Title

Polygenic scores without external summary statistics

Author list

Timothy Shin Heng Mak, 1
Robert Milan Porsch, 1
Shing Wan Choi, 2
Pak Chung Sham, 1, 3, 4

Affliations

1. Centre for Genomic Sciences, University of Hong Kong
2. MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London
3. Department of Psychiatry, University of Hong Kong
4. State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong

Correspondence

Timothy Shin Heng Mak (tshmak@hku.hk)
Pak Chung Sham (pcsham@hku.hk)

Abstract

Polygenic scores (PGS) are estimated scores representing the genetic tendency of an individual for a disease or trait and have become an indispensible tool in a variety of analyses. Typically they are linear combination of the genotypes of a large number of SNPs, with the weights calculated from an external source, such as summary statistics from large meta-analyses. Recently cohorts with genetic data have become very large, rendering external summary statistics superfluous. Making use of raw data in calculating PGS, however, presents us with problems of overfitting. Here we discuss the essence of overfitting as applied to PGS calculations, with one of the consequences being the conflation of genetic correlation with environmental correlation. Our simulations show that the impact of overfitting due to the overlap between the Target and the Validation data (OTV) is much less than overfitting due to the overlap between the Target and the Discovery data (OTD), and that a large sample size can vastly reduce OTD in terms of correlation. However, tests of genetic correlations will still be affected by OTD due to increased power. A proposal called *cross prediction* is offered whereby both OTD and OTV can be avoided when calculating PGS without external summary statistics. Software is made available for implementation of the methods.

Introduction

Polygenic scores, or polygenic risk scores (PGS), have become an indispensible tool in genetic studies (Purcell *et al.*, 2009; Opherk *et al.*, 2014; Stahl *et al.*, 2012; Agerbo *et al.*, 2015; Krapohl *et al.*, 2017, 2016; Byrne *et al.*, 2014; Marquez-Luna *et al.*, 2016; Ruderfer *et al.*, 2013; Socrates *et al.*, 2017; Power *et al.*, 2015; Plomin and von Stumm, 2018; Hagenaars *et al.*, 2016; Ripke *et al.*, 2014). Polygenic scores are routinely calculated in small and large cohorts with genotype data, and they represent individual genetic tendencies for particular traits or diseases. As such they can be used for stratifying individuals into different risk group based on their genetic makeup (Ripke *et al.*, 2014; Stahl *et al.*, 2012; Agerbo *et al.*, 2015; Chatterjee *et al.*, 2016). Potentially, different interventions could be given to individuals with different risks, which is part of the vision in personalized medicine (Tremblay and Hamet, 2013; Lenfant, 2013).

Currently, however, the predictive ability of PGS for complex traits remains considerably lower than the heritability, although with increasing sample sizes and the number of Genome-wide association studies, the power is set to increase (Chatterjee *et al.*, 2013; Wray *et al.*, 2014; Plomin and von Stumm, 2018). Nonetheless, even before the objective of personalized medicine can be achieved, PGS can be used for studying the genetic influence of different phenotypes. By examining the correlation between PGS and various phenotypes, researchers can gather evidence for whether the genetic influence on certain traits were pleiotropic or specific (Domingue *et al.*, 2014; Tesli *et al.*, 2014; Chang *et al.*, 2014; Machiela *et al.*, 2011; Power *et al.*, 2015; Krapohl *et al.*, 2016; Byrne *et al.*, 2014). For example, using PGS, Power *et al.* (2015) showed that genetic tendency for schizophrenia and bipolar disorder were predictive of creativity, confirming earlier suggestions that creativity and tendency towards major psychotic illnesses may share some common roots.

Polygenic scores are calculated as weighted sums of the genotype, with weights typically derived from large cohorts or meta-analyses. A key requirement in the calculation of PGS is that the same individuals should not be used both in the calculation of the weights (in the discovery dataset) and the PGS (in the target dataset). Indeed, preferably, in order to avoid inflation in the assessment of correlation, samples in the discovery and target dataset should not even be related (Wray *et al.*, 2013). Recently, cohorts with genotype data have become very large. Examples of such cohorts include the UK Biobank ($n \approx$ 500,000), the 23andMe cohort (Diogo *et al.*, 2017) ($n \approx$ 600,000), and the deCode cohort (Nielsen *et al.*, 2018) ($n \approx$ 350,000). An important question that surfaces is how we are to calculate PGS in these large datasets. In studies to date using the UK Biobank, for example, following the recommended practice, weights for the PGS were calculated from summary statistics and data external to the cohort (Hagenaars *et al.*, 2016; Liu *et al.*, 2017; Nielsen *et al.*, 2018). However, the exclusion of the target dataset from the calculation of the summary statistics seems wasteful, given they have such large sample sizes. Moreover, there may be phenotypes within these large cohorts which are not available elsewhere.

To address this problem, we studied the statistical issues involved in the use of the same sample in both the discovery and the target datasets, which are in essence those of *overfitting*. Overfitting is defined as the inflation of the correlation of the PGS with the genetic component in the target dataset over a completely independent (unseen) external dataset, and we show that three types of overfitting can be delineated, and we considered their relative impact on PGS calculation. Particular emphasis is placed on the possibility of obtaining false positive results due to overfitting, for example, when correlating an overfitted PGS with a phenotype. Overfitting can lead to apparent genetic correlation even when there is none. We propose a method we call *cross prediction* which avoids overfitting, which is made available publicly as an R package.

Material and Methods

Three types of overfitting in calculating polygenic scores

In their review article, Wray *et al.* (2013) pointed out that if the same individuals were used in both the target dataset and the discovery dataset or if they were related, estimates of the predictive power of PGS would be inflated. Although not specifically mentioned, we believe that this phenomenon can largely be explained by the *overfitting* of the data to the target dataset. Here, we define overfitting to be the inflation of the correlation of the PGS with the genetic component in the target dataset over a completely independent (unseen) external dataset. More precisely, let us assume the following linear model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{1}$$

$$\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{2}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ denotes a vector of phenotype from n independent individuals from the *target* dataset and is determined by a genetic component $\boldsymbol{X\beta}$, and residual environmental effects $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$, with ϵ_i assumed independently and identically distributed. We assume $\boldsymbol{X} = (\boldsymbol{x}'_1, \boldsymbol{x}'_2, \ldots, \boldsymbol{x}'_n)'$ is a n-by-p genotype matrix and $\boldsymbol{\beta}$ a vector of causal effects. We also assume without loss of generality that there is no population stratification because if there is, we assume that $\boldsymbol{y}$ and $\boldsymbol{X}$ have the principal components of $\boldsymbol{X}$ regressed out of them as in Price *et al.* (2006). A PGS for an individual i is an estimate of $\boldsymbol{x}_i\boldsymbol{\beta}$, denoted $\text{PGS}_i = \boldsymbol{x}_i\hat{\boldsymbol{\beta}}$. We define overfitting as

$$\text{Cor}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}, y_i) > \text{Cor}(\boldsymbol{x}_i^E\hat{\boldsymbol{\beta}}, y_i^E). \tag{3}$$

where $(\boldsymbol{x}_i, y_i)$ is a randomly chosen sample from the target dataset, and $(\boldsymbol{x}_i^E, y_i^E)$ is a randomly chosen sample from an independent external dataset. Given the independence of $\boldsymbol{X\beta}$ and $\boldsymbol{\epsilon}$, equation (3) can be expressed as

$$\mathrm{Cor}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i\boldsymbol{\beta}) + \mathrm{Cor}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}, \boldsymbol{\epsilon}_i) > \mathrm{Cor}(\boldsymbol{x}_i^E\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i^E\boldsymbol{\beta}) + \mathrm{Cor}(\boldsymbol{x}_i^E\hat{\boldsymbol{\beta}}, \boldsymbol{\epsilon}_i^E). \tag{4}$$

where $\mathrm{Cor}(\boldsymbol{x}_i^E\hat{\boldsymbol{\beta}}, \boldsymbol{\epsilon}_i^E) = 0$ by definition. A sufficient condition for *no* overfitting is thus

$$\mathrm{Cor}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i\boldsymbol{\beta}) = \mathrm{Cor}(\boldsymbol{x}_i^E\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i^E\boldsymbol{\beta}) \tag{5}$$

$$\mathrm{Cor}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}, \boldsymbol{\epsilon}_i) = 0. \tag{6}$$

The fact that when the target data is used to calculate the summary statistics $\hat{\boldsymbol{\beta}}$, overfitting occurs, can be seen by considering a Directed Acyclic Graph (DAG)[1] showing the relationship between $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}\boldsymbol{\beta}$ (Figure 1a). We see that $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is connected to $\boldsymbol{\epsilon}$ through $\boldsymbol{y}$ and thus expected to be correlated. Moreover, because in general we expect $\mathrm{Cor}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}, y_i) > 0$ and $\mathrm{Cor}(y_i, \epsilon_i) > 0$, we expect $\mathrm{Cor}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}, \epsilon_i) > 0$, resulting in overfitting. In this article, we refer to this type of overfitting as OTD (Overfitting due to the overlap between the Target and the Discovery data).

