A smoothed version of the Lassosum penalty for fitting integrated risk models

Georg Hahn, Dmitry Prokopenko, Sharon M. Lutz, Kristina Mullin, Rudolph E. Tanzi, and Christoph Lange

Abstract

Polygenic risk scores are a popular means to predict the disease risk or disease susceptibility of an individual based on its genotype information. When adding other important epidemiological covariates such as age or sex, we speak of an integrated risk model. Methodological advances for fitting more accurate integrated risk models are of immediate importance to improve the precision of risk prediction, thereby potentially identifying patients at high risk early on when they are still able to benefit from preventive steps/interventions targeted at increasing their odds of survival, or at reducing their chance of getting a disease in the first place. This article proposes a smoothed version of the "Lassosum" penalty used to fit polygenic risk scores and integrated risk models. The smoothing allows one to obtain explicit gradients everywhere for efficient minimization of the Lassosum objective function while guaranteeing bounds on the accuracy of the fit. An experimental section on both Alzheimer's disease and COPD (chronic obstructive pulmonary disease) demonstrates the increased accuracy of the proposed smoothed Lassosum penalty compared to the original Lassosum algorithm, allowing it to draw equal with state-of-the-art methodology such as LDpred2 when evaluated via the AUC (area under the ROC curve) metric.

Keywords: Integrated risk model; Lassosum; Nesterov; Polygenic Risk Scores; Smoothing.

1 Introduction

Polygenic risk scores are a statistical aggregate of risks typically associated with a set of established DNA variants. If only genotype information of an individual is used to predict its risk, we speak of a polygenic risk score. A polygenic risk score with added epidemiological covariates (such as age or sex) is called an integrated risk model (Wand et al., 2020). The goal of both polygenic risk scores and integrated risk models is to predict the disease risk of an individual, that is the susceptibility

to a certain disease. Such scores are usually calibrated on large genome-wide association studies (GWAS) via high-dimensional regression of a fixed set of genetic variants (and additional covariates in case of an integrated risk model) to the outcome. In this article, we focus on the more general case of an integrated risk model.

Since the potential for broad-scale clinical use to identify people at high risk for certain diseases has been demonstrated (Khera et al., 2018), polygenic risk scores and integrated risk models have become a widespread tool for the early identification of patients who are at high risk for a certain disease and who could benefit from intervention measures. However, the accuracy of current polygenic risk scores, measured with the AUC metric (Area under the ROC Curve, where ROC stands for receiver operating characteristic, see Mandrekar (2010)), varies substantially for important diseases. For instance, the AUC achieved by state-of-the-art methods ranges from around 0.8 for type 1 diabetes to around 0.7 for coronary artery disease and schizophrenia (Mak et al., 2017), while for atrial fibrillation the AUC is around 0.64 (Huang and Darbar, 2017), a value which is considered less than acceptable (Mandrekar, 2010; Hosmer and Lemeshow, 2000). For this reason, increasing the accuracy of scores is desirable, which is the focus of the proposed smoothing approach.

Several methodological approaches have been considered in the literature to compute a polygenic risk score or an integrated risk model for a given population, and to predict a given outcome (disease status). For instance, LDpred of Vilhjálmsson et al. (2015) and LDpred2 of Privé et al. (2019) fit a Bayesian model to the effect sizes via Gibbs sampling, and obtain a score via posterior means of the fitted model. The PRS-CS approach of Ge et al. (2020) likewise utilizes a high-dimensional Bayesian regression framework in connection with a continuous shrinkage prior (hence the suffix *CS* for continuous shrinkage) on SNP effect sizes. Fitting genotype data to a disease outcome can also be achieved by means of a simple penalized regression using the least absolute shrinkage and selection operator (Lasso) of Tibshirani (1996), for instance using the *glmnet* package on CRAN, see Friedman et al. (2010, 2020).

One popular way to fit a polygenic risk score is the "Lassosum" approach of Mak et al. (2017). Note that in Mak et al. (2017), no integrated risk models are considered. The Lassosum method is based on a reformulation of the linear regression problem $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $X \in \mathbb{R}^{n\times p}$ denotes SNP data for n individuals and p SNP locations, $\boldsymbol{y} \in \mathbb{R}^n$ denotes a vector of outcomes, $\boldsymbol{\beta} \in \mathbb{R}^p$ is unknown, and $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2 I_n)$ is an n-dimensional, independently and normally distributed

error term with mean zero and some variance $\sigma^2 > 0$ (where I_n denotes the n-dimensional identity matrix). The authors start with the classic Lasso objective function $L(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2 + 2\lambda\|\boldsymbol{\beta}\|_1$, where $\lambda \geq 0$ denotes the Lasso regularization parameter controlling the sparseness of the solution, and rewrite it using the SNP-wise correlation $\boldsymbol{r} = X^\top \boldsymbol{y}$ as

$$L(\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} + (1-s)\boldsymbol{\beta}^\top X_r^\top X_r \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \boldsymbol{r} + s\boldsymbol{\beta}^\top \boldsymbol{\beta} + 2\lambda\|\boldsymbol{\beta}\|_1, \tag{1}$$

where X_r denotes the matrix of genotype data used to derive estimates of LD (linkage disequilibrium), $\lambda \geq 0$ is the Lasso regularization parameter controlling the sparseness of the estimate, and $s \in (0, 1)$ is an additional regularization parameter used to ensure stability and uniqueness of the Lasso solution. Importantly, in Mak et al. (2017) the authors derive an iterative scheme to carry out the minimization of eq. (1) which only requires one column of X_r at a time, thus avoiding the costly computation of the matrix $X_r^\top X_r \in \mathbb{R}^{p \times p}$.

In this work, we consider a different approach for minimizing eq. (1). Using the methodology of Nesterov (2005), we propose to smooth the non-differentiable L_1 penalty in eq. (1), thus allowing us to compute explicit gradients of eq. (1) everywhere. This in turn allows us to efficiently minimize the Lassosum objective function using a quasi-Newton minimization algorithm such as BFGS (Broyden–Fletcher–Goldfarb–Shanno). Besides enabling a more efficient and more accurate computation of the score, our work extends the one of Mak et al. (2017) in that we do not solely consider polygenic risk scores, but the more general integrated risk models. Our approach follows as a special case from Hahn et al. (2020b,a), who propose a general framework to smooth L_1 penalties in a linear regression. Importantly, employing a smoothing approach has a variety of theoretical advantages following directly from Hahn et al. (2020b). Apart from obtaining explicit gradients for fast and efficient minimization, the smoothed objective is convex, thus ensuring efficient minimization, and it is guaranteed that the solution (the fitted integrated risk model) obtained by solving the smoothed Lassosum objective is never further away than a user-specified quantity from the original (unsmoothed) objective of Mak et al. (2017).

We evaluate all aforementioned approaches by computing an integrated risk model in two experimental studies, one on Alzheimer's disease using the summary statistics of Kunkle et al. (2019) and Jansen et al. (2019), and on COPD (chronic obstructive pulmonary disease) using FEV1 data

of Regan E.A. (2010); NHLBI TOPMed (2018). In the first case, the response is binary, whereas in the second study the response is continuous. Our simulations demonstrate that smoothing the Lassosum objective function results in a considerably enhanced performance of the Lassosum approach, allowing it to draw equal with approaches such as LDpred2 or PRScs.

Analogously to the original Lasso of Tibshirani (1996), the L_1 penalty employed in eq. (1) causes some entries of $\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} L(\boldsymbol{\beta})$ to be shrunk to zero exactly (provided the regularization parameter λ is not too small). Therefore, Lassosum performs fitting of the polygenic risk score or integrated risk model and variable selection simultaneously.

