Selecting Covariates for Genome-Wide Association Studies

Erez Dor[1], Ido Margaliot[2], Nadav Brandes[2][0000−0002−0510−2546], Or Zuk[2][0000−0001−7587−2944], Michal Linial[2][0000−0002−9357−4526], and Nadav Rappoport[1][0000−0002−7218−2558]

[1] Ben-Gurion University of the Negev, Beer Sheva 8410501, ISRAEL
{erezdo,nadavrap}@bgu.ac.il

[2] The Hebrew University of Jerusalem, Jerusalem 69121, ISRAEL
{ido.margaliot,nadav.brandes,or.zuk,michal.linial}@mail.huji.ac.il

Abstract. The choice of which covariates to include in a Genome-Wide Association Study (GWAS) is important since it affects the ability to detect true association signal of variants, to correct for confounders and avoid false positives, and the running time of the analysis. Commonly used covariates include age, sex, genotyping batches, genotyping array type, as well as an arbitrary number of Principal Components (PCs) used to adjust for population structure. Despite the importance of this issue, there is no consensus or clear guidelines for the right choice of covariates. Therefore, studies typically employ heuristics for their choice with no clear justification. Here, we explore the dependence of the GWAS analysis results on the choice of covariates for a wide range of quantitative and binary human phenotypes. We propose guidelines for covariates choice based on the phenotype's type (quantitative vs. disease), the heritability, and the disease prevalence, with the goal of maximizing the statistical power to detect true associations and fit accurate polygenic scores while avoiding spurious associations and minimizing computation time. We analyze 36 traits in the UK-Biobank dataset. We show that the genotype batch and assessment center can be safely removed as covariates, thus significantly reducing the GWAS computational burden for these traits.

Keywords: Genome-Wide Association Study (GWAS) · Covariates · Principle Component Analysis · Linkage Disequilibrium · Polygenic Risk Score · Population Genetics · UK Biobank.

1 Introduction

A central goal in performing Genome-Wide Association Study (GWAS) is to identify statistically significant associations between genetic variations and phenotype and thus point to the possible biological mechanisms underlying the studied phenotype. However, GWAS is often prone to multiple uncontrolled confounders and biases (e.g., selection bias and population structure) [29]. The most common genetic variations tested in GWAS studies are Single Nucleotide Polymorphisms (SNPs). The standard practice in GWAS is to test each SNP independently for association with the trait, which may lead to a high rate of false positives when confounders that are correlated with both the trait and the variant are not included in the model as covariates (i.e. variants that are labeled as statistically associated with the phenotypes but are actually false positives) [2]. The routine GWAS protocol suggests including covariates whose purpose is to control for the indirect effects unrelated to the phenotype of interest and eliminate the influence of confounders. These covariates include technical components (e.g. the genomic center and SNP-chip technology used for collecting data) but also covariates of biological and medical importance, such as the sex and age of the individual, that may directly affect the phenotype.

In recent years, increased GWAS sample sizes and improved statistical methods have lead to an interest in using GWAS results for genomic prediction using Polygenic Scores (PGS). These scores, defined as weighted linear combination of risk alleles, may include SNPs that do not reach genome-wide statistical significance individually, but together can improve prediction accuracy, and were demonstrated as effective for predicting individuals at risk for disease [15]. Thousands of PGSs were already fitted and are available in resources such as [16], with the number of variants included in the score ranging from a few dozens to hundreds of thousands. In similar to the search for genome-wide significant SNPs, the fitting of a PGS may also be susceptible to confounders, and the fitted score will vary depending on the covariates included. A major issue of current interest is the transferability of the scores between different scenarios. In particular, the scores may not transfer easily between human populations [1,24,30], mainly due to differences in allele frequencies, LD-structure, and effect size. Moreover, scores may show reduced accuracy even within a single population where most above differences are negligible [21], including in prediction of within-family variation [27], with changes in covariates such as socioeconomic status, age and sex leading to decreased accuracy, possibly due to Gene-by-Environment interactions. These issues highlight the need to understand the possible confounders affecting the

accuracy of the fitted PGSs, and the effect of covariate inclusion on the scores, with the hope that such better understanding will aid in the inclusion of the right covariates when adapting a PGS to a new population or cohort.

Covariates can bias the GWAS results [2], but can also adjust for confounders and prevent spurious associations. For example, population structure has been shown to greatly affect GWAS results [18,13], and including genetic principal components (PCs) as covariates is often used to control for population structure [28,23]. Failure to match cases and controls for the right covariates may also lead to substantial inflation of false positive rate [7,19]. In addition to the effect on the false positive rate, adding covariates may also increase or reduce the statistical power to detect true significant associations [22]. Finally, the addition of covariates to the model comes at a computational price, since multiple regression is performed with the covariates for each SNP repeatedly. Therefore, we may not want to include additional covariates if they do not significantly improve the statistical properties of the analysis.

However, it is often unclear which covariates should be included when performing a GWAS, and what effects will this choice have on the GWAS results [20]. The question of whether under a predetermined setting, a preferred set of covariates should be used is critical to improve the detection power of GWAS while also boosting the accuracy of the findings. To this end, we took an exploratory approach, and performed GWAS for a broad range of traits in the UK-Biobank (UKBB) dataset [5]. For each trait, we performed multiple GWAS with different sets of covariates. We list multiple measures of power and false-positive rates such as the estimated genomic control inflation factor [10] to explore the effect of covariates selection on GWAS. We determine the effect of different covariates for different traits. Specifically, we study how does the heritability of phenotypes influence the effects of different covariates. For binary disease traits (e.g., schizophrenia), we also test the dependence on the disease's prevalence.

In this study, we propose a criterion for selecting a set of covariates by designing quantitative measures that will enable high discovery power, as well as avoid spurious discoveries. We suggest an optimal set of covariates for different scenarios. Specifically, we will be interested in covariates sets that achieve multiple, possibly competing goals: to minimize the genomic inflation, to maximize the prediction performance, while also minimizing run time.

