Within-Family Validation of Polygenic Risk Scores and Complex Trait Prediction

Louis Lello ^{*1,2}, Timothy G. Raben ^{†1}, and Stephen D.H. Hsu ^{‡1,2}

¹Department of Physics and Astronomy, Michigan State University

²Genomic Prediction, Inc., North Brunswick, NJ

March 4, 2020

Abstract

We test a variety of polygenic predictors using tens of thousands of genetic siblings for whom we have SNP genotypes, health status, and phenotype information in late adulthood. Siblings have typically experienced similar environments during childhood, and exhibit negligible population stratification relative to each other. Therefore, the ability to predict differences in disease risk or complex trait values between siblings is a strong test of genomic prediction in humans. We compare validation results obtained using non-sibling subjects to those obtained among siblings and find that typically *most of the predictive power persists in within-family designs*. In the case of disease risk we test the extent to which higher polygenic risk score (PRS) identifies the affected sibling, and also compute Relative Risk Reduction as a function of risk score threshold. For quantitative traits we examine between-sibling differences in trait values as a function of predicted differences, and compare to performance in non-sibling pairs. Example results: Given 1 sibling with normal-range PRS score (<84 percentile) and 1 sibling with high PRS score (top few percentiles), the predictors identify the affected sibling about 70-90% of the time across a variety of disease conditions, including Breast Cancer, Heart Attack, Diabetes, etc. For height, the predictor correctly identifies the taller sibling roughly 80 percent of the time when the (male) height difference is 2 inches or more.

*Corresponding Author; lollou@msu.edu

[†]rabentim@msu.edu

[‡]hsu@msu.edu

1 Introduction

The ability to predict complex human phenotypes, including common disease risks, from DNA alone, is an important advance in genomics and biological science[1, 2]. Sibling comparisons are a powerful method with which to validate genomic prediction in humans. Siblings (i.e., children who share the same mother and father) have typically experienced similar environments while growing up – family social status, exposure to toxins, diet, climate, etc. all tend to be similar [3, 4]. Furthermore, siblings are concordant for ancestry and display negligible differences in population structure.

If a girl grows up to be taller than her sister, with whom she spent the first 18 years of her life, it seems likely at least some of the height difference is due to genetic differences. How much of phenotype difference can we predict from DNA alone? If one of the sisters develops breast cancer later in life, how much of the risk was due to genetic variants that she does not share with her asymptomatic sister? These are fundamental questions in human biology, which we address (at least to some extent) in this paper.

We study two types of predictors: Polygenic Risk Scores (PRS), which estimate the risk of developing a specific common disease condition, and Polygenic Scores (PGS) which predict a quantitative trait such as adult height or bone density. For both types we compare predictive power when applied to non-sibling individuals with power when applied to sibling pairs. (We will sometimes abbreviate sibling by sib for brevity.)

Predictors trained on a large population of non-sibling individuals (see Methods section below) could potentially utilize correlations in the SNP data that arise from environment effects, but are not related to direct genetic causation. Two examples are given below.

1. If environmental conditions in a specific region, such as, e.g., Northern England, affect disease risk, the predictor trained on UK data might assign nonzero effect sizes to SNPs associated with ancestries found in that region – i.e., the predictor learns to use population structure correlated to environmental conditions. These specific SNPs are correlated to disease risk for environmental reasons, but might not have any connection to genetic mechanisms related to the disease. They likely have little power to differentiate between siblings, who experienced similar family conditions and have identical ancestry.
2. It is also possible that some SNP variants affect nurture (the way that parents raise their children). These SNPs could affect the child phenotype via an environmental mechanism under parental control, not a biochemical pathway within the child. This is sometimes referred to as a *genetic nurture* effect [5–9]. Note, siblings raised together would both be affected by parental genetic nurture variants, so these effects are weakened in family designs.

Sibling comparisons reduce the impact of factors such as those described above. We expect some reduction in power when predictors trained in a population of non-sibling individuals are tested among sibs. Sibling validation likely yields a better estimate of truly causal genetic

effects. A more complicated measure of familial relatedness might lead to even better results [10], but we restrict our analyses here to siblings.

For almost all of the predictors studied here, both PRS and PGS, significant power remains even in the sibling tests. Almost all of the predictors we study seem to capture some real genetic effects that cause siblings to differ from each other as adults.

One exception is the Educational Attainment (EA) predictor, which exhibits a very strong reduction in power when applied to sibs. This is not entirely unexpected as effects like the violation of the equal-environment hypothesis should be found for EA, but the effect should be modest [4], and EA can depend on complicated correlations between environment and genes[11]. Interestingly, the corresponding reduction for the Fluid Intelligence predictor is much less than for EA. This is discussed in more detail below.

The results discussed below suggest that almost all of the predictors studied capture some real, direct, genetic effects. These effects survive within-family validity testing, and attenuation of predictive power tends to be modest (most of the power remains in the sibling tests). Our results for height, body mass index (BMI), EA, and Fluid Intelligence are similar recent results found in [12], utilizing data from the Twins Early Development Study (TEDS). As far as we know, this paper is the first to analyze a variety of disease risks using within-family designs.

As shown in earlier work [13, 14], we expect the predictors to improve substantially as more data become available for training. This is conditioned on genotyping that captures a sufficient part of the predictive regions of the genome [15]. It seems clear that with the possible exception of the height phenotype (for which we start to see diminishing returns; most of the common SNP heritability is captured already in the predictor), training is limited by sample size (specifically for risk predictors: number of genotyped cases) and not by algorithm performance or computational resources.

2 Methods and Data

The main dataset we use for training is the 2018 release of the UK Biobank [16, 17]. In previous work, predictors were trained exclusively on genetically British individuals (as identified by principal component analysis [18]), however it has been shown that predictors trained on populations filtered by self-reported ethnicity perform equivalently [14]. In this work, all predictors are trained on the set of self-reported white individuals excluding all individuals who belong to a sibling pair. In each training run, a small fraction of non-sibs is withheld from the training set for validation and model selection, and the set of sibling pairs is used as a final test set.

We construct linear models of genetic predisposition for a variety of disease conditions that

were presented in [14] and linear models of several quantitative human phenotypes, some of which can be found in [13]. The disease condition phenotype data describes a binary case-control status which is defined either by self-report or from a clinical diagnosis.

Polygenic predictors are constructed using compressed sensing [19–22]. It has been demonstrated that SNP matrices of human genome matrices are good compressed sensors: L1 performance guarantee theorems hold and phase transition behavior is observed.

Our primary motivation in this work is the investigation of polygenic predictor performance on sibling pairs. We focus specifically on L1-trained predictors because we understand their training and performance characteristics well. There are many other methods used in the creation of polygenic scores. While we make no claims concerning those other predictors, we suspect that they would perform similarly in within-family validation tests such as those performed here. We do also examine two predictors (for Breast Cancer and Coronary Artery Disease) which were published in Khera et al. [23]. These are indicated as such in the figures and one can compare with L1-trained predictors on similar phenotypes.

For each disease condition, we compute a set of additive effects $\vec{\beta}$ (each component is the effect size for a specific SNP) which minimizes the LASSO objective function:

$$\mathcal{O}(\lambda, \vec{\beta}) = \frac{1}{2} \|\vec{y} - X\vec{\beta}\|^2 + n\lambda \|\vec{\beta}\|_1 \quad (2.1)$$

where n is the number of samples, $\|\dots\|$ means L2 norm (square root of sum of squares), $\|\dots\|_1$ is the L1 norm (sum of absolute values) and the term $\lambda \|\vec{\beta}\|_1$ is a penalization which enforces sparsity of $\vec{\beta}$. The value of the phenotype variable y for case or control status is simply 1 or 0 (respectively). For quantitative phenotypes, y values are z-scored using population means and standard deviations.

The optimization is performed over a set of 50k SNPs which are selected by rank ordering the p-values obtained from single marker regression of the phenotype against the SNPs. The details of this are described in Appendix F.

Predictors are trained using the implementation of the LASSO algorithm from the Scikit-learn Python package [24]. Specifically, the `lassopath` algorithm is called on standardized inputs as it generates the full lasso path. For disease status, we typically use five non-overlapping sets of cases and controls held back from the training set for the purposes of in-sample cross-validation. For each value of λ , there is a particular predictor which is then applied to the cross-validation set, where the polygenic score is defined as (i labels the individual and j labels the SNP)

$$PGS_i \text{ or } PRS_i = \sum_{j=1}^p X_{ij}\beta_j \quad . \quad (2.2)$$

To select a specific value of the penalization λ which defines our final predictor (for final evaluation on out-of-sample testing sets), we choose the λ that maximizes the performance metric in each cross validation set thereby creating 5 different predictors. For case-control phenotypes, the performance metric is AUC, and for quantitative phenotypes, it is the correlation between predicted and actual trait value. This is explained in more detail in Appendices C,F.

