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

Yi Ding[1]*, Kangcheng Hou[1]*, Kathryn S. Burch[1], Sandra Lapinska[2], Florian Privé[3], Bjarni Vilhjálmsson[3], Sriram Sankararaman[1,4,5,6], Bogdan Pasaniuc[1,5,6,7]

1. Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA 90095
2. Department of Statistics and Data Science, Cornell University, Ithaca, NY 14853
3. Department of Economics and Business Economics, National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark
4. Department of Computer Science, UCLA, Los Angeles, CA 90095
5. Department of Computational Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095
6. Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095
7. Department of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095

* - contributed equally

Correspondence: Y.D. (yiding920@ucla.edu); K.H. (houkc@ucla.edu); B.P. (pasaniuc@ucla.edu).

Abstract

Large-scale genome-wide association studies have enabled polygenic risk scores (PRS), which estimate the genetic value of an individual for a given trait. Since PRS accuracy is typically assessed using cohort-level metrics (e.g., R^2), uncertainty in PRS estimates at individual level remains underexplored. Here we show that Bayesian PRS methods can estimate variance of individual PRS and can yield well-calibrated credible intervals for the genetic value of a single individual. For real traits in the UK Biobank (N=291,273 unrelated "white British") we observe large variance in individual PRS estimates which impacts interpretation of PRS-based stratification; for example, averaging across 11 traits, only 1.8% (s.d. 2.4%) of individuals with PRS point estimates in the top decile have their entire 95% credible intervals fully contained in the top decile. To account for this uncertainty, we propose a probabilistic approach to PRS-based stratification that estimates the probability of an individual's genetic value to be above a prespecified threshold. Our results showcase the importance of incorporating uncertainty in individual PRS estimates into subsequent analyses.

Introduction

Polygenic risk scores (PRS) have emerged as the main approach for predicting the genetic component of an individual's phenotype and/or common-disease risk (i.e. genetic value, GV) from large-scale genome-wide association studies (GWAS). Several studies have demonstrated the utility of PRS as estimators of genetic values in genomic research and clinical decision-making—for example, in stratifying patients[1], delivering personalized treatment[2], predicting disease risk[3], forecasting disease trajectories[4,5], and studying shared etiology among traits[6,7]. Increasingly large GWAS sample sizes have improved the predictive value of PRS for several complex traits and diseases[8,9] including breast cancer[3,10], prostate cancer[11], lung cancer[12], coronary artery disease[13], obesity[4], type 1 diabetes[14], type 2 diabetes[15], and Alzheimer's disease[16], thus paving the way for PRS-informed precision medicine.

Under a linear additive genetic model, an individual's genetic value (GV; the estimand of interest for PRS) is the sum of individual's dosage genotypes at causal variants (encoded as the number of copies of the effect allele) 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: (1) pruning/clumping and thresholding (P+T) approaches, which implicitly 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[17,18]; (2) methods that account for LD through regularization of effect sizes, including lassosum[19] and BLUP prediction[20 21]; and (3) Bayesian approaches that explicitly model causal effects and LD to infer the posterior distribution of causal effect sizes to be used as weights in PRS estimation[22–24].

Both the bias and variability of a PRS estimator are critical to assessing its practical utility. Given that most PRS methods select variants (predictors) and estimate their effect sizes, there are two main sources of uncertainty: (1) which variants are causal (i.e. have non-zero effects) and (2) statistical noise in the causal effect estimates due to finite the sample size of GWAS training data. Both types of uncertainty increase as the strength of LD in a region increases and as the sample size of the GWAS training data decreases. As a toy example, consider a region with two variants with nonzero marginal GWAS statistics that are in perfect LD: without additional information, it is impossible to determine whether one or both of the variants are causal. Uncertainty in which variant is causal propagates to uncertainty in the weights used for PRS leading 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 prediction 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 (e.g., stratification) remains unclear.

Motivated by potential clinical applications of PRS in personalized medicine where the goal is estimation of genetic value we focus on evaluating uncertainty in PRS estimates at the level of a single target individual. We assess two metrics of individual PRS uncertainty: (1) the standard deviation of the PRS estimate for individual i, denoted as $sd(\widehat{PRS}_i)$; and (2) the ρ-level credible interval for the genetic value of individual i defined as the interval that contains the genetic value of individual i (GV_i) with ρ (e.g., 95%) probability, denoted as *(ρ GV_i-CI)*. We extend the Bayesian framework of LDpred2[21], a widely used method for PRS estimation, to sample from the PRS posterior distribution to estimate $sd(\widehat{\mathrm{PRS}}_i)$ and ρ GV_i-CI for different values of ρ. In extensive simulations starting from real genotypes in the UK Biobank (N=291,273 individuals, M=459,792 SNPs, unrelated "white British"), we show that ρ GV_i-CI is well-calibrated when target sample matches the training data; $sd(\widehat{\mathrm{PRS}}_i)$ increases as polygenicity (number of causal variants) increases and as heritability and GWAS sample size decrease[26]. Analyzing 11 real traits in the UK Biobank, we observe large uncertainties in individual PRS estimates that greatly impact interpretability of ranking of individuals based on PRS. For example, on average across 11 traits, only 0.6% (s.d. 1.1%) of individuals with PRS point estimates in the top 1% also have their 95% GV_i-CI fully contained in the top 1%. Individuals with PRS point estimates at the 90th percentile in a testing sample can be ranked anywhere between the 48th and 99th percentile in the same testing sample after their 95% credible intervals are taken into account. Finally, we propose a probabilistic approach to PRS-based stratification that incorporates

uncertainty in PRS estimates and we demonstrate how such approaches can enable principled risk stratification under different cost scenarios.

Results

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 $\mathrm{GV}_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, where $\mathbf{x}_i$ is an $M \times 1$ vector of observed genotypes and $\boldsymbol{\beta}$ is the corresponding $M \times 1$ vector of unknown causal effect sizes[27] (Methods). Different PRS methods vary in how they estimate causal effects $\widehat{\boldsymbol{\beta}}$ to construct the estimator $\widehat{\mathrm{PRS}}_i = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}$. Inferential variance in $\widehat{\boldsymbol{\beta}}$ propagates into variance of $\widehat{\mathrm{PRS}}_i$. In this work, we focus on quantifying the inferential uncertainty in $\widehat{\mathrm{PRS}}_i$ and assessing its impact on PRS-based stratification.

To illustrate the impact of statistical noise on $\widehat{\mathrm{PRS}}_i$, consider a toy example of a trait for which the observed marginal GWAS effects at three SNPs are equal (Figure 1). Since SNP 1 is uncorrelated with SNPs 2 and 3 and SNPs 2 and 3 are perfectly correlated (LD = 1), we can conclude that the observed signal is being driven by either two or three of the SNPs (SNPs 1 and 2, SNPs 1 and 3, or all three SNPs). The key challenge is that there is more than one causal effect-size configuration that is consistent with the observed GWAS effects and, in the absence of other information, these causal effect-size configurations are equally probable. In such situations, one can generate different PRS estimates for a single individual from the same GWAS training data. For example, an individual with the genotype $\mathbf{x}_i = (0,1,0)^\top$ can be classified as high-risk or low-risk depending on the underlying causal effect configuration (Figure 1). Thus, point estimates alone ($\widehat{\mathrm{PRS}}_i$) do not provide sufficient information for robust risk stratification.

To explore inferential uncertainty in $\widehat{\mathrm{PRS}}_i$, we sample from the posterior distribution of LDpred2 to estimate inferential variance in $\widehat{\mathrm{PRS}}_i$ (e.g., $sd(\widehat{\mathrm{PRS}}_i)$) and compute credible intervals for GV_i at prespecified confidence levels (e.g., $\rho = 95\%$) (Figure 2). We then propose a probabilistic approach to PRS-based risk stratification that estimates the probability that GV_i is above a given threshold t (Figure 2), and demonstrate how this probability can be used in conjunction with situation-specific cost functions to optimize risk stratification decisions.

