genes

Article

A Smoothed Version of the Lassosum Penalty for Fitting Integrated Risk Models Using Summary Statistics or Individual-Level Data

Georg Hahn [1,*], Dmitry Prokopenko [2], Sharon M. Lutz [1], Kristina Mullin [2], Rudolph E. Tanzi [2], Michael H. Cho [3], Edwin K. Silverman [3], Christoph Lange [1] and on the behalf of the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium [†]

[1] Harvard T.H. Chan School of Public Health, Harvard University, 677 Huntington Ave, Boston, MA 02115, USA; sharon.lutz@channing.harvard.edu (S.M.L.); clange@hsph.harvard.edu (C.L.)

[2] Genetics and Aging Research Unit, McCance Center for Brain Health, Department of Neurology, Massachusetts General Hospital, Boston, MA 02114, USA; dprokopenko@mgh.harvard.edu (D.P.); kmullin1@mgh.harvard.edu (K.M.); tanzi@helix.mgh.harvard.edu (R.E.T.);

[3] Department of Medicine, Brigham and Women's Hospital, Harvard University, Boston, MA 02115, USA; remhc@channing.harvard.edu (M.H.C.); reeks@channing.harvard.edu (E.K.S.)

[*] Correspondence: ghahn@hsph.harvard.edu

[†] Membership of the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium is provided in the Acknowledgments.

Citation: Hahn, G.; Prokopenko, D.; Lutz, S.M.; Mullin, K.; Tanzi, R.E.; Cho, M.H.; Silverman, E.K.; Lange, C.; the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium A Smoothed Version of the Lassosum Penalty for Fitting Integrated Risk Models Using Summary Statistics or Individual-Level Data. *Genes* **2022**, *13*, 112. https://doi.org/10.3390/genes13010112

Academic Editor: Zhongxue Chen

Received: 11 November 2021
Accepted: 4 January 2022
Published: 6 January 2022

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 using either summary statistics or raw data. 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 (for the datasets under consideration), 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 [1]. 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.

As the potential for broad-scale clinical use to identify people at high risk for certain diseases has been demonstrated [2], polygenic risk scores and integrated risk models have become widespread tools for the early identification of patients who are at high risk for a certain disease and who could benefit from intervention measures [3–5]. 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 in [6]), varies substantially across application areas. 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 [7], while for atrial fibrillation the AUC is around 0.64 [8], a value which is considered less than acceptable [6,9]. For this reason, increasing the accuracy of scores is desirable, which is the focus of the proposed smoothing approach.

One popular way to fit a polygenic risk score is the "Lassosum" approach of the authors of [7]. Note that in [7], 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. As in [7], we assume in this article that estimates of the correlations $\boldsymbol{r}$ can be obtained from publicly available summary statistics databases, and that estimates of the LD matrix $X_r^\top X_r$ are obtained from publicly available genotype databases (such as the 1000 Genomes Project). However, the Lassosum objective function can also be used to compute a polygenic risk score using raw data. Importantly, in [7] the authors derive an iterative scheme to carry out the minimization of Equation (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 Equation (1). Using the methodology in [10], we propose to smooth the non-differentiable L_1 penalty in Equation (1), thus allowing us to compute explicit gradients of Equation (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 [7] in that we do not solely consider polygenic risk scores, but the more general integrated risk models. The proposed smoothed Lassosum can be applied to either summary statistics (when using X and $\boldsymbol{r}$ as previously described), as well as individual-level data (when using X and $\boldsymbol{y}$ directly in either the Lasso or Lassosum objective function; in the latter case, $\boldsymbol{y}$ is converted into "correlations" $\boldsymbol{r}$ via $\boldsymbol{r} = X^\top \boldsymbol{y}$).

Our approach follows as a special case from in [11,12], 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 in [11]. 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 [7].

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 [13,14], and one on COPD (chronic obstructive pulmonary disease) using individual-level spirometry data [15]. In the first case, the endpoint is binary, whereas in the second study the endpoint is continuous. Our experiments demonstrate that smoothing the Lassosum objective function results in a considerably enhanced performance of the Lassosum approach for the datasets we consider, allowing it to draw equal with approaches such as LDpred2 [16] or PRScs [17].

Analogously to the original Lasso of [18], the L_1 penalty employed in Equation (1) causes some entries of $\arg\min_{\beta \in \mathbb{R}^p} L(\boldsymbol{\beta})$ 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. A literature review is given in Section 1.1. 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 closes with a discussion in Section 4, and some final remarks are given in the conclusions of Section 5. 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 the function *prsLasso* in the package), available on CRAN [19].

1.1. Literature Review

Several methodological approaches have been considered in the literature to compute a polygenic risk score or an integrated risk model for a given population [20], and to predict a given outcome (disease status).

A simple way to calculate a polygenic risk score is to threshold p-values coming from GWAS summary statistics. If all genetic markers are used, we speak of an unadjusted polygenic risk score. However, if SNPs in linkage disequilibrium (LD) with each other are included in the score, their contribution will be exaggerated, thus making informed LD-pruning of single nucleotide polymorphisms (SNPs) in LD necessary [21]. The selective removal of less significantly related SNPs to reduce LD is called LD-clumping [22]. Such approaches are computationally simple and fast, but have limited accuracy [23]. However, the (optimal) choice of the threshold is an issue, as this determines the number of SNPs to be included [22]. As a result, scores are often constructed for a variety of thresholds [7,24].

Accuracy can be increased by incorporating GWAS summary statistics via Bayesian methods. Notable approaches include LDpred [23] and LDpred2 of [16], which compute a polygenic risk score (but not an integrated risk model) by fitting a Bayesian model to given effect sizes via Gibbs sampling. A score is then obtained by inferring the posterior mean effect size of each marker using a prior on the effect sizes and LD information from an external reference panel. Using a normal mixture model offers enhanced flexibility and accuracy through the incorporation of genome-wide markers and different genetic architectures [25,26]. One weakness of Bayesian methods consists in the choice of the required discrete mixture priors on SNP effect sizes, potentially causing computational issues and inaccurate adjustment for local LD patterns.

PRScs of [27] utilizes a high-dimensional Bayesian regression framework which places a continuous shrinkage prior (thus the suffix *CS* for continuous shrinkage) on SNP effect sizes, an innovation which makes a conjugate block update of the SNP effect sizes in posterior inference possible and which is robust to varying genetic architectures.

SBayesR in [26] is 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 main innovation of MegaPRS [28] consists in the fact that it allows the user to specify how SNPs contribute toward the phenotype. This is done via the specification of a heritability model, which describes how the expected heritability contributed by each SNP varies across the genome. In contrast to current tools which assume that the expected heritability per SNP is constant, the authors show in [28] that realistic heritability models can result in more accurate polygenic risk scores.

