Article

Machine Learning Prediction of Biomarkers from SNPs and of Disease Risk from Biomarkers in the UK Biobank

Erik Widen [1]*, Timothy G. Raben [1], Louis Lello [1,2]* and Stephen D.H. Hsu [1,2]

[1] Michigan State University, Department of Physics and Astronomy, 567 Wilson Rd, East Lansing, MI 48824
[2] Genomic Prediction, Inc., 675 US Highway One, North Brunswick, NJ 08902
* Correspondence: wideneri@msu.edu; (E.W); lellolou@msu.edu; (L.L)

Abstract: We use UK Biobank data to train predictors for 48 blood and urine markers such as HDL, LDL, lipoprotein A, glycated haemoglobin, ... from SNP genotype. For example, our predictor correlates ~ 0.76 with lipoprotein A level, which is highly heritable and an independent risk factor for heart disease. This may be the most accurate genomic prediction of a quantitative trait that has yet been produced (specifically, for European ancestry groups). We also train predictors of common disease risk using blood and urine biomarkers alone (no DNA information). Individuals who are at high risk (e.g., odds ratio of > 5x population average) can be identified for conditions such as coronary artery disease (AUC ~ 0.75), diabetes (AUC ~ 0.95), hypertension, liver and kidney problems, and cancer using biomarkers alone. Our atherosclerotic cardiovascular disease (ASCVD) predictor uses ~ 10 biomarkers and performs in UKB evaluation as well as or better than the American College of Cardiology ASCVD Risk Estimator, which uses quite different inputs (age, diagnostic history, BMI, smoking status, statin usage, etc.). We compare polygenic risk scores (risk conditional on genotype: (risk score | SNPs)) for common diseases to the risk predictors which result from the concatenation of learned functions (risk score | biomarkers) and (biomarker | SNPs).

Keywords: Polygenic Scores, Disease Risk, Machine Learning, Atherosclerotic Cardiovascular Disease, Biomarkers

Citation: Erik Widen; Timothy G. Raben; Louis Lello and Stephen D.H. Hsu . *Preprints* **2021**, *1*, 0. https://doi.org/

Received: preprint
Accepted: preprint
Published: preprint

1. Introduction

Modern machine learning (ML) methods have opened the door to using high dimensional inputs to predict health outcomes and risk. This paper concerns the application of sparse linear ML to genetic and health information in order to make predictions that could be useful in a clinical setting. Recent work has highlighted that ML, especially polygenic predictors, have high potential impact in clinical settings [1–21]. The UK Biobank (UKB)[22] dataset includes single nucleotide polymorphisms (SNP) genotypes, medical diagnosis information, and extensive biomarker information (i.e., 48 quantitative outputs of blood and urine tests) for almost 500k individuals. In this article we describe ML investigations of the correlation structure between these three categories of data. As described in Figure 1, we train:

1. Polygenic Score (PGS) predictors of the quantitative biomarker test results from *SNPs alone*. These functions predict biomarker level conditional on genotype: PGS = (biomarker | SNPs).

   For example, we predict measured lipoprotein A levels from SNPs, achieving a correlation of 0.76 between PGS and actual biomarker level. This may be the most accurate SNP prediction of a complex human trait yet accomplished.

2. Biomarker Risk Scores which predict risk of a specific disease condition *using only measured biomarkers as input*: (risk score | biomarkers).

   For example, our atherosclerotic cardiovascular disease (ASCVD) predictor uses $\sim$10 blood biomarkers to predict disease risk. We show that in UKB validation it predicts disease risk as well as or better than the American College of Cardiology ASCVD Risk Estimator [23,24], which uses quite different inputs such as age, diagnostic history, body mass index (BMI), smoking status, statin usage, etc. Liver and kidney problem

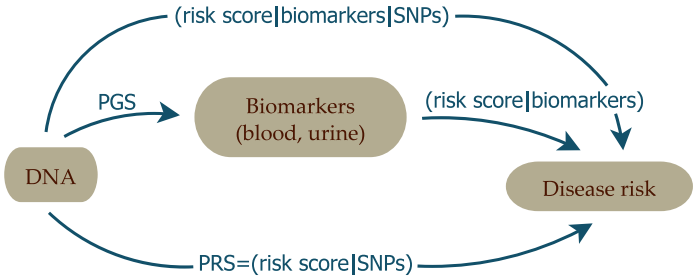


Figure 1. The four different types of predictors appearing in this paper.

risk prediction from biomarkers seems quite promising, based on our results. In total, we investigate predictions for ASCVD, coronary artery disease (CAD), diabetes type I & II, hypertension, very inclusive definitions of kidney and liver problems, and obesity.

3. Finally, by concatenating the predictors in 1 and 2 above, we build functions which map genotype (SNPs) to disease risk, with biomarkers as an intermediate step. We denote these concatenated predictors as: (risk score | biomarkers | SNPs). We emphasize that concatenation (i.e., $F(G(x))$) is *not* the same as training with both biomarkers and SNPs simultaneously used as features. The concatenated predictors *only* require SNPs as input, but use SNP predicted biomarker values as an *intermediate step* in calculation of the predicted disease risk. These functions can be compared to standard Polygenic Risk Scores (PRS) computed directly from SNPs, using disease case status as the training phenotype: PRS = (risk score | SNPs).

   For example, the concatenated function which maps SNPs $\rightarrow$ biomarkers $\rightarrow$ type 2 diabetes risk performs roughly as well as the PRS for type 2 diabetes (Area Under the Receiver operator characteristic Curve, AUC, ~ 0.64).

From our investigations, we conclude that many biomarker levels are not just substantially heritable, but can be predicted with some accuracy from SNPs. This is true despite the fact that levels fluctuate from day to day for a specific individual! We also conclude that disease risk prediction from biomarkers alone, via (risk score | biomarkers), is potentially very powerful, and indeed complementary to existing methods for risk estimation. For example, we show below that existing ASCVD risk predictors use different and complementary information to the biomarkers used in our ASCVD (risk score | biomarkers). Our results suggest that combining this complementary information can lead to stronger prediction and perhaps new insights into heart disease. Significant analyses of the costs and benefits of additional inputs have been performed for the existing ASCVD predictor, which is in clinical use (e.g. [23,24]), including some of the features in our predictor. Our comparison is limited to risk predictor performance and in the UKB cohort only.

We validate all predictors using sibling data: most of the power to differentiate between siblings (either in quantitative trait values or disease risk) persists despite similarity in childhood environments. We also test the fall off in power in distant ancestries (relative to the European training population). The decline for SNP based predictors varies as expected with genetic distance, whereas biomarker prediction does not display this pattern.

Throughout this paper, we refer to the different biomarkers according to the abbreviations listed in Table 1.

Table 1. List of all studied blood and urine markers with abbreviations.

Abbr.	Full name	Abbr.	Full name	Abbr.	Full name
ABC	Basophill count	DBil	Direct bilirubin	PCT	Platelet crit
AEC	Eosinophill count	E2	Oestradiol	phos	Phosphate
ALB	Albumin	GGT	Gamma glutamyltransferase	PLT	Platelet count
ALC	Lymphocyte count	gluc	Glucose	RBC	Red blood cell (erythrocyte) count
ALP	Alkaline phosphatase	HbA1c	Glycated haemoglobin (HbA1c)	RET	Reticulocyte count
ALT	Alanine aminotransferase	HCT	Haematocrit percentage	RF	Rheumatoid factor
AMC	Monocyte count	HDL	HDL cholesterol	SHBG	Sex hormone binding globulin
ANC	Neutrophill count	Hgb	Haemoglobin concentration	T	Testosterone
apoA	Apolipoprotein A	HLSR	High light scatter reticulocyte count	TBil	Total bilirubin
apoB	Apolipoprotein B	IGF1	IGF-1	TG	Triglycerides
AST	Aspartate aminotransferase	K	Potassium in urine	TP	Total protein
Ca	Calcium	LDL	LDL direct	U	Urea
chol	Cholesterol	LpA	Lipoprotein A	UA	Urate
Cr	Creatinine	MA	Microalbumin in urine	UCR	Creatinine (enzymatic) in urine
CRP	C-reactive protein	Na	Sodium in urine	vitD	Vitamin D
CysC	Cystatin C	NRBC	Nucleated red blood cell count	WBC	White blood cell (leukocyte) count

2. Materials and Methods

• Subject data

All research in this paper uses data exclusively from the 2018 UKB release [22,25] and updates (see Supplementary Information for more details). All statements about sex or ancestry refer to the self-reported data within this dataset [26]. There is of course a complicated genetic substructure within each one of these subgroups [27–40], however, it has been repeatedly demonstrated that self-reporting provides sufficiently good data for training purposes [41–45][1]. We refer to the self-reported ancestries labeled white, Asian, Chinese, and black in UKB as European, South-Asian, East-Asian and African, in accordance with the guidelines in [46]. It has been repeatedly confirmed that the power of polygenic predictors is dependent on both the training and testing ancestries, and that generally the power of the prediction falls off as a function of genetic distance [47–50]. All individuals with self-reported admixture were excluded from this study.

• Phenotype data

The phenotypes included in the paper include self-reported UKB statuses, standard (ICD9, ICD10, OPCS3, OPCS4) codes, diagnosed conditions, thresholds, and combinations of all the previous items. Full details of how each phenotype is defined is given in the Supplementary Information.

2.1. Predicting Biomarkers from SNPs

This work primarily focuses on LASSO [51], or compressed sensing [52–55]. LASSO was chosen because it has been repeatedly shown that sparse, linear methods are among the most successful in genetic prediction over a wide variety of traits[11,41,43]. Additionally, sparsity makes application and analysis of the predictors much more computationally efficient. As genetic predictors move into clinical settings, it will undoubtedly be the case that optimal prediction algorithms will vary depending on phenotype and training data, but LASSO currently serves as an excellent jack-of-all-trades. We used LASSO to predict the 48 types of biomarkers listed in Table 1 from SNP data, and denote these type of predictors as (biomarker | SNPs).

• Data and pre-processing

UKB contains data from repeated visits and for samples with more than one measurement of a certain biomarker the average value was taken. These raw measurements were

[1] Genetic prediction in general depends on non-trivial factors including population substructure, size of training sets, algorithms (e.g. sparse vs non-sparse methods), heritability, environmental factors, and etc. Nonetheless, in many instances self-reported identity is sufficient for training.

z-scored for men and women separately and consecutively age corrected by subtracting a linear regression on age-biomarker data obtained from averaging the biomarker value for all samples born the same year (biomarker-age plots are contained in Supplementary Information). The parameters for the pre-processing were determined from training sets with about 340k samples of European ancestry. Evaluation sets of about 20-40k European siblings and all non-European individuals were withheld entirely from training but pre-processed with the same parameters.

The UKB genotypic data were quality controlled by excluding all SNPs with less than 3% call success rate and also those with a minor allele frequency (MAF) < 0.001. All individuals with less than 3% successfully called SNPs were also excluded and, again, any individual with self-reported mixed ancestry was excluded from this study entirely. Furthermore, only autosomal genetic information was used, including SNPs located on chromosomes 1-22 only.