Factors impacting individual PRS uncertainty in simulations

We quantify the degree to which different parameters contribute to uncertainty in individual PRS estimates in simulations starting from real genotypes of unrelated "white British" individuals in the UK Biobank (N=291,273 individuals and M=459,792 SNPs). To avoid overfitting, we partitioned the individuals into disjoint training, validation and testing groups (N_{train}=250,000, $N_{validation}$=20,000, N_{test}=21,273). Training samples were used to estimate PRS weights; validation samples were used to estimate hyperparameters (e.g., heritability and polygenicity) for LDpred2; and testing samples were used to evaluate accuracy (Figure S1) and uncertainty (Methods).

First, we assess the calibration of the ρ-level credible intervals for GV_i estimated by LDpred2. We compared the empirical coverage of the ρ-level credible intervals (proportion of individuals in a single simulation replicate whose ρ GV_i-CI overlaps their true GV_i) to the expected coverage (ρ) across a range of values of ρ. We find that, overall, the ρ GV_i-CI are well-calibrated, albeit slightly mis-calibrated in low-

heritability, high-polygenicity simulations (Figure 3a and Figure S4). For example, across 10 simulation replicates where $h_g^2 = 0.25$ and $p_{\text{causal}} = 1\%$, the 90% GV_i-CIs have an average empirical coverage of 0.91 (s.e.m. 0.005) (Figure 3a). Since individuals with large PRS estimates might have 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 $sd(\widehat{\text{PRS}}_i)$ and their true genetic value (Figure 3b).

We next assess the impact of trait-specific genetic architecture parameters (heritability and polygenicity) on individual PRS uncertainty. First, we fixed heritability and varied polygenicity to find that $sd\left(\widehat{\text{PRS}}_i\right)$ increases with increases from 0.19 to 1.44 when the proportion of causal variants increases from 0.1% to 100% (Figure 3c, Figure S2). Second, we varied the heritability while keeping polygenicity constant; since different heritability and sample size lead to different variances explained by the PRS in the test sample, we scale the individual standard deviation ($sd\left(\widehat{\text{PRS}}_i\right)$) by the standard deviation of PRS point estimates across all tested individuals; we refer to this quantity as "scaled SD" (Methods). We find that the scaled SD decreases with heritability and sample size (Figure 3d, Figure S3). For example, when $h_g^2 = 0.05$ and $p_{\text{causal}} = 0.1\%$, a 5-fold increase in training data sample size (50K to 250K) reduces scaled SD by 6-fold (from 3.02 to 0.51); when $h_g^2 = 0.05$ and $p_{\text{causal}} = 1\%$, the same increase in training data sample size reduces the scaled SD by 5-fold (from 1.99 to 0.39). While the two simulation settings ($h_g^2 = 0.5, p_{\text{causal}} = 1\%$ versus $h_g^2 = 0.05, p_{\text{causal}} = 0.1\%$) yield the same expected variance per causal variant under our simulation framework (i.e. $h_g^2/(M \times p_{\text{causal}})$, see 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.

Individual PRS uncertainty in real data in the UK Biobank

We investigate individual PRS uncertainty across eleven traits in the UK Biobank: hair color, height, body mass index (BMI), bone mass density in the heel (BMD), high-density lipoprotein (HDL), low-density lipoprotein (LDL), cholesterol, igf1, creatinine, red blood cell count (RBC) and white blood cell count (WBC). First we focus on PRS-based risk stratification; we classify test individuals as "high-risk" if their PRS point estimate (the posterior mean of their genetic value) exceeds a prespecified threshold t (i.e. $\widehat{\text{PRS}}_i > t$), where t is set to the 90th PRS percentile obtained from the test-group individuals (Methods). We then partition the "high-risk" individuals into two categories: individuals whose 95% GV_i-CI are fully above the threshold t ("certain high-risk") and individuals whose 95% GV_i-CI contain t ("uncertain high-risk"). Similarly, we classify individuals as "low-risk" if their PRS point estimate lies below the prespecified threshold ($\widehat{\text{PRS}}_i < t$) and we partition these individuals into "certain low-risk" and "uncertain low-risk" based on their 95% GV_i-CI (Figure 4a). At t = 90th percentile and ρ = 95%, only 1.8% (s.d. 2.4%) of "high-risk" individuals (averaged across traits) are deemed "certain high-risk" individuals; the remaining "high-risk" individuals have ρ-level credible intervals that overlap t (Figure 4b, Table 1). On the other hand, 33.7% (s.d. 15.3%) of "low-risk" individuals have ρ-level credible intervals that do not overlap t (Figure 4b, Table 1). Consistent with simulations, we find that uncertainty is higher for traits that are more polygenic[28] (Table 1) with the average standard deviation of $\widehat{\text{PRS}}_i$ ranging between 0.2 to 0.41 across the studied traits (Table S1).

For completeness we investigated the impact of the threshold t, and credible level ρ, on PRS-based stratification uncertainty, defined as the proportion of high-risk individuals classified as "certain high-risk" for a given trait. As expected, the proportion of certain high-risk classifications decreases as ρ increases (Figure 4a). For traits with higher average uncertainty (as defined using the scaled SD) we observe lower rates of certain classifications across all values of ρ. For example, at t =90th and ρ =95%, the proportion of high-risk individuals classified with certainty is 0.2% for BMI (average scaled SD = 1.24) and 8.8% for hair color (average scaled SD = 0.55) (Figure 5a). Height and HDL have similar average levels of uncertainty (average scaled SD of 0.81 for height and 0.82 for HDL) and similar proportions of high-risk individuals classified with certainty. For example, at t =90th and ρ =95%, the proportions of certain classifications among high-risk individuals are 2.1% for both height and HDL (Figure 5a, Table 1). Using a more stringent threshold t amplifies the effect of uncertainty on PRS-based stratification (Figure 5b). For example, for BMI and hair color, the proportion of certain classifications among high-risk individuals drops for all values of ρ when we increase the threshold from t=90th percentile to t=99th percentile (Figure 5b).

We also quantified the impact of inferential variance in $\widehat{\mathrm{PRS}}_i$ on PRS-based ranking of the test-group individuals. Using two different samplings from the PRS posterior for each test individual, we generated two independent rankings for all individuals in the test data and quantified the correlation in the rankings (Figure 4c, Methods). We observe large variability in the rankings across the test data, with the correlation of rankings ranging from 0.47 to 0.82 across the 11 traits. We also estimated 95% credible intervals for the rank of individuals at a given percentile (e.g., 90th) (Table 2, Methods, Supplementary Figure S5) to find high variability in the ranking. For example, in the case of HDL an individual at 90th (99th) percentile based on PRS point estimate can be within 48th to 99th percentiles (79th-99th) with 95% probability when the inferential variance in PRS estimation is taken into consideration (Table 2).

Integrating individual-PRS 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, we propose to use the posterior probability that GV for individual i is above the threshold $\Pr(\mathrm{GV}_i > t)$. This approach is motivated by using the probabilistic framework of LDpred2 to derive the probability that the *absolute genetic risk* of a given individual is greater than a given threshold. We estimate $\Pr(\mathrm{GV}_i > t)$ using Monte Carlo integration within LDpred2 framework (Methods, Figure S6). As a motivating example, two individuals with similar PRS point estimates that happen to lie on either side of a prespecified threshold ($\widehat{\mathrm{PRS}}_i < t$ and $\widehat{\mathrm{PRS}}_j > t$) could have similar probabilities for the *absolute genetic risk* (genetic value) to exceed t (e.g., $\Pr\,(\mathrm{GV}_j > t) = 0.4$ and $\Pr\,(\mathrm{GV}_j > t) = 0.6$) (Figure 2).