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) in [18], for instance, using the glmnet package on CRAN, see in [29,30]. Glmnet is a fast variant of the FISTA proximal gradient algorithm, the current gold standard for minimizing the Lasso objective function [31]. Glmnet is almost identical to FISTA, but performs a cyclic update of all coordinates, whereas FISTA updates all coordinates per iteration, thus making Glmnet faster than FISTA.

More favorable scaling of polygenic risk score computations (in the size of the input data) has also been a focus in the recent literature [32]. Importantly, machine learning has become increasingly popular for constructing polygenic risk scores [33–37], as machine learning approaches do not assume SNP independence or near independence. However, the resulting prediction model cannot be easily interpreted, in contrast to the linear weighting schemes computed by traditional methods. Examples of traditional approaches outperforming machine learning models are also available in the literature [38].

2. Methodology

The Lassosum function of Equation (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 [11,12]. 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 [10], the author introduces a framework to smooth a piecewise affine and convex function $f : \mathbb{R}^q \to \mathbb{R}$, where $q \in \mathbb{N}$. As f is piecewise affine, it can be written for $\boldsymbol{z} \in \mathbb{R}^q$ as

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

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 Equation (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 Equation (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 $Q_k = \left\{ \boldsymbol{w} \in \mathbb{R}^k : \sum_{i=1}^k \boldsymbol{w}_i = 1, \boldsymbol{w}_i \geq 0 \; \forall i = 1, \dots, k \right\} \subseteq \mathbb{R}^k$ 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 non-negative, 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 ([10], Theorem 1). Though several choices of the prox-function ρ are considered in [10], 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 Equation (3) given by

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

which satisfies the uniform bound

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

see [10–12].

2.2. A Smoothed Version of the Lassosum Objective Function

The proposed smoothed Lassosum approach is obtained by applying Nesterov smoothing to the L_1 penalty of the Lassosum objective function, see Equation (1). A detailed study on the behavior of Nesterov smoothing applied to an L_1 penalty using synthetic data can be found in [11].

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 Equation (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 Equation (1) with the approximation in Equation (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 in [11,12], from which the closed-form gradient of the smoothed Lassosum objective function of Equation (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 byL^{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 function optim in R [39].

In Equation (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 Equation (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 in [7] 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 Equation (5) can be re-written as

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

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

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

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 Equation (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 in [7], the Lassosum objective of Equation (1) is equivalent to a Lasso problem, in particular its convexity is preserved. According to Proposition 2 in [11], the smooth approximation of Equation (7) obtained via Nesterov smoothing is strictly convex. As strictly convex functions have one unique minimum, and as 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 Equation (7) can be derived from ([11], 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 ([11], 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 ([11], Proposition 5).

3. Application to Experimental Data

In this section, we evaluate 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) using summary statistics, and one fitting an integrated risk model to continuous outcomes using individual-level data 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. "Lassosum": the Lassosum algorithm of [7], implemented in the R package *lassosum* available on github [40].
2. "LDpred2": the LDpred2 algorithm of [16], implemented in the R package *bigsnpr* on CRAN [41].
3. "PRScs": the PRScs algorithm of [27], available on github [17].

4. “Glmnet”: the standard lasso of [18], implemented in the *Glmnet* package on CRAN [30].
5. “Lasso”: the unsmoothed Lasso of [12], implemented in the R package *smoothedLasso* on CRAN [42].
6. “SmoothedLasso”: the smoothed Lasso of [12], implemented in the R package *smoothedLasso* on CRAN [42]. Both the unsmoothed and smoothed Lasso are included in the experiments to showcase how the unsmoothed (original) and smoothed Lasso compare.
7. “NeuralNetwork”: a neural network implemented with the *Keras* interface [43] to the *Tensorflow* machine learning platform [44]. 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.
8. “SBayesR”: the SBayesR algorithm of [26], implemented in the toolbox *GCTB* [45].
9. “MegaPRS”: we employ the robust version *Bolt Predict* of the MegaPRS algorithm [28] as suggested by the authors. We use default parameters given in the example section of the MegaPRS website (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 [46].
10. “EpiOnly”: we perform a simple linear regression using epidemiological covariates only.

Unless noted otherwise, all aforementioned methods are run with default parameters. 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 Equation (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 $\lambda = 2^{-3}$. This value was chosen in a data-driven way to ensure that the resulting estimates are not too dense (which happens if the regularization parameter is too small), or zero (which happens if the regularization parameter is too large). The Lassosum regularization parameter s in Equation (1) (which ensures stability and uniqueness of solution) was chosen as $s = 0.5$ as recommended in Section 3 of [7], and the smoothing parameter of Section 2.2 was chosen as $\mu = 0.1$, see Section 3 of [12].

3.1. Alzheimer’s Disease 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 in [47]. The HRC-imputed genotype data was downloaded from Partners Biobank [48] (described below). The summary statistics are matched to genotype data for chromosomes 1–22 of 2465 patients available in the Partners Biobank. We considered two sets of summary statistics from two of the largest available AD GWAS: the one of clinically defined AD cases of [13], and the one of AD-by-proxy phenotypes of [14].

The dataset in [13] contains a total of 11,480,632 summary statistics, given by p-value, effect size (denoted as variable "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 in [14] contains a total of 13,367,299 summary statistics in the same format as the one in [13].

Partners Biobank is a hospital-based cohort from the Mass General 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 in [49]. 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 [50,51].

We performed the following quality control steps on the HRC-imputed genotype data from Partners Biobank. Relatedness was assessed with KING [52,53] 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 > 0.0442 (third degree of relatedness or closer) 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 Appendix A). There was a total of 2465 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–thin-count option in PLINK2 [54]. The precise number of 20,000 loci is arbitrary, and was chosen to include a large number of loci while still being able to run all simulations in reasonable time. Although *APOE* variants are known to have a very high effect size for AD, explaining around a quarter of the total heritability [55], 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* [56]. 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 to 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 consist 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). As the data do not exhibit a separation by genomic chip (see Figure A1) we did not include principal components into the model. However, we recommend doing so if a clear separation in the principal component plots is visible.