• Predictor training

Five LASSO predictors were trained on each biomarker using cross-validation, randomly drawing 1000 samples from the training set as validation set for each fold. The latter were used to choose optimal values of the regularization parameter λ (see Supplementary Information). The top performing predictor — as measured by correlation in the European corresponding validation set — from each fold was retained providing some statistics for the uncertainty estimates in the results. More details can be found in the Supplementary Information.

• Evaluation

Each predictor for each biomarker was evaluated on its corresponding evaluation set consisting of $\sim$ 20-40k samples of European ancestry. To test the performance dependence on ancestry we also applied the predictors to the 9k of South-Asian, 1500 of East-Asian, and 7k of African ancestry. In section 3.2.2, we report the correlation between the PGS and the phenotypes as the performance metric for these continuous traits.

Since environmental background, such as life style and diet, and indirect genetic effects have impacts on most of the biomarkers, we conducted a sibling evaluation. Siblings generally have more similar backgrounds than randomly chosen pairs, and are also more genetically similar than unrelated individuals. Retained predictive power among siblings is hence a strong indication of direct genetic effects. Moreover, the amount of lost power as compared to the general population can give some idea of the magnitude of environmental effects, e.g., from childhood environment. (There can also be genetic nurture [56–60] effects that are not analyzed here.) Childhood environments are more similar among siblings than between unrelated individuals, and this comparison gives an indication of whether the instantaneous biomarker measurements in adulthood are sensitive to the effects of childhood environment. To this end, we constructed both random pairs and pairs of genetic siblings within the evaluation set of European ancestry. For each pair, we calculated the difference in phenotype Δ_{phen} and the difference in PGS Δ_{PGS} and compared the correlations between these quantities $\text{corr}(\Delta_{\text{phen}}, \Delta_{\text{PGS}})$ within random and sibling pairs, respectively.

• Genetic architecture

One can define the variance accounted for by each SNP i in a predictor according to

$$\text{variance accounted for by SNP}_i = \beta_i^2(1-f_i)f_i\,, \tag{1}$$

where f_i is the MAF of SNP i. This is described in greater detail in the Supplementary Information and in [42]. We use this alongside Manhattan-plots of the effect sizes β in the results (section 3.1.1) to display the genetic architectures of the top 3 performing (biomarker | SNPs) predictors. Analogous plots for the rest of the (biomarker | SNPs) predictors are contained in the Supplementary Information.

2.2. Methods for disease prediction

We used two approaches to investigate whether biomarkers can be used to predict disease risk, analogous to how blood tests are used clinically.

Approach 1: We trained predictors to predict case/control status directly from phenotypes, i.e., using the direct biomarker measurements as features. We denote this type of predictor (risk score | biomarkers), or biomarker risk score.

Approach 2: Second, we applied these already trained predictors to the *predicted* phenotypes, i.e., using the biomarker PGS output from the SNP-based predictors in section 2.1 as input. As such, we obtain disease risk scores using only SNP data as input. We denote these concatenated predictors (risk score | biomarkers | SNPs).

We evaluated this strategy on eight different condition definitions and we present the details for the two approaches separately. This is done both to display the performance dependence on the two approaches and since the prediction from biomarkers in approach 1 are very interesting in their own right.

2.2.1. Approach 1: Predicting case status from biomarkers

• Condition definitions

Based on the available UKB data, we defined conditions for CAD, cancer, diabetes type 1, diabetes type 2, hypertension, kidney problem, liver problem, and obesity. The detailed definitions for each one of these are to be found in the Supplementary Information. In general, we chose the definitions to be inclusive; kidney (liver) problem for example contains almost all kidney (liver) related problems that are reported in UKB, whereas cancer refers to any type of cancer. Obesity was defined as a BMI over 30. The effects of changing definitions are further discussed in section 4.

• Predictor training

We used 45 out of the 48 biomarkers as input features, dropping E2, MA, and RF due to few available measurements, and taking the first available measurement for each sample. The raw data was pre-processed by sex specific z-scoring and then age correcting by subtracting a linear regression. Using LASSO, we then trained 5 predictors on the case/control status, choosing optimal λ by five-fold cross-validation. The training was done separately for men ($N = 106,656$) and women ($N = 86,193$) and on European ancestry only.

• Evaluation

As was done for the (biomarker | SNPs) in section 2.1, about 40k siblings of European ancestry and all non-European individuals were kept separate from all training and were used as evaluation set. We measured the predictor performance by AUC and by odds ratio plots. Additionally, we conducted sibling tests for the (risk score | biomarkers) predictors to test for environmental effects: we applied the predictors to pairs of siblings with precisely one case and one control and report the fraction of correctly called affected sibling, juxtaposed with the same results for random pairs of one case and one control.

It should be emphasized here that we did not take date of onset into account in this study: disease status was considered on a "life span" (as far as UKB covers) basis such that cases could have onsets both prior to and after the time of the biomarker measurement. Prediction in this sense means what can we predict about current or future case status only knowing a set of momentary biomarker values. Temporal prediction tests (i.e., prospective prediction) are deferred to later work.

2.2.2. Approach 2: Predicting case status from PGS of biomarkers

To form predictors taking SNP data as input, we concatenated the PGS predictors from section 2.1 with the biomarker predictors from approach 1, what we call (risk score |

biomarkers | SNPs). The disease predictors (risk score | biomarkers) were taken as is from Approach 1 and applied to the z-scored PGS output of the predictors in section 2.1. No further training was done and the performance was evaluated as for and compared with the (risk score | biomarkers) predictors.

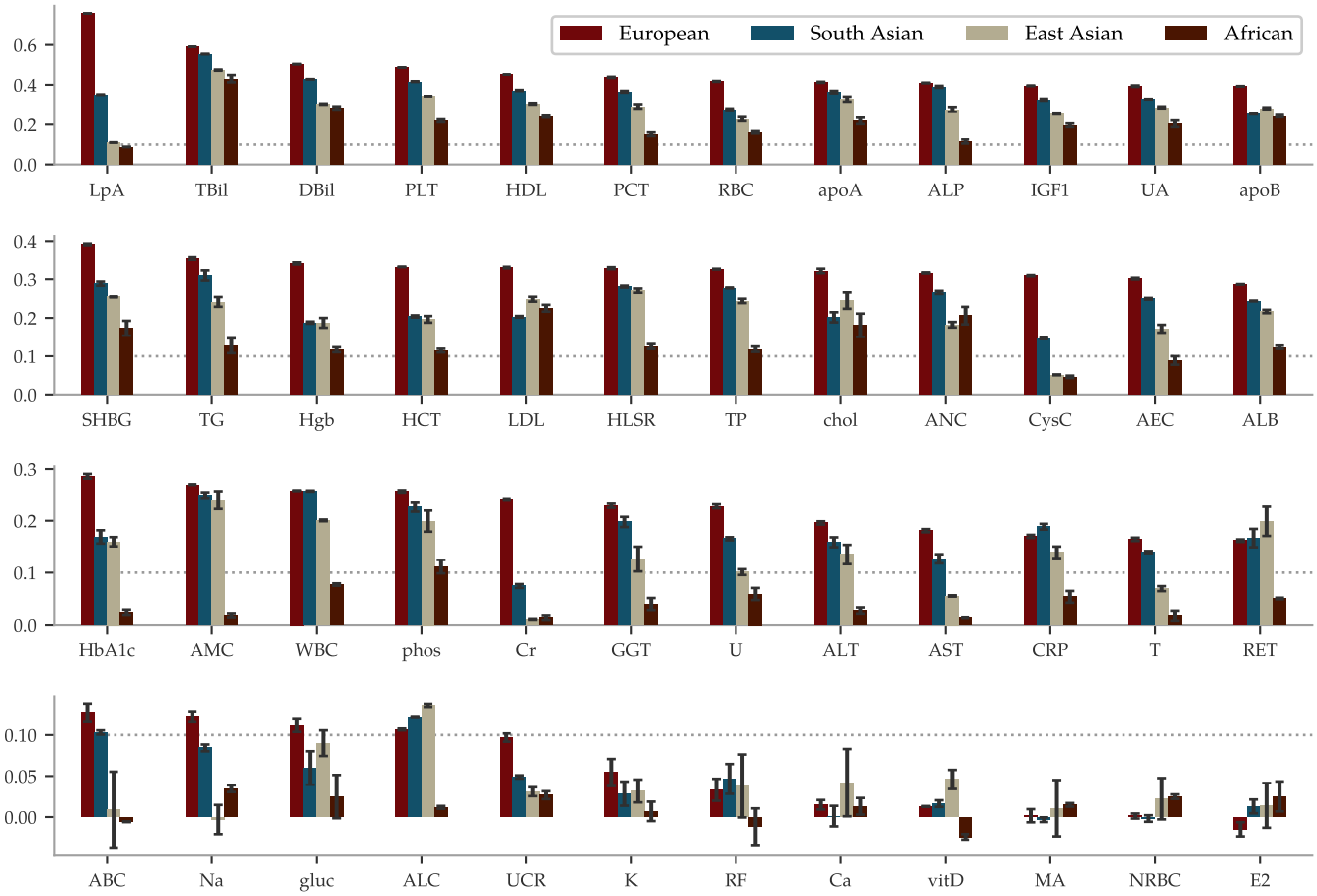


Figure 2. Correlations between PGS and phenotype vary from very strong to effectively zero, depending on the biomarker, and fall off with genetic distance from the training population. The mean of the PGS-phenotype correlation for evaluation sets are listed for all 48 biomarkers, ordered according to the results within Europeans — the ancestry for the training population. The error bars represent ± the standard deviation for 5 different predictors trained on slightly different training sets. The dotted line is there to aid graphical comparisons across the rows. The LASSO predictor of lipoprotein A achieves a correlation of 0.759 within European ancestry, which is the highest correlation for a polygenic trait we are aware of. The correlation fall-off for the other ancestries generally follows the order European > South Asian > East Asian > African. Note that the sample sizes for these ancestries are much smaller.

2.3. Comparison with ASCVD Risk Estimator

The ASCVD Risk Estimator [24] is a widely used tool to aid clinicians in risk estimations of and preventative care against atherosclerotic cardiovascular disease. We used this well-established resource to benchmark the approach of (risk score | biomarkers) predictors by training a predictor on this condition specifically. ASCVD aggregates several sub-diagnoses and exists in different versions. Hard ASCVD includes acute coronary syndromes, death by coronary heart disease, a history of myocardial infarction, and fatal and non-fatal stroke. A more general (extended) ASCVD definition additionally includes stable or unstable angina, coronary or other arterial revascularization, transient ischemic attack, and peripheral arterial disease presumed to be of atherosclerotic origin. We used a UKB specific extended definition, detailed in the Supplementary Information. The ASCVD

Risk Estimator requires the input: age, sex, race, systolic and diastolic blood pressure, total cholesterol, HDL, LDL, history of diabetes, smoking status, time since quit smoking (if applicable), whether on hypertension treatment, whether on a statin, and whether on aspirin. It can also use previous data for follow-ups but we restricted our analysis to "first visit patients" only. All of these data fields can be found in some form in the UKB (the exact field choices are listed in the Supplementary Information).

