Received: 14 June 2016 | Revised: 20 February 2017 | Accepted: 14 March 2017

DOI: 10.1002/gepi.22050

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

Polygenic scores via penalized regression on summary statistics

Timothy Shin Heng Mak[1] | **Robert Milan Porsch**[2] | **Shing Wan Choi**[2] | **Xueya Zhou**[2] | **Pak Chung Sham**[1,2,3]

[1]Centre for Genomic Sciences, University of Hong Kong, Hong Kong

[2]Department of Psychiatry, University of Hong Kong, Hong Kong

[3]State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Hong Kong

Correspondence
Timothy Shin Heng Mak, Centre for Genomic Sciences, The University of Hong Kong, 1/F, 5 Sassoon Road, Hong Kong.
Email: tshmak@hku.hk

Pak Chung Sham, Centre for Genomic Sciences, The University of Hong Kong, 6/F, 5 Sassoon Road, Hong Kong.
Email: pcsham@hku.hk

Grant sponsor: Hong Kong Research Grants Council General Research Fund; Grant numbers: 776513M, HKU 776412M, 17128515; Grant sponsor: Hong Kong Research Grants Council Theme-Based Research Scheme; Grant numbers: T12-705/11, T12/708/12N, T12C-714/14-R; Grant sponsor: National Science Foundation of China – Research Grants Council of Hong Kong; Grant number: N_HKU736/14; Grant sponsor: European Network of National Schizophrenia Networks Studying Gene-Environment Interactions (EU-GEI).

ABSTRACT

Polygenic scores (PGS) summarize the genetic contribution of a person's genotype to a disease or phenotype. They can be used to group participants into different risk categories for diseases, and are also used as covariates in epidemiological analyses. A number of possible ways of calculating PGS have been proposed, and recently there is much interest in methods that incorporate information available in published summary statistics. As there is no inherent information on linkage disequilibrium (LD) in summary statistics, a pertinent question is how we can use LD information available elsewhere to supplement such analyses. To answer this question, we propose a method for constructing PGS using summary statistics and a reference panel in a penalized regression framework, which we call `lassosum`. We also propose a general method for choosing the value of the tuning parameter in the absence of validation data. In our simulations, we showed that pseudovalidation often resulted in prediction accuracy that is comparable to using a dataset with validation phenotype and was clearly superior to the conservative option of setting the tuning parameter of `lassosum` to its lowest value. We also showed that lassosum achieved better prediction accuracy than simple clumping and *P*-value thresholding in almost all scenarios. It was also substantially faster and more accurate than the recently proposed LDpred.

KEYWORDS

polygenic score, LASSO, elastic net, linkage disequilibrium, summary statistics

1 | INTRODUCTION

A vast number of twin studies as well as recent genome-wide association studies have demonstrated that a large proportion of the variance in liability to common diseases and human traits is due to genetic differences between individuals (Bulik-Sullivan et al., 2015; Polderman et al., 2015; Yang et al., 2011). These studies have also made clear that only a very small proportion of the total genetic contribution can be unambiguously attributed to variation in particular loci of the genome. The vast majority of such genetic contribution is thus spread across the huge landscape of the genome, with many loci each contributing a small, almost undetectable effect on the phenotypes (Dudbridge, 2013, 2016). One important source of evidence toward this conclusion is from studies that examined the association of polygenic predictors of diseases/traits, where it has been repeatedly found that Single Nucleotide Polymorphism (SNPs) that are not themselves significantly associated with the phenotypes can, by being aggregated as a score, be very significantly associated with the phenotypes in different samples (Agerbo et al., 2015; Byrne et al., 2014; Chang et al., 2014; Evans, Visscher, & Wray, 2009; Machiela et al., 2011; Martin, O'Donovan, Thapar, Langley, & Williams, 2015; Purcell et al., 2009; Ripke et al., 2013; Speliotes et al., 2010; Stahl et al., 2012; Wei et al., 2009). A particular remarkable demonstration is that persons with such *polygenic scores* (PGS) for schizophrenia at the top 10 percentile of the population can be at more than 10 times the risk of having the disease than those at the bottom 10 percentile (Agerbo et al., 2015; Ripke et al., 2014) raising hope that one

day a person's risk for many common disease can be accurately assessed simply by the examination of one's genome.

Thus, there is considerable interest in the calculation of such PGS in GWAS and genome-wide meta-analyses, where they are also known as risk scores (Domingue et al., 2014; Ripke et al., 2013), polygenic risk scores (e.g., Agerbo et al., 2015; Byrne et al., 2014; Dudbridge, 2013; Euesden, Lewis, & O'Reilly, 2015), and allelic scores (Burgess & Thompson, 2013; Evans et al., 2013). In a typical application, a unique PGS is assigned to each individual based on the person's genotype. The score summarizes the genetic contribution to a particular disease or phenotype for that individual given his/her genotype. They are then used for testing of complex genetic contribution due to multiple loci or even the entire genome, or the examination of genetic correlation, or are used as a covariate for the adjustment of genetic effects in a multiple regression model (Wray et al., 2014).

From a statistical perspective, PGS are weighted sums of the genotypes of a set of SNPs. In most applications of PGS, the weights are usually the SNPs' individual regression coefficients with the phenotype (e.g., Euesden et al., 2015; Purcell et al., 2009; Wray et al., 2014). A critical issue is the total number of SNPs that should be included in the PGS. Although it is usually advisable to use a liberal P-value cutoff in the selection of SNPs to be included, the optimal P-value cutoff is generally unknown (Wray et al., 2014). As a result, in many studies, PGS are constructed using a number of thresholds (Byrne et al., 2014; Chang et al., 2014; Martin et al., 2015; Purcell et al., 2009; Ripke et al., 2014), and there is at least one piece of software developed to facilitate this (Euesden et al., 2015). Generally, we focus on the P-value threshold that achieves the highest correlation/association with the phenotypes in a validation dataset that contains a measure of the phenotype under study. This approach, however, becomes less useful if the phenotype is not available in the target dataset. Recently, Mak, Kwan, Campbell, and Sham (2016) sought to overcome this problem by downweighting the usual weights by the SNPs' local true discovery rate, where the additional downweighting or shrinkage factor can be estimated using a data-driven approach. Although P-value thresholds were not needed, they showed that this leads to comparable predictive performance with the best P-value threshold.

Another issue with this standard approach to PGS calculation is that there is no account taken of the fact that SNPs are in linkage disequilibrium (LD) with each other. If SNPs of a particular locus that are in high LD with one another are all included in the score, the contribution to the PGS due to that locus will be exaggerated in the score. For this reason, it is often recommended that SNPs be *pruned* before the application of PG scoring, such that highly correlated SNPs within a locus will have one or more removed (Purcell et al., 2009). Such an approach, however, may well reduce the predictive power of the PGS, as SNPs that are most predictive of the phenotype may be pruned away. A recent suggestion that has become very popular is that of clumping, which selectively removes less significantly related SNPs to reduce LD (Wray et al., 2014).