In Figure 1b we see that if we use an external discovery dataset for estimating $\boldsymbol{\beta}$, $\mathrm{Cor}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}, \epsilon_i) = 0$, because the path between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{y}$ is broken. Moreover, if the external discovery sample $\boldsymbol{x}^D$, $\boldsymbol{x}^E$, and $\boldsymbol{x}$ are all drawn from the same population, $\mathrm{Cor}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i\boldsymbol{\beta}) = \mathrm{Cor}(\boldsymbol{x}_i^E\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i^E\boldsymbol{\beta})$ and overfitting is avoided. However, overfitting can still occur if we use the target dataset for selecting the tuning parameters or p-value thresholds, which appears to be common in practice (e.g. Hagenaars *et al.*, 2016; Socrates *et al.*, 2017; Stahl *et al.*, 2017). This situation is illustrated in Figure 1c where there are now arrows pointing to $\hat{\boldsymbol{\beta}}$ from $\boldsymbol{X}$ and $\boldsymbol{y}$. Moreover, the fact that we generally choose the tuning parameter or p-value threshold that maximizes the correlation between the PGS and the phenotype means that there is a Winner's curse such that the apparent correlation between the PGS and the phenotype is higher than it would be in an external dataset. In this article we refer to overfitting due to the target data being used in validation OTV (Overfitting due to the overlap between the Target and the Validation data).

Finally, let us note that the inflation of correlation as cautioned by Wray *et al.* (2013) concerns not only the overlapping of samples. Rather, Wray *et al.* (2013) pointed out that inflation of correlation was likely if the target dataset were genetically related to the discovery dataset. We illustrate this situation in Figure 1d, where correlations are expected between $\boldsymbol{x}$ and $\boldsymbol{x}^D$. Here, although we still have $\mathrm{Cor}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}, \epsilon_i) = 0$, we cannot expect $\mathrm{Cor}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i\boldsymbol{\beta}) = \mathrm{Cor}(\boldsymbol{x}_i^E\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i^E\boldsymbol{\beta})$, leading to overfitting. However, in this article we do not concern ourselves with this type of overfitting, since our primary aim is to

[1] For readers who are unfamiliar, a DAG can be seen as a graphical representation of the probabilistic dependency of the different variables, and its interpretation is grounded in probability theory (Pearl, 2000). Two variables are 'connected' if a line can be traced through the graph connecting the two variables, except when a 'collider' is present along the path that connects the two. A 'collider' is a variable within a path where the two edges connecting it are both arrows pointing towards it, such as teh variables $\boldsymbol{y}$, $\hat{\boldsymbol{\beta}}$, and $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, in Figure 1a. Probabilistically, variables that are connected are expected to be dependent and correlated. Variables that are not connected are not dependent and thus not correlated.

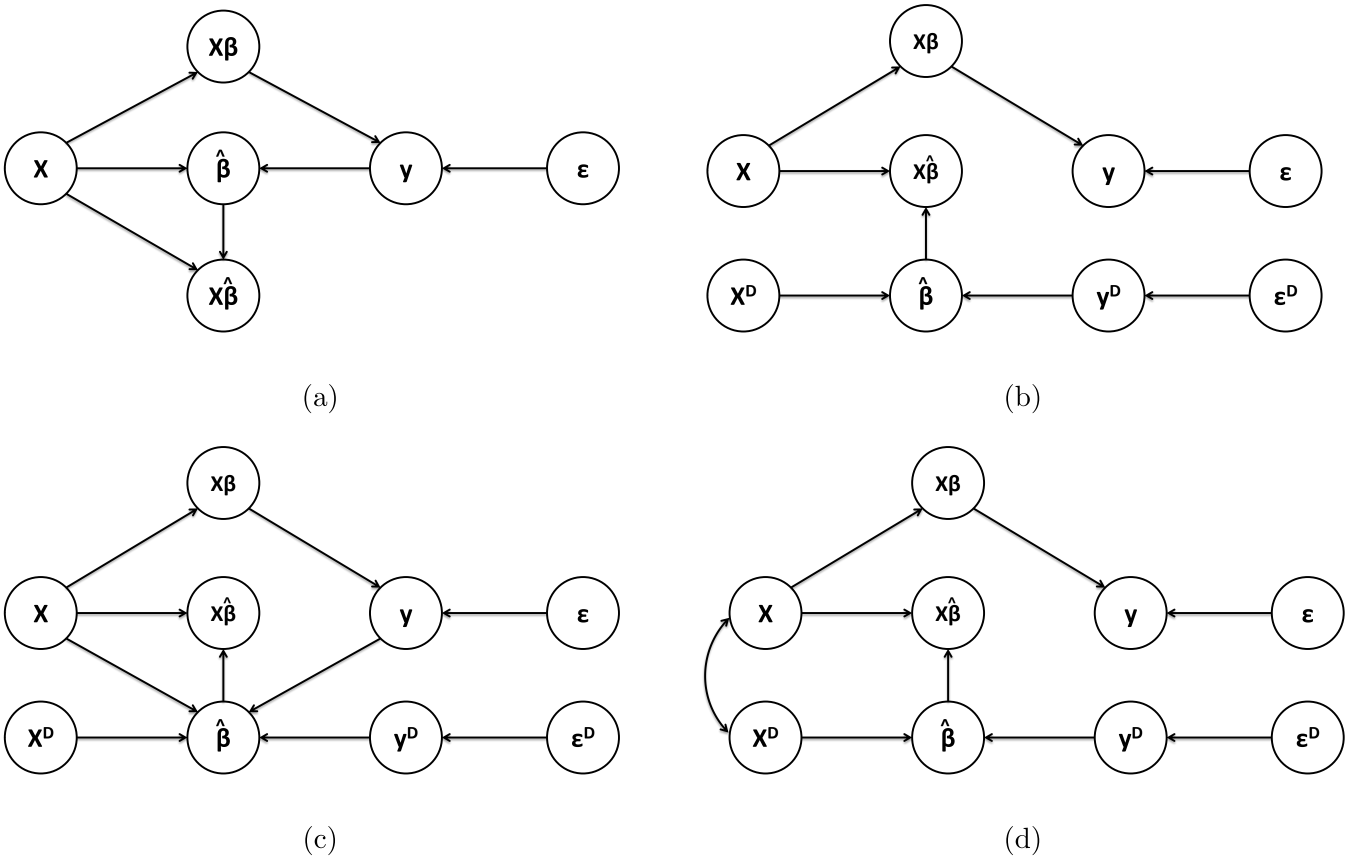


Figure 1: DAGs illustrating the relationship between the different variables in PGS estimation (a) when the target data is also used in the estimation of $\boldsymbol{\beta}$, (b) when a separate discovery dataset $(\boldsymbol{X}^D, \boldsymbol{y}^D)$ is used, (c) when the target dataset is used in choosing the tuning paramter or the best $\hat{\boldsymbol{\beta}}$ among a set of different $\hat{\boldsymbol{\beta}}$s, and (d) when the target dataset is genetically related to the discovery dataset.

derive a method for calculating PGS in a large cohort. In other words, we define our $\boldsymbol{x}^E$ to be the target dataset $\boldsymbol{x}$ and do not worry if the target dataset gives an unbiased estimate of the predictive power of the PGS. What we are concerned with, is when $\text{Cor}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}, \epsilon_i) > 0$, as happens under OTD and OTV, there will be potential bias in using overfitted PGS in assessing genetic correlation, since the apparent correlation between PGS and phenotypes could in fact be due to environmental correlations. Below we introduce a method to avoid such possible false positive conclusions when assessing genetic correlations between different phenotypes.