The outputs of the ASCVD Risk Estimator are (up to) three risk estimates: 10 year risk, lifetime risk, and optimal risk, all given as a percentage. Since our UKB data only cover approximately 10 years from the first biomarker measurement, we exclusively used the 10 year risk output. We applied the underlying function of the ASCVD Risk Estimator to the corresponding data in UKB and obtained a 10 year risk estimate for 358,650 individuals for whom we also had an ASCVD case/control status. Strictly speaking, the ASCVD Risk Estimator was developed for North American cohorts and and based on hard ASCVD but, as seen in section 3.3, performed very well also in the cohorts of the UKB using the extended definition. Note, however, the current comparison is not intended as a rigorous test for deployment.

We then trained a (risk score | biomarkers) predictor on case/control status, analogously to approach 1 in section 2.2, but using ordinary linear regression on the z-scored biomarker measurements. This outputs a risk *score* which we mapped to absolute risk *estimates in percentages* as follows. The risk scores obtained from applying the predictor on the training data were binned and, within each bin, the disease prevalence was calculated from the case/control statuses as an estimated risk for samples with the corresponding risk scores. This discrete mapping was then made continuous using rolling averages and linear interpolation. For details see Supplementary Information.

2.3.1. Combination of predictor from biomarkers and the ASCVD Risk Estimator

In the results section 3.3, we show that the ASCVD (risk score | biomarkers) predictor and the ASCVD Risk Estimator are making complementary predictions. We therefore also tested a combination of them. We made a linear regression on all the input features from the two predictors combined (48 continuous and 8 discrete variables), z-scoring the discrete variables from the ASCVD Risk Estimator input so that everything was on the same scale. In addition, we made a second regression also including the *output* of the ASCVD Risk Estimator to capture the non-linearities within that function. These regressions were made and evaluated on the same training and evaluation sets as for the (risk score | biomarkers) predictors.

3. Results

3.1. Predicting Biomarkers from SNPs

The performance of the (biomarker | SNPs) predictors ranges from the highest phenotype-PGS correlation for a polygenic predictor we are aware of to no predictive power whatsoever. We present the results in order of correlation within European ancestry in Figure 2. The best performing predictor is for lipoprotein A at a correlation of ~ 0.76. This is not too surprising as lipoprotein A levels are well-known to be highly heritable [61–64], related to the LPA gene and other loci [65–74], and thus do not greatly vary by life style or environment [2]. Yet, it is a striking example of predictive power. After lipoprotein A, we find correlations almost evenly distributed within the correlation range 0.1-0.59 and a group of 7 almost uncorrelated biomarkers at the bottom. In the same Figure 2, we have included the performance within the non-European ancestries. Being trained on European ancestry only, the predictors suffer the now familiar [49,50] fall-off pattern according to genetic distance, with performance generally being successively worse for South-Asian, East-Asian, and African ancestries.

[2] Lipoprotein A has long been studied because of its association with CAD, atherosclerotic risk, liver problems, metabolism, and even cancer. Further discussion can be found in the review [75]

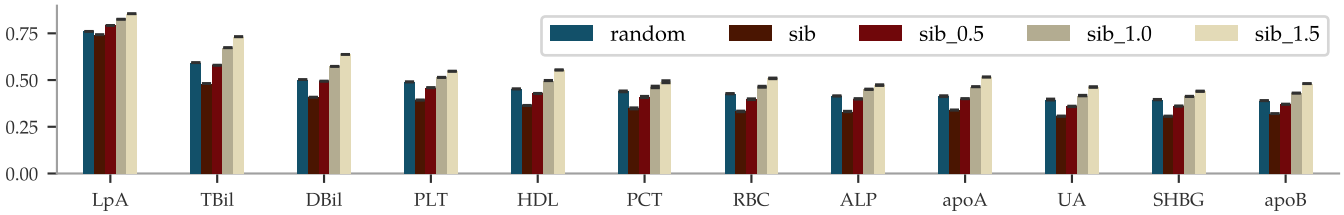


Figure 3. Sibling comparisons of correlation between difference in phenotype and difference in PGS, i.e., $\mathrm{corr}(\Delta_{\mathrm{phen}}, \Delta_{\mathrm{PGS}})$, show that most of the correlation is retained also for pairs that share similar environmental backgrounds. UKBs $\tilde{4}$0k siblings of European ancestry were paired either randomly or as genetic siblings and were used as test set. The correlations between the pairs' differences in phenotype and their differences in PGS was then calculated for each biomarker, ordered above from strongest to weakest correlation. The error bars indicate $\pm$ the standard deviations for 5 predictors trained on slightly different training sets. The additional three bars ▮ labeled sib 0.5, sib 1.0 and sib 1.5, are the results when restricting to siblings with phenotype differences larger than 0.5, 1 and 1.5 standard deviations, respectively. Two siblings are likely to have more similar environmental backgrounds than random pairs, affecting the similarity of late-life biomarker measurements independently from (direct) genetic effects. This could explain the decreased correlation for siblings as compared to random pairs. Yet, the remaining correlations are strong evidence that the predictors capture some direct genetic effects on the biomarkers. The comprehensive figure for all biomarkers can be found in the Supplementary Information.

The results from the sibling comparison can be seen in Figure 3. On average, there is a $\sim$ 26% drop in correlation when comparing differences within random pairs and differences within sibling pairs. The figure also shows that siblings that are separated by more than 0.5, 1.0, and 1.5 times the standard deviation in phenotype are predicted with increased correlation. The sibling comparisons for the other biomarkers can be found in the Supplementary Information.

3.1.1. Genetic Architecture

Polygenic predictors have shown to usually use information spread over the entire genome, even when enforcing sparsity [11,41,42,45]. In Figure 4, we illustrate the genetic architectures behind three of the top performing (biomarker | SNPs) predictors with Manhattan plots of the effect sizes β and the variance accounted for in eq. (1), accumulated across chromosomes 1-22 (the Supplementary Information contains figures for all biomarkers). It shows that both biomarkers with a few very strong loci and biomarkers with an evenly distributed dependence can be predicted well. Let us make a few remarks on the top 5 performing predictors (see Supplementary Information for the direct bilirubin and platelet count plots):

- The lipoprotein A predictor is as expected totally dominated by the single locus on chromosome 6, the gene carrying its name LPA.
- The total bilirubin predictor is very similar to the one for direct bilirubin. GWASes have implicated many variants on all but chromosome 15 (according to a GWAS Catalog[76] trait search) but most have a very minor impact on our predictor. For example, [77] reported a locus on chromosome 19 but although there are groups of moderately large β in this region, the entire chromosome 19 does not account for more than $\sim$ 1% of total variance in our predictors.
- GWASes for direct bilirubin in the literature [77,78] are generally dominated by variants in gene UGT1A on chromosome 2. The LASSO predictors pick these up too. In addition, there is another $\sim$ 17% variance accounted for by the locus at chromosome 12, also known[78]. Chromosomes 6 and 19 account for $\sim$ 1% variance each and have no generally listed loci. The β_i with the largest magnitude corresponds to SNP rs908327 on chromosome 1. It has SNPs in linkage disequilibrium (LD) that have been

linked to triglycerides[79] but not directly to bilirubin, to our knowledge. It has a very small MAF, however, and does not account for much variance.

- The predictor for platelet count is very polygenic with the variance accounted for almost evenly distributed across all 22 chromosomes. Chromosome 12 provides a small deviation from this pattern, accounting for $\sim$ 14% of the variance, partly due to a locus near one end.
- The predictor for HDL is also highly polygenic. Previous GWASes have recorded loci at all but chromosome 13, which has no large magnitude β_i but still accounts for $\sim$ 1% of the total variance.

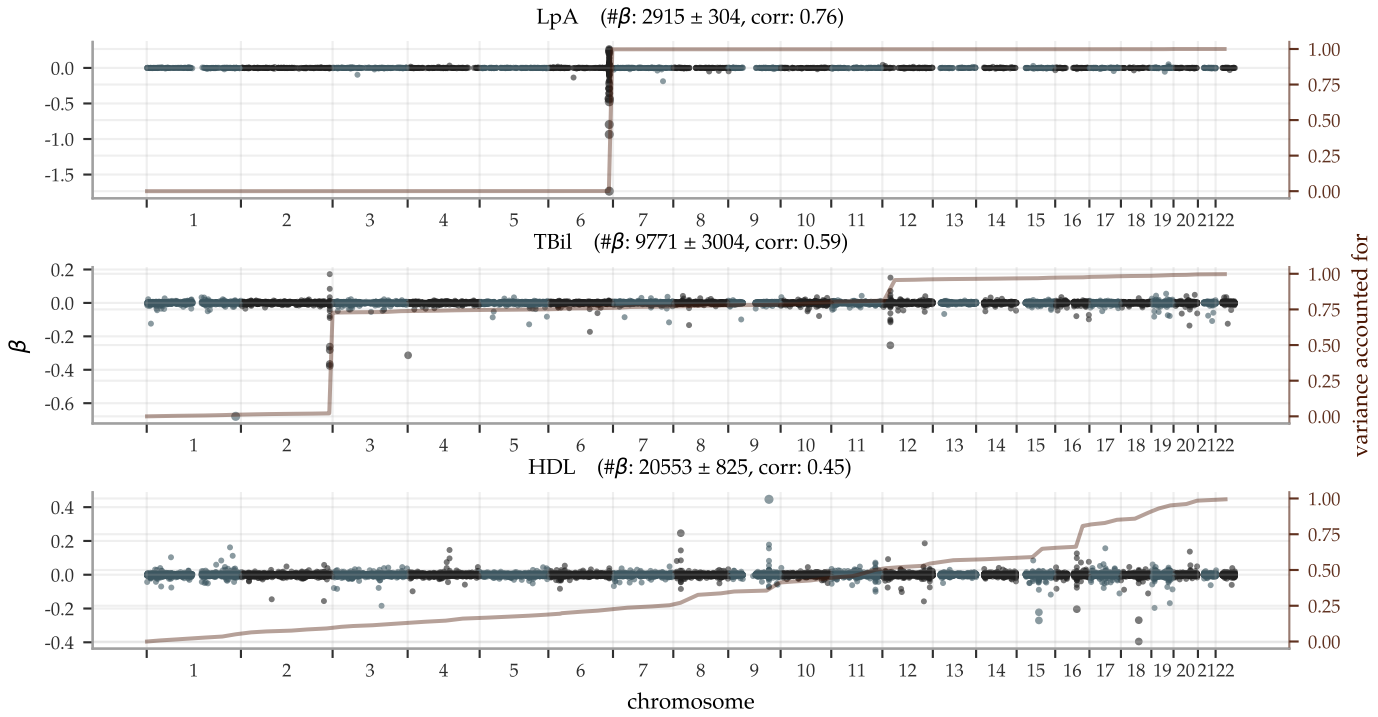


Figure 4. Manhattan plots of LASSO β — superimposed with the aggregate single snp variance accounted for — show both highly localized as well as widely polygenic architectures. The predictor for Lipoprotein A is almost entirely determined by the well-known gene LPA in chromosome 6; the top 50 SNPs in this region account for $\sim$ 95% of the aggregate single SNP variance. In contrast, HDL has an almost uniform distribution of the variance accounted for across all the 22 autosomal chromosomes, despite some loci with high magnitude β-coefficients. (The difference being due to the MAF in equation (1).) The most significant genetic loci are discussed further in the main text. The plot titles include the achieved PGS-phenotype correlation and mean number of non-zero $\beta \pm$ the standard deviation for the 5 predictors trained on each trait.