In principle, various machine learning methods or Bayesian methods can be applied in the construction of PGS, as they have been applied in the estimation of breeding values in animal studies (Abraham, Kowalczyk, Zobel, & Inouye, 2013; Erbe et al., 2012; Habier, Fernando, Kizilkaya, & Garrick, 2011; Meuwissen, Hayes, & Goddard, 2001; Ogutu, Schulz-Streeck, & Piepho, 2012; Pirinen, Donnelly, & Spencer, 2013; Szymczak et al., 2009; Zhou, Carbonetto, & Stephens, 2013). These methods do not require the assumption of SNP independence or near independence, and have been shown to perform better than simple PGS in simulation settings. However, their disadvantage is that they cannot be applied to summary statistics. Researchers without access to large datasets are thus unable to take advantage of the power offered by these studies or meta-analyses. A recent development in this direction is Vilhjálmsson et al. (2015). The authors proposed an approximate Bayesian method known as LDpred that calculates PGS based on summary statistics, using LD information from a reference panel. Such a development is particularly welcome due to the ready availability of summary statistics from many consortia, often calculated from tens to hundreds of thousands of individuals.

In this paper, we present an alternative method based on penalized regression. It is a deterministic method and a convex optimization problem, and as such does not suffer from problems of nonconvergence, which is a possible problem with LDpred. It is also substantially faster than LDpred, and in our simulations achieved near-best prediction performance across a wide variety of scenarios. As a side observation, it was also found that LDpred did not achieve the improved prediction performance claimed by the authors in our simulations. As with any machine learning approach, our proposed method requires the choice of a tuning parameter. This is particularly difficult when we do not have raw data and hence cannot perform cross-validation. Here, we offer a solution that can potentially be applied more generally. The approach is presented in the methods section and we assessed its performance by simulation studies. Insights gained from the simulations are discussed.

2 | MATERIALS AND METHODS

2.1 | The LASSO problem in terms of summary statistics

Given a linear regression problem

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

where $\boldsymbol{X}$ denotes an n-by-p data matrix, and $\boldsymbol{y}$ a vector of observed outcomes, the LASSO (Tibshirani, 1996) is a popular method for deriving estimates of $\boldsymbol{\beta}$ and predictors of (future observations of) $\boldsymbol{y}$, especially in the case where p (the number of predictors/columns in $\boldsymbol{X}$) is large and when it is reasonable to assume that many $\boldsymbol{\beta}$ are 0. LASSO obtains estimates of $\boldsymbol{\beta}$ (weights in the linear combination of $\boldsymbol{X}$) given $\boldsymbol{y}$ and $\boldsymbol{X}$ by minimizing the objective function

$$f(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\lambda||\boldsymbol{\beta}||_1^1 \quad (2)$$

$$= \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + 2\lambda||\boldsymbol{\beta}||_1^1, \quad (3)$$

where $||\boldsymbol{\beta}||_1^1 = \sum_i |\beta_i|$ denote the $\mathcal{L}_1$ norm of $\boldsymbol{\beta}$, for a particular fixed value of λ. In general, depending on λ, a proportion of the β_i are given the estimate of 0. It is also a specific instance of *penalized regression* where the usual least square formulation of the linear regression problem is augmented by a penalty, in this case $2\lambda||\boldsymbol{\beta}||_1^1$. LASSO lends itself to being used for estimation of $\boldsymbol{\beta}$ in the event where only summary statistics are available, because if $\boldsymbol{X}$ represent standardized genotype data and $\boldsymbol{y}$ standardized phenotype, divided by $\sqrt{n}$, then Equation (3) can be written as

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{R}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{r} + 2\lambda||\boldsymbol{\beta}||_1^1, \quad (4)$$

where $\boldsymbol{r} = \boldsymbol{X}^T\boldsymbol{y}$ represents the SNP-wise correlation between the SNPs and the phenotype, and $\boldsymbol{R} = \boldsymbol{X}^T\boldsymbol{X}$ is the LD matrix, a matrix of correlations between SNPs. As we can obtain estimates of $\boldsymbol{r}$ from summary statistics databases that are publicly available for major diseases/phenotypes (see, e.g., the list from Pasaniuc & Price, 2016) and LD hub (`http://ldsc.broadinstitute.org/`), and estimates of LD ($\boldsymbol{R}$) from publicly available genotype such as the 1000 Genome database (1000 Genomes Project Consortium, 2015), Equation (4) suggests a method for deriving PGS weights as estimates of $\boldsymbol{\beta}$ by minimizing $f(\boldsymbol{\beta})$.

An issue that surfaces when we substitute $\boldsymbol{R}$ and $\boldsymbol{r}$ with the estimates derived from publicly available data is that the genotype $\boldsymbol{X}$ used to estimate $\boldsymbol{R}$ and $\boldsymbol{r}$ will in general be different. In particular, it will be more appropriate to write $\boldsymbol{R} = \boldsymbol{X}_r^T\boldsymbol{X}_r$ to indicate that the genotype used to derive estimates of LD ($\boldsymbol{X}_r$) will not in general be the same as the genotype that gave rise to the correlations $\boldsymbol{r}$. Writing Equation (4) as

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}_r^T\boldsymbol{X}_r\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + 2\lambda||\boldsymbol{\beta}||_1^1, \quad (5)$$

however, would imply that (5) is no longer a LASSO problem, because it is no longer a penalized least squares problem. A minimum to (5) can still be sought, although the solutions would often be unstable and nonunique, since $\boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}_r^T\boldsymbol{X}_r\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}$ will not generally have a finite minimum.

A natural solution to this problem is to *regularize* Equation (5). In particular, if we replace $\boldsymbol{X}_r^T\boldsymbol{X}_r$ with $\boldsymbol{R}_s = (1-s)\boldsymbol{X}_r^T\boldsymbol{X}_r + s\boldsymbol{I}$, for some $0 < s < 1$, then

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{R}_s\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{r} + 2\lambda||\boldsymbol{\beta}||_1^1, \quad (6)$$

will be equivalent to a LASSO problem.

Proof. *First, we note that* $\boldsymbol{y}^T\boldsymbol{y}$ *is a constant and thus replacing it with any other constant will not change the solution.* $\boldsymbol{R}_s$ *is necessarily positive definite for* $0 < s < 1$. *This means that there always exists* $\boldsymbol{W}$ *and* $\boldsymbol{v}$ *such that*

$$\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{R}_s, \quad \boldsymbol{W}^T\boldsymbol{v} = \boldsymbol{r}. \quad (7)$$

Substituting (7) *into* (6) *and replacing* $\boldsymbol{y}^T\boldsymbol{y}$ *with* $\boldsymbol{v}^T\boldsymbol{v}$, *we see that* (6) *can be written in a form such as* (2) *and is therefore a LASSO problem.* □

Expanding (6) into

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + (1-s)\boldsymbol{\beta}^T\boldsymbol{X}_r^T\boldsymbol{X}_r\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{r} + s\boldsymbol{\beta}^T\boldsymbol{\beta} + 2\lambda||\boldsymbol{\beta}||_1^1, \quad (8)$$

we note that (8) encompasses a number of submodels as special cases. For example, when $s = 1$, estimates of $\boldsymbol{\beta}$ will be equivalent to applying a "soft" threshold to the univariate correlation summary statistics $\boldsymbol{r}$ (as opposed to the "hard" thresholds using P-values.) In particular,

$$\hat{\beta}_i^{s=1} = \begin{cases} \text{sign}(r_i)(|r_i| - \lambda) & \text{if } |r_i| - \lambda > 0 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

(Zou & Hastie, 2005). Note that because there is a monotonic relationship between univariate P-values and unsigned correlation coefficients (coming from the monotonic relationship between correlation coefficients and t-statistics with $n-2$ degrees of freedom, Equation (15)), soft-thresholding using correlation coefficients can be expected to be very similar to P-value thresholding. Another feature is that when $\lambda = 0$, the problem is similar to applying ridge regression to estimate $\boldsymbol{\beta}$, except for a constant scaling value. In most cases, the scale of a PGS is irrelevant, since it is almost never directly used in genomic risk prediction without appropriate scaling (e.g., in So, Kwan, Cherny, & Sham, 2011). For a particular choice of s, therefore, Equation (8) results in genomic BLUP (best linear unbiased predictors; de Los Campos, Vazquez, Fernando, Klimentidis, & Sorensen, 2013). When $\lambda = 0$ and $s = 1$, the estimated PGS becomes equivalent to simply using the entire set of correlation estimates without shrinkage or subset selection.

