Statistical learning for sparser fine-mapped polygenic models

Maj, Carlo; Staerk, Christian; Borisov, Oleg; Klinkhammer, Hannah; Yeung, Ming Wai; Krawitz, Peter; Mayr, Andreas

Published in:
Genetic epidemiology

DOI:
10.1002/gepi.22495

Publication date:
2022

Citation for published version (APA):
Maj, C., Staerk, C., Borisov, O., Klinkhammer, H., Yeung, M. W., Krawitz, P., & Mayr, A. (2022). Statistical learning for sparser fine-mapped polygenic models: The prediction of LDL-cholesterol. *Genetic epidemiology*, *46*(8), 589-603. https://doi.org/10.1002/gepi.22495

Received: 24 August 2021 | Revised: 3 June 2022 | Accepted: 7 June 2022

DOI: 10.1002/gepi.22495

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

Statistical learning for sparser fine-mapped polygenic models: The prediction of LDL-cholesterol

Carlo Maj[1,2] | **Christian Staerk**[3] | **Oleg Borisov**[1] | **Hannah Klinkhammer**[1,3] | **Ming Wai Yeung**[1,4] | **Peter Krawitz**[1] | **Andreas Mayr**[3]

[1]Institute for Genomic Statistics and Bioinformatics, Medical Faculty, University Bonn, Bonn, Germany

[2]Centre for Human Genetics, University of Marburg, Marburg, Germany

[3]Institute for Medical Biometry, Informatics and Epidemiology, Medical Faculty, University Bonn, Bonn, Germany

[4]Department of Cardiology, University of Groningen, Groningen, The Netherlands

Correspondence
Carlo Maj, Institute for Genomic Statistics and Bioinformatics, Medical Faculty, University Bonn, Bonn, Germany.
Email: carlo.maj@gmail.com

Funding information
BONFOR-program of the Medical Faculty University of Bonn,
Grant/Award Number: O-147.0002

Abstract

Polygenic risk scores quantify the individual genetic predisposition regarding a particular trait. We propose and illustrate the application of existing statistical learning methods to derive sparser models for genome-wide data with a polygenic signal. Our approach is based on three consecutive steps. First, potentially informative loci are identified by a marginal screening approach. Then, fine-mapping is independently applied for blocks of variants in linkage disequilibrium, where informative variants are retrieved by using variable selection methods including boosting with probing and stochastic searches with the Adaptive Subspace method. Finally, joint prediction models with the selected variants are derived using statistical boosting. In contrast to alternative approaches relying on univariate summary statistics from genome-wide association studies, our three-step approach enables to select and fit multivariable regression models on large-scale genotype data. Based on UK Biobank data, we develop prediction models for LDL-cholesterol as a continuous trait. Additionally, we consider a recent scalable algorithm for the Lasso. Results show that statistical learning approaches based on fine-mapping of genetic signals result in a competitive prediction performance compared to classical polygenic risk approaches, while yielding sparser risk models.

KEYWORDS

boosting, polygenic score, stochastic search, UK Biobank, variable selection

1 | INTRODUCTION

Polygenic risk scores (PRS) represent quantification of the individual genetic predisposition for a particular phenotype (e.g., a clinical trait). PRS have already been developed for a large variety of human traits and have been integrated into prediction models for common diseases together with traditional risk factors and clinical variables (Kachuri et al., 2020; Khera et al., 2018).

Polygenic modeling has to deal with two major issues: the high-dimensionality of the genetic signal (typically characterized by millions of common variants) and the

Carlo Maj and Christian Staerk contributed equally to this study.

presence of high correlations among variants (i.e., linkage disequilibrium [LD]) within the genome (Ardlie et al., 2002). Therefore, a classical approach to PRS is based on the cumulative marginal effects from simple univariate regression models for individual variants, incorporating a variant selection filter according to the level of LD and to the level of significance with respect to the considered phenotype. This traditional method—usually referred to as clumping plus thresholding (C+T)—is still one of the most popular approaches, given its computational efficiency (Euesden et al., 2015). More refined methods adapt penalized regression methods for summary statistics from genome-wide association studies (GWAS), including lassosum (Mak et al., 2017) implementing Lasso regression and penRegSum (Pattee & Pan, 2020) considering an elastic net penalty. Alternatively, Bayesian approaches can be employed to induce shrinkage of effect estimates, such as LDpred (Vilhjálmsson et al., 2015) and PRS-CS (Ge et al., 2019).

All these methods are based on summary statistics from GWAS, typically incorporating a particular way of external reference LD-panel matching to account for LD (Vilhjálmsson et al., 2015). While such approaches facilitate the exchange of research data and the computational feasibility, particularly regarding memory issues, a major limitation of using summary statistics is that the resulting polygenic scores do not fully utilize the joint information of the variants regarding the phenotype of interest. Furthermore, while *trans*-ethnic GWAS indicate that causal variants are mainly shared across populations (Li & Keating, 2014), the poor generalization of classical PRS models to different populations is largely driven by different allele frequencies and LD-patterns across populations (Wang et al., 2020). Therefore, it can be hypothesized that PRS based only on the most informative ("fine-mapped") variants may be less sensitive to LD differences among populations (Weissbrod et al., 2020).

From a statistical point of view, a more effective approach would be to apply modern multivariable regression models directly on the original genotype data, instead of relying on univariate summary statistics. Multivariable regression models enable the joint estimation of associations and could lead to improved fine-mapping (i.e., the identification of potentially causal variants; Benner et al., 2016). Although there exists a huge literature on statistical prediction models for high-dimensional data (large number of predictors p related to small sample size n), their application for PRS is often hindered by computational aspects and memory restrictions due to the size of the data sets (large p and large n). Recently, Qian et al. (2020) proposed a so-called batch screening iterative Lasso (BASIL) algorithm to apply penalized regression for large GWAS cohorts like UK Biobank, showing that one can compute the full Lasso solution path for the original genotype data by successively solving lower-dimensional subproblems without requiring to load the full data set into memory.

In this study, we propose and illustrate an alternative technique to allow for the application of existing statistical learning methods for multivariable regression on original genotype data, by exploiting the characteristic correlation structure between the variants in the genome. In particular, variants within the same genomic region are likely to be correlated due to LD, while variants in distant genomic regions can be considered as independent, implying that variable selection methods can focus on the identification of the most informative variants within each region of high LD. Thus, in the proposed approach we first split the genome into independent LD blocks (chunks), then apply modern variable selection techniques on relevant blocks to finemap the genetic signal regarding the particular phenotype, and finally combine the selected variants to fit a joint prediction model using statistical boosting. Particularly, we illustrate our approach with several different statistical learning methods for variant selection (fine-mapping) in the independent blocks. As the first alternative, we make use of a stochastic search algorithm called the Adaptive Subspace (AdaSub) method, aiming to select the best model according to an ℓ_0-type information criterion by adaptively solving lower-dimensional subproblems (Staerk et al., 2021). As the second alternative, we consider statistical boosting in combination with probing (Thomas et al., 2017), yielding automatic stopping and sparse prediction models. Finally, we use the recently proposed BASIL algorithm (Qian et al., 2020) to compute Lasso estimates based on chunk-based prefiltered variants and compare it to the direct application of the Lasso without prefiltering.

We apply the proposed approach on UK Biobank data with $n = 487{,}410$ samples and $p = 9{,}812{,}717$ variants, considering LDL-cholesterol as a continuous trait characterized by a polygenic architecture. Our results suggest that the statistical learning methods AdaSub and probing yield sparser and more interpretable polygenic prediction models, while showing a competitive prediction performance on independent test data compared to classical C+T and Bayesian approaches based on summary statistics. The direct application of the Lasso on full genotype data yields the best prediction performance with considerably larger numbers of selected variants compared to the sparse fine-mapped models via AdaSub and boosting with probing. However, we find that the Lasso still yields similar prediction accuracy with less selected variants when applied only on chunk-based pre-filtered variants. In a simulation study we investigate the robustness of the different methods regarding alterations of LD-patterns and genotyping/imputation errors. The evaluation of the final scores on data from other populations indicates potential benefits regarding the generalizability of the developed fine-mapped models.