This article is structured as follows. Section 2 introduces the smoothed Lassosum objective function and discusses its minimization, the theoretical guarantees it comes with, and its drawbacks. Section 3 evaluates the proposed approach, the original Lassosum approach, as well as additional state-of-the-art methods in two experimental studies on both Alzheimer's disease and COPD. The article concludes with a discussion in Section 4. The appendix contains two figures showing plots of principal components for the genotype dataset employed in Section 3.1.

The methodology of this article is implemented in the R package *smoothedLasso* (see function *prsLasso* in the package), available on CRAN (Hahn et al., 2020c).

2 Methodology

The Lassosum function of eq. (1) consists of a smooth part, given by $\boldsymbol{y}^\top\boldsymbol{y} + (1-s)\boldsymbol{\beta}^\top X_r^\top X_r\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\boldsymbol{r} + s\boldsymbol{\beta}^\top\boldsymbol{\beta}$, and a non-smooth part, the L_1 penalty $2\lambda\|\boldsymbol{\beta}\|_1$. Only the latter needs smoothing, which we achieve with the help of Nesterov smoothing introduced in Section 2.1. Section 2.2 applies the Nesterov methodology to Lassosum and introduces our proposed smoothed Lassosum objective function. The proposed smoothed Lassosum actually follows from the more general framework of Hahn et al. (2020a,b). We demonstrate this in Section 2.3, where we also state the theoretical guarantees following from the framework.

2.1 Brief overview of Nesterov smoothing

In Nesterov (2005), the author introduces a framework to smooth a piecewise affine and convex function $f : \mathbb{R}^q \to \mathbb{R}$, where $q \in \mathbb{N}$. Since f is piecewise affine, it can be written for $\boldsymbol{z} \in \mathbb{R}^q$ as

$$f(\boldsymbol{z}) = \max_{i=1,\ldots,k} \left(A[\boldsymbol{z}, 1]^\top \right)_i, \tag{2}$$

using $k \in \mathbb{N}$ linear pieces (components), where $[\boldsymbol{z}, 1] \in \mathbb{R}^{q+1}$ denotes the vector obtained by concatenating $\boldsymbol{z}$ and the scalar 1. In eq. (2), the linear coefficients of each of the k linear pieces are summarized as a matrix $A \in \mathbb{R}^{k \times (q+1)}$ (with the constant coefficients being in column $q+1$).

The author then introduces a smoothed version of eq. (2) as

$$f^\mu(\boldsymbol{z}) = \max_{\boldsymbol{w} \in Q_k} \left\{ \langle A[\boldsymbol{z}, 1]^\top, \boldsymbol{w} \rangle - \mu \rho(\boldsymbol{w}) \right\}, \tag{3}$$

where $Q_k = \left\{ \boldsymbol{w} \in \mathbb{R}^k : \sum_{i=1}^k \boldsymbol{w}_i = 1, \boldsymbol{w}_i \geq 0 \ \forall i = 1, \ldots, k \right\} \subseteq \mathbb{R}^k$ is the unit simplex in k dimensions. The parameter $\mu \geq 0$ controls the smoothness of the approximation f^μ to f, called the Nesterov smoothing parameter. Larger values of μ result in a stronger smoothing effect, while the choice $\mu = 0$ recovers $f^0 = f$. The function ρ is called the proximity function (or prox-function) which is assumed to be nonnegative, continuously differentiable, and strongly convex.

Importantly, f^μ is both smooth for any $\mu > 0$ and uniformly close to f, that is the approximation error is uniformly bounded as

$$\sup_{\boldsymbol{z} \in \mathbb{R}^q} |f(\boldsymbol{z}) - f^\mu(\boldsymbol{z})| \leq \mu \sup_{\boldsymbol{w} \in Q_k} \rho(\boldsymbol{w}) = O(\mu),$$

see (Nesterov, 2005, Theorem 1). Though several choice of the prox-function ρ are considered in Nesterov (2005), we fix one particular choice (called the entropy prox-function) in the remainder of the article for the following reasons: (a) The different prox-functions are equivalent in that all choices yield the same theoretical guarantee and performance; and (b) the entropy prox-function leads to a closed-form expression of eq. (3) given by

$$f_e^\mu(\boldsymbol{z}) = \mu \log \left(\frac{1}{k} \sum_{i=1}^k e^{\frac{\left(A[\boldsymbol{z},1]^\top\right)_i}{\mu}} \right), \tag{4}$$

which satisfies the uniform bound

$$\sup_{\boldsymbol{z}\in\mathbb{R}^q} |f(\boldsymbol{z}) - f_e^{\mu}(\boldsymbol{z})| \leq \mu \log(k), \tag{5}$$

see Nesterov (2005) and Hahn et al. (2020a,b).

2.2 A smoothed version of the Lassosum objective function

As observed at the beginning of Section 2, it suffices to smooth the non-differentiable penalty $2\lambda\|\boldsymbol{\beta}\|_1$ of the Lassosum objective function, where $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{p} |\boldsymbol{\beta}_i|$. To this end, we apply Nesterov smoothing to each absolute value independently.

We observe that the absolute value can be expressed as piecewise affine function with $k = 2$ components, given by $f(z) = \max\{-z, z\} = \max_{i=1,2} \left(A[z, 1]^{\top}\right)_i$, where

$$A = \begin{pmatrix} -1 & 0 \\ 1 & 0 \end{pmatrix}$$

and $z \in \mathbb{R}$ is a scalar. Substituting this specific choice of A into eq. (4) leads to a smoothed approximation of the absolute value given by

$$f_e^{\mu}(z) = \mu \log\left(\frac{1}{2}e^{-z/\mu} + \frac{1}{2}e^{z/\mu}\right). \tag{6}$$

Substituting the absolute value in the L_1 norm in eq. (1) with the approximation in eq. (6) results in a *smoothed version of the Lassosum objective function*, given by

$$L^{\mu}(\boldsymbol{\beta}) = \boldsymbol{y}^{\top}\boldsymbol{y} + (1-s)\boldsymbol{\beta}^{\top}X^{\top}X\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\top}\boldsymbol{r} + s\boldsymbol{\beta}^{\top}\boldsymbol{\beta} + 2\lambda\sum_{i=1}^{p} f_e^{\mu}(\boldsymbol{\beta}_i). \tag{7}$$

The first derivative of f_e^{μ} is explicitly given by

$$\frac{\partial}{\partial z} f_e^{\mu}(z) = \frac{-e^{-z/\mu} + e^{z/\mu}}{e^{-z/\mu} + e^{z/\mu}} =: g_e^{\mu}(z),$$

see also Hahn et al. (2020a,b), from which the closed-form gradient of the smoothed Lassosum

objective function of eq. (7) immediately follows as

$$\frac{\partial}{\partial \boldsymbol{\beta}} L^{\mu} = (1-s)2(X^{\top}X)\boldsymbol{\beta} - 2\boldsymbol{r} + 2s\boldsymbol{\beta} + 2\lambda \sum_{i=1}^{p} g_e^{\mu}(\boldsymbol{\beta}_i).$$

Using the smoothed version of the Lassosum objective function, given by L^{μ}, and its explicit gradient $\frac{\partial}{\partial \boldsymbol{\beta}} L^{\mu}$, an integrated risk model can easily be computed by minimizing L^{μ} using a quasi-Newton method such as BFGS (Broyden–Fletcher–Goldfarb–Shanno), implemented in the funcition *optim* in *R* (R Core Team, 2014).