Moreover, (8) is simply an elastic net problem (Zou & Hastie, 2005), and thus can be solved using fast coordinate descent algorithms (Friedman, Hastie, & Tibshirani, 2010) for many values of λ at a time. In particular, using this algorithm,

it is not necessary to compute the p-by-p matrix $\boldsymbol{X}_r^T\boldsymbol{X}_r$, which would be extremely memory-consuming even for tens of thousands of SNPs. Denoting $\tilde{\boldsymbol{X}} = \sqrt{1-s}\boldsymbol{X}_r$, the solution to the minimization of Equation (8) can be obtaining by iteratively updating β_i as

$$\beta_i^{(t)} = \begin{cases} \text{sign}(u_i^{(t)})|u_i^{(t)} - \lambda|/(\tilde{\boldsymbol{X}}_i^T\tilde{\boldsymbol{X}}_i + s) & \text{if } |u_i^{(t)}| - \lambda > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

$$u_i^{(t)} = r_i - \tilde{\boldsymbol{X}}_i^T(\tilde{\boldsymbol{X}}\boldsymbol{\beta}^{(t-1)} - \tilde{\boldsymbol{X}}_i\beta_i^{(t-1)}) \tag{11}$$

A more detailed proof of Equations (10) and (11) is given in the supplementary materials. An R package that carries out the estimation of $\boldsymbol{\beta}$ is made available at `https://github.com/tshmak/lassosum`. We made special effort to allow estimation to be done directly on PLINK 1.9 (Chang et al., 2015) `.bed` files, eliminating the need to load large genotype matrices into R.

2.2 | Selection of tuning parameters

As with standard elastic net problems, in any application, λ and s need to be chosen. Generally, in the presence of a validation dataset, we can choose λ by maximizing the correlation of the PGS with the validation phenotype data, just as it has been done in the choice of a P-value cutoff points in standard PGS calculations (Euesden et al., 2015; Wray et al., 2014). In principle, we can use this method to choose a suitable value for s also, although repeating the estimation over different values of s is much more time-consuming. Thus in this paper, we set s to a few chosen values and examined whether they are sufficient in arriving at a PGS with reasonable prediction accuracy.

A more pressing problem is that validation phenotypes are not often available. And here we try to simulate this procedure in the following manner, which we refer to as *pseudovalidation* in this paper. First, note that the correlation between a $PGS(\lambda) \equiv \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_\lambda$ and the phenotype $\tilde{\boldsymbol{y}}$ in a new "test" dataset with standardized genotype $\tilde{\boldsymbol{X}}$ is

$$Corr(PGS(\lambda), \tilde{\boldsymbol{y}}) = \frac{\boldsymbol{\beta}_\lambda^T\tilde{\boldsymbol{X}}^T\boldsymbol{P}\tilde{\boldsymbol{y}}}{\sqrt{\boldsymbol{\beta}_\lambda^T\tilde{\boldsymbol{X}}^T\boldsymbol{P}\tilde{\boldsymbol{X}}\boldsymbol{\beta}_\lambda\tilde{\boldsymbol{y}}^T\boldsymbol{P}\tilde{\boldsymbol{y}}}}, \tag{12}$$

where $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n$ is the mean-centering matrix.

In the absence of validation data, $\tilde{\boldsymbol{y}}$ is unavailable. Our solution is to substitute $\hat{\boldsymbol{r}}$ for $\tilde{\boldsymbol{X}}^T\boldsymbol{P}\tilde{\boldsymbol{y}}$, where $\hat{\boldsymbol{r}}$ is a shrunken estimate of the $\boldsymbol{r}$, the observed correlation coefficient vector. Since $\tilde{\boldsymbol{X}}^T\boldsymbol{P}\tilde{\boldsymbol{y}}$ can be interpreted as a correlation coefficient only if $\tilde{\boldsymbol{X}}$ is a standardized genotype matrix and $\tilde{\boldsymbol{y}}$ standardized phenotype, we replace $\tilde{\boldsymbol{X}}$ with its standardized version, $\tilde{\boldsymbol{X}}_0$, and discard the constant $\tilde{\boldsymbol{y}}^T\boldsymbol{P}\tilde{\boldsymbol{y}}$ term, so as to maximize the function

$$f(\lambda) = \frac{\boldsymbol{\beta}_\lambda^T\hat{\boldsymbol{r}}}{\sqrt{\boldsymbol{\beta}_\lambda^T\tilde{\boldsymbol{X}}_0^T\tilde{\boldsymbol{X}}_0\boldsymbol{\beta}_\lambda}} \tag{13}$$

over λ. Here, following Mak et al. (2016), we calculated

$$\hat{r}_i = r_i(1 - \text{fdr}_i), \tag{14}$$

where fdr_i is the local false discovery rate of SNP i. Although Mak et al. (2016) estimated fdr_i using maximum likelihood and a nonparametric kernel density estimator, we found that Strimmer (2008) provided a fast, nonparametric estimator for fdr_i that is constrained to be monotonic decreasing with $|r_i|$, and it is this approach that we have implemented in the simulations.

2.3 | Some notes on application

In the above, we have assumed that the SNP-wise correlations ($\boldsymbol{r}$) will be available from the summary statistics. When these are not available, we suggest *pseudocorrelation* estimates $\tilde{r}_i$ be derived by converting P-values to correlation, using the monotonic relationship between t-statistics and correlations:

$$\tilde{r}_i = \frac{t_i}{\sqrt{n - 1 + t_i^2}}. \tag{15}$$

In our simulations, this resulted in almost identical estimates as using actual (Pearson's product moment) correlations (supplementary Fig. S1).

Another issue is that in the theory given above, we assume that $\boldsymbol{X}$ and $\boldsymbol{y}$ have been standardized such that $\boldsymbol{r}$ represent the correlation coefficients between the genotype and the phenotype. We note that such standardization can be justified by the fact that the LASSO is often performed on standardized variables (Hastie, Tibshirani, & Friedman, 2009; Li, Gui, Kwan, Bao, & Sham, 2012; Yi, Breheny, Imam, Liu, & Hoeschele, 2014). However, when it comes to the construction of PGS, we ought to use unstandardized coefficients as weights. To convert standardized coefficients to unstandardized ones, we can simply use the formula

$$\beta_i^{\text{unstandardized}} = r_i\frac{\text{sd}(\boldsymbol{y})}{\text{sd}(\boldsymbol{X}_i)}. \tag{16}$$

However, since $\text{sd}(\boldsymbol{y})$ and $\text{sd}(\boldsymbol{X}_i)$ are generally unavailable, we can use $\text{sd}(\tilde{\boldsymbol{y}})$ and $\text{sd}(\tilde{\boldsymbol{X}}_i)$ from the validation data instead. Using these also prevents any SNP from undue influence in the overall PGS due to the division of $\text{sd}(\tilde{\boldsymbol{X}}_i)$ close to 0, since a SNP's variance contribution is proportional to its variance and the square of the coefficients.