In this study we propose recommendations for the choice of covariates as a piece of practical advice when dealing with major human quantitative and binary traits in the UK biobank data. By examining many different sets of covariates for dozens of phenotypes in the UK-Biobank dataset, we find that for this dataset,

the assessment center and genotyping batch can be excluded from the covariates set without compromising GWAS performance. However, for binary traits, there seems to be some effect of the genotype batch and the assessment center when estimating PGS.

2 Methods

2.1 GWAS model

Consider a covariates matrix $Z \in \mathbb{R}_{n\times q}$ with n individuals (rows) and q covariates $Z_1, .., Z_q$ divided into groups $S_1, .., S_K$ (for example, S_1 may contain the Principal Components, S_2 all categorical dummy variables representing assessment center etc.). Consider also the genotypes matrix $X \in \mathbb{R}_{n\times p}$ with SNPs $X_1, .., X_p$, and the phenotypes matrix $Y \in \mathbb{R}_{n\times m}$ with phenotypes $Y_1, .., Y_m$.

We assume a linear model relating a single quantitative phenotype Y to known genetic and non-genetic covariates:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j=1}^{q} \alpha_j z_{ij} + \epsilon_i \ , \ \forall i = 1, .., n \tag{1}$$

Where ϵ_i is an additive noise variable representing environmental effects and other unaccounted-for factors, for individual i. In this model, a SNP j is termed as true (false) causal if $\beta_j \neq 0$ ($\beta_j = 0$). A false causal SNP declared as significant in a GWAS analysis is termed false positive. For disease phenotypes, a similar model is defined using logistic or probit regression.

2.2 The Dataset Used

We used the UKBB dataset which includes genotypic and phenotypic data of about 500,000 subjects [5]. We analyzed 36 phenotypes representing a variety of phenotypes with known genetic contribution, for 19 continuous phenotypes and 17 binary disease phenotypes (See Figure 1 and Table 1 for a detailed list). Data from 488,377 samples was used. Samples without the phenotype of interest were filtered out.

For the UKBB dataset, we examined $q = 168$ possible covariates divided into $K = 6$ groups: Genotyping batch, assessment center, sex, age, First 5 PCs, next 35 PCs. We chose to test nine subsets of covariates. Twenty-two assessment centers were represented by 21 binary covariates (dummy variables). Similarly, 106 genotyping batches were represented by 105 binary covariates (see Table 2).

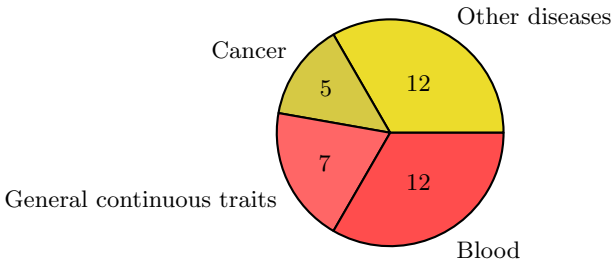

Fig. 1: The number of traits in UKBB that were analyzed in this study. In yellow - seventeen binary traits (diseases) and in red - nineteen continuous traits such as blood tests and other physiological measurements. A complete list is shown in Table 1.

Table 1: List of phenotypes used in the current study. Prevalence and mean/SE were calculated from the UKBB data in this study. Heritability estimates were taken from Neale lab heritability browser at https://nealelab.github.io/UKBB_ldsc/index.html.

Phenotype	**Type**	**Prevalence (%)**		**Heritability***	**ICD10 code**
Asthma	Binary	0.471		0.1090	J45
Bipolar disorder	Binary	0.271		0.7560	F31
Breast cancer	Binary	2.784		0.1100	C50
Chronic lymphocytic leukemia	Binary	0.193		-0.1050	C91
Colorectal cancer	Binary	0.676		0.1200	C18
Crohn and colitis	Binary	1.222		0.2410	[K50, K51]
Epithelial ovarian cancer	Binary	0.266		-0.0486	C56
Hypertension	Binary	22.461		0.0789	I10
Lung cancer	Binary	0.676		0.1170	C34
Melanoma	Binary	0.611		0.0813	C43
Multiple sclerosis	Binary	0.361		0.1170	G35
Parkinson's disease	Binary	0.391		-0.0582	G20
Rheumatoid arthritis	Binary	1.338		0.0007	[M05, M06]
Schizophrenia	Binary	0.143		0.2590	F20
Stroke	Binary	1.018		0.0326	I63
Sudden cardiac arrest	Binary	0.321		0.1460	I46
Type 2 diabetes	Binary	5.404		0.1990	E11
		Mean (units)	**Standard Error (SE)**		**UKBB Field ID**
BMI	Continuous	27.403 (Kg/m^2)	0.007497	0.2480	21001
Diastolic blood pressure	Continuous	82.251 (mmHg)	0.017229	0.1430	4079
Eosinophil counts	Continuous	0.173 (10^9 cells/Litre)	0.000217	0.1840	30150
Height	Continuous	168.729 (cm)	0.014467	0.4850	50
High light scatter reticulocyte count	Continuous	0.018 (10^{12} cells/Litre)	0.000016	0.2480	30300
High light scatter reticulocyte percentage of red cells	Continuous	0.399 (%)	0.000520	0.2480	30290
Hip circumference	Continuous	103.449 (cm)	0.014309	0.2232	49
Lymphocyte counts	Continuous	1.951 (10^9 cells/Litre)	0.001846	0.2100	30120
Mean corpuscular hemoglobin	Continuous	31.547 (picograms)	0.002882	0.2530	30050
Menarche age at onset	Continuous	12.566 (years)	0.005964	0.2090	2714
Monocyte count	Continuous	0.478 (10^9 cells/Litre)	0.000349	0.2300	30130
Neutrophil count	Continuous	4.247 (10^9 cells/Litre)	0.002239	0.1640	30140
Platelet count	Continuous	253.401 (10^9 cells/Litre)	0.094977	0.3080	30080
Red blood cell count	Continuous	4.510 (10^{12} cells/Litre)	0.000646	0.2340	30010
Red cell distribution width	Continuous	13.470 (%)	0.001510	0.2170	30070
Reticulocyte count	Continuous	0.060 (10^{12} cells/Litre)	0.000062	0.2270	30250
Systolic blood pressure	Continuous	140.204 (mmHg)	0.031777	0.1510	4080
Waist circumference	Continuous	90.345 (cm)	0.021079	0.2060	48
White blood cell count	Continuous	6.891 (10^9 cells/Litre)	0.003266	0.1910	30000

Table 2: List of subsets of covariates and number of covariates in each.