In the following experiments, we considered the datasets of [13,14] separately and extracted SNP weights based on corresponding effect sizes. This gives us the three quantities $X, \boldsymbol{r}, \boldsymbol{y}$ required to fit the Lassosum model of Equation (1). Next, we only consider a proportion $p \in \{0.1, \ldots, 0.9\}$ of the pool of indices of the Partners genotyped subjects as a training dataset (selected uniformly at random), that is a proportion p of the rows of X and corresponding entries of $\boldsymbol{y}$ ($\boldsymbol{r}$ is updated using X and $\boldsymbol{y}$) and fit an integrated risk model using Equation (1) to these training subjects with the aforementioned methods. In the case of the neural network, we use the training dataset to tune its hyperparameters. Finally, we evaluate the performance of all methods on the unseen proportion $1 - p$ of the data, that is we compute an estimate of the outcome $\boldsymbol{y}$ with the help of Equation (1) on the unseen data, and compare the outcome estimate to the true outcomes. 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 [13]. 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.

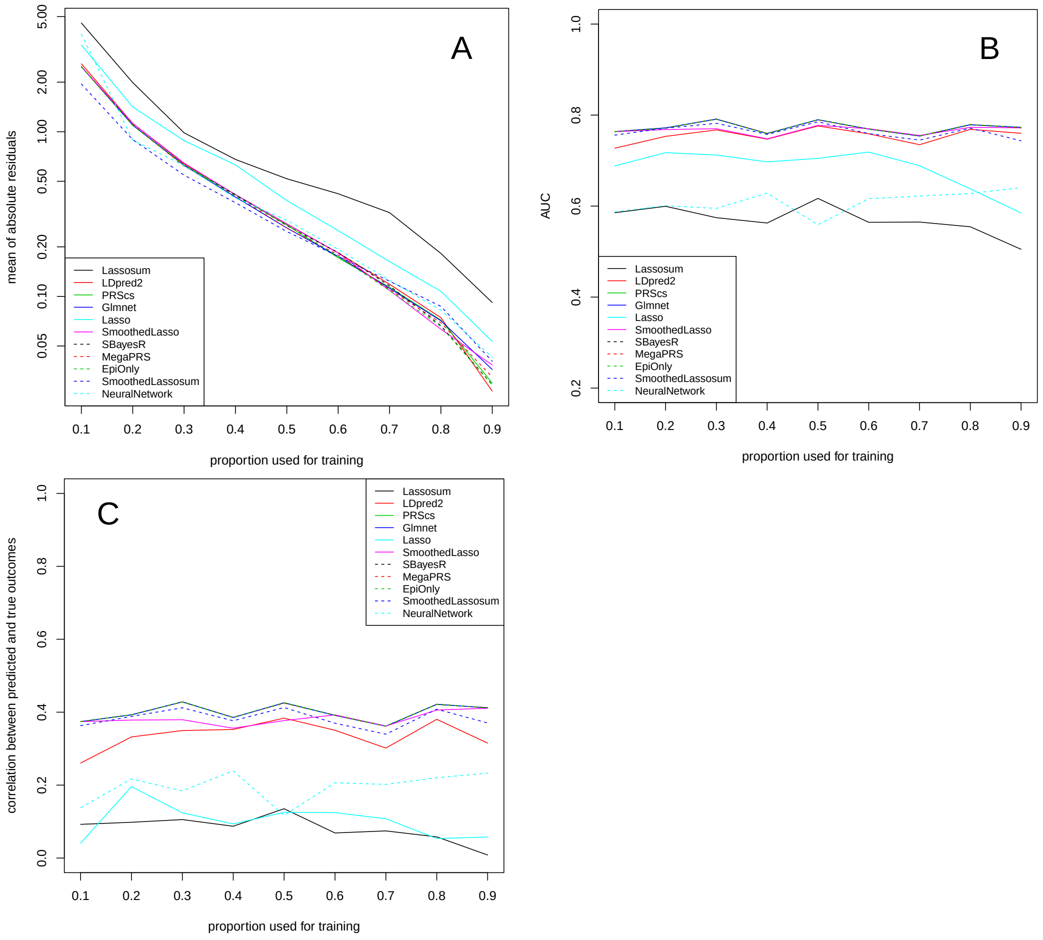

Figure 1. Dataset of clinically defined AD cases of [13]. Mean of absolute residuals (**A**), AUC (**B**), and correlation between predicted and true outcomes (**C**) as a function of the proportion of data used for training. Plotted with jittering.

Second, the AUC is very high (reaching almost 0.80) for all methods apart from Lassosum, Lasso, 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. This is in line with previous observations that the AUC is invariant to the prior class probabilities [57]. 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. This is likely a result of the binary cross-entropy loss, which implicitly tunes the behavior towards low residuals. Tensorflow allows for the specification of other loss functions (such as the mean absolute error loss or AUC), though for a binary response the binary cross-entropy loss is a natural choice. 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 can be explained with the observation that neural networks typically have many more parameters than conventional methods and thus traditionally require larger amounts of data to be trained on. For instance, the number of nodes per layer, the activation function, dropout rate, etc. per layer all depend on how many layers were chosen in the first place can be freely tuned, thus quickly resulting in large numbers of parameters.

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 [7], 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 epidemiological covariates (see Section 4 for more details).

The results for the dataset of [14], reported in Figure 2, are almost identical to the ones for the dataset of [13] 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.

