99.0 (0.17)	45.8 (4.0)	99.8 (0.04)
HDL	41.3 (2.8)	98.7 (0.18)	72.3 (4.1)	99.9 (0.04)
LDL	49.1 (2.4)	98.6 (0.19)	77.7 (3.5)	99.9 (0.04)
Cholesterol	45.1 (2.8)	98.6 (0.19)	74.9 (3.8)	99.9 (0.04)
IGF-1	33.2 (2.4)	98.8 (0.17)	63.0 (4.1)	99.9 (0.04)
Creatinine	28.0 (2.4)	98.9 (0.17)	54.7 (4.3)	99.9 (0.04)
RBC	34.5 (2.7)	98.8 (0.17)	64.4 (4.5)	99.9 (0.04)
WBC	28.2 (2.0)	98.9 (0.17)	56.0 (3.9)	99.9 (0.04)
BMD	26.0 (2.2)	98.9 (0.18)	52.5 (4.1)	99.9 (0.04)
Hypertension	17.7 (1.8)	99.0 (0.17)	36.6 (3.4)	99.8 (0.05)
CVD	15.5 (1.9)	99.0 (0.18)	32.3 (3.8)	99.8 (0.06)
Average	34.2 (12.9)	98.8 (0.03)	61.0 (16.6)	99.9 (0)

Note: We estimated the 95% posterior ranking CIs for individuals at the 90th and 99th percentiles of the testing population PRS estimates. Mean and s.d. were calculated from the 95% posterior ranking intervals of individuals whose point estimates lay within 0.5% of the stratification threshold (213 individuals between the 89.5th and 90.5th percentiles for t = 90th percentile and between the 98.5th and 99.5th percentiles for t = 99th percentile).

CHAPTER 3

Polygenic scoring accuracy varies across the genetic ancestry continuum

3.1 Introduction

Polygenic scores (PGSs) have limited portability across different groupings of individuals (for example, by genetic ancestries and/or social determinants of health), preventing their equitable use[70, 75, 76]. PGS portability has typically been assessed using a single aggregate population-level statistic (for example, R^2)[2], ignoring inter-individual variation within the population. Here, using a large and diverse Los Angeles biobank (ATLAS, $n = 36,778$)[77] along with the UK Biobank[74] (UKBB, $n = 487,409$), we show that PGS accuracy decreases individual-to-individual along the continuum of genetic ancestries[78] in all considered populations, even within traditionally labelled 'homogeneous' genetic ancestries. The decreasing trend is well captured by a continuous measure of genetic distance (GD) from the PGS training data: Pearson correlation of -0.95 between GD and PGS accuracy averaged across 84 traits. When applying PGS models trained on individuals labelled as white British in the UKBB to individuals with European ancestries in ATLAS, individuals in the furthest GD decile have 14% lower accuracy relative to the closest decile; notably, the closest GD decile of individuals with Hispanic Latino American ancestries show similar PGS performance to the furthest GD decile of individuals with European ancestries. GD is significantly correlated with PGS estimates themselves for 82 of 84 traits, further emphasizing the importance of incorporating the continuum of genetic ancestries in PGS interpretation. Our results highlight the need to move away from discrete genetic ancestry clusters towards the continuum of genetic ancestries when considering PGSs. Using two large biobank datasets, a study shows that the accuracy of polygenic scores decreases as a function of relatedness at the individual level when modeling genetic ancestry as a continuum.

PGSs—estimates of an individual's genetic predisposition for complex traits and diseases (that is, genetic liability; also referred to as genetic value)—have garnered tremendous attention recently across a wide range of fields, from personalized genomic medicine[2, 25, 26, 79] to disease risk prediction and prevention[4, 14, 15, 80] to socio-genomics[21, 76]. However, the variation in PGS performance across different genetic ancestries and/or socio-demographic features (for example, sex, age, and social determinants of health)[75] poses a critical equity barrier that has prevented widespread adoption of PGSs. Similar portability issues have also been reported for non-genetic clinical models[81–83]. The interpretation and application of PGSs are further complicated by the conflation of genetic ancestries with social constructs such as nationality, race, and/or ethnicity. Here we investigate PGS performance across genetically inferred ancestry (GIA), which describes the genetic similarity of an individual to a reference dataset (for example, 1000 Genomes[84]) as inferred by methods such as principal component analysis (PCA); GIAs do not represent the full genetic diversity of human populations.

Genetic prediction and its accuracy (or reliability) have been extensively studied in agricultural settings with a focus on breeding programs[85–88]. At the population level, PGS accuracy can be expressed as a function of heritability, training sample size, and the number of markers used in the predictor in single[52, 89, 90] or multi-population settings with or without effect size heterogeneity[91]. At the individual level, accuracy of genetic prediction from pedigree data[39, 92, 93] can be derived as a function of the inverse of the coefficient matrix of mixed-models equations, whereas accuracy of genetic prediction using whole-genome genetic data can be derived similarly, with the pedigree matrix replaced with the genomic relationships matrix[86–88, 91, 94, 95] among training and testing individuals. Simulations guided by dairy breeding programs showcase that genomic prediction accuracy varies with genetic relatedness of the testing individual to the training data[96, 97] as well as across generations, owing to the decay of genetic relationships[98].

In humans, PGS performance evaluation has traditionally relied on population-level accuracy metrics (for example, R^2) [2, 75]. PGS accuracy decays as the target populations become more dissimilar from the training data using either relatedness [99, 100] or continental or subcontinental ancestry groupings [22, 70, 101, 102]; the decay may be explained by differences in linkage disequilibrium, minor allele frequencies and/or heterogeneity in genetic effects due to gene–gene and gene–environment interactions [69]. However, population-level metrics of accuracy provide only an aggregate (average) metric for all individuals in the population, thus implicitly assuming some level of homogeneity across individuals [2, 75, 103]. Homogeneous populations are an idealized concept that only roughly approximate human data; human diversity exists along a genetic ancestry continuum without clearly defined clusters and with various correlations between genetic and socio-environmental factors [68, 78, 103–106]. Grouping individuals into discrete GIA clusters obscures the impact of individual variation on PGS accuracy. This is evident among individuals with recently admixed genomes for which genetic ancestries vary individual-to-individual and locus-to-locus in the genome. For example, a single population-level PGS accuracy estimated across all African Americans overestimates PGS accuracy for African Americans with large proportions of African GIA [102]; likewise, coronary artery disease PGS performs poorly in Hispanic individuals with high proportions of African GIA [107]. The genetic ancestry continuum affects PGS accuracy even in traditionally labelled 'homogeneous' or 'non-admixed' populations. For example, PGS accuracy decays across a gradient of subcontinental ancestries within Europe as the target cohorts become more genetically dissimilar from the PGS training data [101, 106]. Assessing PGS accuracy using population-level metrics is further complicated by technical issues in assigning individuals to discrete clusters of GIA. Different algorithms and/or reference panels may assign the same individual to different clusters [101, 103, 108], leading to different PGS accuracies. Moreover, many individuals are not assigned to any cluster owing to limited reference panels used for genetic ancestry inference [77, 101], leaving such individuals outside PGS characterization. This poses equity concerns as it limits PGS applications only to individuals within well-defined GIAs.

Here we leverage classical theory [39, 92, 93] and methods that characterize PGS performance at the level of a single target individual [109] to evaluate the impact of the genetic ancestry continuum on PGS accuracy. We use simulations and real-data analyses to show that PGS accuracy decays continuously individual-to-individual across the genetic continuum as a function of GD from the PGS training data; GD is defined as a PCA projection of the target individual on the training data used to estimate the PGS weights. We leverage a large and diverse Los Angeles biobank at the University of California, Los Angeles [77] (ATLAS, $n = 36,778$) along with the UK Biobank [74] (UKBB, $n = 487,409$) to investigate the interplay between genetic ancestries and PGS for 84 complex traits and diseases. The accuracy of PGS models trained on individuals labelled as white British (WB; see Methods for naming convention used in this work) in the UKBB ($n = 371,018$) is negatively correlated with GD for all considered traits (average Pearson $R = -0.95$ across 84 traits), demonstrating pervasive individual variation in PGS accuracy. The negative correlation remains significant even when restricted to traditionally defined GIA clusters (ranging from $R = -0.43$ for East Asian GIA to $R = -0.85$ for the African American GIA in ATLAS). On average across the 84 traits, when rank-ordering individuals according to distance from training data, PGS accuracy decreases by 14% in the furthest versus closest decile in the European GIA. Notably, the furthest decile of individuals of European ancestries showed similar accuracy to the closest decile of Hispanic Latino individuals. Characterizing PGS accuracy across the continuum allows the inclusion of individuals unassigned to any GIA (6% of all ATLAS), thus allowing more individuals to be included in PGS applications. Finally, we explore the relationship between GD and PGS estimates themselves. Of 84 PGSs, 82 show significant correlation between GD and PGS with 30 showing opposite correlation (GD, trait) versus (GD, PGS); we exemplify the importance of incorporating GD in interpretation of PGSs using height and neutrophils in the ATLAS data. Our results demonstrate the need to incorporate the genetic ancestry continuum in assessing PGS performance and/or bias.

3.2 Results

3.2.1 Overview of the Study

PGS accuracy has conventionally been assessed at the level of discrete GIA clusters using population-level metrics of accuracy. Individuals from diverse genetic backgrounds are routinely grouped into discrete GIA clusters using computational inference methods such as PCA [110] and/or admixture analysis [111] (Fig. 3.5a). Population-level metrics of PGS accuracy are then estimated for each GIA cluster and generalized to everyone in the cluster (Fig. 3.5b). This approach has three major limitations: the inter-individual variability within each cluster is ignored; the GIA cluster boundary is sensitive to algorithms and reference panels used for clustering; and a substantial proportion of individuals may not be assigned to any GIA owing to a lack of reference panels for genetic ancestry inference (for example, individuals of uncommon or admixed ancestries).

Here we evaluate PGS accuracy across the genetic ancestry continuum at the level of a single target individual. We model the phenotype of individual i as $y_i = x_i^T\beta + \epsilon_i$, in which x_i is an $M \times 1$ vector of standardized genotypes for M variants, β is an $M \times 1$ vector of standardized causal effects, and ϵ_i is random noise. Under a random effects model, genetic liability $g_i = x_i^T\beta$ and its PGS estimate $\hat{g}_i = E(x_i^T\beta|D)$ are random variables for which the randomness comes from β and training data $D = (X_{\text{train}}, y_{\text{train}})$. We define the individual PGS accuracy as the correlation of an individual's genetic liability and PGS estimate with the following equation in consistency with classical theory[41, 92, 95]:

$$r_i^2(g_i, \hat{g}_i) = \frac{\text{cov}_{\beta,D}(g_i, \hat{g}_i)^2}{\text{var}_\beta(g_i)\text{var}_{\beta,D}(\hat{g}_i)} = 1 - \frac{E_D(\text{var}_{\beta|D}(x_i^T\beta))}{\text{var}_\beta(x_i^T\beta)} \tag{3.1}$$

Under an infinitesimal assumption for which all variants are causal and drawn from a normal distribution $N(0, \sigma_\beta^2)$, the analytical form of PGS accuracy can be derived as:

$$r_i^2(g_i, \hat{g}_i) = 1 - \frac{\sigma_e^2 \sum_{j=1}^{J} \frac{1}{\lambda_j} x_i^T v_j v_j^T x_i}{\sigma_\beta^2 x_i^T x_i} = 1 - \frac{\sigma_e^2}{\sigma_\beta^2} \frac{\sum_{j=1}^{J} \frac{1}{\lambda_j} x_i^T v_j v_j^T x_i}{x_i^T x_i}$$

in which σ_β^2 is per single nucleotide polymorphism (SNP) heritability; σ_e^2 is the variance of residual environmental noise; v_j and λ_j are the jth eigenvector and eigenvalues of training genotype data, and J is the total number of eigenvectors. $\sum_{j=1}^{J} \frac{1}{\lambda_j} x_i^T v_j v_j^T x_i$ is the squared Mahalanobis distance of the testing individual i from the center of the training genotype data on its principal component (PC) space, and $x_i^T x_i$ is the sum of squared genotypes across all variants. Empirically, the ratio of the squared Mahalanobis distance to the sum of squared genotypes is highly correlated with the Euclidean distance of the individual from the training data on that PC space ($R = 1$, $P < 2.2 \times 10^{-16}$ in the UKBB). Given that this metric of accuracy is highly dependent on the GD from the training data, we term it the panel distance r_i^2. In practice, we use LDpred2 to estimate $E_D(\text{var}_{\beta|D}(x_i^T \beta))$ (refs. [33, 109]) and approximate $\text{var}_\beta(x_i^T \beta)$ as the heritability of the phenotype[93] (Methods). As a continuous GD, we use $d_i = \sqrt{\sum_{j=1}^{J} (x_i^T v_j)^2}$ with J set to 20 (Fig. 1c,d and Methods). We note two caveats of individual PGS accuracy: first, the genetic effects are assumed to be the same for all individuals regardless of their genetic ancestry background; second, the SNPs used for PGS training may not fully capture trait heritability. Therefore, the metric we proposed here is an upper bound of genetic prediction accuracy.

3.2.2 PGS performance is calibrated in simulations

First, we evaluated calibration of the posterior variance of genetic liability $E_D(\text{var}_{\beta|D}(x_i^T \beta))$ estimated by LDpred2 for individuals at various GDs from the UKBB WB training data by checking the calibration of the 90% credible intervals (Fig. 3.5a). We simulated 100 phenotypes at heritability $h_g^2 = 0.25$ and proportion of causal variants $p_{\text{causal}} = 1\%$ for all individuals in the UKBB, assuming shared causal variants and homogeneous causal effect sizes for individuals from various genetic backgrounds (Methods). Overall, the 90% credible intervals are approximately well calibrated (that is, the 90% credible interval overlaps with the true genetic liability across 90 of 100 replicates, for all individuals, regardless of their GD from the training population or GIA labels; Fig. 3.5a). For example, when individuals are binned into 10 deciles based on their GD from the training population, the average empirical coverage of the 90% credible intervals is 89.7% (s.d. 2.6%) for individuals from the closest decile (composed of 96.9% individuals labelled as WB, 3.1% labelled as PL under a discrete view of ancestries; see detailed naming convention in) compared to the average empirical coverage of 82.4% (s.d. 4.6%) for individuals from the furthest decile (composed of 19.9% individuals labelled as CB and 80.1% labelled as NG).

Next, we investigated the impact of GD on individual-level PGS accuracy. As expected, the width of the credible interval increases linearly with GD, reflecting reduced predictive accuracy for the PGS (Fig. 3.5b). The average width of the 90% credible interval is 1.83 in the furthest decile of GD, a 1.8-fold increase over the average width in the closest decile of GD. In contrast to the credible interval width, the individual-level PGS accuracy $\hat{r}_i^2$ decreases with GD from the training data (Fig. 3.5c); the average estimated accuracy of individuals in the closest decile GD is fourfold higher than that of individuals in the furthest decile. Even among the most homogeneous grouping of individuals traditionally labeled as WB, we observe a 5% relative decrease in accuracy for individuals at the furthest decile of GD as compared to those in the closest decile. Similar results are observed when using a population-level PGS metric of accuracy, albeit at the expense of binning individuals according to GD; we find a high degree of concordance between the average $\hat{r}_i^2$ within the bin and the population-level R^2 estimated within the bin (Fig. 3.5d). Similarly, we observe a high consistency between average $\hat{r}_i^2$ and squared correlation between PGS and simulated phenotypes ($R = 0.86, P < 10^{-10}$). Taken together, our results show that the 90% credible intervals remain calibrated for individuals that are genetically distant from the training population at the expense of wider credible intervals, and $\hat{r}_i^2$ captures the PGS accuracy decay across GD.

To demonstrate that the continuous accuracy decay is not specific to PGS models trained on European ancestries, we conducted further analyses using a non-European training dataset composed of individuals of NG and CB GIAs (we grouped the two GIAs to attain sufficient sample size for simulations). We simulated a high signal-to-noise trait by setting $h_g^2 = 0.8$ and proportion of causal variants $p_{\text{causal}} = 1\%$ and 0.1% with 56,539 SNPs on chromosome 10 alone. We trained PGS models on 5,000 individuals from the NG and CB GIA clusters and applied the models to the remaining testing individuals. The coverage of the 90% credible intervals was invariant to GD despite slight miscalibration. The 90% credible interval width increased and individual PGS accuracy decreased when the testing individual was further away from the training data. This trend is consistent with the observed decrease in empirical accuracy computed as squared correlation between PGS and genetic value as GD increases.