The third issue concerns the difference between the SNPs with summary statistics and the SNPs that are included in the reference panel. Often the reference panel may not contain all SNPs with summary statistics. Equivalently, there may be no variation within the panel for some SNPs. In LDpred, these SNPs are discarded by default. However, we think that this is not necessary, as it may result in the removal of SNPs that are predictive of the disease/phenotype. An intuitive approach to dealing with these SNPs is that we treat them as if they were all mutually independent and apply soft-thresholding as in (9). Equivalently, we let $\boldsymbol{X}_{ri}$ for these SNPs to be a vector of 0, and we augment Equation (8) by a term $(1-s)\boldsymbol{\beta}_0^T\boldsymbol{\beta}_0$,

$$f(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + (1-s)\boldsymbol{\beta}^T\boldsymbol{X}_r^T\boldsymbol{X}_r\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{r} + s\boldsymbol{\beta}^T\boldsymbol{\beta} + (1-s)\boldsymbol{\beta}_0^T\boldsymbol{\beta}_0 + 2\lambda||\boldsymbol{\beta}||_1^1, \quad (17)$$

where $\boldsymbol{\beta}_0$ denotes the subvector of $\boldsymbol{\beta}$ whose $\text{sd}(\boldsymbol{X}_i) = 0$, such that the total ridge penalty for these parameters is 1.

The fourth issue concerns the application of pseudovalidation to clumped data. We proposed above that $\hat{\boldsymbol{r}}$ be estimated using (14) and that the local false discovery rates be estimated using the procedure of Strimmer (2008). An important point is that the method assumes that a sizeable proportion of the $\boldsymbol{r}$ are in fact null. Under clumping, this may not necessarily be the case, and we therefore suggest estimating fdr_i and hence $\hat{r}_i$ *before* applying clumping.

2.4 | Simulation studies

We performed a number of simulation studies to assess the performance of our proposed method, which we refer to in this paper as `lassosum`. In our first simulation study, we used the Wellcome Trust Case Control Consortium (WTCCC) Phase 1 data for seven diseases. We filtered variants and participants using the following QC criteria: genotype rate >0.99, minor allele frequency >0.01, missing genotype per individual <0.01, SNP rsID included in the 1000 Genome project (Phase 3, release May 2013) genotype data, with matching reference and alternative alleles, on top of the QC done by the original researchers (Wellcome Trust Case Control Consortium, 2007). This resulted in 358,179 SNPs and 15,603 individuals, of which 2,859 were controls. In our first set of simulations, we ignored the phenotype data and generated our own based on the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (18)$$

where $\boldsymbol{X}$ is the unstandardized genotype matrix, and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ represents random error. The distribution of the causal effects $\boldsymbol{\beta} \equiv vec(\{\beta_i\}) \equiv vec(\{\beta_{jk}\})$ is generated using a similar scheme to that described in Vilhjálmsson et al. (2015):

$$\beta_{jk} \sim \begin{cases} N(0,1) & \text{with probability } \pi_j \\ 0 & \text{with probability } 1-\pi_j, \end{cases} \quad (19)$$

$$\pi_j \sim Beta(P(\text{causal}), 1 - P(\text{causal})) \quad (20)$$

where j denotes genomic regions and k indices SNPs within the region and i is a general index for all SNPs in the database, and p is the expected proportion of causal SNPs across the genome (note $E(\pi_j) = P(\text{causal})$). Genomic regions were defined using the 1,725 LD blocks obtained from the 1000 Genomes European (EUR) subpopulation, as provided by Berisa and Pickrell (2015).

We derived standardized $\boldsymbol{\beta}$ as

$$\beta_i^0 = \beta_i \frac{\hat{\text{sd}}(\boldsymbol{X}_i)}{\hat{\text{sd}}(\boldsymbol{y})}, \quad (21)$$

and observed correlation coefficients as

$$\boldsymbol{r}_j \sim N(\hat{\boldsymbol{R}}_j\boldsymbol{\beta}_j^0, \hat{\boldsymbol{R}}_j/n), \quad (22)$$

where $\hat{\boldsymbol{R}}$ is the observed correlation matrix of the jth region from the genotype $\boldsymbol{X}$ and n is the sample size. We set $\sigma^2 = \hat{\text{Var}}(\boldsymbol{X}\boldsymbol{\beta})\frac{1-h^2}{h^2}$ and $h^2 = 0.5$ in our calculation of $\boldsymbol{y}$.

We randomly chose two 1,000 samples as two test datasets. In the first dataset $\boldsymbol{X}^{(1)}$, validation and pseudovalidation were performed to determine the optimal value of λ. This choice of λ and/or s was applied in the other test dataset $\boldsymbol{X}^{(2)}$ in the assessment of prediction accuracy. Prediction accuracy was assessed by the correlation of the PGS with the true predictor $\boldsymbol{X}^{(2)}\boldsymbol{\beta}$. Except when assessing the performance of using different reference panels, we used the first test dataset $\boldsymbol{X}^{(1)}$ as the reference panel also.

In assessing the impact of using different reference panels, we let the 1000 Genome East Asian (EAS) subpopulation ($n = 503$) be our test dataset. We compared the performance of using four different reference panels: (1) the original sample that generated the summary statistics, (2) a sample of 1,000 from the WTCCC, (3) the EUR subpopulation from the 1000 Genome project, and (4) the EAS subpopulation from the 1000 Genome project.

The above simulations were repeated 10 times and were compared with the approach of P-value thresholding (with and without clumping) and LDpred. For clumping, we used a window of 250 kb and an R^2 of $\{0.1, 0.2, 0.5, 0.8\}$. (see supplementary Note for a brief explanation of clumping.) For P-value thresholding, we used the set of P values $\{5e^{-8}, 1e^{-5}, 1e^{-4}, 1e^{-3}, 0.0015, 0.002, 0.0025, \ldots, 0.995, 1\}$ as possible P-value thresholds. For LDpred, we used the set of proportion of causal SNPs $\{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1\}$. The size of the window

for LD calculation was calculated as the number of SNPs in the dataset divided by 3,000, as recommended in the LDpred paper. For P-value thresholding and LDpred, we used a validation dataset as well as pseudovalidation to select the best threshold and proportion of causal SNPs, respectively.

Because summary statistics are often calculated from large sample sizes and for a large number (often around 10 million) of SNPs, we also attempted to carry out simulations using a larger dataset. In particular, we wanted to see whether clumping is an efficient strategy for data reduction, as the speed of `lassosum` suffers with such a large number of SNPs. For this purpose, we first identified SNPs from the summary statistics derived in the meta-analysis of Okada et al. (2014) for rheumatoid arthritis (RA) that were common with those in the 1000 Genome dataset. We then generated our own summary statistics using the above method (Equations (18)–(22)), using the EUR subsample of the 1000 Genome dataset as a base. This resulted in a dataset of 8,270,298 SNPs. We used the EUR subsample as the reference panel and the EAS subsample of the 1000 Genome dataset as the test sample to assess the predictive performance.