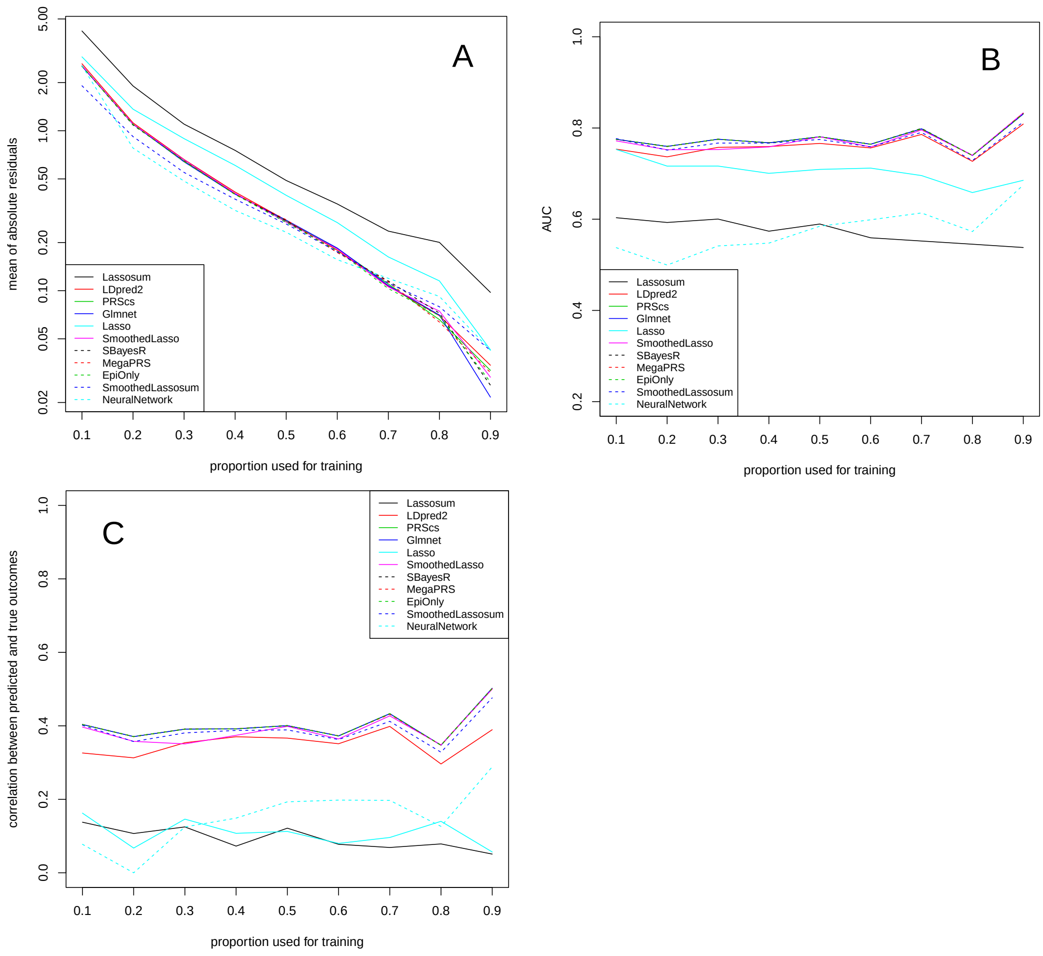

Figure 2. Dataset of AD-by-proxy phenotypes of [14]. Mean of absolute residuals (**A**), AUC (**B**), and correlation between predicted and true outcomes (**C**) as a function of the proportion of data used for training. Plotted with jittering.

The similarity between Figures 1 and 2 is expected. The two experiments differ only in the way the response (AD status) is defined. The response provided in [13] consists of clinically defined AD cases, while the one in [14] 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 datasets considered in Section 3.1 are 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 in [15], a case–control study of COPD in current and former smokers which has been sequenced as part of the TOPMed Project.

The dataset we consider contains TOPMed WGS data of smokers with COPD, selected as having an age at enrollment of 45–80 years, a smoking history of at least 10 pack-years, non-Hispanic White or non-Hispanic African American descent, and a diagnosis of COPD Stages 2, 3, and 4 by GOLD criteria (post-bronchodilator FEV1/FVC < 0.70 and FEV1 $< 80\%$ predicted), where FEV1 is defined as the air volume in liters a person can exhale during the first second of a forced expiration, and FEV1/FVC (also called Tiffeneau–Pinelli index) is the proportion of a person's vital capacity that they are able to expire in the first second of forced expiration (FEV1) to the full forced vital capacity (FVC), see in [58]. We focus on chromosome 15 and consider the risk loci for spirometric measures which have been identified in [59]. Overall, we consider 8881 loci of 3495 individuals.

We aim to predict the raw FEV1 value from the WGS data and four epidemiological covariates, that is, in this section we fit an integrated risk model using individual-level data only. The final data used for the computation of the integrated risk models consists of the raw FEV1 value (the quantity $\boldsymbol{y}$ in Equation (1)) and the 8881 loci plus age, sex, pack-years of smoking, and height (the quantity X in Equation (1) from which $\boldsymbol{r} = X^{\top}\boldsymbol{y}$ can be computed). As in Section 3.1, we use a classic training (proportion $p \in (0,1)$) and validation (proportion $1 - p$) setup. Precisely, we only consider a randomly drawn pool of proportion p of the rows of X and corresponding entries of $\boldsymbol{y}$ for fitting the integrated risk model using Equation (1). After fitting, we compute an estimated outcome by evaluating Equation (1) on the unseen rows of X and entries of y, allowing us to compare predicted and true outcomes. 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 3A 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 better, showing a good and robust performance throughout all training proportions, although they do not reach the performance of methods such as LDpred2 or PRScs. 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 3B, 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 (see Section 4 for more details).

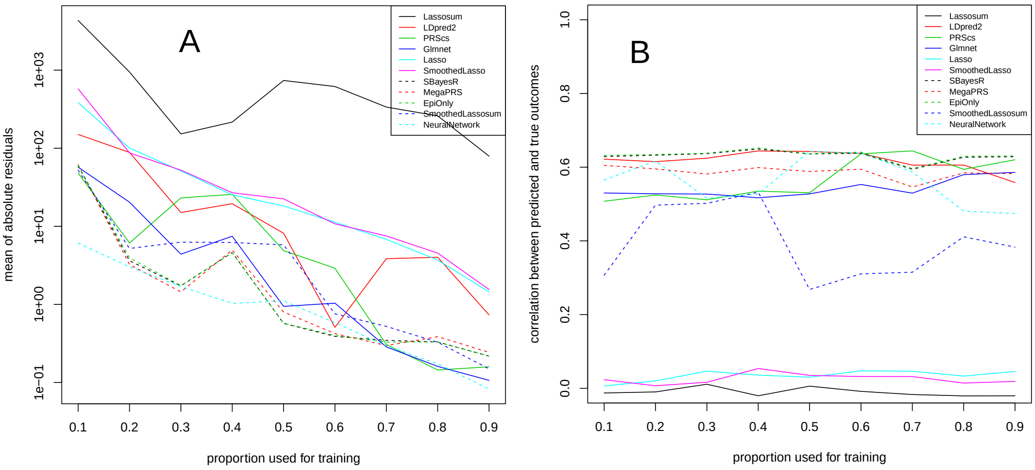

Figure 3. Dataset of the COPD study of [15]. Mean of absolute residuals (**A**) and correlation between predicted and true outcomes (**B**) as a function of the proportion of data used for training.

4. Discussion

This article considered the calculation of an integrated risk model by minimizing a smoothed version of the Lassosum objective function (see Equation (1)) introduced in [7]. 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. Our proposed smoothed Lassosum approach can be applied to both summary statistics and individual-level data.

