

Advanced polygenic prediction models via statistical boosting

Doctoral thesis
to obtain a doctorate (PhD)
from the Faculty of Medicine
of the University of Bonn

Hannah Klinkhammer

from Simmerath

2025

Written with authorization of
the Faculty of Medicine of the University of Bonn

First reviewer: Prof. Dr. Peter M. Krawitz
Second reviewer: Priv.-Doz. Dr. Christoph Engel

Day of oral examination: 18.02.2025

From the Institute of Genomic Statistics and Bioinformatics
and the Institute of Medical Biometry, Informatics and Epidemiology

Table of Contents

List of abbreviations	4
1 Abstract	5
2 Introduction and aims with references	6
2.1 Thesis outline	7
2.1.1 Risk stratification in moderate penetrance genes	8
2.1.2 Statistical boosting for polygenic prediction models	9
2.1.3 Transferability of polygenic risk scores	10
2.2 References	10
3 Publications	13
3.1 Publication 1: Ability of a polygenic score to refine colorectal cancer risk in Lynch syndrome	14
3.2 Publication 2: A statistical boosting framework for polygenic risk scores based on large-scale genotype data	15
3.3 Publication 3: Genetic prediction modeling in large cohort studies via boosting targeted loss functions	32
3.4 Publication 4: Generalizability of polygenic prediction models: how is the R^2 defined on test data?	52
4 Discussion with references	63
4.1 Conclusion	65
4.2 References	65
5 Acknowledgments	70

List of abbreviations

CRC	Colorectal cancer
GWAS	Genome-wide association study
LD	Linkage disequilibrium
PRS	Polygenic risk score
SNP	Single-nucleotide polymorphism

1 Abstract

In times of growing availability of large biobanks with extensive genetic data, polygenic prediction modeling has gained importance and aims at capturing an individual's genetic predisposition to specific, often complex, traits. In contrast to monogenic diseases, complex traits are typically characterized by a limited genetic signal that is distributed across many genetic loci and based on common variants exhibiting only low to medium effect sizes. Additionally, common variants in close proximity are often highly correlated (linkage disequilibrium), increasing the statistical complexity of polygenic prediction modeling.

The aim of this cumulative dissertation is to enable advanced statistical modeling of polygenic risk scores (PRS) based on individual-level genotype data from large cohort studies. The first work underlines the potential of PRS to partly explain incomplete penetrance in monogenic conditions by analyzing patients diagnosed with Lynch syndrome, a monogenic condition increasing the risk for colorectal cancer. Here, PRS showed a higher potential for risk stratification in individuals with a variant in moderate penetrance genes compared to individuals with an affected high penetrance gene. PRS are commonly based on univariate effect estimates from genome-wide association studies. In the second work of this dissertation, the new statistical boosting framework *snptest* is introduced. The algorithm incorporates an additional batch-building step which substantially decreases the search space in each boosting iteration. By iteratively working on batches of variants, computational challenges are therefore solved, and multivariable modeling of PRS directly from individual-level genotype data via statistical boosting is made feasible for the first time. The third work in this dissertation extends the *snptest* framework to be applicable not only to Gaussian and binary data but also to time-to-event data, count data and quantile regression. Finally, the last included work emphasizes the importance of a thorough performance assessment of PRS. In this context, a major challenge of PRS is addressed, namely a strongly decreased prediction performance in out-of-target data, e.g. individuals of different ancestry than the training population. All research articles have been published in international peer-reviewed journals (see *Publications 1 - 4*).