As expected, we find a tight correlation between $\widehat{\mathrm{PRS}}_j$ and $\Pr\left(\mathrm{GV_j} > t\right)$ across individuals in the test data (Figure S7). However, $\Pr\left(\mathrm{GV_j} > t\right)$ 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 a high-risk versus low-risk classification (Methods). As an example of utility of the probabilities consider three cost functions which relate the relative costs of false positive versus false negative diagnoses: (a) equal cost for each FP and FN diagnosis ($C_{FP} = C_{FN} = 1$); (b) 3x higher cost for FP diagnoses ($C_{FP} = 3$, $C_{FN} = 1$); and (c) 3x higher cost for FN diagnoses ($C_{FP} = 1$, $C_{FN} = 3$). For an individual with $\Pr\,(\mathrm{GV}_i > t) = 0.6$, probability of a FP versus FN diagnosis is 0.4 versus 0.6, respectively. Thus, the expected costs from FP diagnoses ($\Pr(\mathrm{FP}) \times \mathrm{C_{FP}}$) under each scenario are (a) 0.4, (b) 1.2, and (c) 0.4; the

expected costs of FN diagnoses ($\Pr(\mathrm{FN}) \times \mathrm{C_{FN}}$) are (a) 0.6, (b) 0.6, and (c) 1.8. Therefore, the classification for this individual that minimizes the expected cost under each scenario is (a) high-risk, (b) low-risk, and (c) high-risk. Assuming the same three cost functions as above, we find that the optimal decision threshold on $\Pr(\mathrm{GV}_i > t)$ that maximizes the utility of the cost/gain models differs under the three functions. For $C_{FP} = C_{FN} = 1$, both the estimated cost curve and true cost curve achieve minimum cost at threshold = 0.5. For $C_{FP} = 3$, $C_{FN} = 1$, the estimated optimum is 0.25 and the true optimum is 0.3. For $C_{FP} = 1$, $C_{FN} = 3$, the estimated optimum is 0.75 and the true optimum is 0.7. More notably, assuming the probabilities are well-calibrated, we can estimate the expected cost with the individual probability of being at high risk, with the estimated cost curve being very close to the true cost curve despite slight inflation (Figure 6).

Discussion

In this work, we demonstrate that uncertainty in PRS estimates at the individual level can have a large impact on subsequent analyses such as PRS-based risk stratification. We propose a general procedure for obtaining estimates of individual-PRS uncertainty which can be applied to a wide range of existing PRS methods. Among 11 real traits in the UK Biobank, we find that even with GWAS sample sizes on the order of hundreds of thousands of individuals, there is considerable uncertainty in individual-PRS estimates (i.e. large ρ-level credible intervals) that can impair the reliability of PRS-based stratification. To account for individual-PRS uncertainty we propose a probabilistic approach to stratification and demonstrate how the probabilistic framework can be used in conjunction with situation-specific cost functions to optimize PRS-based decision-making.

Our work is complementary to methods that aim to improve cohort-level metrics of PRS accuracy such as prediction R^2 or the area under the receiver operating characteristic (AUROC). On the application side, we use the example of genetic risk stratification to showcase the importance of incorporating individual uncertainty, which allows us to estimate individual absolute and relative genetic risks without a validation sample that is normally required to estimate absolute risks. As the individualized absolute risk estimates (genetic values) do not depend on a validation sample, and are individualized, we believe they could be robust leads to our proposed probabilistic genetic risk stratification, which can be seen as a principled approach for genetic risk stratification in clinical settings.

We conclude with several caveats and future directions. First, we quantify individual-PRS uncertainty by extending LDpred2[21], which is just one of many existing Bayesian methods that can be adapted for the same purpose (e.g., SBayesR[24], PRS-CS[28] and AnnoPred[29]). Extensions of other methods, including analogous procedures for P+T (PRSice-2[30]) and regularization-based approaches (lassosum[19] and BLUP prediction[20] [21]), could also be investigated.

Second, while we find broad evidence that both trait-specific genetic architecture parameters (e.g., heritability, polygenicity) and individual-specific genomic features (e.g., cumulative number of effect alleles) can impact individual-PRS uncertainty, both sources of uncertainty merit further exploration. For example, we perform simulations under a model in which each causal variant explains an equal portion of total SNP-heritability but, in reality, genetic architecture can vary significantly among different traits. Does individual-PRS uncertainty change if both monogenic and polygenic disease risk factors[31,32] are used for PRS estimation? We find a small albeit significant correlation between an individual's cumulative number of effect alleles and their individual-PRS uncertainty, but is this increased uncertainty driven by certain categories of effect

alleles (e.g., effect alleles with MAF > 5%)? We leave a more detailed investigation of the various sources contributing to individual-PRS uncertainty for future work.

Third, we perform all simulations and real data analyses using genotyped SNPs (MAF > 1% on the UK Biobank Axiom Array). Since the array is designed such that the genotyped SNPs tag most of the signal from unobserved SNPs, the SNPs (predictors) used in our real data analyses likely capture most of the SNP-heritability for each trait. However, it is unclear whether individual-PRS uncertainty would increase or decrease if imputed data were used instead of genotyped SNPs. Moreover, for many diseases, the largest GWAS are only available as summary statistics (estimates of marginal effects and their standard errors). It's important to assess whether there's larger uncertainty for causal effects inferred from summary statistics which leads to higher variability in estimated PRS. We conjecture that changes in uncertainty will also vary across traits depending on factors such as the number of SNPs (predictors) included in the PRS; the resolution of the credible sets generated by sampling causal configurations; and differences in LD tagging between predictor SNPs and causal SNPs as well as among predictor SNPs. A comparison of individual-PRS uncertainty with respect to array data, imputed data, and summary statistics merits thorough investigation in future work.

Lastly, 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 significantly across populations[33,34]. 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[35,36].We therefore leave a detailed exploration of individual-PRS uncertainty with respect to ancestry as future work.

Methods

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

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

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

Here, 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 (i.e., polygenicity). Let $\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS}}$ and $\widehat{\mathbf{R}}$ represent GWAS marginal effects and LD matrix computed from GWAS samples. By combining the prior probability $p(\boldsymbol{\beta} | h_g^2, p_{\mathrm{causal}})$

and the likelihood of observed data $p(\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS}}|\boldsymbol{\beta}, \widehat{\mathbf{R}})$, we can compute a posterior distribution as $p(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS}}, \widehat{\mathbf{R}}, h_g^2, p_{\mathrm{causal}})$. The posterior distribution is intractable and therefore LDpred2 uses Markov Chain Monte Carlo (MCMC) to obtain posterior samples from $p(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS}}, \widehat{\mathbf{R}}, h_g^2, p_{\mathrm{causal}})$. For simplicity, we use $\widetilde{\boldsymbol{\beta}} \sim p(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS}}, \widehat{\mathbf{R}}, h_g^2, p_{\mathrm{causal}})$ to refer to the samples from the posterior distribution, and use $p(\widetilde{\boldsymbol{\beta}})$ to refer to $p(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS}}, \widehat{\mathbf{R}}, h_g^2, p_{\mathrm{causal}})$ whenever context is clear. The posterior samples of the causal effects are summarized using the expectation $\mathbb{E}[\widetilde{\boldsymbol{\beta}}] = \int \widetilde{\boldsymbol{\beta}} p(\widetilde{\boldsymbol{\beta}}) d\widetilde{\boldsymbol{\beta}}$, leading to $\widehat{\mathrm{PRS}}_i = \mathbf{x}_i^\top \mathbb{E}[\widetilde{\boldsymbol{\beta}}]$.

Unlike existing methods that summarize the posterior samples of causal effects into the expectation and then estimate $\widehat{\mathrm{PRS}}_i$, we sample from the posterior of PRS_i to construct a ρ level credible interval of genetic value (ρ GV_i-CI) for each individual. Bernstein-von Mises theorem provides the basis that under certain conditions, such constructed Bayesian credible interval will asymptotically be of coverage probability ρ[37]. This property of the Bayesian credible interval provides intuitive explanation of the uncertainty. Concretely, we obtain B MCMC samples from the posterior distribution of causal effects $p(\widetilde{\boldsymbol{\beta}})$: $\widetilde{\boldsymbol{\beta}}^{(1)}, \widetilde{\boldsymbol{\beta}}^{(2)}, \dots, \widetilde{\boldsymbol{\beta}}^{(B)}$. Then we compute a PRS estimate for individual i from each sample of $p(\widetilde{\boldsymbol{\beta}})$: $\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}^{(1)}, \mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}^{(2)}, \dots, \mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}^{(B)}$ to approximate the posterior distribution of $\mathrm{PRS_i}$ ($p(\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}})$). From the B samples of posterior, we obtain empirical $\frac{1-\rho}{2}$ and $\frac{1-\rho}{2}$ quantiles as lower and upper bound estimates of ρ GV_i-CI (Figure 2b). As B goes to infinity, such Monte Carlo estimates converge to the $[Q_{(1-\rho)/2}(\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}), Q_{(1+\rho)/2}(\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}})]$, where $Q_\alpha(\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}})$ represents the α-quantile (here, $\alpha = (1-\rho)/2, (1+\rho)/2$) for distribution of $p(\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}})$. Similarly, we summarize the posterior samples using the second moment to estimate $sd(\widehat{\mathrm{PRS}}_i) = sd(\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}})$. In practice, we used $B = 500$ as that leads to stable results. Although in this work we focus on LDpred2, the above described procedure is generalizable to a wide range of Bayesian methods (e.g., SBayesR[24], PRS-CS[28] and AnnoPred[29]). Methods that are not based on Bayesian principle could potentially use Bootstrap to obtain individual uncertainty intervals[38].