In eq. (7), the quantity X is not limited to contain only genotype information. Any data on the individuals (including additional epidemiological covariates) to compute the integrated risk model can be summarized in X. The other quantities in eq. (7) are the outcome $\boldsymbol{y}$ (either binary/discrete or continuous), the correlations $\boldsymbol{r} = X^{\top}\boldsymbol{y}$, and the additional regularization parameter $s \in (0, 1)$ introduced by Mak et al. (2017) used to ensure stability and uniqueness of the Lasso solution.

2.3 Theoretical guarantees

Using the fact that the absolute value can be expressed as a piecewise affine function with $k = 2$, see Section 2.2, the error bound of eq. (5) can be re-written as

$$\sup_{z \in \mathbb{R}} |f(z) - f_e^{\mu}(z)| \leq \mu \log(2). \tag{8}$$

Since in our proposed smoothed version of eq. (7), only the non-smooth L_1 contribution of the original Lassosum objective function of eq. (1) has been replaced, the bound of eq. (8) immediately carries over to a bound on the smoothed Lassosum. In particular,

$$\sup_{\boldsymbol{\beta} \in \mathbb{R}^p} |L(\boldsymbol{\beta}) - L_e^{\mu}(\boldsymbol{\beta})| \leq \sup_{\boldsymbol{\beta} \in \mathbb{R}^p} 2\lambda \left| \sum_{i=1}^{p} |\boldsymbol{\beta}_i| - \sum_{i=1}^{p} g_e^{\mu}(\boldsymbol{\beta}_i) \right| \leq 2\lambda p \mu \log(2). \tag{9}$$

For a given computation of an integrated risk model, the Lasso parameter $\lambda > 0$ and the dimension p are fixed by the problem specification. According to eq. (9), this allows one to make the approximation error of our proposed smoothed Lassosum to the original Lassosum arbitrarily small as the smoothing parameter $\mu \to 0$.

As stated in Section 2.1 of Mak et al. (2017), the Lassosum objective of eq. (1) is equivalent to

a Lasso problem, in particular its convexity is preserved. According to Proposition 2 in Hahn et al. (2020b), the smooth approximation of eq. (7) obtained via Nesterov smoothing is strictly convex. Since strictly convex functions have one unique minimum, and since a closed-form gradient $\frac{\partial}{\partial \boldsymbol{\beta}} L^{\mu}$ of L^{μ} is available (see Section 2.2), this makes the minimization of our proposed smoothed Lassosum in lieu of the original Lassosum very appealing.

Furthermore, two additional properties of eq. (7) can be derived from (Hahn et al., 2020b, Section 4.3). First, the $\arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} L^{\mu}(\boldsymbol{\beta})$ is continuous with respect to the supremum norm (Hahn et al., 2020b, Proposition 4), which implies that the minimum of our proposed smoothed Lassosum L^{μ} converges to the one of the original Lassosum as $\mu \to 0$. Second, in addition to this qualitative statement, the error between the minimizers of the smoothed and original Lassosum function can be quantified a priori (Hahn et al., 2020b, Proposition 5).

3 Experiments

The proposed smoothed Lassosum approach is obtained by applying Nesterov smoothing to the L_1 penalty of the Lassosum objective function, see eq. (1). A detailed study on the behavior of Nesterov smoothing applied to an L_1 penalty using synthetic data can be found in Hahn et al. (2020b).

In this section, we evaluate the the performance of our proposed smoothed Lassosum approach of Section 2.2 in two experimental studies, one fitting an integrated risk model to binary outcomes in the context of Alzheimer's disease (Section 3.1), and one fitting an integrated risk model to continuous outcomes in the context of COPD (Section 3.2). We benchmark our smoothed Lassosum approach, which we refer to as "SmoothedLassosum", against the following state-of-the-art approaches:

1. We benchmark against the original Lassosum of Mak et al. (2017), implemented in the R package *lassosum* (Mak et al., 2020), and refer to it as "Lassosum".

2. LDpred and LDpred2 (Vilhjálmsson et al., 2015; Privé et al., 2019) compute a polygenic risk score (but not integrated risk model) by inferring the posterior mean effect size of each marker by using a prior on effect sizes and LD information from an external reference panel. To run

LDpred2, we employed the implementation in the R package *bigsnpr* on CRAN (Privé et al., 2020) using the Gibbs sampler to estimate effect sizes. We will refer to this algorithm as "LDpred2".

3. PRScs of Ge et al. (2019) utilizes a high-dimensional Bayesian regression framework which places a continuous shrinkage prior on SNP effect sizes, an innovation which the authors claim is robust to varying genetic architectures. We use the implementation of Ge et al. (2020) and refer to it as "PRScs".

4. We employ a simple penalized regression using the Lasso of Tibshirani (1996) to fit the genotype data to disease outcome. We employ the *Glmnet* package on CRAN, see Friedman et al. (2010, 2020). We will refer to this method as "Glmnet".

5. We employ the unsmoothed Lasso of Hahn et al. (2020a), implemented in the R package *smoothedLasso* on CRAN (Hahn et al., 2020d). We refer to this method as "Lasso".

6. Similarly, we also employ the smoothed Lasso of Hahn et al. (2020a), which is likewise implemented in *smoothedLasso* on CRAN (Hahn et al., 2020d). We refer to this method as "SmoothedLasso".

7. We train a neural network with the *Keras* interface (Falbel et al., 2020a) to the *Tensorflow* machine learning platform (Falbel et al., 2020b). We train a network with four layers, having 20, 8, 4 and 2 nodes. We employ the LeakyReLU activation function, a dropout rate of 0.1, a validation splitting rate of 0.1, the *he_normal* truncated normal distribution for kernel initialization, and kernel, bias and activity regularization with L_1 penalty. The last layer employs the sigmoid (for Section 3.1) or ReLU (for Section 3.2) activation functions. The model is compiled for binary crossentropy loss (for Section 3.1) or mean absolute error loss (for Section 3.2) using the Adam optimizer, evaluated with the AUC (for Section 3.1) or the mean squared error (for Section 3.2) using 1000 epochs. We refer to the neural network as "NeuralNetwork".

8. We employ *SBayesR* of Lloyd-Jones et al. (2019), a linear regression likelihood which takes into account GWAS summary statistics and a reference LD correlation matrix, and is coupled

to a finite mixture of normal priors on the genetic effects. The normal priors allow one to incorporate sparsity and to perform Bayesian posterior inference on the model parameters, such as genetic effects, variance components and mixing proportions. The method is implemented in the toolbox *GCTB* (Zeng et al., 2020). We employ SBayesR with default parameters and refer to it as "SBayesR".

9. We run *MegaPRS* of Zhang et al. (2021). In particular, we employ the robust version *Bolt Predict*, as suggested by the authors, using default parameters given the example section of MegaPRS (a cross validation proportion of 0.1, the *--ignore-weights* option and a power parameter of -0.25). MegaPRS is implemented in the *LDAK* package (Speed, 2021). We refer to it as "MegaPRS".

10. We use epidemiological covariates only in a simple linear regression fit to the response. We refer to this as "EpiOnly".

The Lassosum, LDpred2, PRScs, SBayesR, and MegaPRS algorithms are only designed to fit polygenic risk scores, but not integrated risk models. To include epidemiological covariates for these methods (and thus fit an integrated risk model), we first perform a linear regression of the epidemiological covariates to the outcome, and then run the aforementioned methods on the residuals. Importantly, in order to apply Lassosum with epidemiological covariates, we additionally have to recompute the SNP-wise correlation $\boldsymbol{r} = X^\top \boldsymbol{y}$ as in eq. (1) using the residuals in place of $\boldsymbol{y}$.