3.2. Predicting Disease Risk

The results for the disease risk predictors are divided into sections corresponding to the (risk score | biomarkers) from approach 1 and (risk score | biomarkers | SNPs) from approach 2, respectively.

3.2.1. Predicting case status from biomarkers

The performance of the (risk score | biomarkers) predictors was evaluated and are reported as AUCs and odds ratio plots in Figure 5. With training optimized for European ancestry, we regard the results for this ancestry as the main results and provide the performance in other ancestries for reference. The results vary with the condition. Within European ancestry, they range from an AUC of .53 (.60) for cancer for women (men) up to $\sim$ .95 for diabetes type 1 (both sexes). As a comparison, we report below on an ASCVD predictor with an AUC of $\sim$ .76 which performs risk prediction as well as or better than

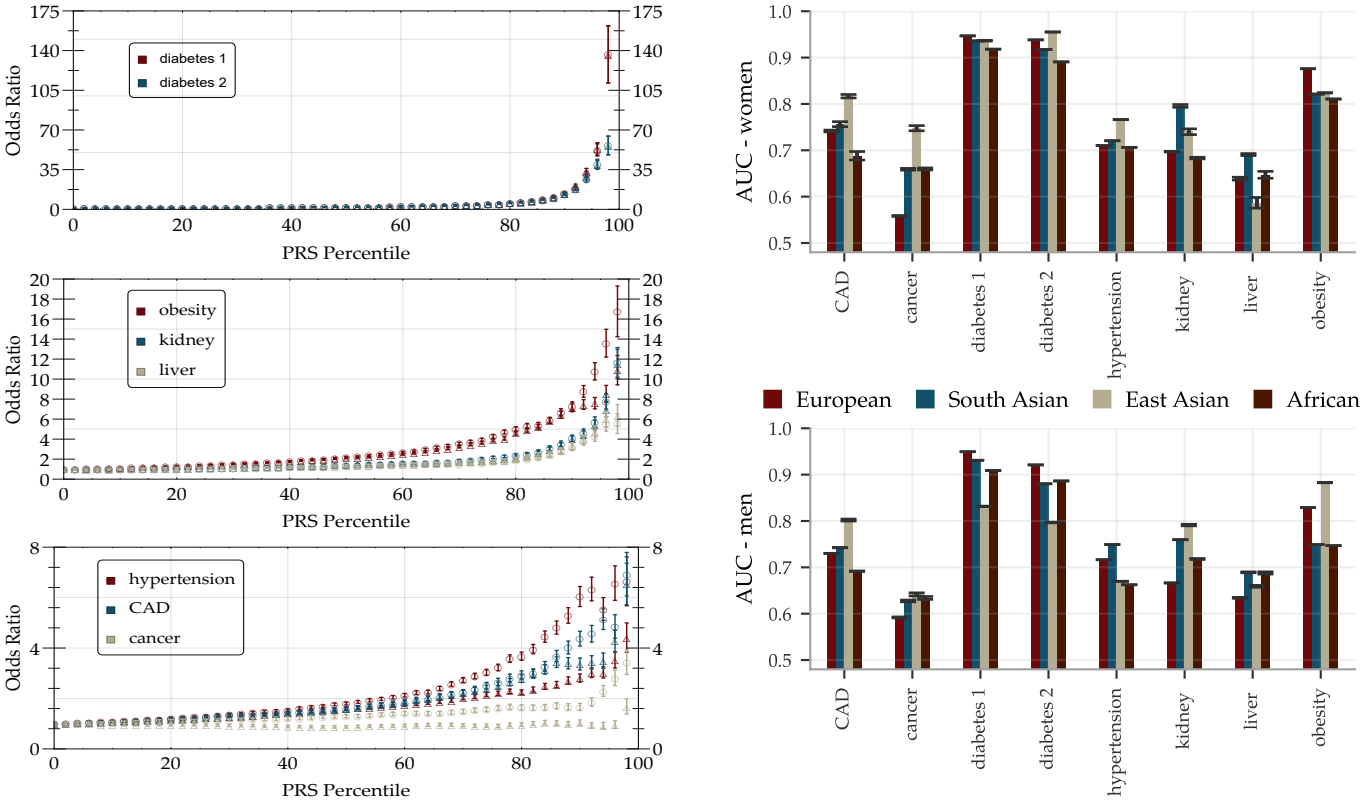


Figure 5. The predictive power of (risk score | biomarkers) can single out high risk individuals with over 10x odds ratio for many traits, and AUCs > 0.7 for most traits including tests across ancestry. Left: inclusive odds ratio (OR) plots for diabetes type 1/2, obesity, kidney problem, liver problem, hypertension, CAD, and any cancer trained and validated on the European population. Horizontal axis indicates individuals at that percentile *and above* in PRS. Marker ○ is for predictors trained and validated on men and marker △ for predictors trained and validated on women. Error bars represent the standard error of the mean value with a contribution coming from computing the OR and a contribution from including 5 predictors. **Right:** AUCs for (risk score | biomarkers) predictors separately trained on men and women. All predictors are trained on the European population and then validated on European, South Asian, East Asian, and African populations. The error bars indicate the standard deviations for 5 different predictors and do not reflect the significant uncertainties arising from limited available statistics (sample sizes are listed in Supplementary Information).

the American College of Cardiology ASCVD Risk Estimator. We discuss this in detail below in section 4. The odds ratio plots show a wide range of results that also vary with condition. Figure 5 separates conditions into groups based on the odds ratios of the high risk outliers. The strength of the diabetes predictors is probably due to their use of blood biomarkers (e.g. HbA1c) which are standard diagnostic indicators for diabetes. That this standard diagnostic indicator is so highly ranked lends confidence to the results of the general methodology.

There are some differences in performance for men and women, most notably in cancer (possibly due to sex specific cancer variants). The differences are condition specific and viewed across all conditions the performance is similar. We delay a more detailed analysis of these differences to future study. The reported performance variations across the different ancestries are notably smaller and show less of a consistent pattern than what is the usual case for prediction from genetic information; this is expected since predicting from biomarkers stays on a higher biological level and does not involve issues such as LD patterns and tag SNPs etc. Note, however, that these results are limited by the available statistics, see Supplementary Information for the case/control numbers for each ancestry.

In Figure 6, we also include two examples of the LASSO coefficients for CAD and type 2 diabetes. For CAD, we find mostly well-known biomarkers with the highest weight, such as LDL, apolipoprotein B, total cholesterol and HDL. However, for women cystatin C

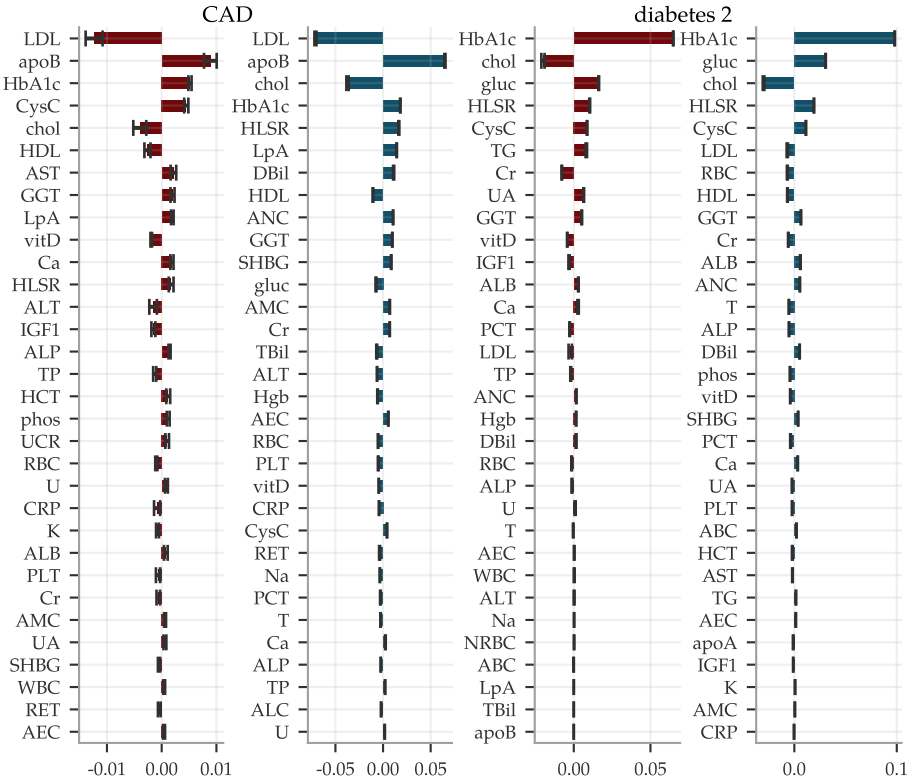


Figure 6. Predictors for phenotypes like CAD and type 2 diabetes from biomarkers are dominated by a top few inputs. Relative weights of each biomarker within predictors for CAD and type 2 diabetes. ■ women and ■ men while error bars indicate ± standard deviations from the mean of five predictors. The most impactful biomarkers are very well-known but we highlight cystatin C as surprisingly frequent among the moderately strong coefficients. Corresponding plots for all condition predictors are shown in the Supplementary Information.

appears at fourth place, which to our knowledge is not often used in this context. Cystatin C also is the fifth most influential biomarker in the diabetes type 2 predictor for both sexes, while these predictors are dominated by the standard biomarker glycated haemoglobin. In fact, cystatin C is among the more important biomarkers for most of our predictors. Coefficients for all conditions are listed in the Supplementary Information.

We investigated the presence of non-linear effects for (risk score | biomarkers) by extending the input features with all possible quadratic interactions among the seven most influential biomarkers for each condition. We saw no effect on the performance in either direction and conclude that the effects of the biomarkers on all the listed conditions appear to be linear to very good approximation.

3.2.2. Predicting case status from PGS of biomarkers

The concatenated predictors (risk score | biomarkers | SNPs) suffer a significant drop in performance, as can be seen in Figure 7. The imprecise predictors (biomarker | SNPs) introduce a lot of noise and, exacerbated further by the uncertainty in the (risk score | biomarkers) predictors, the concatenation does in general not lead to meaningful predictions. A notable exception are the diabetes predictors. The combination of reasonably correlated PGS for the most important biomarkers and the exceptionally high AUCs for these predictors lead to an average AUC of $\sim$ .63 for the type 2 diabetes (risk score | biomarkers | SNPs) predictor. This is comparable to what we have achieved in the past by training SNP-based LASSO directly on type 2 diabetes status[11]. Furthermore, the two different types of predictors (risk score | biomarkers | SNPs) and (risk score | SNPs) capture somewhat complementary information, as shown in Figure 8. The sum of the two types of risk scores reaches an AUC of $\sim$ .67. It is unclear why the use of biomarkers as an intermediate step adds additional information relative to training directly with SNPs as features and case status as the phenotype. We leave this as an interesting topic for future research.