Simulations. We design 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 UK Biobank [39]. We excluded SNPs with MAF < 0.01 and genotype missingness > 0.01, and those SNPs that fail the Hardy-Weinberg test at significance threshold 10^{-7}, which leaves us 459,792 SNPs. We preserve "white British individual", with self-reported British white ancestry and filter pairs of individuals with kinship coefficient < $1/2^{(9/2)}$) [39]. We further filtered individuals who are outliers for genotype heterozygosity and/or missingness, and obtained 291,273 individuals for all analyses.

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

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

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

Finally, given the phenotypes $\mathbf{y} = (y_1, \dots, y_N)^\top$ and genotypes $\mathbf{X}$, we simulate the GWAS marginal association statistics with $\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS}} = \frac{1}{N}\mathbf{X}^\top \mathbf{y}$. We simulate the data using a wide range of parameters, $h_g^2 \in \{0.05, 0.1, 0.25, 0.5, 0.8\}$, $p_{\mathrm{causal}} \in \{0.001, 0.01, 0.1, 1\}$, 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 rest of 21,273 individuals as the testing population, as the usual practice for the PRS model building process. When investigating how sample sizes in the training cohort change PRS uncertainty, we vary the sample sizes in the training population in 20,000, 50,000, 100,000, 150,000, and 250,000, while holding the validation population and testing population as intact, to enable a fair comparison between sample sizes.

Real data analysis. We performed real data analysis with 11 real traits from UK Biobank, including hair color, height, body mass index (BMI), bone mass density in the heel (BMD), high density lipoprotein (HDL), low density lipoprotein (LDL), cholesterol, igf1, creatinine, red blood cell count (RBC) and white blood cell count (WBC). 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. The random partition was repeated five times to average of 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 phenotype in 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 are first estimated from the training population, and applied to phenotype from training, validation and testing population respectively. We evaluate accuracy of PRS estimates in validation and testing groups by Pearson correlation between PRS estimates and adjusted phenotypes.

PRS analysis using LDpred2. We run LDpred2 for both simulation and real data analysis with the following settings. We calculate the in-sample LD with functions provided by the LDpred2 package, using the window size parameter of 3cM. We estimate the heritability $h^2_{\text{chr}_i}, i = 1, \dots, 22$ for each chromosome with built-in constrained LD score regression[40] function. We run LDpred2-grid per chromosome with a grid of six polygenicity parameters p_{causal} from 10^{-4} to 10^{-1} equally spaced on a log space, three heritability parameters $0.9h^2_{\text{chr}_i}, h^2_{\text{chr}_i}, 1.1\ h^2_{\text{chr}_i}$ and sparsity option enabled. We choose the model with the highest R^2 between the predicted posterior mean and the (adjusted) phenotype on validation set as best model to apply to testing data. We extract 500 posterior samples of causal effects $\widetilde{\boldsymbol{\beta}}^{(1)}, \widetilde{\boldsymbol{\beta}}^{(2)}, \dots, \widetilde{\boldsymbol{\beta}}^{(500)}$ after 100 burn-in iterations from MCMC sampler of the model to approximate posterior distribution of causal effects. For each individual with genotype $\mathbf{x}_i$, we calculate $\mathbf{x}_i^\top\ \widetilde{\boldsymbol{\beta}}^{(1)}, \mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}^{(2)}, \dots, \mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}^{(500)}$ to approximate GV posterior distribution for individual i. We then calculate summary statistics of GV posterior distribution, including the posterior mean ($\widehat{\text{PRS}}_i$), ρ level credible interval (ρ GV_i-CI) and probability of above threshold t ($\Pr(\text{GV}_i > \text{t})$).

Calculating and evaluating the coverage. We evaluate the coverage properties of ρ GV_i-CI in simulation: we check whether $\mathbb{P}\left(\mathbf{x}_i^\top \boldsymbol{\beta} \in \left[Q_{(1-\rho)/2}\left(\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}\right), Q_{(1+\rho)/2}\left(\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}\right)\right]\right) = \rho$. To evaluate this property, for each simulated dataset, we calculate the frequency of the true genetic risk lies in the predicted interval, i.e., the frequency of $\mathbf{x}_i^\top \boldsymbol{\beta} \in \left[Q_{(1-\rho)/2}\left(\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}\right), Q_{(1+\rho)/2}\left(\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}\right)\right]$ for every individual in the testing population, for $\rho \in \{0.1, 0.2, \dots, 1.0\}$. This property provides us an intuitive understanding of the predicted interval: for an individual with a predicted interval $\left[Q_{(1-\rho)/2}\left(\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}\right), Q_{(1+\rho)/2}\left(\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}\right)\right]$, its true genetic risk is expected to be in this interval with a probability ρ.

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

Posterior individual ranking interval. The relative rank of individual PRS $\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}^{(\mathrm{b})}$ in the population $\mathbf{x}_j^\top \widetilde{\boldsymbol{\beta}}^{(\mathrm{b})}, j = 1, \dots, N$ varies across different MCMC samplings of posterior causal effects. To evaluate the uncertainty of ranking for individual i, we compute $r_i^{(b)}$ as the quantile of $\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}^{(\mathrm{b})}$ in the population $\mathbf{x}_j^\top \widetilde{\boldsymbol{\beta}}^{(\mathrm{b})}, j = 1, \dots, N$ for each of the $b = 1, \dots, B$ posterior samples to approximate posterior distribution of the relative rank. We can obtain ρ-level credible intervals of ranking as $[Q_{(1-\rho)/2}(r_i), Q_{(1+\rho)/2}(r_i)]$ for each individual i. To assess the uncertainty of ranking for individuals at 90 (99) percentile threshold based on PRS estimates, we select individuals within 1 percentile of thresholds (89.5-90.5%, 98.5-99.5%) and compute mean and standard deviation for lower and upper bound of ρ=95% posterior ranking interval, across the selected individuals.

PRS rank correlation between different MCMC samplings. With the B posterior causal effects samples $\widetilde{\boldsymbol{\beta}}^{(1)}, \widetilde{\boldsymbol{\beta}}^{(2)}, \dots, \widetilde{\boldsymbol{\beta}}^{(B)}$, and N individuals in the testing population $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$, we compute PRS for each individual, $\mathbf{x}_1^\top \widetilde{\boldsymbol{\beta}}^{(b)}, \dots, \mathbf{x}_N^\top \widetilde{\boldsymbol{\beta}}^{(b)}$ and its relative rank in the population $r_1^{(b)}, \dots, r_N^{(b)}$ for each posterior sample $\widetilde{\boldsymbol{\beta}}^{(b)}$. Then for each pair of different b_1-th,b_2-th posterior samples, $\widetilde{\boldsymbol{\beta}}^{(b_1)}, \widetilde{\boldsymbol{\beta}}^{(b_2)}$, we calculate the spearman correlation between $r_1^{(b_1)}, \dots, r_N^{(b_1)}$ and $r_1^{(b_2)}, \dots, r_N^{(b_2)}$, representing the variability of the ranks across MCMC samplings. We compute the rank correlation for 1000 pairs of different MCMC samplings, and get the distribution of the rank correlation.