Note that Glmnet, as well as Lasso and SmoothedLasso, can be applied in two ways. First, they can be applied to both the epidemiological covariates and genotype information in one go, given all information is summarized in the design matrix. Second, they can likewise be applied to residuals after regressing out all epidemiological covariates. For consistency with the way the Lassosum, LDpred2, PRScs, SBayesR, and MegaPRS algorithms are applied, we also employ Glmnet, Lasso and SmoothedLasso to residuals after regressing out all epidemiological covariates. Throughout the section, we fix the Lasso regularization parameter at $\lambda = 2^{-3}$, the Lassosum regularization parameter s in eq. (1) at $s = 0.5$ (this parameter is used to ensure stability and uniqueness of solution), and the smoothing parameter of Section 2.2 at $\mu = 0.1$.

3.1 Alzheimer study

We performed training and testing of different PRS algorithms using summary statistics for Alzheimer's disease (AD), together with genotype data imputed on the Haplotype Reference Consortium (HRC), see McCarthy et al. (2016). The HRC-imputed genotype data was downloaded from Partners Biobank (Partners, 2020) (described below). The summary statistics are matched to genotype data for chromosomes 1–22 of $2,465$ patients available in the Partners Biobank. As initial training weights we considered two sets of summary statistics from two largest available AD GWAS: the one of clinically defined AD cases of Kunkle et al. (2019), and the one of AD-by-proxy phenotypes of Jansen et al. (2019).

The dataset of Kunkle et al. (2019) contains a total of $11,480,632$ summary statistics, given by p-value, effect size (beta), and standard deviation of the effect size. Each entry is characterized by its chromosome number, position on the chromosome, as well as the effect allele and non-effect allele. The dataset of Jansen et al. (2019) contains a total of $13,367,299$ summary statistics in the same format as the one of Kunkle et al. (2019).

Partners Biobank is a hospital-based cohort from the MassGeneral Brigham (MGB) hospitals. This cohort includes collected DNA from consented subjects linked to electronic health records. We have obtained a subset in April 2019, which included AD cases and controls. Cases were defined as subjects who were diagnosed with AD based on the International Statistical Classification of Diseases and Related Health Problems (ICD-10), see World Health Organization (2021). Controls were selected as individuals of age 60 and greater, who had no family history of AD, no diagnosed disease of nervous system (coded as G00-G99 in ICD-10), no mental and behavioral disorders (coded as F01-F99 in ICD-10), and a Charlson Age-Comorbidity Index of 2, 3, or 4 (Charlson et al., 1994; Karlson et al., 2016).

We performed the following quality control steps on the HRC-imputed genotype data from Partners Biobank. Relatedness was assessed with KING (Manichaikul et al., 2010; Chen, 2021) and population structure was assessed with principal components. Principal components were calculated on a pruned subset (PLINK2 parameters: --indep-pairwise 50 5 0.05) of common variants (MAF > 0.1). We excluded subjects which had a KING kinship coefficient > 0.0438 (third degree of relatedness or less) and which were at least 5 standard deviations away from the mean value of

the inbreeding coefficient. We kept only self-reported non-hispanic white (NHW) individuals and excluded outliers, defined as subjects which are at least 5 standard deviations away from the mean value of each of the ten principal components (see Section A). There was a total of $2,465$ subjects (481 cases) left for analysis.

To compare performance across both datasets, we determined the set of variants which are found in both datasets, as well as in the genotype data of the Partners Biobank. We randomly selected $20,000$ loci with the *--thincount* option in PLINK2 (Purcell and Chang, 2020). Although *APOE* variants are known to have a very high effect size for AD, explaining around a quarter of the total heritability (Zhang et al., 2020), including the *APOE* region in a polygenic risk score or integrated risk model has been shown to be insufficient to account for the large risk attributed to *APOE* (Ware et al., 2020). To fine tune our integrated risk models on other *non-APOE* variants with much smaller effect sizes and good prediction power, we decided to keep *APOE* status as a separate predictor. At the same time, we made sure that the extended *APOE* region (from $45,000,000 - 46,000,000$bp on chromosome 19) is excluded while the two *APOE* loci 19:45411941:T:C and 19:45412079:C:T are kept in the data. This leaves $18,038$ loci.

The final data used for the computation of the integrated risk models consists of these $18,038$ loci, as well as the following epidemiological covariates: age, sex, and *APOE* status with classes "none" (encoded as 0), "single e4" (encoded as 1), or "e4/e4" (encoded as 2).

In the following experiments, we considered the datasets of Kunkle et al. (2019) and Jansen et al. (2019) separately and extracted SNP weights based on corresponding effect sizes. Next, we withhold a proportion $p \in \{0.1, \ldots, 0.9\}$ of the pool of Partners genotyped subjects as a validation dataset to fit an integrated risk model with the aforementioned methods, or to tune the hyperparameters of the neural network. Finally, we evaluated their performance on the unseen proportion of the data $(1-p)$. We report the mean of absolute residuals $\frac{1}{n}\sum_{i=1}^{n}|r_i|$ (where n is the number of subjects in the validation set and r_i is the residual for subject i), the AUC (Area under the ROC Curve), and the correlation between predicted and true outcomes.

Figure 1 shows results for the dataset of Kunkle et al. (2019). A series of observations are noteworthy. First, the mean of absolute residuals decreases with an increasing proportion of the data used for training, as expected.

Second, the AUC is very high (reaching almost 0.80) for all methods apart from Lassosum, Lasso,

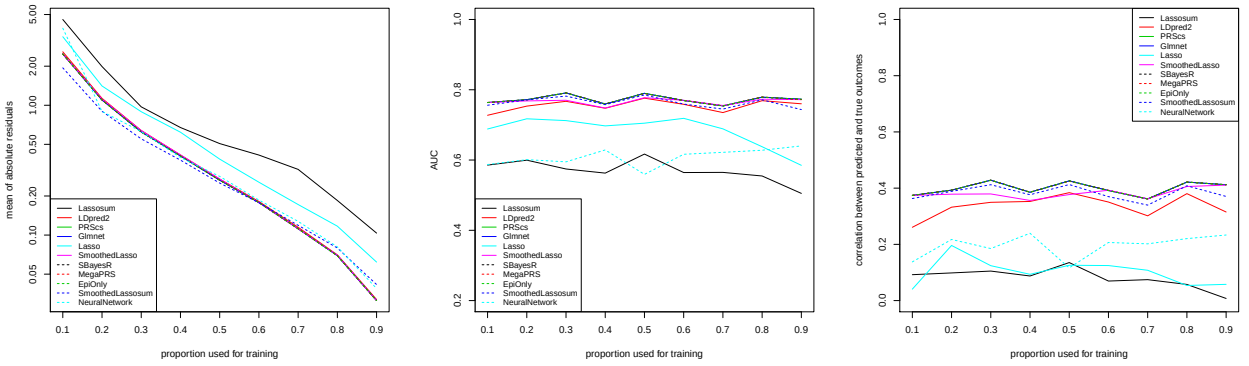

Figure 1: Dataset of clinically defined AD cases of Kunkle et al. (2019). Mean of absolute residuals (left), AUC (middle), and correlation between predicted and true outcomes (right) as a function of the proportion of data used for training. The behavior of most methods is similar to the one of LDpred2 or Glmnet.

and NeuralNetwork. Interestingly, it is much less affected than the residuals by the proportion of data used for training and stays essentially constant for all training proportions. A similar picture is observed when looking at the correlation between predicted and true outcomes, which is roughly equally high for all methods apart from Lassosum, Lasso, and NeuralNetwork. After training, NeuralNetwork achieves a very low mean of absolute residuals, though its AUC and its correlation between predicted and true outcomes somewhat lacks behind the other methods. NeuralNetwork does manage to achieve an increased performance for higher proportions of training data (in both the AUC metric and with respect to the correlation between predicted and true outcomes). This is sensible, as neural nets traditionally need large amounts of data to be trained on.