2 Introduction and aims with references

Within the last decade, the importance of genetics for predicting disease risk and treatment responses has grown rapidly as genetic data have become increasingly available. Large cohort studies such as the UK Biobank (Bycroft et al., 2018) provide access to a wide range of genetic as well as phenotypic data to the scientific community. Monogenic conditions are most often caused by rare variants, e.g. in the genes *BRCA1* and *BRCA2* leading to an increased risk of breast cancer (Kuchenbaecker et al., 2017). In contrast, complex traits are usually characterized by a polygenic signal that is distributed across many genetic loci. The associated genetic variants are often common variants such as single-nucleotide polymorphisms (SNPs) which only have small to medium effect sizes. Polygenic risk scores (PRS) have been developed to aggregate those effects and aim to predict an individual genetic liability to a certain trait. By capturing the polygenic signal, PRS can help to stratify individuals with respect to their disease risk and to partly explain incomplete penetrance observed in many monogenic variants (Khera et al., 2018; Hassanin et al., 2021; Hassanin et al., 2023).

Genotype data encompassing common variants are high-dimensional (number of variants $p >$ number of individuals n) and include linkage disequilibrium (LD), i.e., variants that are in close proximity to each other are highly correlated. Therefore, PRS are typically derived from univariate effect estimates of the variants based on summary statistics from genome-wide association studies (GWAS, Choi et al., 2020). Subsequently, statistical approaches are deployed to correct for LD ranging from a basic clumping and thresholding approach (PRSice, Choi and O'Reilly, 2019), to approaches that rely on LD reference panels and are based on penalized regression such as lassosum (Mak et al., 2017) or on Bayesian theory such as PRSs (Ge et al., 2019) and LDpred (Privé et al., 2020). However, from a statistical point of view, prediction performance can often be improved by multivariable regression approaches that can yield sparse models by incorporating variable selection. Consequently, it is desirable to apply such an approach simultaneously to all variants to construct a single sparse prediction model. Nevertheless, applying multivariable regression to genotype data presents great challenges.

In addition to the high-dimensionality, genotype data from large cohort studies are expansive (large n and large p) and often cannot be processed easily in common statistical programs such as R, because they exceed the maximum available memory or increase the computational burden. While the first challenge of high-dimensionality can be addressed via regularized regression approaches or similar statistical learning algorithms, the latter proves more complex. In this context, Qian et al. (2020) developed a batch-based approach to derive PRS based on individual-level genotype data via lasso regression.

Statistical boosting is a statistical learning algorithm that can serve as an alternative to penalized regression approaches (Bühlmann and Hothorn, 2007; Mayr and Hofner, 2018). Originating from machine learning, boosting has two main components: (a) the loss function ρ defining the optimization problem and (b) the base-learners, i.e., weak learners that are iteratively combined to a single strong learner. For statistical boosting, the loss function is typically the negative log-likelihood of the assumed model and the base-learners are univariate regressions for each of the p variables, reaching from simple linear regressions to more complex, non-linear regressions. In each boosting iteration, the negative gradient of the loss function $-\frac{\partial}{\partial \mu} \rho$ is evaluated at the current predictor μ and each base-learner is fitted to this negative gradient. Then, the best fitting base-learner with respect to loss reduction is selected and the model is updated accordingly. Iteratively combining weak learners makes the algorithm applicable to high-dimensional data. Finally, the algorithm is stopped after a maximum number of boosting iterations or based on another stopping criterion, inducing implicit variable selection and typically yielding sparse models. The modular nature of statistical boosting provides a high flexibility and thus a wide range of modeling choices. However, due to the iterative fashion and the refitting of every base-learner in each boosting iteration, statistical boosting is computationally expensive and usually restricted to a medium sized number of covariates.

2.1 Thesis outline

This cumulative dissertation aims at developing advanced modeling techniques to derive PRS from individual-level genotype data of large cohort studies.

The first article illustrates the potential of PRS to stratify disease risk within individuals with monogenic predispositions for colorectal cancer (*Publication 1*). In the second and third article, the statistical boosting framework *snptest* is introduced (*Publication 2*) to derive PRS via multivariable statistical models. This framework is extended to adequately model different types of outcomes including time-to-event and count data by implementing various loss functions (*Publication 3*). Finally, the fourth article addresses one of the main challenges in PRS modeling, that is, the transferability of PRS to out-of-target populations (*Publication 4*).