We further evaluated the impact of the number of PCs used for calculating GD on its ability to capture accuracy decay. We varied the number of PCs (J) from 1 to 20 and observed that the correlation between GD and individual accuracy ($-\text{cor}(d_i, r_i^2(g_i, \hat{g}_i))$) increases when more PCs are used for computing GD, but no further improvement is observed when $J > 15$ for any GIA clusters or the whole biobank. Therefore, we set $J = 20$ for simplicity. We also explored average squared genetic relationship from training data as an alternative metric of GD and found that it is a better prediction of accuracy decay within each GIA cluster. However, because this metric relies on individual-level training data that are usually not available, we choose to use PCA-based GD for convenience.

3.2.3 PGS accuracy across the genetic continuum

Having validated our approach in simulations, we next turn to empirical data. For illustration purposes, we use height as an example, focusing on the ATLAS biobank as the target population with PGS trained on the 371,018 WB individuals from the UKBB . Other traits show similar trends and are presented in the next sections. PGS accuracy at the individual level varies with GD across the entire biobank as well as within each GIA cluster (Fig. 3.5). For example, GD strongly correlates with PGS accuracy of individuals in the GIA cluster labeled as Hispanic Latino American (HL, $R = -0.84$) and African American (AA, $R = -0.88$) in ATLAS. Notably, GD correlates with PGS accuracy even in non-admixed GIA clusters with correlations as -0.66, -0.66, and -0.35 for European American (EA), South Asian American (SAA), or East Asian American (EAA) GIA clusters, respectively. Similar qualitative results are also observed when applying PGS to test data from the UKBB;

significant negative correlations are found between GD and individual PGS accuracy in all of the subcontinental GIA clusters in the UKBB, with correlation coefficients ranging from $R = -0.031$ for the WB cluster to $R = -0.62$ for the CB cluster.

This trend is consistent with the observed decrease in empirical accuracy computed as squared correlation between PGS and genetic value as GD increases.

We further evaluated the impact of the number of PCs used for calculating GD on its ability to capture accuracy decay. We varied the number of PCs (J) from 1 to 20 and observed that the correlation between GD and individual accuracy ($-\mathrm{cor}(d_i, r_i^2(g_i, \hat{g}_i))$ increases when more PCs are used for computing GD, but no further improvement is observed when $J > 15$ for any GIA clusters or the whole biobank. Therefore, we set $J = 20$ for simplicity. We also explored average squared genetic relationship from training data as an alternative metric of GD and found that it is a better prediction of accuracy decay within each GIA cluster. However, because this metric relies on individual-level training data that are usually not available, we choose to use PCA-based GD for convenience.

3.2.4 PGS accuracy varies across the genetic continuum

Having validated our approach in simulations, we next turn to empirical data. For illustration purposes, we use height as an example, focusing on the ATLAS biobank as the target population with PGS trained on the 371,018 WB individuals from the UKBB (Methods); other traits show similar trends and are presented in the next sections. PGS accuracy at the individual level varies with GD across the entire biobank as well as within each GIA cluster (Fig. 3.5). For example, GD strongly correlates with PGS accuracy of individuals in the GIA cluster labelled as Hispanic Latino American (HL, $R = -0.84$) and African American (AA, $R = -0.88$) in ATLAS. Notably, GD correlates with PGS accuracy even in non-admixed GIA clusters with correlations as -0.66, -0.66, and -0.35 for European American (EA), South Asian American (SAA), or East Asian American (EAA) GIA clusters, respectively. Similar qualitative results are also observed when applying PGS to test data from the UKBB; significant negative correlations are found between GD and individual PGS accuracy in all of the subcontinental GIA clusters in the UKBB.

Next, we focused on the impact of GD on PGS accuracy across all ATLAS individuals regardless of GIA clustering ($R = -0.96, P < 10^{-10}$; Fig. 3.5b). Notably, we find a strong overlap of PGS accuracies across individuals from different GIA clusters demonstrating the limitation of using a single cluster-specific metric of accuracy. For example, when rank-ordering by GD, we find that the individuals from the closest GD decile in the HL cluster have similar estimated accuracy to the individuals from the furthest GD decile in EA cluster (average $\hat{r}_i^2$ of 0.71 versus 0.71). This shows that GD enables identification of HL individuals with similar PGS performance to the EA cluster thus partly alleviating inequities due to limited access to accurate PGS. Most notably, GD can be used to evaluate PGS performance for individuals that cannot be easily clustered by current genetic inference methods (6% of ATLAS; Fig. 3.5b) partly owing to limitations of reference panels and algorithms for assigning ancestries. Among this traditionally overlooked group of individuals, we find the GD ranging from 0.02 to 0.64 and their corresponding estimated PGS accuracy $\hat{r}_i^2$ ranging from 0.63 to 0.21. In addition to evaluating PGS accuracy with respect to the genetic liability, we also evaluated accuracy with respect to the residual height after regressing out sex, age and PC1-10 on the ATLAS from the actual measured trait. Using equally spaced bins across the GD continuum, we find that correlation between PGS and the measured height tracks significantly with GD ($R = -0.92, P = 1.1 \times 10^{-8}$; Fig. 3.5c).

3.2.5 PGS accuracy decay is pervasive

Having established the coupling of GD with PGS accuracy in simulations and for height, we next investigate whether this relationship is common across complex traits using PGSs for a broad set of 84 traits (Supplementary Table 1). We find consistent and pervasive correlations of GD with PGS accuracy across all considered traits in both ATLAS and the UKBB (Fig. 3.4). For example, the correlations between GD and individual PGS accuracy range from -0.71 to -0.97 with an average of -0.95 across the 84 PGSs in ATLAS with similar results observed in the UKBB. Traits with sparser genetic architectures and fewer non-zero weights in the PGS have a lower correlation between GD and PGS accuracy; we reason that this is because GD represents genome-wide genetic variation patterns that may not reflect a limited number of causal SNPs well. For example, PGS for lipoprotein A (log_lipoA) has the lowest estimated polygenicity (0.02%) among the 84 traits and has the lowest correlation in ATLAS (-0.71) and the UKBB (-0.85). By contrast, we observe a high correlation between GD and PGS accuracy (>0.9) for all traits with an estimated polygenicity >0.1%.

Next, we show that the fine-scale population structure accountable for the individual PGS accuracy variation is also prevalent within the traditionally defined genetic ancestry group. For example, in ATLAS we find that 501 of 504 (84 traits across 6 GIA clusters) trait-ancestry pairs have significant associations between GD and individual PGS accuracy after Bonferroni correction. In the UKBB, we find 572 of the 756 (84 traits across 9 subcontinental GIA clusters) trait-ancestry pairs have significant associations between GD and PGS accuracy after Bonferroni correction. We also find that a more stringent definition of homogeneous GIA clusters results in a lower correlation magnitude. Empirical analyses of PGS accuracy show a similar trend. When averaging across 84 traits, we find that the empirical accuracy decreases with increased GD across GIA clusters as reported by previous studies[33]. Further analyses based on GD bins show the decreasing trend at a finer scale.

3.2.6 PGS varies across the genetic continuum

We have focused so far on investigating the relationship between GD (d_i) and PGS accuracy ($\hat{r}_i^2$). Next, we evaluate the impact of GD on PGS estimates ($\hat{g}_i$) themselves. We find a significant correlation between GD and PGS estimates for 82 of 84 traits, with correlation coefficients ranging from $R = -0.52$ to $R = 0.74$; this broad range of correlations is in stark contrast with the consistently observed negative correlation between GD and PGS accuracy. To better understand whether the coupling of PGS with GD is due to stratification or true signal, we compared the correlation of GD with PGS estimates ($\text{cor}(d_i, \hat{g}_i)$) to the correlation of GD with measured phenotype values ($\text{cor}(d_i, y_i)$). We find a wide range of couplings reflecting trait-specific signals; for 30 traits, GD correlates in opposite directions with PGS versus phenotype; for 40 traits, GD correlates in the same directions with PGS versus phenotype but differs in correlation magnitudes. For example, GD shows opposite and significantly different correlations for PGS versus trait for years of education (years_of_edu, $\text{cor}(y_i, d_i)$ = 0.03, $\text{cor}(\hat{g}_i, d_i)$ = -0.18). Other traits, such as hair colour, show a highly consistent impact of GD on PGS versus trait (darker_hair, $\text{cor}(y_i, d_i) = 0.59$, $\text{cor}(\hat{g}_i, d_i) = 0.74$), whereas for monocyte percentage, GD shows different magnitudes albeit with the same directions (monocyte_perc, $\text{cor}(y_i, d_i)$ = -0.03, $\text{cor}(\hat{g}_i, d_i)$ = -0.52). Moreover, GD correlates with PGS and phenotype even within the same GIA cluster, and the correlation patterns vary across clusters.

The correlation of GD with phenotype and PGS is also observed in ATLAS. For example, both height phenotype and height PGS vary along GD in ATLAS (Fig. 3.5); this holds true even when restricting analysis to the EA genetic ancestry cluster (Supplementary Fig. 1). This is consistent with genetic liability driving difference in phenotypes but could also be explained by residual population stratification. For neutrophil counts, phenotype and PGS vary in opposite directions with respect to GD across the ATLAS (Fig. 3.5), although the trend is similar for phenotype and PGS in the EA GIA clusters (Supplementary Fig. 1). This could be explained by genetic liability driving signal in Europeans with stratification for other groups. Neutrophil counts have been reported to vary greatly across ancestry groups with reduced counts in individuals of African ancestries [112]. In ATLAS, we observe a negative correlation (-0.04) between GD and neutrophil counts in agreement with the previous reports, whereas GD is positively correlated (0.08) with PGS estimates—genetically distant individuals traditionally labeled as African American having higher PGS than average. The opposite directions in phenotype–distance and PGS–distance correlations are partly attributed to the Duffy-null SNP rs2814778 on chromosome 1q23.2. This variant is strongly associated with neutrophil counts among individuals traditionally identified as African ancestry, but it is rare and excluded in our training data. This exemplifies the potential bias in PGS due to non-shared causal variants and emphasizes ancestral diversity in genetic studies. As PGS can vary across GD either as a reflection of true signal (that is, genetic liability varying with ancestry) or owing to biases in PGS estimation ranging from unaccounted residual population stratification to incomplete data (for example, partial ancestry-specific tagging of causal effects), our results emphasize the need to consider GD in PGS interpretation beyond adjusting for PGS r_i^2.

3.3 Discussion

In this work, we have shown that PGS accuracy varies from individual to individual and proposed an approach to personalize PGS metrics of performance. We used a PCA-based GD [101] from the centre of training data to describe an individual's unique location on the genetic ancestry continuum and showed that individual PGS accuracy tracks well with GD. The continuous decay of PGS performance as the target individual becomes further away from the training population is pervasive across traits and ancestries. We highlight the variability in PGS performance along the continuum of genetic ancestries, even within traditionally defined homogeneous populations. As the genetic ancestries are increasingly recognized as continuous rather than discrete [68, 78, 103–106], the individual-level PGS accuracy provides a powerful tool to study PGS performance across diverse individuals to enhance the utility of PGS. For example, by using individual-level PGS accuracy, we can identify individuals from Hispanic Latino GIA who have similar PGS accuracy to individuals of European GIA, thus partly alleviating inequities due to lack of access to accurate PGS.

Simulation and real-data analyses show that individual PGS accuracy is highly correlated with GD, in alignment with existing works showing that decreased similarity (measured by relatedness, linkage disequilibrium and/or minor allele frequency differences, fixation index (Fst) and so on) [69, 113] between testing individuals and training data is a major contributor to PGS accuracy decay. However, practical factors that may affect transferability, such as genotype–environment interaction and population-specific causal variants, are not modelled in the calculation of individual PGS accuracy and this is left for future work.

Our results emphasize the importance of PGS training in diverse ancestries [114] as it can provide advantages for all individuals. Broadening PGS training beyond European ancestries can lead to improved accuracy in genetic effect estimation particularly for variants with higher frequencies in non-European data. It can also increase PGS portability by reducing the GD from target to training data. However, increased diversity may also bring challenges to statistical modelling; for example, differences in genetic effects may correlate with environment factors and could bias genetic risk prediction. To address these challenges, more sophisticated statistical methods are needed that can effectively leverage ancestrally diverse populations to train PGS [76, 115–117]. Concerted global effort and equitable collaborations are also crucial to increase the sample size of underrepresented individuals as part of an effort to reduce health disparities across ancestries [114, 118].

We highlight the pervasive correlation between PGS estimates and GD of varying magnitude and sign as compared to the correlation between phenotype and GD. This provides a finer resolution of the mean shift of PGS estimates across genetic ancestry groupings [22]. The correlation between GD and PGS estimates can arise from bias and/or true biological difference, and more effort is needed to investigate the PGS bias in the context of genetic ancestry continuum.

We note several limitations and future directions of our work. First, our proposed individual PGS accuracy is an upper bound of true accuracy and should be interpreted only in terms of the additive heritability captured by SNPs included in the model. Missing heritability [119, 120] and misspecification of the heritability model along with population-specific causal variants and effect sizes may further decrease real accuracy. For example, the prediction accuracy for neutrophil count is overestimated among African American individuals because the Duffy-null SNP rs2814778 [112] is not captured in the UKBB WB training data. Future work could investigate the impact of the population-specific components of genetic architecture on the calibration of PGS accuracy. Second, we approximate the variance of genetic liability in the denominator of equation (1) with heritability and set a fixed value for all individuals. Preliminary results show that replacing the denominator with a Monte Carlo estimation of genetic liability variance recapitulates the accuracy decay in estimated PGS accuracy, albeit the correlation is slightly reduced. Third, individual PGS accuracy evaluates how well the PGS estimates the genetic liability instead of phenotype. Quantifying the individual accuracy of PGS with respect to phenotype can be achieved by also modelling non-genetic factors for proper calibration. Fourth, limited by sample size, we combined GIA groups as a training set in simulation experiments to replicate PGS accuracy decay; this is not an optimal strategy for data analysis as the population structure in the training data may confound the true genetic effects and reduce prediction accuracy. We leave a more comprehensive investigation of non-European PGS training data for future work. Sixth, although we advocate for the use of continuous genetic ancestry, we trained our PGS models on a discrete GIA cluster of WB because current PGS methods rely on discrete genetic ancestry groupings. We leave the development of PGS training methods that are capable of modelling continuous ancestries as future work. Finally, we highlight that, just like PGS, the traditional clinical risk assessment may suffer from limited portability across diverse populations [83]. For examples, the pooled cohort equation overestimates atherosclerotic cardiovascular disease risk among non-European populations [81]; and a traditional clinical breast cancer risk model developed in the European population in the USA overestimated the breast cancer risk among older Korean women [82]. Here we focus on genetic prediction portability owing to the wide interest and attention from both the research community and society. We emphasize that improving the portability of traditional clinical risk factor models in diverse populations is an essential component of health equity and requires thorough investigation.

3.4 Methods

3.4.1 Individual PGS accuray

3.4.1.1 Model setup

We model the phenotype of an individual with a standard linear model $y_i = x_i^\top \beta + \epsilon_i$, in which x_i is an $M \times 1$ vector of standardized genotypes (centred and standardized with respect to the allele frequency in the training population for both training and testing individuals), β is an

$M \times 1$ vector of standardized genetic effects, and ϵ_i is random noise. Under a random effects model, β is a vector of random variable sampled from a prior distribution $p(\beta)$ that differs under different genetic architecture assumptions[120] and PGS methods[33, 36, 59, 121]. The PGS weights $\hat{\beta} = E_{\beta|D}(\beta)$ are estimated to be the posterior mean given the observed data D ($D = (X_{\text{train}}, y_{\text{train}})$ with access to individual-level genotype, X_{train}, and phenotype, y_{train}; or $D = (\hat{\beta}_{\text{GWAS}}, \hat{R})$ with access to marginal association statistics $\hat{\beta}_{\text{GWAS}}$ and LD matrix $\hat{R}$, in which GWAS stands for genome-wide association study). The genetic liability ($g_i = x_i^T \beta$) of an individual i is estimated to be $\hat{g}_i = E_{\beta|D}(x_i^T \beta)$, the uncertainty of which is estimated as the posterior variance of genetic liability $\text{var}(\hat{g}_i) = \text{var}_{\beta|D}(x_i^T \beta)$ (ref. [109]).