The training computations were performed using the super-computing cluster in the Michigan State University High Performance Computing Center.

3 Sibling Differences in Case/Control Phenotypes

The goal of this work is to study the effectiveness of polygenic predictors using siblings in the UK Biobank. We expect the predictor performance within families to be diminished to some extent compared to the general population because of shared environments, shared genetics, genetic nurture, and other confounding factors. For all traits (case/control and quantitative), predictors are trained, validated and tested on individuals who self-report as some form of "white ancestry" - i.e., British, Irish, or other white (note this terminology is from UK Biobank data tables). From this group, all individuals for whom there is at least one sibling match are set aside for use in the sibling test set. This is described in appendix C.

Training is conducted using the remaining individuals, for whom there is no sibling in UK Biobank. For each trait, 1000 randomly selected (non-sibling) individuals are set aside (not used in the training), but are used for cross-validation and model selection. For case-control phenotypes, there are 500 cases and 500 controls making up the 1000. (For Breast, Prostate, and Testicular Cancer the corresponding numbers are 100 and 100, due to smaller datasets.) This process is repeated 5 times to generate a set of 5 predictors so that statistical fluctuations associated with the training process (mean and variance) can be estimated. We do not report the performance metrics on the validation sets as they are quantitatively similar to that of the final test set - see [14] for an example of this.

For all traits, we make use of L1 penalized regression as described in [13, 14]. Previous work has shown this to be an effective method of generating polygenic predictors [13, 14]. The typical outputs of a LASSO run are the regularization parameters and a vector of SNP weights - this is discussed at length in refs [13, 14, 23] where we use the scikit-learn package instead of a custom implementation [24]. We include results from publicly available predictors for Breast Cancer and Coronary Artery Disease from Khera et al. [23] - scoring from these predictors is described in Appendix B.

The first quantity which is calculated for a predictor is an overall performance metric: for case/control phenotypes this corresponds to AUC; for quantitative phenotypes we focus on

the correlation coefficient between predicted and actual phenotypes. The test set is composed of all individuals who are within a sibling pair in the UKB - the performance metric on this test set matches previous results from the literature [13, 14] and sets the baseline of comparison.

3.1 Sibling Call Rates - Case | Control

A first test of polygenic scores in the affected sibling context can be made by simply computing the frequency at which the higher PRS sibling corresponds to the affected individual. We restrict the test set to all sibling pairs with one affected sibling and one unaffected sibling - i.e., we exclude sibling pairs where both are cases or controls. Within this set, we compute the fraction of the time in which the sibling with higher PRS is the case. The results are given in table 1. As a baseline comparison, we also compute the fraction called correctly using an equal number of *non-sibling* case-control pairs randomly drawn from the total sibling set.

In Appendix D we perform a similar analysis for trios of siblings.

3.2 Case identification for high risk sibling

Here we consider sibling pairs with one affected (case) and one control. Further, we focus on the subset of pairs in which one sibling has a high PRS score and the other a PRS score in the normal range (i.e., less than +1 SD above average). In other words, exactly one of the sibs is a "high risk" outlier, and we wish to know how often it is the outlier that is a case.

The previous analysis focused on the identification of the case in a sibling pair by selecting the larger polygenic score even if the difference was very small. Because the absolute risk rises nonlinearly in the tails of the PRS distribution (i.e., for outliers), we do not expect strong prediction results when comparing two individuals in the normal PRS range (i.e., drawn from the middle of the population distribution). In this analysis, summarized in table 2, one sibling is labeled high risk and the other sibling is normal risk as defined by PRS. In all cases, normal risk is defined as in being in the 84th percentile or below ($< +1$ SD in PRS), while we vary the threshold used to define high risk ($> +1.5$ SD, $+2.0$ SD, $+2.5$ SD, etc.).

As we restrict to sibling pairs with a larger risk differential, the predictions of which sibling is the case become more accurate (albeit still noisy). In other words: given that one of two siblings is affected, when one sibling is normal risk in PRS but the other sibling is in the top few percentile of risk - the larger PRS sibling will be increasingly likely to be the affected sibling as the difference in PRS becomes larger.

Condition	N pairs (single case)	Sibling Case Higher PRS Fraction	Random Case Higher PRS Fraction
Asthma	3,948	0.618 (0.004)	0.633 (0.005)
Atrial Fibrillation	332	0.620 (0.033)	0.636 (0.013)
Basal Cell Carcinoma	431	0.599 (0.014)	0.616 (0.009)
Breast Cancer (LASSO-L1)	583	0.585 (0.020)	0.586 (0.014)
Breast Cancer (Khera)	583	0.557 (–)	0.601 (–)
Coronary Artery Disease (LASSO-L1)	1,072	0.556 (0.012)	0.579 (0.017)
Coronary Artery Disease (Khera)	1,073	0.596 (–)	0.614 (–)
Gallstones	700	0.592 (0.006)	0.622 (0.014)
Glaucoma	440	0.593 (0.013)	0.602 (0.015)
Gout	631	0.627 (0.007)	0.661 (0.003)
Heart Attack	900	0.593 (0.012)	0.603 (0.006)
High Cholesterol	4,291	0.596 (0.005)	0.632 (0.002)
Hypothyroidism	2,031	0.658 (0.003)	0.699 (0.005)
Hypertension	6,931	0.627 (0.002)	0.645 (0.001)
Malignant Melanoma	360	0.547 (0.040)	0.592 (0.013)
Prostate Cancer	106	0.642 (0.015)	0.650 (0.034)
Testicular Cancer	24	0.575 (0.090)	0.607 (0.133)
Type 1 Diabetes	290	0.646 (0.019)	0.669 (0.006)
Type 2 Diabetes	1,594	0.595 (0.005)	0.620 (0.001)

Table 1: Polygenic predictors tested on sibling pairs. The first column is the number of sibling pairs with one affected and one unaffected sibling. The second column is the average and standard deviation (over 5 predictors) of the fraction in which the case has higher PRS. Third column gives results for non-sibling (random) pairs. Quantities in red have uncertainties in the central value larger than 10% due to low statistics.

Condition	N pairs	1 sib 93rd	N	1 sib 97th	N	1 sib 99th
Asthma	402	0.758 (0.021)	110	0.782 (0.039)	17	0.882 (0.078)
Atrial Fibrillation	37	0.757 (0.071)	20	0.750 (0.097)	9	1.0 (–)
Basal Cell Carcinoma	54	0.648 (0.065)	23	0.826 (0.079)	6	0.667 (0.192)
Breast Cancer (LASSO-L1)	45	0.662 (0.071)	11	0.545 (0.150)	2	0.5 (0.353)
Breast Cancer (Khera)	52	0.596 (0.068)	12	0.583 (0.142)	3	0.667 (0.272)
Coronary Artery Disease (LASSO-L1)	131	0.613 (0.043)	48	0.424 (0.071)	10	0.5 (0.158)
Coronary Artery Disease (Khera)	109	0.706 (0.044)	38	0.816 (0.063)	5	0.800 (0.179)
Gallstones	212	0.720 (0.031)	158	0.697 (0.037)	85	0.686 (0.050)
Glaucoma	30	0.720 (0.082)	9	0.667 (0.157)	1	–
Gout	70	0.743 (0.052)	37	0.784 (0.068)	16	0.875 (0.083)
Heart Attack	68	0.685 (0.056)	16	0.688 (0.116)	4	–
High Cholesterol	441	0.660 (0.023)	130	0.662 (0.042)	28	0.786 (0.078)
Hypothyroidism	282	0.780 (0.025)	109	0.890 (0.030)	32	0.906 (0.052)
Hypertension	757	0.726 (0.016)	229	0.777 (0.027)	53	0.811 (0.054)
Malignant Melanoma	30	0.600 (0.089)	10	0.600 (0.155)	2	1.0 –
Prostate Cancer	8	0.875 (0.117)	0	–	0	–
Testicular Cancer	0	–	0	–	0	–
Type 1 Diabetes	41	0.805 (0.062)	28	0.893 (0.058)	17	0.824 (0.092)
Type 2 Diabetes	137	0.772 (0.036)	37	0.816 (0.064)	8	0.75 (0.153)

Table 2: Predictors tested on sibling pairs with a single case, where one sibling is high risk (+1.5, +2, +2.5 SD or 93rd, 97.5th, 99th percentile) and the other is normal risk (< +1 SD or < 85th percentile). The first column is the number of pairs used. The second column is the fraction of pairs where the high risk sibling is the case. 1 SD binomial errors given in parenthesis. Quantities in red have uncertainties in the central value larger than 10% due to low statistics.