The sibling evaluation of the disease risk predictors, described in section 2.2, is reported in Figure 9. The fraction of sibling pairs with one case and one control called

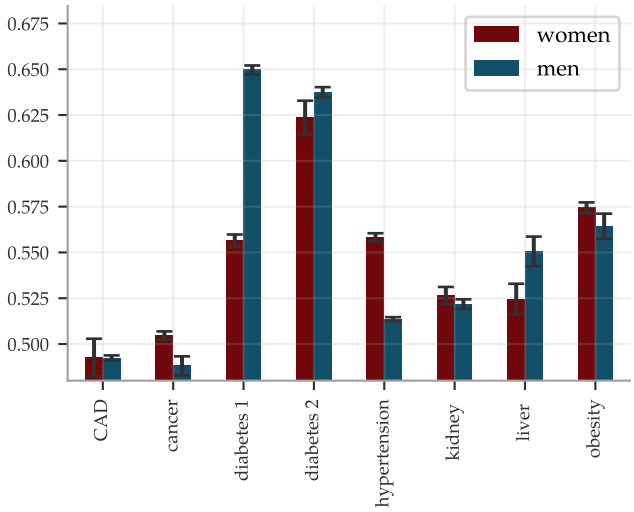


Figure 7. AUCs for (risk score | biomarkers | SNPs) predictors drop significantly as compared to (risk score | biomarkers) in Figure 5 and only the diabetes predictors reach par with other methods. The predictors were evaluated on 9016 (9607) white women (men) and the error bars indicate ± the standard deviation for 5 different predictors.

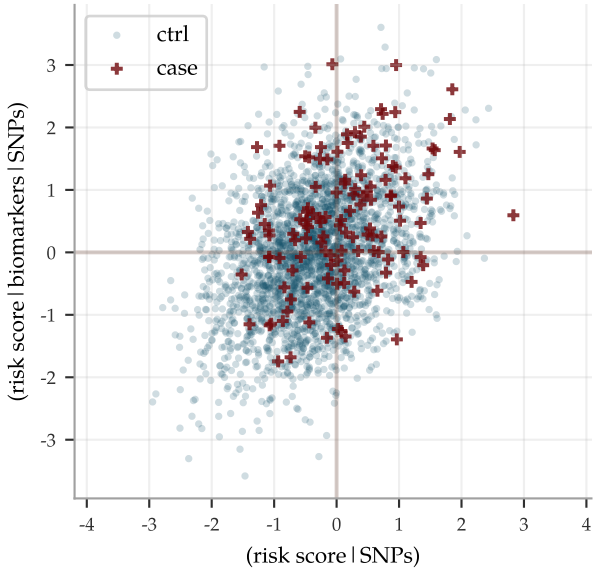


Figure 8. Risk scores predicted from SNPs (risk score | SNPs) and from PGS of biomarkers (risk score | biomarkers | SNPs) do not always agree, here exemplified by type 2 diabetes data for men. Both predictors predict case status directly from SNPs alone. Their outputs correlate ∼ 0.37 with a linear regression coefficient of ∼ 0.39. In the noise, they capture some complementary information: the sum of the risk scores achieves an AUC of ∼ 0.67 while the SNP and PGS based predictors individually achieve AUCs of ∼ 0.63 and ∼ 0.65, respectively.

correctly ranged from pure chance for cancer and liver problems, while reaching ∼ 0.9 for diabetes type 1 and 2, using the (risk score | biomarkers) predictors. The accuracy dropped significantly for the (risk score | biomarkers | SNPs) predictors, as expected; no predictor of this type reached a correctly called fraction above 0.6.

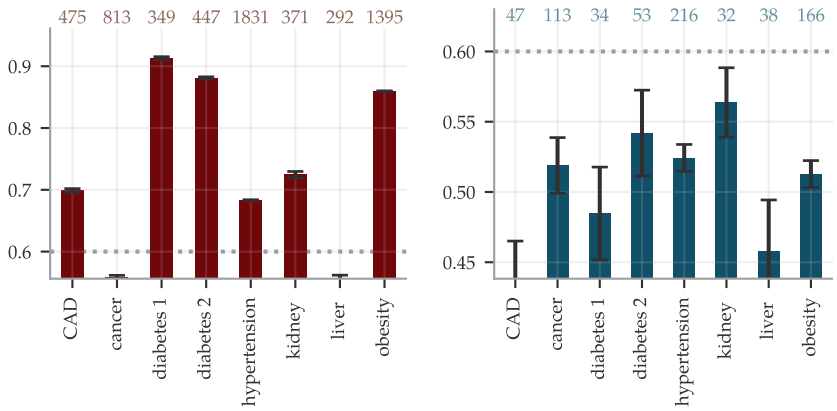
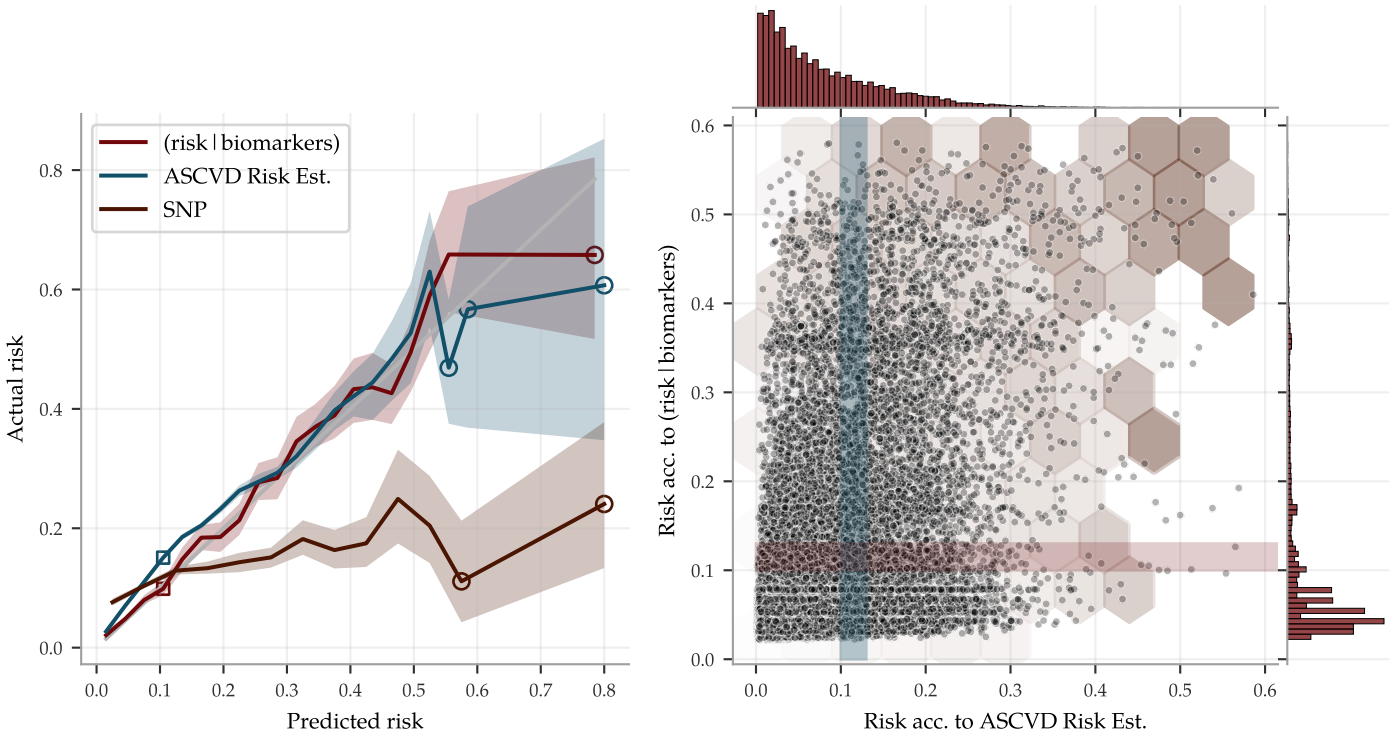


Figure 9. The fractions of sibling pairs with precisely one case and one control called correctly are generally high for ■ (risk score | biomarkers) but not much better than chance when predicting from genotypes using ■ (risk score | biomarkers | SNPs). The pairs were considered correctly called if the PRS was higher for the affected sibling, without any restriction on the size of the separation. Number of included sibling pairs differed for the two types of predictors and are listed at the top. The error bars indicate ± the standard deviation for five different predictors for (risk score | biomarkers) and for 5×5 concatenation combinations of predictors in the (risk score | biomarkers | SNPs).

3.3. Comparison with ASCVD Risk Estimator

To illustrate the performance of the (risk score | biomarkers) predictor for ASCVD and to compare it with the ASCVD Risk Estimator, we used the risk percentage output as described in section 2.3. The ASCVD Risk Estimator was built using American cohorts of separately European and African ancestry. Due to the similarities with the UKB population, we deemed it could be applied somewhat fairly to the entire UKB, whereas we used the withheld evaluation set of $\sim$ 40k of European ancestry for the (risk score | biomarkers) predictor. The result is shown in Figure 10, in which the predicted risks were binned and the actual disease prevalence within each bin was calculated, labeled "Actual risk". Both predictors give very accurate risk estimates, with increasing uncertainty for individuals with high predicted risk. However, although they do assign correct risk estimates for bins taken as a whole, they do not always agree on who is at low versus high risk. The scatter plot in Figure 10 shows their individual distributions and occasional disagreements. Their partially complementary predictions are further highlighted in the risk heat map in Figure 10 and utilized below in a combined predictor.



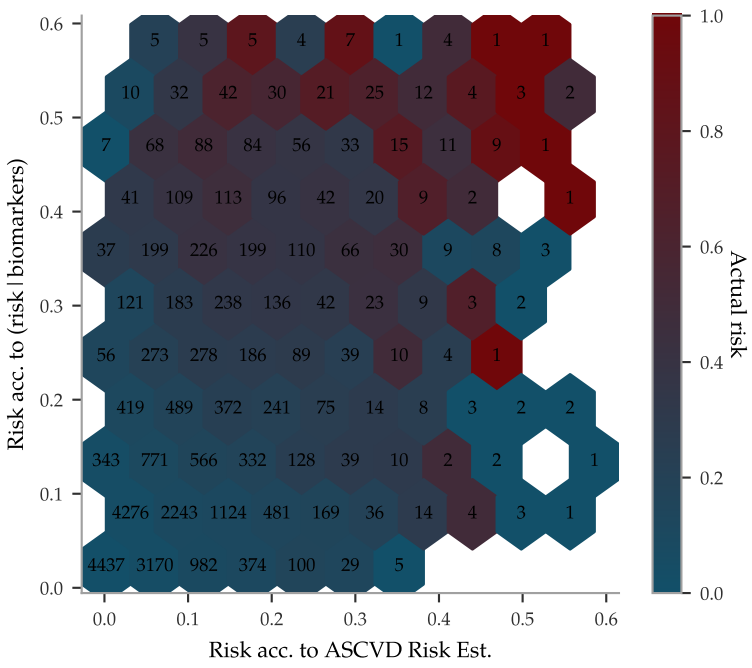


Figure 10. The ASCVD (risk score | biomarkers) and the ASCVD Risk Estimator both make accurate risk predictions but with partially complementary information. Left: Predicted risk by (risk score | biomarkers), the ASCVD Risk Estimator and a (risk score | SNPs) predictor were binned and compared to the actual disease prevalence within each bin. The gray 1:1 line indicates perfect prediction. Shaded regions are 95% confidence intervals obtained from 100 fold bootstrap estimates of the prevalence in each bin. The ASCVD Risk Estimator was applied to 340k UKB samples while the others were applied to an evaluation set of 28k samples, all of European ancestry. **Upper right** shows a scatter plot and distributions of the risk predicted by (risk score | biomarkers) versus the risk predicted by the ASCVD Risk Estimator for the 28k Europeans in the evaluation set. The (risk score | biomarkers) distribution has a longer tail of high predicted risk, providing the tighter confidence interval in this region. The left plot y-axis is the actual prevalence within the horizontal and vertical cross-sections, as illustrated with the shaded bands corresponding to the hollow squares to the left. Notably, both predictors perform well despite the differences in assigned stratification. The hexagons are an overlay of the **lower right** heat map of actual risk within each bin (numbers are bin sizes). Both high risk edges have varying actual prevalence but with a very strong enrichment when the two predictors agree.