An experimental study on Alzheimer's disease and COPD demonstrates that our smoothed Lassosum improves upon the original Lassosum of [7], 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 (for the datasets under consideration). The reduced performance of Lassosum we observe in the real data applications 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), which occurs, for instance, when regressing out epidemiological covariates and using the residuals as input to Lassosum. In particular, in its original formulation in [7], Lassosum only considers genotype data X, and the incorporation of additional covariates is not possible in the Lassosum R package [40]. Moreover, although recomputing the SNP-wise correlation $\boldsymbol{r} = X^\top \boldsymbol{y}$ in Equation (1) and using them in place of $\boldsymbol{y}$ is a valid approach, 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. For instance, it is not guaranteed any more that each entry of $\boldsymbol{r}$ lays in the open interval $(-1, +1)$ for arbitrary y, and it is not straightforward how to transform the input to comply with this condition. In contrast, our smoothed Lassosum works well for both epidemiological covariates and continuous responses.

Using an L_1 penalty in Equation (1) has the advantage that, in analogy to the original Lasso of [18], 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 Equation (1) satisfying $|\beta_i| < \tau$ for some threshold τ are set to zero, although this might also cause a decrease in predictive performance [60]. Determining an optimal threshold, as well as the trade-offs incurred compared to working with dense polygenic risk scores, remains for future research.

5. Conclusions

We observe that for the prediction of Alzheimer's disease and COPD, computing an integrated risk model involving genetic information provides little benefit in addition to using epidemiological covariates only. In [61], the authors show that the odds ratio achieved by current polygenic risk scores is too small to warrant their usage as a screening method, and that it would be equally sensible to offer the intervention regardless, given it is effective and inexpensive. In the case of Alzheimer's disease and COPD, this means that the usage of an integrated risk model is only sensible for costly treatments.

Additional genotype data used in the simulations is available from the Partners Biobank [48]. The summary statistics in [13,14] used in the simulations are available online, see in [62,63].

Author Contributions: G.H. conducted all experiments and took care of writing the manuscript. D.P. supplied the datasets as well as the figures in the appendix. S.M.L., K.M., R.E.T., M.H.C., and E.K.S. gave technical advice. C.L. proofread the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The methodology work in this paper was funded by Cure Alzheimer's Fund. The COPDGene project was supported by Award Number R01 HL089856 and Award Number R01 HL089897 from the National Heart, Lung, and Blood Institute.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Genome Sequencing for "NHLBI TOPMed: Genetic Epidemiology of COPD (COPDGene) Funded by the National Heart, Lung, and Blood Institute (NHLBI) in the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program" (phs000951.v3.p3) was performed at the Northwest Genomics Center (HHSN268201600032I, 3R01HL089856-08S1, HHSN268201600032I) as well as the Broad Institute Genomics Platform (HHSN 268201500014C, HHSN268201500014C), see Table A1. 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.

Acknowledgments: For COPDGene Phase 3, Grant Support and Disclaimer: The 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. COPD Foundation Funding: COPDGene is also supported by the COPD Foundation through contributions made to an Industry Advisory Board that has included AstraZeneca, Bayer Pharmaceuticals, Boehringer-Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer, and Sunovion. COPDGene Investigators—Core Units: *Administrative Center*: James D. Crapo; Edwin K. Silverman; Barry J. Make; Elizabeth A. Regan. *Genetic Analysis Center*: Terri H. Beaty; Peter J. Castaldi; Michael H. Cho; Dawn L. DeMeo; Adel El Boueiz; Marilyn G. Foreman; Auyon Ghosh; Lystra P. Hayden; Craig P. Hersh; Jacqueline Hetmanski; Brian D. Hobbs; John E. Hokanson; Wonji Kim; Nan Laird; Christoph Lange; Sharon M. Lutz; Merry-Lynn McDonald; Dmitry Prokopenko; Matthew Moll; Jarrett Morrow; Dandi Qiao; Elizabeth A. Regan; Aabida Saferali; Phuwanat Sakornsakolpat; Edwin K. Silverman; Emily S. Wan; Jeong Yun. *Imaging Center*: Juan

Pablo Centeno; Jean-Paul Charbonnier; Harvey O. Coxson; Craig J. Galban; MeiLan K. Han; Eric A. Hoffman, Stephen Humphries; Francine L. Jacobson; Philip F. Judy; Ella A. Kazerooni; Alex Kluiber; David A. Lynch; Pietro Nardelli; John D. Newell, Jr.; Aleena Notary; Andrea Oh; Elizabeth A. Regan; James C. Ross; Raul San Jose Estepar; Joyce Schroeder; Jered Sieren; Berend C. Stoel; Juerg Tschirren; Edwin Van Beek; Bram van Ginneken; Eva van Rikxoort; Gonzalo Vegas Sanchez-Ferrero; Lucas Veitel; George R. Washko; Carla G. Wilson. *PFT QA Center, Salt Lake City, UT*: Robert Jensen. *Data Coordinating Center and Biostatistics, National Jewish Health, Denver, CO*: Douglas Everett; Jim Crooks; Katherine Pratte; Matt Strand; Carla G. Wilson. *Epidemiology Core, University of Colorado Anschutz Medical Campus, Aurora, CO*: John E. Hokanson; Erin Austin; Gregory Kinney; Sharon M. Lutz; Kendra A. Young. *Mortality Adjudication Core*: Surya P. Bhatt; Jessica Bon; Alejandro A. Diaz; MeiLan K. Han; Barry Make; Susan Murray; Elizabeth Regan; Xavier Soler; Carla G. Wilson. *Biomarker Core*: Russell P. Bowler; Katerina Kechris; Farnoush Banaei-Kashani. COPDGene Investigators—Clinical Centers: *Ann Arbor VA*: Jeffrey L. Curtis; Perry G. Pernicano. *Baylor College of Medicine, Houston, TX*: Nicola Hanania; Mustafa Atik; Aladin Boriek; Kalpatha Guntupalli; Elizabeth Guy; Amit Parulekar. *Brigham and Women's Hospital, Boston, MA*: Dawn L. DeMeo; Craig Hersh; Francine L. Jacobson; George Washko. *Columbia University, New York, NY*: R. Graham Barr; John Austin; Belinda D'Souza; Byron Thomashow. *Duke University Medical Center, Durham, NC*: Neil MacIntyre, Jr.; H. Page McAdams; Lacey Washington. *HealthPartners Research Institute, Minneapolis, MN*: Charlene McEvoy; Joseph Tashjian. *Johns Hopkins University, Baltimore, MD*: Robert Wise; Robert Brown; Nadia N. Hansel; Karen Horton; Allison Lambert; Nirupama Putcha. *Lundquist Institute for Biomedical Innovation at Harbor UCLA Medical Center, Torrance, CA*: Richard Casaburi; Alessandra Adami; Matthew Budoff; Hans Fischer; Janos Porszasz; Harry Rossiter; William Stringer. *Michael E. DeBakey VAMC, Houston, TX*: Amir Sharafkhaneh; Charlie Lan. *Minneapolis VA*: Christine Wendt; Brian Bell; Ken M. Kunisaki. *Morehouse School of Medicine, Atlanta, GA*: Eric L. Flenaugh; Hirut Gebrekristos; Mario Ponce; Silanath Terpenning; Gloria Westney. *National Jewish Health, Denver, CO*: Russell Bowler; David A. Lynch. *Reliant Medical Group, Worcester, MA*: Richard Rosiello; David Pace. *Temple University, Philadelphia, PA*: Gerard Criner, MD; David Ciccolella; Francis Cordova; Chandra Dass; Gilbert D'Alonzo; Parag Desai; Michael Jacobs; Steven Kelsen; Victor Kim; A. James Mamary; Nathaniel Marchetti; Aditi Satti; Kartik Shenoy; Robert M. Steiner; Alex Swift; Irene Swift; Maria Elena Vega-Sanchez. *University of Alabama, Birmingham, AL*: Mark Dransfield; William Bailey; Surya P. Bhatt; Anand Iyer; Hrudaya Nath; J. Michael Wells. *University of California, San Diego, CA*: Douglas Conrad; Xavier Soler; Andrew Yen. *University of Iowa, Iowa City, IA*: Alejandro P. Comellas; Karin F. Hoth; John Newell, Jr.; Brad Thompson. *University of Michigan, Ann Arbor, MI*: MeiLan K. Han; Ella Kazerooni; Wassim Labaki; Craig Galban; Dharshan Vummidi. *University of Minnesota, Minneapolis, MN*: Joanne Billings; Abbie Begnaud; Tadashi Allen. *University of Pittsburgh, Pittsburgh, PA*: Frank Sciurba; Jessica Bon; Divay Chandra; Joel Weissfeld. *University of Texas Health, San Antonio, San Antonio, TX*: Antonio Anzueto; Sandra Adams; Diego Maselli-Caceres; Mario E. Ruiz; Harjinder Singh.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Principal Component Plots