We repeat this calculation for a set of pairs in which no individual is paired with his or her sibling. This is done using the sibling population by randomizing the pairings. We generate random pairs of non-sibling individuals with exactly one case per pair. Further, we consider the subset of pairs in which one member of the pair is normal risk ($\text{PRS} < +1 \text{ SD}$), while the other is high risk. We then compute the probability that the high risk individual is the affected individual. Results are given in table 3.

Condition	N pairs	1 sib 93rd	N	1 sib 97th	N	1 sib 99th
Asthma	777	0.749 (0.016)	289	0.794 (0.026)	72	0.889 (0.050)
Atrial Fibrillation	76	0.734 (0.057)	44	0.723 (0.079)	24	0.783 (0.108)
Basal Cell Carcinoma	65	0.708 (0.067)	34	0.706 (0.098)	9	0.889 (0.211)
Breast Cancer (LASSO-L1)	132	0.638 (0.050)	50	0.720 (0.077)	10	0.800 (0.205)
Breast Cancer (Khera)	143	0.678 (0.044)	60	0.683 (0.071)	23	0.696 (0.124)
Coronary Artery Disease (LASSO-L1)	117	0.634 (0.054)	40	0.640 (0.092)	7	0.429 (0.247)
Coronary Artery Disease (Khera)	187	0.711 (0.037)	78	0.719 (0.060)	22	0.773 (0.124)
Gallstones	210	0.695 (0.035)	149	0.718 (0.041)	65	0.723 (0.066)
Glaucoma	42	0.757 (0.079)	16	0.788 (0.125)	3	1.0 (0.457)
Gout	115	0.852 (0.041)	69	0.870 (0.054)	40	0.850 (0.078)
Heart Attack	121	0.645 (0.048)	45	0.622 (0.085)	14	0.571 (0.166)
High Cholesterol	881	0.712 (0.016)	340	0.738 (0.026)	108	0.769 (0.048)
Hypothyroidism	505	0.844 (0.018)	240	0.879 (0.025)	79	0.911 (0.044)
Hypertension	1883	0.755 (0.010)	727	0.812 (0.016)	222	0.820 (0.029)
Malignant Melanoma	64	0.597 (0.078)	29	0.655 (0.109)	17	0.706 (0.149)
Prostate Cancer	37	0.784 (0.089)	12	0.750 (0.184)	9	0.778 (0.221)
Testicular Cancer	—	—	—	—	—	—
Type 1 Diabetes	86	0.849 (0.049)	56	0.911 (0.056)	35	0.914 (0.076)
Type 2 Diabetes	230	0.764 (0.030)	75	0.853 (0.052)	23	0.870 (0.110)

Table 3: Predictors tested on non-sibling (random) pairs w/ a single case where one is high risk (+1.5, +2, +2.5 SD above or 93rd, 97.5th, 99th percentile) and the other is normal risk ($< +1$ Standard Deviation or < 85 th percentile). The first column is the number of pairs. The second column is the fraction of pairs where the high risk individual is the case. 1 SD binomial errors given in parenthesis. Quantities in red have uncertainties in the central value larger than 10% due to low statistics.

Comparing tables 2 and 3 suggests higher prediction accuracy for non-sibling pairs of individuals. The difference in accuracy is slightly inflated by the fact that the normal risk individuals in the related (sib) pairs tend to cluster closer to the +1 SD PRS upper limit than those in the non-sibling pairs. This is because, conditional on having a high-risk sibling,

the distribution of PRS scores is shifted to larger than average values. Nevertheless, we see that the success fractions are not very different between the two tables, and almost always overlap within one standard deviation uncertainty.

These results suggest that polygenic prediction works almost as well between siblings as in unrelated individuals.

In figure 1, we repeat the analysis from the tables using a continuously varying threshold (in z-score) to define the high risk set of individuals. As the threshold z-score increases the fraction of cases called correctly also increases. We display the results for Affected Sibling Pairs (ASP) as well as non-sibling pairs of individuals where each pairing consists of one case and one control. There is some reduction in accuracy for sibling pairs versus non-sibling pairs, as expected.

The error estimates in the figures and tables are generated as follows. We display the larger of two contributions to the uncertainty in determining the fraction called correctly (vertical axis): One results from the standard deviation among the 5 predictors we generate for each trait. The other results from sampling error (i.e., having only a finite number of pairs in which to estimate the fraction called correctly). The second source of error is a Clopper–Pearson interval with a 68% confidence value.

Figure 1 is given specifically as an example - similar plots are generated for all conditions. These are shown in appendix G.

3.3 Population prevalence: Relative Risk Reduction

Polygenic scores can be used to identify subsets of the population who are at high or low risk for a given condition. This information can be used to better allocate resources for, e.g., screening or prevention. In [14], it is proposed that polygenic scores could lead to more effective detection and intervention for a wide variety of health conditions (e.g., breast cancer). The early detection of disease conditions could then lead to a net cost reduction and better health outcomes. Throughout this section we are specifically focused on *genetic* risks, but these results could be incorporated into more complete risk models as found in [25].

In this section, we investigate how population prevalence varies as we exclude high and low risk individuals from the group. The population prevalence can be interpreted as the probability that a randomly selected individual will develop the condition, conditional on either having 1. PRS below some upper limit (left panel in figures – a low risk population defined by PRS) or 2. PRS above some lower limit (right panel in figures – a high risk population defined by PRS).

The upper panels in figures 2-5 display the results for randomly selected individuals from

Figure 1: Predictors tested on random (non-sibling) pairs and affected sibling pairs with a single case. One individual is high risk (with z-score given on the horizontal axis) and the other is normal risk ($PRS < +1$ SD). The error estimates are explained in the text.

the general population. The orange line in both panels represents the disease prevalence in the entire testing set (general population).

Figure 2: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Type 2 Diabetes.

These plots are meant to be illustrative. Similar plots are shown for each of the disease conditions we study in appendix H.

We demonstrate the utility of PRS in the context of a known family history by repeating the previous calculation on a restricted Affected Sibling Pair (ASP) testing set. In the lower panels of figures 2-5 we compute the same disease prevalence as in the upper panels, but *for individuals with an affected sibling*. That is, all cases and all controls used in the calculation have an affected sibling; the existence of this affected sibling defines the population analyzed

Figure 3: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Breast Cancer.

Figure 4: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Hypertension.

Figure 5: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Heart Attack.

as one with higher than normal risk. The values in the lower panels of figures 2-5 therefore reflect an overall higher risk due to the family history of the individuals. However, the results show that low PRS individuals have reduced risk *relative to others with a similar family history*. Given two individuals A and B, where A has an affected sibling A' and B has an affected sibling B', the graphs show that between A and B, the one with higher PRS has a higher probability of having the condition. The green line in both panels represents the disease prevalence in the entire testing set – the population of individuals with an affected sibling.

For some of the disease conditions with small rate of incidence, we did not have enough data to directly estimate risk as a function of PRS for sibs in the ASP population – i.e., there are not enough sib pairs in which *both* are cases. However, we typically did have enough data to estimate mean and standard deviation in PRS for affected and unaffected individuals conditional on each individual having an affected sib. (Less data is required to estimate a mean and SD than to map out an entire curve bin by bin.) Assuming that the distribution of cases and controls are both approximately Gaussian in PRS (something we verified to be true for conditions for which we have more data), this allows us to compute the implied risk as a function of PRS. We include this predicted risk as a function of PRS (see green curves) on all prevalence plots involving the ASP populations. The results are shown in the corresponding figures in Appendix H.

Figure 2 is meant to be illustrative and similar plots for all conditions are given in appendix H. We include the predicted prevalence as a function of score - the predicted prevalence is calculated assuming that cases and controls are normally distributed (a mixed Gaussian distribution). The means, standard deviations and total numbers of cases and controls are the only (six) parameters needed for the predicted curve - these are calculated directly from the data. See [14] for a more in depth discussion.

3.4 Identification within Affected Sibling Pairs (ASP)

To assess the degree to which discriminatory power is altered within affected families for case/control phenotypes, we calculate the AUC amongst the full testing set (i.e., a proxy for the general population; for convenience we used all individuals with a sibling) and amongst a cohort of affected sibling pairs (ASP; all cases or controls must have a sibling who is also a case). The ASP cohort is constructed by restricting the testing set to all sibling pairs as follows: controls have at least one sibling which is a case; cases must also have at least one other sibling which is a case - i.e., in this new test set, all cases and controls have at least one affected sibling. The difference in the AUC between the entire population and the affected sibling testing sets are given in table 4.