Finally, we assessed the performance of `lassosum` using real summary statistics from large meta-analyses. Summary statistics were downloaded from five publicly available resources: Bipolar disorder (`https://www.med.unc.edu/pgc`, Sklar et al. (2011), $n(\text{cases}) = 7,481, n(\text{controls}) = 9,250$), coronary artery disease (`http://www.cardiogramplusc4d.org`, Nikpay et al. (2015), $n(\text{cases}) = 60,801, n(\text{controls}) = 123,504$), Crohn's disease (`http://ibdgenetics.org/downloads.html`, Liu et al. (2015), $n(\text{cases}) = 22,575, n(\text{controls}) = 46,693$), RA (`http://plaza.umin.ac.jp/~yokada/datasource/software.htm/`, Okada et al. (2014), $n(\text{cases}) = 14,361, n(\text{controls}) = 43,923$), and Type 2 diabetes (`http://diagram-consortium.org/`, Mahajan et al. (2014), $n(\text{cases}) = 26,488, n(\text{controls}) = 83,964$). The performance of PGS derived using `lassosum` and other methods were assessed using the WTCCC data. Because all these meta-analyses included the WTCCC as one of the studies, PGS derived using these summary statistics directly would overfit the data. To overcome this problem, we attempted to isolate the non-WTCCC components of the summary statistics by reversing the fixed-effects meta-analysis equations:

$$\beta_{\text{meta}} = \frac{\beta_s/\sigma_s^2 + \beta_{\bar{s}}/\sigma_{\bar{s}}^2}{1/\sigma_s^2 + 1/\sigma_{\bar{s}}^2} \tag{23}$$

$$\frac{1}{\sigma_{\text{meta}}^2} = \frac{1}{\sigma_s^2} + \frac{1}{\sigma_{\bar{s}}^2}, \tag{24}$$

where β_s and σ_s denote the log odds ratio and standard error from the WTCCC study and $\beta_{\bar{s}}$ and $\sigma_{\bar{s}}$ the contribution to the meta-analysis apart from WTCCC. SNPs with negative $\sigma_{\bar{s}}^2$ were set to have zero effect size. P-values were derived from $2(1 - \Psi^{-1}(|\beta_{\bar{s}}/\sigma_{\bar{s}}|))$ and converted to correlations using (15). Prediction accuracy of the summary statistics-derived PGS were assessed by the area under the ROC curve (AUC) statistic when used to predict disease status in the WTCCC dataset with the relevant disease and the 2,859 controls. The testing sample was also used as the reference panel. In all the above analyses, we carried out estimation by LD blocks as defined by Berisa and Pickrell (2015).

3 | RESULTS

Our WTCCC simulations were performed with summary statistics sample sizes of 10,000, 50,000, and 250,000, respectively. We used two values for $P(\text{causal})$, the expected proportion of causal SNPs: 0.1 and 0.01. $P(\text{causal}) = 0.01$ represents a scenario where there are fewer causal SNPs and effect sizes are larger. Conversely $P(\text{causal}) = 0.1$ represents a scenario where causal SNPs have smaller effect sizes and are more spread out over the genome. Supplementary Figure S2 displays the performance of `lassosum` with different values of λ for one of the simulations. It can be seen that in all the simulation scenarios, the general pattern is that predictive performance increases with λ up to a point and then decreases, often rapidly. Using a validation dataset or alternatively pseudovalidation is usually effective in helping us select a value of λ that is close to the optimal. Comparing different values of s, the shrinkage parameter, we see that the maximum attainable correlation is generally lower for $s = 1$, the scenario where `lassosum` reduces to soft-thresholding, that is, where information on LD is ignored, except when $n = 10,000$ and $P(\text{causal}) = 0.1$. In addition, $s = 0.5$ and $s = 0.2$ usually gives better performance than $s = 0.9$.

In Figure 1A, we give the average prediction performance over 10 simulations, comparing the use of pseudovalidation and a validation dataset with phenotype data as well as using the minimum λ value of 0.001. We use $\lambda = 0.001$ for comparison because it is shown in supplementary Figure S2 that in general the prediction performance of `lassosum` approaches a constant as λ tends to 0, whereas when λ approaches 1, the performance drops sharply. Thus, using λ close to 0 represents a conservative, safe option, and as noted before $\lambda = 0$ is equivalent to ridge regression. When $s = 0.2$ or 0.5, the performance of pseudovalidation was very similar to using a real validation phenotype. Both approaches were clearly superior to the conservative option of setting $\lambda = 0.001$. When $s = 0.9$ or $s = 1$, pseudovalidation was still clearly superior to setting $\lambda = 0.001$ for $n = 10,000$ and $n = 50,000$ and $P(\text{causal}) = 0.01$. In all simulations, the performance of P-value thresholding was similar to the use of `lassosum` with $s = 1$. Thus, "soft-thresholding" and "hard-thresholding"