2 | METHODS

2.1 | Data preprocessing

This study has been conducted using the UK Biobank Resource under Application Number 81202. The UK Biobank (Bycroft et al., 2018) is a large-scale cohort study covering a huge prospective sample ($n > 500{,}000$) of the British general population, including both genotype as well as phenotype (health-related outcomes) data. In this study we focus on LDL-cholesterol as the phenotype of interest. Only individuals with both self-reported white British origin (Field 21000) and Caucasian origin according to the principal components provided by UK Biobank (Field 22006) were used in the training cohort. Variants with a genotyping rate of less than 99%, or which had a minor allele frequency (MAF) < 1% were removed. Variants not in Hardy–Weinberg equilibrium ($p < 10^{-6}$) were also excluded.

All data preprocessing steps were performed with PLINK 1.9 (Chang et al., 2015). The models were trained considering imputed dosages provided by UK Biobank after filtering for MAF > 0.01 and post imputation info >0.8. UK Biobank is enriched of related individuals; specifically, 147,731 individuals were inferred to be related up to third degree or closer (Bycroft et al., 2018). We applied no filter on relatedness as removing related individuals would have led to a nontrivial decrease of the sample size. Previous work had shown that sample filtering according to the coefficients of relationship (kinship) leads to similar PRS associations (Lello et al., 2018).

Since one of the aims of the present work was the evaluation of the generalizability of the derived PRS, no population-based filter was applied for the test data set. In total 318,258 samples were included in the training cohort and 150,521 samples were considered as an independent test cohort. To reduce the memory demand and allow for parallelization, we split the genome into independent LD blocks (i.e., chunks) according to the ldetect method as described in Berisa and Pickrell (2016). Specifically, we considered the 1,703 blocks in autosomal chromosomes identified by ldetect considering the European population of 1K genome. As the proposed approaches are based on fine-mapping of the significant regions, first a univariate linear-regression association test has been performed between the variant dosages and LDL-cholesterol (Field 30,780), respectively. The chunks with at least one genome-wide significant association (i.e., p value of association lower than 5×10^{-8}) were then further processed to fine-map the genetic signal. Within each chunk we additionally filtered for suggestively significant variants ($p < 10^{-5}$), compare also Fan and Lv (2008) and Hoffman et al. (2013).

2.2 | Fine-mapping of variants

2.2.1 | Stochastic search with AdaSub

As the first fine-mapping approach, we considered ℓ_0-type information criteria to identify the most informative variants in each of the chunks. Such variable selection criteria provide a natural trade-off between goodness-of-fit and model complexity, by explicitly penalizing the number of variants included in the model. As an important example, the extended Bayesian information criterion (EBIC; J. Chen & Chen, 2008) has been proposed for high-dimensional data situations with many possible covariates (variants) and has shown to yield variable selection consistency even when the number of covariates p exceeds the sample size n. In particular, for a subset of variants indexed by $S \subseteq \{1, \ldots, p\}$, the EBIC_γ is given by

$$\text{EBIC}_\gamma(S) = n \log\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_S + \boldsymbol{x}_i{}'\hat{\boldsymbol{\beta}}_S - y_i\right)^2\right) + (\log(n) + 2\gamma\log(p))|S|, \quad (1)$$

where $y_i \in \mathbb{R}$ denotes the observed phenotype and $\boldsymbol{x}_i \in \{0, 1, 2\}^p$ the genotype for subjects $i = 1, \ldots, n$. Furthermore, $\hat{\mu}_S$ denotes the estimated intercept and $\hat{\beta}_S \in \mathbb{R}^p$ the least squares estimate under model S (i.e., $\hat{\beta}_j = 0$ for $j \notin S$). The EBIC_γ penalizes the number $|S|$ of selected variants with the factor $\log(n) + 2\gamma\log(p)$, where n refers to the sample size of the training data and p to the total number of variants in the training data. The additional constant $\gamma \in [0, 1]$ controls the induced sparsity, with $\gamma = 1$ resulting in the sparsest models and $\gamma = 0$ corresponding to the classical Bayesian information criterion. Based on the EBIC_γ, the "optimal" set $\hat{S}$ of variants is defined by the one which yields the smallest criterion value among all possible sets $S \subseteq \{1, \ldots, p\}$ of variants, that is

$$\text{EBIC}_\gamma(\hat{S}) = \min_{S \subseteq \{1, \ldots, p\}} \text{EBIC}_\gamma(S). \quad (2)$$

Due to the combinatorial nature of the optimization problem (2), it is computationally challenging to find the model minimizing the EBIC_γ. Here we make use of the AdaSub method (Staerk et al., 2021) as a stochastic search algorithm, which is based on reducing the high-dimensional search problem (2) to low-dimensional subproblems

$$\mathrm{EBIC}_{\gamma}(S^{[m]}) = \min_{S \subseteq V^{[m]}} \mathrm{EBIC}_{\gamma}(S), \tag{3}$$

for smaller random subspaces $V^{[m]} \subseteq \{1, \ldots, p\}$, which can be solved exactly with branch-and-bound algorithms (Furnival & Wilson, 2000). After each iteration m of AdaSub, the information from the solved subproblems (3) is adaptively combined, so that variables $x^{(j)}$ which have already proven to be informative in many of the solved sub-problems receive a larger sampling probability

$$P(j \in V^{[m+1]}) = \frac{q + K\sum_{i=1}^{m} 1_{S^{[i]}}(j)}{p + K\sum_{i=1}^{m} 1_{V^{[i]}}(j)}, \; j = 1, \ldots, p, \tag{4}$$

in the subsequent stochastic search (where 1_S denotes the indicator function for a set S). The initial expected size of the sampled subspaces is given by q (with $q \ll p$), while the adaptation rate of the algorithm is controlled by the parameter $K > 0$. Under certain conditions it can be guaranteed that AdaSub converges against the optimal solution of the original problem (Staerk et al., 2021). Particularly in sparse and highly correlated data situations it has been observed that AdaSub shows favorable variable selection properties in comparison to ℓ_1-type regularization methods such as the Lasso (Tibshirani, 1996). As high correlations occur frequently among variants within the same genomic region due to linkage disequilibrium, the AdaSub method is a well-suited candidate to identify sparse sets of informative variants explaining the genomic signal for a particular phenotype.

For each chunk, we independently applied AdaSub using the EBIC_{γ} with constant $\gamma = 1$ as the selection criterion. As the penalty of the EBIC_1 incorporates the total number of considered variants, it accounts for the fact that AdaSub is only applied on the prefiltered chunks and the resulting multiple testing issue. In AdaSub we initialized the expected search size $q = 5$ and the parameter $K = 2000$ controlling the adaptation rate of the algorithm. For each chunk, the best model identified after 10,000 iterations of AdaSub was returned as the set of selected variants.

2.2.2 | Statistical gradient boosting with probing

As an alternative to the explicit regularization imposed by information criteria, we implemented a statistical boosting algorithm (Mayr et al., 2014) in combination with probing (Thomas et al., 2017) for early stopping, providing implicit regularization and variable selection. The concept of boosting emerged from machine-learning (Freund & Schapire, 1996), where it is often used in combination with trees as base-learners to form a powerful and flexible classifier (T. Chen & Guestrin, 2016). For statistical boosting approaches, univariate regression functions are implemented as base-learners and are iteratively fitted to the current gradients of the loss function, yielding a gradient descent procedure in function space (Bühlmann & Hothorn, 2007).