Third, using epidemiological covariates only in a simple linear regression fit seems to perform very well on this dataset. This seems to suggest that actually, the response is well explained by the genetic factor of *APOE* status as well as the other non-genetic factors (such as age), and that the remaining genetic information is rather negligible for prediction.

Fourth, our proposed SmoothedLassosum considerably improves upon Lassosum of Mak et al. (2017), now drawing equal with state-of-the-art methodology such as LDpred2 with respect to e.g. the AUC measure. Moreover, our proposed SmoothedLassosum achieves a considerably improved mean of absolute residuals compared to Lassosum, and a state-of-the-art correlation between predicted and true outcomes. The reason for the reduced performance of Lassosum is not fully understood. However, it is likely related to the fact that Lassosum is not designed to incorporate

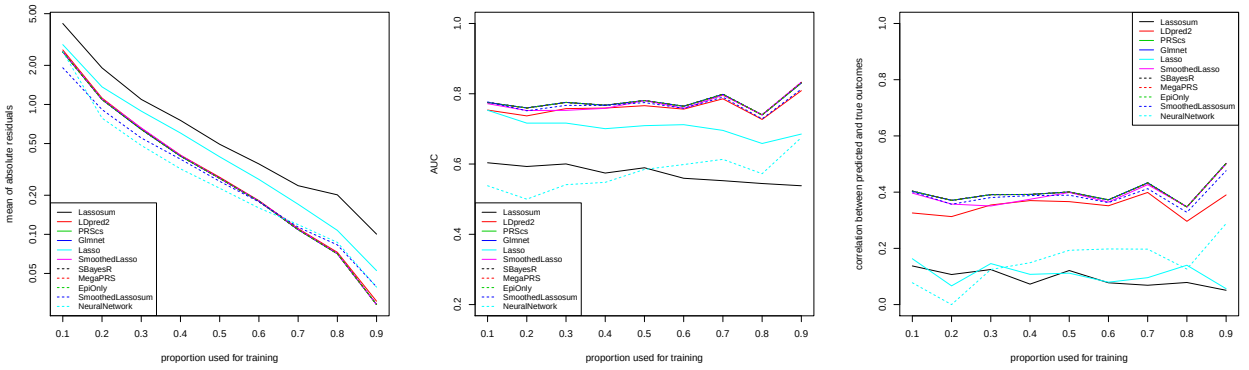

Figure 2: Dataset of AD-by-proxy phenotypes of Jansen et al. (2019). Mean of absolute residuals (left), AUC (middle), and correlation between predicted and true outcomes (right) as a function of the proportion of data used for training. The behavior of most methods is similar to the one of LDpred2 or Glmnet.

epidemiological covariates (see Section 4 for more details).

The results for the dataset of Jansen et al. (2019), reported in Figure 2, are almost identical to the ones for the dataset of Kunkle et al. (2019) in Figure 1. In particular, the Lassosum, Lasso, and NeuralNetwork algorithms generally have the weakest performance on this dataset, while the other methods perform equally well. Importantly, SmoothedLassosum considerably improves upon Lassosum by achieving a mean of absolute residuals, AUC, and correlation between predicted and true outcomes that is similar to the others methods.

The very similar behavior of all methods is expected. The two experiments differ only in the way the response (AD status) is defined. The response provided in Kunkle et al. (2019) consists of clinically defined AD cases, while the one of Jansen et al. (2019) contains AD-by-proxy phenotypes which are based on 13 independent GWS loci having a strong genetic correlation of (at least) 0.81 with the AD status.

3.2 COPD study

The dataset considered in Section 3.1 is characterized through binary outcomes. In this section, we consider a continuous response in the context of Chronic Obstructive Pulmonary Disease (COPD). To be precise, we look at the COPDGene study of Regan E.A. (2010), a case-control study of COPD in current and former smokers (Silverman et al., 1998, 2000) which has been sequenced as part of the TOPMED Project. The data we employed are available through dbGaP (NHLBI

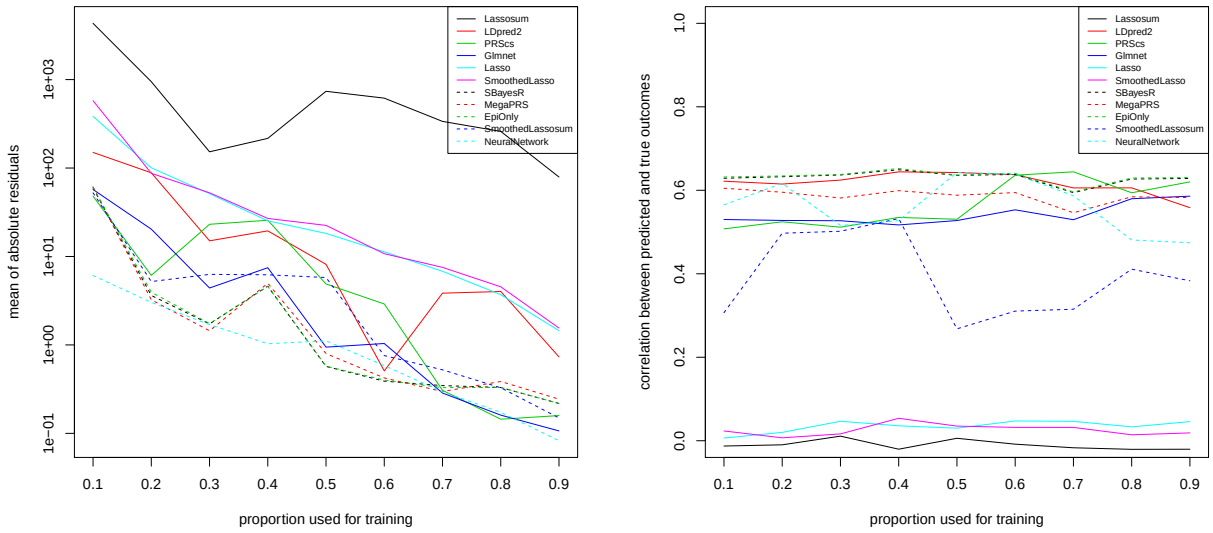

Figure 3: Dataset of AD-by-proxy phenotypes of Jansen et al. (2019). Mean of absolute residuals (left) and correlation between predicted and true outcomes (right) as a function of the proportion of data used for training.

TOPMed, 2018).

Chronic obstructive pulmonary disease (COPD) is the third leading cause of death in the United States (NHLBI TOPMed, 2018). The dataset we consider contains subjects with severe COPD, defined as having a *FEV1* ratio of $< 40\%$ predicted at an early age (< 53 years) without alpha-1 antitrypsin deficiency (a known Mendelian risk factor for COPD). The *FEV1* ratio, also called the Tiffeneau-Pinelli index describes the proportion of lung volume that a person can exhale within the first second of a forced expiration in a spirometry (pulmonary function) test. We focus on chromosome 15 and consider the risk loci for spirometric measures which have been identified in Lutz et al. (2015). Overall, we consider $8,881$ loci for $3,495$ individuals.

The genetic information is then matched to four epidemiological covariates. The final data used for the computation of the integrated risk models consists of the $8,881$ loci, as well as age, sex, pack-years of smoking, and height in centimeters. We aim to predict FEV1 from this data, again using a classic training (proportion $p \in (0,1)$) and validation (proportion $1-p$) setup. We apply all algorithms as outlined in Section 3. As the AUC is only defined for a categorical response, we only report the mean of absolute residuals and the correlation between predicted and true outcomes.