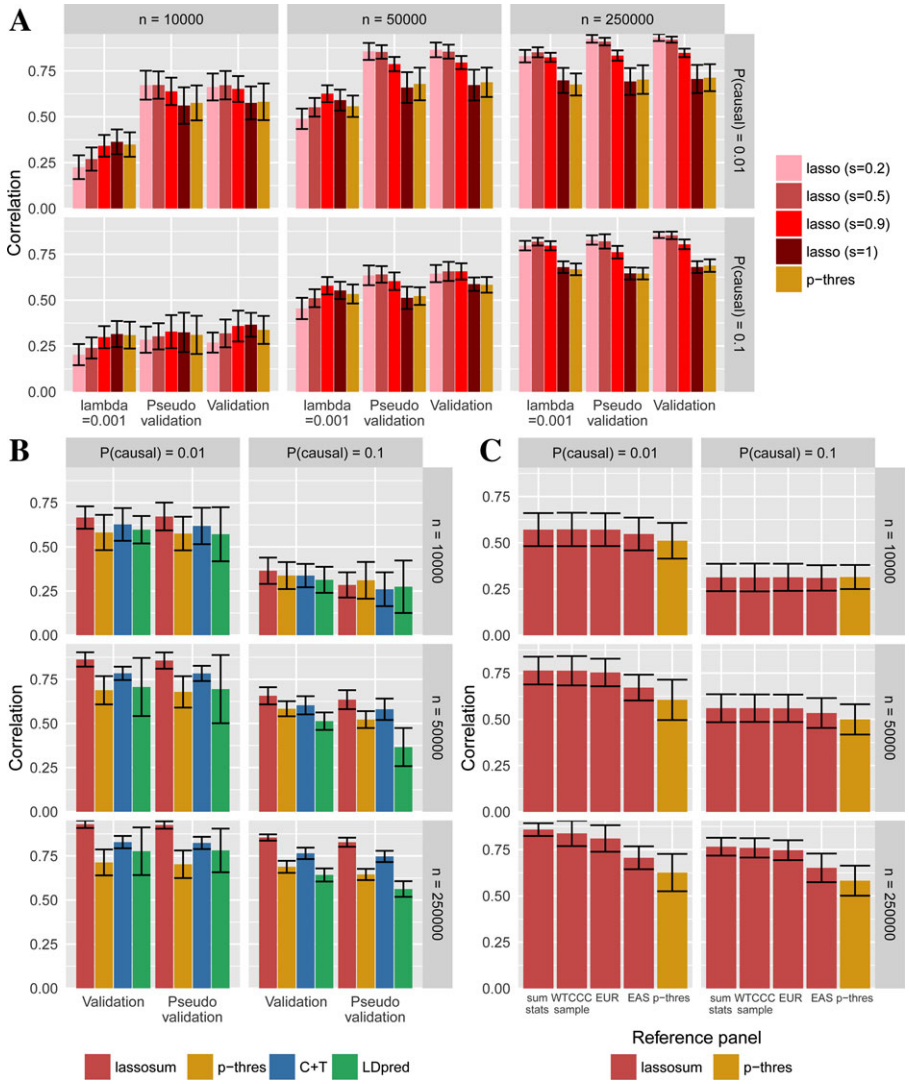

FIGURE 1 In all of the plots, mean and standard deviation of the correlation of the PGS with the true predictor are plotted. (A) Comparing the use of a validation dataset with phenotype data and pseudovalidation in selecting the tuning parameter λ. (B) Comparing the performance `lassosum`, P-value thresholding (p-thres), P-value thresholding with clumping (C + T), and LDpred. (C) The effect of using different reference panels on `lassosum`. sum stats: The same data from which the summary statistics were simulated, WTCCC sample: a sample of 1,000 from the WTCCC; EUR, European; EAS, East Asian reference panel from 1000G

appeared to give similar performance. We also observed that `lassosum` with $s = 0.2$ or $s = 0.5$ tended to give the best performance overall. In our implementation of `lassosum`, the computation time for $s = 0.2, 0.5$, and 0.9 were similar (supplementary Figs. S4 and S5). Thus, it is reasonable to maximize over s also using either a validation phenotype or pseudovalidation when using `lassosum`. In Figure 1B, we compare the performance of `lassosum` with clumping and P-value thresholding, as well as with LDpred. For `lassosum`, we optimized over both λ and $s = \{0.2, 0.5, 0.9, 1\}$. For comparison, we optimized over P-value thresholds and clumping $R^2 = \{0.1, 0.2, 0.5, 0.8, \text{no clumping}\}$. For LDpred, we optimized over $P(\text{causal}) = \{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1\}$. For P-value thresholding, clumping led to a noticeable increase in prediction accuracy, except when $P(\text{causal}) = 0.1$ and $n = 10,000$. However, in all scenarios, `lassosum` was superior to clumping and thresholding. The result was similar whether the method was optimized using a validation dataset or pseudovalidation. We found that LDpred did not appear to have the claimed advantage over P-value thresholding in our simulations. At first, we thought this might be because the size of the reference sample used was only 1,000, smaller than the recommended size of at least 2,000 in the paper. However, we found that the performance of LDpred did not improve even when the sample size of the reference panel (and test panels) were set to 5,000 (supplementary Fig. S3).

A possible criticism of our simulations so far is that we performed `lassosum` by LD blocks defined by Berisa and Pickrell (2015), while the summary statistics were also simulated by the same LD blocks. To address this issue, we repeated

the analysis using blocks with roughly the same number of SNPs spread uniformly across the genome. The number of blocks were made equal to the number of blocks given by Berisa and Pickrell (2015), but the boundaries were different. This would allow lassosum to adjust for LD within blocks, but not LD across blocks in the boundary regions. We also compared it to the scenario when lassosum was carried out by chromosomes. The results are presented in supplementary Figure S4. It can be seen that lassosum by LD blocks and uniform blocks had nearly identical predictive performance. Thus, the advantage that lassosum had in our simulations by sharing the same blocks by which the summary statistics were generated was negligible. The relative poor performance of lassosum when carried out by chromosomes is likely due to confounding by chance correlations between SNPs over long distances that are not in fact in LD.

In Figure 1C, we examined the effect of using different reference panels when using lassosum. We generated the summary statistics using the entire WTCCC sample, and used four different reference panels for our LD information: (1) the original WTCCC sample that generated the summary statistics, (2) a sample of 1,000 from the WTCCC, (3) the EUR subpopulation from the 1000 Genome project, and (4) the EAS subpopulation from the 1000 Genome project, which also served as the test sample. It was found that for the small sample size ($n = 10,000$) scenario the use of the different reference panels made relatively little difference to predictive performance. However, as sample size increased, using the true sample that generated the summary statistics led to noticeably improved predictive performance. For many scenarios, using the 1000 Genome EUR sample as the reference panel led to a similar performance as using the original summary statistic sample. A clear advantage for using the summary statistics sample was only shown in the scenario with the most power ($n = 250,000$ and $P(\text{causal}) = 0.01$). Using the wrong (EAS) reference sample was clearly inferior when the sample size was above 50,000, but it was still better than simple P-value thresholding.

Next, we examined the performance of lassosum in a larger simulated dataset with around 8 million SNPs, with a focus on clumping, to see whether prefiltering by clumping can be an effective method in reducing the number of SNPs in the analysis. The sample size for the summary statistics was set to 200,000. Six levels of clumping ($r^2 = 0.01$, 0.05, 0.1, 0.2, 0.5, and 0.8) were applied to the data, using a window size of 250 kb, resulting in around 190,000, 330,000, 430,000, 610,000, 1,170,000, and 1,940,000 SNPs respectively. (The actual number depends on the simulations.) We did not perform LDpred for $r^2 > 0.2$ because it was too time and memory intensive. In Figure 2A, we present the results from this simulation. Here, we see that clumping was beneficial in improving prediction performance for P-value thresholding, and the best performance was achieved with an r^2 of 0.5 or 0.8. For lassosum, performance decreased with increasing level of clumping (decreasing r^2). lassosum with no clumping gave the best performance overall. LDpred performed poorly in this simulation, likely because the reference panel size was too small.

In Figure 2B, we present the results for using real summary statistics from five large meta-analyses to predict phenotypes in the WTCCC data. In all cases, the use of pseudovalidation resulted in a PGS that is close to the maximum AUC across all tuning parameters, and was clearly superior to using $\lambda = 0.001$. For BD, CAD, CD, and RA, the performance of lassosum, LDpred, and clumping and thresholding were similar, although a slightly higher AUC was observed for lassosum. For T2D, the maximum AUC was surprisingly achieved by P-value thresholding without clumping.

In supplementary Figures S5 and S6, we plot the average time taken to run lassosum on our computer cluster, using 1 core for each analysis. In general, running times for different values of s were similar, although lower values of s led to slightly longer running times. However, running times increased exponentially both with the number of participants (supplementary Fig. S5) and the number of SNPs (supplementary Fig. S6). Nonetheless, it was still substantially faster than LDpred. Although LDpred typically requires hours to run, lassosum took only minutes.

4 | DISCUSSION

In this paper, we have proposed the calculation of PGS using a penalized regression approach using summary statistics and examined its performance in simulation experiments. Our proposed approach, lassosum, in general appeared to give better prediction than P-value thresholding with or without clumping as well as the recently proposed LDpred, for which we failed to demonstrate the claimed superior performance over P-value thresholding. Clumping was beneficial for P-value thresholding in most scenarios but not for lassosum. In some scenarios, clumping actually decreases the predictive power of P-value thresholding, such as in our simulations with $P(\text{causal}) = 0.1$ and $n = 10,000$.