More formally, the gradient vector $\boldsymbol{u}^{[m]}$ at iteration $m = 1, \ldots, m_{\text{stop}}$ is the first derivative of the loss function $\rho(y_i, \eta_i)$ w.r.t. the model η evaluated at the previous iteration $m - 1$, that is

$$\begin{aligned}\boldsymbol{u}^{[m]} &= \left(u_i^{[m]}\right)_{i=1,\ldots,n} \\ &= \left(-\frac{\partial}{\partial \eta_i}\rho(y_i, \eta_i)\bigg|_{\eta_i = \hat{\eta}^{[m-1]}(x_i)}\right)_{i=1,\ldots,n}.\end{aligned} \tag{5}$$

In the context of linear regression, η_i represents the linear predictor for observation i at iteration $m - 1$:

$$\hat{\eta}^{[m-1]}(x_i) = \hat{\mu}^{[m-1]} + \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}^{[m-1]}. \tag{6}$$

The gradient vector $\boldsymbol{u}^{[m]}$ is then fitted one-by-one to the base-learners $h_j(x^{(j)}), j = 1, \ldots, p$, which typically represent univariate regression functions for the different covariates $x^{(1)}, \ldots, x^{(p)}$. In the case of the L_2 loss given by $\rho(y_i, \eta_i) = (y_i - \eta_i)^2$, this procedure basically leads to a component-wise refitting of residuals by the base-learners. Although all base-learners are fitted to $\boldsymbol{u}^{[m]}$, only the best-fitting base-learner $h_{j^*}\left(x^{(j^*)}\right)$ is selected and its fit is added to the current model via a small step-length (e.g., sl = 0.1):

$$\hat{\eta}^{[m]} = \hat{\eta}^{[m-1]} + \mathrm{sl} \cdot h_{j^*}\left(x^{(j^*)}\right). \tag{7}$$

In the framework of classical linear regression with univariate linear base-learners $h_j(x^{(j)}) = \mu + \beta_j x^{(j)}$ for $j = 1, \ldots, p$, this leads to

$$\begin{aligned}\hat{\beta}_{j^*}^{[m]} &= \hat{\beta}_{j^*}^{[m-1]} + \mathrm{sl} \cdot h_{j^*}\left(x^{(j^*)}\right) \quad \text{and} \\ \hat{\beta}_{j}^{[m]} &= \hat{\beta}_{j}^{[m-1]} \quad \text{for } j \neq j^*.\end{aligned} \tag{8}$$

As in every iteration only the best-fitting base-learner $h_{j^*}(x^{(j^*)})$ is selected, variables that have not been included in any of the selected base-learners are effectively excluded from the final model when the algorithm is stopped. As a result, statistical boosting yields multivariable regression models while incorporating automated variable selection for potentially high-dimensional data situations, where classical statistical inference procedures become infeasible (Mayr et al., 2014).

For each chunk, we applied statistical boosting separately using linear models with single variants as base-learners. Instead of tuning the stopping iteration with respect to the prediction accuracy via resampling techniques, we incorporated a probing approach (Wu et al., 2007): for each variant $x^{(j)}$, we additionally included a base-learner with a shadow-variable $\tilde{x}^{(j)}$ (probe). Each probe is a randomly permuted sibling of an original variant and, due to the permutation, not associated with the outcome. Once the first base-learner corresponding to one of these probes $\tilde{x}^{(1)}, \ldots, \tilde{x}^{(p)}$ is (falsely) selected, the boosting procedure is stopped. This combination of statistical boosting with probing has shown to yield particularly sparse models with low numbers of false-positives, shifting the focus of the tuning procedure from prediction accuracy to variable selection (Thomas et al., 2017).

2.2.3 | Lasso via the BASIL algorithm

Qian et al. (2020) recently proposed a batch screening iterative Lasso (BASIL) algorithm to apply the Lasso (Tibshirani, 1996) directly on large-scale genotype data. The Lasso is defined as a solution to the penalized regression problem

$$\min_{\mu \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \left(\mu + \boldsymbol{x}_i' \boldsymbol{\beta} - y_i\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{9}$$

with penalty parameter $\lambda \geq 0$ controlling the sparsity and shrinkage of regression coefficients induced by the ℓ_1-regularization. For large-scale applications with a high memory demand, the BASIL algorithm enables the construction of the full Lasso path (for decreasing λ) by iteratively working on smaller adaptive batches of genomic variants. In more detail, the algorithm consists of three repetitive steps: in a screening step, variants which are most correlated with the current residuals are added to an active set of variants. Then, the Lasso is fitted only on the active set for a consecutive range of λ values. Finally, the found solutions are checked for validity using the Karush–Kuhn–Tucker conditions, before those three steps are repeated. By checking the validity of each solution, the exact Lasso path is retrieved (Qian et al., 2020). The final Lasso estimate is derived by choosing the penalty parameter λ yielding the best prediction performance on a validation set.

2.3 | Final estimation and comparison to classical PRS

After identifying the informative variants separately for the different chunks using the two fine-mapping approaches AdaSub and boosting with probing based on the training cohort with $n = 281{,}843$ samples, we combined the respective variants in a PRS by fitting a multivariable regression model on the training cohort via statistical boosting (component-wise L_2 boosting, Bühlmann & Yu, 2003). This procedure was performed once for the variants selected via stochastic search with AdaSub and once for variants selected via boosting with probing. In both cases, we fit the complete PRS, estimating new coefficients $\hat{\beta}_j$ to allow those estimates to take the combined multivariable effects of the selected variants also among different chunks into account. In this final estimation step, we led the boosting algorithm converge by fixing a large number of 10,000 iterations. We employed the Lasso using the BASIL algorithm (Qian et al., 2020) for two different sets of variants. First, similarly as in Qian et al. (2020), we applied the Lasso on all genome-wide genotyped variants. Second, similarly to AdaSub and boosting with probing, we applied the Lasso only on the chunk-based prefiltered variants (see Section 2.1). For both sets, we first computed the Lasso path using 87.5% of our training cohort for the fitting and 12.5% of our training cohort as validation data. We then refitted the Lasso on the complete training data for the penalty parameter λ corresponding to the highest prediction accuracy on the validation data.

The derived models were also compared with the classical PRS obtained with clumping plus thresholding (C+T), considering the full genome-wide signal (without prefiltering and fine-mapping of variants). In the PRS derived via C+T, the regression coefficients $\hat{\beta}_j$ reflect the univariate (marginal) associations between the allele dosages and the phenotype (as also derived from GWAS summary statistics). Different p value thresholds were considered (i.e., 25 values equally distributed in the log scale between 5×10^{-8} and 0.05) and also different LD thresholds for clumping were used (i.e., eight correlation values equally distributed between 0.1 and 0.8), as implemented in PRSice (Euesden et al., 2015). Moreover, to further evaluate the potential influence of model sparsity, we also computed genome-wide PRS based on genotyped variants, assuming a genetic architecture in which all variants are causal. To this aim, we implemented the infinitesimal model of LDpred2 based on the shrinkage of effect sizes according to heritability estimation (Privé et al., 2020). In addition, we also derived a

PRS based on continuous shrinkage of effect sizes according to the UKB–EUR reference LD panel as implemented in PRS-CS (Ge et al., 2019). Similarly to the C+T methods, both LDpred2 and PRS-CS are based on univariate (marginal) associations which are usually available as summary statistics from GWAS.

As LDL blood levels are strongly influenced by lipid-lowering drugs, we adjusted LDL values by a factor of 0.684 in individuals taking statins as estimated in a recent work (Sinnott-Armstrong et al., 2021). The final PRS for individual i is then computed as the weighted sum of effect alleles:

$$\widehat{\mathrm{PRS}}_i = \sum_{j \in S} \hat{\beta}_j x_i^{(j)}, \qquad (10)$$

where $\hat{\beta}_j$ is the estimated weight for variant j (obtained from the univariate association or derived from the final multivariable models after fine-mapping), $x_i^{(j)}$ is the corresponding genotype for individual i and S denotes the respective set of selected (fine-mapped) variants.

As a sensitivity analysis regarding the robustness of the different PRS models on deviations from the target population, we also performed two permutation-based simulations on parts of the test data with British ancestry. In both simulation scenarios, we altered randomly selected windows of 1000 variants each, leading in total to 1%–25% deviating variants across the genome. To check for the effect of an alteration of correlations between nearby variants, in the first scenario we permuted the location of the variants inside the selected windows on the test observations. To investigate the robustness regarding genotyping/imputation errors, in the second simulation scenario we again consider the variants inside the randomly selected windows, but, instead of their locations, we jointly permute their values across observations, effectively "knocking-out" their effect on the outcome for these test observations. In contrast to the first scenario, in the second scenario the LD-patterns within the windows are not altered by the permutations since the values of the variants in each window are permuted together across observations. For both simulation scenarios we assessed the relative performance of the PRS compared to their performance on the original data.