Table 4 shows that, as expected, prediction accuracy is typically higher in groups of non-sibling individuals. However, the decrease in AUC when working with high risk families (i.e., where at least one sib is affected) is typically modest.

Condition	All Siblings		ASPs	
	N cases / ctrls	AUC - All	N cases / ctrls	AUC - ASP
Asthma	4,519 / 35,511	0.630 (0.002)	944 / 3,877	0.628 (0.003)
Atrial Fibrillation	327 / 39,703	0.624 (0.004)	16 / 330	0.577 (0.019)
Basal Cell Carcinoma	415 / 39,615	0.626 (0.007)	16 / 428	0.528 (0.024)
Breast Cancer (LASSO)	963 / 22,204	0.585 (0.016)	52 / 583	0.567 (0.015)
Breast Cancer (Khera)	963 / 22,242	0.608 (–)	52 / 583	0.573 (–)
Coronary Artery Disease (LASSO)	1,058 / 38,972	0.582 (0.017)	70 / 1,069	0.570 (0.019)
Coronary Artery Disease (Khera)	1,059 / 39,049	0.621 (–)	70 / 1,070	0.617 (–)
Gallstones	690 / 39,340	0.638 (0.003)	40 / 699	0.586 (0.015)
Glaucoma	422 / 39,608	0.592 (0.012)	26 / 439	0.602 (0.030)
Gout	601 / 39,429	0.660 (0.004)	29 / 631	0.653 (0.010)
Heart Attack	889 / 39,141	0.602 (0.006)	60 / 898	0.618 (0.025)
High Cholesterol	5,240 / 34,790	0.632 (0.002)	1,351 / 4,203	0.622 (0.002)
Hypertension	10,524 / 29,506	0.648 (0.001)	4,296 / 6,719	0.635 (0.001)
Hypothyroidism	2,152 / 37,878	0.709 (0.002)	319 / 1,997	0.685 (0.007)
Malignant Melanoma	334 / 39,696	0.585 (0.007)	2 / 359	–
Prostate Cancer	262 / 16,601	0.644 (0.014)	20 / 106	0.654 (0.030)
Testicular Cancer	57 / 16,806	0.631 (0.012)	0 / 24	–
Type 1 Diabetes	277 / 39,753	0.676 (0.003)	12 / 290	0.643 (0.018)
Type 2 Diabetes	1,692 / 38,338	0.617 (0.005)	235 / 1,576	0.599 (0.014)

Table 4: Polygenic predictors tested on sibling pairs. The first column gives the number of cases/controls and the AUC for the entire sibling cohort (proxy for general population). The second column gives the number of cases/controls and the AUC for subset of cohort in which all pairs have at least one affected sibling (ASP). Quantities in parentheses are standard deviations amongst 5 predictors.

4 Sibling Differences in Quantitative Traits

4.1 Performance difference: siblings vs non-sibling pairs

We now turn to prediction of quantitative phenotypes. To evaluate performance one typically computes the correlation between predicted and actual phenotypes: $\rho(PGS, y)$ where PGS and y are the predicted phenotype from polygenic score and the measured phenotype respectively.

In comparing within-family performance to performance in the general (non-sibling) population, it is useful to consider pairwise differences in actual phenotype and predicted phenotype (polygenic score): Δy and ΔPGS . For example Δy could be the (z-scored) difference in height between the two in the pair, and ΔPGS the (z-scored) difference in predicted heights (or PGS score).

We compute the correlation between phenotype and score difference, $\rho(\Delta PGS, \Delta y)$, for pairs of siblings and for pairs of non-sibling individuals. The results are given in table 5. Figure 6 provides a specific example - results are shown for all traits considered in appendix I.

Trait	N pairs	$\rho(\Delta PGS, \Delta y)$ siblings	$\rho(\Delta PGS, \Delta y)$ non-sibling
BMI	21,556	0.271 (0.003)	0.345 (0.005)
Educational Attainment	21,352	0.089 (0.001)	0.256 (0.003)
Body Fat Percentage	20,990	0.245 (0.002)	0.319 (0.001)
Fluid Intelligence	4,968	0.165 (0.004)	0.264 (0.005)
Heel Bone Density	10,133	0.345 (0.001)	0.415 (0.002)
Standing Height	21,418	0.545 (0.001)	0.614 (0.001)
Platelet Count	20,534	0.393 (0.002)	0.490 (0.002)

Table 5: Polygenic predictors tested on sibling pairs and non-sibling (random) pairs. First column is number of pairs, second and third are correlation between difference in predicted phenotype and actual phenotype for sibs and non-sibling pairs.

Educational Attainment (EA) shows an especially large within-family attenuation in performance relative to the other predictors. This has been noticed in other studies [26]. The results suggest that at least some of the observed power in polygenic prediction of EA among non-sibling individuals comes from effects such as subtle population stratification (perhaps correlated to environmental conditions or family socio-economic status)[8], genetic nurture [9], or other environmental-genetic correlations [5–7]. Interestingly, the decrease in power seems to be not as large for the phenotype Fluid Intelligence (measured in UKB using a brief 12 item cognitive test).

Figure 6: Difference in phenotype (vertical axis) and difference in polygenic score (horizontal axis) for pairs of individuals. Red dots are sibling pairs and blue dots are random (non-sibling) pairs.

4.2 Rank order accuracy: siblings vs non-sibling pairs

We can further compare within-family effectiveness of quantitative trait predictors by estimating the probability of predicting rank order – e.g., which sib is taller – using PGS.

First, how often does the higher PGS sibling have the larger value of the actual phenotype? We restrict the analysis to only those pairs of siblings whose phenotypes are known and then compute the fraction of the time in which rank order by PGS agrees with rank order in phenotype. The results are listed in table 6.

Similar results for trios of siblings are presented in Appendix D.

Trait	N pairs	Fraction called (Sibling pairs)	Fraction called (Random pairs)
BMI	21,556	0.588 (0.001)	0.614 (0.003)
Educational Attainment	21,352	0.528 (0.001)	0.591 (0.001)
Body Fat Percentage	20,990	0.583 (0.002)	0.606 (0.001)
Fluid Intelligence	4,968	0.558 (0.003)	0.592 (0.006)
Heel Bone Density	10,133	0.627 (0.002)	0.657 (0.003)
Standing Height	21,418	0.684 (0.001)	0.718 (0.001)
Platelet Count	20,534	0.646 (0.001)	0.679 (0.001)

Table 6: Rank ordering by polygenic score. The first column gives the number of sibling pairs, the second and third columns give the fraction called correctly (higher PGS individual has greater phenotype value) in sibling / non-sibling pairs. Quantities in parenthesis are standard deviations.

4.3 Rank order accuracy as a function of phenotype difference

In the previous calculation many of the failures to correctly predict rank order result from the two individuals in the pair having very close values of the phenotype. To further investigate, we consider accuracy of rank order prediction as a function of actual phenotype difference in the pair. As expected, probability of correct rank ordering increases with actual difference in phenotype.

PGS from sets of 5 trained predictors are z-scored based on the testing population. The identification of pairs with phenotypic difference larger than x (value shown on horizontal axis of figure 7) is based upon the average score value across the five predictors. This selects the sub-cohort with large phenotypic difference. Then the fraction called correct is calculated for each of the 5 polygenic scores. This fraction, for 0.5, 1, and 1.5 standard deviation difference in phenotype, can be found in Table 7. The quoted error is computed as the larger of the standard deviation resulting from the 5 different predictors, and the statistical sampling error (Clopper-Pearson interval) in estimating the probability p in a binomial distribution. (See earlier discussion in Sec. 3.2) To clarify: the first error contribution is intrinsic to the construction of the predictor (different training runs create slightly different predictors), the second error contribution always arises when estimating the (success) probability p from a finite sample of N datum.

Trait	N pairs	$\Delta y > 0.5$	N	$\Delta y > 1.0$	N	$\Delta y > 1.5$
Body Mass Index	13,376	0.623 (0.004)	7,387	0.658 (0.006)	3,815	0.694 (0.007)
Educational Attainment	11,532	0.545 (0.005)	8,020	0.548 (0.006)	5,402	0.554 (0.007)
Body Fat Percentage	13,836	0.613 (0.004)	8,075	0.642 (0.005)	4,234	0.670 (0.007)
Fluid Intelligence	3,288	0.570 (0.009)	1,875	0.585 (0.011)	952	0.605 (0.016)
Heel Bone Density	6,365	0.674 (0.006)	3,435	0.716 (0.008)	1,724	0.734 (0.011)
Standing Height	12,689	0.762 (0.004)	6,184	0.835 (0.005)	2,488	0.890 (0.006)
Platelet Count	13,039	0.702 (0.004)	7,253	0.745 (0.005)	3,591	0.769 (0.007)

Table 7: Predictors tested on sibling pairs where a phenotype difference is larger than some value (+0.5, +1.0, +1.5 Standard Deviations difference; the adjusted phenotypes are described in the appendix). The first column is the number of pairs and the second column is the fraction of pairs where higher PGS corresponds to greater phenotype value.