Compared with LDpred, we showed that lassosum is not only more accurate but also a lot faster. Running lassosum on a reference panel of around 300,000 SNPs and 1,000 individuals typically takes only several minutes without parallel processing. Even when using a reference panel with 8 million SNPs and 500 participants, lassosum took around 15 min without parallel processing for each value of s. The time taken was similar to that for clumping in PLINK 1.9 and therefore lassosum had similar speed to clumping and P-value thresholding when run with a small reference sample size. Increasing the sample size of the reference panel will

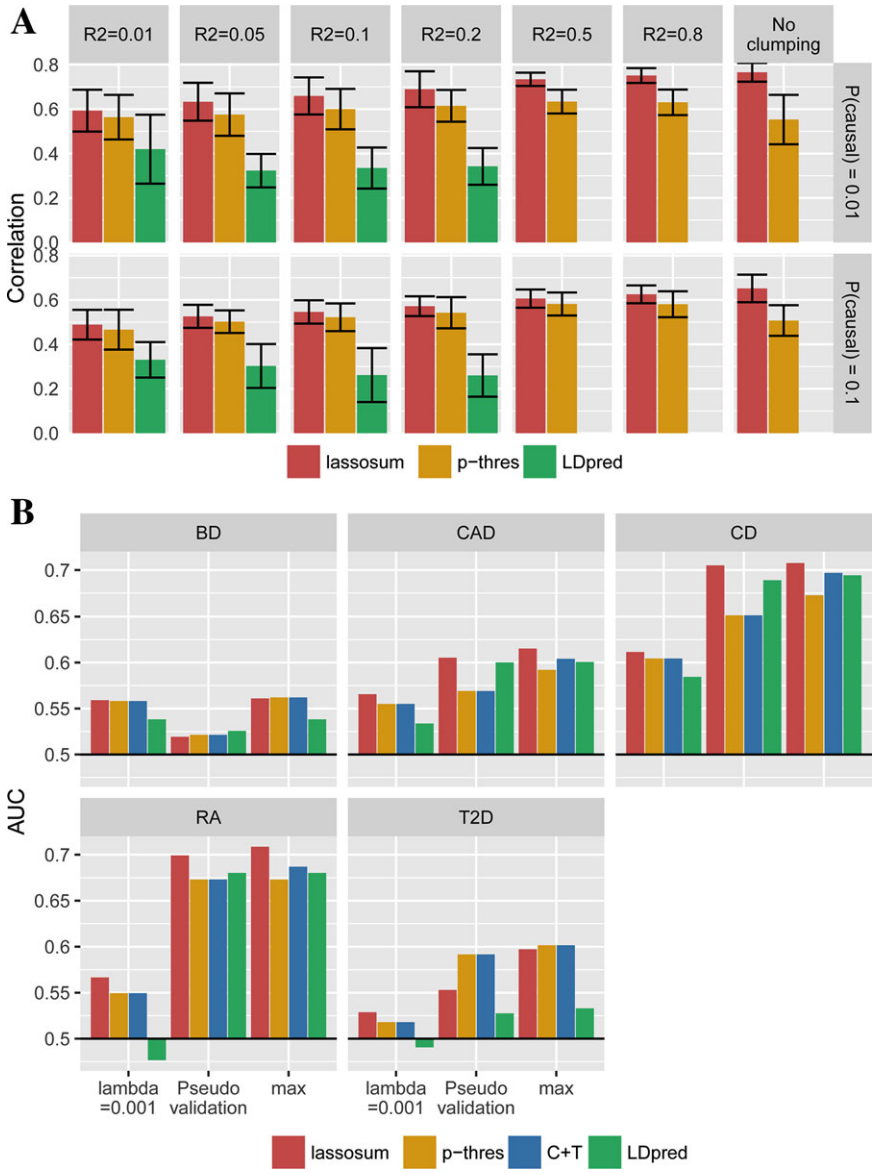

FIGURE 2 (A) Performance of `lassosum` in a large simulated dataset with $n = 200,000$ using different clumping levels in relation to *P*-value thresholding and LDpred. Mean and standard deviation of the AUC of the PGS with the true disease status are plotted. (B) Performance of `lassosum` vs. other methods when using real summary statistics data from meta-analyses. Predictive accuracy was assessed by prediction in the WTCCC dataset after the contribution from WTCCC was removed from the summary statistics. p-thres, *P*-value thresholding without clumping; C + T: *P*-value thresholding with clumping

generally increase prediction accuracy also, although this comes at a cost of exponentially increasing running times. In our simulations, we found that gains in prediction accuracy from a larger reference panel were usually modest. We are currently working on a parallel implementation of `lassosum` and this should be available by the time the article is accepted for publication.

Another contribution from this paper is the method of pseudovalidation, which can be applied to any PGS method that requires a tuning parameter. We showed that it is effective in selecting a parameter value that is close to the optimum. Not surprisingly, having a validation dataset with phenotype data generally provides an even more reliable method for selecting the tuning parameter. However, in the event where this is unavailable, pseudovalidation offers an alternative. Recently, PGS were often used to assess genetic correlation between two diseases. Oftentimes, the tuning parameter (or *P*-value threshold) used in the PGS was chosen by maximizing over the correlation of the PGS with another disease (e.g., Krapohl et al., 2015). We have not examined the performance of using this approach to select the tuning parameter, although it is likely that there will be bias in estimation of correlations due to winner's curse.

Although we have focused on the performance of `lassosum` as a method, we note that it is more generally an instance of penalized regression. Potentially, other penalties

can be used in place of $\lambda||\boldsymbol{\beta}||_1^1$ in Equation (2), which can lead to even better prediction. We chose the LASSO penalty because of its simplicity. Other similar methods that can also be solved using the fast coordinate descent method of Friedman, Hastie, Höfling, and Tibshirani (2007) include the non-negative garotte, LAD-LASSO, and Grouped LASSO.

Some limitations of the present study are worth bearing in mind when considering these results. For example, summary statistics may be inflated due to population stratification in the data where they are generated. As summary statistics are often derived from meta-analyses, it is also possible that there is underlying heterogeneity in effect sizes. How these impact PGS calculation is currently unknown.

Recently, methods for conducting GWAS have moved beyond the single-disease paradigm. Often, multiple related diseases are analyzed together to give improved power for detection of GWAS signals (Andreassen et al., 2013; Chung, Yang, Li, Gelernter, & Zhao, 2014; Korte et al., 2012; Li, Yang, Gelernter, & Zhao, 2014; Zhou & Stephens, 2014). Frequently, these new methods operate in the Bayesian framework resulting in Bayes factor or posterior probability of associations (or alternatively local false discovery rates) for each SNPs. In principle, we can translate these into *P*-values (Stephens & Balding, 2009) and thus make use of additional information to improve PGS predictive performance. Likewise, additional information gained in the consideration of functional annotations of the genome (Kichaev et al., 2014; Pickrell, 2014; Schork et al., 2013) can be incorporated similarly. The simplicity of `lassosum` makes it an ideal framework from which more complex methods can be developed.

ACKNOWLEDGMENTS

We would like to thank Dr. Johnny S. H. Kwan for pointing out to us the work by Strimmer (2008). We would also like to thank two anonymous referees for their comments in improving this paper. We acknowledge financial support from the Hong Kong Research Grants Council General Research Fund (776513M, HKU 776412M, 17128515), the Hong Kong Research Grants Council Theme-Based Research Scheme (T12-705/11, T12/708/12N, T12C-714/14-R), the National Science Foundation of China – Research Grants Council of Hong Kong (N_HKU736/14), and the European Community Seventh Framework Programme Grant on European Network of National Schizophrenia Networks Studying Gene-Environment Interactions (EU-GEI).