Covariates	Number of covariates
$\emptyset$	0
{Age}	1
{Sex}	1
{Sex, Age}	2
{First 5 PCs}	5
{Sex, Age, 40 PCs}	42
{Sex, Age, Assessment Center, 40 PCs}	63
{Sex, Age, Batch, 40 PCs}	147
{Sex, Age, Batch, Assessment Center, 40 PCs}	168

GWAS execution GWAS was performed using Plink2 [6][26] using "`--glm`" command. Covariates were standardized using the "`--covar-variance-standardize`" flag.

2.3 Evaluation metrics

As ground truth for GWAS is not available, we considered multiple evaluation metrics for assessing the quality and optimality of covariates' selection.

Polygenic scores accuracy For each pair of a phenotype and a covariates subset, we trained a PGS model using the GWAS's summary statistics (additive effect size β_j and significance level ($Pval_j$) of every SNP j) with this set of covariates. PGS was computed using PRSice-2 software [12]. The score is estimated by removing SNPs in linkage disequilibrium (LD) and by thresholding the p-values, where an optimal p-value threshold is chosen to optimize the prediction accuracy of the resulting PGS.

We used as an evaluation metric the percent of phenotypic variance explained (R^2) by the PGS [11], a metric indicative of the prediction quality of the PGS model trained based on the GWAS results with a particular covariates set [3]. R^2 was computed by running PRSice-2 using all samples. The same subset of covariates that was used to estimate the effect sizes was also used for estimating R^2 of the PGS. In total, we trained 324 PGS models (36 phenotypes $\times$ 9 covariates sets). The full R^2 represents the variance explained by both the PGS and the covariates.

Genomic control inflation factor To control for inflation of discoveries in GWAS, [10] introduced the λ inflation factor statistic and proposed a method called 'genomic control', which utilizes this statistic to correct for false positive signals. We used the λ inflation factor as an evaluation metric that measures the discrepancy between the empirical p-value distribution and the null $Uniform(0,1)$ distribution. This statistic is defined as the scaled median of the individual SNPs' χ^2 test statistics, with $\lambda = 1$ indicating a complete agreement and a higher value indicating a large number of significant associations that can be due to confounders and/or a true polygenic signal. Genomic λ inflation factor was computed using PLINK-2 [6].

Linkage Disequilibrium Score Regression's Intercept We run LD Score Regression (LDSC) [4] to discriminate between confounders and a true polygenic score for each pair of phenotype and covariates-subset. The measure of interest was the intercept of LDSC-regression (LDSCI). LDSCI provides an estimate of the confounder effect [4] based on a simple yet powerful idea. Since SNPs are correlated due to linkage-disequilibrium, the signal observed in GWAS for a single SNP can be a proxy for the signal of other neighboring SNPs. The LD-score of a SNP measures the cumulative correlations between this SNP and neighboring SNPs. For a true polygenic signal that is spread across many SNPs, we expect, on average more causal SNPs nearby a SNP with a higher LD-score, hence a linear relationship between the χ^2 association statistic of a SNP and its LD-score. In contrast, we expect confounders such as population structure to affect SNPs more uniformly and independently of their LD-score. Therefore, the true polygenic signal is correlated with the LD-score of a SNP and is reflected in the slope of the regression line between the LD-score and the χ^2 association statistic. In contrast, the intercept of this regression analysis reflects the inflation due to confounders. This metric measures the level of spurious associations in a GWAS, with $LDSCI = 1$ indicative of no confounding, and values above 1 indicate confounding (see [17] for additional details). In contrast to the genomic control λ, this metric is not inflated by a true polygenic signal that is correlated with the LD-level of individual SNPs. For computing LDSC we removed strand-ambiguous SNPs, and used the European population from the 1,000 Genomes project as a reference panel [8] and for computing the LD scores.