3 | RESULTS

3.1 | Fine-mapping

The considered polygenic risk approaches for LDL-cholesterol yield substantially different numbers of selected variants in the final models (see Table 1). Here, we specifically focus on the selected variants by the fine-mapping approaches via AdaSub and statistical boosting with probing as well as by the classical genome-wide C+T approach. In particular, the C+T approach selects 1588 variants (best fitting model obtained with p value threshold of 5.25×10^{-5} and clumping R^2 of 0.1), boosting with probing selects 792 variants and AdaSub with EBIC_1 leads to only 108 selected variants.

In Figure 1, the top significant locus (referring to univariate associations with LDL-cholesterol) is displayed on chromosome 16, highlighting the variants that were selected for the final PRS with the statistical learning methods AdaSub and probing as well as the

TABLE 1 Results of the covariate-only model (M_c) and the univariate and multivariable polygenic models based on genome-wide and chunk-based prefiltered variants for the prediction of LDL-cholesterol

Method	Input	N variants	R^2 for EUR	R^2 for AFR	R^2 for EAS	R^2 for SAS
Covariate-only model	—	0	0.030	0.022	0.041	0.00042
Univariate polygenic models						
C+T	Genome-wide	1588	0.083 (100%)	0.033 (39.1%)	0.045 (54.1%)	0.012 (14.3%)
LDpred2-inf	Genome-wide	1,048,692	0.079 (100%)	0.028 (35.0%)	0.040 (50.6%)	0.011 (14.2%)
PRS-CS	Genome-wide	1,110,740	0.122 (100%)	0.048 (39.5%)	0.074 (60.3%)	0.022 (18.0%)
Multivariable polygenic models						
Boosting with Probing	Prefiltered	792	0.180 (100%)	0.161 (89.4%)	0.098 (54.4%)	0.062 (34.4%)
AdaSub with EBIC_1	Prefiltered	108	0.163 (100%)	0.145 (88.9%)	0.095 (58.3%)	0.053 (32.5%)
Lasso	Genome-wide	12,492	0.204 (100%)	0.104 (50.8%)	0.108 (52.3%)	0.072 (35.2%)
Lasso	Prefiltered	1821	0.197 (100%)	0.164 (83.2%)	0.117 (59.2%)	0.079 (40.1%)

Note: The number of selected variants (N variants) for the final models (sparsity) and the prediction performance (R^2) on the test set of 120,242 Europeans (EUR) as well as on three populations different from the training set (6706 Africans, AFR; 2073 East Asians, EAS; 7788 South Asians, SAS). For each of the methods, the relative prediction performance compared to the European population is provided in parentheses.

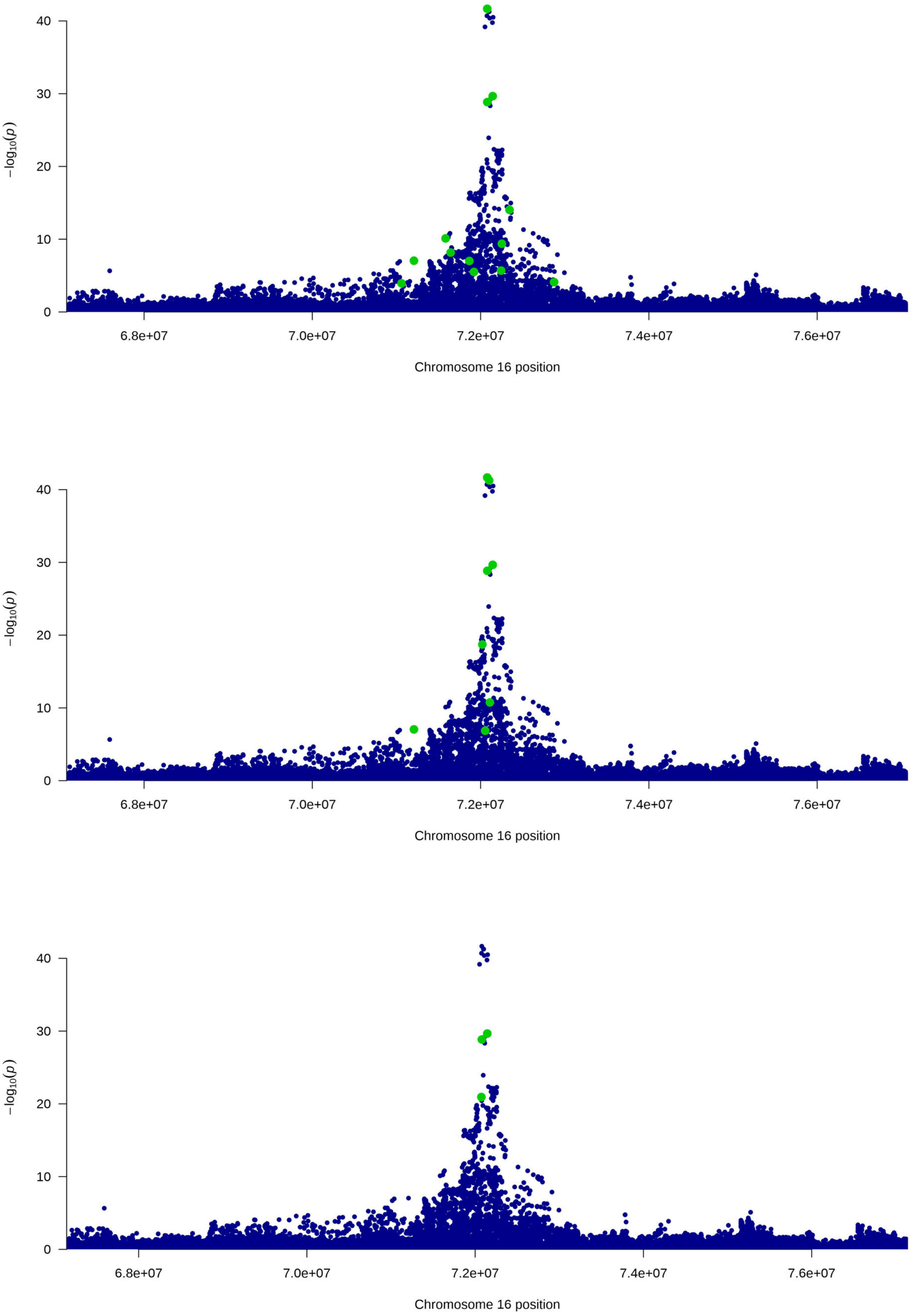

FIGURE 1 Regional association signals for top significant locus in chromosome 16 (p values correspond to univariate associations between variants and LDL-cholesterol). Variants selected in the final models (the top, middle, and bottom figures correspond to C+T, boosting with probing, and AdaSub with EBIC_1) are highlighted in green.