We repeat this calculation for non-related individuals, by simply randomizing the pairings so that individuals are no longer paired with their siblings. We then perform the same operations: select pairs where the phenotype difference is larger than a certain value and then compute the fraction of pairs where the high PGS individual has a larger value. This is illustrated in table 8.

The comparison between non-sibling pairs and sibling pairs is shown in Fig. 7, where we display the fraction identified correctly for sibling pairs and for randomly paired individuals,

Trait	N pairs	$\Delta y > 0.5$	N	$\Delta y > 1.0$	N	$\Delta y > 1.5$
Body Mass Index	14,816	0.652 (0.004)	9,185	0.686 (0.005)	5,399	0.723 (0.006)
Educational Attainment	13,973	0.618 (0.004)	10,521	0.630 (0.005)	7,866	0.646 (0.005)
Body Fat Percentage	14,892	0.637 (0.004)	9,771	0.668 (0.005)	5,794	0.702 (0.006)
Fluid Intelligence	3,495	0.611 (0.008)	2,337	0.625 (0.010)	1,394	0.648 (0.013)
Heel Bone Density	6,985	0.707 (0.005)	4,410	0.758 (0.006)	2,505	0.790 (0.008)
Standing Height	15,365	0.777 (0.003)	9,994	0.837 (0.004)	5,828	0.884 (0.004)
Platelet Count	14,443	0.724 (0.004)	9,131	0.774 (0.004)	5,251	0.805 (0.005)

Table 8: Predictors tested on non-sibling pairs where a phenotype difference is larger than some value (+0.5, +1.0, +1.5 Standard Deviations difference; the adjusted phenotypes are described in the appendix). The first column is the number of pairs and the second column is the fraction of pairs where higher PGS corresponds to greater phenotype value.

allowing the threshold phenotype difference to vary continuously. The difference between the blue and orange lines represents the difference between predictive power amongst non-sibling and related individuals.

Figure 7: Probability of PGS correctly identifying the individual with larger phenotype value (vertical axis). Horizontal axis shows absolute difference in phenotypes. The blue line is for sibling pairs, the orange line is for randomized (non-sibling) pairs.

Figure 7 is given specifically as an example - similar plots are generated for all continuous traits which are discussed in this paper. These are shown in appendix J. The loss of power in polygenic predictors is expected, but these calculations illustrate the central point that polygenic predictors can still reliably improve the identification of individuals (or rank ordering) when large phenotypic differentials exist.

5 Discussion

Siblings have typically experienced similar environments during childhood, and exhibit negligible population stratification relative to each other. The ability to predict differences in disease risk or complex trait values between siblings provides an important validation of polygenic predictors. We compared validation results obtained using non-sibling subjects to those obtained among siblings, and found that most of the predictive power persists in within-family designs.

In the case of disease risk we tested the extent to which higher polygenic risk score (PRS) identifies the affected sibling, and also estimated Relative Risk Reduction as a function of risk score threshold. For quantitative traits we studied between-sibling differences in trait values as a function of predicted differences, and compared to performance in non-sibling pairs.

Our focus was not primarily on the absolute level of prediction, but rather on the comparison between results in non-sibling pairs versus sibling pairs. For most predictors the observed reduction in power tends to be modest. The largest decline in power is observed for the quantitative trait Educational Attainment (but see Fluid Intelligence in contrast).

We emphasize that predictors trained on even larger datasets will likely have significantly stronger performance than the ones analyzed here [13, 14]. As we elaborated in earlier work, where many of these predictors were first investigated, their main practical utility at the moment is in the identification of outliers who may be at exceptionally high (or low) risk for a specific disease condition. The results here confirm that high risk score outliers are indeed at elevated risk, even compared to their (normal range score) siblings.

The sibling results presented in this paper, together with the many out of sample validations of polygenic scores that continue to appear in the literature, suggest that genomic prediction in humans is a robust and important advance that will lead to improvements in translational medicine as well as deep insights into human genetics.

Acknowledgments LL, TR, and SH acknowledge support from the Office of the Senior Vice-President for Research and Innovation at Michigan State University. Computational resources provided by the MSU High-Performance Computing Center. The authors acknowledge acquisition of datasets via UK Biobank Main Application 15326.

Competing Interests Stephen Hsu a shareholder of Genomic Prediction, Inc. (GP), and serves on its Board of Directors. Louis Lello is an employee and shareholder of GP. Tim Raben has no commercial interests relevant to the research.

A Genotype Quality Control

The main dataset used in this work is the 2018 release of the UK Biobank (the 2018 version corrected some issues with imputation, included sex chromosomes, etc). In 2018, the UK Biobank (UKB) re-released the dataset representing approximately 500,000 individuals genotyped on two Affymetrix platforms - approximately 50,000 samples on the UKB BiLEVE Axiom array and the remainder on the UKB Biobank Axiom array. The genotype information was collected for 488,377 individuals for 805,426 SNPs which were then subsequently imputed to a much larger number of SNPs. In all predictor training, we restricted our analysis to self-report White individuals (as defined using self-reported ethnic background as surveyed by UK Biobank) [18] - namely participants with the code 1, 1001, 1002 or 1003 in the Ethnic Background field.

In our analyses, we do not use the imputed SNP set for training, but rather build all predictors from the directly genotyped markers. Our previous work has indicated that polygenic scores constructed from imputed data perform slightly less well. The imputed SNP set is used only for testing predictors generated by other authors (Khera et al. [23]). Further, we restrict our analysis to the autosomal chromosomes on every trait as we have found that inclusion of the sex chromosomes does not provide a significant performance enhancement for the traits studied here. From the called SNPs, further quality control was performed using PLINK version 1.9 [27]. After combining individual chromosomes to a single BED file, SNPs and samples which had missing call rates exceeding 3% were excluded and SNPs with minor allele frequency below 0.1% were also removed so to avoid rare variants. This resulted in 663,533 SNPs and 487,048 people - filtering this set by self-reported ancestry results in 459,063 remaining individuals.

B External Predictors

We use Breast Cancer and Coronary Artery Disease predictors that were published in [23]. To apply these predictors, we use the imputed genomes directly from the UK Biobank without any additional QC. We again restrict to the self-report white sibling cohort for all testing. For each individual, we scan through each chromosome and generate a score using the SNP weight column from the predictor file for every SNP which is present in the imputed chromosome. Each individual chromosome score is combined and then the final result is z-scored based on the controls in the cohort (so that the score is centered near zero with unit variance). For Coronary Artery Disease, this set gave 40,108 individuals and for Breast Cancer, after restricting to only females, gave 23,205 individuals.

C Sibling Identification, Training and Testing Sets

The UK Biobank performed an initial relatedness analysis on all participants, but familial relationships among UK Biobank participants were not directly recorded. The analysis was done by the UK Biobank, calculating kinship coefficients and IBS0 using the KING software [18]; the details can be found in the original UK Biobank paper [18]. The UK Biobank provides a total of 107,162 related pairs of individuals with kinship coefficients and IBS0.

To identify sibling pairs, we implement the same filters which the UK Biobank used: parent-offspring and sibling relationships have an expected relatedness coefficient of 0.25 and parent-offspring relationship are distinguished by those with $IBS0 < 0.0012$. We select all pairs with $IBS0 > 0.0012$ and kinship coefficient > 0.176 , yielding a set of 22,667 pairs of participants - this agrees with table S3 of the UK Biobank quality control supplementary information [18]. An in-depth analysis of the relationships can be found in the UK Biobank supplementary material. After restricting this group to individuals who survived the genotype quality control and also self-reported white ancestry, 40,030 individuals remained in this set and formed 21,671 pairs of siblings.

These sibling pairs are assumed to be first-degree siblings - each pair is assumed to have the same mother and father. However, amongst this set there exist situations where a participant is listed as sibling with two others, but the two are not listed as a sibling with each other. For example, if 3 individuals are labeled A, B, C then the following situation exists: A/B form a sibling pair, A/C form a sibling pair, but the pair B/C is not in the list of pairs. In this example, we do NOT include the B/C pair in our calculations with sibling pairs - we maintain the pairing which matches with the UK Biobank. However, when forming trios we group A,B,C together as a trio.