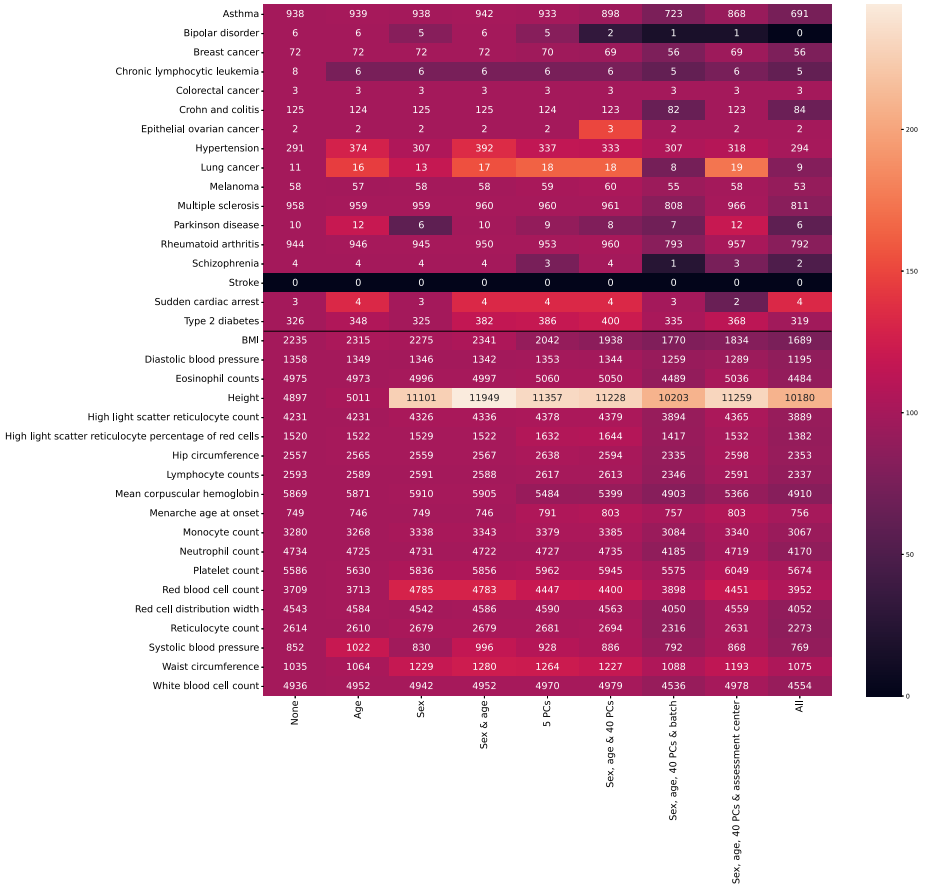

Fig. 2: The number of genome-wide significant SNPs for each phenotype-covariates combination. The color of each cell represents the percentage of significant SNPs compared to the base value on the leftmost column, which is the number of significant SNPs without any covariates. The top rows are binary traits and the bottom rows are continuous.

3 Results

A diverse set of 36 phenotypes was selected for this study covering a variety of traits (Figure 1): 17 disease traits of different prevalence (five cancer diagnoses and twelve diagnoses of other diseases) and 19 continuous phenotypes (twelve blood test results and seven other continuous physiological measurements). The number of cases for the different binary traits ranges from 62,000 for hypertension to only 127 for systemic sclerosis. More details about the phenotypes are shown in Table 1. For each phenotype, we executed multiple GWAS with nine different sets of confounders, as shown in Table 2.

3.1 Genome-wide Significant SNPs

A possible measure of GWAS success is the number of genome-wide significant SNPs. We first recorded this number for each combination of covariates and phenotypes at $\alpha = 5 \times 10^{-8}$ [9]. The number of genome-wide significant SNPs for each trait-covariates combination is shown in Figure 2. The color of each cell represents the percentage of significant SNPs compared with the base value on the left-most column, which is the number of significant SNPs without any covariates.

3.2 Genomic Inflation factor

We computed λ inflation factor across all (phenotype, covariates-subset) pairs. Usually, adding covariates to the association test decreases this inflation factor. As a rule of thumb, $\lambda < 1.1$ is considered acceptable [31]. Therefore, we hypothesized that the subset of covariates that minimizes inflation is more adequate as it controls for the correct confounders of the population. However, adding more covariates may lead to overcorrection and missing true SNPs while lowering the λ value. The λ inflation factors for each run are summarized in Table 3.

3.3 Polygenic Score Accuracy

We evaluated the predictive power of the different GWAS results as used in Polygenic risk scores. For every phenotype and covariates-subset pair, we add the same set of covariates when computing the PGS's R^2. We found that for most binary phenotypes (12 out of 17), using all covariates maximizes the R^2. The second best covariates subset was 'without genotype batch'. The subset of covariates 'without genotype batch and assessment center' achieved R^2 within

Table 3: λ inflation factor for each set of covariates for each phenotype, using the Plink2 software. Upper part contains binary phenotypes and lower part contain continuous phenotypes. Minimal value in each row is in bold face. In each row, the values within 105% from the minimal values are highlighted in magenta. AC-Assessment center; Batch-Genotype measurement batch.

	None	Age	Sex	Sex & Age	5Pcs	W/O Batch & AC	W/O AC	W/O Batch	All
Asthma	1.162	1.162	1.162	1.165	1.159	1.143	**1.14**	1.143	**1.14**
Bipolar Disorder	**1.02**	**1.02**	**1.02**	**1.02**	**1.02**	**1.02**	1.023	**1.02**	**1.02**
Breast Cancer	1.083	**1.08**	1.083	**1.08**	**1.08**	**1.08**	1.083	**1.08**	**1.08**
Chronic Lymphocytic Leukemia	**1.023**	**1.023**	**1.023**	**1.023**	1.026	**1.023**	**1.023**	**1.023**	**1.023**
Colorectal Cancer	1.047	1.047	1.047	1.047	1.044	**1.041**	1.044	**1.041**	1.044
Crohn And Colitis	1.047	1.047	1.047	1.047	1.044	1.044	**1.038**	1.044	**1.038**
Epithelial Ovarian Cancer	1.005	**1.002**	1.005	**1.002**	**1.002**	**1.002**	1.005	**1.002**	1.005
Hypertension	1.398	1.418	1.389	1.418	1.396	1.355	1.351	1.355	**1.337**
Lung Cancer	1.029	1.029	1.029	1.029	**1.02**	**1.02**	1.026	**1.02**	1.023
Melanoma	1.032	1.032	**1.029**	1.032	1.032	1.032	1.032	1.032	1.032
Multiple Sclerosis	**1.026**	1.029	1.029	1.029	**1.026**	**1.026**	1.032	**1.026**	1.032
Parkinson Disease	1.029	1.029	1.032	**1.023**	1.026	1.032	1.029	1.032	1.029
Rheumatoid Arthritis	1.056	1.056	1.056	1.056	**1.053**	**1.053**	**1.053**	**1.053**	**1.053**
Schizophrenia	1.008	1.008	1.011	1.011	**1.002**	1.008	1.011	1.008	1.011
Stroke	1.014	**1.011**	**1.011**	1.014	**1.011**	**1.011**	**1.011**	**1.011**	**1.011**
Sudden Cardiac Arrest	**1.005**	**1.005**	**1.005**	**1.005**	**1.005**	1.008	1.008	1.008	**1.005**
Type 2 Diabetes	1.22	1.22	1.217	1.217	1.217	1.21	**1.207**	1.21	**1.207**
Bmi	1.893	1.901	1.897	1.91	1.832	1.816	1.797	1.816	**1.785**
Diastolic Blood Pressure	1.46	1.464	1.46	1.457	1.457	1.446	1.446	1.446	**1.432**
Eosinophil Counts	1.403	1.403	1.407	1.407	1.375	1.372	**1.369**	1.372	1.372
Height	**1.988**	2.034	2.67	2.842	2.26	2.172	2.159	2.172	2.142
High Light Scatter Reticulocyte Count	1.533	1.536	1.544	1.547	1.507	1.5	1.496	1.5	**1.482**
High Light Scatter Reticulocyte Percentage Of Red Cells	1.256	1.256	1.256	1.256	1.23	1.23	1.227	1.23	**1.22**
Hip Circumference	1.741	1.741	1.741	1.741	1.737	1.73	1.722	1.73	**1.71**
Lymphocyte Counts	1.24	1.243	1.24	1.24	1.227	1.22	**1.214**	1.22	**1.214**
Mean Corpuscular Hemoglobin	1.562	1.577	1.577	1.588	1.386	**1.369**	**1.369**	**1.369**	1.372
Menarche Age At Onset	1.425	1.421	1.425	1.421	1.414	**1.403**	**1.403**	**1.403**	**1.403**
Monocyte Count	1.31	1.317	1.317	1.32	1.307	1.307	**1.3**	1.307	**1.3**
Neutrophil Count	1.507	1.507	1.507	1.507	1.496	1.482	1.475	1.482	**1.471**
Platelet Count	1.668	1.656	1.675	1.672	1.592	1.581	**1.573**	1.581	**1.573**
Red Blood Cell Count	1.649	1.649	1.73	1.733	1.581	1.573	**1.566**	1.573	1.57
Red Cell Distribution Width	1.3	1.303	1.303	1.303	1.3	1.303	1.297	1.303	**1.293**
Reticulocyte Count	1.29	1.29	1.293	1.293	1.29	1.283	1.28	1.283	**1.276**
Systolic Blood Pressure	**1.449**	1.482	**1.449**	1.489	1.485	1.471	1.467	1.471	1.457
Waist Circumference	**1.607**	1.615	1.718	1.733	1.733	1.714	1.702	1.714	1.691
White Blood Cell Count	1.507	1.507	1.504	1.507	1.489	1.478	1.471	1.478	**1.464**