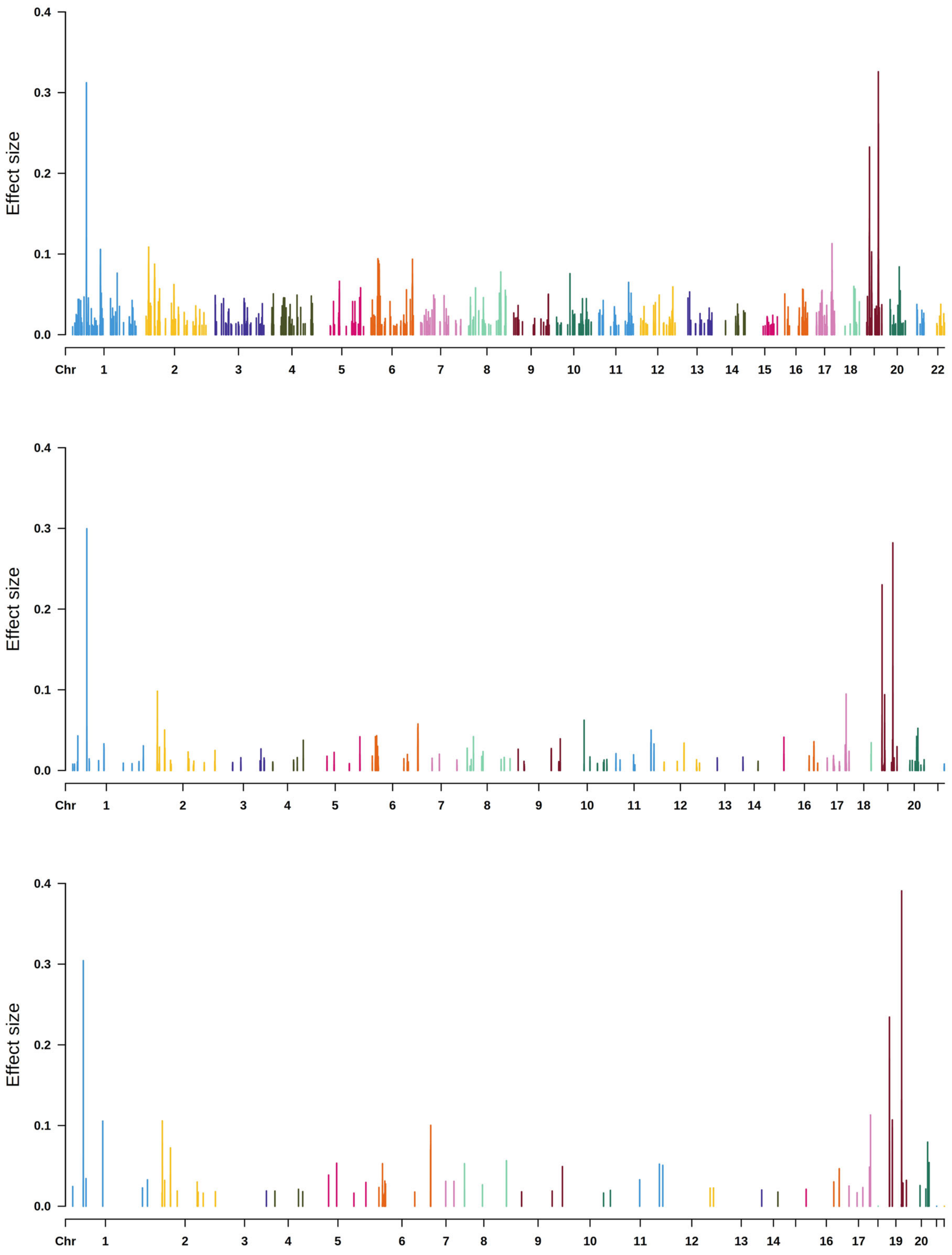

FIGURE 2 Absolute value of regression coefficients in the considered PRS models (the top, middle, and bottom figures correspond to C+T, boosting with probing, and AdaSub with EBIC_1) for LDL-cholesterol.

classical C+T approach. It can be observed that the statistical learning approaches for fine-mapping tend to select less variants than the classical C+T approach. Furthermore, this example illustrates that AdaSub with EBIC_1 does not necessarily select the variant with the highest marginal association (smallest *p* value).

Figure 2 displays the final estimated absolute regression coefficients $|\hat{\beta}_j|$ on their corresponding chromosome.

The top variants (largest absolute coefficients) in all three models are mainly located in genomic loci corresponding to the localization of well-known cholesterol-associated genes (main signal colocalize with LDL receptor and APOE genes in chromosome 19, a second peak is present in chromosome 1 at the level of the PCSK9 gene whose inhibition leads to a decrease of LDL; Sabatine, 2019).

The comparison of the included variants across the considered models shows that several variants were selected by all three approaches (mainly leading variants in significant loci). Concerning the two fine-mapped PRS, 9 and 36 variants were selected in the region surrounding the LDL receptor gene (1 MB upstream and downstream) using AdaSub with EBIC_1 and boosting with probing, respectively. Annotation using variant effect predictor (McLaren et al., 2016) revealed that four out of the nine variants selected by AdaSub were located in the regulatory region and one located in the splicing region for LDL. Notably, an upstream variant rs12151108 was previously reported in GWAS on LDL level among African population (Gurdasani et al., 2019) as well as in combined East Asian/European population (Nielsen et al., 2020). In addition to selecting eight out of nine variants included in the AdaSub model, boosting with probing selected 28 further variants covering 11 genes (Supporting Information: Table 1).

When comparing the sizes of estimated coefficients for the selected loci (see Figure 2), one can generally observe a very similar pattern among the three models, while there is a tendency toward smaller (shrunk) regression coefficients for the two PRS that were finally estimated via statistical boosting (see Section 2.3). This is particularly true for the PRS based on variants selected via probing, which may be due to the larger number of selected variants compared to the one selected by AdaSub while using the same number of $m_{\text{stop}} = 10{,}000$ iterations for the final boosting fit. Overall, the two fine-mapping methods via statistical learning identify less variants for PRS than the classical C+T approach, which could also increase the interpretability of the underlying models.

Of note, no fine-mapping takes place when applying LDpred2-inf and PRS-CS since these methods assume a genetic architecture in which all variants are casual, leading to omnigenic models based on all overlapping variants between the analyzed target data and the reference LD data set. On the other hand, multivariable Lasso regression via the BASIL algorithm (Qian et al., 2020) yields a final model with 12,492 selected variants based on genotyped data, which is sparser compared to the omnigenic models, but considerably larger compared to the fine-mapped AdaSub and probing models. Compared to the genome-wide analysis, the application of the Lasso based on only the chunk-based prefiltered variants results in a sparser model with 1821 selected variants, which is, however still larger than both the AdaSub and probing models.

3.2 | Prediction performance

A sparser PRS model might be advantageous for interpretation; however, an important aim of PRS modeling remains prediction. To assess the prediction performance of the differently derived polygenic scores for LDL, we computed the R^2 value given by the squared correlation between the observed and the predicted phenotypes on the European test cohort. The test cohort was composed of the remaining $n = 120{,}495$ self-reported British individuals with Caucasian genetic origin that were not used in the training cohort.

We considered full models (M_f) including the estimated $\widehat{\text{PRS}}$:

$$M_f : \text{Y} = \alpha + \gamma_{\text{PRS}}\widehat{\text{PRS}} + \gamma_1 \text{PC}_1 + \cdots + \gamma_{10}\text{PC}_{10} + \gamma_{\text{sex}}\,\text{sex} + \gamma_{\text{age}}\,\text{age} + \epsilon \quad (11)$$

and the corresponding covariate-only model (M_c):

$$M_c : \text{Y} = \alpha + \gamma_1 \text{PC}_1 + \cdots + \gamma_{10}\text{PC}_{10} + \gamma_{\text{sex}}\,\text{sex} + \gamma_{\text{age}}\,\text{age} + \epsilon. \quad (12)$$

Here, the variable Y denotes the outcome (LDL level as a continuous phenotype) and PC_k represents the kth principal component for $k = 1, \ldots, 10$. The R^2 attributable to the PRS (partial R^2) is defined as the difference between the R^2 of the full model M_f and the R^2 of the covariate-only model M_c.

There is large heterogeneity both in the sparsity of the final polygenic models M_f and in the reached prediction performance in terms of R^2 on the European population (see Table 1). Overall comparisons with the covariate-only model M_c suggest that the prediction accuracy for LDL-cholesterol is largely driven by genetic predisposition as summarized by PRS. In particular, the sparse multivariable AdaSub and probing approaches based on chunk-based fine-mapping yield a better prediction performance compared to the univariate C+T, LDpred2, and PRS-CS approaches based on summary statistics, which include more variants in their final models. Previous work on classical penalized-regression approaches (Qian et al., 2020) has already indicated that, in the presence of large population-based omics data enabling the training of multivariable models, it is possible to outperform univariate approaches based on summary statistics.

Also, in our analysis, multivariable Lasso regression applied on all genotyped variants reached the best prediction performance on the European population, further indicating that multivariable regression models are favorable compared to methods based on univariate summary statistics. As the Lasso selects substantially larger numbers of variants than the fine-mapping approaches via AdaSub and probing, this illustrates that the identification of sparser models based on the most informative variants can yield lower prediction performance compared to larger models including many variants with lower effect sizes. However, compared to the direct application of the Lasso on all genotyped variants, the Lasso still yields very similar prediction accuracy with less selected variants when applied only on the chunk-based prefiltered variants. Overall, the competitive performances that are obtained using smaller numbers of variants suggests that fine-mapping approaches via multivariable statistical learning are able to detect the most predictive variants which can then be further analyzed for biological interpretation.