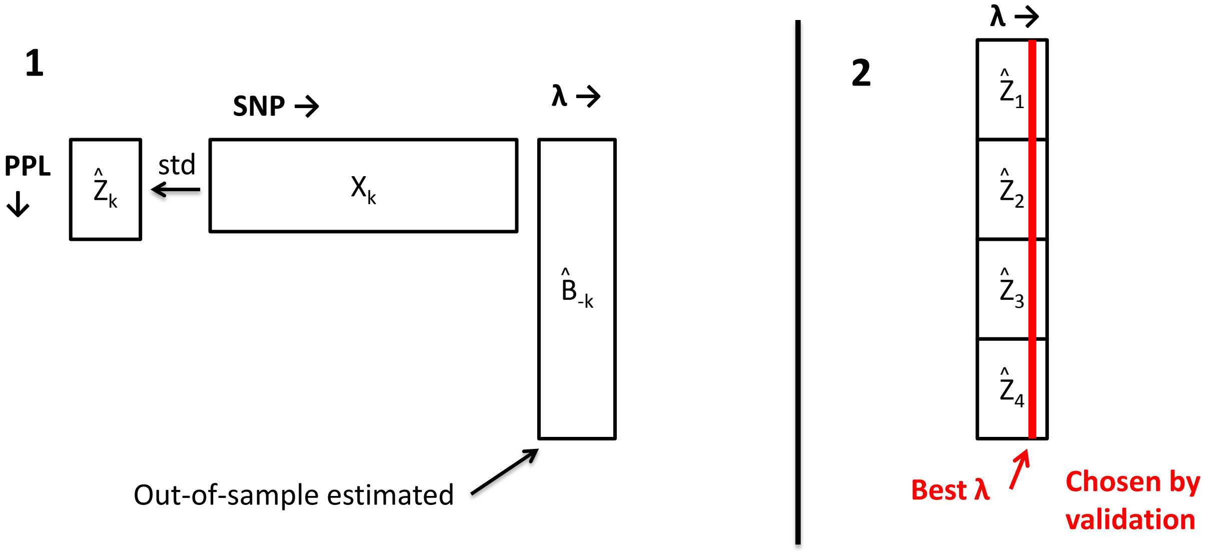
Cross-prediction as a method to overcome overfitting

As already noted above, overfitting can be avoided by breaking the path connecting $\boldsymbol{y}$ to $\hat{\boldsymbol{\beta}}$. One way to do this in practice is to use an independent discovery dataset for estimating $\boldsymbol{\beta}$ (Figure 1(b)). When faced with a large target dataset which we also want to use as our discovery dataset, we can repeat this procedure in a cross-validation-like manner, i.e. we split the data into a number of folds, and repeatedly estimate $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ for the different folds, using the remaining folds for discovery. We call this procedure *cross-prediction*, to distinguish it from the more familiar procedure of cross-validation where fold-splitting is used only for choosing tuning parameters (Varma and Simon, 2006; Abraham *et al.*, 2013; de Maturana *et al.*, 2014). If external summary statistics are available, these can also be meta-analysed with those calculated from the discovery folds. Moreover, in cross-prediction, we propose that PGS from different folds be standardized before being stacked together in forming the final PGS. This is so that the correlation of any variable with the resulting stacked PGS can represent the average correlation between the particular variable and the fold-specific PGS (Appendix A). Moreover, we prove that stacking the fold-specific PGS in this way preserves independence between individual elements of $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\epsilon}$ (Appendix B).

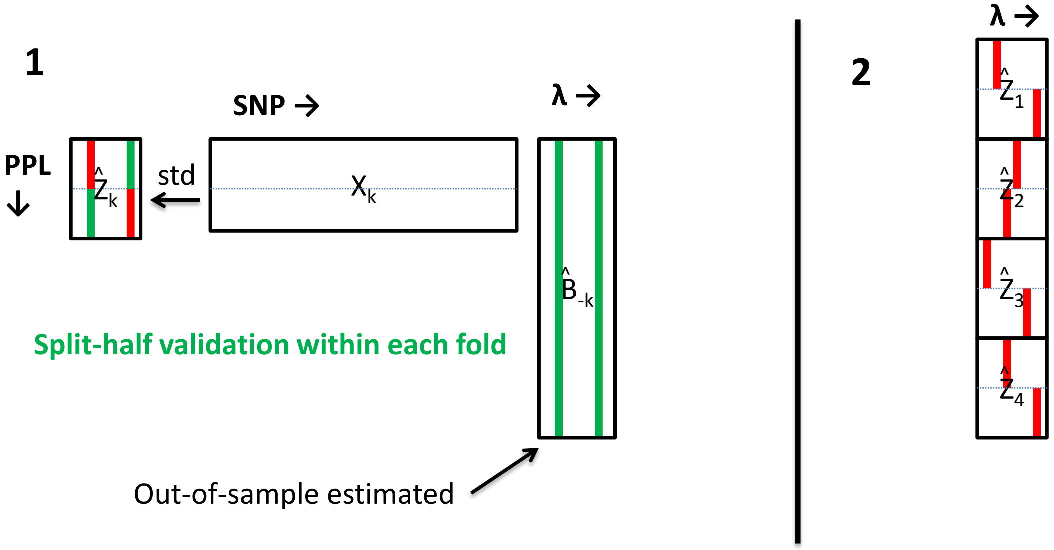
In this section, we present two specific methods (Figure 2) for the implementing cross-prediction using the following notations. Let the overall genotype matrix be $\boldsymbol{X}$, the overall phenotype vector be $\boldsymbol{y}$, and the overall error be $\boldsymbol{\epsilon}$, with the three related as in (1). We assume without loss of generality that there are no covariates for adjustment. If covariates adjustment is needed, simply let the covariates be regressed out of $\boldsymbol{X}$ and $\boldsymbol{y}$. We assume the data $\boldsymbol{\Omega} = \{\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\epsilon}\}$ are split into N folds, such that $\boldsymbol{X} = (\boldsymbol{X}'_1, \boldsymbol{X}'_2, \ldots, \boldsymbol{X}'_N)'$, $\boldsymbol{y} = (\boldsymbol{y}'_1, \boldsymbol{y}'_2, \ldots, \boldsymbol{y}'_N)'$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}'_1, \boldsymbol{\epsilon}'_2, \ldots, \boldsymbol{\epsilon}'_N)'$. Let $\boldsymbol{\Omega}_{-k} = \{\boldsymbol{X}_{-k}, \boldsymbol{y}_{-k}, \boldsymbol{\epsilon}_{-k}\}$ denote the data without the k^{th} fold. For estimating $\boldsymbol{\beta}$, we use straightforward SNP-wise correlation coefficients together with `lassosum` (Mak *et al.*, 2017) to derive m different estimates of $\boldsymbol{\beta}$ indexed by λ. Selection of the optimal λ differs between the two methods and is explained below.

Method 1

1. Use $\boldsymbol{X}_{-k}, \boldsymbol{y}_{-k}$ to estimate $\boldsymbol{\beta}$ (for all values of λ). Let $\hat{\mathbf{B}}_k$ be the p-by-m estimated matrix of $\boldsymbol{\beta}$



(a) Method 1: $\hat{Z}$ is stacked before λ is chosen by validation.



(b) Method 2: Validation is performed first to choose λ (green). The best column within $\hat{Z}$ is then chosen, although the values chosen for stacking are from a different half (red) to that used for validation (green).

Figure 2: Cross-prediction by (a) Method 1, and (b) Method 2

2. Calculate the PGS, $\boldsymbol{X}_k\hat{\mathbf{B}}_k$ for the k^{th} fold, and let $\hat{\boldsymbol{Z}}_k$ be the column-standardized version of $\boldsymbol{X}_k\hat{\mathbf{B}}_k$
3. Repeat step 1 to 2 for all folds and stack together the $\hat{\boldsymbol{Z}}_k$, such that $\hat{\boldsymbol{Z}} = (\hat{\boldsymbol{Z}}_1^{'}, \hat{\boldsymbol{Z}}_2^{'}, \ldots, \hat{\boldsymbol{Z}}_N^{'})'$.
4. Validate $\hat{\boldsymbol{Z}}$ against $\boldsymbol{y}$, i.e. choose the column in $\hat{\boldsymbol{Z}}$ with the highest correlation with $\boldsymbol{y}$, and denote this by $\hat{\boldsymbol{\mu}}$.