The Appendix lists additional publications resulting from projects at the Institute for Genomic Statistics and Bioinformatics and the Institute for Medical Biometry, Informatics and Epidemiology during the years of this PhD. Alongside the PRS work, a major project was Gestalt-Matcher, which focused on facial analysis for patients with rare genetic disorders.

2.1.1 Risk stratification in moderate penetrance genes

Publication 1 serves as a motivating example of PRS as a risk modifier in moderate penetrance genes. Lynch syndrome is caused by pathogenic variants in mismatch repair genes such as *MLH1*, *MSH2* and *MSH6* and is the main cause of hereditary colorectal cancer (CRC), which accounts for approximately 5 % of all CRC cases (Czene et al., 2002). However, the penetrance of the variants differs between the affected genes, i.e., variants in *MLH1* are linked to a greater increase in CRC risk than variants in *MSH6* (Dominguez-Valentin et al., 2020). To investigate whether PRS can explain part of the reduced penetrance in some genes, *Publication 1* analyzed individuals diagnosed with Lynch syndrome and population controls in two studies, both using a published GWAS-based PRS for CRC as a covariable in Cox proportional hazard models with random effects. The results were combined in a fixed-effect meta-analysis. They included a positive correlation of PRS and cumulative lifetime risk in the moderate penetrance gene *MSH6* (non-significant), but not in the high penetrance gene *MLH1*. This suggests that PRS can add genetic information in cases of incomplete penetrance. While this publication used a published GWAS-based PRS, risk stratification could potentially be improved through more accurate multivariable PRS.

2.1.2 Statistical boosting for polygenic prediction models

To improve polygenic prediction models by usage of multivariable regression approaches and by exploiting the flexibility of boosting, our aim was to develop an adapted statistical boosting algorithm that can directly be applied on the complete individual-level genotype data. A first approach included a pre-filtering step substantially reducing the number of considered variants and worked on pre-defined LD-blocks (Maj et al., 2022). However, this approach incorporated several manual steps and was computationally demanding. Therefore, in *Publication 2* we introduced *snpboost*, a new framework to boost PRS directly from individual-level genotype data. Similar to the approach for lasso regression of Qian et al. (2020), we incorporated a batch-building step, substantially reducing the search space in each boosting iteration and thus the computational costs. The iterative algorithm now incorporates an outer loop, in which a batch of the $p_{\text{batch}} \ll p$ variants with the highest correlation to the current residual is created, and an inner loop, in which statistical boosting is applied on this batch of variants. A correlation-based stopping criterion for the inner loop prevents over-fitting on a single batch of variants. Additionally, we simultaneously monitor the loss on a validation set. The final model is reached if the validation loss has not decreased for b_{stop} batches. For more technical details we refer to the method section of *Publication 2*.

Publication 2 focused on continuous traits modeled via an L_2 -loss and binary traits modeled via a log-loss. We demonstrated that our approach yields sparse polygenic prediction models with better prediction accuracy than many GWAS-based PRS tools and comparable performance to the multivariable lasso regression approach by Qian et al. (2020). However, in biomedical research, outcomes of interest can also be neither continuous nor binary, but time-to-event data or count data. Since those outcomes exhibit particular challenges such as censoring and skewed distributions, it is important to model them adequately. Thus, in *Publication 3* we extended the *snpboost* framework by incorporating targeted loss functions for different optimization problems. Particularly, we also implemented a loss function to perform quantile regression allowing us to directly model the quantile of a distribution and therefore construct individual prediction intervals from genetic data.

2.1.3 Transferability of polygenic risk scores

Despite the advances made in polygenic prediction modeling, there are still challenges that need to be overcome. Besides further modeling choices such as non-linear effects or gene-gene and gene-environment interactions, one of the main challenges remains the transferability of PRS across different populations. There is an over-representation of individuals with European ancestry in the currently available large cohort studies (Sirugo et al., 2019). Therefore, most PRS are derived from European ancestry and their prediction performance strongly decreases for other ancestries such as African or Asian ancestries (Bitarello and Mathieson, 2020). *Publication 4* illustrates this effect for three traits, four ancestries and three PRS methods including *snpboost*. Particularly, based on different definitions of the R^2 measure on test data, the work highlights that for integrating PRS into the clinical practice it is not only important to assess the discriminative performance (R^2 as a correlation based measure) but also investigate possible miscalibration (R^2 as a prediction error based measure).