3.4.1.2 Definition of individual PGS accuracy

We define individual PGS accuracy as the squared correlation between an individual's genetic liability, g_i, and its PGS estimate, $\hat{g}_i$, following the general form in ref. [92]:

$$r_i^2 = \frac{\text{cov}_{\beta,D}(g_i, \hat{g}_i)^2}{\text{var}_{\beta,D}(g_i)\text{var}_{\beta,D}(\hat{g}_i)} = \frac{\text{var}_D(x_i^\top \hat{\beta})^2}{\text{var}_\beta(x_i^\top \beta)\text{var}_D(x_i^\top \hat{\beta})} \tag{3.2}$$

Here we are interested in the PGS accuracy of a given individual; therefore, the genotype is treated as a fixed variable, and genetic effects are treated as a random variable. We note that a random effects model is essential; otherwise, $\text{cov}_{\beta,D}(g_i, \hat{g}_i)$ and $\text{var}_{\beta,D}(g_i)$ are 0. Under a random effects model, both the genetic liability and PGS estimate for individual i are random variables. The randomness of $g_i = x_i^\top \beta$ comes from the randomness in β, and the randomness of $\hat{g}_i = x_i^\top \hat{\beta}$ comes from the randomness of both β and the training data D. Individual PGS accuracy measures the correlation between g_i and $\hat{g}_i$, which can be computed with the following equation:

$$r_i^2 = 1 - \frac{E_D(\text{var}_{\beta|D}(x_i^\top \beta))}{\text{var}_\beta(x_i^\top \beta)}$$

In which $\text{var}_{\beta|D}(x_i^\top \beta)$ is the posterior variance of genetic liability given the training data, and $\text{var}_\beta(x_i^\top \beta)$ is the genetic variance. The equation is derived as follows.

First, we show that under the random effects model, $\mathrm{cov}_{\beta,D}(x_i^\top\hat{\beta}, x_i^\top\beta) = \mathrm{var}_D(x_i^\top\hat{\beta})$ (in which $\hat{\beta} = E_{\beta|D}(\beta)$) following equation 5.149 in ref. [40]:

$$\begin{aligned}
\mathrm{cov}_{\beta,D}(\hat{\beta}, \beta^\top) &= E_{\beta,D}(\hat{\beta}\beta^\top) - E_{\beta,D}(\hat{\beta})E_{\beta,D}(\beta^\top) \\
&= E_D(E_{\beta|D}(\hat{\beta}\beta^\top)) - E_{D,\beta}(\hat{\beta})E_D(E_{\beta|D}(\beta^\top)) \\
&= E_D(E_{\beta|D}(E_{\beta|D}(\beta)\beta^\top)) - E_D(E_{\beta|D}(\beta))E_D(E_{\beta|D}(\beta^\top)) \\
&= E_D(E_{\beta|D}(\beta)E_{\beta|D}(\beta^\top)) - E_D(E_{\beta|D}(\beta))E_D(E_{\beta|D}(\beta^\top)) \\
&= \mathrm{var}_D(E_{\beta|D}(\beta)) \\
&= \mathrm{var}_D(\hat{\beta})
\end{aligned}$$

Multiplying x_i on both sides of the equation, we obtain:

$$x_i^\top \mathrm{cov}_{\beta,D}(\hat{\beta}, \beta)x_i = x_i^\top \mathrm{var}_D(\hat{\beta})x_i$$

$$\mathrm{cov}_{\beta,D}(x_i^\top\hat{\beta}, x_i^\top\beta) = \mathrm{var}_D(x_i^\top\hat{\beta}) \tag{3.3}$$

Equation (3) also implies the slope from regression of observed phenotypic values (or true genetic liability) on the estimated PGS equal to 1, which offers an alternative way to assess the calibration of PGS as done in refs. [36, 40].

$$\mathrm{slope} = \frac{\mathrm{cov}(x_i^\top\hat{\beta}, y_i)}{\mathrm{var}(x_i^\top\hat{\beta})} = \frac{\mathrm{cov}(x_i^\top\hat{\beta}, x_i^\top\beta + \epsilon_i)}{\mathrm{var}(x_i^\top\hat{\beta})} = \frac{\mathrm{var}(x_i^\top\hat{\beta})}{\mathrm{var}(x_i^\top\hat{\beta})} = 1$$

Next, by applying the law of total variance, we show that:

$$\mathrm{var}_{\beta,D}(g_i) = \mathrm{var}_{\beta,D}(x_i^T\beta) = E_D(\mathrm{var}_{\beta|D}(x_i^T\beta)) + \mathrm{var}_D(E_{\beta|D}(x_i^T\beta))$$

$$\mathrm{var}_D(x_i^\top\hat{\beta}) = \mathrm{var}_{\beta,D}(x_i^\top\beta) - E_D(\mathrm{var}_{\beta|D}(x_i^\top\beta)) \tag{3.4}$$

Third, we derive the correlation between g_i and $\hat{g}_i$ as:

$$\begin{aligned}
r_i^2 &= \frac{\text{cov}_{\beta,D}(g_i, \hat{g}_i)^2}{\text{var}_{\beta,D}(g_i)\text{var}_{\beta,D}(\hat{g}_i)} \\
&= \frac{\text{var}_D(x_i^\top \hat{\beta})^2}{\text{var}_\beta(x_i^\top \beta)\text{var}_D(x_i^\top \hat{\beta})} \text{ by applying equation (3)} \\
&= \frac{\text{var}_D(x_i^\top \hat{\beta})}{\text{var}_\beta(x_i^\top \beta)} \\
&= \frac{\text{var}_{\beta,D}(x_i^\top \beta) - E_D(\text{var}_{\beta|D}(x_i^\top \beta))}{\text{var}_\beta(x_i^\top \beta)} \text{ by applying equation (4)} \\
&= 1 - \frac{E_D(\text{var}_{\beta|D}(x_i^\top \beta))}{\text{var}_\beta(x_i^\top \beta)}
\end{aligned}$$

The above equation is widely used in animal breeding theory to compute the reliability of estimated breeding value for each individual [93]. In this work, we use individual PGS uncertainty $\text{var}(\hat{g}_i) = \text{var}_{\beta|D}(x_i^\top \beta)$ as an unbiased estimator of $E_D(\text{var}_{\beta|D}(x_i^\top \beta))$. We also use estimated heritability to approximate $\text{var}_\beta(x_i^\top \beta)$ in simulations in which the phenotype has unit variance. In real-data analysis, as the phenotype does not necessarily have unit variance, we approximate $\text{var}_\beta(x_i^\top \beta)$ by scaling the estimated heritability with the residual phenotypic variance in the training population after regressing GWAS covariates including sex, age, and precomputed UKBB PC1-16 (Data-Field 22009).

3.4.1.3 Analytical form of individual PGS accuracy under infinitesimal assumption

Without loss of generality, we assume a prior distribution of genetic effects as follows:

$$p(\beta|\sigma_\beta^2) = MVN(0, \sigma_\beta^2 I_M)$$

where M is the number of genetic variants. With access to individual genotype, X_{train}, and phenotype, y_{train}, data, the likelihood of the data is

$$p(y_{\text{train}}|X_{\text{train}}, \beta, \sigma_{\text{e}}^2) = MVN(X_{\text{train}}\beta, \sigma_{\text{e}}^2 I_N)$$

where N is the training sample size. The posterior distribution of genetic effects given the data is proportional to the product of the prior and the likelihood:

$$\begin{aligned} p(\beta|X_{\text{train}}, y_{\text{train}}, \sigma_\beta^2, \sigma_{\text{e}}^2) &\propto p(\beta|\sigma_\beta^2)p(y_{\text{train}}|X_{\text{train}}, \beta, \sigma_{\text{e}}^2) \\ &\propto MVN(0, \sigma_\beta^2 I_M)MVN(X_{\text{train}}\beta, \sigma_{\text{e}}^2 I_N) \\ &\propto MVN(\mu_\beta, \sigma_\beta) \end{aligned}$$

in which $\mu_\beta = (\frac{\sigma_{\text{e}}^2}{\sigma_\beta^2} I_M + X_{\text{train}}^\top X_{\text{train}})^{-1} X_{\text{train}}^\top y_{\text{train}}$ and $\mathit{\Sigma}_\beta = \sigma_{\text{e}}^2(\frac{\sigma_{\text{e}}^2}{\sigma_\beta^2} I_M + X_{\text{train}}^\top X_{\text{train}})^{-1}$. This form is equivalent to the solution of random effects in the best linear unbiased prediction with the pedigree matrix or genetic relationship matrix[39, 95].

For a new target individual, the posterior variance of the genetic liability is:

$$\text{var}(x_i^\top \beta|x_i, X_{\text{train}}, y_{\text{train}}, \sigma_\beta^2, \sigma_{\text{e}}^2) = x_i^\top \mathit{\Sigma}_\beta x_i = \sigma_{\text{e}}^2 x_i^\top \left(\frac{\sigma_{\text{e}}^2}{\sigma_\beta^2} I_M + X_{\text{train}}^\top X_{\text{train}}\right)^{-1} x_i$$

After carrying out eigendecomposition on $X_{\text{train}}^\top X_{\text{train}} = \sum_{j=1}^J \lambda_j v_j v_j^\top$, we can rewrite

$$\left(\frac{\sigma_{\text{e}}^2}{\sigma_\beta^2} I_M + X_{\text{train}}^\top X_{\text{train}}\right)^{-1} = \left(\frac{\sigma_{\text{e}}^2}{\sigma_\beta^2} I_M + \sum_{j=1}^J \lambda_j v_j v_j^\top\right)^{-1} = \sum_{j=1}^J \left(\frac{\sigma_{\text{e}}^2}{\sigma_\beta^2} + \lambda_j\right)^{-1} v_j v_j^\top$$

in which λ_j and v_j correspond to the jth eigenvalue and unit-length eigenvector of the training genotype, X_{train}.

Thus, we can rewrite the posterior variance of genetic liability as

$$\text{var}(x_i^\top \beta|x_i, X_{\text{train}}, y_{\text{train}}, \sigma_\beta^2, \sigma_{\text{e}}^2) = \sigma_{\text{e}}^2 \sum_{j=1}^J \left(\frac{\sigma_{\text{e}}^2}{\sigma_\beta^2} + \lambda_j\right)^{-1} x_i^\top v_j v_j^\top x_i$$

Replacing $E_D(\text{var}_{\beta|D}(x_i^\top \beta))$ in equation (2) with the analytical form of $\text{var}(x_i^\top \beta|x_i, X_{\text{train}}, y_{\text{train}}, \sigma_\beta^2, \sigma_{\text{e}}^2)$, we get

$$r_i^2 = 1 - \frac{\text{var}(x_i^\top \beta|x_i, X_{\text{train}}, y_{\text{train}}, \sigma_\beta^2, \sigma_{\text{e}}^2)}{\text{var}(x_i^\top \beta)} = 1 - \frac{\sigma_{\text{e}}^2 \sum_{j=1}^J \left(\frac{\sigma_{\text{e}}^2}{\sigma_\beta^2} + \lambda_j\right)^{-1} x_i^\top v_j v_j^\top x_i}{\sigma_\beta^2 x_i^\top x_i}$$

As the eigenvalue of $X_{\text{train}}^\top X_{\text{train}}$ increases linearly with training sample size N (ref. 68), at the UKBB-level sample size (for example, $N = 371,018$ for our UKBB WB training data), the eigenvalues for the top PCs are usually larger than the ratio of environmental noise variance and genetic variance $\frac{\sigma_e^2}{\sigma_\beta^2}$. Thus, we can further approximate the analytical form with:

$$r_i^2 = 1 - \frac{\sigma_{\text{e}}^2 \sum_{j=1}^{J} \frac{1}{\lambda_j} x_i^\top v_j v_j^\top x_i}{\sigma_\beta^2 x_i^\top x_i} = 1 - \frac{\sigma_{\text{e}}^2}{\sigma_\beta^2} \frac{\sum_{j=1}^{J} \frac{1}{\lambda_j} x_i^\top v_j v_j^\top x_i}{x_i^\top x_i}$$

The term $\sum_{j=1}^{J} \frac{1}{\lambda_j} x_i^\top v_j v_j^\top x_i$ is the squared Mahalanobis distance of the testing individual i from the centre of the training genotype data on its PC space and $x_i^\top x_i$ is the sum of squared genotype across all variants. Empirically, the ratio between the two is highly correlated with the Euclidean distance of the individual from the training data on that PC space ($R = 1$, P value $< 2.2 \times 10^{-16}$ in the UKBB).

3.4.2 Genetic distance (GD)

3.4.2.1 Definition of GD

The GD is defined as the Euclidean distance between a target individual and the centre of training data on the PC space of training data.

$$d_i = \sqrt{\sum_{j=1}^{J} (x_i^\top v_j - \bar{x}_{\text{train}} v_j)^2} = \sqrt{\sum_{j=1}^{J} (x_i^\top v_j)^2}$$

in which d_i is the GD of a testing individual i from the training data, x_i is an $M \times 1$ standardized genotype vector for testing individual i, v_j is the jth eigenvector for the genotype matrix of training individuals, $\bar{x}_{\text{train}}$ is the average genotype in the training population ($\bar{x}_{\text{train}} v_j = 0$ given that the genotypes are centred with respect to the allele frequency in the training population), and J is set to 20.

3.4.2.2 GD from PGS training data

To compute the GD of testing individuals from the training population, we carry out PCA on the 371,018 UKBB WB training individuals and project the 48,586 UKBB testing individuals and 36,778 ATLAS testing individuals on the PC space. We start from the 979,457 SNPs that are overlapped in UKBB and ATLAS. First, we carry out LD pruning with plink2 (`--indep-pairwise 1000 50 0.05`) and exclude the long-range LD regions. Next, we carry out PCA analysis with flashpca2[122] on the 371,018 UKBB WB training individuals to obtain the top 20 PCs. Then, we project the remaining 48,586 UKBB individuals that are not included in the training data and 36,778 ATLAS individuals onto the PC space of training data by using SNP loadings (`--outload loadings.txt`) and their means and standard deviations (`--outmeansd meansd.txt`) output from flashpca2. In the end, we compute the GD for each individual as the Euclidean distance of their PCs from the centre of training data with the equation

$$d_i = \sqrt{\sum_{j=1}^{20} (\mathrm{pc}_{ij})^2}$$

in which pc_{ij} is the jth PC of individual i.

3.4.3 Ancestry ascertainment

3.4.3.1 Ancestry ascertainment in UKBB

The UKBB individuals are clustered into nine subcontinental GIA clusters—WB (white British), PL (Poland), IR (Iran), IT (Italy), AS (Ashkenazi), IN (India), CH (China), CB (Caribbean) and NG (Nigeria)—based on the top 16 precomputed PCs (Data-Field 22009) as described in ref.[33]. First, UKBB participants are grouped by country of origin (Data-Field 20115) and the centre of each country on the PC space is computed as the geometric median for all countries, which serves as a proxy for the centre for each subcontinental ancestry. The centre of Ashkenazi GIA is determined using a dataset from ref.[101]. Second, we reassign each individual to one of the nine GIA groups on the basis of their Euclidean distance to the centres on the PC space, as the self-reported country of origin does not necessarily match an individual's genetic ancestry. The genetic ancestry of an individual is labelled as unknown if its distance to any genetic ancestry centre is larger than one-eighth of the maximum distance between any pairs of subcontinental ancestry clusters. We are able to cluster 91% of the UKBB participants into 411,018 WB, 4,127 PL, 1,169 IR, 6,499 IT, 2,352 AS, 1,798 CH, 2,472 CB and 3,894 NG. GIAs are not necessarily reflective of the full genetic diversity of a particular region but reflect only the diversity present in the UKBB individuals.

3.4.3.2 Ancestry Ascertainment in ATLAS

The ATLAS individuals are clustered into five GIA clusters—European Americans (EA), Hispanic Latino Americans (HL), South Asian Americans (SAA), East Asian Americans (ESA) and African Americans (AA)—as described in ref. [108] on the basis of their proximity to 1000 Genome super populations on the PC space. First, we filter the ATLAS-typed genotypes with plink2 by Mendel error rate (`plink --me 1 1 -set-me-missing`), founders (`--filter-founders`), minor allele frequency (`-maf 0.15`), genotype missing call rate (`--geno 0.05`) and Hardy-Weinberg equilibrium test P value (`-hwe 0.001`). Next, ATLAS genotypes were merged with the 1000 Genomes phase 3 dataset. Then, linkage disequilibrium (LD) pruning was carried out on the merged dataset (`--indep 200 5 1.15 --indep-pairwise 100 5 0.1`). The top 10 PCs were computed with the flashpca2 (ref. [122]) software with all default parameters. Next, we use the super population label and PCs of the 1000 Genome individuals to train the K-nearest neighbours model to assign genetic ancestry labels to each ATLAS individual. For each ancestry cluster, we run the K-nearest neighbours model on the pair of PCs that capture the most variation for each genetic ancestry group: the European, East Asian and African ancestry groups use PCs 1 and 2, the Admixed American group uses PCs 2 and 3, and the South Asian group uses PCs 4 and 5. In each analysis, we use tenfold cross-validation to select the k hyper-parameter from $k = 5, 10, 15, 20$. If an individual is assigned to multiple ancestries with probability larger than 0.5 or is not assigned to any clusters, their ancestry is labelled as unknown. We label the five 1000 Genome super population as EA for Europeans, HL for Admixed Americans, SAA for South Asians, AA for Africans and ESA for East Asians. We can cluster 95% of the ATLAS participants into 22,380 EA, 6,973 HL, 625 SAA, 3,331 EAA and 1,995 AA, and the ancestry of 2,332 individuals is labelled as unknown.