Results of this experiment are given in Figure 3. We observe that measurements are overall

more unstable than in Section 3.1, though as usual, the mean of absolute residuals in Figure 3 (left) decreases with an increasing proportion of the data used for training.

Lassosum is again not performing at its best, which is likely related to the fact that we are aiming to predict a continuous response (see Section 4 for more details). The Lasso and SmoothedLasso approaches are performing average. Together with LDpred2 and PRScs, our proposed SmoothedLassosum approach performs very well and again considerably improves upon the original Lassosum. Glmnet is again one of the best methods together with SBayesR, MegaPRS, though a fit of epidemiological covariates only also seems to have high predictive power. NeuralNetwork seems to be very suited in this experiment to learn the continuous FEV1 responses from the input data.

The correlation between predicted and true outcomes, shown in Figure 3 (right), confirms that most state-of-the-art algorithms achieve a comparable correlation of around 0.6. The performance of our SmoothedLassosum is slightly worse than those methods with regards to the correlation between predicted and true outcomes, though it again considerably improves upon Lassosum (as well as Lasso and SmoothedLasso) which seem to have difficulties to predict the continuous FEV1 response from this data.

4 Discussion

This article considered the calculation of an integrated risk model by minimizing a smoothed version of the Lassosum objective function (see eq. (1)) introduced in Mak et al. (2017). Utilizing a smoothing approach circumvents the non-differentiability of the L_1 penalty of Lassosum, thus allowing for an efficient minimization with quasi-Newton algorithms.

An experimental study on Alzheimer's disease and COPD demonstrates that our smoothed Lassosum improves upon the original Lassosum of Mak et al. (2017), measured with respect to the mean of absolute residuals, the AUC, and the correlation between predicted and true outcomes, thus making it draw equal in accuracy with state-of-the-art approaches. The reduced performance of Lassosum we observe in the simulations is likely attributed to the fact that (a) Lassosum is not designed to incorporate epidemiological covariates in integrated risk models, and (b) Lassosum is not designed for continuous responses (as in the COPD study). In particular, although recomputing the SNP-wise correlation $\boldsymbol{r} = X^\top \boldsymbol{y}$ in eq. (1) and using them in place of $\boldsymbol{y}$ is a valid approach and

an admissible input to the Lassosum objective function, the distribution of residuals is different from the one of the original binary response (without regressing out the covariates), which might cause a suboptimal behavior of the Lassosum algorithm. In contrast, our smoothed Lassosum works well in those cases.

Using an L_1 penalty in eq. (1) has the advantage that, in analogy to the original Lasso of Tibshirani (1996), computing $\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} L(\boldsymbol{\beta})$ performs both regression of the polygenic risk score or integrated risk model and variable selection simultaneously. One potential drawback of our proposed smoothed Lassosum is that it yields dense minimizers (i.e., unused predictors are not necessarily shrunk to zero), meaning that the variable selection property is not preserved. This is not necessarily a disadvantage, as usually the fitted models are only used for risk prediction, for which our dense models achieve a high accuracy. Moreover, other widespread methods such as neural networks likewise do not provide variable selection. If necessary, sparseness can be restored after estimation via thresholding, meaning that all entries β_i of the estimate $\boldsymbol{\beta}$ of eq. (1) satisfying $|\beta_i| < \tau$ for some threshold τ are set to zero. Determining an optimal threshold remains for future research.

Conflict of Interest

The authors declare no conflict of interest.

Funding

The methodology work in this paper was funded by Cure Alzheimer's Fund.

Acknowledgements

This work involved the use of the Enterprise Research Infrastructure & Services (ERIS) at Massachusetts General Hospital. We thank the MGB/Partners HealthCare Biobank for providing genomic and health information data.

Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). See the TOPMed Omics Support Table

TOPMed Accession #	TOPMed Project	Parent Study Short Name	TOPMed Phase	Omics Center	Omics Support	Omics Type
phs000951	COPD	COPDGene	1	NWGC	3R01HL089856-08S1	WGS
phs000951	COPD	COPDGene	2	Broad Genomics	HHSN268201500014C	WGS
phs000951	COPD	COPDGene	2.5	Broad Genomics	HHSN268201500014C	WGS
phs000951	COPD	COPDGene	4	NWGC	HHSN268201600032I	RNASeq
phs000951	COPD	COPDGene	5	NWGC	HHSN268201600032I	Methylomics

Table 1: TOPMed Omics Support Table. Broad Genomics = Broad Institute Genomics Platform; Broad Metabolomics = Broad Institute and Beth Israel Metabolomics Platform; Keck MGC = Keck Molecular Genomics Core Facility; NWGC = Northwest Genomics Center.

(Table 1) for study specific omics support information. Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

Parent Study-specific Acknowledgements:

- NHLBI TOPMed: Childhood Asthma Management Program (CAMP)
- NHLBI TOPMed: Genetic Epidemiology of COPD Study (COPDGene). The COPDGene project described was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. The COPDGene project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion. A full listing of COPDGene investigators can be found at: `http://www.copdgene.org/directory`

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The genotype data used in the simulations is available from the Partners Biobank (Partners, 2020). The summary statistics of Kunkle et al. (2019) and Jansen et al. (2019) used in the simulations are available online, see NIAGADS (2016) and CTG Lab (2021).

The data that support the findings of this study are openly available in "NHLBI TOPMed: Boston Early-Onset COPD Study in the National Heart, Lung, and Blood Institute (NHLBI) Trans-Omics for Precision Medicine (TOPMed) Program" at `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000946.v3.p1`.

References

Charlson, M., Szatrowski, T., Peterson, J., and Gold, J. (1994). Validation of a combined comorbidity index. *J Clin Epidemiol*, 47(11):1245–51.

Chen, W.-M. (2021). KING: Kinship-based INference for Gwas. `https://kingrelatedness.com/`.

CTG Lab (2021). Summary statistics for Alzheimer's dementia from Iris Jansen et al., 2019. `https://ctg.cncr.nl/software/summary_statistics`.

Falbel, D., Allaire, J., Chollet, F., RStudio, Google, Tang, Y., Bijl, W. V. D., Studer, M., and Keydana, S. (2020a). keras: R Interface to 'Keras'. R-package version 2.3.0.0: `https://cran.r-project.org/package=keras`.

Falbel, D., Allaire, J., RStudio, Tang, Y., Eddelbuettel, D., Golding, N., Kalinowski, T., and Google (2020b). tensorflow: R Interface to 'TensorFlow'. R-package version 2.2.0: `https://cran.r-project.org/package=tensorflow`.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., and Qian, J. (2020). glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. R-package version 4.0: `https://cran.r-project.org/package=glmnet`.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun*, 10:1776.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2020). PRS-CS: a polygenic prediction method that infers posterior SNP effect sizes under continuous shrinkage (CS) priors using GWAS summary statistics and an external LD reference panel. `https://github.com/getian107/PRScs`.

Hahn, G., Lutz, S., Laha, N., Cho, M., Silverman, E., and Lange, C. (2020a). A fast and efficient smoothing approach to LASSO regression and an application in statistical genetics: polygenic risk scores for Chronic obstructive pulmonary disease (COPD). *bioRxiv:2020.03.06.980953*, pages 1–20.

Hahn, G., Lutz, S. M., Laha, N., and Lange, C. (2020b). A framework to efficiently smooth L1 penalties for linear regression. *bioRxiv:2020.09.17.301788*, pages 1–35.

Hahn, G., Lutz, S. M., Laha, N., and Lange, C. (2020c). smoothedLasso: Smoothed LASSO Regression via Nesterov Smoothing. R-package version 1.4: `https://cran.r-project.org/package=smoothedLasso`.