2.2 References

- Bitarello BD, Mathieson I. Polygenic Scores for Height in Admixed Populations. In: G3 Genes|Genomes|Genetics 2020; 10 (11): 4027–4036
- Bühlmann P, Hothorn T. Boosting Algorithms: Regularization, Prediction and Model Fitting. In: Statistical Science 2007; 22 (4): 477–505
- Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, 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. The UK Biobank resource with deep phenotyping and genomic data. In: Nature 2018; 562 (7726): 203–209
- Choi SW, Mak TSH, O’Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. In: Nature Protocols 2020; 15 (9): 2759–2772
- Choi SW, O’Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. In: GigaScience 2019; 8 (7): 1–6

- Czene K, Lichtenstein P, Hemminki K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish family-cancer database. In: *International Journal of Cancer* 2002; 99 (2): 260–266
- Dominguez-Valentin M, Sampson JR, Seppälä TT, Broeke SW ten, Plazzer JP, Nakken S, Engel C, Aretz S, Jenkins MA, Sunde L, Bernstein I, Capella G, Balaguer F, Thomas H, Evans DG, Burn J, Greenblatt M, Hovig E, Vos tot Nederveen Cappel WH de, Sijmons RH, Bertario L, Tibiletti MG, Cavestro GM, Lindblom A, Della Valle A, Lopez-Köstner F, Gluck N, Katz LH, Heinimann K, Vaccaro CA, Büttner R, Görgens H, Holinski-Feder E, Morak M, Holzapfel S, Hüneburg R, Knebel Doeberitz Mv, Loeffler M, Rahner N, Schackert HK, Steinke-Lange V, Schmiegel W, Vangala D, Pylvänäinen K, Renkonen-Sinisalo L, Hopper JL, Win AK, Haile RW, Lindor NM, Gallinger S, Le Marchand L, Newcomb PA, Figueiredo JC, Thibodeau SN, Wadt K, Therkildsen C, Okkels H, Ketabi Z, Moreira L, Sánchez A, Serra-Burriel M, Pineda M, Navarro M, Blanco I, Green K, Laloo F, Crosbie EJ, Hill J, Denton OG, Frayling IM, Rødland EA, Vasen H, Mints M, Neffa F, Esperon P, Alvarez K, Kariv R, Rosner G, Pinero TA, Gonzalez ML, Kalfayan P, Tjandra D, Winship IM, Macrae F, Möslin G, Mecklin JP, Nielsen M, Møller P. Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from the Prospective Lynch Syndrome Database. In: *Genetics in Medicine* 2020; 22 (1): 15–25
- Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. In: *Nature Communications* 2019; 10 (1): 1776
- Hassanin E, May P, Aldisi R, Spier I, Forstner AJ, Nöthen MM, Aretz S, Krawitz P, Bobbili DR, Maj C. Breast and prostate cancer risk: The interplay of polygenic risk, rare pathogenic germline variants, and family history. In: *Genetics in Medicine* 2021: 576–585
- Hassanin E, Spier I, Bobbili DR, Aldisi R, Klinkhammer H, David F, Dueñas N, Hüneburg R, Perne C, Brunet J, Capella G, Nöthen MM, Forstner AJ, Mayr A, Krawitz P, May P, Aretz S, Maj C. Clinically relevant combined effect of polygenic background, rare pathogenic germline variants, and family history on colorectal cancer incidence. In: *BMC Medical Genomics* 2023; 16 (1): 42

- Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. In: *Nature Genetics* 2018; 50 (9): 1219–1224
- Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips KA, Mooij TM, Roos-Blom MJ, Jervis S, Leeuwen FE van, Milne RL, Andrieu N, et al. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. In: *JAMA* 2017; 317 (23): 2402–2416
- Maj C, Staerk C, Borisov O, Klinkhammer H, Wai Yeung M, Krawitz P, Mayr A. Statistical learning for sparser fine-mapped polygenic models: The prediction of LDL-cholesterol. In: *Genetic Epidemiology* 2022; 46 (8): 589–603
- Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. In: *Genetic Epidemiology* 2017; 41 (6): 469–480
- Mayr A, Hofner B. Boosting for statistical modelling-A non-technical introduction. In: *Statistical Modelling* 2018; 18 (3–4): 365–384
- Privé F, Arbel J, Vilhjálmsdóttir BJ. LDpred2: better, faster, stronger. In: *Bioinformatics* 2020; 36 (22–23): 5424–5431
- Qian J, Tanigawa Y, Du W, Aguirre M, Chang C, Tibshirani R, Rivas MA, Hastie T. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. In: *PLOS Genetics* 2020; 16 (10): e1009141
- Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. In: *Cell* 2019; 177 (1): 26–31

3 Publications

3.1 Publication 1: Ability of a polygenic score to refine colorectal cancer risk in Lynch syndrome

Dueñas N*, Klinkhammer H*, Bonifaci N*, Spier I, Mayr A, Hassanin E, Diez-Villanueva A, Moreno V, Pineda M, Maj C, Capellà G, Aretz S, Brunet J. Ability of a polygenic score to refine colorectal cancer risk in Lynch syndrome. In: Journal of Medical Genetics 2023: 60: 1044-1051

*These authors are joint first authors.

<https://doi.org/10.1136/jmg-2023-109344>

Supplementary data can be found at:

<https://doi.org/10.1136/jmg-2023-109344>

3.2 Publication 2: A statistical boosting framework for polygenic risk scores based on large-scale genotype data

Klinkhammer H, Staerk C, Maj C, Krawitz PM, Mayr A. A statistical boosting framework for polygenic risk scores based on large-scale genotype data. In: *Frontiers in Genetics* 2023; 13: 1076440.

<https://doi.org/10.3389/fgene.2022.1076440>

Supplementary information can be found at:

<https://doi.org/10.3389/fgene.2022.1076440>

Implementations are available on GitHub:

<https://github.com/hklinkhammer/snpboost>



OPEN ACCESS

EDITED BY

Angelo Facchiano,
National Research Council (CNR), Italy

REVIEWED BY

Wenlan Chen,
St. Jude Children's Research Hospital,
United States
Tianyuan Lu,
McGill University, Canada

*CORRESPONDENCE

Hannah Klinkhammer,
✉ klinkhammer@imbie.uni-bonn.de

SPECIALTY SECTION

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

RECEIVED 21 October 2022

ACCEPTED 20 December 2022

PUBLISHED 10 January 2023

CITATION

Klinkhammer H, Staerk C, Maj C,
Krawitz PM and Mayr A (2023). A statistical
boosting framework for polygenic risk
scores based on large-scale
genotype data.
Front. Genet. 13:1076440.
doi: 10.3389/fgene.2022.1076440

COPYRIGHT

© 2023 Klinkhammer, Staerk, Maj, Krawitz
and Mayr. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

A statistical boosting framework for polygenic risk scores based on large-scale genotype data

Hannah Klinkhammer^{1,2*}, Christian Staerk¹, Carlo Maj^{2,3},
Peter Michael Krawitz² and Andreas Mayr¹

¹Institute for Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Bonn, Germany, ²Institute for Genomic Statistics and Bioinformatics, Medical Faculty, University of Bonn, Bonn, Germany, ³Center for Human Genetics, University of Marburg, Marburg, Germany