Method 2

1. As in Method 1, Use $\boldsymbol{X}_{-k}, \boldsymbol{y}_{-k}$ to estimate $\boldsymbol{\beta}$ (for all values of λ). Let $\hat{\mathbf{B}}_k$ be the p-by-m estimated matrix of $\boldsymbol{\beta}$
2. Split the remaining fold into 2, such that $\boldsymbol{X}_k = (\boldsymbol{X}_{k1}', \boldsymbol{X}_{k2}')'$.
3. Use one half to validate the estimated $\hat{\mathbf{B}}_k$, i.e., to choose the column in $\hat{\mathbf{B}}_k$ which maximizes the correlation between $\boldsymbol{y}_{kl}$ and $\hat{\boldsymbol{Z}}_{kl}$, the column standardized version of $\boldsymbol{X}_{kl}\hat{\mathbf{B}}_k$. Let $\hat{\boldsymbol{\beta}}_{kl}$ denote the column thus identified.
4. Form the PGS using the other half and the chosen $\hat{\boldsymbol{\beta}}_{kl}$. Let this be $\hat{\boldsymbol{\mu}}_{k,3-l} = \boldsymbol{X}_{k,3-l}\hat{\boldsymbol{\beta}}_{kl}$.
5. Repeat step 3 to 4, reversing the role of the two halves.
6. Repeat step 1 to 5 for all other folds.
7. Stack together the $\hat{\boldsymbol{\mu}}_{k,3-l}$ to form the final PGS, i.e. $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_{k1}', \hat{\boldsymbol{\mu}}_{k2}', \ldots, \hat{\boldsymbol{\mu}}_{N2}')'$.

Both Method 1 and Method 2 avoid OTD by separating the (sub-)dataset that estimate $\boldsymbol{\beta}$ from the (sub-)dataset where the PGS is calculated. However, OTV still remains in Method 1. Method 2 avoids both OTD and OTV by further separating the validation data from the data where the PGS is calculated. However, Method 2 uses smaller samples for validation, and may lead to sub-optimal choices of λ. Our simulation experiments as discussed below shed more light on the performance of these methods in practice.

Implementation

The method proposed in this study is implemented an updated version of the `lassosum` package (`https://github.com/tshmak/lassosum`) (Mak *et al.*, 2017). Thus, in this framework, $\hat{\mathbf{B}}$ is estimated in two stage. In the first stage, univariate summary statistics (correlation coefficients) are estimated from the raw data. This stage can be performed efficiently using the highly optimized `plink` software (Chang *et al.*, 2015). If available, summary statistics from external sources can also be meta-analysed together with those in the calculated from the data. In the second stage, LASSO or elastic net estimates are derived using `lassosum`. A major advantage of this approach over performing LASSO or elastic net on the raw dataset is that one can use only a subset of the data as the reference panel (while using the entire dataset for the summary statistics). When dealing with data the size of the UK Biobank it becomes virtually impossible to carry out elastic net or LASSO in the usual manner with moderate computing resources, particularly as we need to repeat calculations for the N folds. Using a subset of the data for the reference panel is a compromise between subsetting the entire dataset and not subsetting at all. In our simulation, a random sample of 1000 is taken as the reference panel.

Simulations

In the first set of simulations, we examined how severe a problem overfitting is in a relatively small dataset. We used the Wellcome Trust Case Control Consortium (WTCCC) dataset of 15,605 individuals and 359,973 SNPs (after QC). We randomly sampled 1,000 individuals. 1,000 SNPs were randomly assigned to be causal. The causal effects followed a distribution of

$$\beta_i \sim N(0, 1) \tag{7}$$

The phenotype was simulated as in equation (1), although to avoid unnecessary complication we assumed no covariates. Further, we assumed a heritability of 0.5, i.e.,

$$\epsilon_i \sim N(0, \hat{\text{Var}}(\boldsymbol{X\beta})) \tag{8}$$

We examined the correlation of the estimated PGS with the $\boldsymbol{X\beta}$. We used 2-fold and 5-fold cross-validation as well as the method of *cross-prediction* described above (both OTD and OTV). We also examined the correlation of the PGS in an unrelated sample of 1,000 individuals from the same data.

As we aim to apply this method in a dataset the size of the UK Biobank, we also performed simulations on the UK Biobank data (UKBB). First, we extracted 7,185,952 variants with MAF $\geq$ 0.01 from the white British subset of the data ($n = 300, 163$). Here, we simulated our 'true' PGS using 2,000 causal SNPs chosen randomly, with the β_i following the distribution of (7). In the UKBB simulations, we assumed our phenotype was binary and simulated according to the liability threshold model:

$$z_i = \begin{cases} 1 & \text{if } y_i > t \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where t is a threshold determined by the prevalence of the disease and $\boldsymbol{z} = (z_1, z_2, \dots, z_n)'$ is the phenotype. n =1,000, 10,000, and 100,000 individuals were randomly chosen from the dataset. When $n = 1,000$, the prevalence of disease was set as 0.1. When $n = 10,000$, prevalence was either 0.01 or 0.1. When $n = 100,000$, the prevalence was either 0.001, 0.01, or 0.1.

In the second set of simulations, we examined whether overfitting may lead to the detection of genetic correlation between variables when there is none. Here, we simulated two phenotypes using (1), with $\boldsymbol{\beta}$

simulated as before for both the WTCCC and the UKBB datasets. Denoting the two phenotypes by $\boldsymbol{y_1} = \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$ and $\boldsymbol{y_2} = \boldsymbol{X}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2$, we further constrained $\boldsymbol{X}\boldsymbol{\beta}_1$ and $\boldsymbol{X}\boldsymbol{\beta}_2$ to have a correlation of 0 using a procedure described in Appendix C. We then simulated $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ such that $\mathrm{Cor}(\epsilon_{1i}, \epsilon_{2i}) = 0.5$. Thus, the two phenotypes had environmental correlation but not genetic correlation. We examined whether the distribution of the p-values when testing for correlation between $\boldsymbol{X}\hat{\boldsymbol{\beta}}_1$ and $\boldsymbol{X}\hat{\boldsymbol{\beta}}_2$ as well as between $\boldsymbol{X}\hat{\boldsymbol{\beta}}_1$ and $\boldsymbol{y}_2$ were significantly different from a null distribution.

WTCCC simulations were repeated 20 times and UKBB simulations were repeated 10 times.

Results

Figure 3 shows the comparison of correlation between the estimated PGS and $\boldsymbol{X}\boldsymbol{\beta}$ using cross-validation (CV) and cross-prediction (CP) from the WTCCC simulation. The most striking result is that the correlation of the CV PGS in the dataset that generated the summary statistics (internal) was vastly higher than in an external dataset, suggesting OTD overfitting had a large impact on the fit in CV. This overfitting was much reduced when using CP. However, a small degree of overfitting could still be observed in CP (Method 1), showing the much smaller impact of OTV. 5-fold CP was slightly more predictive than 2-fold CP. Figure 4 shows the same with the UKBB data, using different sample sizes and prevalence in a liability threshold model. When the sample size was 1,000 or 10,000, the same pattern of overfitting was observed as in the WTCCC example. However, when the sample size was 100,000, the impact of OTD was also much less, and OTV was hardly noticeable.

Figure 5 shows the results of the second set of simulations using the WTCCC dataset. Similar to the previous set, the CV PGS_1 was highly correlated with $\boldsymbol{X}\boldsymbol{\beta}_1$. The correlation with $\boldsymbol{y}_1$ was nearly near 0.9, considerably higher than the true correlation between $\boldsymbol{y}_1$ and $\boldsymbol{X}\boldsymbol{\beta}_1$, set at $\sqrt{0.5} \approx 0.7$ in these simulations. However, a very high correlation with $\boldsymbol{\epsilon}_1$ showed that this was in large part due to correlation with $\boldsymbol{\epsilon}_1$. This correlation with $\boldsymbol{\epsilon}_1$ spilled over into correlation with $\boldsymbol{\epsilon}_2$, and in turn with $\boldsymbol{y}_2$, since $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ were correlated. Thus we see that apparent genetic correlation between a PGS calculated using CV could in fact be due to environmental correlations. These overfits were largely abated when using CP instead of CV, although as in the first set of simulation, the correlation of CP (Method 1) PGS_1 with $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ was slightly higher than expected. Figure 6 gives qq-plot of the p-values of a regression of $\boldsymbol{y}_2$ on the estimated PGS. Inflation of the p-values were observed with CV but not with CP (both Method 1 and Method 2), suggesting a negligible impact of OTV.