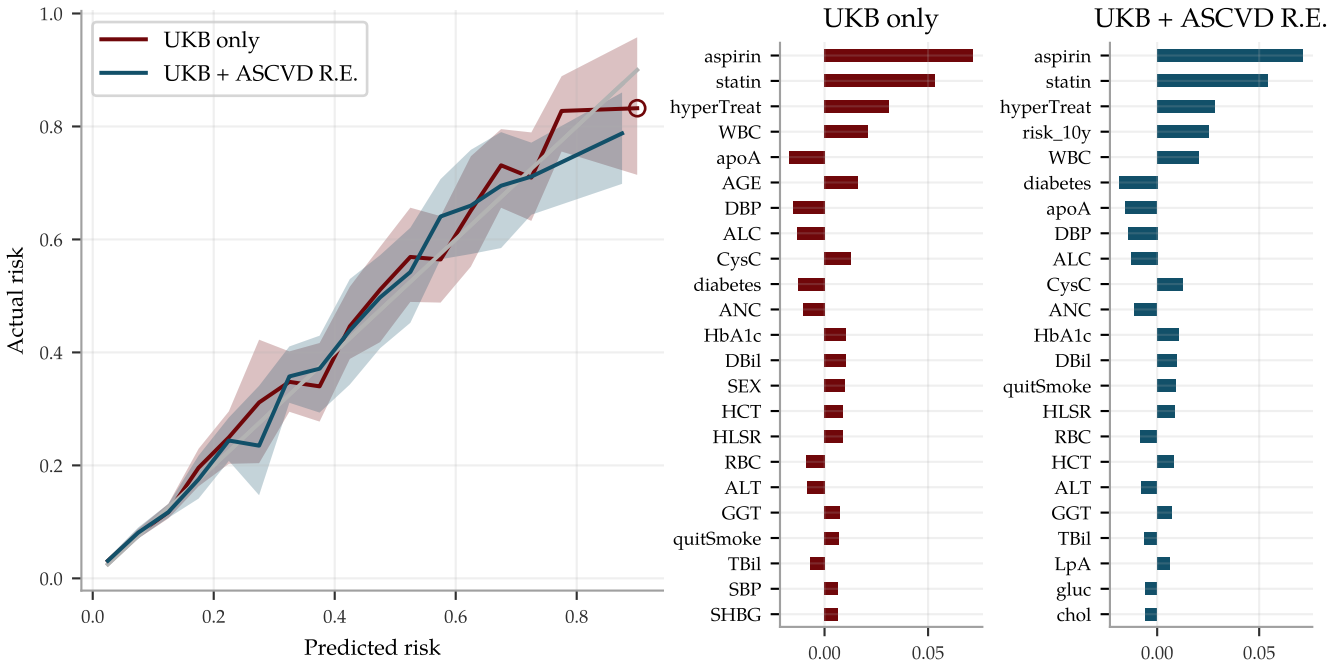


Figure 12. The risk prediction using both 45 biomarkers and all the ASCVD Risk Estimator input improves performance as compared to Figure 10, in particular for high risk individuals, and is very good all the way up to risk levels of 80%. The figure compares two predictors: ■ a combined ASCVD predictor using all 45 biomarkers plus all the input fields (age, sex, etc.) used by the ASCVD Risk Estimator, using UKB data only, and ■ a predictor using the same input plus the ASCVD Risk Estimator *output*, labeled UKB + ASCVD R.E. The latter does not perform notably better, although the ASCVD Risk Estimator output "risk_10y" corresponds to the fourth strongest coefficient. Both perform better than both the (risk score | biomarkers) and ASCVD Risk Estimator individually, confirming their complementary nature shown in the heat map, Figure 10. The shaded areas in the left panel again indicate 95% confidence intervals obtained by 100 fold bootstrap calculations of the actual prevalence in each risk bin. Figures with all coefficients can be found in the Supplementary Information.

3.3.1. Combination of predictor from biomarkers and the ASCVD Risk Estimator

Since the ASCVD Risk Estimator and the (risk score | biomarkers) predictor use different input and give complementary predictions, we combined them into a a very reliable risk predictor, superseding both the former. The risk estimates are compared with actual disease prevalence in Figure 12 for two versions of the combined predictor: *(1)* a linear regression on the biomarkers and all of the input going into the ASCVD Risk Estimator, and *(2)* a similar regression but also including the *output* of the ASCVD Risk Estimator. Their top coefficients are listed in the same figure.

4. Discussion

UK Biobank data include about 500k individuals, for each of whom the following are recorded: SNP genotype, biomarker (blood, urine) test results, and case status for most common disease conditions. We have explored the pattern of correlations between these three distinct data types using machine learning.

We have shown that SNPs can be used to predict quantitative values of biomarkers by training new polygenic scores (PGS) for biomarker prediction. We note that the day to day fluctuation of these biomarker levels suppresses the quality of prediction. A more stable phenotype (e.g., average value of biomarker measured on multiple occasions) would probably be even better predicted from SNPs alone.

As is typical for current genomic predictors, we find predictive power falls off significantly with genetic distance from the (European) training population. This highlights the importance of increasing ancestry diversity in genetic data collection. As genetic

predictors begin to find clinical applications, lack of diversity can exacerbate healthcare inequalities[48,80] (a larger list of associated ethical issues is highlighted in [44]).

We showed that biomarkers can be used as input to predict common disease risk. Some of these (risk score | biomarkers) predictors (e.g., ASCVD, diabetes) are very strong and may even surpass risk predictors in widespread clinical use. The combined predictor trained using both biomarkers and ASCVD Risk Estimator inputs clearly outperforms the latter in our comparison, at least for individuals at very high risk. It should be emphasized here that we did not perform the careful evidence review nor the statistical analysis that underlie the ASCVD Risk Estimator [23] and our comparison did not take into account the time of diagnosis. As such, our ASCVD predictor presented here is merely a comparative example and is not intended for clinical use in its current form. Yet, this naive approach performs remarkably well, utilizing the large statistical power of the UKB.

In the case of liver and kidney disease, we are not aware of other quantitative risk predictors that can be evaluated from biomarkers alone. Our results suggest that further research in this direction is warranted.

We note that (risk score | biomarkers) prediction quality does not exhibit the pattern of fall-off with genetic distance as previously found with genomic predictors[3]. For example, CAD and ASCVD predictors work well in all major ancestry groups despite using a European training sample. Further investigation is needed.

We studied concatenated predictor functions, which map SNPs to biomarkers to risk. In general, there were significant declines in performance. The magnitudes of these declines were perhaps expected for correlation chains of generic, high dimensional, vectors with similar pairwise correlations. Of the (risk score | biomarkers | SNPs) predictors, only the type 2 diabetes predictor performs well: AUC of $\sim$ .63. This is in fact comparable to what we have achieved in the past by training SNP-based LASSO directly on type 2 diabetes status. Furthermore, the two different types of predictors (risk score | biomarkers | SNPs) and (risk score | SNPs) capture somewhat complementary information, as shown in Figure 8. The sum of the two types of risk scores reaches an AUC of $\sim$ .67. It is unclear why the use of biomarkers as an intermediate step adds additional information relative to training directly with SNPs as features and case status as the phenotype. We leave this as an interesting topic for future research.

Author Contributions: Authors, listed alphabetically according to last name, contributed in the following ways: conceptualization, L.L., S.H. and E.W.; methodology, L.L., T.G.R., S.H., E.W.; software, L.L., T.G.R., and E.W.; validation, L.L., T.G.R., E.W.; formal analysis, L.L., T.G.R., E.W.; investigation, L.L., T.G.R., E.W.; resources, S.H.; data curation, L.L., T.G.R., and E.W.; writing — original draft preparation, L.L., T.G.R., S.H., and E.W.; writing — review and editing, L.L., T.G.R., S.H., E.W.; visualization, T.G.R. and E.W.; supervision, S.H.; project administration, S.H.; funding acquisition, S.H.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The predictors described in the paper are available to other researchers upon request.

Acknowledgments: Computational resources provided by the Michigan State University High-Performance Computing Center. The authors thank Dr. Andrew Siskind (UC Irvine) and Dr. Pui-Yan Kwok (Academia Sinica UC San Francisco) for fruitful discussions and correspondence. The authors acknowledge acquisition of datasets via UK Biobank Main Application 15326.

Conflicts of Interest: Stephen Hsu is a founder, shareholder and serves on the Board of Directors of Genomic Prediction, Inc. (GP). Louis Lello is an employee and shareholder of GP. These roles had no impact in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results. EW and TR declare no competing interests.

[3] Previous GWAS and PGS studies generally see a fall off behavior, but there are occasional exceptions (e.g. [81]).