3.4.4 Simulations

3.4.4.1 Simulation setup

We use simulations on all UKBB individuals with 1,054,151 UKBB HapMap 3 SNPs to investigate the impact of GD from training data on the various metrics of PGS. We fix the proportion of causal SNPs $p_{\text{causal}} = 0.01$ and heritability as $h_g^2 = 0.25$. The simulated genetic effects and phenotype are generated as follows. First, we randomly sample

$$\beta_m \sim \begin{cases} N\left(0, \frac{h_g^2}{\text{var}(x_m) M p_{\text{causal}}}\right) & c_j = 1, \text{with probability } p_{\text{causal}} \\ 0 & c_j = 0, \text{with probability } 1 - p_{\text{causal}} \end{cases} \tag{3.5}$$

in which $\text{var}(x_m)$ is the variance of allele counts for SNP m among all UKBB individuals. Second, we compute the genetic liability for each individual as $g_i = \sum_{m=1}^{M} x_{im}\beta_m$ and randomly sample environmental noise $\epsilon_i \sim N(0, 1 - h_g^2)$. Third, we generate phenotype as $y_i = g_i + \epsilon_i$. We repeat the process 100 times to generate 100 sets of genetic liability and phenotypes.

3.4.4.2 Calibration of credible interval in simulation

We run the LDpred2 model on 371,018 WB training individuals for the 100 simulation replicates. In each simulation, for individual with genotype x_i, we compute $x_i^T \widetilde{\beta}_r^{(1)}, x_i^T \widetilde{\beta}_r^{(2)}, \ldots, x_i^T \widetilde{\beta}_r^{(B)}$ to approximate their posterior distribution of genetic liability, generate a 90% credible interval $\text{CI} - g_{ir}$ (90% credible interval of genetic liability of the ith individual in rth replication) with 5% and 95% quantile of the distribution and check whether their genetic liability is contained in the credible interval $I(g_{ir} \in \text{CI} - g_{ir})$. We compute the empirical coverage for each individual as the mean across the 100 simulation replicates $\text{coverage}_i = \frac{1}{100}\sum_{r=1}^{100} I(g_{ir} \in \text{CI} - g_{ir})$.

3.4.5 LDpred2 PGS model training

The PGS models were trained on 371,018 UKBB individuals labelled as WB with the LDpred2 [33] method for both simulation and real-data analysis. For simulation analysis, we use 1,054,151 UKBB HapMap 3 variants. For real-data analysis, we use 979,457 SNPs that are overlapped in UKBB HapMap 3 variants and ATLAS imputed genotypes.

First, we obtain GWAS summary statistics by carrying out GWAS on the training individuals with plink2 using sex, age and precomputed PC1-16 as covariates. Second, we calculate the in-sample LD matrix with the function `snp_cor` from the R package bigsnpr[123]. Next, we use the GWAS summary statistics and LD matrix as input for the `snp_ldpred2_auto` function in bigsnpr to sample from the posterior distribution of genetic effect sizes. Instead of using a held-out validation dataset to select hyperparameters p (proportion of causal variants) and h^2 (heritability), `snp_ldpred2_auto` estimates the two parameters from data with the Markov chain Monte Carlo (MCMC) method directly. We run 10 chains with different initial sparsity p from 10^{-4} to 1 equally spaced in log space. For all chains, we set the initial heritability as the LD score regression heritability[73] estimated by the built-in function `snp_ldsc`. We carry out quality control of the 10 chains by filtering out chains with estimated heritability that is smaller than 0.7 times the median heritability of the 10 chains or with estimated sparsity that is smaller than 0.5 times the median sparsity or larger than 2 times the median sparsity. For each chain that passes filtering, we remove the first 100 MCMC iterations as burn-in and thin the next 500 iterations by selecting every fifth iteration to reduce autocorrelation between MCMC samples. In the end, we obtain an $M \times B$ matrix $[\widetilde{\beta}^{(1)}, \widetilde{\beta}^{(2)}, ..., \widetilde{\beta}^{(B)}]$, in which each column of the matrix $\widetilde{\beta}^{(b)}$ is a sample of posterior causal effects of the M SNPs. Owing to the quality control of MCMC chains, the total number of posterior samples B ranges from 500 to 1,000.

3.4.6 Calculation of PGS and accuracy

We use the score function in plink2 to compute the PGS for 48,586 and 36,778 testing individuals in UKBB and ATLAS, respectively. For each $\widetilde{\beta}^{(b)}$, we compute the PGS for each individual i as $x_i^\top \widetilde{\beta}^{(b)}$ with plink2 (`--score`). For each individual with genotype x_i, we compute $x_i^\top \widetilde{\beta}^{(1)}, x_i^\top \widetilde{\beta}^{(2)}, ..., x_i^\top \widetilde{\beta}^{(B)}$ to approximate its posterior distribution of genetic liability. The genotype $x_i^\top$ is centred to the average allele count (`--read-freq`) in training data to reduce the uncertainty from the unmodelled intercept. We estimate the PGS with the posterior mean of the genetic liability as $\hat{g}_i = E_{\beta|D}(x_i^\top \beta) = \frac{1}{B} \sum_{b=1}^{B} x_i^\top \widetilde{\beta}^{(b)}$. We estimate the individual-level PGS uncertainty as $\text{var}(\hat{g}_i) = \text{var}_{\beta|D}(x_i^\top \beta) = \frac{1}{B} \sum_{b=1}^{B} (x_i^\top \widetilde{\beta}^{(b)} - \hat{g}_i)^2$. The individual-level PGS accuracy is calculated as $\hat{r_i^2} = 1 - \frac{\text{var}(\hat{g}_i)}{h_g^2}$ for simulation (h_g^2 is the heritability estimated by the LDpred2 model) and as $\hat{r_i^2} = 1 - \frac{\text{var}(\hat{g}_i)}{h_g^2 \text{var}(y_{\text{train}} - \hat{y}_{\text{train}})}$ for real-data analysis, in which $\text{var}(y_{\text{train}} - \hat{y}_{\text{train}})$ is the variance of residual phenotype in training data after regressing out GWAS covariates.

3.5 Figures

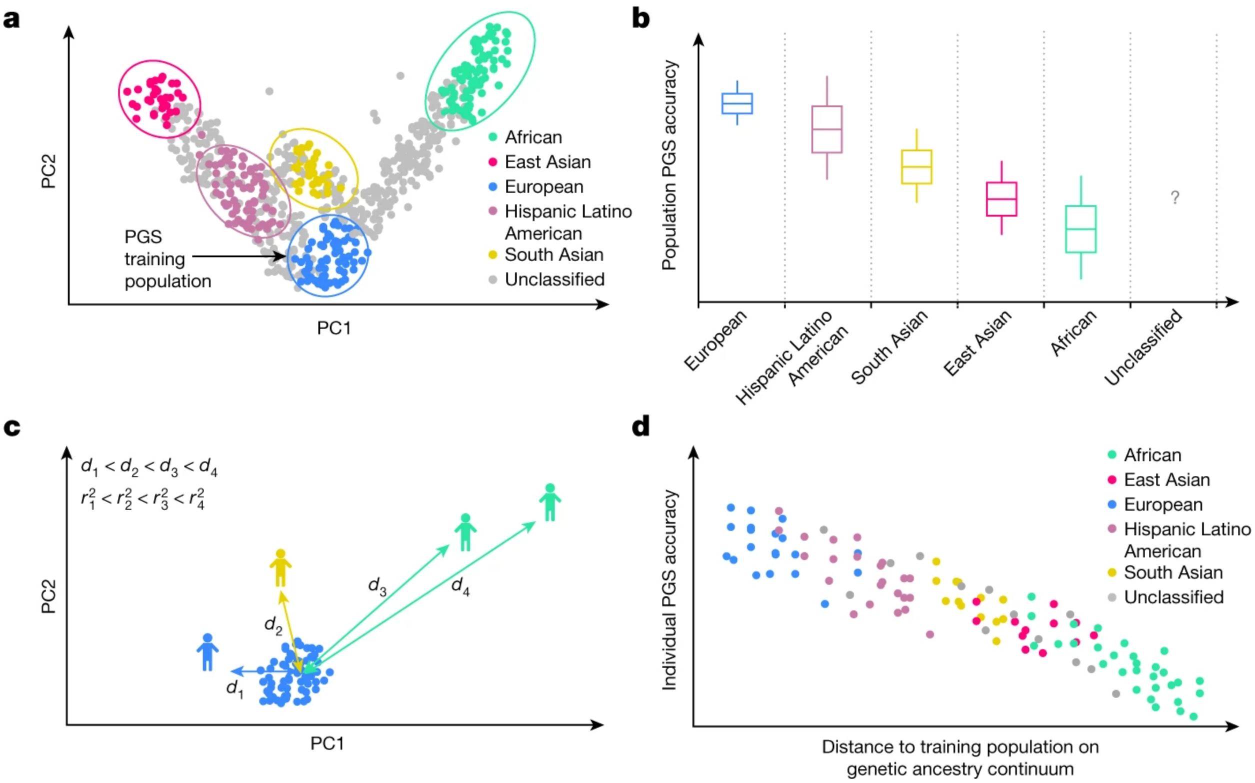
a
PC2
PC1
PGS
training
population
● African
● East Asian
● European
● Hispanic Latino American
● South Asian
● Unclassified
b
Population PGS accuracy
European
Hispanic Latino American
South Asian
East Asian
African
Unclassified
?
c
PC2
PC1
$d_1 < d_2 < d_3 < d_4$
$r_1^2 < r_2^2 < r_3^2 < r_4^2$
d_1
d_2
d_3
d_4
d
Individual PGS accuracy
Distance to training population on genetic ancestry continuum
● African
● East Asian
● European
● Hispanic Latino American
● South Asian
● Unclassified

Figure 3.1: **Illustration of population-level versus individual-level PGS accuracy.(a)** Discrete labelling of GIA with PCA-based clustering. Each dot represents an individual. The circles represent arbitrary boundaries imposed on the genetic ancestry continuum to divide individuals into different GIA clusters. The colour represents the GIA cluster label. The grey dots are individuals who are left unclassified. **(b)** Schematic illustrating the variation of population-level PGS accuracy across clusters. The box plot represents the PGS accuracy (for example, R^2) measured at the population level. The question mark emphasizes that the PGS accuracy for unclassified individuals is unknown owing to the lack of a reference group. Grey dashed lines emphasize the categorical nature of GIA clustering. **(c)** Continuous labelling of everyone's unique position on the genetic ancestry continuum with a PCA-based GD. The GD is defined as the Euclidean distance of an individual's genotype from the centre of the training data when projected on the PC space of training genotype data. Everyone has their own unique GD, d_i, and individual PGS accuracy, r_i^2. **(d)** Individual-level PGS accuracy decays along the genetic ancestry continuum. Each dot represents an individual and its colour represents the assigned GIA label. Individuals labelled with the same ancestry spread out on the genetic ancestry continuum, and there are no clear boundaries between GIA clusters. This figure is illustrative and does not involve any real or simulated data.

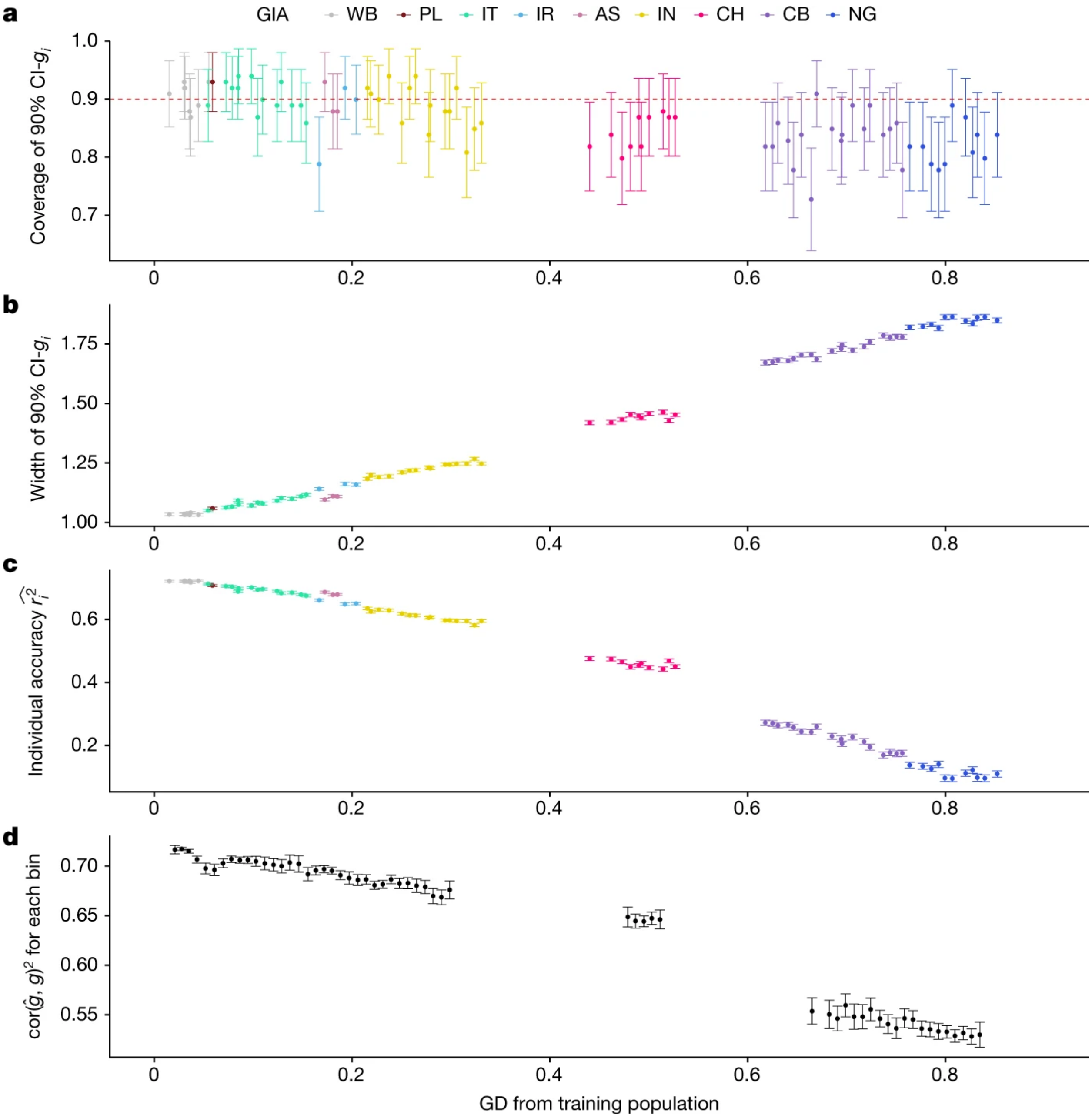
a
GIA
WB
PL
IT
IR
AS
IN
CH
CB
NG
Coverage of 90% CI-g_i
1.0
0.9
0.8
0.7
0
0.2
0.4
0.6
0.8
b
Width of 90% CI-g_i
1.75
1.50
1.25
1.00
0
0.2
0.4
0.6
0.8
c
Individual accuracy $\widehat{r_i^2}$
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
d
cor($\hat{g}$, g)^2 for each bin
0.70
0.65
0.60
0.55
0
0.2
0.4
0.6
0.8
GD from training population

Figure 3.2: **PGS performance is calibrated across GD in simulations using UKBB data.** **(a)** The 90% credible intervals of genetic liability (CI-g_i) are well calibrated for testing individuals at all GDs. The red dashed line represents the expected coverage of the 90% CI-g_i. Each dot represents a randomly selected UKBB testing individual. For each dot, the x-axis is its GD from the training data, the y-axis is the empirical coverage of the 90% CI-g_i calculated as the proportion of simulation replicates for which the 90% CI-g_i contain the individual's true genetic liability, and the error bars represent the mean ± 1.96 standard error of the mean (s.e.m.) of the empirical coverage calculated from 100 simulations. **(b)** The width of the 90% CI-g_i increases with GD. For each dot, the y-axis is the average width of the 90% CI-g_i across 100 simulation replicates, and the error bars represent ± 1.96 s.e.m. **(c)** Individual PGS accuracy decreases with GD. For each dot, the y-axis is the average individual-level PGS accuracy across 100 simulation replicates, and the error bars represent ± 1.96 s.e.m. **(d)** Population-level metrics of PGS accuracy recapitulate the decay in PGS accuracy across the genetic continuum. All UKBB testing individuals are divided into 100 equal-interval bins based on their GD. The x-axis is the average GD for the bin, and the y-axis is the squared correlation between genetic liability and PGS estimates for the individuals within the bin. The dot and error bars represent the mean and ± 1.96 s.e.m from 100 simulations, respectively.