Figure 7 shows the same simulation in the UKBB dataset. When $n = 1,000$ or $n = 10,000$, a similar pattern of overfitting was observed. When $n = 100,000$ the degree of OTV was visibly less. Although the CV PGS_1 was still clearly associated with $\boldsymbol{\epsilon_1}$, its association with $\boldsymbol{\epsilon_2}$ and $\boldsymbol{y}_2$ was very low. Figure 8 gives the qq plots from the regression of $\boldsymbol{y}_2$ on PGS_1. Although the correlation between $\boldsymbol{y}_2$ and PGS_1 was apparently low in Figure 5, we see that an inflation of p-values is still very much evident when using

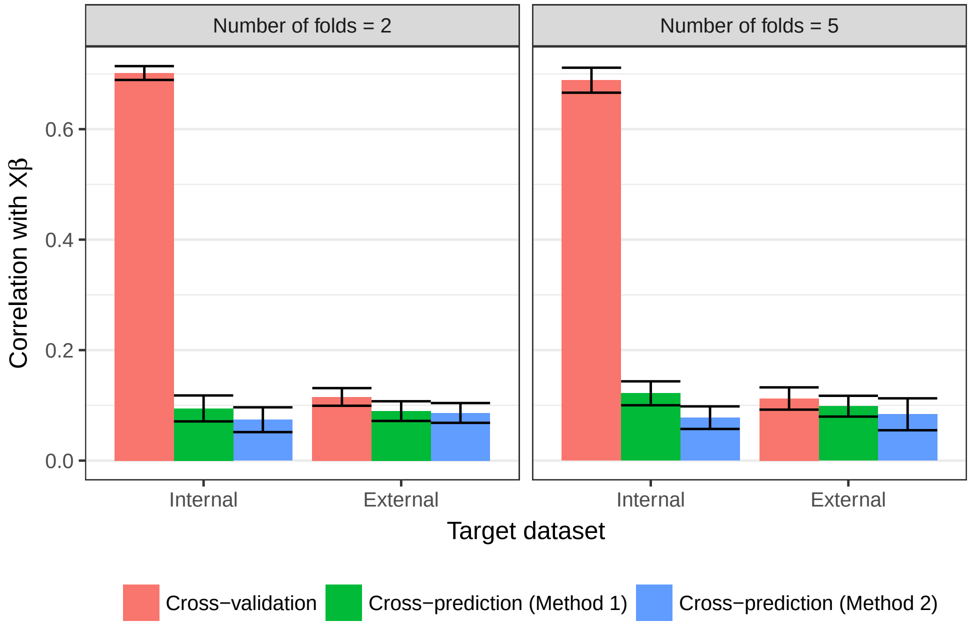


Figure 3: Comparison of correlation of PGS with $\boldsymbol{X\beta}$ in the simulated data based on the WTCCC ($n = 1,000$): cross-validation versus PGS from cross-prediction. Error bars represent the 95% confidence intervals.

CV.

Timings

Disregarding time taken to generate summary statistics, running CP on the WTCCC dataset took on average 1,000 seconds for 2 folds and 2,300 seconds for 5 folds without parallelization. Running CP on the UKBB dataset took on average 3,100 seconds (5 folds) when parallelized over 9 threads. Note that results for Method 1 and Method 2 CP and CV can all be collected in one go. Note also that the WTCCC simulations used $s = (0.2, 0.5, 0.9, 1)$ while the UKBB simulations used $s = (0.5, 1)$ as parameters in `lassosum`. When $s = 1$, `lassosum` is much faster than when $s < 1$. Running CP in the UKBB data using the same settings as in WTCCC would take around 3 times longer.

Discussion

In this article, we have given a theoretical account of overfitting when using the same sample both for calculating summary statistics and PGS. It is shown that overfitting can be due to the target dataset overlapping with the discovery dataset (OTD) and/or the validation dataset (OTV). It can also be due to genetically related samples between the target and the discovery dataset (Type 3). We propose *cross-prediction* (CP) as a practical method for overcoming this bias. Two methods for implementing cross-

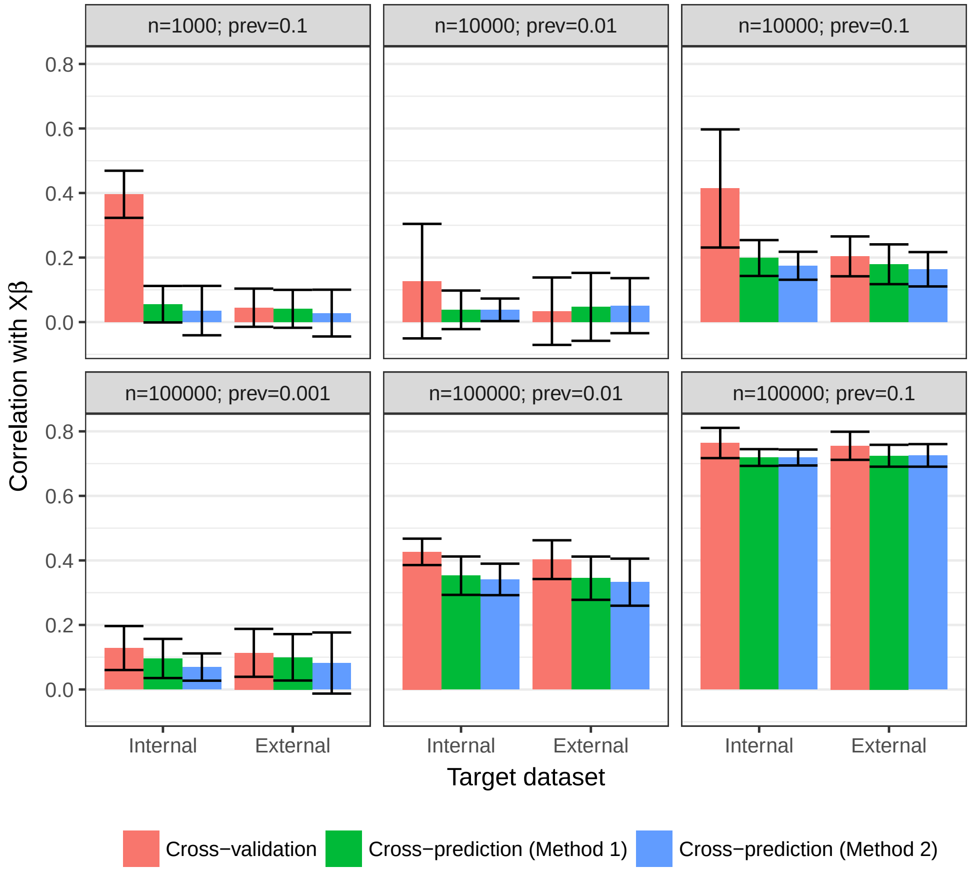


Figure 4: Comparison of correlation of PGS with $\boldsymbol{X\beta}$ in the simulated data based on the UKBB: cross-validation versus PGS from cross-prediction. 5-fold CV and CP were used. Error bars represent the 95% confidence intervals.

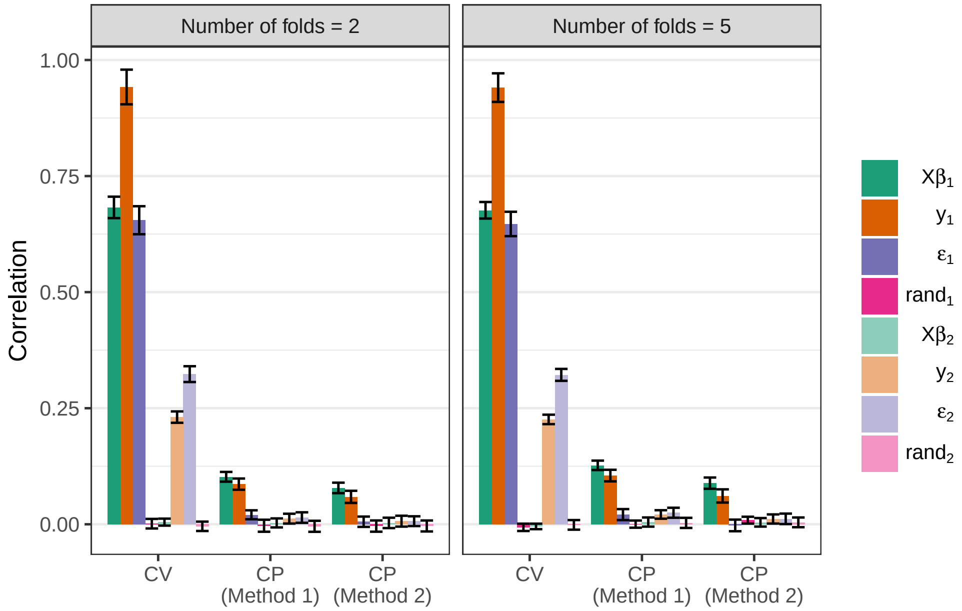


Figure 5: Correlation of estimated PGS$_1$ with $(\boldsymbol{y}_1, \boldsymbol{X\beta}_1, \boldsymbol{\epsilon}_1, \boldsymbol{y}_2, \boldsymbol{X\beta}_2, \boldsymbol{\epsilon}_2)$ in the second simulation based on the WTCCC data ($n = 1,000$). $\boldsymbol{X\beta}_1$ and $\boldsymbol{X\beta}_2$ were uncorrelated but $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ were. rand$_1$ and rand$_2$ were randomly generated Normal variables as controls. Error bars represent the 95% confidence intervals.