Probabilistic risk stratification. We define the notion of probabilistic framework for risk stratification based on posterior distribution of GV_i. Given a pre-specified threshold t, for every individual, we can calculate the posterior probability of the genetic risk larger than the given threshold t, $\Pr(\mathrm{GV_i} > t)$, with Monte Carlo integration as

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

We use the previous simulation settings to show that this probability is well calibrated. For each simulation, we divide the individuals based on their posterior probability of being at high risk into 10 bins with $\{0, 0.1, \dots, 1.0\}$ as breaks. For each bin, we calculate 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 is expected to be the same as theoretical probability.

Utility analysis. The individualized posterior distribution of genetic value provides extra information for patient stratification. We consider a scenario that there is a cost associated for decision that (1) classify an individual with low genetic risk into a high genetic risk category, C_{FP}, where FP represents false positive. (2) classify an individual with high genetic risk into a low genetic risk category, C_{FN}, where FN represents

false negative. For an individual with posterior probability $\Pr(\mathrm{GV}_i > t)$, we want to decide an action, whether to classify this individual to be at high genetic risk, and perform further screening. If we classify this individual as high risk, we will have probability $1 - \Pr(\mathrm{GV}_i > t)$, that this individual is in fact low risk, inducing an expected cost $C_{\mathrm{FP}}\big(1 - \Pr(\mathrm{GV}_i > t)\big)$. Conversely, if we classify this individual as low risk, we will have probability $\Pr(\mathrm{GV}_i > t)$ that this individual will be in the high genetic risk, inducing an expected cost $C_{FN}\Pr(\mathrm{GV}_i > t)$. To minimize the expected cost, we would decide according to which action leads to the least cost. The critical value in this scenario is $\frac{C_{\mathrm{FN}}}{C_{\mathrm{FP}}+C_{\mathrm{FN}}}$: if $\Pr(\mathrm{GV}_i > t) > \frac{C_{\mathrm{FN}}}{C_{\mathrm{FP}}+C_{\mathrm{FN}}}$, we would choose to classify this individual as high risk, otherwise low risk.

Data availability

The individual-level genotype and phenotype data are available by application from the UKBB http://www.ukbiobank.ac.uk/.

URLs

LDPred2 software implementing individual PRS credible intervals:
https://privefl.github.io/bigsnpr/articles/LDpred2.html

Acknowledgments

This research was conducted using the UK Biobank Resource under application 33297. We thank the participants of UK Biobank for making this work possible. This work was funded in part by NIH awards nos: R01HG009120, R01MH115676. R01HG006399, and U01CA194393.The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Figures and Tables

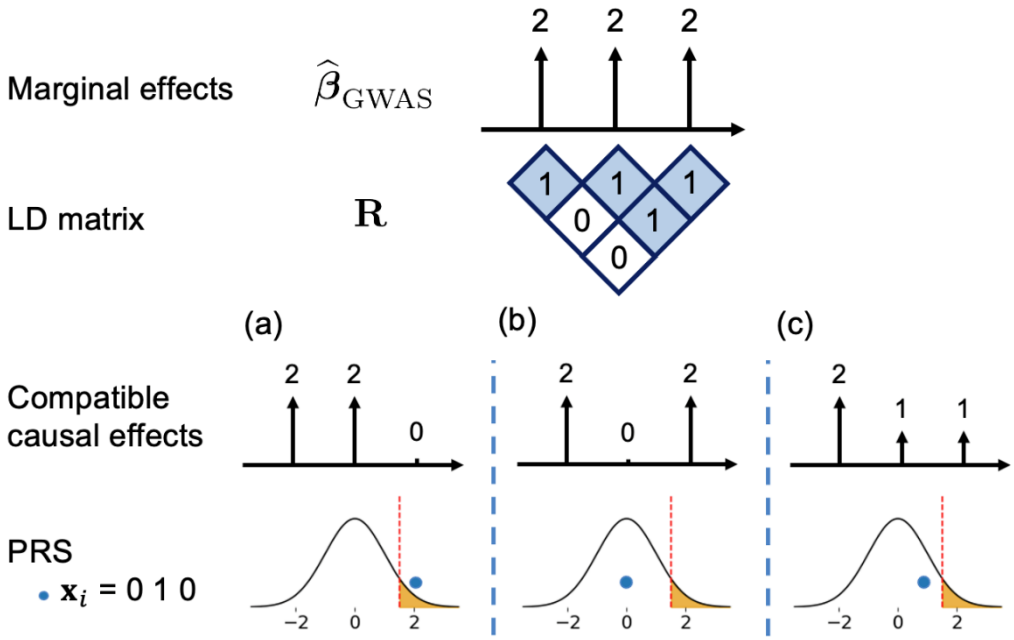

Figure 1. Toy example to illustrate how LD introduces uncertainty into PRS estimates. Three SNPs have the same GWAS marginal effects, $\widehat{\boldsymbol{\beta}}_{\mathrm{GWAS}} = (2,2,2)^{\top}$ and SNPs 2 and 3 are in perfect LD. (a-c) represent three causal effect configurations that are compatible with observed GWAS marginal effects: (a) SNP 2 is causal with an effect size of 2; (b) SNP 3 is causal with an effect size 2; and (c) both SNPs 2 and 3 are causal with effect sizes of 1. If the high-risk threshold is set to $t = 1.5$, a given individual with the genotype $\mathbf{x}_i = (0,1,0)^{\top}$would have PRS point estimates of 2 (high risk), 0 (low risk), and 1 (low risk) under each configuration, respectively, where the risk threshold (dotted red line) is defined with respect to the population PRS distribution. The blue dot represents the PRS point estimate of an individual with $\mathbf{x}_i = (0,1,0)^{\top}$.