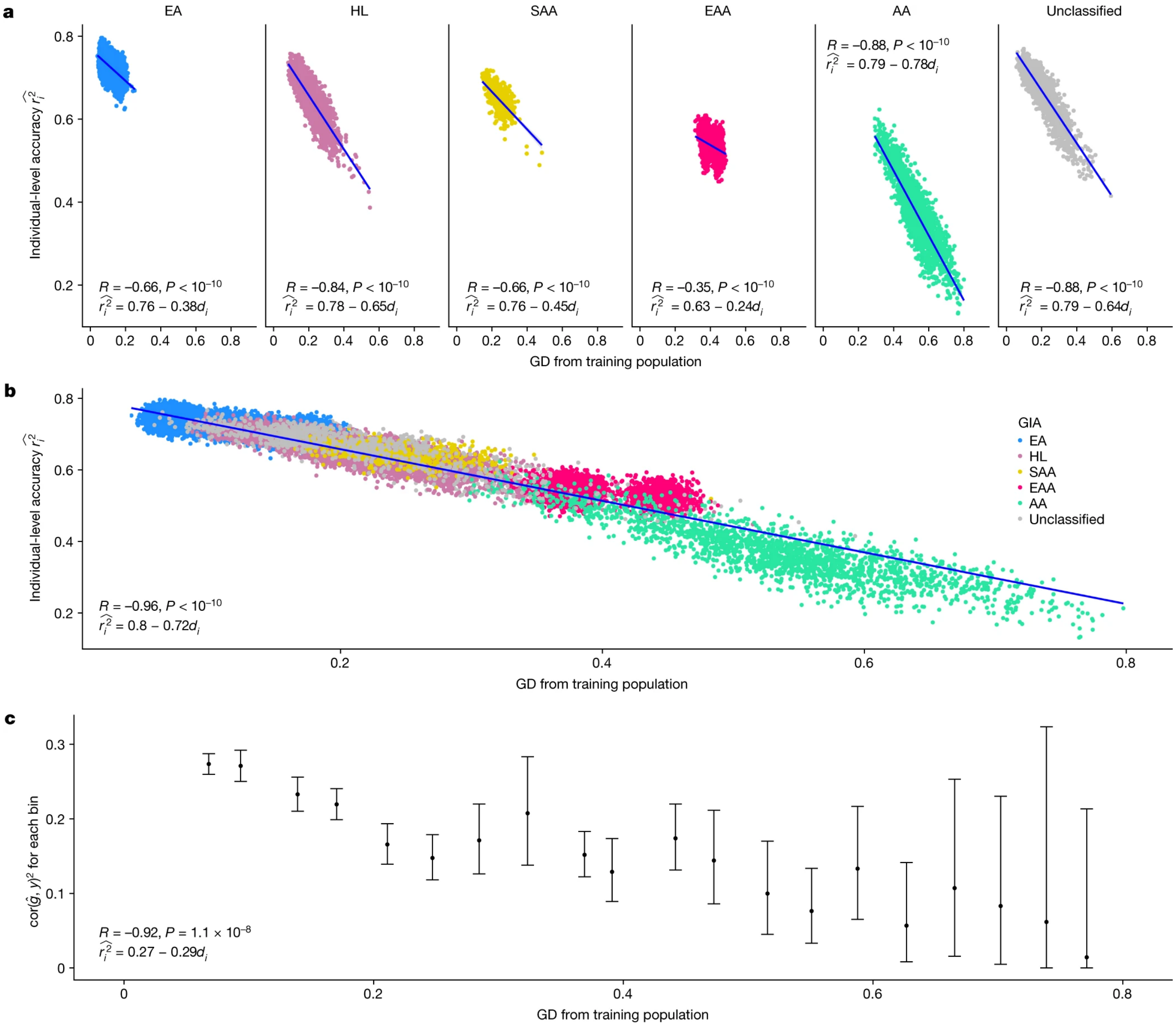
a
EA
HL
SAA
EAA
AA
Unclassified
Individual-level accuracy $\widehat{r_i^2}$
0.8
0.6
0.4
0.2
$R = -0.66$, $P < 10^{-10}$
$\widehat{r_i^2} = 0.76 - 0.38d_i$
$R = -0.84$, $P < 10^{-10}$
$\widehat{r_i^2} = 0.78 - 0.65d_i$
$R = -0.66$, $P < 10^{-10}$
$\widehat{r_i^2} = 0.76 - 0.45d_i$
$R = -0.35$, $P < 10^{-10}$
$\widehat{r_i^2} = 0.63 - 0.24d_i$
$R = -0.88$, $P < 10^{-10}$
$\widehat{r_i^2} = 0.79 - 0.78d_i$
$R = -0.88$, $P < 10^{-10}$
$\widehat{r_i^2} = 0.79 - 0.64d_i$
0 0.2 0.4 0.6 0.8
GD from training population
b
Individual-level accuracy $\widehat{r_i^2}$
0.8
0.6
0.4
0.2
GIA
EA
HL
SAA
EAA
AA
Unclassified
$R = -0.96$, $P < 10^{-10}$
$\widehat{r_i^2} = 0.8 - 0.72d_i$
0.2
0.4
0.6
0.8
GD from training population
c
$\mathrm{cor}(\hat{g}, y)^2$ for each bin
0.3
0.2
0.1
0
$R = -0.92$, $P = 1.1 \times 10^{-8}$
$\widehat{r_i^2} = 0.27 - 0.29d_i$
0
0.2
0.4
0.6
0.8
GD from training population

Figure 3.3: **The individual-level accuracy for height PGS decreases across the genetic ancestry continuum in ATLAS.** **(a)** Individual PGS accuracy decreases within both homogeneous and admixed genetic GIA clusters. Each dot represents a testing individual from ATLAS. For each dot, the x-axis represents its distance from the training population on the genetic continuum; the y-axis represents its PGS accuracy. The colour represents the GIA cluster. **(b)** Individual PGS accuracy decreases across the entire ATLAS. **(c)** Population-level PGS accuracy decreases with the average GD in each GD bin. All ATLAS individuals are divided into 20 equal-interval GD bins. The x-axis is the average GD within the bin, and the y-axis is the squared correlation between PGS and phenotype for individuals in the bin; the dot and error bar show the mean and 95% confidence interval from 1,000 bootstrap samples. R and P refer to the correlation between GD and PGS accuracy and its significance, respectively, from two-sided Pearson correlation tests without adjustment for multiple hypothesis testing. Any P value below 10^{-10} is shown as $P < 10^{-10}$. EA, European American; HL, Hispanic Latino American; SAA, South Asian American; EAA, East Asian American; AA, African American.

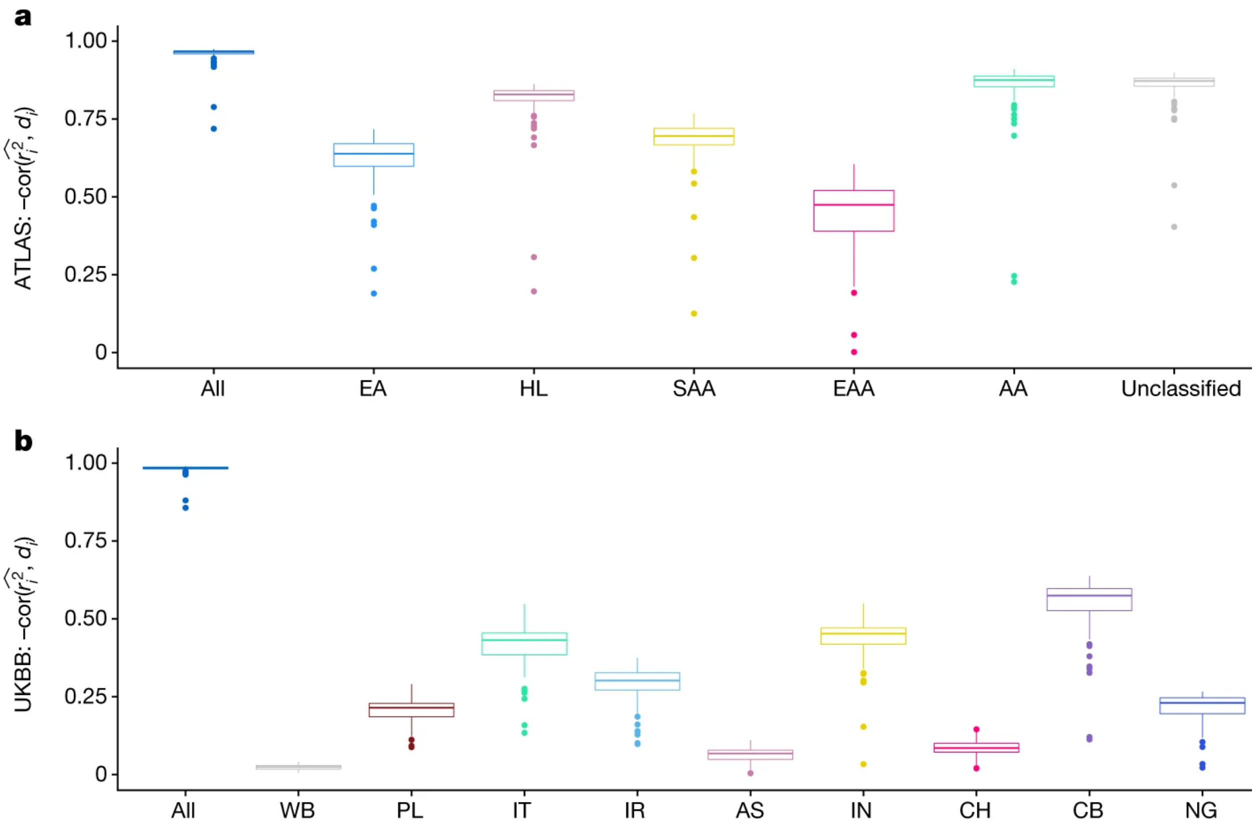

Figure 3.4: **The correlation between individual PGS accuracy and GD is pervasive across 84 traits across ATLAS and the UKBB. (a)** The distribution of correlation between individual PGS accuracy and GD for 84 traits in ATLAS. **(b)** The distribution of correlation between individual PGS accuracy and GD for 84 traits in the UKBB. Each box plot contains 84 points corresponding to the correlation between PGS accuracy and GD within the GIA group specified by the x-axis for each of the 84 traits. The box shows the first, second and third quartiles of the 84 correlations, and the whiskers extend to the minimum and maximum estimates located within 1.5 x IQR from the first and third quartiles, respectively. Numerical results are reported in Supplementary Tables 2 and 3.

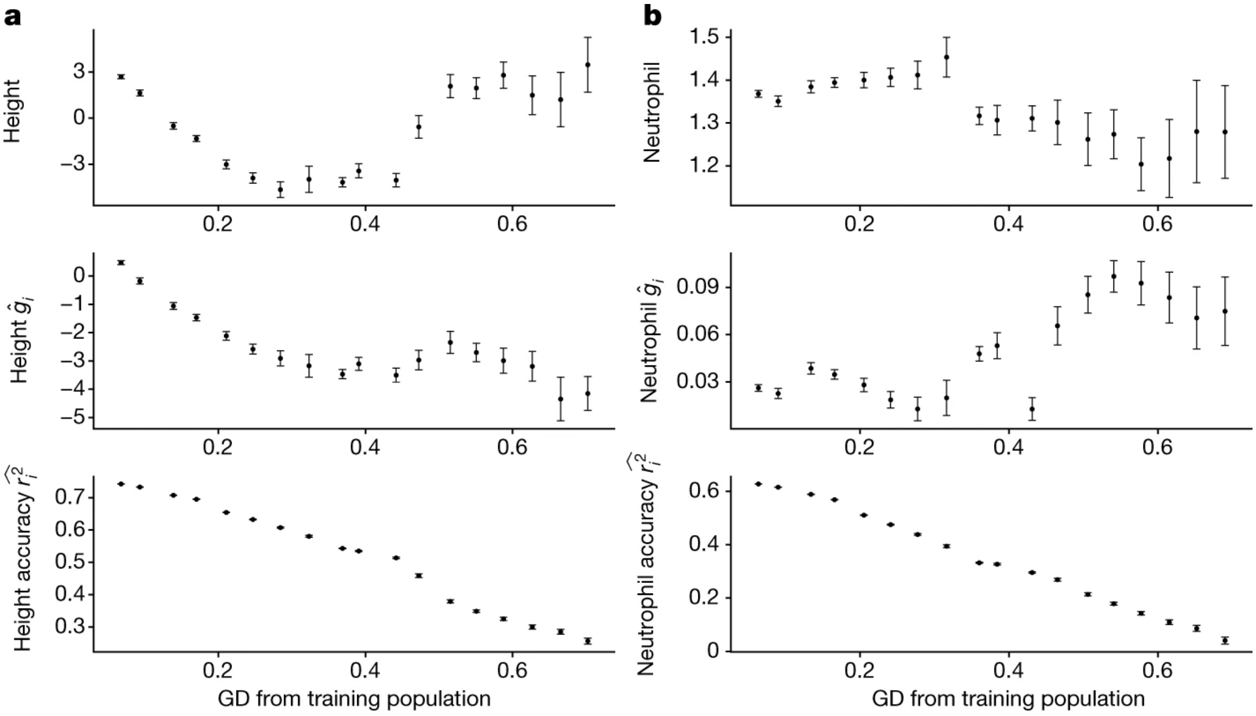

Figure 3.5: **Measured phenotype, PGS estimates and accuracy vary across ATLAS. (a)** Variation of height phenotype, PGS estimates and accuracy across different GD bins in ATLAS. **(b)** Variation of log neutrophil count phenotype, PGS estimates and accuracy across different GD bins in ATLAS. The 36,778 ATLAS individuals are divided into 20 equal-interval GD bins. Bins with fewer than 50 individuals are not shown owing to large s.e.m. All panels share the same layout: the x axis is the average GD within the bin; the y axis is the average phenotype (top), PGS (middle) and individual PGS accuracy (bottom); the error bars represent ± 1.96 s.e.m.

CHAPTER 4

Conclusion

In my thesis, I focus on evaluating the performance of PGS at the individual level by assessing individual PGS uncertainty (Chapter 2) and accuracy (Chapter 3), without relying on any arbitrary population discretizations. Furthermore, I also demonstrate that it is the increased genetic distance, rather than population membership, that primarily leads to the observed decrease in PGS performance (Chapter 3).

The first part of my research suggests that despite the potential utility of PGS at the population level, such as in cancer screening programs, the substantial uncertainty at the individual level poses challenges in utilizing PGS for informed clinical decision-making in personalized medicine. For instance, consider an individual whose PGS for cardiovascular disease falls at the 90th percentile of a reference population, its true genetic risk may range widely from the 15th to the 99th percentile. To address this issue, we proposed a probabilistic approach to stratification, which enhances PGS-based decision-making by optimizing stratification and downstream medical treatment based on individual-specific cost functions.

Importantly, the second part of my research reveals a continuous decay in PGS performance (marked by increased individual level uncertainty and decreased individual level accuracy) as targets diverge genetically from the training population. Although the limited portability of PGS across different ancestral populations has been demonstrated by many previous studies, our study reveals, for the first time, the striking individual-to-indiviudal decerease of PGS performance along the genetic ancestry continumm, even within traditionally labelled 'homogeneous' genetic ancestries. Through theoretical derivations and empirical analyses, we further demonstrate that the increased genetic distance of a target individual is the cause of decrased accuracy rather than population membership. For example, when applying PGS models trained on individuals labelled as white British in the UKBB to individuals with European ancestries in ATLAS, individuals in the furthest genetic distance decile have 14% lower accuracy relative to the closest decile; while the Hispanic Latino American individuals situated at a similar genetic distance to this group of European individuals displayed comparable PGS performance.

To conclude, while PGS is increasingly recognoized as a promising tool for personalized medicine, my reasearch underscores the critical need of evaluating and interpreting PGS at the individual level - the core unit in the realm of personalized medicine, for its appropriate and equitable application. Our findings of continuous variation of PGS accuracy whithin and across groupings challenge the traditional assumptions of inter-group heterogeneity and within-group homogeneity prevalent in population-based methods, highlight the importance of conceptulazing individual based on its unique profiles (as opposed to population labels) in healthcare and call for the increased diversity in genetic studies to ensure more accurate and inclusive healthcare solutions.

Bibliography

[1] Malika Kumar Freund, Kathryn S Burch, Huwenbo Shi, Nicholas Mancuso, Gleb Kichaev, Kristina M Garske, David Z Pan, Zong Miao, Karen L Mohlke, Markku Laakso, Päivi Pajukanta, Bogdan Pasaniuc, and Valerie A Arboleda. Phenotype-Specific enrichment of mendelian disorder genes near GWAS regions across 62 complex traits. *Am. J. Hum. Genet.*, 103(4):535–552, October 2018.