REFERENCES

1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Abraham, G., Kowalczyk, A., Zobel, J., & 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, P. F., Vilhjálmsson, B. J., Pedersen, C. B., Mors, O., Børglum A. D., … Mortensen, P. B. (2015). Polygenic risk score, parental socioeconomic status, family history of psychiatric disorders, and the risk for schizophrenia. *JAMA Psychiatry*, *72*(7), 635–641.

Andreassen, O. A., Thompson, W. K., Schork, A. J., Ripke, S., Mattingsdal, M., Kelsoe, J. R., … Dale, A. M. (2013). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genetics*, *9*(4), e1003455.

Berisa, T., & Pickrell, J. K. (2015). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, *32*(2), 283–285.

Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., … Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*(3), 291–295.

Burgess, S., & Thompson, S. G. (2013). Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology*, *42*(4), 1134–1144.

Byrne, E. M., Carrillo-Roa, T., Penninx, B. W. J. H., Sallis, H. M., Viktorin, A., Chapman, B., … Wray, N. R. (2014). Applying polygenic risk scores to postpartum depression. *Archives of Women's Mental Health*, *17*(6), 519–528.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 1–16.

Chang, S. C., Glymour, M. M., Walter, S., Liang, L., Koenen, K. C., Tchetgen, E. J., … Kubzansky, L. D. (2014). Genome-wide polygenic scoring for a 14-year long-term average depression phenotype. *Brain and Behavior*, *4*(2), 298–311.

Chung, D., Yang, C., Li, C., Gelernter, J., & Zhao, H. (2014). GPA: A statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genetics*, *10*(11), e1004787.

de Los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., & Sorensen, D. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genetics*, *9*(7), e1003608.

Domingue, B. W., Belsky, D. W., Harris, K. M., Smolen, A., McQueen, M. B., & Boardman, J. D. (2014). Polygenic risk predicts obesity in both white and black young adults. *PloS One*, *9*(7), e101596.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, *9*(3), e1003348.

Dudbridge, F. (2016). Polygenic epidemiology. *Genetic Epidemiology*, *40*(4), 268–272.

Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., … Goddard, M. E. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, *95*(7), 4114–4129.

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic risk score software. *Bioinformatics* (Advanced Access),*31*(9), 1466–1468.

Evans, D. M., Brion, M. J. A., Paternoster, L., Kemp, J. P., McMahon, G., Munafò, M., … Smith, G. D. (2013). Mining the human phenome using allelic scores that index biological intermediates. *PLoS Genetics*, *9*(10), e1003919.

Evans, D. M., Visscher, P. M., & Wray, N. R. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human Molecular Genetics*, *18*(18), 3525–3531.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, *1*(2), 302–332.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1–22.

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*, (186).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.

Kichaev, G., Yang, W. Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., … Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics*, *10*(10) e1004722.

Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., & Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, *44*(9), 1066–1071.

Krapohl, E., Euesden, J., Zabaneh, D., Pingault, J. B., Rimfeld, K., von Stumm, S., … Plomin, R. (2015). Phenome-wide analysis of genome-wide polygenic scores. *Molecular Psychiatry* (May), *21*(9), 1–6.

Li, C., Yang, C., Gelernter, J., & Zhao, H. (2014). Improving genetic risk prediction by leveraging pleiotropy. *Human Genetics*, *133*(5), 639–650.

Li, M. X., Gui, H. S., Kwan, J. S. H., Bao, S. Y., & Sham, P. C. (2012). A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Research*, *40*(7), e53.

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., … Weersma, R. K. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, *47*(9), 979–989.

Machiela, M. J., Chen, C. Y., Chen, C., Chanock, S. J., Hunter, D. J., & Kraft, P. (2011). Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genetic Epidemiology*, *35*(6), 506–514.

Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., … Morris, A. P. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, *46*(3), 234–44.

Mak, T. S. H., Kwan, J. S. H., Campbell, D. D., & Sham, P. C. (2016). Local true discovery rate weighted polygenic scores using GWAS summary data. *Behavior Genetics*, *46*(4), 573–582.

Martin, J., O'Donovan, M. C., Thapar, A., Langley, K., & Williams, N. (2015). The relationship between common and rare genetic variants in ADHD. *Translational Psychiatry*, *5*, e506.

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Nikpay, M., Goel, A., Won, H. H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., Hopewell, J. C., *et al.* (2015). A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, *47*(10), 1121–1130.

Ogutu, J. O., Schulz-Streeck, T., & Piepho, H. P. (2012). Genomic selection using regularized linear regression models: Ridge regression, LASSO, elastic net and their extensions. *BMC Proceedings*, *6*(Suppl 2), S10.

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., … Plenge, R. M. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, *506*(7488), 376–381.

Pasaniuc, B., & Price, A. L. (2016). Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, *18*(2), 117–127. http://doi.org/10.1038/nrg.2016.142

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics*, *94*(4), 559–573.

Pirinen, M., Donnelly, P., & Spencer, C. C. A. (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Annals of Applied Statistics*, *7*(1), 369–390.

Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, *47*(7), 702–709.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., & Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748–752.

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K. H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., Collier, D. A., Huang, H., *et al.* (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*, 421–427.

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., … Sullivan, P. F. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*, *45*(10), 1150–1159.

Schork, A. J., Thompson, W. K., Pham, P., Torkamani, A., Roddey, J. C., Sullivan, P. F., … Dale, A. M. (2013). All SNPs are not created equal: Genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genetics*, *9*(4), e1003449.

Sklar, P., Ripke, S., Scott, L., Andreassen, O., Cichon, S., Craddock, N., … Purcell, S. M. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics*, *43*(10), 977–983.

So, H. C., Kwan, J. S. H., Cherny, S. S., & Sham, P. C. (2011). Risk prediction of complex diseases from family history and known

susceptibility loci, with applications for cancer screening. *American Journal of Human Genetics*, *88*(5), 548–565.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., … Loos, R. J. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, *42*(11), 937–948.

Stahl, E. A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B. F., Kraft, P., … Plenge, R. M. (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature Genetics*, *44*(5), 483–489.

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, *10*(10), 681–690.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, *9*(303), 1.

Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., König, I. R., Zhang, H., & Sun, Y. V. (2009). Machine learning in genome-wide association studies. *Genetic Epidemiology*, *33*(Suppl 1), S51–S57.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B (Methodological)*, *58*(1), 267–288.

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., … Price, A. L. (2015). modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American Journal of Human Genetics*, *97*(4), 576–592.

Wei, Z., Wang, K., Qu, H. Q., Zhang, H., Bradfield, J., Kim, C., … Hakonarson, H. (2009). From disease association to risk assessment: An optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genetics*, *5*(10), e1000678.

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678.

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A., Dudbridge, F., & Middeldorp, C. M. (2014). Research Review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, *55*(10), 1068–1087.

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., Visscher, P. M. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*, *43*(6), 519–525.

Yi, H., Breheny, P., Imam, N., Liu, Y., & Hoeschele, I. (2014). Penalized multi-marker versus single-marker regression methods for genome-wide association studies of quantitative traits. *Genetics*, *199*(1), 205–222.

Zhou, X., Carbonetto, P., & Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics*, *9*(2), e1003264.

Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, *11*(4), 407–409.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* 2017;00:1–12. https://doi.org/10.1002/gepi.22050