Hahn, G., Lutz, S. M., Laha, N., and Lange, C. (2020d). smoothedLasso: Smoothed LASSO Regression via Nesterov Smoothing. R-package version 1.5: `https://cran.r-project.org/package=smoothedLasso`.

Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd Ed. Chapter 5, John Wiley and Sons, New York, NY.

Huang, H. and Darbar, D. (2017). Genetic Risk Scores for Atrial Fibrillation: Do they Improve Risk Estimation? *Can J Cardiol*, 33(4):422–424.

Jansen, I., Savage, J., Watanabe, K., Bryois, J., Williams, D., Steinberg, S., Sealock, J., Karlsson, I., Hägg, S., Athanasiu, L., Voyle, N., Proitsi, P., Witoelar, A., Stringer, S., Aarsland, D., Almdahl, I., Andersen, F., Bergh, S., Bettella, F., ..., and Posthuma, D. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genet*, 51:404–413.

Karlson, E. W., Boutin, N. T., Hoffnagle, A. G., and Allen, N. L. (2016). Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations. *J Pers Med*, 6(1):1–11.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50:1219–1224.

Kunkle, B., Grenier-Boley, B., Sims, R., Bis, J., Damotte, V., Naj, A., Boland, A., Vronskaya, M., van der Lee, S., Amlie-Wolf, A., Bellenguez, C., Frizatti, A., Chouraki, V., Martin, E., Sleegers, K., Badarinarayan, N., Jakobsdottir, J., Hamilton-Nelson, K., Moreno-Grau, S., ..., and Alzheimer Disease Genetics Consortium (ADGC); European Alzheimer's Disease Initiative (EADI); Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium (CHARGE); Genetic and Environmental Risk in AD/Defining Genetic, Polygenic and Environmental Risk for Alzheimer's Disease Consortium (GERAD/PERADES) (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet*, 51:414–430.

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., Metspalu, A., Wray, N. R., Goddard, M. E., Yang, J., and Visscher, P. M. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nature Communications*, 10(5086).

Lutz, S. M., Cho, M. H., Young, K., Hersh, C. P., Castaldi, P. J., McDonald, M.-L., Regan, E., Mattheisen, M., DeMeo, D. L., Parker, M., Foreman, M., Make, B. J., Jensen, R. L., Casaburi, R., Lomas, D. A., Bhatt, S. P., Bakke, P., Gulsvik, A., Crapo, J. D., Beaty, T. H., Laird, N. M., Lange, C., Hokanson, J. E., Silverman, E. K., ECLIPSE Investigators, and COPDGene Investigators (2015). A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genetics*, 16(138):1–11.

Mak, T., Porsch, R., Choi, S., Zhou, X., and Sham, P. (2017). Polygenic scores via penalized regression on summary statistics. *Genet Epidemiol*, 41(6):469–480.

Mak, T., Porsch, R., Choi, S., Zhou, X., and Sham, P. (2020). Lassosum: a method for computing LASSO/Elastic Net estimates of a linear regression problem given summary statistics from GWAS and Genome-wide meta-analyses. `https://github.com/tshmak/lassosum`.

Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., Veldink, J., ..., and Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*, 48(10):1279–83.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Math. Program. Ser. A*, 103:127–152.

NHLBI TOPMed (2018). Boston Early-Onset COPD Study in the National Heart, Lung, and Blood Institute (NHLBI) Trans-Omics for Precision Medicine (TOPMed) Program.

NIAGADS (2016). NG00075 - IGAP Rare Variant Summary Statistics - Kunkle et al. (2019). `https://www.niagads.org/datasets/ng00075`.

Partners (2020). Partners healthcare biobank. `https://biobank.partners.org`.

Privé, F., Arbel, J., and Vilhjálmsson, B. J. (2019). LDpred2: better, faster, stronger. *Bioinformatics*, btaa1029.

Privé, F., Blum, M., and Aschard, H. (2020). bigsnpr: Analysis of Massive SNP Arrays. R-package version 1.5.2: `https://cran.r-project.org/package=bigsnpr`.

Purcell, S. and Chang, C. (2020). PLINK2 (v2.00, 31 Aug 2020). `www.cog-genomics.org/plink/2.0/`.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Stat Comp, Vienna, Austria.

Regan E.A., e. a. (2010). Genetic epidemiology of COPD (COPDGene) study design. *COPD*, 7(7):32–43.

Silverman, E., Chapman, H., Drazen, J., Weiss, S., Rosner, B., Campbell, E., O'Donnell, W., Reilly, J., Ginns, L., Mentzer, S., Wain, J., and Speizer, F. (1998). Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease; Risk to relatives for airflow obstruction and chronic bronchitis. *Am J Respir Crit Care Med*, 157(6):1770–8.

Silverman, E., Weiss, S., Drazen, J., Chapman, H., Carey, V., Campbell, E., Denish, P., Silverman, R., Celedon, J., Reilly, J., Ginns, L., and Speizer, F. (2000). Gender-related differences in severe, early-onset chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*, 162(6):2152–8.

Speed, D. (2021). Megaprs. `http://dougspeed.com/prediction/`.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *J Roy Stat Soc B Met*, 58(1):267–288.

Vilhjálmsson, B., Yang, J., Finucane, H., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P., Bhatia, G., Do, R., Hayeck, T., Won, H., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study, Kathiresan, S., Pato, M., Pato, C., Tamimi, R., Stahl, E., Zaitlen, N., ..., and Price, A. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet*, 97(4):576–592.

Wand, H., Lambert, S. A., Tamburro, C., Iacocca, M. A., O'Sullivan, J. W., Sillari, C., Kullo, I. J., Rowley, R., Dron, J. S., Brockman, D., Venner, E., McCarthy, M. I., Antoniou, A. C., Easton, D. F., Hegele, R. A., Khera, A. V., Chatterjee, N., Kooperberg, C., Edwards, K., ..., and Wojcik, G. (2020). Improving reporting standards for polygenic scores in risk prediction studies. *bioRxiv:2020.04.23.20077099*, pages 1–19.

Ware, E. B., Faul, J. D., Mitchell, C. M., and Bakulski, K. M. (2020). Considering the APOE

locus in Alzheimer's disease polygenic scores in the Health and Retirement Study: a longitudinal panel study. *BMC Medical Genomics*, 13(164):1–13.

World Health Organization (2021). International Statistical Classification of Diseases and Related Health Problems (ICD). `https://www.who.int/standards/classifications/classification-of-diseases`.

Zeng, J., Yang, J., Zhang, F., Zheng, Z., Lloyd-Jones, L., and Goddard, M. (2020). GCTB: A tool for Genome-wide Complex Trait Bayesian analysis. `https://cnsgenomics.com/software/gctb/#Overview`.

Zhang, Q., Privé, F., Vilhjálmsson, B., and Speed, D. (2021). Improved genetic prediction of complex traits from individual-level data or summary statistics. *bioRxiv:2020.08.24.265280*, pages 1–15.

Zhang, Q., Sidorenko, J., Couvy-Duchesne, B., Marioni, R. E., Wright, M. J., Goate, A. M., Marcora, E., lin Huang, K., Porter, T., Laws, S. M., Australian Imaging Biomarkers and Lifestyle (AIBL) Study, Sachdev, P. S., Mather, K. A., Armstrong, N. J., Thalamuthu, A., Brodaty, H., Yengo, L., Yang, J., Wray, N. R., McRae, A. F., and Visscher, P. M. (2020). Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. *Nature Communications*, 11(4799):1–11.