[2] Samuel A Lambert, Gad Abraham, and Michael Inouye. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.*, 28(R2):R133–R142, November 2019.

[3] Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F O'Reilly. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.*, 15(9):2759–2772, September 2020.

[4] Amit V Khera, Mark Chaffin, Kaitlin H Wade, Sohail Zahid, Joseph Brancale, Rui Xia, Marina Distefano, Ozlem Senol-Cosar, Mary E Haas, Alexander Bick, Krishna G Aragam, Eric S Lander, George Davey Smith, Heather Mason-Suares, Myriam Fornage, Matthew Lebo, Nicholas J Timpson, Lee M Kaplan, and Sekar Kathiresan. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell*, 177(3): 587–596.e9, April 2019.

[5] Allison Meisner, Prosenjit Kundu, Yan Dora Zhang, Lauren V Lan, Sungwon Kim, Disha Ghandwani, Parichoy Pal Choudhury, Sonja I Berndt, Neal D Freedman, Montserrat Garcia-Closas, and Nilanjan Chatterjee. Combined utility of 25 disease and risk factor polygenic risk scores for stratifying risk of All-Cause mortality. *Am. J. Hum. Genet.*, 107(3):418–431, September 2020.

[6] Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, Jonathan P Tyrer, Ting-Huei Chen, Qin Wang, Manjeet K Bolla, Xin Yang, Muriel A Adank, Thomas Ahearn, Kristiina Aittomäki, Jamie Allen, Irene L Andrulis, Hoda Anton-Culver, Natalia N Antonenkova, Volker Arndt, Kristan J Aronson,

Paul L Auer, Päivi Auvinen, Myrto Barrdahl, Laura E Beane Freeman, Matthias W Beckmann, Sabine Behrens, Javier Benitez, Marina Bermisheva, Leslie Bernstein, Carl Blomqvist, Natalia V Bogdanova, Stig E Bojesen, Bernardo Bonanni, Anne-Lise Børresen-Dale, Hiltrud Brauch, Michael Bremer, Hermann Brenner, Adam Brentnall, Ian W Brock, Angela Brooks-Wilson, Sara Y Brucker, Thomas Brüning, Barbara Burwinkel, Daniele Campa, Brian D Carter, Jose E Castelao, Stephen J Chanock, Rowan Chlebowski, Hans Christiansen, Christine L Clarke, J Margriet Collée, Emilie Cordina-Duverger, Sten Cornelissen, Fergus J Couch, Angela Cox, Simon S Cross, Kamila Czene, Mary B Daly, Peter Devilee, Thilo Dörk, Isabel Dos-Santos-Silva, Martine Dumont, Lorraine Durcan, Miriam Dwek, Diana M Eccles, Arif B Ekici, A Heather Eliassen, Carolina Ellberg, Christoph Engel, Mikael Eriksson, D Gareth Evans, Peter A Fasching, Jonine Figueroa, Olivia Fletcher, Henrik Flyger, Asta Försti, Lin Fritschi, Marike Gabrielson, Manuela Gago-Dominguez, Susan M Gapstur, José A García-Sáenz, Mia M Gaudet, Vassilios Georgoulias, Graham G Giles, Irina R Gilyazova, Gord Glendon, Mark S Goldberg, David E Goldgar, Anna González-Neira, Grethe I Grenaker Alnæs, Mervi Grip, Jacek Gronwald, Anne Grundy, Pascal Guénel, Lothar Haeberle, Eric Hahnen, Christopher A Haiman, Niclas Håkansson, Ute Hamann, Susan E Hankinson, Elaine F Harkness, Steven N Hart, Wei He, Alexander Hein, Jane Heyworth, Peter Hillemanns, Antoinette Hollestelle, Maartje J Hooning, Robert N Hoover, John L Hopper, Anthony Howell, Guanmengqian Huang, Keith Humphreys, David J Hunter, Milena Jakimovska, Anna Jakubowska, Wolfgang Janni, Esther M John, Nichola Johnson, Michael E Jones, Arja Jukkola-Vuorinen, Audrey Jung, Rudolf Kaaks, Katarzyna Kaczmarek, Vesa Kataja, Renske Keeman, Michael J Kerin, Elza Khusnutdinova, Johanna I Kiiski, Julia A Knight, Yon-Dschun Ko, Veli-Matti Kosma, Stella Koutros, Vessela N Kristensen, Ute Krüger, Tabea Kühl, Diether Lambrechts, Loic Le Marchand, Eunjung Lee, Flavio Lejbkowicz, Jenna Lilyquist, Annika Lindblom, Sara Lindström, Jolanta Lissowska, Wing-Yee Lo, Sibylle Loibl, Jirong Long, Jan Lubiński, Michael P Lux, Robert J MacInnis, Tom Maishman, Enes Makalic, Ivana Maleva Kostovska, Arto Mannermaa, Siranoush Manoukian, Sara Margolin, John

W M Martens, Maria Elena Martinez, Dimitrios Mavroudis, Catriona McLean, Alfons Meindl, Usha Menon, Pooja Middha, Nicola Miller, Fernando Moreno, Anna Marie Mulligan, Claire Mulot, Victor M Muñoz-Garzon, Susan L Neuhausen, Heli Nevanlinna, Patrick Neven, William G Newman, Sune F Nielsen, Børge G Nordestgaard, Aaron Norman, Kenneth Offit, Janet E Olson, Håkan Olsson, Nick Orr, V Shane Pankratz, Tjoung-Won Park-Simon, Jose I A Perez, Clara Pérez-Barrios, Paolo Peterlongo, Julian Peto, Mila Pinchev, Dijana Plaseska-Karanfilska, Eric C Polley, Ross Prentice, Nadege Presneau, Darya Prokofyeva, Kristen Purrington, Katri Pylkäs, Brigitte Rack, Paolo Radice, Rohini Rau-Murthy, Gad Rennert, Hedy S Rennert, Valerie Rhenius, Mark Robson, Atocha Romero, Kathryn J Ruddy, Matthias Ruebner, Emmanouil Saloustros, Dale P Sandler, Elinor J Sawyer, Daniel F Schmidt, Rita K Schmutzler, Andreas Schneeweiss, Minouk J Schoemaker, Fredrick Schumacher, Peter Schürmann, Lukas Schwentner, Christopher Scott, Rodney J Scott, Caroline Seynaeve, Mitul Shah, Mark E Sherman, Martha J Shrubsole, Xiao-Ou Shu, Susan Slager, Ann Smeets, Christof Sohn, Penny Soucy, Melissa C Southey, John J Spinelli, Christa Stegmaier, Jennifer Stone, Anthony J Swerdlow, Rulla M Tamimi, William J Tapper, Jack A Taylor, Mary Beth Terry, Kathrin Thöne, Rob A E M Tollenaar, Ian Tomlinson, Thérèse Truong, Maria Tzardi, Hans-Ulrich Ulmer, Michael Untch, Celine M Vachon, Elke M van Veen, Joseph Vijai, Clarice R Weinberg, Camilla Wendt, Alice S Whittemore, Hans Wildiers, Walter Willett, Robert Winqvist, Alicja Wolk, Xiaohong R Yang, Drakoulis Yannoukakos, Yan Zhang, Wei Zheng, Argyrios Ziogas, ABCTB Investigators, kConFab/AOCS Investigators, NBCS Collaborators, Alison M Dunning, Deborah J Thompson, Georgia Chenevix-Trench, Jenny Chang-Claude, Marjanka K Schmidt, Per Hall, Roger L Milne, Paul D P Pharoah, Antonis C Antoniou, Nilanjan Chatterjee, Peter Kraft, Montserrat García-Closas, Jacques Simard, and Douglas F Easton. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.*, 104(1):21–34, January 2019.

[7] Tyler M Seibert, Chun Chieh Fan, Yunpeng Wang, Verena Zuber, Roshan Karuna-

muni, J Kellogg Parsons, Rosalind A Eeles, Douglas F Easton, Zsofia Kote-Jarai, Ali Amin Al Olama, Sara Benlloch Garcia, Kenneth Muir, Henrik Grönberg, Fredrik Wiklund, Markus Aly, Johanna Schleutker, Csilla Sipeky, Teuvo Lj Tammela, Børge G Nordestgaard, Sune F Nielsen, Maren Weischer, Rasmus Bisbjerg, M Andreas Røder, Peter Iversen, Tim J Key, Ruth C Travis, David E Neal, Jenny L Donovan, Freddie C Hamdy, Paul Pharoah, Nora Pashayan, Kay-Tee Khaw, Christiane Maier, Walther Vogel, Manuel Luedeke, Kathleen Herkommer, Adam S Kibel, Cezary Cybulski, Dominika Wokolorczyk, Wojciech Kluzniak, Lisa Cannon-Albright, Hermann Brenner, Katarina Cuk, Kai-Uwe Saum, Jong Y Park, Thomas A Sellers, Chavdar Slavov, Radka Kaneva, Vanio Mitev, Jyotsna Batra, Judith A Clements, Amanda Spurdle, Manuel R Teixeira, Paula Paulo, Sofia Maia, Hardev Pandha, Agnieszka Michael, Andrzej Kierzek, David S Karow, Ian G Mills, Ole A Andreassen, Anders M Dale, and PRACTICAL Consortium*. Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. *BMJ*, 360:j5757, January 2018.

[8] Juncheng Dai, Jun Lv, Meng Zhu, Yuzhuo Wang, Na Qin, Hongxia Ma, Yong-Qiao He, Ruoxin Zhang, Wen Tan, Jingyi Fan, Tianpei Wang, Hong Zheng, Qi Sun, Lijuan Wang, Mingtao Huang, Zijun Ge, Canqing Yu, Yu Guo, Tong-Min Wang, Jie Wang, Lin Xu, Weibing Wu, Liang Chen, Zheng Bian, Robin Walters, Iona Y Millwood, Xi-Zhao Li, Xin Wang, Rayjean J Hung, David C Christiani, Haiquan Chen, Mengyun Wang, Cheng Wang, Yue Jiang, Kexin Chen, Zhengming Chen, Guangfu Jin, Tangchun Wu, Dongxin Lin, Zhibin Hu, Christopher I Amos, Chen Wu, Qingyi Wei, Wei-Hua Jia, Liming Li, and Hongbing Shen. Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in chinese populations. *The Lancet Respiratory Medicine*, 7(10):881–891, October 2019.

[9] Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, and Sekar Kathiresan. Genome-wide polygenic scores for common diseases

identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.*, 50(9): 1219–1224, September 2018.

[10] James W Harrison, Divya Sri Priyanka Tallapragada, Alma Baptist, Seth A Sharp, Seema Bhaskar, Kalpana S Jog, Kashyap A Patel, Michael N Weedon, Giriraj R Chandak, Chittaranjan S Yajnik, and Richard A Oram. Type 1 diabetes genetic risk score is discriminative of diabetes in non-europeans: evidence from a study in india. *Sci. Rep.*, 10(1):9450, June 2020.

[11] Kristi Läll, Reedik Mägi, Andrew Morris, Andres Metspalu, and Krista Fischer. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genet. Med.*, 19(3):322–329, March 2017.

[12] Qian Zhang, Julia Sidorenko, Baptiste Couvy-Duchesne, Riccardo E Marioni, Margaret J Wright, Alison M Goate, Edoardo Marcora, Kuan-Lin Huang, Tenielle Porter, Simon M Laws, Colin L Masters, Ashley I Bush, Christopher Fowler, David Darby, Kelly Pertile, Carolina Restrepo, Blaine Roberts, Jo Robertson, Rebecca Rumble, Tim Ryan, Steven Collins, Christine Thai, Brett Trounson, Kate Lennon, Qiao-Xin Li, Fernanda Yevenes Ugarte, Irene Volitakis, Michael Vovos, Rob Williams, Jenalle Baker, Alyce Russell, Madeline Peretti, Lidija Milicic, Lucy Lim, Mark Rodrigues, Kevin Taddei, Tania Taddei, Eugene Hone, Florence Lim, Shane Fernandez, Stephanie Rainey-Smith, Steve Pedrini, Ralph Martins, James Doecke, Pierrick Bourgeat, Jurgen Fripp, Simon Gibson, Hugo Leroux, David Hanson, Vincent Dore, Ping Zhang, Samantha Burnham, Christopher C Rowe, Victor L Villemagne, Paul Yates, Sveltana Bozin Pejoska, Gareth Jones, David Ames, Elizabeth Cyarto, Nicola Lautenschlager, Kevin Barnham, Lesley Cheng, Andy Hill, Neil Killeen, Paul Maruff, Brendan Silbert, Belinda Brown, Harmid Sohrabi, Greg Savage, Michael Vacher, Perminder S Sachdev, Karen A Mather, Nicola J Armstrong, Anbupalam Thalamuthu, Henry Brodaty, Loic Yengo, Jian Yang, Naomi R Wray, Allan F McRae, Peter M Visscher, and Australian Imaging Biomarkers and Lifestyle (AIBL) Study. Risk prediction of late-onset

alzheimer's disease implies an oligogenic architecture. *Nat. Commun.*, 11(1):4799, September 2020.

[13] Leo P Sugrue and Rahul S Desikan. What are polygenic scores and why are they important? *JAMA*, 321(18):1820–1821, May 2019.

[14] Pradeep Natarajan, Robin Young, Nathan O Stitziel, Sandosh Padmanabhan, Usman Baber, Roxana Mehran, Samantha Sartori, Valentin Fuster, Dermot F Reilly, Adam Butterworth, Daniel J Rader, Ian Ford, Naveed Sattar, and Sekar Kathiresan. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation*, 135 (22):2091–2101, May 2017.

[15] Andrew Lee, Nasim Mavaddat, Amber N Wilcox, Alex P Cunningham, Tim Carver, Simon Hartley, Chantal Babb de Villiers, Angel Izquierdo, Jacques Simard, Marjanka K Schmidt, Fiona M Walter, Nilanjan Chatterjee, Montserrat Garcia-Closas, Marc Tischkowitz, Paul Pharoah, Douglas F Easton, and Antonis C Antoniou. BOADICEA: a comprehensive breast cancer risk prediction modelincorporating genetic and nongenetic risk factors. *Genet. Med.*, 21(8):1708–1718, August 2019.

[16] George Hindy, Krishna G Aragam, Kenney Ng, Mark Chaffin, Luca A Lotta, Aris Baras, Regeneron Genetics Center, Isabel Drake, Marju Orho-Melander, Olle Melander, Sekar Kathiresan, and Amit V Khera. Genome-Wide polygenic score, clinical risk factors, and Long-Term trajectories of coronary artery disease. *Arterioscler. Thromb. Vasc. Biol.*, 40(11):2738–2746, November 2020.

[17] Jin K. Park and Christine Y. Lu. Polygenic scores in the direct-to-consumer setting: Challenges and opportunities for a new era in consumer genetic testing. *Journal of Personalized Medicine*, 13(4), 2023. ISSN 2075-4426. doi: 10.3390/jpm13040573. URL https://www.mdpi.com/2075-4426/13/4/573.

[18] Linda Kachuri, Nilanjan Chatterjee, Jibril Hirbo, Daniel J Schaid, Iman Martin,

Iftikhar J Kullo, Eimear E Kenny, Bogdan Pasaniuc, Polygenic Risk Methods in Diverse Populations (PRIMED) Consortium Methods Working Group, John S Witte, and Tian Ge. Principles and methods for transferring polygenic risk scores across global populations. *Nat. Rev. Genet.*, August 2023.