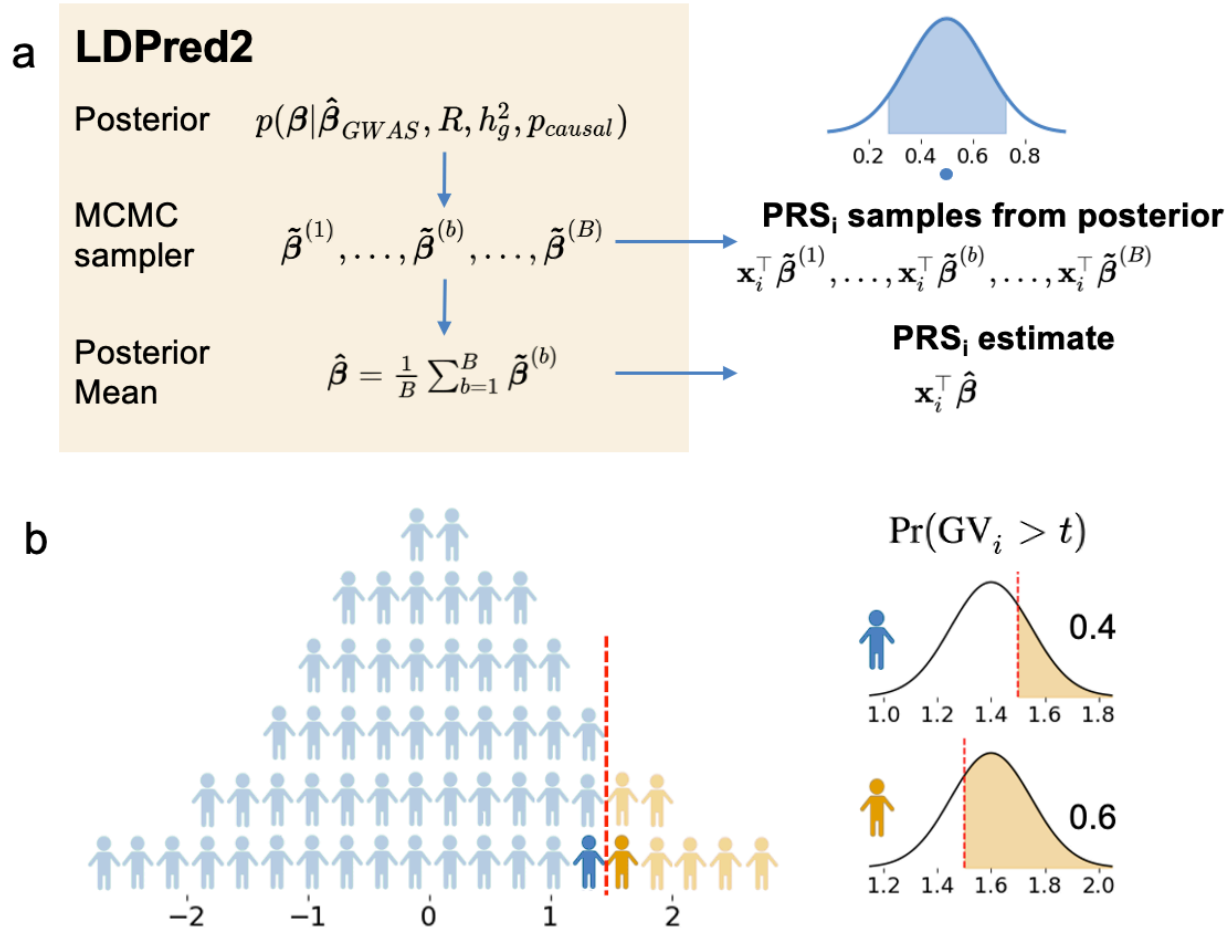

Figure 2. Framework of extracting uncertainty from Bayesian methods and probabilistic individual stratification. (a) Procedure to obtain uncertainty from LDpred2. LDpred2 uses MCMC to sample from posterior effect distribution given GWAS marginal effects, LD matrix and prior distribution. It outputs posterior mean effects and use that to predict posterior mean genetic value as PRS estimate. Our framework samples posterior samples of genetic effects to approximate the posterior distribution of genetic value. The density plot represents posterior distribution of GV for an individual. The shaded area represents a ρ-level credible interval. The dot represents posterior mean. (b) Probabilistic risk stratification framework. Given a threshold t, Instead of dividing individuals into high risk ($\widehat{\mathrm{PRS}}_i > t$) and low risk ($\widehat{\mathrm{PRS}}_i \leq t$) dichotomously (left), probabilistic risk stratification assigns each individual a probability of high risk $\Pr(\mathrm{GV}_i > t)$ (right).

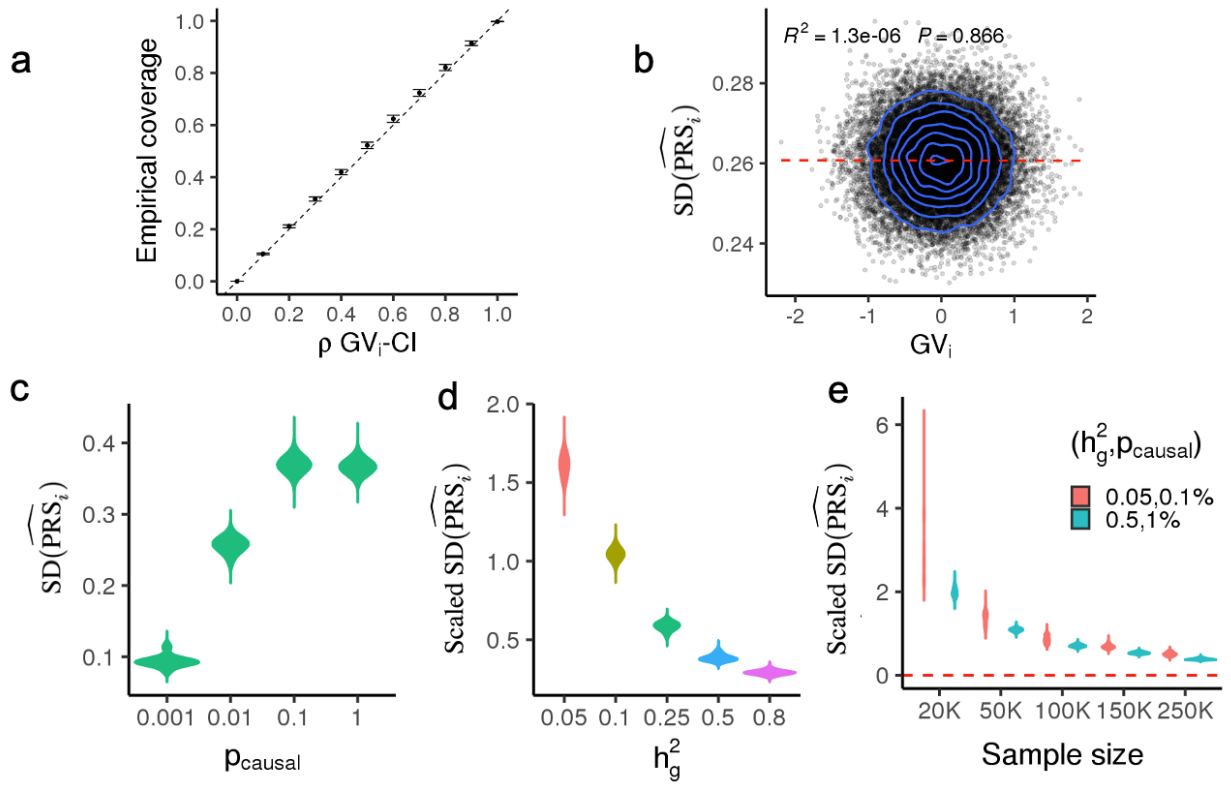

Figure 3. Genetic architecture (polygenicity (p_{causal}), SNP-heritability (h_g^2), and GWAS sample sizes) impacts uncertainty in PRS estimates in simulations. (a) Individual credible intervals are well-calibrated ($h_g^2 = 0.25$, $p_{\text{causal}} = 1\%$). Empirical coverage is calculated as the proportion of ρ-level credible intervals that contain true genetic risk for one simulation. The error bars represent 1.96 standard errors of the mean calculated from 10 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 standard deviation of individual PRS estimate ($sd(\widehat{\text{PRS}}_i)$) (c) Distribution of individual PRS uncertainty estimates with respect to polygenicity ($p_{causal} \in \{0.0001, 0.01, 0.1, 1\}, h_g^2 = 0.25$). Each violin plot represents $sd(\widehat{\text{PRS}}_i)$ for 21,273 testing individuals across 10 simulations. (d) Distribution of individual PRS uncertainty estimates with respect to heritability ($h_g^2 \in \{0.05, 0.1, 0.25, 0.5, 0.8\}, p_{causal} = 0.01$). Each violin plot represents scaled $sd(\widehat{\text{PRS}}_i)$ for 21,273 testing individuals across 10 simulation replicates. (e) Distribution of individual uncertainty estimates with respect to training GWAS sample. Each violin plot represents scaled $sd(\widehat{\text{PRS}}_i)$ of individual prs for 21,273 testing individuals across 10 simulation replicates.