Amongst the initial 22,667 sibling pairs, we find 1,051 trios, 66 quartets, and 11 quintets. Due to the small number of groups larger than 3, we only use pairs and triplets in our analysis. After excluding individuals who do not self-report as white or pass quality control, 982 trios remain. The set of 40,030 individuals who are present in sibling pairs are used as a final testing set for all analyses and are excluded from the 459,063 genotyped individuals in all training. This results in training sets consisting of 419,033 individuals - from this set, smaller validation sets are chosen for model selection.

D Trio analysis

Case-control phenotypes

After inferring familial units, it was found that there are very few larger families in the UK Biobank - only trios are present in large enough quantity to warrant investigation. We performed risk-reduction analysis similar to that for the sibling pairs, in which we consider only trios with one affected sibling and then compute the fraction of the time in which the highest or lowest PGS sibling is a case. We note that the expected result under random selection would be 1/3. In all cases, the highest (lowest) PGS sibling is more (less) likely to be the affected sibling than random chance would produce. Table 9 summarizes these results.

Condition	N Trios (single case)	High PGS	Low PGS	Random Selection
Asthma	221	0.417 (0.011)	0.247 (0.010)	0.356 (0.043)
Atrial Fibrillation	18	0.344 (0.082)	0.211 (0.046)	0.289 (0.099)
Basal Cell Carcinoma	25	0.416 (0.046)	0.240 (0.0)	0.264 (0.061)
Breast Cancer	20	0.390 (0.147)	0.270 (0.084)	0.420 (0.084)
Coronary Artery Disease	62	0.326 (0.039)	0.277 (0.031)	0.316 (0.065)
Gallstones	45	0.409 (0.020)	0.276 (0.040)	0.396 (0.029)
Glaucoma	36	0.444 (0.052)	0.244 (0.066)	0.32 (0.050)
Gout	51	0.435 (0.049)	0.231 (0.038)	0.314 (0.069)
Heart Attack	59	0.434 (0.026)	0.207 (0.033)	0.305 (0.052)
High Cholesterol	224	0.458 (0.014)	0.246 (0.014)	0.354 (0.027)
Hypothyroidism	128	0.491 (0.017)	0.206 (0.010)	0.330 (0.035)
Hypertension	303	0.452 (0.010)	0.224 (0.008)	0.326 (0.024)
Malignant Melanoma	22	0.345 (0.041)	0.273 (0.045)	0.273 (0.045)
Prostate Cancer	2	—	—	—
Testicular Cancer	1	—	—	—
Type 1 Diabetes	21	0.419 (0.021)	0.210 (0.043)	0.390 (0.092)
Type 2 Diabetes	96	0.444 (0.035)	0.229 (0.016)	0.308 (0.047)

Table 9: Polygenic predictors tested on sibling trios. The first column is the number of sibling trios with one affected and two unaffected siblings. The next three columns are the probabilities (and standard deviation over 5 predictors) that the higher PRS / lowest PRS / randomly chosen sibling are the case.

Quantitative phenotypes

After identifying family structure in the UKB, we find that only trios are present in large enough quantity to warrant discussion. We performed a similar analysis to that for the

sibling pairs, in which we consider all trios with measured phenotypes and then compute the fraction of the time in which the highest or lowest PGS sibling has the highest / lowest phenotype value. We note that the expected result under random selection would be $1/3$. We also compute the fraction of trios for which the correct *order* is assigned; the expected result under random selection would be $1/6$ for this question. In all cases, the highest (lowest) PGS sibling is more likely to have the largest (smallest) phenotypic value. Also, the ordering of siblings is more likely to be correct than random chance would produce. Table 10 summarizes these results.

Sibling Trios Trait	N trios	Fraction High called	Fraction Low Called	Fraction Order Called
Body Mass Index	975	0.418 (0.004)	0.429 (0.004)	0.229 (0.002)
Educational Attainment	966	0.350 (0.009)	0.374 (0.006)	0.191 (0.007)
Body Fat Percentage	934	0.421 (0.005)	0.413 (0.006)	0.230 (0.009)
Fluid Intelligence	188	0.391 (0.008)	0.400 (0.015)	0.216 (0.010)
Heel Bone Density	440	0.495 (0.006)	0.464 (0.009)	0.248 (0.009)
Standing Height	971	0.556 (0.002)	0.538 (0.006)	0.345 (0.008)
Platelet Count	913	0.517 (0.006)	0.504 (0.005)	0.306 (0.004)
Random Trios Trait	N trios	Fraction High called	Fraction Low Called	Fraction Order Called
Body Mass Index	975	0.454 (0.010)	0.478 (0.004)	0.267 (0.007)
Educational Attainment	966	0.425 (0.013)	0.432 (0.006)	0.233 (0.019)
Body Fat Percentage	934	0.459 (0.010)	0.445 (0.005)	0.257 (0.007)
Fluid Intelligence	188	0.462 (0.022)	0.415 (0.016)	0.268 (0.012)
Heel Bone Density	440	0.517 (0.005)	0.515 (0.007)	0.315 (0.007)
Standing Height	971	0.590 (0.003)	0.596 (0.003)	0.382 (0.001)
Platelet Count	913	0.569 (0.003)	0.554 (0.008)	0.350 (0.006)

Table 10: Polygenic predictors tested on sibling trios and random trios. The first column gives the number of trios, columns 2-4 give the probabilities that the largest / smallest phenotype value, and order, are correctly identified using PGS. Quantities in parenthesis are standard deviations. Note that all quantities exceed the expected values of $1/3$ (0.33) for selection and $1/6$ (0.167) for rank ordering.

E Phenotype Quality Control

Case/Control Phenotypes

In this section, we describe how cases and controls are identified. We extract the following conditions and use as a case/control phenotype: Atrial Fibrillation, Asthma, Basal Cell Carcinoma, Breast Cancer, Coronary Artery Disease, Gallstones, Glaucoma, Gout, Heart Attack, High Cholesterol, Hypothyroidism, Hypertension, Malignant Melanoma, Prostate

Cancer, Testicular Cancer, Type 1 Diabetes and Type 2 Diabetes. All conditions were identified using the fields "Non cancer illness code (self-reported)", "Cancer code (self-reported)" and "Diagnoses primary ICD10" or "Diagnoses secondary ICD10". All individuals who are not labelled as a case for a specific condition are then labelled as a control.

We used the field "Non-Cancer Illness Code (self-reported)" to identify cases and controls for the following: Atrial Fibrillation, Asthma, Gallstones, Glaucoma, Gout, Heart Attack, High Cholesterol, Hypothyroidism, Hypertension. Specifically, cases were identified by selecting individuals with the codes in "Non cancer illness code (self-reported)": asthma:1111, atrial fibrillation:1471, gallstones:1162, glaucoma:1277, gout:1466, heart attack:1075, high cholesterol:1473, hypothyroidism/myxoedema:1226, hypertension:1065. We used codes in the the field "Cancer code (self-reported)" to identify cases of the following: breast cancer:1002, basal cell carcinoma:1061, malignant melanoma:1059, prostate cancer:1044, testicular cancer:1045.

To select Type 1 Diabetes cases, we identify individuals based on a doctor's diagnosis using the fields "Diagnoses primary ICD10" or "Diagnoses secondary ICD10". Specifically, any individual with ICD10 code E10.0-E10.9 (Insulin-dependent diabetes mellitus) in the Main Diagnosis or Secondary Diagnosis field is labelled as a case. To select Type 2 Diabetes cases in UKB, we identify individuals based on a doctor's diagnosis using the fields Diagnoses primary ICD10 or Diagnoses secondary ICD10. Specifically, any individual with ICD10 code E11.0-E11.9 (Non-insulin-dependent diabetes mellitus) in the Main Diagnosis or Secondary Diagnosis field is labelled as a case. To select for Coronary Artery Disease, we select based on ICD10 criterion identified in [23], specifically we select any individual with any of the following ICD10 codes (corresponding to various diagnoses of angina, myocardial infarctions or ischaemic heart disease): I20.0, I20.1, I20.8, I20.9, I21.0 - I21.4, I21.9, I21.X, I22.0, I22.1, I22.8, I22.9, I23.5, I23.6, I23.8, I24.0, I25.0 - I25.6, I25.8, I25.9.

After identifying cases and controls in the whole UKB population, we restricted our training set to self-reported white and our testing set to self-reported white but within a sibling pair. For sex-specific traits (breast cancer / prostate cancer), we restricted our analysis to individuals who are of the appropriate sex. From the training set, we then select a random 500 cases and 500 controls to use for validation and model selection - with the exception of breast, prostate and testicular cancers where we choose 100/100 for validation due to the smaller training set. The number of cases and controls identified in this manner for training / validation are listed in Table 11. The number of cases / controls for the final corresponding test sets are listed in the main body.