[19] Jodell E. Linder, Aimee Allworth, Harris T. Bland, Pedro J. Caraballo, Rex L. Chisholm, Ellen Wright Clayton, David R. Crosslin, Ozan Dikilitas, Alanna DiVietro, Edward D. Esplin, Sophie Forman, Robert R. Freimuth, Adam S. Gordon, Richard Green, Maegan V. Harden, Ingrid A. Holm, Gail P. Jarvik, Elizabeth W. Karlson, Sofia Labrecque, Niall J. Lennon, Nita A. Limdi, Kathleen F. Mittendorf, Shawn N. Murphy, Lori Orlando, Cynthia A. Prows, Luke V. Rasmussen, Laura Rasmussen-Torvik, Robb Rowley, Konrad Teodor Sawicki, Tara Schmidlen, Shannon Terek, David Veenstra, Digna R. Velez Edwards, Devin Absher, Noura S. Abul-Husn, Jorge Alsip, Hana Bangash, Mark Beasley, Jennifer E. Below, Eta S. Berner, James Booth, Wendy K. Chung, James J. Cimino, John Connolly, Patrick Davis, Beth Devine, Stephanie M. Fullerton, Candace Guiducci, Melissa L. Habrat, Heather Hain, Hakon Hakonarson, Margaret Harr, Eden Haverfield, Valentina Hernandez, Christin Hoell, Martha Horike-Pyne, George Hripcsak, Marguerite R. Irvin, Christopher Kachulis, Dean Karavite, Eimear E. Kenny, Atlas Khan, Krzysztof Kiryluk, Bruce Korf, Leah Kottyan, Iftikhar J. Kullo, Katie Larkin, Cong Liu, Edyta Malolepsza, Teri A. Manolio, Thomas May, Elizabeth M. McNally, Frank Mentch, Alexandra Miller, Sean D. Mooney, Priyanka Murali, Brenda Mutai, Naveen Muthu, Bahram Namjou, Emma F. Perez, Megan J. Puckelwartz, Tejinder Rakhra-Burris, Dan M. Roden, Elisabeth A. Rosenthal, Seyedmohammad Saadatagah, Maya Sabatello, Dan J. Schaid, Baergen Schultz, Lynn Seabolt, Gabriel Q. Shaibi, Richard R. Sharp, Brian Shirts, Maureen E. Smith, Jordan W. Smoller, Rene Sterling, Sabrina A. Suckiel, Jeritt Thayer, Hemant K. Tiwari, Susan B. Trinidad, Theresa Walunas, Wei-Qi Wei, Quinn S. Wells, Chunhua Weng, Georgia L. Wiesner, Ken Wiley, Adam Gordon, Agboade Sobowale, Aimee Allworth, Akshar Patel, Alanna DiVietro, Alanna Strong, Alborz Sherafati,

Alborz Sherfati, Alex Bick, Alexandra Miller, Alka Chandel, Alyssa Rosenthal, Amit Khera, Amy Kontorovich, Andrew Beck, Andy Beck, Angelica Espinoza, Anna Lewis, Anya Prince, Atlas Khan, Ayuko Iverson, Bahram Namjou Khales, Barbara Benoit, Becca Hernan, Ben Kallman, Ben Kerman, Ben Shoemaker, Benjamin Satterfield, Beth Devine, Bethany Etheridge, Blake Goff, Bob Freimuth, Bob Grundmeier, Brenae Collier, Brenda Mutai, Brett Harnett, Brian Chang, Brian Piening, Brittney Davis, Bruce Korf, Candace Patterson, Carmen Demetriou, Casey Ta, Catherine Hammack, Catrina Nelson, Caytie Gascoigne, Chad Dorn, Chad Moretz, Chris Kachulis, Christie Hoell, Christine Cowles, Christoph Lange, Chunhua Weng, Cindy Prows, Cole Brokamp, Cong Liu, Courtney Scherr, Crystal Gonzalez, Cynthia Ramirez, Daichi Shimbo, Dan Roden, Daniel Schaid, Dave Kaufman, David Crosslin, David Kochan, David Veenstra, Davinder Singh, Dean Karavite, Debbie Abrams, Devin Absher, Digna Velez Edwards, Eden Haverfield, Eduardo Morales, Edward Esplin, Edyta Malolepsza, Ehsan Alipour, Eimear Kenny, Elisabeth Rosenthal, Eliza Duvall, Elizabeth McNally, Elizabeth Bhoj, Elizabeth Cohn, Elizabeth Hibler, Elizabeth Karlson, Ellen Clayton, Emily Chesnut, Emily DeFranco, Emily Gallagher, Emily Soper, Emma Perez, Erin Cash, Eta Berner, Fei Wang, Firas Wehbe, Francisco Ricci, Frank Mentch, Gabriel Shaibi, Gail Jarvik, George Hahn, George Hripcsak, Georgia Wiesner, Gillian Belbin, Gio Davogustto, Girish Nadkarni, Haijun Qiu, Hakon Hakonarson, Hana Bangash, Hannah Beasley, Hao Liu, Heide Aungst, Hemant Tiwari, Hillary Duckham, Hope Thomas, Iftikhar Kullo, Ingrid Holm, Isabelle Allen, Iuliana Ionita-Laza, Jacklyn Hellwege, Jacob Petrzelka, Jacqueline Odgis, Jahnavi Narula, Jake Petrzelka, Jalpa Patel, James Cimino, James Meigs, James Snyder, Janet Olson, Janet Zahner, Jeff Pennington, Jen Pacheco, Jennifer Allen Pacheco, Jennifer Morse, Jeremy Corsmo, Jeritt Thayer, Jim Cimino, Jingheng Chen, Jocelyn Fournier, Jodell Jackson, Joe Glessner, Joel Pacyna, Johanna Smith, John Connolly, John Lynch, John Shelley, Jonathan Mosley, Jordan Nestor, Jordan Smoller, Jorge Alsip, Joseph Kannry, Joseph Sutton, Josh Peterson, Joshua Smith, Julia Galasso, Julia Smith, Julia Wynn, Justin Gundelach, Justin Starren, Karmel Choi, Kate Mittendorf, Katherine Anderson, Katherine Bonini, Kathleen Leppig, Kathleen Muenzen, Katie Larkin, Kelsey Stuttgen, Ken Wiley, Kenny Nguyen, Kevin Dufendach, Kiley Atkins, Konrad Sawicki, Kristjan Norland, Krzysztof Kiryluk, Laura Beskow, Laura Rasmussen-Torvik, Leah Kottyan, Li Hsu, Lifeng Tian, Lisa Mahanta, Lisa Martin, Lisa Wang, Lizbeth Gomez, Lorenzo Thompson, Lori Orlando, Lucas Richter, Luke Rasmussen, Lynn Petukhova, Lynn Seabolt, Madison O'Brien, Maegan Harden, Malia Fullerton, Margaret Harr, Mark Beasley, Marta Guindo, Martha Horike, Martha Horike-Pyne, Marwah Abdalla, Marwan Hamed, Mary Beth Terry, Mary Maradik, Matt Wyatt, Matthew Davis, Matthew Lebo, Maureen Smith, Maya del Rosario, Maya Sabatello, Meckenzie Behr, Meg Roy-Puckelwartz, Mel Habrat, Melanie Myers, Meliha Yetisgen, Merve Iris, Michael DaSilva, Michael Preuss, Michelle McGowan, Mingjian Shi, Minoli Perera, Minta Thomas, Mitch Elkind, Mohammad Abbass, Mohammad Saadatagah, Molly Hess, Molly Maradik, Nataraja "RJ" Vaitinadin, Nataraja Vaitinadin, Naveen Muthu, Neil Netherly, Niall Lennon, Ning Shang, Nita Limdi, Noah Forrest, Noheli Romero, Nora Robinson, Noura Abul-Husn, Omar Elsekaily, Ozan Dikilitas, Patricia Kovatch, Patrick Davis, Paul Appelbaum, Paul Francaviglia, Paul O'Reilly, Paulette Chandler, Pedro Caraballo, Peter Tarczy-Hornoch, Pierre Shum, Priya Marathe, Priyanka Murali, Qiping Feng, Quinn Wells, Rachel Atchley, Radhika Narla, Rene Barton, Rene Sterling, Rex Chisholm, Richard Green, Richard Sharp, Riki Peters, Rita Kukafka, Robb Rowley, Robert Freimuth, Robert Green, Robert Winter, Roger Mueller, Ruth Loos, Ryan Irvin, Sabrina Suckiel, Sajjad Hussain, Samer Sharba, Sandy Aronson, Sarah Jones, Sarah Knerr, Scott Nigbur, Scott Weiss, Sean Mooney, Shannon Terek, Sharon Aufox, Sharon Nirenberg, Shawn Murphy, Sheila O'Byrne, Shing Wang (Sam) Choi, Sienna Aguilar, S.T. Bland, Stefanie Rodrigues, Stephanie Ledbetter, Stephanie Rutledge, Stuart James Booth, Su Xian, Susan Brown Trinidad, Suzanne Bakken, Tara Schmidlen, Tejinder Rakhra-Burris, Teri Manolio, Tesfaye Mersha, Theresa Walunas, Thevaa Chandereng, Thomas May, Tian Ge, Todd Edwards, Tom Kaszemacher, Valentina Hernandez, Valerie Willis, Vemi Desai, Vimi Desai, Virginia Lorenzi, Vivian Gainer, Wei-Qi Wei, Wendy Chung, Wu-Chen Su, Xiao Chang, Yiqing Zhao,

Yuan Luo, Yufeng Shen, and Josh F. Peterson. Returning integrated genomic risk and clinical recommendations: The emerge study. *Genetics in Medicine*, 25(4): 100006, 2023. ISSN 1098-3600. doi: https://doi.org/10.1016/j.gim.2023.100006. URL https://www.sciencedirect.com/science/article/pii/S1098360023000023.

[20] Nina Mars, Sini Kerminen, Yen-Chen A. Feng, Masahiro Kanai, Kristi Läll, Laurent F. Thomas, Anne Heidi Skogholt, Pietro della Briotta Parolo, Benjamin M. Neale, Jordan W. Smoller, Maiken E. Gabrielsen, Kristian Hveem, Reedik Mägi, Koichi Matsuda, Yukinori Okada, Matti Pirinen, Aarno Palotie, Andrea Ganna, Alicia R. Martin, and Samuli Ripatti. Genome-wide risk prediction of common diseases across ancestries in one million people. *Cell Genomics*, 2(4):100118, 2022. ISSN 2666-979X. doi: https://doi.org/10.1016/j.xgen.2022.100118. URL https://www.sciencedirect.com/science/article/pii/S2666979X22000428.

[21] Polygenic Risk Score Task Force of the International Common Disease Alliance. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat. Med.*, 27(11):1876–1884, November 2021.

[22] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.*, 100(4):635–649, April 2017.

[23] L Duncan, H Shen, B Gelaye, J Meijsen, K Ressler, M Feldman, R Peterson, and B Domingue. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.*, 10(1):1–9, July 2019.

[24] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.*, 51(4):584–591, March 2019.

[25] Ali Torkamani, Nathan E Wineinger, and Eric J Topol. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.*, 19(9):581–590, May 2018.

[26] Ruowang Li, Yong Chen, Marylyn D Ritchie, and Jason H Moore. Electronic health records and polygenic risk scores for predicting disease risk. *Nat. Rev. Genet.*, 21(8): 493–502, March 2020.

[27] Nilanjan Chatterjee, Jianxin Shi, and Montserrat García-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17(7):392–406, 2016.

[28] Naomi R Wray, Sang Hong Lee, Divya Mehta, Anna A E Vinkhuyzen, Frank Dudbridge, and Christel M Middeldorp. Research review: Polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry*, 55(10):1068–1087, October 2014.

[29] Lars G Fritsche, Stephen B Gruber, Zhenke Wu, Ellen M Schmidt, Matthew Zawistowski, Stephanie E Moser, Victoria M Blanc, Chad M Brummett, Sachin Kheterpal, Gonçalo R Abecasis, and Bhramar Mukherjee. Association of polygenic risk scores for multiple cancers in a phenome-wide study: Results from the michigan genomics initiative. *Am. J. Hum. Genet.*, 102(6):1048–1061, June 2018.

[30] Christopher C Chang, Carson C Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4:7, February 2015.

[31] Timothy Shin Heng Mak, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.*, 41(6):469–480, September 2017.

[32] Doug Speed and David J Balding. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.*, 24(9):1550–1557, September 2014.

[33] Florian Privé, Julyan Arbel, and Bjarni J Vilhjálmsson. LDpred2: better, faster, stronger. *Bioinformatics*, 36(22-23):5424–5431, April 2021.

[34] Gerhard Moser, Sang Hong Lee, Ben J Hayes, Michael E Goddard, Naomi R Wray, and Peter M Visscher. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet.*, 11(4):e1004969, April 2015.

[35] Bjarni J Vilhjálmsson, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, Tristan Hayeck, Hong-Hee Won, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study, Sekar Kathiresan, Michele Pato, Carlos Pato, Rulla Tamimi, Eli Stahl, Noah Zaitlen, Bogdan Pasaniuc, Gillian Belbin, Eimear E Kenny, Mikkel H Schierup, Philip De Jager, Nikolaos A Patsopoulos, Steve McCarroll, Mark Daly, Shaun Purcell, Daniel Chasman, Benjamin Neale, Michael Goddard, Peter M Visscher, Peter Kraft, Nick Patterson, and Alkes L Price. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.*, 97(4):576–592, October 2015.

[36] Luke R Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E Kemper, Huanwei Wang, Zhili Zheng, Reedik Magi, Tõnu Esko, Andres Metspalu, Naomi R Wray, Michael E Goddard, Jian Yang, and Peter M Visscher. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nat. Commun.*, 10(1):1–11, November 2019.

[37] Miriam S Udler, Jonathan Tyrer, and Douglas F Easton. Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. *Genet. Epidemiol.*, 34(5):463–468, July 2010.

[38] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.*, 19(8):491–504, August 2018.

[39] Michael Lynch, Bruce Walsh, and Others. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998.

[40] Daniel Sorensen and Daniel Gianola. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer New York.

[41] Gregor Gorjanc, Piter Bijma, and John M Hickey. Reliability of pedigree-based and genomic evaluations in selected populations. *Genet. Sel. Evol.*, 47:65, August 2015.

[42] C R Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2):423–447, June 1975.

[43] G Su, B Guldbrandtsen, V R Gregersen, and M S Lund. Preliminary investigation on reliability of genomic estimated breeding values in the danish holstein population. *J. Dairy Sci.*, 93(3):1175–1183, March 2010.

[44] I Misztal and G R Wiggans. Approximation of prediction error variance in Large-Scale animal models. *J. Dairy Sci.*, 71:27–32, June 1988.

[45] K Meyer. Approximate accuracy of genetic evaluation under an animal model. *Livestock Production Science*, 21(2):87–100, February 1989.

[46] J Jamrozik, L R Schaeffer, and G B Jansen. Approximate accuracies of prediction from random regression models. *Livestock Production Science*, 66(1):85–92, September 2000.

[47] B Tier and K Meyer. Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. *J. Anim. Breed. Genet.*, 121(2):77–89, April 2004.

[48] John M Hickey, Roel F Veerkamp, Mario P L Calus, Han A Mulder, and Robin Thompson. Estimation of prediction error variances via monte carlo sampling methods using different formulations of the prediction error variance. *Genet. Sel. Evol.*, 41(1):23, February 2009.

[49] Simon Klau, Marie-Laure Martin-Magniette, Anne-Laure Boulesteix, and Sabine Hoffmann. Sampling uncertainty versus method uncertainty: A general framework with applications to omics biomarker selection. *Biom. J.*, 62(3):670–687, May 2020.

[50] P Bycott and J Taylor. A comparison of smoothing techniques for CD4 data measured with error in a time-dependent cox proportional hazards model. *Stat. Med.*, 17(18): 2061–2077, September 1998.

[51] Jaime E Hart, Xiaomei Liao, Biling Hong, Robin C Puett, Jeff D Yanosky, Helen Suh, Marianthi-Anna Kioumourtzoglou, Donna Spiegelman, and Francine Laden. The association of long-term exposure to PM2.5 on all-cause mortality in the nurses' health study and the impact of measurement-error correction. *Environ. Health*, 14:38, May 2015.

[52] Naomi R Wray, Jian Yang, Ben J Hayes, Alkes L Price, Michael E Goddard, and Peter M Visscher. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.*, 14(7):507–515, July 2013.

[53] Kelsey E Grinde, Qibin Qi, Timothy A Thornton, Simin Liu, Aladdin H Shadyab, Kei Hang K Chan, Alexander P Reiner, and Tamar Sofer. Generalizing polygenic risk scores from europeans to Hispanics/Latinos. *Genet. Epidemiol.*, 43(1):50–62, February 2019.

[54] Jian Zeng, Ronald de Vlaming, Yang Wu, Matthew R Robinson, Luke R Lloyd-Jones, Loic Yengo, Chloe X Yap, Angli Xue, Julia Sidorenko, Allan F McRae, Joseph E Powell, Grant W Montgomery, Andres Metspalu, Tonu Esko, Greg Gibson, Naomi R Wray, Peter M Visscher, and Jian Yang. Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.*, 50(5):746–753, May 2018.

[55] Julian J Faraway. Practical regression and anova using R. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.394.2244&rep=rep1&type=pdf, 2002. Accessed: 2021-3-26.

[56] Frank Dudbridge. Criteria for evaluating risk prediction of multiple outcomes. *Stat. Methods Med. Res.*, 29(12):3492–3510, December 2020.

[57] Kathleen F Kerr, Zheyu Wang, Holly Janes, Robyn L McClelland, Bruce M Psaty, and Margaret S Pepe. Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology*, 25(1):114–121, January 2014.

[58] D R Cox. Regression models and life-tables. *J. R. Stat. Soc.*, 34(2):187–202, January 1972.

[59] Tian Ge, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nat. Commun.*, 10(1):1776, April 2019.

[60] Yiming Hu, Qiongshi Lu, Ryan Powles, Xinwei Yao, Can Yang, Fang Fang, Xinran Xu, and Hongyu Zhao. Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput. Biol.*, 13(6):e1005589, June 2017.