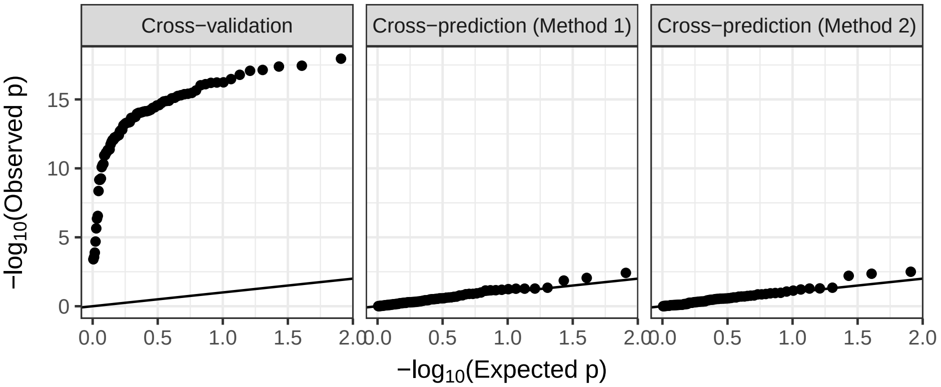


Figure 6: qq-plot of p-values when regressing $\boldsymbol{y}_2$ on the estimated PGS$_1$ in the WTCCC simulation. Data from simulations with 2 and 5 folds were combined together to increase samples.

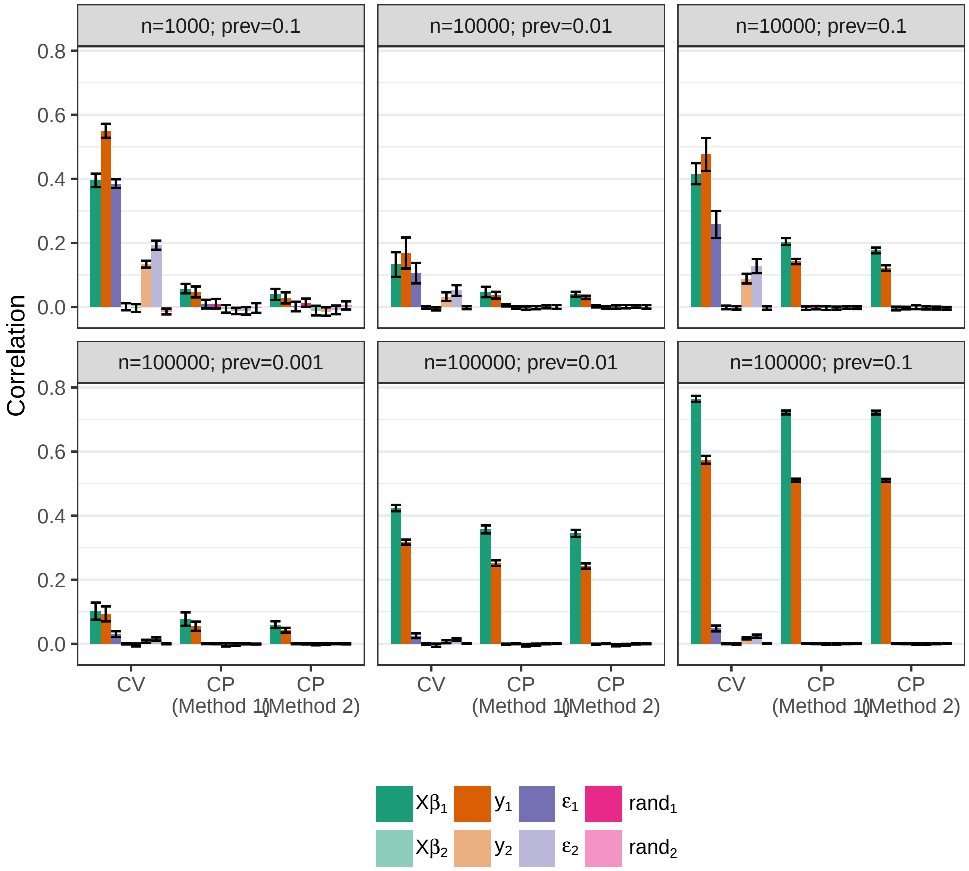


Figure 7: Correlation of estimated PGS_1 with $(\boldsymbol{y}_1, \boldsymbol{X\beta}_1, \boldsymbol{\epsilon}_1, \boldsymbol{y}_2, \boldsymbol{X\beta}_2, \boldsymbol{\epsilon}_2)$ in the second simulation based on the UKBB data. $\boldsymbol{X\beta}_1$ and $\boldsymbol{X\beta}_2$ were uncorrelated but $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ were. rand_1 and rand_2 were randomly generated Normal variables as controls. Error bars represent the 95% confidence intervals.

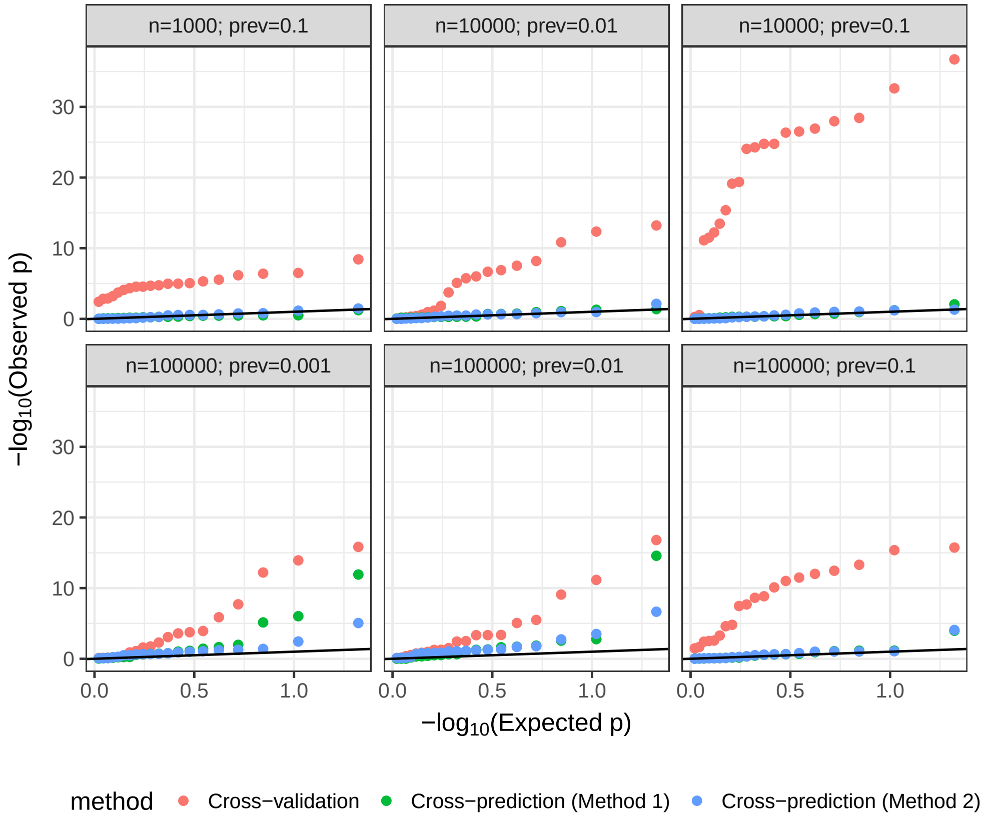


Figure 8: qq-plot of p-values when regressing $\boldsymbol{y}_2$ on the estimated PGS_1 in the UKBB simulation.

prediction have been made available publicly in an R package (`https://github.com/tshmak/lassosum`). CP (Method 1) has greater power, but does not eliminate Type 2 overfitting. CP (Method 2) completely eliminates overfitting, but is less powerful. Simulations suggested the impact of OTV is relatively small, and although the impact of OTD was large in smaller samples, in sample size approaching that of the UK Biobank, the impact of OTD in terms of inflation in correlation was also small. However, p-value calculations can still be inflated due to the larger power associated with larger samples.