3.3 | Generalizability

One could hypothesize that sparser PRS might be more robust toward deviations from the target population. To test the generalizability of the derived PRS models, as test cohort we considered the complete UK Biobank data after removing samples used in the training data set. As in similar works, to adjust for population stratification, we first fit a linear regression model considering the first 10 principal components (PC) in the training data set

$$\widehat{\mathrm{PRS}} = \alpha + \delta_1 \mathrm{PC}_1 + \cdots + \delta_{10} \mathrm{PC}_{10} + \epsilon \tag{13}$$

to predict PRS in the test cohort only based on the genetic ancestry (Fahed et al., 2020; Khera et al., 2019). The residualized PRS are then used to fit the full model with covariates. The prediction performance was evaluated by splitting the complete test cohort in different ethnic groups according to the individual genetic background. The estimation of the genetic ethnicity via PC projection with respect to 1000 Genomes Project samples was performed using the `bigsnpr` package (Privé et al., 2018). Samples were assigned to one of the five 1000 Genomes Project superpopulations (European: EUR, African: AFR, East Asian: EAS, South Asian: SAS, American: AMR) according to the Euclidean distance in the 10 PC space with respect to the population centers as described in Privé (2020). Only cohorts with more than 1000 individuals were included in the analysis (i.e., the AMR superpopulation which included only 230 samples was excluded).

The R^2 values for the different PRS as well as for the covariate-only model on all considered populations are reported in in Table 1 (13,585 samples were excluded from this analysis because they did not cluster to any superpopulation due to a likely partial mixed ancestry origin). Overall, one can observe that the highest total R^2 values are reached on the European population (after adjusting for population stratification). On the other hand, all PRS models performed substantially worse in non-European populations, though to a different extent. In particular, the sparse fine-mapped AdaSub and probing models tend to yield a lower reduction of prediction performance in out-of-target populations compared to the univariate PRS approaches: for example, for the African population, the three univariate PRS approaches C+T, LDpred2 and PRS-CS achieve only between 35% and 39.5% of their prediction performance on the European population, while the fine-mapping approaches AdaSub and probing yield 88.9% and 89.4% of their respective performance on the European population. Furthermore, the Lasso model without prefiltering appears to be less robust regarding the generalizability on different populations compared to the sparser Lasso model based on chunk-based prefiltered variants: for example, for the African population, the genome-wide Lasso model achieves a relative prediction performance of 50.8% compared to its performance on the European population, while the sparser Lasso model based on prefiltering yields a relative performance of 83.2%.

Despite these indicative results, the general pattern between sparsity and generalizability remains less clear across the different approaches implemented to model the polygenic architecture of LDL. The differences of PRS prediction across populations might be due to population-specific allele-frequencies, LD patterns, and effect sizes (possibly capturing also gene–environment interactions). While the presence of differences in allele frequencies could be overcome by matching the individual genetic scores with population-specific PRS distributions, different LD patterns and the heterogeneity of variant effects across populations are more difficult to assess. Concerning population-specific effect sizes, access to genome-wide association studies performed in different ancestries would be required. However, the presence of variable LD patterns can be simulated by altering the correlations between variants.

The simulation results obtained via permuting the position of variants inside randomly selected windows of size 1000 (leading in total to 1%, 5%, 10%, 25% permuted variants, respectively) revealed that the prediction performance is indeed strongly affected by the alteration of the correlation structure across nearby variants (Supporting Information: Table 2). This tendency is present for all models; as expected, larger amounts of changes in local variant correlations tend to imply lower generalizability. This is in line with the typical low performance of PRS models that are obtained in the African population which is characterized by a high level of allelic heterogeneity (Duncan et al., 2019). However, this is not a general rule: for instance, concerning the prediction of LDL-cholesterol, we obtained the lowest performance in UK Biobank for individuals with East Asian ethnic background (EAS), replicating what has also been recently observed in another work (Tanigawa et al., 2022). These findings further highlight the complexity of the issue of PRS generalizability, which is likely to depend on a combination of factors, among which population-specific LD patterns play a major role.

Noteworthy, the simulation results based on the perturbation of the local correlation structure showed that for larger variations nonsparse models like LDpred2-inf and PRS-CS can be more robust than sparse models. On the other hand, the second simulation scenario with joint permutations of variant values inside windows across observations (effectively “knocking-out” these variant effects on the outcome, e.g., representing genotyping errors or genotyping missingness) revealed that the sparse models obtained by AdaSub and probing tended to be more robust compared to the nonsparse models of LDpred2-inf and PRS-CS (Supporting Information: Table 3). Overall, these results suggest that further analyses are required to investigate the hypothesis that sparser and more carefully fine-mapped models tend to be more robust regarding the generalizability to different populations (Weissbrod et al., 2020).

4 | DISCUSSION

In this study, we have proposed and illustrated the application of existing statistical learning approaches for sparser fine-mapped polygenic risk models. These methods take advantage of the full genotype data and incorporate modern statistical modeling approaches as well as data-driven variable selection strategies in the fitting of PRS.

PRS are usually constructed via estimated effects from simple linear models representing the cumulative univariate/marginal effects of many common variants from GWAS (Choi et al., 2020; Wand et al., 2021). One of the major methodological limitations of relying only on univariate summary statistics is that the joint information and interdependence of effects of multiple variants cannot be fully assessed. Our proposed approaches, in contrast, directly apply multivariable regression models on the actual genotype data. For genetic fine-mapping, we consider the recently developed stochastic search method AdaSub (Staerk et al., 2021), as well as statistical gradient, boosting (Bühlmann & Hothorn, 2007). We employ these methods to yield particularly sparse models by incorporating the ℓ_0-penalized EBIC$_1$ (J. Chen & Chen, 2008) for AdaSub and by enforcing early stopping via probing for the statistical boosting algorithm (Thomas et al., 2017). To overcome the high computational burden and memory demand of applying these existing methods on large-scale data (large p and large n), we first split the genome into independent LD blocks incorporating a chunk-based screening approach and then identify the most informative variants for the phenotype inside these chunks. Afterward, the selected variants are combined into a final multivariable regression model—again fitted by a statistical boosting algorithm. Additionally, we consider the recent BASIL algorithm (Qian et al., 2020) to fit the ℓ_1-penalized Lasso (Tibshirani, 1996) directly on the full genotype data as well as on chunk-based prefiltered variants to further encourage the sparsity of the resulting model.

The proposed statistical learning methods based on AdaSub and probing are able to yield sparser and hence more interpretable multivariable risk models than classical methods based on univariate summary statistics. Even though the final models fitted via statistical boosting after fine-mapping with AdaSub and probing were not optimized for prediction performance, they showed a competitive performance on UK Biobank data regarding the prediction of LDL-cholesterol compared to classical PRS methods based on summary statistics such as C+T and different Bayesian approaches. With the sparse, fine-mapped AdaSub and probing models, around 13%–15% of the variability of LDL-cholesterol in our test cohort could be explained via genetic predisposition in the matched EUR population, after adjusting for age, sex and population stratification. Instead, the univariate C+T approach explains only around 5% of the LDL variability, a result in line with the $R^2 = 0.054$ obtained in the FinnGen cohort with a much larger PRS model including around 6M variants (Ripatti et al., 2020).

The Lasso fitted on full genotype data via the recently proposed BASIL algorithm (Qian et al., 2020) is able to yield even improved prediction performance on test data, which further demonstrates that the estimation and selection of variants in multivariable regression models is favorable compared to methods based on univariate summary statistics. However, the Lasso selected considerably more variants in the polygenic score for LDL-cholesterol compared to the fine-mapping approaches via AdaSub and probing. We further demonstrated that the number of selected variants in the Lasso PRS model can be substantially reduced when the Lasso is applied on the chunk-based prefiltered variants, resulting in a sparser model with still very competitive prediction performance.