between 98% and 100% of the maximal R^2 for all phenotypes except for Colorectal cancer for which it achieved 62% of the maximal R^2 of 0.58 which was found only using all covariates (Table 4).

For quantitative phenotypes, removing either the genotype batch or the assessment center improved the R^2 compared to using all covariates, and one of the subsets 'without genotype batch and assessment center' and 'without genotype batch' yielded the maximal R^2 for all 19 phenotypes.

Table 4: Percent variance explained (R^2) of PGS for each set of covariates for each phenotype. The R^2 indicates the predictive power of each set of covariates. The upper part contains binary phenotypes and lower part contain continuous phenotypes. Highest value in each phenotype (row) is marked in bold. In each row, the values within 99.5% from the maximal values are highlighted. AC-Assessment center; Batch-Genotype measurement batch

	None	Age	Sex	Sex & Age 5Pcs		W/O Batch & AC	W/O AC	W/O Batch	All
Asthma	0.176907	0.183716	0.176426	0.183139	0.181878	0.198238	0.195514	**0.199345**	0.197075
Bipolar Disorder	0.759259	0.758495	0.758699	0.758565	0.758842	0.760468	0.766036	0.763049	**0.768458**
Breast Cancer	0.60117	0.603884	0.648812	0.651716	0.639314375	0.653294	0.650205	**0.654144**	0.65129
Chronic Lymphocytic Leukemia	0.680206	0.684271	0.689766286	0.686704	0.680713	0.689766286	0.701258	0.692259	**0.702953**
Colorectal Cancer	0.036229757	0.0407673	0.00455048	0.0421024	0.00345742	0.036229757	0.0562653	0.0481774	**0.058288**
Crohn And Colitis	0.731435	0.731939	0.731601	0.731964	0.732347	0.733512	0.734327	0.734396	**0.735227**
Epithelial Ovarian Cancer	0.667381	0.669405	0.685326	0.687685	0.667482	0.690308	0.697756	0.692375	**0.700252**
Hypertension	0.321378	0.448498	0.333611	0.45842	0.408663	0.539991	0.525548	**0.544823**	0.532295
Lung Cancer	0.708452125	0.706394	0.699145	0.706558	0.702632	0.711418	0.713269	0.713089	**0.715112**
Melanoma	0.614372	0.615796	0.614348	0.615731	0.615237	0.618309	0.621796	0.623592	**0.627046**
Multiple Sclerosis	0.759828	0.759462	0.760724	0.761164	0.75906	0.763045	0.76848	0.764093	**0.769575**
Parkinson Disease	0.763048	0.77316	0.763549	0.767823	0.759515	0.775876	0.780987	0.777738	**0.782861**
Rheumatoid Arthritis	0.703014	0.706695	0.706145	0.709938	0.703443	0.711657	0.710824	**0.713176**	0.712635
Schizophrenia	0.768676	0.768461	0.768991	0.769128	0.768402	0.773321	0.785785	0.776503	**0.789035**
Stroke	0.71972	0.729285	0.723017	0.731454	0.720696	0.73358	0.734189	0.734599	**0.735387**
Sudden Cardiac Arrest	0.733326	0.737344	0.73822	0.741809	0.733651	0.743628	0.747937	0.745696	**0.749957**
Type 2 Diabetes	0.64917575	0.649	0.635014	0.655753	0.626909	0.661072	0.651036	**0.66176**	0.652862
BMI	0.266313	0.263294	0.266814	0.264066	0.37541	**0.392171**	0.380359	0.39215	0.382912
Diastolic Blood Pressure	0.267545	0.268902	0.283461	0.284174	0.259099	0.29183	0.282401	**0.296418**	0.28872
Eosinophil Counts	0.281054	0.2804	0.281799	0.281464	0.337361	**0.345903**	0.331365	0.343617	0.331781
Height	0.165523	0.171083	0.634123	0.648699	0.221728	0.76901	0.764686	**0.770324**	0.766625
High Light Scatter Reticulocyte Count	0.286089	0.283968	0.290115	0.288014	0.351981	**0.368696**	0.35588	0.367896	0.357491
High Light Scatter Reticulocyte Percentage Of Red Cells	0.277094	0.27524	0.276706	0.274556	0.354123	**0.35809**	0.344698	0.358026	0.345653
Hip Circumference	0.351245	0.353305	0.351682	0.353441	0.369623	0.377247	0.363526	**0.378476**	0.367524
Lymphocyte Counts	0.228428	0.228727	0.232479	0.231641	0.257995	**0.264321**	0.246929	0.262593	0.247778
Mean Corpuscular Hemoglobin	0.182728	0.180437	0.188824	0.18723	0.361246	0.380889	0.367301	**0.381347**	0.369923
Menarche Age At Onset	0.138983	0.139413	0.147028	0.147387	0.144194	**0.156847**	0.152624	0.156502	0.152928
Monocyte Count	0.32473	0.319854	0.338878	0.333125	0.332496	**0.370327**	0.350578	0.36496	0.352213
Neutrophil Count	0.317717	0.316808	0.317443	0.316219	0.34126	**0.351543**	0.337514	0.350101	0.338207
Platelet Count	0.257853	0.269839	0.309315	0.32005	0.388477	0.450706	0.438915	**0.457588**	0.447697
Red Blood Cell Count	0.189942	0.189336	0.40777	0.406255	0.293416	0.550335	0.54028	**0.551001**	0.542628
Red Cell Distribution Width	0.308922	0.318027	0.308787	0.318387	0.322714	**0.331795**	0.315935	0.330266	0.316305
Reticulocyte Count	0.295111	0.295526	0.301908	0.301956	0.29343	0.305233	0.292988	**0.305493**	0.294866
Systolic Blood Pressure	0.265741	0.344743	0.280878	0.353923	0.240541	0.36858	0.360118	**0.372083**	0.365203
Waist Circumference	0.359712	0.365055	0.498278	0.497498	0.290535	0.515721	0.505542	**0.516025**	0.50826
White Blood Cell Count	0.261694	0.260435	0.261774	0.260086	0.304578	**0.314863**	0.301933	0.314067	0.301941