Quantitative phenotypes

In this section, we describe how quantitative traits are standardized before training. We focus on Standing Height, Body Mass Index, Body Fat percentage, Heel Bone Mineral Density, Platelet Count, Educational Attainment and Fluid Intelligence score. For Heel Bone Mineral Density, we used field 3148 (there are several other possible fields that could be combined

Condition	Cases (train)	Controls (train)	Cases (val)	Controls (val)
Asthma	48,875	369,158	500	500
Atrial Fibrillation	3,095	414,938	500	500
Basal Cell Carcinoma	3,795	414,238	500	500
Breast Cancer *	9,459	216,339	100	100
Coronary Artery Disease	11,264	406,769	500	500
Gallstones	6,769	411,264	500	500
Glaucoma	4,264	413,769	500	500
Gout	5,712	412,321	500	500
Heart Attack	9,455	408,578	500	500
High Cholesterol	53,603	364,430	500	500
Hypertension	110,893	307,140	500	500
Hypothyroidism	20,518	397,515	500	500
Malignant Melanoma	2,911	415,122	500	500
Prostate Cancer *	3,275	189,560	100	100
Testicular Cancer *	650	192,185	100	100
Type 1 Diabetes	2,345	415,688	500	500
Type 2 Diabetes	18,097	399,936	500	500

Table 11: Table of number of cases and controls in training and testing sets. Traits with (*) are trained and tested only on a single sex.

for this - i.e. 3148 / 4105 / 4124 / etc.).

For all traits, we calculate the mean and standard deviation for Genetically British males / females and then z-score the phenotypes appropriately. After z-scoring based on sex, we fit a trendline to everyone born between 1938 and 1968, this is then subtracted from the z-scored values. This is done to correct for any population level changes over time. Educational Attainment was converted into a quantitative trait using the ISCED scale, similar to that used by the SSGAC collaboration [28]. Specifically, the codes in field "Qualifications" were converted to to years of education via the following mapping: (1,2,3,4,5,6,-7,-3) -> (20,13,10,10,19,15,7,NA). All other quantitative traits were read directly from their corresponding fields without conversion. All quantitative phenotypes are adjusted in this manner and we use this sex / age adjusted phenotype in all calculations.

F Model Training Algorithm

In all calculations, we use the implementation of LASSO regression (Least Absolute Shrinkage and Selection Operator) found in Scikit Learn for Python 3 [24]. Given the large datasets, we use the lassopath algorithm which outputs the set of λ and $\vec{\beta}$ along the regularization pathway so we can choose λ on our own validation sets. For completeness, we outline the general optimization procedure executed to obtain the SNP weights.

Trait	Train Set	Validation Set	Testing Set	$\rho(y, PGS)_{test}$
Body Mass Index	416,630	1,000	39,927	0.342 (0.005)
Body Fat Percentage	410,413	1,000	39,395	0.327 (0.001)
Bone Mineral Density	238,703	1,000	23,980	0.427 (0.001)
Educational Attainment	414,241	1,000	39,735	0.251 (0.002)
Fluid Intelligence	156,351	1,000	14,189	0.259 (0.002)
Height	415,455	1,000	39,784	0.621 (0.001)
Platelet Count	406,451	1,000	38,951	0.487 (0.002)

Table 12: Table of number of samples used in training, validation and testing sets. Samples with missing values are excluded. The average correlation between phenotypes and PGS for the testing set is given.

First, we perform single marker regression using the PLINK software [27]. From this, the top 50k SNPs by p-value are selected and this subset of SNPs is used in the LASSO run. The BED matrix is loaded into memory using the Pandas-Plink package [29]. The data is standardized and lassopath is run with $nstep = 200$ and $eps = 0.04$ ($eps = 0.01$ for continuous traits).

Given a set of samples $i = 1, 2, \dots, n$ with a set of p SNPs, the phenotype y_i and state of the j^{th} SNP, X_{ij} , are observed. X_{ij} is an $n \times p$ matrix which contains the number of copies of the minor allele and any missing values are replaced with the SNP average. L_1 penalized regression, LASSO, seeks to minimize the objective function

$$\mathcal{O}_\lambda(\vec{\beta}) = \frac{1}{2} \|\vec{y} - X\vec{\beta}\|^2 + n\lambda \|\vec{\beta}\|_1 \quad (\text{F.1})$$

where $\|\vec{v}\|_1 = \sum_i^n |v_i|$ is the L_1 norm, $\|\vec{v}\| = \sum_i^n v_i^2$ is the L_2 norm and λ is a adjustable hyperparameter. The solution is given in terms of the soft-thresholding function as

$$\begin{aligned} S(z, \gamma) &= \text{sgn}(z) \max(|z| - \gamma, 0) \\ \beta_j^* &= \frac{1}{\sum_{i=1}^n X_{ij}^2} S \left(\sum_{i=1}^n \left[X_{ij} y_i - \sum_{k \neq j} X_{ij} X_{ik} \beta_k \right], n\lambda \right) \end{aligned} \quad (\text{F.2})$$

The penalty term affects which elements of $\vec{\beta}$ have non-zero entries. The value of λ is first chosen to be the maximum value such that all β_i are zero, and it is then decreased, allowing more nonzero components in the predictor. For each value of λ , $\vec{\beta}^*(\lambda_n)$ is obtained using the previous values of $\vec{\beta}^*(\lambda_{n-1})$ (warm start) and coordinate descent. The Donoho-Tanner phase transition [30] describes how much data is required to recover the true nonzero components of the linear model and suggests, e.g., for SNP heritability $h^2 \sim 0.5$ that we expect to recover the true signal with s SNPs when the number of samples is $n \sim 30s - 100s$ (see [31, 32]). For a more complete description of the algorithm, consult

the documentation available at https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.lasso_path.html.

G High vs Normal risk call rates.

Here we focus pairs in which one member is a case and the other is a control. We further restrict to the subset of pairs where one sib has a high PRS score and the other a PRS score in the normal range (i.e., less than +1 SD above average). In other words, exactly one of the pair is a high risk outlier, and we wish to know how often it is the outlier that is affected.

As we restrict to sibling pairs with a larger risk differential, the predictions for which sibling is the case become more accurate (albeit noisy). In other words: given that one of two siblings is affected, when one sibling is normal risk in PRS but the other sibling is in the top few percentile of risk - the larger PRS sibling will be increasingly likely to be the affected sibling as the difference in PRS becomes larger.

We repeat this calculation for non-related individuals. We generate random pairs of non-sibling individuals with exactly one case per pair. Further, we consider the subset of pairs in which one member of the pair is normal risk ($\text{PRS} < +1 \text{ SD}$), while the other is high risk. We then compute the probability that the high risk individual is the affected individual. Results are given in table 3.

The error estimates in the figures are generated as follows. We display the larger of two contributions to the uncertainty in determining the fraction called correctly (vertical axis): one results from the SD among the 5 predictors we generate for each trait, the other results from sampling error (i.e., having only a finite number of pairs in which to compute the fraction called correctly). This is known as the Clopper–Pearson interval; in the case that the probability p of calling the case correctly is not too close to 0 or 1, and N pairs used, the one standard deviation sampling uncertainty is given by roughly $\sqrt{p(1-p)/N}$.

Figure 8: Predictors tested on random (non-sibling) pairs and affected sibling pairs with a single case. One individual is high risk (with z-score given on the horizontal axis) and the other is normal risk ($\text{PRS} < +1 \text{ SD}$). The error estimates are explained in the text.

Figure 9: Predictors tested on random (non-sibling) pairs and affected sibling pairs with a single case. One individual is high risk (with z-score given on the horizontal axis) and the other is normal risk ($PRS < +1$ SD). The error estimates are explained in the text. (the label “Khera” distinguishes the LASSO generated predictor vs that from ref [23].)

Figure 10: Predictors tested on random (non-sibling) pairs and affected sibling pairs with a single case. One individual is high risk (with z-score given on the horizontal axis) and the other is normal risk ($PRS < +1$ SD). The error estimates are explained in the text.

Figure 11: Predictors tested on random (non-sibling) pairs and affected sibling pairs with a single case. One individual is high risk (with z-score given on the horizontal axis) and the other is normal risk ($PRS < +1$ SD). The error estimates are explained in the text.

Figure 12: Predictors tested on random (non-sibling) pairs and affected sibling pairs with a single case. One individual is high risk (with z-score given on the horizontal axis) and the other is normal risk (PRS < +1 SD). The error estimates are explained in the text.