Additionally, results of the sparser scores tended to be more robust when applied to different populations. The dependency of PRS performance on ancestry (Curtis, 2018) and the inherent disadvantage for individuals from populations with less genetic data (Duncan et al., 2019) or with multiethnic origin are urgent practical and ethical problems for the application of PRS in clinical practice (Lewis & Green, 2021). As our proposed techniques might not be the ultimate solution to these problems, further research is warranted on robust methods for developing PRS models in the presence of multiethnic populations and for incorporating scores in distant populations (Grinde et al., 2019; Ji et al., 2021).

Our simulation results showed that the model predictions are strongly affected by the LD structure. However, this is likely to be only one of the potential parameters influencing the PRS generalizability which is also depending on the underlying genetic architecture. For highly polygenic traits it might also be an advantage in terms of robustness to build large PRS models as they might be less sensitive to large variability of LD-patterns which may specifically occur in targeted significantly associated loci (Mars et al., 2022). Since population-specific LD is probably one of the major drivers of PRS differences across populations, the lack of PRS generalizability can be at least partially addressed by using ancestry-matched LD-reference panels (Ruan et al., 2022). Hopefully, with the increasing availability of data sets including individuals from different ancestries there will be further potentials to refine the PRS models; indeed, fine-mapping on multipopulations training cohorts can improve the generalizability of cross-population PRS models as recently shown by Weissbrod et al. (2022).

Limitations of our approach include that the computational burden is still relatively high and high-performance computing clusters are necessary to apply the proposed techniques on large cohorts. Furthermore, the considered fine-mapping methods are only feasible with access to full genotype data, which are associated with a higher memory demand and lower availability compared to summary statistics that are used for classical techniques. However, it can be expected that with the development of large population-based cohorts, such as UK Biobank (Bycroft et al., 2018), FinnGen (Locke et al., 2019), or Biobank Japan (Matoba et al., 2020), research will increasingly have access to full genotype data. Furthermore, this study focused on fine-mapping of variants in linkage disequilibrium, aiming to select only the most informative variants for LDL-cholesterol. Yet, in general, polygenic risk models with the best prediction performance may not be sparse depending on the considered phenotype (cf. Qian et al., 2020; Tanigawa et al., 2022). Future research should be targeted at adapting the considered methods when the main focus is not on the selection of variants but on the prediction performance.

Our three-step PRS approach, consisting of (1) screening, (2) fine-mapping (variant selection) and (3) final estimation via statistical boosting, has similarities with the recent batch screening approach for the Lasso by Qian et al. (2020), which we also incorporated into our framework as an alternative selection and multivariable estimation method after the initial screening. An important feature of this study is that we specifically aimed at fine-mapping in chunks of highly correlated variants to obtain particularly sparse PRS models. In contrast, the Lasso approach of Qian et al. (2020) employs variant selection and estimation in a single step derived from the ℓ_1-regularized optimization problem based on all genotyped variants. As a consequence, while our fine-mapping approaches via AdaSub and boosting with probing yield sparser PRS models, the joint estimation and selection of the Lasso yields improved predictions in the considered situation. Nevertheless, it has been shown that, in many practical situations, statistical boosting with implicit penalization can yield very similar coefficient paths as the direct penalization in the Lasso (Hepp et al., 2016), indicating that there may be room for improved predictions when considering a direct application of boosting without chunk-based fine-mapping via AdaSub and probing.

In future research, we want to extend the considered boosting and stochastic search methods for their efficient application on large-scale genotype data without considering prefiltered chunks for fine-mapping, aiming for an improved prediction performance at the potential cost of less sparse models. In this context, the main advantage of statistical boosting compared to the Lasso is the modular nature of the algorithm (Hofner et al., 2014; Mayr et al., 2014): any base-learner can be easily combined with any type of objective function. This leads to vast possibilities to extend the algorithm to further advanced statistical modeling techniques. In the current work, we have focused on the most classical combination of linear base-learners with the L_2 loss (Bühlmann & Yu, 2003), leading to linear regression models. Future research will be focused on considering other combinations of base-learners and loss functions. The most obvious choices are loss functions leading to logistic regression (for binary traits) and time-to-event models (for time-to-event traits). However, also the application of more robust loss functions (e.g., L_1) or objective functions that might be better suited to identify patients with a particularly high-risk (e.g., the check-function leading to quantile regression) might be promising.

ACKNOWLEDGMENTS

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Genome-wide genotyping data and phenotypic data from the UK Biobank are available upon successful project application to the UK Biobank (http://www.ukbiobank.ac.uk/about-biobank-uk/). Polygenic risk score models will be made accessible to the research community through PGS catalog (https://www.pgscatalog.org/).

ORCID

Carlo Maj http://orcid.org/0000-0002-9559-1725
Christian Staerk http://orcid.org/0000-0003-0526-0189
Hannah Klinkhammer https://orcid.org/0000-0003-3752-1275
Andreas Mayr http://orcid.org/0000-0001-7106-9732

REFERENCES

Ardlie, K. G., Kruglyak, L., & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, *3*(4), 299–309.

Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S., & Pirinen, M. (2016). Finemap: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, *32*(10), 1493–1501.

Berisa, T., & Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, *32*(2), 283–285.

Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting (with discussion). *Statistical Science*, *22*, 477–522.

Bühlmann, P., & Yu, B. (2003). Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association*, *98*(462), 324–339.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation plink: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742–015.

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp 785–794).

Choi, S. W., Mak, T. S.-H., & O'Reilly, P. F. (2020). Tutorial: A guide to performing polygenic risk score analyses. *Nature Protocols*, *15*(9), 2759–2772.

Curtis, D. (2018). Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatric Genetics*, *28*(5), 85–89.

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., & Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, *10*(1), 1–9.

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

Fahed, A. C., Wang, M., Homburger, J. R., Patel, A. P., Bick, A. G., Neben, C. L., Lai, C., Brockman, D., Philippakis, A., Ellinor, P. T., Cassa, C. A., Lebo, M., Ng, K., Lander, E. S., Zhou, A. Y., Kathiresan, S., & Khera, A. V. (2020). Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nature Communications*, *11*(1), 1–9.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849–911.

Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning Theory* (pp 148–156). Morgan Kaufmann Publishers Inc.

Furnival, G. M., & Wilson, R. W. (2000). Regressions by leaps and bounds. *Technometrics*, *42*(1), 69–79.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., & Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, *10*(1), 1–10.

Grinde, K. E., Qi, Q., Thornton, T. A., Liu, S., Shadyab, A. H., Chan, K. H. K., Reiner, A. P., & Sofer, T. (2019). Generalizing polygenic risk scores from Europeans to Hispanics/Latinos. *Genetic Epidemiology*, *43*(1), 50–62.

Gurdasani, D., Carstensen, T., Fatumo, S., Chen, G., Franklin, C. S., Prado-Martinez, J., Bouman, H., Abascal, F., Haber, M., Tachmazidou, I., Mathieson, I., Ekoru, K., DeGorter, M. K., Nsubuga, R. N., Finan, C., Wheeler, E., Chen, L., Cooper, D. N., Schiffels, S., … Sandhu, M. S. (2019). Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell*, *179*(4), 984–1002.

Hepp, T., Schmid, M., Gefeller, O., Waldmann, E., & Mayr, A. (2016). Approaches to regularized regression—A comparison between gradient boosting and the lasso. *Methods of Information in Medicine*, *55*(5), 422–430.

Hoffman, G. E., Logsdon, B. A., & Mezey, J. G. (2013). Puma: A unified framework for penalized multiple regression analysis of GWAS data. *PLoS Computational Biology*, *9*(6), e1003101.

Hofner, B., Mayr, A., Robinzonov, N., & Schmid, M. (2014). Model-based boosting in R: A hands-on tutorial using the R package mboost. *Computational Statistics*, *29*(1), 3–35.

Ji, Y., Long, J., Kweon, S.-S., Kang, D., Kubo, M., Park, B., Shu, X.-O., Zheng, W., Tao, R., & Li, B. (2021). Incorporating European GWAS findings improve polygenic risk prediction accuracy of breast cancer among East Asians. *Genetic Epidemiology*, *45*(5), 471–484.