3.4 LDSCI

We next calculated the LD-Score regression's intercept (LDSCI) using LDSC v1.0.1 [4]. The LDSCI values for each run are summarized in Table 5. Overall,

the results are similar to the result of the genomic control metric. For binary traits the effect of covariates on the LDSCI metric is minimal for most traits except Asthma, Hypertension, and Lung cancer, where the exclusion of Principal Components decreases the LDSCI. For continuous traits, the inclusion of Principal Components seems more critical and may decrease substantially the LDSCI metric.

Table 5: Intercept of LDSC for each set of covariates for each phenotype, using the LDSC software. Upper part contains binary phenotypes and lower part contains continuous phenotypes. Minimal value in each row is in bold face. In each row, the values within 105% from the minimal values are highlighted in magenta. AC-Assessment center; Batch-Genotype measurement batch.

	None	Age	Sex	Sex & Age	5Pcs	W/O Batch & AC	W/O AC	W/O Batch	All
Asthma	1.017	1.017	1.018	1.018	1.016	1.005	1.004	1.004	**1.003**
Bipolar Disorder	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	0.994	0.991	0.994
Breast Cancer	1.014	1.013	1.014	1.013	1.013	**1.012**	**1.012**	**1.012**	1.013
Chronic Lymphocytic Leukemia	**0.981**	0.982	0.982	0.982	0.982	0.982	0.985	0.982	0.984
Colorectal Cancer	0.999	1.0	1.0	1.0	**0.997**	0.998	0.998	**0.997**	0.998
Crohn And Colitis	1.028	1.028	1.028	1.028	1.028	**1.027**	1.029	**1.027**	1.029
Epithelial Ovarian Cancer	0.986	**0.985**	0.986	**0.985**	0.986	0.986	0.988	0.986	0.988
Hypertension	1.091	1.089	1.087	1.088	1.071	1.046	1.048	**1.043**	1.045
Lung Cancer	0.999	1.002	0.999	1.002	0.994	**0.993**	**0.993**	0.994	0.994
Melanoma	1.013	1.013	1.013	1.013	1.011	1.011	**1.007**	1.01	**1.007**
Multiple Sclerosis	1.028	1.028	1.028	1.028	1.028	1.027	**1.026**	1.027	**1.026**
Parkinson Disease	**0.993**	**0.993**	**0.993**	0.994	0.995	0.994	0.998	0.995	0.999
Rheumatoid Arthritis	1.019	1.019	1.02	1.02	**1.018**	**1.018**	1.019	**1.018**	**1.018**
Schizophrenia	0.991	0.991	0.992	0.991	0.99	0.99	**0.989**	0.991	0.99
Stroke	1.003	1.002	1.003	1.003	1.002	**1.001**	1.004	1.002	1.005
Sudden Cardiac Arrest	1.0	1.001	**0.999**	1.0	1.0	1.0	1.001	1.001	1.001
Type 2 Diabetes	1.033	1.033	1.033	1.034	1.034	1.031	1.029	1.029	**1.027**
Bmi	1.144	1.15	1.149	1.154	1.1	1.083	1.084	**1.078**	1.079
Diastolic Blood Pressure	1.063	1.064	1.066	1.066	1.065	1.056	1.055	1.053	**1.05**
Eosinophil Counts	1.134	1.135	1.136	1.137	1.106	1.105	**1.1**	1.105	1.101
Height	1.392	1.428	1.664	1.757	1.367	1.296	1.294	1.291	**1.287**
High Light Scatter Reticulocyte Count	1.096	1.099	1.104	1.106	1.076	1.073	1.067	1.072	**1.066**
High Light Scatter Reticulocyte Percentage Of Red Cells	1.048	1.049	1.048	1.05	1.028	1.025	**1.02**	1.024	1.022
Hip Circumference	1.105	1.105	1.105	1.105	1.099	1.091	1.089	1.088	**1.087**
Lymphocyte Counts	1.06	1.06	1.058	1.059	1.051	1.048	1.047	1.047	**1.046**
Mean Corpuscular Hemoglobin	1.22	1.229	1.23	1.24	1.093	**1.077**	1.081	1.079	1.082
Menarche Age At Onset	1.037	1.037	1.037	1.037	1.028	**1.017**	1.019	1.02	1.018
Monocyte Count	1.08	1.085	1.081	1.087	1.077	1.076	**1.075**	1.077	1.076
Neutrophil Count	1.092	1.094	1.093	1.095	1.086	1.077	**1.072**	1.075	**1.072**
Platelet Count	1.216	1.206	1.214	1.208	1.138	1.13	**1.126**	1.132	1.129
Red Blood Cell Count	1.252	1.253	1.267	1.27	1.157	1.15	**1.145**	1.151	1.146
Red Cell Distribution Width	1.036	1.035	1.034	1.035	1.034	1.033	1.03	1.031	**1.029**
Reticulocyte Count	1.023	1.023	1.025	1.025	1.025	1.022	1.022	**1.018**	**1.018**
Systolic Blood Pressure	1.089	1.096	1.089	1.096	1.093	1.085	1.086	**1.079**	1.08
Waist Circumference	**1.061**	1.066	1.083	1.091	1.087	1.077	1.078	1.071	1.071
White Blood Cell Count	1.116	1.118	1.116	1.119	1.1	1.091	**1.088**	1.092	**1.088**