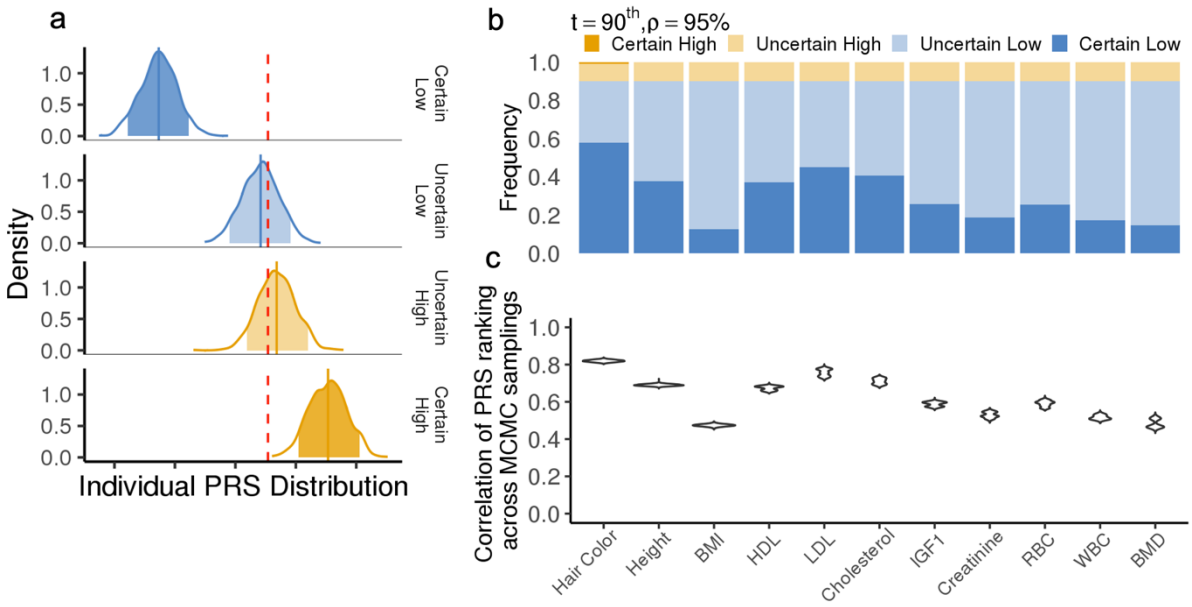

Figure 4. Uncertainty in real data and its influence on genetic risk stratification. (a) Example of posterior PRS distribution for individuals with certain low risk (dark blue), uncertain low risk (light blue), uncertain high risk (light yellow), high risk (dark yellow) classification. Each density plot is a smoothed posterior HDL PRS distribution of an individual randomly chosen from that category. The solid vertical line is posterior mean. The shading area is 95% credible interval. The red dotted line is classification threshold. (b) Distribution of classification categories across 11 traits (t=90%, ρ=95%). Each bar plot represents the distribution of four classification categories in testing individuals for one trait. The frequency is averaged among five random partition repeats of the whole dataset. (c) Correlation of PRS rankings of test individuals across pairs of MCMC samplings from posterior. For each trait, we sample two pairs of effects from posterior distribution of causal effects, rank all individuals in the test data based on PRS predicted by the first or the second sample from posterior and compute correlation. The distribution of correlation of the PRS rankings for each trait across 1,000 pairs of MCMC sampling and five random partition repeats is shown.

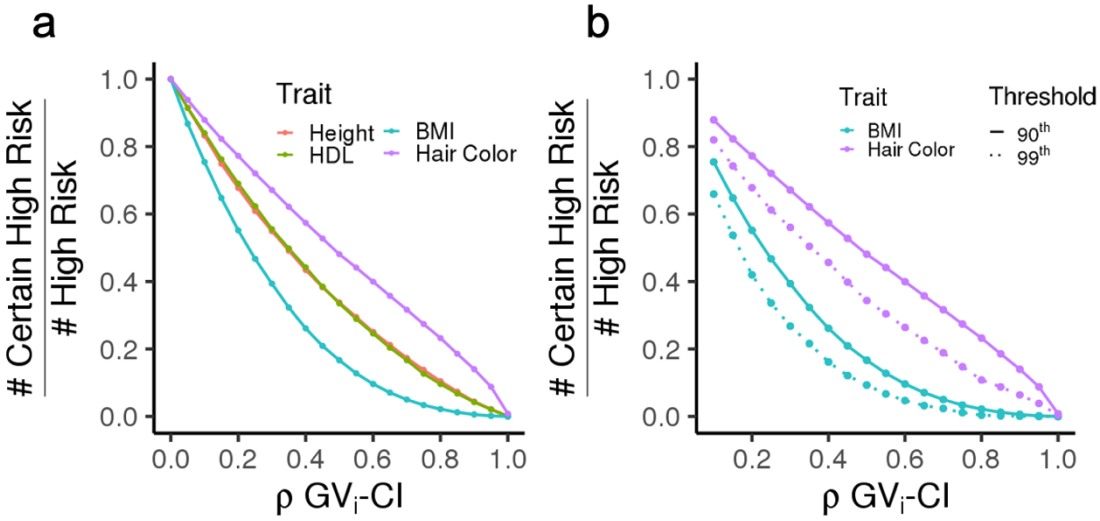

Figure 5. Impact of threshold t, credible set level ρ on stratification uncertainty. (a) Proportion of certain classification among high risks across four representative traits and credible interval sizes from 0 to 1 with 0.05 as increment. The stratification threshold t is fixed at 90%. (b) Proportion of certain classification among high risks across two representative traits, and two different stratification thresholds ($t = 90\%, 99\%$).

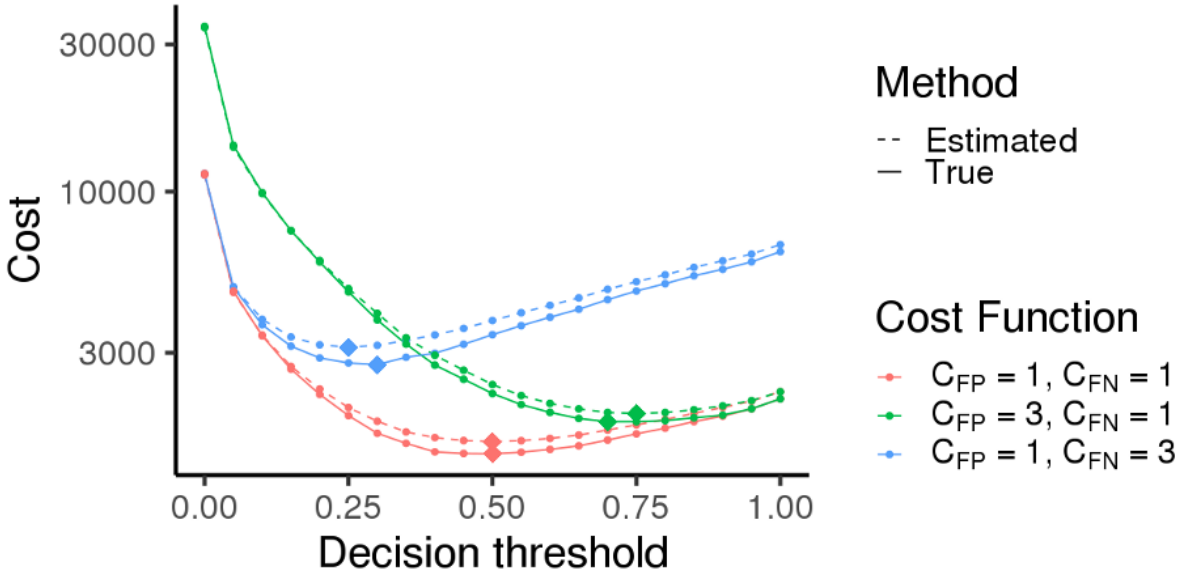

Figure 6. Flexible cost optimization with probabilistic individual stratification under various cost functions. Each color corresponds to one cost function: (a) equal cost for each FP and FN diagnosis (C_{FP} = C_{FN} = 1, red); (b) 3x higher cost for FP diagnoses (C_{FP} = 3, C_{FN} = 1, green); and (c) 3x higher cost for FN diagnoses (C_{FP} = 1, C_{FN} = 3, blue). Each line is a cost optimization curve for varying probability thresholds to classify high risk individuals. Solid line represents cost calculated using true genetic risk and dotted line represents cost estimated from individual probability of high risk. Diamond symbols represent the optimal threshold for each cost optimization curve. (simulation parameter $h_g^2 = 0.25, p_{\text{causal}} = 1\%$).