Figures A1 and A2 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 A1, 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 A2; this should not affect the results. Figure A3 shows a projection of our dataset onto the reference 1000 Genomes populations with *bigsnpr*, indicating that our dataset is clustered within the European population of 1000 Genomes.

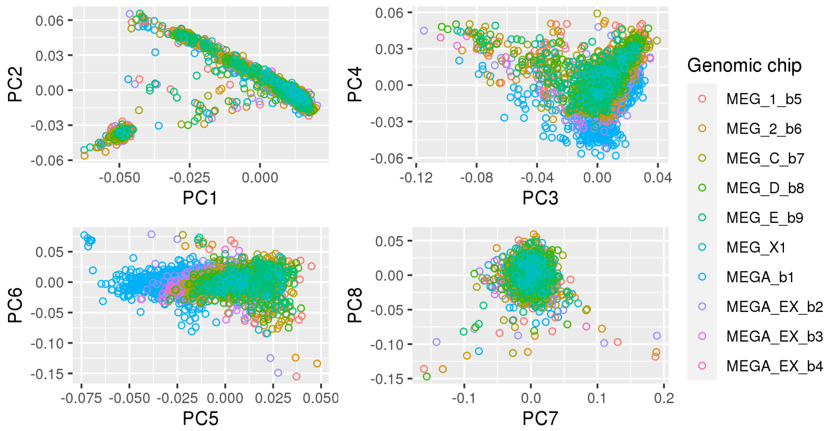

Figure A1. First eight principal components of the HRC-imputed genotype data downloaded from Partners Biobank. Stratification by genomic chip.

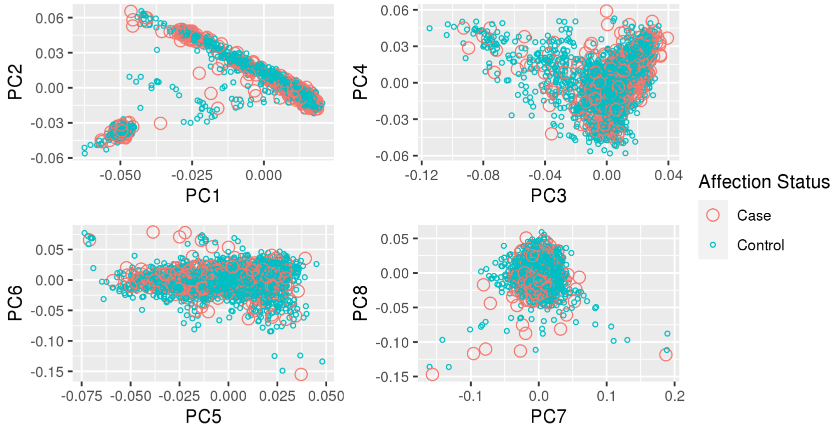

Figure A2. First eight principal components of the HRC-imputed genotype data downloaded from Partners Biobank. Stratification by affection status.

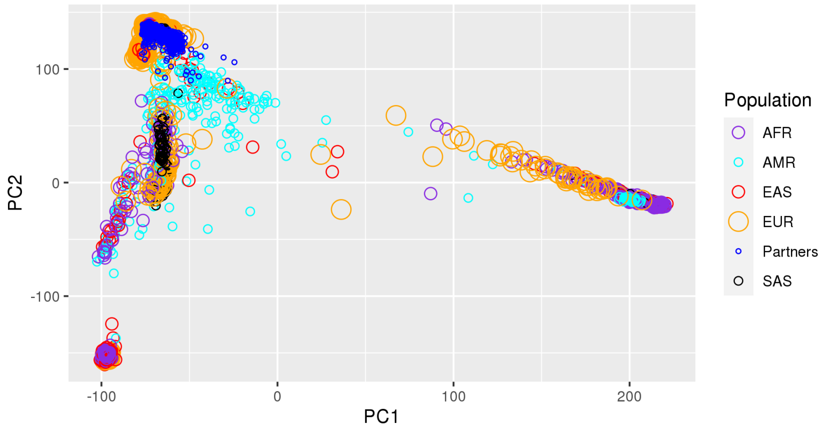

Figure A3. Using *bigsnpr* we have projected our dataset onto the reference 1000 Genomes populations. As the readers can see now, our dataset is clustered within the European population of 1000 Genomes. The population acronyms are: AFR (Africa), AMR (America), EAS (East Asia), EUR (Europe), SAS (South Asian).

Table A1. TOPMed Omics Support Table. Broad Genomics = Broad Institute Genomics Platform. NWGC = Northwest Genomics Center.