We suggest using CP (Method 2) for examining genetic correlations in large samples, although in smaller samples, CP (Method 1) can be used, and if a significant result is obtained, we further carry out analysis using CP (Method 2) for verification. In situations where overfitting is not an issue (e.g. in patient risk stratification), cross validation should be used since it has the most power. Due to the vast demand on computational resources, we have not carried out an exhaustive simulation on the influence of the number of folds in performance, although it appears from our simulation that using 5 folds has greater power than using 2 folds. In general, larger number of folds means more information is captured by the summary statistics, but it also means the validation data sets has a smaller sample size when performing CP (Method 2), and a higher computational burden.

We have not discussed overfitting due to other kinds of overfitting in order to keep our focus. We note that while CP (Method 2) may have avoided overfitting when being used to assess genetic correlations, they can over-estimate prediction power if used within a sample that is related (Wray *et al.*, 2013). Moreover, we caution that when samples are very related, for example, in twin or family studies, then their environmental components are also likely to be correlated. When $\text{Cor}(\epsilon_i^D, \epsilon_i) > 0$, it is likely that $\text{Cor}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}, \epsilon_i) > 0$, resulting in overfitting. One way to overcome this is to define the folds such that families are not split across different folds.

Another note of caution concerns the use of PGS in genetic correlation calculations. Usually genetic correlations can be assessed by examining the relationship between the PGS and various phenotypes. However, in principle we can also examine the correlation between PGS calculated for different phenotypes. We note that overfitting can still occur when correlating different PGS calculated using CP. This is because in CP we try to keep the discovery and the target samples separate. However, when two PGS are both calculated using CP, their discovery samples can overlap, leading to overfitting.

We conclude with a number of suggestions for future work. First, depending on the number of folds use, a proportion of the sample is left out in the calculation of the summary statistics. It is unsure whether there can be a procedure that uses all data and also avoids OTD and OTV. Secondly, the current procedure is stochastic as the folds are randomly defined. The resulting PGS is also not a linear predictor in that it is not calculated as a linear combination of $\boldsymbol{X}$. Rather it is a mixture of different linear combinations. This has the disadvantage that theoretical properties of the PGS are less easily obtained. In principle, it is possible to find estimates of $\boldsymbol{\beta}$ such that when multiplied with $\boldsymbol{X}$, equals the CP PGS as calculated in our study. However, in our preliminary simulations, these estimates of $\boldsymbol{\beta}$

had very poor performance in external validation and we have not pursued this approach further. It is also possible in principle to extend this work further to the case where the number of folds used equals the sample size, such that we have a jackknife-like procedure for cross-prediction. This approach has not been studied. Thirdly, calculation of PGS using cross-prediction is currently very time consuming for large cohorts, as it involves repeating the procedure for the N folds. Performing informed pruning (clumping) (Euesden *et al.*, 2015) on SNPs before CP is a possible remedy which has not been tested in the current study.

Acknowledgement

This research has been conducted using the UK Biobank Resource (project ID 23079, 28732).

Web resource

`lassosum` (`https://github.com/tshmak/lassosum`)

List of Figures

A Standardizing PGS within fold before stacking approximates the average correlation of the PGS with another variable

Let $\boldsymbol{x} = (\boldsymbol{x}'_1, \boldsymbol{x}'_2, \ldots, \boldsymbol{x}'_N)'$ denote a stacked column of PGS, and $\boldsymbol{y}$ a column of phenotype. Further assume $\boldsymbol{x}$ is standardized within fold, such that $\mathbf{1}'\boldsymbol{x}_k = \mathbf{0}$ and $\boldsymbol{x}'_k\boldsymbol{x}_k = n_k$, and that $\boldsymbol{y}$ is standardized such that $\mathbf{1}'\boldsymbol{y} = \mathbf{0}$ and $\boldsymbol{y}'\boldsymbol{y} = n = \sum_k n_k$ without loss of generality. The correlation of $\boldsymbol{x}$ with $\boldsymbol{y}$ is $\boldsymbol{x}'\boldsymbol{y}/n$. Let the standard deviation of $\boldsymbol{y}$ within fold k be $1/s_k$. We have

$$\frac{\boldsymbol{x}'\boldsymbol{y}}{n} = \sum_k \frac{\boldsymbol{x}'_k\boldsymbol{y}_k s_k}{n_k}\frac{n_k}{s_k n} \tag{10}$$

where $\frac{\boldsymbol{x}'_k\boldsymbol{y}_k s_k}{n_k}$ is the fold-specific correlation. Thus, $\frac{\boldsymbol{x}'\boldsymbol{y}}{n}$ is a weighted average of the fold-specific correlation with weights $\frac{n_k}{s_k n}$. In general s_k approximates 1, such that the weights are approximately optimal.

B Proof that $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ remain independent of with $\boldsymbol{\epsilon}$ after stacking

As in the main text, we assume that $\boldsymbol{X} = (\boldsymbol{X}'_1, \boldsymbol{X}'_2, \ldots, \boldsymbol{X}'_N)'$, $\boldsymbol{y} = (\boldsymbol{y}'_1, \boldsymbol{y}'_2, \ldots, \boldsymbol{y}'_N)'$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}'_1, \boldsymbol{\epsilon}'_2, \ldots, \boldsymbol{\epsilon}'_N)'$. Denote $\boldsymbol{z}_k = \boldsymbol{X}_k\hat{\boldsymbol{\beta}}$. From Figure 1(b), we establish that $\boldsymbol{z}_k$ is independent of $\boldsymbol{\epsilon}$ if $\hat{\boldsymbol{\beta}}$ is derived from a different fold from $\boldsymbol{z}_k$. It follows that the i^{th} element of $\boldsymbol{z}_k$, denoted z_{ki} is independent of the i^{th} element of $\boldsymbol{\epsilon}$, within a particular fold $\mathcal{F}$. In notation:

$$f_{z_i,\epsilon_i|\mathcal{F}}(z_i, \epsilon_i) = f_{z_i|\mathcal{F}}(z_i) f_{\epsilon_i|\mathcal{F}}(\epsilon_i) \tag{11}$$

Proof: $f_{z_i,\epsilon_i}(z_i, \epsilon_i) = f_{\epsilon_i}(\epsilon_i) f_{z_i}(z_i)$.

$$\begin{aligned} f_{z_i,\epsilon_i}(z_i, \epsilon_i) &= \sum_{\mathcal{F}} p(\mathcal{F}) f_{z_i,\epsilon_i|\mathcal{F}}(z_i, \epsilon_i) && (12) \\ &= \sum_{\mathcal{F}} p(\mathcal{F}) f_{z_i|\mathcal{F}}(z_i) f_{\epsilon_i|\mathcal{F}}(\epsilon_i) && (13) \end{aligned}$$

Now, because ϵ_i are assumed $i.i.d.$ regardless of fold, we have

$$\begin{aligned} f_{\epsilon_i|\mathcal{F}}(\epsilon_i) &= f_{\epsilon_i}(\epsilon_i) && (14) \\ f_{z_i,\epsilon_i}(z_i, \epsilon_i) &= f_{\epsilon_i}(\epsilon_i) \sum_{\mathcal{F}} p(\mathcal{F}) f_{z_i|\mathcal{F}}(z_i) && (15) \\ &= f_{\epsilon_i}(\epsilon_i) f_{z_i}(z_i) && (16) \end{aligned}$$

completing the proof. □

C A procedure for simulation two PGS with zero correlation

We seek $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ such that

$$\boldsymbol{\beta}_1'\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta}_2 = 0 \tag{17}$$

$$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}'/n \tag{18}$$

where $\boldsymbol{X}$ is of dimension n by p, with $n < p$.

Letting $x = \boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta}_1$, we randomly generate $\boldsymbol{\gamma}$. Letting $\boldsymbol{\beta}_2$ be the residuals of regression $\boldsymbol{\gamma}$ on x, i.e., if

$$\boldsymbol{\beta}_2 = \boldsymbol{\gamma} - \boldsymbol{x}\boldsymbol{x}'\gamma/\boldsymbol{x}'\boldsymbol{x} \tag{19}$$

$\boldsymbol{X}\boldsymbol{\beta}_2$ would be uncorrelated with $\boldsymbol{X}\boldsymbol{\beta}_1$.