A Principal component plots

Figures 4 and 5 show the first eight principal components of the HRC-imputed genotype data downloaded from Partners Biobank. All individuals we kept in the dataset are self-reported non-hispanic white (NHW) individuals. We excluded outliers which are at least 5 standard deviations away from the mean value of each of the ten principal components. In Figure 4 we observe a negligible amount of stratification based on the genotyping chip, but given the even distribution of cases/controls across chips displayed in Figure 5, this should not affect the results.

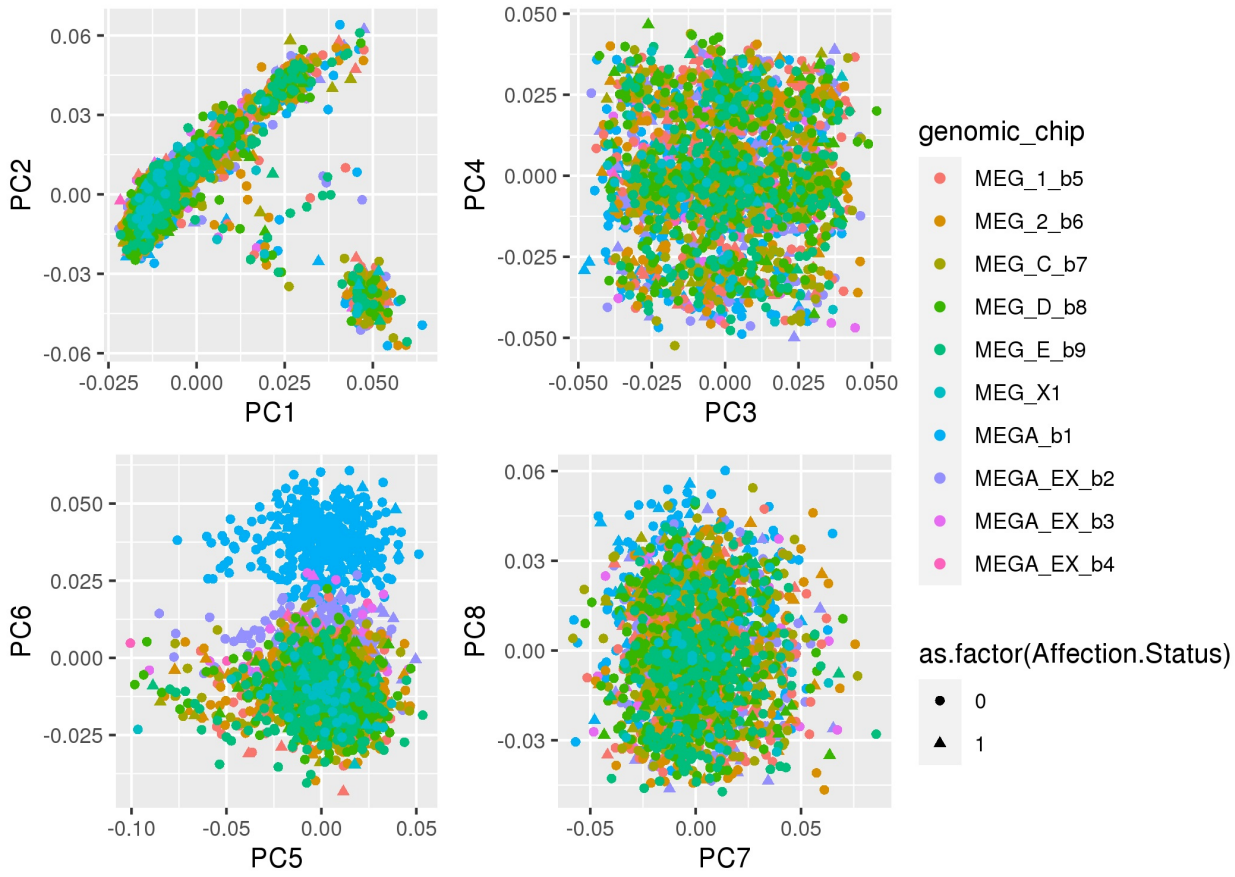

Figure 4: First eight principal components of the HRC-imputed genotype data downloaded from Partners Biobank. Stratification by genomic chip and affection status.

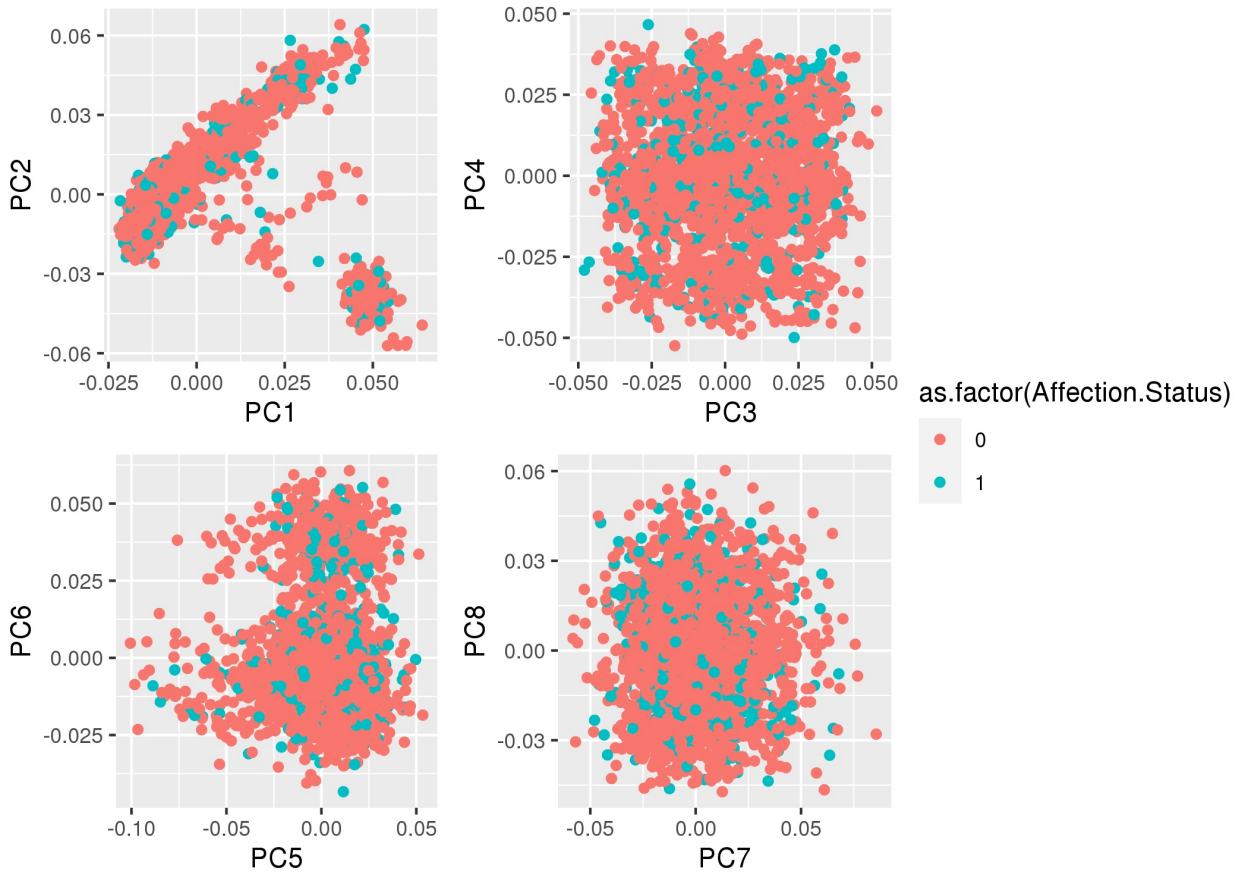

Figure 5: First eight principal components of the HRC-imputed genotype data downloaded from Partners Biobank. Stratification by affection status.