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

References

1. 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.; et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature* **2021**, *591*, 211–219. [CrossRef]
2. Khera, A.V.; Chaffin, M.; Aragam, K.G.; Haas, M.E.; Roselli, C.; Choi, S.H.; Natarajan, P.; Lander, E.S.; Lubitz, S.A.; Ellinor, P.T.; et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **2018**, *50*, 1219–1224. [CrossRef]
3. Duncan, L.; Shen, H.; Gelaye, B.; Meijsen, J.; Ressler, K.; Feldman, M.; Peterson, R.; Domingue, B. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **2019**, *10*, 3328. [CrossRef]
4. Knowles, J.; Ashley, E. Cardiovascular disease: The rise of the genetic risk score. *PLoS Med.* **2018**, *15*, e1002546. [CrossRef] [PubMed]
5. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **2014**, *511*, 421–427. [CrossRef]
6. Mandrekar, J.N. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *J. Thorac. Oncol.* **2010**, *5*, 1315–1316. [CrossRef] [PubMed]
7. Mak, T.; Porsch, R.; Choi, S.; Zhou, X.; Sham, P. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **2017**, *41*, 469–480. [CrossRef]
8. Huang, H.; Darbar, D. Genetic Risk Scores for Atrial Fibrillation: Do they Improve Risk Estimation? *Can. J. Cardiol.* **2017**, *33*, 422–424. [CrossRef] [PubMed]
9. Hosmer, D.; Lemeshow, S. *Applied Logistic Regression*, 2nd ed.; John Wiley and Sons: New York, NY, USA, 2000; pp. 160–164. Chapter 5.
10. Nesterov, Y. Smooth minimization of non-smooth functions. *Math. Program. Ser. A* **2005**, *103*, 127–152. [CrossRef]
11. Hahn, G.; Lutz, S.M.; Laha, N.; Lange, C. A framework to efficiently smooth L1 penalties for linear regression. *bioRxiv* **2020**, 1–35. [CrossRef]
12. Hahn, G.; Lutz, S.; Laha, N.; Cho, M.; Silverman, E.; Lange, C. A fast and efficient smoothing approach to LASSO regression and an application in statistical genetics: Polygenic risk scores for Chronic obstructive pulmonary disease (COPD). *Stat. Comput.* **2021**, *31*, 35. [CrossRef]
13. Kunkle, B.; Grenier-Boley, B.; Sims, R.; Bis, J.; Damotte, V.; Naj, A.; Boland, A.; Vronskaya, M.; van der Lee, S.; Amlie-Wolf, A.; et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **2019**, *51*, 414–430. [CrossRef] [PubMed]
14. Jansen, I.; Savage, J.; Watanabe, K.; Bryois, J.; Williams, D.; Steinberg, S.; Sealock, J.; Karlsson, I.; Hägg, S.; Athanasiu, L.; et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **2019**, *51*, 404–413. [CrossRef]
15. Regan, E.; Hokanson, J.; Murphy, J.; Make, B.; Lynch, D.; Beaty, T.; Curran-Everett, D.; Silverman, E.; Crapo, J. Genetic epidemiology of COPD (COPDGene) study design. *COPD* **2010**, *7*, 32–43. [CrossRef]
16. Privé, F.; Arbel, J.; Vilhjálmsson, B.J. LDpred2: Better, faster, stronger. *Bioinformatics* **2019**. [CrossRef]
17. Ge, T.; Chen, C.Y.; Ni, Y.; Feng, Y.C.A.; Smoller, J.W. 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. 2020. Available online: https://github.com/getian107/PRScs (accessed on 8 May 2021).
18. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. B Meter.* **1996**, *58*, 267–288. [CrossRef]
19. Hahn, G.; Lutz, S.M.; Laha, N.; Lange, C. smoothedLasso: Smoothed LASSO Regression via Nesterov Smoothing. 2020. R-Package Version 1.4. Available online: https://cran.r-project.org/src/contrib/Archive/smoothedLasso/smoothedLasso_1.4.tar.gz (accessed on 8 May 2021).
20. Choi, S.W.; Mak, T.S.H.; O'Reilly, P.F. Tutorial: A guide to performing polygenic risk score analyses. *Nat. Protoc.* **2020**, *15*, 2759–2772. [CrossRef]
21. Purcell, S.; Wray, N.; Stone, J.; Visscher, P.; O'Donovan, M.C.; Sullivan, P.; Sklar, P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **2009**, *460*, 748–752. [PubMed]