H Population prevalence changes

In this appendix, we present the results on how population prevalence varies as we exclude high and low risk individuals from the group. The population prevalence can be interpreted as the probability that a randomly selected individual will develop the condition, conditional on either having 1. PRS below some upper limit (left panel in figures) or 2. PRS above some lower limit (right panel in figures).

We demonstrate the utility of PRS in the context of a known family history by repeating this calculation on a restricted ASP testing set. We compute the same disease prevalence but using only individuals with an affected sibling. In this test set, all cases and all controls have an affected sibling (see earlier discussion). The values therefore reflect an overall higher risk due to the family history of the individuals.

Here we illustrate for each disease studied, how the population prevalence changes by excluding individuals with PRS above / below a threshold given on the horizontal axis.

Figure 13: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Asthma.

Figure 14: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Atrial Fibrillation.

Figure 15: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Basal Cell Carcinoma.

Figure 16: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Breast Cancer.

Figure 17: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Breast Cancer with PRS generated by the predictor from [23].

Figure 18: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Coronary Artery Disease.

Figure 19: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Coronary Artery Disease with PRS generated from the predictor in [23].

Figure 20: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Type 1 Diabetes.

Figure 21: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Type 2 Diabetes.

Figure 22: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Gallstones.

Figure 23: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Glaucoma.

Figure 24: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Gout.

Figure 25: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Heart Attack.

Figure 26: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is High Cholesterol.

Figure 27: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Hypertension.

Figure 28: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Hypothyroidism.

Figure 29: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Malignant Melanoma.

Figure 30: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Prostate Cancer.

Figure 31: Exclusion of individuals above (left panel) and below (right panel) a z-score threshold (horizontal axis) with resulting group prevalence shown on the vertical axis. The left panel shows risk reduction in a low PRS population, the right panel shows risk enhancement in a high PRS population. Top figures are results in the general population, bottom figures are the Affected Sibling Pair (ASP) population (i.e., variation of risk with PRS among individuals with an affected sib). Phenotype is Testicular Cancer – there was not enough data for the lower panels.

I Sibling Pair vs Random Pair score-phenotype difference correlations

In this section we show the correlation between phenotype and score difference, $\rho(\Delta PGS, \Delta y)$, for pairs of siblings and for pairs of non-sibling individuals.

Figure 32: Difference in phenotype (vertical axis) and difference in polygenic score (horizontal axis) for pairs of individuals. Red dots are sibling pairs and blue dots are random (non-sibling) pairs.

Figure 33: Difference in phenotype (vertical axis) and difference in polygenic score (horizontal axis) for pairs of individuals. Red dots are sibling pairs and blue dots are random (non-sibling) pairs.

J Large phenotypic difference call rates in siblings vs non-sibling.

We consider accuracy of rank order prediction as a function of actual phenotype difference in sets of pairs. The identification of pairs with phenotypic difference larger than x (value shown on horizontal axis of figures) is based upon the average score value across the five predictors. This selects the sub-cohort with large phenotypic difference. Then the fraction called correct is done for each of the 5 polygenic scores. The quoted error is then computed as the larger of the standard deviation resulting from the 5 different predictors, and the statistical sampling error in estimating the probability p in a binomial distribution. This is done for sibling pairs and randomly paired individuals.

Figure 34: Probability of PGS correctly identifying the individual with larger phenotype value (vertical axis). Horizontal axis shows absolute difference in phenotypes. The blue line is for sibling pairs, the orange line is for randomized (non-sibling) pairs.

Figure 35: Probability of PGS correctly identifying the individual with larger phenotype value (vertical axis). Horizontal axis shows absolute difference in phenotypes. The blue line is for sibling pairs, the orange line is for randomized (non-sibling) pairs.

References

1. Polderman, T. J. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics* **47**, 702 (2015) (cit. on p. 2).
2. Boomsma, D., Busjahn, A. & Peltonen, L. Classical twin studies and beyond. *Nature reviews genetics* **3**, 872–882 (2002) (cit. on p. 2).
3. Jelenkovic, A. *et al.* Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts. *Scientific reports* **6**, 1–13 (2016) (cit. on p. 2).
4. Felson, J. What can we learn from twin studies? A comprehensive evaluation of the equal environments assumption. *Social science research* **43**, 184–199 (2014) (cit. on pp. 2, 3).
5. Wertz, J. *et al.* Using DNA from mothers and children to study parental investment in children’s educational attainment. *Child development* (2019) (cit. on pp. 2, 18).
6. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424–428 (2018) (cit. on pp. 2, 18).
7. Bates, T. C. *et al.* 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* **21**, 73–83 (2018) (cit. on pp. 2, 18).
8. Belsky, D. W. *et al.* Genetic analysis of social-class mobility in five longitudinal studies. *Proceedings of the National Academy of Sciences* **115**, E7275–E7284 (2018) (cit. on pp. 2, 18).
9. 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* **64**, 187–215 (2018) (cit. on pp. 2, 18).
10. Young, A. I. *et al.* Relatedness disequilibrium regression estimates heritability without environmental bias. *Nature Genetics* **50**, 1304–1310 (2018) (cit. on p. 3).
11. Daw, J., Guo, G. & Harris, K. M. Nurture net of nature: Re-evaluating the role of shared environments in academic achievement and verbal intelligence. *Social science research* **52**, 422–439 (2015) (cit. on p. 3).
12. Selzam, S. *et al.* Comparing within-and between-family polygenic score prediction. *The American Journal of Human Genetics* **105**, 351–363 (2019) (cit. on p. 3).
13. Lello, L. *et al.* Accurate genomic prediction of human height. *Genetics* **210**, 477–497 (2018) (cit. on pp. 3–6, 22).
14. 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* **9**, 2019 (2019) (cit. on pp. 3–6, 10, 16, 22).
15. Yong, S. Y., Raben, T. G., Lello, L. & Hsu, S. D. Genetic Architecture of Complex Traits and Disease Risk Predictors. *bioRxiv* (2020) (cit. on p. 3).
16. *UK Biobank* Accessed: 2017-07-21. <http://www.ukbiobank.ac.uk/> (cit. on p. 3).

17. Bycroft, C., Freeman, C. & Petkova, D. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (cit. on p. 3).
18. Bycroft, C. *et al.* Genome-wide genetic data on 500,000 UK Biobank participants. *bioRxiv*. eprint: <https://www.biorxiv.org/content/early/2017/07/20/166298.full.pdf>. <https://www.biorxiv.org/content/early/2017/07/20/166298> (2017) (cit. on pp. 3, 23, 24).
19. Vattikuti, S., Lee, J. J., Chang, C. C., Hsu, S. D. & Chow, C. C. Applying compressed sensing to genome-wide association studies. *GigaScience* **3**, 10 (2014) (cit. on p. 4).
20. Ho, C. M. & Hsu, S. D. Determination of nonlinear genetic architecture using compressed sensing. *GigaScience* **4**, 44 (2015) (cit. on p. 4).
21. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**, 76–82 (2011) (cit. on p. 4).
22. Vilhjálmsson, B. J. *et al.* Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics* **97**, 576–592 (2015) (cit. on p. 4).
23. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics* **50**, 1219 (2018) (cit. on pp. 4, 5, 23, 27, 32, 41, 43).
24. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011) (cit. on pp. 4, 5, 28).
25. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* **17**, 392 (2016) (cit. on p. 10).
26. Mostafavi, H. *et al.* Variable prediction accuracy of polygenic scores within an ancestry group. *eLife* **9**, e48376 (2020) (cit. on p. 18).
27. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575. <https://doi.org/10.1086/519795> (Sept. 2007) (cit. on pp. 23, 29).
28. *Social Science Genetic Association Consortium: Data* <https://www.thessgac.org/data> (cit. on p. 28).
29. Horta, D. Pandas-Plink. <https://pypi.org/project/pandas-plink/> (cit. on p. 29).
30. Donoho, D. & Tanner, J. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**, 4273–4293 (2009) (cit. on p. 29).
31. Vattikuti, S., Lee, J. J., Chang, C. C., Hsu, S. D. H. & Chow, C. C. Applying compressed sensing to genome-wide association studies. *GigaScience* **3**, 10. ISSN: 2047-217X. <http://dx.doi.org/10.1186/2047-217X-3-10> (2014) (cit. on p. 29).

32. Ho, C. M. & Hsu, S. D. Determination of nonlinear genetic architecture using compressed sensing. *GigaScience* **4**. <https://doi.org/10.1186/s13742-015-0081-6> (Sept. 2015) (cit. on p. 29).