However, we want a fixed number of values in $\boldsymbol{X}\boldsymbol{\beta}_2$ to be zero. To achieve this, we generate $\boldsymbol{\gamma}$ in such a way so that $\boldsymbol{x}'\boldsymbol{\gamma} = \boldsymbol{x}'\boldsymbol{x}$. In this way,

$$\boldsymbol{\beta}_2 = \boldsymbol{\gamma} - \boldsymbol{x} \tag{20}$$

Moreover, letting $\mathcal{C}$ denote the set of i where $\beta_i = 0$, and $\not{\mathcal{C}}$ its complement, we set

$$\delta_i \sim N(0,1) \tag{21}$$

$$\gamma_i = \begin{cases} x_i & \text{if } \beta_i = 0 \\ \delta_i c & \text{otherwise} \end{cases} \tag{22}$$

$$c = \boldsymbol{x}'_{\not{\mathcal{C}}}\boldsymbol{x}_{\not{\mathcal{C}}}/\boldsymbol{x}'_{\not{\mathcal{C}}}\boldsymbol{\delta}_{\not{\mathcal{C}}} \tag{23}$$

This ensures $\sum_{i\in\not{\mathcal{C}}} x_i\delta_i c = \boldsymbol{x}'_{\not{\mathcal{C}}}\boldsymbol{x}_{\not{\mathcal{C}}}$, and thus $\boldsymbol{x}'\boldsymbol{\gamma} = \boldsymbol{x}'\boldsymbol{x}$.

References

Abraham G, Kowalczyk A, Zobel J, and Inouye M (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic epidemiology*, 37(2), 184–95

Agerbo E, Sullivan PF, Vilhjálmsson BJ, Pedersen CB, Mors O, Børglum AD, Hougaard DM, Hollegaard MV, Meier S, Mattheisen M *et al.* (2015). Polygenic Risk Score, Parental Socioeconomic Status, Family History of Psychiatric Disorders, and the Risk for Schizophrenia: A Danish Population-Based Study and Meta-analysis. *JAMA psychiatry*, 72(7), 635–41

Byrne EM, Carrillo-Roa T, Penninx BWJH, Sallis HM, Viktorin A, Chapman B, Henders AK, Pergadia ML, Heath AC, Madden PAF *et al.* (2014). Applying polygenic risk scores to postpartum depression. *Archives of Women's Mental Health*, 17(6), 519–528

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, and Lee JJ (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 1–16

Chang SC, Glymour MM, Walter S, Liang L, Koenen KC, Tchetgen EJ, Cornelis MC, Kawachi I, Rimm E, and Kubzansky LD (2014). Genome-wide polygenic scoring for a 14-year long-term average depression phenotype. *Brain and behavior*, 4(2), 298–311

Chatterjee N, Shi J, and García-Closas M (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*

Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, and Park JHH (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genetics*, 45(4), 400–5, 405e1–3

de Maturana EL, Chanok SJ, Picornell AC, Rothman N, Herranz J, Calle ML, García-Closas M, Marenne G, Brand A, Tardón A *et al.* (2014). Whole genome prediction of bladder cancer risk with the Bayesian LASSO. *Genetic epidemiology*, 38(5), 467–76

Diogo D, Tian C, Franklin C, Alanne-Kinnunen M, March M, Spencer C, Vangjeli C, Weale M, Mattsson H, Kilpelainen E *et al.* (2017). Phenome-wide association studies (PheWAS) across large "real-world data" population cohorts support drug target validation. *bioRxiv*, 1–37

Domingue BW, Belsky DW, Harris KM, Smolen A, McQueen MB, and Boardman JD (2014). Polygenic risk predicts obesity in both white and black young adults. *PloS one*, 9(7), e101596

Euesden J, Lewis CM, and O'Reilly PF (2015). PRSice: Polygenic Risk Score software. *Bioinformatics*, 31(9), 1466–1468

Hagenaars SP, Harris SE, Davies G, Hill WD, Liewald DC, Ritchie SJ, Marioni RE, Fawns-Ritchie C, Cullen B, Malik R *et al.* (2016). Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112 151) and 24 GWAS consortia. *Molecular Psychiatry*, 21(11), 1624–1632

Krapohl E, Euesden J, Zabaneh D, Pingault JB, Rimfeld K, von Stumm S, Dale PS, Breen G, O'Reilly PF, and Plomin R (2016). Phenome-wide analysis of genome-wide polygenic scores. *Molecular Psychiatry*, 21(9), 1188–1193

Krapohl E, Patel H, Newhouse S, Curtis CJ, von Stumm S, Dale PS, Zabaneh D, Breen G, O'Reilly PF, and Plomin R (2017). Multi-polygenic score approach to trait prediction. *Molecular Psychiatry*, (May), 1–7

Lenfant C (2013). Prospects of personalized medicine in cardiovascular diseases. *Metabolism: clinical and experimental*, 62 Suppl 1, S6–10

Liu JZ, Erlich Y, and Pickrell JK (2017). Case-control association mapping by proxy using family history of disease. *Nature Genetics*, 49(3), 325–331

Machiela MJ, Chen CY, Chen C, Chanock SJ, Hunter DJ, and Kraft P (2011). Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genetic Epidemiology*, 35(6), 506–514

Mak TSH, Porsch RM, Choi SW, Zhou X, and Sham PC (2017). Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, (February), 1–12

Marquez-Luna C, Consortium TSTD, and Price AL (2016). Multi-ethnic polygenic risk scores improve risk prediction in diverse populations. *bioRxiv*, 051458

Nielsen JB, Thorolfsdottir RB, Fritsche LG, Zhou W, Skov MW, Graham SE, Herron TJ, McCarthy S, Schmidt EM, Sveinbjornsson G *et al.* (2018). Genome-wide association study of 1 million people identifies 111 loci for atrial fibrillation. *bioRxiv*

Opherk C, Gonik M, Duering M, Malik R, Jouvent E, Hervé D, Adib-Samii P, Bevan S, Pianese L, Silvestri S *et al.* (2014). Genome-wide genotyping demonstrates a polygenic risk score associated with white matter hyperintensity volume in CADASIL. *Stroke; a journal of cerebral circulation*, 45(4), 968–72

Pearl J (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York

Plomin R and von Stumm S (2018). The new genetics of intelligence. *Nature Reviews Genetics*

Power RA, Steinberg S, Bjornsdottir G, Rietveld CA, Abdellaoui A, Nivard MM, Johannesson M, Galesloot TE, Hottenga JJ, Willemsen G *et al.* (2015). Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nature Neuroscience*, 18(7), 953–955

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick Na, and Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8), 904–9

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, and Sklar P (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256), 748–52

Ripke S, Neale BM, Corvin A, Walters JTR, Farh KH, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H *et al.* (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511, 421–427

Ruderfer DM, Fanous AH, Ripke S, McQuillin A, Amdur RL, Gejman PV, O'Donovan MC, Andreassen OA, Djurovic S, Hultman CM *et al.* (2013). Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Molecular Psychiatry*, 19(9), 1017–1024

Socrates A, Bond T, Karhunen V, Auvinen J, Rietveld C, Veijola J, Jarvelin MR, and O&#039;Reilly P (2017). Polygenic risk scores applied to a single cohort reveal pleiotropy among hundreds of human phenotypes. *bioRxiv*

Stahl E, Forstner A, McQuillin A, Ripke S, Ophoff R, Scott L, Cichon S, Andreassen OA, Sklar P, Kelsoe J *et al.* (2017). Genomewide association study identifies 30 loci associated with bipolar disorder. *bioRxiv*

Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, Kraft P, Chen R, Kallberg HJ, Kurreeman FAS *et al.* (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature genetics*, 44(5), 483–9

Tesli M, Espeseth T, Bettella F, Mattingsdal M, Aas M, Melle I, Djurovic S, and Andreassen OA (2014). Polygenic risk score and the psychosis continuum model. *Acta Psychiatrica Scandinavica*, 130(4), 311–317

Tremblay J and Hamet P (2013). Role of genomics on the path to personalized medicine. *Metabolism: clinical and experimental*, 62 Suppl 1, S2–5

Varma S and Simon R (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1), 91

Wray NR, Lee SH, Mehta D, Vinkhuyzen AA, Dudbridge F, and Middeldorp CM (2014). Research Review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, 55(10), 1068–1087

Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, and Visscher PM (2013). Pitfalls of predicting complex traits from SNPs. *Nature reviews Genetics*, 14(7), 507–15