Kachuri, L., Graff, R. E., Smith-Byrne, K., Meyers, T. J., Rashkin, S. R., Ziv, E., Witte, J. S., & Johansson, M. (2020). Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nature Communications*, *11*(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., & Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, *50*(9), 1219–1224.

Khera, A. V., Chaffin, M., Zekavat, S. M., Collins, R. L., Roselli, C., Natarajan, P., Lichtman, J. H., D'Onofrio, G., Mattera, J., Dreyer, R., Spertus, J. A., Taylor, K. D., Psaty, B. M., Rich, S. S., Post, W., Gupta, N., Gabriel, S., Lander, E., Chen, Y.-D. I., … Kathiresan, S. (2019). Whole-genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. *Circulation*, *139*(13), 1593–1602.

Lello, L., Avery, S. G., Tellier, L., Vazquez, A. I., de Los Campos, G., & Hsu, S. D. (2018). Accurate genomic prediction of human height. *Genetics*, *210*(2), 477–497.

Lewis, A. C., & Green, R. C. (2021). Polygenic risk scores in the clinic: New perspectives needed on familiar ethical issues. *Genome Medicine*, *13*(1), 1–10.

Li, Y. R., & Keating, B. J. (2014). Trans-ethnic genome-wide association studies: Advantages and challenges of mapping in diverse populations. *Genome Medicine*, *6*(10), 1–14.

Locke, A. E., Steinberg, K. M., Chiang, C. W., Service, S. K., Havulinna, A. S., Stell, L., Pirinen, M., Abel, H. J., Chiang, C. C., Fulton, R. S., Jackson, A. U., Kang, C. J., Kanchi, K. L., Koboldt, D. C., Larson, D., Nelson, J., Nicholas, T. J., Pietilä, A., Ramensky, V., … Freimer, N. B. (2019). Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature*, *572*(7769), 323–328.

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., & Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, *41*(6), 469–480.

Mars, N., Kerminen, S., Feng, Y.-C. A., Kanai, M., Läll, K., Thomas, L. F., Skogholt, A. H., della Briotta Parolo, P., Neale, B. M., Smoller, J. W., Gabrielsen, M. E., Hveem, K., Mägi, R., Matsuda, K., Okada, Y., Pirinen, M., Palotie, A., Ganna, A., Martin, A. R., & Ripatti, S. (2022). Genome-wide risk prediction of common diseases across ancestries in one million people. *Cell Genomics*, *2*(4), 100118.

Matoba, N., Akiyama, M., Ishigaki, K., Kanai, M., Takahashi, A., Momozawa, Y., Ikegawa, S., Ikeda, M., Iwata, N., Hirata, M., Matsuda, K., Murakami, Y., Kubo, M., Kamatani, Y., & Okada, Y. (2020). GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits. *Nature Human Behaviour*, *4*(3), 308–316.

Mayr, A., Binder, H., Gefeller, O., & Schmid, M. (2014). The evolution of boosting algorithms—From machine learning to statistical modelling. *Methods of Information in Medicine*, *53*(6), 419–427.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., & Cunningham, F. (2016). The ensemble variant effect predictor. *Genome Biology*, *17*(1), 1–14.

Nielsen, J. B., Rom, O., Surakka, I., Graham, S. E., Zhou, W., Roychowdhury, T., Fritsche, L. G., Taliun, S. A. G., Sidore, C., Liu, Y., Gabrielsen, M. E., Skogholt, A. H., Wolford, B., Overton, W., Zhao, Y., Chen, J., Zhang, H., Hornsby, W. E., Acheampong, A., … Hveem, K. (2020). Loss-of-function genomic variants highlight potential therapeutic targets for cardiovascular disease. *Nature Communications*, *11*(1), 1–12.

Pattee, J., & Pan, W. (2020). Penalized regression and model selection methods for polygenic scores on summary statistics. *PLoS Computational Biology*, *16*(10), e1008271.

Privé, F. (2020). Ancestry inference and grouping from principal component analysis of genetic data. bioRxiv. https://www.biorxiv.org/content/early/2020/10/26/2020.10.06.328203, https://doi.org/10.1101/2020.10.06.328203

Privé, F., Arbel, J., & Vilhjálmsson, B. J. (2020). Ldpred2: Better, faster, stronger. *Bioinformatics*, *36*(22–23), 5424–5431.

Privé, F., Aschard, H., Ziyatdinov, A., & Blum, M. G. (2018). Efficient analysis of large-scale genome-wide data with two R packages: Bigstatsr and bigsnpr. *Bioinformatics*, *34*(16), 2781–2787.

Qian, J., Tanigawa, Y., Du, W., Aguirre, M., Chang, C., Tibshirani, R., Rivas, M. A., & Hastie, T. (2020). A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK biobank. *PLoS Genetics*, *16*(10), e1009141.

Ripatti, P., Rämö, J. T., Mars, N. J., Fu, Y., Lin, J., Söderlund, S., Benner, C., Surakka, I., Kiiskinen, T., Havulinna, A. S., Palta, P., Freimer, N. B., Widen, E., Salomaa, V., Tukiainen, T., Pirinen, M., Palotie, A., Taskinen, M. R., Ripatti, S., & FinnGen. (2020). Polygenic hyperlipidemias and coronary artery disease risk. *Circulation: Genomic and Precision Medicine*, *13*(4), e002725.

Ruan, Y., Lin, Y.-F., Feng, Y.-C. A., Chen, C.-Y., Lam, M., Guo, Z., He, L., Sawa, A., Martin, A. R., Qin, S., Huang, H., & Ge, T. (2022). Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics*, *54*, 573–580.

Sabatine, M. S. (2019). Pcsk9 inhibitors: clinical evidence and implementation. *Nature Reviews Cardiology*, *16*(3), 155–165.

Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G. R., Wainberg, M., Ollila, H. M., Kiiskinen, T., Havulinna, A. S., Pirruccello, J. P., Qian, J., Shcherbina, A., Rodriguez, F., Assimes, T. L., Agarwala, V., Tibshirani, R., Hastie, T., … FinnGen (2021). Genetics of 35 blood and urine biomarkers in the UK biobank. *Nature Genetics*, *53*(2), 185–194.

Staerk, C., Kateri, M., & Ntzoufras, I. (2021). High-dimensional variable selection via low-dimensional adaptive learning. *Electronic Journal of Statistics*, *15*(1), 830–879.

Tanigawa, Y., Qian, J., Venkataraman, G., Justesen, J. M., Li, R., Tibshirani, R., Hastie, T., & Rivas, M. A. (2022). Significant sparse polygenic risk scores across 813 traits in UK Biobank. medRxiv. https://www.medrxiv.org/content/early/2022/01/27/2021.09.02.21262942, https://doi.org/10.1101/2021.09.02.21262942

Thomas, J., Hepp, T., Mayr, A., & Bischl, B. (2017). Probing for sparse and fast variable selection with model-based boosting. *Computational and Mathematical Methods in Medicine*, *2017*, 1421409.

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

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., Hayeck, T., Won, H. 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., … Price, A. L. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, *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., … Wojcik, G. L. (2021). Improving reporting standards for polygenic scores in risk prediction studies. *Nature*, *591*(7849), 211–219.

Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P. M., & Yengo, L. (2020). Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nature Communications*, *11*(1), 1–9.

Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A. P., Van De Geijn, B., Reshef, Y., Márquez-Luna, C., O'Connor, L., Pirinen, M., Finucane, H. K., & Price, A. L. (2020). Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nature Genetics*, *52*(12), 1355–1363.

Weissbrod, O., Kanai, M., Shi, H., Gazal, S., Peyrot, W. J., Khera, A. V., Okada, Y., Martin, A. R., Finucane, H. K., & Price, A. L. (2022). Leveraging fine-mapping and multi-population training data to improve cross-population polygenic risk scores. *Nature Genetics*, *54*, 450–458.

Wu, Y., Boos, D. D., & Stefanski, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association*, *102*(477), 235–243.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Maj, C., Staerk, C., Borisov, O., Klinkhammer, H., Yeung, M. W., Krawitz, P., & Mayr, A. (2022). Statistical learning for sparser fine-mapped polygenic models: The prediction of LDL-cholesterol. *Genetic Epidemiology*, 1–15. https://doi.org/10.1002/gepi.22495