22. Wray, N.; Lee, S.; Mehta, D.; Vinkhuyzen, A.; Dudbridge, F.; Middeldorp, C. Research Review: Polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry* **2014**, *55*, 1068–1087. [CrossRef]
23. Vilhjálmsson, B.; Yang, J.; Finucane, H.; Gusev, A.; Lindström, S.; Ripke, S.; Genovese, G.; Loh, P.; Bhatia, G.; Do, R.; et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **2015**, *97*, 576–592. [CrossRef] [PubMed]
24. Mak, T.; Kwan, J.; Campbell, D.; Sham, P. Local true discovery rate weighted polygenic scores using GWAS summary data. *Behav. Genet.* **2016**, *46*, 573–582. [CrossRef]
25. Zhang, Y.; Qi, G.; Park, J.H.; Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* **2018**, *50*, 1318–1326. [CrossRef] [PubMed]
26. Lloyd-Jones, L.R.; Zeng, J.; Sidorenko, J.; Yengo, L.; Moser, G.; Kemper, K.E.; Wang, H.; Zheng, Z.; Magi, R.; Esko, T.; et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **2019**, *10*, 5086. [CrossRef]
27. Ge, T.; Chen, C.Y.; Ni, Y.; Feng, Y.C.A.; Smoller, J.W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **2019**, *10*, 1776. [CrossRef] [PubMed]
28. Zhang, Q.; Privé, F.; Vilhjálmsson, B.; Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.* **2021**, *12*, 1–9. [CrossRef]
29. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]
30. Friedman, J.; Hastie, T.; Tibshirani, R.; Narasimhan, B.; Tay, K.; Simon, N.; Qian, J. glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. 2020. R-Package Version 4.0. Available online: https://cran.r-project.org/package=glmnet (accessed on 8 May 2021).
31. Beck, A.; Teboulle, M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imaging Sci.* **2009**, *2*, 183–202. [CrossRef]
32. Choi, S.W.; O'Reilly, P.F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **2019**, *8*, giz082. [CrossRef]
33. Zhang, W.; Tang, J.; Wang, N. Using the Machine Learning Approach to Predict Patient Survival from High-Dimensional Survival Data. In Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, China, 15–18 December 2016; pp. 1–5.
34. Mamaniab, N.M. Machine Learning techniques and Polygenic Risk Score application to prediction genetic diseases. *Adv. Distrib. Comput. Artif. Intell.* **2020**, *9*, 5–14.
35. Badré, A.; Zhang, L.; Muchero, W.; Reynolds, J.C.; Pan, C. Deep neural network improves the estimation of polygenic risk scores for breast cancer. *J. Hum. Genet.* **2021**, *66*, 359–369. [CrossRef]
36. Huang, S.; Ji, X.; Cho, M.; Joo, J.; Moore, J. DL-PRS: A novel deep learning approach to polygenic risk scores. *BMC Bioinform.* **2021**. [CrossRef]
37. Peng, J.; Li, J.; Han, R.; Wang, Y.; Han, L.; Peng, J.; Wang, T.; Hao, J.; Shang, X.; Wei, Z. A Deep Learning-based Genome-wide Polygenic Risk Score for Common Diseases Identifies Individuals with Risk. *medRxiv* **2021**. [CrossRef]
38. Gola, D.; Erdmann, J.; Müller-Myhsok, B.; Schunkert, H.; König, I.R. Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. *Genet. Epidemiol.* **2020**, *44*, 125–138. [CrossRef]
39. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Stat Comp.: Vienna, Austria, 2014.
40. Mak, T.; Porsch, R.; Choi, S.; Zhou, X.; Sham, P. Lassosum: A Method for Computing LASSO/Elastic Net Estimates of a Linear Regression Problem Given Summary Statistics from GWAS and Genome-Wide Meta-Analyses. 2020. Available online: https://github.com/tshmak/lassosum (accessed on 8 May 2021).
41. Privé, F.; Blum, M.; Aschard, H. bigsnpr: Analysis of Massive SNP Arrays. 2020. R-Package Version 1.5.2. Available online: https://cran.r-project.org/package=bigsnpr (accessed on 8 May 2021).
42. Hahn, G.; Lutz, S.M.; Laha, N.; Lange, C. smoothedLasso: Smoothed LASSO Regression via Nesterov Smoothing, 2020. R-Package Version 1.5. Available online: https://cran.r-project.org/package=smoothedLasso (accessed on 8 May 2021).
43. Falbel, D.; Allaire, J.; Chollet, F.; Studio, R.; Tang, Y.; Bijl, W.V.D.; Studer, M.; Keydana, S. keras: R Interface to 'Keras'. 2020. R-Package Version 2.3.0.0. Available online: https://cran.r-project.org/package=keras (accessed on 8 May 2021).
44. Falbel, D.; Allaire, J.; Studio, R.; Tang, Y.; Eddelbuettel, D.; Golding, N.; Kalinowski, T. Tensorflow: R Interface to 'TensorFlow'. 2020. R-Package Version 2.2.0. Available online: https://cran.r-project.org/package=tensorflow (accessed on 8 May 2021).
45. Zeng, J.; Yang, J.; Zhang, F.; Zheng, Z.; Lloyd-Jones, L.; Goddard, M. GCTB: A Tool for Genome-Wide Complex Trait Bayesian Analysis. 2020. Available online: https://cnsgenomics.com/software/gctb/#Overview (accessed on 8 May 2021).
46. Speed, D. MegaPRS. 2021. Available online: http://dougspeed.com/prediction/ (accessed on 8 May 2021).
47. McCarthy, S.; Das, S.; Kretzschmar, W.; Delaneau, O.; Wood, A.R.; Teumer, A.; Kang, H.M.; Fuchsberger, C.; Danecek, P.; Sharp, K.; et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **2016**, *48*, 1279–1283.
48. Partners. Partners Healthcare Biobank. 2020. Available online: https://biobank.partners.org (accessed on 8 May 2021).
49. World Health Organization. *International Statistical Classification of Diseases and Related Health Problems (ICD)*; World Health Organization: Geneva, Switzerland. 2021. Available online: https://www.who.int/standards/classifications/classification-of-diseases (accessed on 8 May 2021).
50. Charlson, M.; Szatrowski, T.; Peterson, J.; Gold, J. Validation of a combined comorbidity index. *J. Clin. Epidemiol.* **1994**, *47*, 1245–1251. [CrossRef]

51. Karlson, E.W.; Boutin, N.T.; Hoffnagle, A.G.; Allen, N.L. Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations. *J. Pers. Med.* **2016**, *6*, 2. [CrossRef]
52. Manichaikul, A.; Mychaleckyj, J.C.; Rich, S.S.; Daly, K.; Sale, M.; Chen, W.M. Robust relationship inference in genome-wide association studies. *Bioinformatics* **2010**, *26*, 2867–2873. [CrossRef]
53. Chen, W.M. KING: Kinship-Based INference for Gwas. 2021. Available online: https://kingrelatedness.com/ (accessed on 8 May 2021).
54. Purcell, S.; Chang, C. PLINK2 (v2.00, 31 Aug 2020). 2020. Available online: www.cog-genomics.org/plink/2.0/ (accessed on 8 May 2021).
55. 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.; et al. Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. *Nat. Commun.* **2020**, *11*, 4799. [CrossRef] [PubMed]
56. Ware, E.B.; Faul, J.D.; Mitchell, C.M.; Bakulski, K.M. Considering the APOE locus in Alzheimer's disease polygenic scores in the Health and Retirement Study: A longitudinal panel study. *BMC Med. Genom.* **2020**, *13*, 164. [CrossRef]
57. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
58. NHLBI TOPMed. Genetic Epidemiology of COPD (COPDGene) Funded by the National Heart, Lung, and Blood Institute (NHLBI) in the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program. 2018. Available online: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000951.v5.p5 (accessed on 13 October 2021)
59. 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.; et al. A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genet.* **2015**, *16*, 138. [CrossRef] [PubMed]
60. Bolli, A.; Domenico, P.D.; Bottà, G. Software as a Service for the Genomic Prediction of Complex Diseases. 2019. Available online: http://xxx.lanl.gov/abs/10.1101/763722 (accessed on 8 May 2021).
61. Wald, N.J.; Old, R. The illusion of polygenic disease risk prediction. *Genet. Med.* **2019**, *21*, 1705–1707. [CrossRef] [PubMed]
62. NIAGADS. NG00075—IGAP Rare Variant Summary Statistics—Kunkle et al. (2019). 2016. Available online: https://www.niagads.org/datasets/ng00075 (accessed on 8 May 2021).
63. CTG Lab. Summary Statistics for Alzheimer's Dementia from Iris Jansen et al., 2019. 2021. Available online: https://ctg.cncr.nl/software/summary_statistics (accessed on 8 May 2021).