[61] Shing Wan Choi and Paul F O'Reilly. PRSice-2: Polygenic risk score software for biobank-scale data. *Gigascience*, 8(7), July 2019.

[62] Karoline B Kuchenbaecker, Lesley McGuffog, Daniel Barrowdale, Andrew Lee, Penny Soucy, Joe Dennis, Susan M Domchek, Mark Robson, Amanda B Spurdle, Susan J Ramus, Nasim Mavaddat, Mary Beth Terry, Susan L Neuhausen, Rita Katharina Schmutzler, Jacques Simard, Paul D P Pharoah, Kenneth Offit, Fergus J Couch, Georgia Chenevix-Trench, Douglas F Easton, and Antonis C Antoniou. Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. *J. Natl. Cancer Inst.*, 109(7), July 2017.

[63] Akl C Fahed, Minxian Wang, Julian R Homburger, Aniruddh P Patel, Alexander G Bick, Cynthia L Neben, Carmen Lai, Deanna Brockman, Anthony Philippakis, Patrick T Ellinor, Christopher A Cassa, Matthew Lebo, Kenney Ng, Eric S Lander, Alicia Y Zhou, Sekar Kathiresan, and Amit V Khera. Polygenic background modifies

penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.*, 11(1): 3635, August 2020.

[64] Ali Pazokitoroudi, Alec M Chiu, Kathryn S Burch, Bogdan Pasaniuc, and Sriram Sankararaman. Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data. *Am. J. Hum. Genet.*, 108(5):799–808, May 2021.

[65] V Hivert, J Sidorenko, F Rohart, M E Goddard, J Yang, and others. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *bioRxiv*, 2020.

[66] Andy Dahl, Khiem Nguyen, Na Cai, Michael J Gandal, Jonathan Flint, and Noah Zaitlen. A robust method uncovers significant Context-Specific heritability in diverse complex traits. *Am. J. Hum. Genet.*, 106(1):71–91, January 2020.

[67] Huanwei Wang, Futao Zhang, Jian Zeng, Yang Wu, Kathryn E Kemper, Angli Xue, Min Zhang, Joseph E Powell, Michael E Goddard, Naomi R Wray, Peter M Visscher, Allan F McRae, and Jian Yang. Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK biobank. *Sci Adv*, 5(8):eaaw3538, August 2019.

[68] Genevieve L Wojcik, Mariaelisa Graff, Katherine K Nishimura, Ran Tao, Jeffrey Haessler, Christopher R Gignoux, Heather M Highland, Yesha M Patel, Elena P Sorokin, Christy L Avery, Gillian M Belbin, Stephanie A Bien, Iona Cheng, Sinead Cullina, Chani J Hodonsky, Yao Hu, Laura M Huckins, Janina Jeff, Anne E Justice, Jonathan M Kocarnik, Unhee Lim, Bridget M Lin, Yingchang Lu, Sarah C Nelson, Sung-Shim L Park, Hannah Poisner, Michael H Preuss, Melissa A Richard, Claudia Schurmann, Veronica W Setiawan, Alexandra Sockell, Karan Vahi, Marie Verbanck, Abhishek Vishnu, Ryan W Walker, Kristin L Young, Niha Zubair, Victor Acuña-Alonso, Jose Luis Ambite, Kathleen C Barnes, Eric Boerwinkle, Erwin P Bottinger, Carlos D Bustamante, Christian Caberto, Samuel Canizales-Quinteros, Matthew P

Conomos, Ewa Deelman, Ron Do, Kimberly Doheny, Lindsay Fernández-Rhodes, Myriam Fornage, Benyam Hailu, Gerardo Heiss, Brenna M Henn, Lucia A Hindorff, Rebecca D Jackson, Cecelia A Laurie, Cathy C Laurie, Yuqing Li, Dan-Yu Lin, Andres Moreno-Estrada, Girish Nadkarni, Paul J Norman, Loreall C Pooler, Alexander P Reiner, Jane Romm, Chiara Sabatti, Karla Sandoval, Xin Sheng, Eli A Stahl, Daniel O Stram, Timothy A Thornton, Christina L Wassel, Lynne R Wilkens, Cheryl A Winkler, Sachi Yoneyama, Steven Buyske, Christopher A Haiman, Charles Kooperberg, Loic Le Marchand, Ruth J F Loos, Tara C Matise, Kari E North, Ulrike Peters, Eimear E Kenny, and Christopher S Carlson. Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 570(7762):514–518, June 2019.

[69] Ying Wang, Jing Guo, Guiyan Ni, Jian Yang, Peter M Visscher, and Loic Yengo. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.*, 11(1):3865, July 2020.

[70] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.*, 51(4):584–591, April 2019.

[71] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

[72] Bradley Efron and R J Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1 edition, January 1993.

[73] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, and Schizophrenia Working Group of the Psychiatric Genomics Consortium. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, 47(3):291–295, March 2015.

[74] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018.

[75] Hakhamanesh Mostafavi, Arbel Harpak, Ipsita Agarwal, Dalton Conley, Jonathan K Pritchard, and Molly Przeworski. Variable prediction accuracy of polygenic scores within an ancestry group. *Elife*, 9:e48376, January 2020.

[76] Ying Wang, Kristin Tsuo, Masahiro Kanai, Benjamin M Neale, and Alicia R Martin. Challenges and opportunities for developing more generalizable polygenic risk scores. *Annu. Rev. Biomed. Data Sci.*, 5(1):293–320, August 2022.

[77] Ruth Johnson, Yi Ding, Vidhya Venkateswaran, Arjun Bhattacharya, Kristin Boulier, Alec Chiu, Sergey Knyazev, Tommer Schwarz, Malika Freund, Lingyu Zhan, Kathryn S Burch, Christa Caggiano, Brian Hill, Nadav Rakocz, Brunilda Balliu, Christopher T Denny, Jae Hoon Sul, Noah Zaitlen, Valerie A Arboleda, Eran Halperin, Sriram Sankararaman, Manish J Butte, Clara Lajonchere, Daniel H Geschwind, and Bogdan Pasaniuc. Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS community health initiative. *Genome Med.*, 14(1): 1–23, September 2022.

[78] Anna C F Lewis, Santiago J Molina, Paul S Appelbaum, Bege Dauda, Anna Di Rienzo, Agustin Fuentes, Stephanie M Fullerton, Nanibaa' A Garrison, Nayanika Ghosh, Evelynn M Hammonds, David S Jones, Eimear E Kenny, Peter Kraft, Sandra S-J Lee, Madelyn Mauro, John Novembre, Aaron Panofsky, Mashaal Sohail, Benjamin M Neale, and Danielle S Allen. Getting genetic ancestry right for science and society. *Science*, 376(6590):250–252, 2022.

[79] Iftikhar J Kullo, Cathryn M Lewis, Michael Inouye, Alicia R Martin, Samuli Ripatti,

and Nilanjan Chatterjee. Polygenic scores in biomedical research. *Nat. Rev. Genet.*, March 2022.

[80] Diana O Perkins, Loes Olde Loohuis, Jenna Barbee, John Ford, Clark D Jeffries, Jean Addington, Carrie E Bearden, Kristin S Cadenhead, Tyrone D Cannon, Barbara A Cornblatt, Daniel H Mathalon, Thomas H McGlashan, Larry J Seidman, Ming Tsuang, Elaine F Walker, and Scott W Woods. Polygenic risk score contribution to psychosis prediction in a target population of persons at clinical high risk. *Am. J. Psychiatry*, 177(2):155–163, February 2020.

[81] Fatima Rodriguez, Sukyung Chung, Manuel R Blum, Adrien Coulet, Sanjay Basu, and Latha P Palaniappan. Atherosclerotic cardiovascular disease risk prediction in disaggregated asian and hispanic subgroups using electronic health records. *J. Am. Heart Assoc.*, 8(14):e011874, July 2019.

[82] Yon Ho Jee, Chi Gao, Jihye Kim, Seho Park, Sun Ha Jee, and Peter Kraft. Validating breast cancer risk prediction models in the korean cancer prevention Study-II biobank. *Cancer Epidemiol. Biomarkers Prev.*, 29(6):1271–1277, June 2020.

[83] Paul D Myers, Kenney Ng, Kristen Severson, Uri Kartoun, Wangzhi Dai, Wei Huang, Frederick A Anderson, and Collin M Stultz. Identifying unreliable predictions in clinical risk models. *NPJ Digit Med*, 3:8, January 2020.

[84] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[85] T H Meuwissen, B J Hayes, and M E Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, April 2001.

[86] P M VanRaden. Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91 (11):4414–4423, November 2008.

[87] Mike Goddard. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 136(2):245–257, June 2009.

[88] M E Goddard, B J Hayes, and T H E Meuwissen. Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed. Genet.*, 128(6): 409–421, December 2011.

[89] Hans D Daetwyler, Beatriz Villanueva, and John A Woolliams. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*, 3(10):e3395, October 2008.

[90] Peter M Visscher, Jian Yang, and Michael E Goddard. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by yang et al. (2010). *Twin Res. Hum. Genet.*, 13(6):517–524, December 2010.

[91] Yvonne C J Wientjes, Roel F Veerkamp, Piter Bijma, Henk Bovenhuis, Chris Schrooten, and Mario P L Calus. Empirical and deterministic accuracies of across-population genomic prediction. *Genet. Sel. Evol.*, 47(1):5, February 2015.

[92] Douglas Scott Falconer. *Introduction to Quantitative Genetics*. Longman Scientific & Technical, 1989.

[93] Bruce Walsh and Michael Lynch. Evolution and selection of quantitative traits. In *Evolution and Selection of Quantitative Traits*. Oxford University Press, 1 edition, September 2018.

[94] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, 42(7):565–569, July 2010.

[95] H Ben Zaabza, E A Mäntysaari, and I Strandén. Using monte carlo method to include polygenic effects in calculation of SNP-BLUP model reliability. *J. Dairy Sci.*, 103(6): 5170–5182, June 2020.

[96] M Pszczola, T Strabel, H A Mulder, and M P L Calus. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.*, 95(1):389–400, January 2012.

[97] Yvonne C J Wientjes, Roel F Veerkamp, and Mario P L Calus. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*, 193(2):621–631, February 2013.

[98] D Habier, R L Fernando, and J C M Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397, December 2007.

[99] S Hong Lee, W M Shalanee P Weerasinghe, Naomi R Wray, Michael E Goddard, and Julius H J van der Werf. Using information of relatives in genomic prediction to apply effective stratified medicine. *Sci. Rep.*, 7:42091, February 2017.

[100] Buu Truong, Xuan Zhou, Jisu Shin, Jiuyong Li, Julius H J van der Werf, Thuc D Le, and S Hong Lee. Efficient polygenic risk scores for biobank scale data by exploiting phenotypes from inferred relatives. *Nat. Commun.*, 11(1):3074, June 2020.

[101] Florian Privé, Hugues Aschard, Shai Carmi, Lasse Folkersen, Clive Hoggart, Paul F O'Reilly, and Bjarni J Vilhjálmsson. Portability of 245 polygenic scores when derived from the UK biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.*, 109(1):12–23, January 2022.

[102] Bárbara D Bitarello and Iain Mathieson. Polygenic scores for height in admixed populations. *G3*, 10(11):4027–4036, November 2020.

[103] Graham Coop. Genetic similarity and genetic ancestry groups. July 2022.

[104] Iain Mathieson and Aylwyn Scally. What is ancestry? *PLoS Genet.*, 16(3):e1008624, March 2020.

[105] Talia Krainc and Agustín Fuentes. Genetic ancestry in precision medicine is reshaping the race debate. *Proceedings of the National Academy of Sciences*, 119(12): e2203033119, 2022.

[106] Gillian M Belbin, Sinead Cullina, Stephane Wenric, Emily R Soper, Benjamin S Glicksberg, Denis Torre, Arden Moscati, Genevieve L Wojcik, Ruhollah Shemirani, Noam D Beckmann, Ariella Cohain, Elena P Sorokin, Danny S Park, Jose-Luis Ambite, Steve Ellis, Adam Auton, CBIPM Genomics Team, Regeneron Genetics Center, Erwin P Bottinger, Judy H Cho, Ruth J F Loos, Noura S Abul-Husn, Noah A Zaitlen, Christopher R Gignoux, and Eimear E Kenny. Toward a fine-scale population health monitoring system. *Cell*, 184(8):2068–2083.e11, April 2021.

[107] Shoa L Clarke, Rose D L Huang, Austin T Hilliard, Catherine Tcheandjieu, Julie Lynch, Scott M Damrauer, Kyong-Mi Chang, Philip S Tsao, and Themistocles L Assimes. Race and ethnicity stratification for polygenic risk score analyses may mask disparities in hispanics. *Circulation*, 146(3):265–267, July 2022.

[108] Ruth Johnson, Yi Ding, Arjun Bhattacharya, Sergey Knyazev, Alec Chiu, Clara Lajonchere, Daniel H Geschwind, and Bogdan Pasaniuc. The UCLA ATLAS community health initiative: Promoting precision health research in a diverse biobank. *Cell Genomics*, 3(1):100243, January 2023.

[109] Yi Ding, Kangcheng Hou, Kathryn S Burch, Sandra Lapinska, Florian Privé, Bjarni Vilhjálmsson, Sriram Sankararaman, and Bogdan Pasaniuc. Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. *Nat. Genet.*, 54(1):30–39, December 2021.

[110] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38(8):904–909, July 2006.

[111] David H Alexander and Kenneth Lange. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, 12:246, June 2011.

[112] David Reich, Michael A Nalls, W H Linda Kao, Ermeg L Akylbekova, Arti Tandon, Nick Patterson, James Mullikin, Wen-Chi Hsueh, Ching-Yu Cheng, Josef Coresh, Eric Boerwinkle, Man Li, Alicja Waliszewska, Julie Neubauer, Rongling Li, Tennille S Leak, Lynette Ekunwe, Joe C Files, Cheryl L Hardy, Joseph M Zmuda, Herman A Taylor, Elad Ziv, Tamara B Harris, and James G Wilson. Reduced neutrophil count in people of african descent is due to a regulatory variant in the duffy antigen receptor for chemokines gene. *PLoS Genet.*, 5(1):e1000360, January 2009.

[113] Marco Scutari, Ian Mackay, and David Balding. Using genetic distance to infer the accuracy of genomic prediction. *PLoS Genet.*, 12(9):e1006288, September 2016.

[114] Segun Fatumo, Tinashe Chikowore, Ananyo Choudhury, Muhammad Ayub, Alicia R Martin, and Karoline Kuchenbaecker. A roadmap to increase diversity in genomic studies. *Nat. Med.*, 28(2):243–250, February 2022.

[115] Yunfeng Ruan, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, Zhenglin Guo, Lin He, Akira Sawa, Alicia R Martin, Shengying Qin, Hailiang Huang, and Tian Ge. Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.*, 54(5):573–580, May 2022.

[116] Jeffrey P Spence, Nasa Sinnott-Armstrong, Themistocles L Assimes, and Jonathan K Pritchard. A flexible modeling and inference framework for estimating variant effect sizes from GWAS summary statistics. April 2022.

[117] Haoyu Zhang, Jianan Zhan, Jin Jin, Jingning Zhang, Thomas U Ahearn, Zhi Yu, Jared O'Connell, Yunxuan Jiang, Tony Chen, Montserrat Garcia-Closas, Xihong Lin, Bertram L Koelsch, and Nilanjan Chatterjee. Novel methods for multi-ancestry polygenic prediction and their evaluations in 3.7 million individuals of diverse ancestry. April 2022.

[118] Alicia R Martin, Rocky E Stroud, 2nd, Tamrat Abebe, Dickens Akena, Melkam Alemayehu, Lukoye Atwoli, Sinéad B Chapman, Katelyn Flowers, Bizu Gelaye, Stella Gichuru, Symon M Kariuki, Sam Kinyanjui, Kristina J Korte, Nastassja Koen, Karestan C Koenen, Charles R J C Newton, Ana Maria Olivares, Sam Pollock, Kristianna Post, Ilina Singh, Dan J Stein, Solomon Teferra, Zukiswa Zingela, and Lori B Chibnik. Increasing diversity in genomics requires investment in equitable partnerships and capacity building. *Nat. Genet.*, 54(6):740–745, June 2022.

[119] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

[120] Kangcheng Hou, Kathryn S Burch, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, Yue Wu, Sriram Sankararaman, and Bogdan Pasaniuc. Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat. Genet.*, 51(8):1244–1251, July 2019.

[121] Qianqian Zhang, Florian Privé, Bjarni Vilhjálmsson, and Doug Speed. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.*, 12(1):4192, July 2021.

[122] Gad Abraham, Yixuan Qiu, and Michael Inouye. FlashPCA2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics*, 33(17):2776–2778, September 2017.

[123] Florian Privé, Hugues Aschard, Andrey Ziyatdinov, and Michael G B Blum. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*, 34(16):2781–2787, August 2018.