3.5 Comparison across covariate subsets

Figure 3 shows the distribution of the evaluation metrics for each covariate subset across the phenotypes, thus enabling a high-level view of the effects of covariate subsets on the evaluation metrics across phenotypes. For example, for continuous phenotypes, LDSCI decreases when adding more covariates and saturates when 40 PCs are included. A similar observation can be seen when considering PGS's R^2.

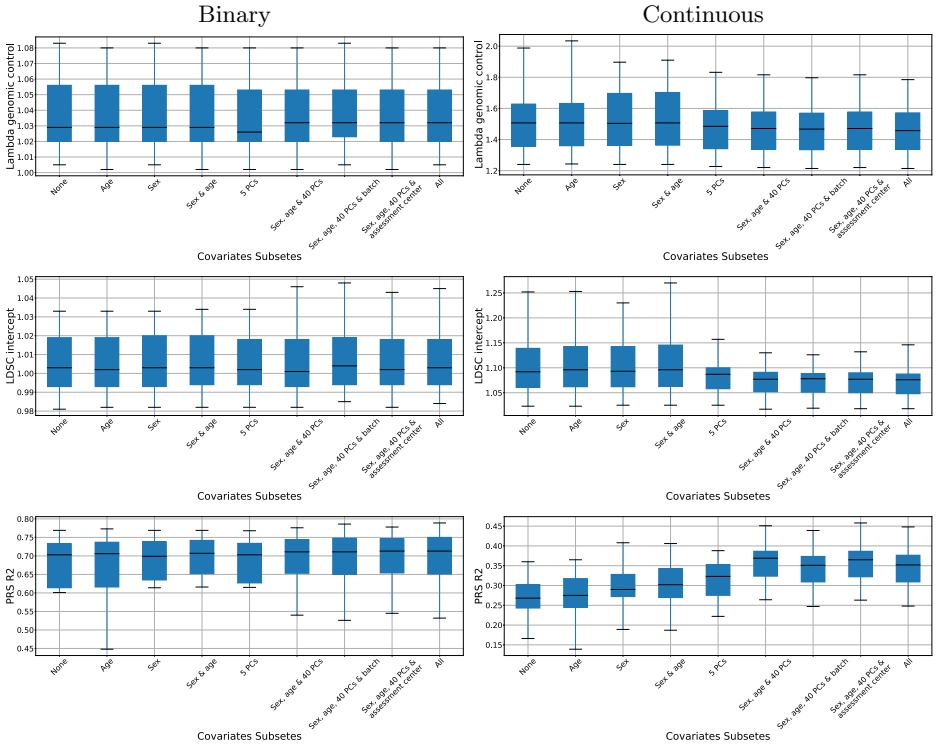

Fig. 3: **Evaluation metrics summary phenotypes.** Top row: λ inflation factor; Middle row: LDSCI, Bottom row: PGS's R^2. Left columns: binary traits, Right columns: continuous traits. The box plots are per subset of covariates across the different phenotypes.

4 Conclusions

In this study, we conducted an empirical evaluation of the set of covariates included in a GWAS study on various metrics representing the GWAS results.

Based on our exploratory analysis, we set to determine a practical recommendation for the choice of covariates to include in the GWAS analysis. The goal is to minimize the $LDSCI$, λ-control, and running time while maximizing the number of genome-wide significant discoveries and the PGS R^2. Based on these criteria, we recommend that PGS estimations include age, sex, and all 40 principal components as covariates for both binary and quantitative traits. It balanced between getting a λ inflation factor close to the minimum, and minimized LDSCI, while PGS R^2 is maximized (Figure 3). For binary traits, the effect of the genotype batch and assessment center seems to be more critical in terms of the PGS's R^2, and we recommend including all covariates of the analysis if this is the goal of the study. That makes sense if genotyping batch and assessment center have no correlation with the genotype, but are correlated with the trait.

While our recommendations are applicable to the UKBB dataset, our empirical approach can be utilized to suggest the set of covariates for other GWAS studies in other cohorts as well, with differences in population structure, sample size, case-control balances, etc. Specifically, our approach suggests balancing the running time and the statistical properties of the results. The running time is quadratic in the number of covariates and could increase substantially, especially when including multiple PCs and dummy variables with many categories such as batches and assessment centers - hence it is desirable to limit the number of covariates if there is no evidence for significant changes in the GWAS results.