References

1. Wray, N.R.; Yang, J.; Goddard, M.E.; Visscher, P.M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS genetics* **2010**, *6*.
2. Veenstra, D.L.; Roth, J.A.; Garrison Jr, L.P.; Ramsey, S.D.; Burke, W. A formal risk-benefit framework for genomic tests: facilitating the appropriate translation of genomics into clinical practice. *Genetics in Medicine* **2010**, *12*, 686.
3. Amir, E.; Freedman, O.C.; Seruga, B.; Evans, D.G. Assessing women at high risk of breast cancer: a review of risk assessment models. *JNCI: Journal of the National Cancer Institute* **2010**, *102*, 680–691. Oxford University Press.
4. Euesden, J.; Lewis, C.M.; O'reilly, P.F. PRSice: polygenic risk score software. *Bioinformatics* **2014**, *31*, 1466–1468.
5. Abraham, G.; Tye-Din, J.A.; Bhalala, O.G.; Kowalczyk, A.; Zobel, J.; Inouye, M. Accurate and Robust Genomic Prediction of Celiac Disease Using Statistical Learning. *PLOS Genetics* **2014**, *10*, 1–15. doi:10.1371/journal.pgen.1004137.
6. Priest, J.R.; Ashley, E.A. Genomics in clinical practice, 2014.
7. Jacob, H.J.; Abrams, K.; Bick, D.P.; Brodie, K.; Dimmock, D.P.; Farrell, M.; Geurts, J.; Harris, J.; Helbling, D.; Joers, B.J.; others. Genomics in clinical practice: lessons from the front lines. *Science translational medicine* **2013**, *5*, 194cm5–194cm5.
8. Shieh, Y.; Shieh, Y.; Hu, D.; Ma, L.; Huntsman, S.; Gard, C.C.; Leung, J.W.T.; Tice, J.A.; Vachon, C.M.; Cummings, S.R.; Kerlikowske, K.; Ziv, E. Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast Cancer Research and Treatment* **2016**, *159*, 513–525.
9. Bowdin, S.; Gilbert, A.; Bedoukian, E.; Carew, C.; Adam, M.P.; Belmont, J.; Bernhardt, B.; Biesecker, L.; Bjornsson, H.T.; Blitzer, M.; others. Recommendations for the integration of genomics into clinical practice. *Genetics in Medicine* **2016**, *18*, 1075.
10. Chatterjee, N.; Shi, J.; García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* **2016**, *17*, 392.
11. Lello, L.; Raben, T.G.; Yong, S.Y.; Tellier, L.C.; Hsu, S.D.H. Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate cancer. *Sci Rep* **2019**, *9*, 1–16. [PMC6814833].
12. Khera, A.V.; Chaffin, M.; Aragam, K.G.; Haas, M.E.; Roselli, C.; Choi, S.H.; Natarajan, P.; Lander, E.S.; Lubitz, S.A.; Ellinor, P.T.; Kathiresan, S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* **2018**, *50*, 1219.
13. Liu, L.; Kiryluk, K. Genome-wide polygenic risk predictors for kidney disease. *Nature Reviews Nephrology* **2018**, *14*, 723–724.
14. Torkamani, A.; Wineinger, N.E.; Topol, E.J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **2018**, *19*, 581.
15. Khera, A.V.; Chaffin, M.; Wade, K.H.; Zahid, S.; Brancale, J.; Xia, R.; Distefano, M.; Senol-Cosar, O.; Haas, M.E.; Bick, A.; Aragam, K.G.; Lander, E.S.; Smith, G.D.; Mason-Suares, H.; Fornage, M.; Lebo, M.; Timpson, N.J.; Kaplan, L.M.; Kathiresan, S. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell* **2019**, *177*, 587–596. [PMC6661115].
16. Nelson, H.D.; Pappas, M.; Cantor, A.; Haney, E.; Holmes, R. Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer in women: updated evidence report and systematic review for the US Preventive Services Task Force. *Jama* **2019**, *322*, 666–685.
17. Meisner, A.; Kundu, P.; Zhang, Y.D.; Lan, L.V.; Kim, S.; Ghandwani, D.; Pal Choudhury, P.; Berndt, S.I.; Freedman, N.D.; Garcia-Closas, M.; Chatterjee, N. Combined Utility of 25 Disease and Risk Factor Polygenic Risk Scores for Stratifying Risk of All-Cause Mortality. *American Journal of Human Genetics* **2020**, *107*, 418–431. doi:10.1016/j.ajhg.2020.07.002.
18. Lewis, C.M.; Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome medicine* **2020**, *12*, 1–11.
19. Lewis, A.C.; Green, R.C. Polygenic risk scores: from research tools to clinical instruments. *Genome medicine* **2021**, *13*, 14.
20. Kulm, S.; Marderstein, A.; Mezey, J.; Elemento, O. A systematic framework for assessing the clinical impact of polygenic risk scores. *medRxiv* **2021**, pp. 2020–04.
21. Wray, N.R.; Lin, T.; Austin, J.; McGrath, J.J.; Hickie, I.B.; Murray, G.K.; Visscher, P.M. From basic science to clinical application of polygenic risk scores: a primer. *JAMA psychiatry* **2021**, *78*, 101–109.
22. Bycroft, C.; Freeman, C.; Petkova, D. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*, 203–209.
23. Lloyd-Jones, D.M.; Braun, L.T.; Ndumele, C.E.; Smith Jr, S.C.; Sperling, L.S.; Virani, S.S.; Blumenthal, R.S. Use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic cardiovascular disease: a special report from the American Heart Association and American College of Cardiology. *Circulation* **2019**, *139*, e1162–e1177.
24. ASCVD Risk Estimator Plus. Available online: http://tools.acc.org/ASCVD-Risk-Estimator-Plus/#!/calculate/estimate/ (accessed on 29-03-2021).
25. UK Biobank.
26. Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L.T.; Sharp, K.; Motyer, A.; Vukcevic, D.; Delaneau, O.; O'Connell, J.; Cortes, A.; Welsh, S.; McVean, G.; Leslie, S.; Donnelly, P.; Marchini, J. Genome-wide genetic data on 500,000 UK Biobank participants. *bioRxiv* **2017**, [https://www.biorxiv.org/content/early/2017/07/20/166298.full.pdf]. doi:10.1101/166298.
27. Price, A.L.; Patterson, N.J.; Plenge, R.M.; Weinblatt, M.E.; Shadick, N.A.; Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **2006**, *38*, 904–909.
28. Novembre, J.; Johnson, T.; Bryc, K.; Kutalik, Z.; Boyko, A.R.; Auton, A.; Indap, A.; King, K.S.; Bergmann, S.; Nelson, M.R.; others. Genes mirror geography within Europe. *Nature* **2008**, *456*, 98–101.
29. Mathieson, I.; McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nature genetics* **2012**, *44*, 243–246.

30. Bhatia, G.; Gusev, A.; Loh, P.R.; Finucane, H.; Vilhjálmsson, B.J.; Ripke, S.; Purcell, S.; Stahl, E.; Daly, M.; de Candia, T.R.; others. Subtle stratification confounds estimates of heritability from rare variants. *BioRxiv* **2016**, p. 048181.
31. Dandine-Roulland, C.; Bellenguez, C.; Debette, S.; Amouyel, P.; Génin, E.; Perdry, H. Accuracy of heritability estimations in presence of hidden population stratification. *Scientific reports* **2016**, *6*, 1–10.
32. Guo, J.; Wu, Y.; Zhu, Z.; Zheng, Z.; Trzaskowski, M.; Zeng, J.; Robinson, M.R.; Visscher, P.M.; Yang, J. Global genetic differentiation of complex traits shaped by natural selection in humans. *Nature communications* **2018**, *9*, 1–9.
33. Rosenberg, N.A.; Edge, M.D.; Pritchard, J.K.; Feldman, M.W. Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. *Evolution, medicine, and public health* **2019**, *2019*, 26–34.
34. Sohail, M.; Maier, R.M.; Ganna, A.; Bloemendal, A.; Martin, A.R.; Turchin, M.C.; Chiang, C.W.; Hirschhorn, J.; Daly, M.J.; Patterson, N.; Neale, B.; Mathieson, I.; Reich, D.; Sunyaev, S.R. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife* **2019**, *8*, e39702.
35. Barton, N.; Hermisson, J.; Nordborg, M. Population genetics: Why structure matters. *Elife* **2019**, *8*, e45380.
36. Berg, J.J.; Harpak, A.; Sinnott-Armstrong, N.; Joergensen, A.M.; Mostafavi, H.; Field, Y.; Boyle, E.A.; Zhang, X.; Racimo, F.; Pritchard, J.K.; others. Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* **2019**, *8*, e39725.
37. Wray, N.R.; Kemper, K.E.; Hayes, B.J.; Goddard, M.E.; Visscher, P.M. Complex trait prediction from genome data: contrasting EBV in livestock to PRS in humans: genomic prediction. *Genetics* **2019**, *211*, 1131–1141.
38. Bitarello, B.D.; Mathieson, I. Polygenic scores for height in admixed populations. *G3: Genes, Genomes, Genetics* **2020**, *10*, 4027–4036.
39. Trochet, H.; Hussin, J. Fine-scale population structure confounds genetic risk scores in the ascertainment population. *bioRxiv* **2020**.
40. Refoyo-Martínez, A.; Liu, S.; Jørgensen, A.M.; Jin, X.; Albrechtsen, A.; Martin, A.R.; Racimo, F. How robust are cross-population signatures of polygenic adaptation in humans? *BioRxiv* **2021**, pp. 2020–07.
41. Lello, L.; Avery, S.G.; Tellier, L.; Vazquez, A.I.; de los Campos, G.; Hsu, S.D. Accurate genomic prediction of human height. *Genetics* **2018**, *210*, 477–497. [PMC6216598].
42. Yong, S.Y.; Raben, T.G.; Lello, L.; Hsu, S.D. Genetic Architecture of Complex Traits and Disease Risk Predictors. *Scientific Reports* **2020**, *10*. [PMC7374622].
43. Lello, L.; Raben, T.G.; Hsu, S.D.H. Sibling validation of polygenic risk scores and complex trait prediction. *Scientific Reports* **2020**, *10*, 13190. [PMC7411027], doi:10.1038/s41598-020-69927-7.
44. Raben, T.G.; Lello, L.; Widen, E.; Hsu, S.D.H. From Genotype to Phenotype: polygenic prediction of complex human traits, 2021, [arXiv:q-bio.GN/2101.05870].
45. Privé, F.; Aschard, H.; Blum, M. Efficient Implementation of Penalized Regression for Genetic Risk Prediction. *Genetics* **2019**, *212*, 65–74.
46. Wand, H.; Lambert, S.A.; Tamburro, C.; Iacocca, M.A.; O'Sullivan, J.W.; Sillari, C.; Kullo, I.J.; Rowley, R.; Dron, J.S.; Brockman, D.; Venner, E.; McCarthy, M.I.; Antoniou, A.C.; Easton, D.F.; Hegele, R.A.; Khera, A.V.; Chatterjee, N.; Kooperberg, C.; Edwards, K.; Vlessis, K.; Kinnear, K.; Danesh, J.N.; Parkinson, H.; Ramos, E.M.; Roberts, M.C.; Ormond, K.E.; Khoury, M.J.; Janssens, A.C.J.; Goddard, K.A.; Kraft, P.; MacArthur, J.A.; Inouye, M.; Wojcik, G.L. Improving reporting standards for polygenic scores in risk prediction studies, 2020. doi:10.1101/2020.04.23.20077099.
47. Carlson, C.S.; Matise, T.C.; North, K.E.; Haiman, C.A.; Fesinmeyer, M.D.; Buyske, S.; Schumacher, F.R.; Peters, U.; Franceschini, N.; Ritchie, M.D.; Duggan, D.J.; Spencer, K.L.; Dumitrescu, L.; Eaton, C.B.; Thomas, F.; Young, A.; Carty, C.; Heiss, G.; Le Marchand, L.; Crawford, D.C.; Hindorff, L.A.; Kooperberg, C.L.; for the PAGE Consortium. Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLOS Biology* **2013**, *11*, 1–11. doi:10.1371/journal.pbio.1001661.
48. Martin, A.R.; Gignoux, C.R.; Walters, R.K.; Wojcik, G.L.; Neale, B.M.; Gravel, S.; Daly, M.J.; Bustamante, C.D.; Kenny, E.E. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics* **2017**, *100*, 635–649.
49. Huang, H.; Ruan, Y.; Feng, Y.C.A.; Chen, C.Y.; Lam, M.; Sawa, A.; Martin, A.; Qin, S.; Ge, T. Improving Polygenic Prediction in Ancestrally Diverse Populations.
50. Privé, F.; Aschard, H.; Carmi, S.; Folkersen, L.; Hoggart, C.; O'Reilly, P.F.; Vilhjálmsson, B.J. High-resolution portability of 245 polygenic scores when derived and applied in the same cohort. *medRxiv* **2021**.
51. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **1996**, *58*, 267–288.
52. Donoho, D.L.; Tanner, J. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proceedings of the National Academy of Sciences* **2005**, *102*, 9446–9451. doi:10.1073/pnas.0502269102.
53. Donoho, D.; Stodden, V. Breakdown Point of Model Selection When the Number of Variables Exceeds the Number of Observations. The 2006 IEEE International Joint Conference on Neural Network Proceedings. IEEE, 2006. doi:10.1109/ijcnn.2006.246934.
54. Donoho, D.L.; Maleki, A.; Montanari, A. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences* **2009**, *106*, 18914–18919. doi:10.1073/pnas.0909892106.
55. Donoho, D.L.; Tanner, J. Precise Undersampling Theorems. *Proceedings of the IEEE* **2010**, *98*, 913–924. doi:10.1109/jproc.2010.2045630.