Trait	PRS < t ("Low")		PRS > t ("High")	
	# Certain	*# Certain/ (#Certain + # Uncertain)*	*# Certain*	*# Certain/ (#Certain + # Uncertain)*
		t = 90th		
Hair color	12326.0 (261.3)	64.4 (1.4) %	186.8 (20.1)	8.8 (0.9) %
Height	8037.4 (135.6)	42.0 (0.7) %	45.6 (7.5)	2.1 (0.4) %
Body mass index (BMI)	2669.2 (46.1)	13.9 (0.2) %	3.6 (1.3)	0.2 (0.1) %
High density lipoprotein (HDL)	7893.8 (295.5)	41.2 (1.5) %	44.0 (8.5)	2.1 (0.4) %
Low density lipoprotein (LDL)	9576.6 (527.6)	50.0 (2.8) %	56.4 (13.4)	2.7 (0.6) %
Cholesterol	8658.0 (376.9)	45.2 (2.0) %	44.2 (11.2)	2.1 (0.5) %
IGF1	5479.8 (332.7)	28.6 (1.7) %	13.0 (2.5)	0.6 (0.1) %
Creatinine	4006.0 (331.8)	20.9 (1.7) %	5.6 (3.6)	0.3 (0.2) %
Red blood cell count (RBC)	5443.8 (444.2)	28.4 (2.3) %	14.6 (4.4)	0.7 (0.2) %
White blood cell count (WBC)	3671.6 (348.8)	19.2 (1.8) %	4.0 (2.8)	0.2 (0.1) %
Bone mass density in heel (BMD)	3107.2 (615.3)	16.2 (3.2) %	3.2 (1.3)	0.2 (0.1) %
Average (s.d.)	**6442.7 (2937.2)**	**33.7 (15.3) %**	**38.3 (50.8)**	**1.8 (2.4) %**
		t= 99th		
Hair color	18941.2 (162.5)	89.9 (0.8) %	8.2 (2.4)	3.8 (1.1) %
Height	16383.2 (146.4)	77.8 (0.7) %	1.4 (1.1)	0.7 (0.5) %
Body mass index (BMI)	9592.4 (206.2)	45.5 (1.0) %	0.0 (0.0)	0.0 (0.0) %
High density lipoprotein (HDL)	16119.6 (297.7)	76.5 (1.4) %	0.4 (0.5)	0.2 (0.3) %
Low density lipoprotein (LDL)	16805.0 (479.3)	79.8 (2.3) %	1.4 (0.5)	0.7 (0.3) %
Cholesterol	16374.0 (467.3)	77.7 (2.2) %	1.2 (0.8)	0.6 (0.4) %
IGF1	13973.6 (371.1)	66.4 (1.8) %	0.4 (0.5)	0.2 (0.3) %
Creatinine	11885.8 (484.0)	56.4 (2.3) %	0.2 (0.4)	0.1 (0.2) %
Red blood cell count (RBC)	13908.2 (406.3)	66.0 (1.9) %	0.6 (0.9)	0.3 (0.4) %
White blood cell count (WBC)	11686.2 (605.9)	55.5 (2.9) %	0.0 (0.0)	0.0 (0.0) %
Bone mass density in heel (BMD)	10807.8 (1192.0)	51.3 (5.7) %	0.0 (0.0)	0.0 (0.0) %
Average (s.d.)	**14225.2 (2812.0)**	**67.5 (13.4) %**	**1.3 (2.3)**	**0.6 (1.1) %**

Table 1. PRS-based individual stratification uncertainty across 11 complex traits in UK Biobank. We quantified PRS-based stratification uncertainty in testing individuals for eleven complex traits at two stratification thresholds t = 90th and t = 99th. The uncertainty is determined with 95% credible interval ($\rho = 95\%$). For each trait, the number of certain low (high) risk and proportion of certain low (high) risk among low (high) risk diagnoses are reported as an average from five random partition repeats of the whole dataset. Numbers in the brackets are standard deviation of corresponding metric across five random partition repeats.

Trait	t = 90th		t = 99th	
	Lower bound	Upper bound	Lower bound	Upper bound
Hair color	64.2 (1.3)	97.6 (0.2)	90.4 (1.8)	99.8 (.06)
Height	50.6 (1.4)	98.3 (0.1)	81.9 (2.6)	99.9 (.04)
Body mass index (BMI)	31.0 (1.5)	98.9 (0.1)	59.3 (3.5)	99.9 (.04)
High density lipoprotein (HDL)	48.9 (1.7)	98.4 (0.1)	79.9 (2.9)	99.9 (.04)
Low density lipoprotein (LDL)	54.0 (2.4)	98.4 (0.2)	82.0 (3.2)	99.9 (.04)
Cholesterol	50.8 (2.0)	98.4 (0.2)	80.4 (3.2)	99.9 (.04)
IGF1	41.1 (1.9)	98.7 (0.1)	72.5 (3.4)	99.9 (.04)
Creatinine	35.7 (2.0)	98.8 (0.1)	65.9 (3.6)	99.9 (.04)
Red blood cell count (RBC)	40.8 (2.0)	98.7 (0.1)	72.5 (3.4)	99.9 (.04)
White blood cell count (WBC)	34.6 (1.7)	98.8 (0.1)	65.3 (3.8)	99.9 (.04)
Bone mass density in heel (BMD)	31.8 (2.5)	98.9 (0.1)	61.6 (4.6)	99.9 (.04)
Average (s.d.)	**44 (11)**	**98.5 (.03)**	**73.8 (9.9)**	**99.9 (0)**

Table 2. Average 95% posterior ranking interval for individuals at stratification threshold for 11 traits. We estimated the 95% ranking credible interval for individuals at 90th and 99th threshold of testing population PRS estimates. Mean and standard deviation are calculated from 95% posterior ranking intervals of one percent testing individuals around corresponding stratification threshold (213 individuals for each threshold within 89.5-90.5 percentile for t = 90th, 98.5-99.5 percentile for t = 99th respectively)

References

1. Sugrue, L. P. & Desikan, R. S. What are polygenic scores and why are they important? *JAMA* **321**, 1820–1821 (2019).
2. Natarajan, P. *et al.* Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation* **135**, 2091–2101 (2017).
3. Lee, A. *et al.* BOADICEA: a comprehensive breast cancer risk prediction modelincorporating genetic and nongenetic risk factors. *Genet. Med.* **21**, 1708–1718 (2019).
4. Khera, A. V. *et al.* Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell* **177**, 587-596.e9 (2019).
5. Hindy, G. *et al.* Genome-Wide Polygenic Score, Clinical Risk Factors, and Long-Term Trajectories of Coronary Artery Disease. *Arterioscler. Thromb. Vasc. Biol.* **40**, 2738–2746 (2020).
6. Wray, N. R. *et al.* Research review: Polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry* **55**, 1068–1087 (2014).
7. Fritsche, L. G. *et al.* Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *Am. J. Hum. Genet.* **102**, 1048–1061 (2018).
8. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28**, R133–R142 (2019).
9. Meisner, A. *et al.* Combined Utility of 25 Disease and Risk Factor Polygenic Risk Scores for Stratifying Risk of All-Cause Mortality. *Am. J. Hum. Genet.* **107**, 418–431 (2020).
10. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).
11. Seibert, T. M. *et al.* Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. *BMJ* **360**, (2018).
12. Dai, J. *et al.* 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**, 881–891 (2019).
13. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
14. Harrison, J. W. *et al.* Type 1 diabetes genetic risk score is discriminative of diabetes in non-Europeans: evidence from a study in India. *Sci. Rep.* **10**, 9450 (2020).
15. Läll, K., Mägi, R., Morris, A., Metspalu, A. & Fischer, K. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genet. Med.* **19**, 322–329 (2017).
16. Zhang, Q. *et al.* Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. *Nat. Commun.* **11**, 4799 (2020).
17. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
18. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
19. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology* vol. 41 469–480 (2017).
20. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **24**, 1550–1557 (2014).

21. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. 2020.04.28.066720 (2020) doi:10.1101/2020.04.28.066720.
22. Moser, G. *et al.* Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet.* **11**, e1004969 (2015).
23. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
24. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).
25. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
26. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).
27. Grinde, K. E. *et al.* Generalizing polygenic risk scores from Europeans to Hispanics/Latinos. *Genet. Epidemiol.* **43**, 50–62 (2019).
28. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).
29. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
30. Hu, Y. *et al.* Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput. Biol.* **13**, e1005589 (2017).
31. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**, (2019).
32. Kuchenbaecker, K. B. *et al.* Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. *J. Natl. Cancer Inst.* **109**, (2017).
33. Fahed, A. C. *et al.* Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.* **11**, 3635 (2020).
34. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
35. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).
36. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
37. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
38. Vaart, A. W. van der. *Asymptotic Statistics*. (Cambridge University Press, 1998).
39. Efron, B. & Tibshirani, R. J. *An introduction to the bootstrap*. (Chapman and Hall/CRC, 1994).
40. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
41. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).