A recent study [25] explored the choice of covariates for GWAS in the UKBB dataset. However, this study focused only on Principal Components and their effect on population structure, concluding that 16-18 PCs should be taken for the UKBB population. The authors did not consider the effect of phenotype on the choice of covariates. Our analysis extends these findings, showing that multiple PCs are indeed required, but also exploring the effect on the phenotype on the choice of covariates, the differences between binary traits and quantitative traits, and the inclusion of additional covariates such as batch and assessment center.

5 Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 56774.

References

1. Abdellaoui, A., Verweij, K.J., Nivard, M.G.: Geographic confounding in genome-wide association studies. BioRxiv (2021)
2. Aschard, H., Vilhjálmsson, B.J., Joshi, A.D., Price, A.L., Kraft, P.: Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. The American Journal of Human Genetics **96**(2), 329–339 (2015)
3. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., Duncan, L., Perry, J.R., Patterson, N., Robinson, E.B., et al.: An atlas of genetic correlations across human diseases and traits. Nature genetics **47**(11), 1236–1241 (2015)
4. Bulik-Sullivan, B., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M.: LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics **47**(3), 291–295 (2015)
5. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al.: The UK biobank resource with deep phenotyping and genomic data. Nature **562**(7726), 203–209 (2018)
6. 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 **4**(1) (02 2015). https://doi.org/10.1186/s13742-015-0047-8, https://doi.org/10.1186/s13742-015-0047-8, s13742-015-0047-8
7. Cohen, J.: Statistical power analysis for the behavioral sciences. Routledge (2013)
8. Consortium, .G.P.: An integrated map of genetic variation from 1,092 human genomes. Nature **491**(7422), 56–65 (2012)
9. De Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J., Altshuler, D.: Efficiency and power in genetic association studies. Nature Genetics **37**(11), 1217–1223 (2005)
10. Devlin, B., Roeder, K.: Genomic control for association studies. Biometrics **55**(4), 997–1004 (1999)
11. Dudbridge, F.: Power and predictive accuracy of polygenic risk scores. PLoS Genetics **9**(3), e1003348 (2013)
12. Euesden, J., Lewis, C.M., O'Reilly, P.F.: Prsice: polygenic risk score software. Bioinformatics **31**(9), 1466–1468 (2015)
13. Freedman, M.L., Reich, D., Penney, K.L., McDonald, G.J., Mignault, A.A., Patterson, N., Gabriel, S.B., Topol, E.J., Smoller, J.W., Pato, C.N., et al.: Assessing the impact of population stratification on genetic association studies. Nature genetics **36**(4), 388–393 (2004)
14. Ge, T., Chen, C.Y., Neale, B.M., Sabuncu, M.R., Smoller, J.W.: Phenome-wide heritability analysis of the uk biobank. PLoS genetics **13**(4), e1006711 (2017)
15. 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., et al.: Genome-wide poly-

genic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics **50**(9), 1219–1224 (2018)

16. Lambert, S.A., Gil, L., Jupp, S., Ritchie, S.C., Xu, Y., Buniello, A., McMahon, A., Abraham, G., Chapman, M., Parkinson, H., et al.: The polygenic score catalog as an open database for reproducibility and systematic evaluation. Nature Genetics **53**(4), 420–425 (2021)
17. Lee, J.J., McGue, M., Iacono, W.G., Chow, C.C.: The accuracy of LD score regression as an estimator of confounding and genetic correlations in genome-wide association studies. Genetic epidemiology **42**(8), 783–795 (2018)
18. Marchini, J., Cardon, L.R., Phillips, M.S., Donnelly, P.: The effects of human population structure on large genetic association studies. Nature genetics **36**(5), 512–517 (2004)
19. Maxwell, S.E., Delaney, H.D., Kelley, K.: Designing experiments and analyzing data: A model comparison perspective. Routledge (2017)
20. Mefford, J., Witte, J.S.: The covariate's dilemma. PLoS Genetics **8**(11), e1003096 (2012)
21. Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J.K., Przeworski, M.: Variable prediction accuracy of polygenic scores within an ancestry group. Elife **9**, e48376 (2020)
22. Pirinen, M., Donnelly, P., Spencer, C.C.: Including known covariates can reduce power to detect genetic effects in case-control studies. Nature genetics **44**(8), 848–851 (2012)
23. 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 **38**(8), 904–909 (2006)
24. Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., O'Reilly, P.F., Vilhjálmsson, B.J.: Portability of 245 polygenic scores when derived from the uk biobank and applied to 9 ancestry groups from the same cohort. The American Journal of Human Genetics **109**(1), 12–23 (2022)
25. Privé, F., Luu, K., Blum, M.G., McGrath, J.J., Vilhjálmsson, B.J.: Efficient toolkit implementing best practices for principal component analysis of population genetic data. Bioinformatics **36**(16), 4449–4457 (2020)
26. Purcell, S., Chang, C.: Plink 2.00 alpha (2020), https://www.cog-genomics.org/plink/2.0/
27. Selzam, S., Ritchie, S.J., Pingault, J.B., Reynolds, C.A., O'Reilly, P.F., Plomin, R.: Comparing within-and between-family polygenic score prediction. The American Journal of Human Genetics **105**(2), 351–363 (2019)
28. Tucker, G., Price, A.L., Berger, B.: Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select. Genetics **197**(3), 1045–1049 (07 2014). https://doi.org/10.1534/genetics.114.164285, https://doi.org/10.1534/genetics.114.164285

29. Vilhjálmsson, B.J., Nordborg, M.: The nature of confounding in genome-wide association studies. Nature Reviews Genetics **14**(1), 1–2 (2013)
30. Wang, Y., Namba, S., Lopera, E., Kerminen, S., Tsuo, K., Läll, K., Kanai, M., Zhou, W., Wu, K.H., Favé, M.J., et al.: Global biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. Cell Genomics p. 100241 (2023)
31. Yamaguchi-Kabata, Y., Nakazono, K., Takahashi, A., Saito, S., Hosono, N., Kubo, M., Nakamura, Y., Kamatani, N.: Japanese population structure, based on snp genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. The American Journal of Human Genetics **83**(4), 445–456 (2008)