56. Wertz, J.; Moffitt, T.E.; Agnew-Blais, J.; Arseneault, L.; Belsky, D.W.; Corcoran, D.L.; Houts, R.; Matthews, T.; Prinz, J.A.; Richmond-Rakerd, L.S.; others. Using DNA from mothers and children to study parental investment in children's educational attainment. *Child development* **2019**, *00*, 1–17.
57. Kong, A.; Thorleifsson, G.; Frigge, M.L.; Vilhjalmsson, B.J.; Young, A.I.; Thorgeirsson, T.E.; Benonisdottir, S.; Oddsson, A.; Halldorsson, B.V.; Masson, G.; others. The nature of nurture: Effects of parental genotypes. *Science* **2018**, *359*, 424–428.
58. Bates, T.C.; Maher, B.S.; Medland, S.E.; McAloney, K.; Wright, M.J.; Hansell, N.K.; Kendler, K.S.; Martin, N.G.; Gillespie, N.A. The nature of nurture: Using a virtual-parent design to test parenting effects on children's educational attainment in genotyped families. *Twin Research and Human Genetics* **2018**, *21*, 73–83.
59. Belsky, D.W.; Domingue, B.W.; Wedow, R.; Arseneault, L.; Boardman, J.D.; Caspi, A.; Conley, D.; Fletcher, J.M.; Freese, J.; Herd, P.; others. Genetic analysis of social-class mobility in five longitudinal studies. *Proceedings of the National Academy of Sciences* **2018**, *115*, E7275–E7284.
60. Trejo, S.; Domingue, B.W. Genetic nature or genetic nurture? Introducing social genetic parameters to quantify bias in polygenic score analyses. *Biodemography and Social Biology* **2018**, *64*, 187–215.
61. Boerwinkle, E.; Leffert, C.C.; Lin, J.; Lackner, C.; Chiesa, G.; Hobbs, H.H.; others. Apolipoprotein (a) gene accounts for greater than 90% of the variation in plasma lipoprotein (a) concentrations. *The Journal of clinical investigation* **1992**, *90*, 52–60.
62. Kraft, H.; Köchl, S.; Menzel, H.; Sandholzer, C.; Utermann, G. The apolipoprotein (a) gene: a transcribed hypervariable locus controlling plasma lipoprotein (a) concentration. *Human genetics* **1992**, *90*, 220–230.
63. Austin, M.; Sandholzer, C.; Selby, J.; Newman, B.; Krauss, R.; Utermann, G. Lipoprotein (a) in women twins: heritability and relationship to apolipoprotein (a) phenotypes. *American journal of human genetics* **1992**, *51*, 829.
64. Rao, F.; Schork, A.J.; Maihofer, A.X.; Nievergelt, C.M.; Marcovina, S.M.; Miller, E.R.; Witztum, J.L.; O'Connor, D.T.; Tsimikas, S. Heritability of biomarkers of oxidized lipoproteins: twin pair study. *Arteriosclerosis, thrombosis, and vascular biology* **2015**, *35*, 1704–1711.
65. Frank, S.L.; Klisak, I.; Sparkes, R.S.; Mohandas, T.; Tomlinson, J.E.; McLean, J.W.; Lawn, R.M.; Lusis, A.J. The apolipoprotein (a) gene resides on human chromosome 6q26–27, in close proximity to the homologous gene for plasminogen. *Human genetics* **1988**, *79*, 352–356.
66. Drayna, D.T.; Hegele, R.A.; Hass, P.E.; Emi, M.; Wu, L.L.; Eaton, D.L.; Lawn, R.M.; Williams, R.R.; White, R.L.; Lalouel, J.M. Genetic linkage between lipoprotein (a) phenotype and a DNA polymorphism in the plasminogen gene. *Genomics* **1988**, *3*, 230–236.
67. Lindahl, G.; Gersdorf, E.; Menzel, H.J.; Duba, C.; Cleve, H.; Humphries, S.; Utermann, G. The gene for the Lp (a)-specific glycoprotein is closely linked to the gene for plasminogen on chromosome 6. *Human genetics* **1989**, *81*, 149–152.
68. Clarke, R.; Peden, J.F.; Hopewell, J.C.; Kyriakou, T.; Goel, A.; Heath, S.C.; Parish, S.; Barlera, S.; Franzosi, M.G.; Rust, S.; others. Genetic variants associated with Lp (a) lipoprotein level and coronary disease. *New England Journal of Medicine* **2009**, *361*, 2518–2528.
69. Tsimikas, S.; Hall, J.L. Lipoprotein (a) as a potential causal genetic risk factor of cardiovascular disease: a rationale for increased efforts to understand its pathophysiology and develop targeted therapies. *Journal of the American College of Cardiology* **2012**, *60*, 716–721.
70. Nikpay, M.; Goel, A.; Won, H.H.; Hall, L.M.; Willenborg, C.; Kanoni, S.; Saleheen, D.; Kyriakou, T.; Nelson, C.P.; Hopewell, J.C.; others. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature genetics* **2015**, *47*, 1121.
71. Kettunen, J.; Demirkan, A.; Würtz, P.; Draisma, H.H.; Haller, T.; Rawal, R.; Vaarhorst, A.; Kangas, A.J.; Lyytikäinen, L.P.; Pirinen, M.; others. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature communications* **2016**, *7*, 1–9.
72. Mack, S.; Coassin, S.; Rueedi, R.; Yousri, N.A.; Seppälä, I.; Gieger, C.; Schönherr, S.; Forer, L.; Erhart, G.; Marques-Vidal, P.; others. A genome-wide association meta-analysis on lipoprotein (a) concentrations adjusted for apolipoprotein (a) isoforms. *Journal of lipid research* **2017**, *58*, 1834–1844.
73. Schmidt, K.; Kraft, H.G.; Parson, W.; Utermann, G. Genetics of the Lp (a)/apo (a) system in an autochthonous Black African population from the Gabon. *European journal of human genetics* **2006**, *14*, 190–201.
74. Hoekstra, M.; Chen, H.Y.; Rong, J.; Dufresne, L.; Yao, J.; Guo, X.; Tsai, M.Y.; Tsimikas, S.; Post, W.S.; Vasan, R.S.; others. Genome-wide association study highlights APOH as a novel locus for lipoprotein (a) levels—brief report. *Arteriosclerosis, Thrombosis, and Vascular Biology* **2021**, *41*, 458–464.
75. Schmidt, K.; Noureen, A.; Kronenberg, F.; Utermann, G. Structure, function, and genetics of lipoprotein (a). *Journal of lipid research* **2016**, *57*, 1339–1359.
76. Buniello, A.; MacArthur, J.A.L.; Cerezo, M.; Harris, L.W.; Hayhurst, J.; Malangone, C.; McMahon, A.; Morales, J.; Mountjoy, E.; Sollis, E.; Suveges, D.; Vrousgou, O.; Whetzel, P.L.; Amode, R.; Guillen, J.A.; Riat, H.S.; Trevanion, S.J.; Hall, P.; Junkins, H.; Flicek, P.; Burdett, T.; Hindorff, L.A.; Cunningham, F.; Parkinson, H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* **2019**, *47*, D1005–D1012.
77. Coltell, O.; Asensio, E.M.; Sorlí, J.V.; Barragán, R.; Fernández-Carrión, R.; Portolés, O.; Ortega-Azorín, C.; Martínez-Lacruz, R.; González, J.I.; Zanón-Moreno, V.; Gimenez-Alba, I.; Fitó, M.; Ros, E.; Ordovas, J.M.; Corella, D. Genome-wide association study

(GWAS) on bilirubin concentrations in subjects with metabolic syndrome: Sex-specific gwas analysis and gene-diet interactions in a mediterranean population. *Nutrients* **2019**, *11*. doi:10.3390/nu11010090.

78. Bielinski, S.J.; Chai, H.S.; Pathak, J.; Talwalkar, J.A.; Limburg, P.J.; Gullerud, R.E.; Sicotte, H.; Klee, E.W.; Ross, J.L.; Kocher, J.P.A.; Kullo, I.J.; Heit, J.A.; Petersen, G.M.; De Andrade, M.; Chute, C.G. Mayo genome consortia: A genotype-phenotype resource for genome-wide association studies with an application to the analysis of circulating bilirubin levels. *Mayo Clinic Proceedings* **2011**, *86*. doi:10.4065/mcp.2011.0178.
79. Kathiresan, S.; Manning, A.K.; Demissie, S.; D'Agostino, R.B.; Surti, A.; Guiducci, C.; Gianniny, L.; Burtt, N.P.; Melander, O.; Orho-Melander, M.; Arnett, D.K.; Peloso, G.M.; Ordovas, J.M.; Cupples, L.A. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Medical Genetics* **2007**, *8*. doi:10.1186/1471-2350-8-S1-S17.
80. Martin, A.R.; Kanai, M.; Kamatani, Y.; Okada, Y.; Neale, B.M.; Daly, M.J. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics* **2019**, *51*, 584. PMC6563838.
81. Loos, R.J.; Yeo, G.S. The bigger picture of FTO—the first GWAS-identified obesity gene. *Nature Reviews Endocrinology* **2014**, *10*, 51–61.
82. Vattikuti, S.; Lee, J.J.; Chang, C.C.; Hsu, S.D.; Chow, C.C. Applying compressed sensing to genome-wide association studies. *GigaScience* **2014**, *3*, 2047–217X.
83. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, 2009.
84. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
85. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **2015**, *4*, s13742–015. PMC4342193.
86. Horta, D. Pandas-Plink. Available online: https://pypi.org/project/pandas-plink/ (accessed on 29-03-2021).
87. Kadie, C.M. PySNPTools. Available online: https://pypi.org/project/pysnptools/ (accessed on 30-03-2021).