Polygenic risk scores (PRS) evaluate the individual genetic liability to a certain trait and are expected to play an increasingly important role in clinical risk stratification. Most often, PRS are estimated based on summary statistics of univariate effects derived from genome-wide association studies. To improve the predictive performance of PRS, it is desirable to fit multivariable models directly on the genetic data. Due to the large and high-dimensional data, a direct application of existing methods is often not feasible and new efficient algorithms are required to overcome the computational burden regarding efficiency and memory demands. We develop an adapted component-wise L_2 -boosting algorithm to fit genotype data from large cohort studies to continuous outcomes using linear base-learners for the genetic variants. Similar to the snpnet approach implementing lasso regression, the proposed snpboost approach iteratively works on smaller batches of variants. By restricting the set of possible base-learners in each boosting step to variants most correlated with the residuals from previous iterations, the computational efficiency can be substantially increased without losing prediction accuracy. Furthermore, for large-scale data based on various traits from the UK Biobank we show that our method yields competitive prediction accuracy and computational efficiency compared to the snpnet approach and further commonly used methods. Due to the modular structure of boosting, our framework can be further extended to construct PRS for different outcome data and effect types—we illustrate this for the prediction of binary traits.

KEYWORDS

polygenic risk score (PRS), high-dimensional data, variable selection, boosting, GWAS—genome-wide association study, prediction

1 Introduction

In times of next-generation sequencing and decreasing costs for whole genome sequencing, the amount of available genotype data has increased dramatically in recent years, giving rise to new genetic insights (Beesley et al., 2020; National Human Genome Research Institute, 2021).

Polygenic risk scores (PRS) measure the individual genetic liability to a certain trait and can provide relevant information in the context of disease-risk stratification. In contrast to high-impact monogenic variants, which are mostly rare and have a high effect size, PRS are derived from common variants such as single-nucleotide polymorphisms (SNPs) with low or medium effect sizes. Polygenic effects could also explain part of the incomplete penetrance seen in many identified monogenic variants, as for example in the genes BRCA1 and BRCA2 both leading to a highly increased risk of breast cancer (Kuchenbaecker et al., 2017). Recent studies on the UK Biobank suggest that high-impact monogenic variants, PRS and family history could contribute

additively to the risk of developing breast and prostate cancer (Hassanin et al., 2021). Despite these findings, PRS still lack to explain relevant parts of the estimated heritability of many traits.

PRS are typically derived as a sum of risk allele counts weighted by univariate effect estimates of the measured variants based on summary statistics from genome-wide association studies (GWAS) (Choi et al., 2020). Despite several approaches to account for linkage disequilibrium (LD, referring to the correlation structure between variants) and for the selection of informative variants (Euesden et al., 2014; Vilhjálmsson et al., 2015; Mak et al., 2017; Privé et al., 2021), the univariate structure of the estimation cannot fully account for interdependencies between the variants. For example, lassosum (Mak et al., 2017) adopts an L_1 penalty term and solves a lasso-like problem while only using summary statistics and a LD reference panel. However, as published summary statistics and LD reference panels are most often based on different samples, lassosum can generally only approximate the full lasso path. A natural extension of using effect estimates from univariate models could hence be to fit a single multivariable model. While this approach seems natural from a methodological perspective, a direct application of existing methods is typically infeasible due to the high dimensionality of the genotype data, which can easily exceed the available computer memory. Recently, some approaches have been proposed to overcome this computational burden (Privé et al., 2018; Qian et al., 2020; Maj et al., 2022). In particular, Qian et al. proposed the so-called batch screening iterative lasso (BASIL) algorithm to fit the lasso on the complete original genotype data (Tibshirani, 1996; Qian et al., 2020; Li et al., 2022). The algorithm works on subsets of variants and computes the complete lasso path in an iterative fashion. Apart from the lasso, the algorithm can also be extended to other penalized regression methods such as the relaxed lasso (Meinshausen, 2007) or the elastic net (Zou and Hastie, 2005). In this context, Qian et al. were able to demonstrate that multivariable regularized PRS models fitted via the BASIL algorithm outperform the classical GWAS-based PRS for various traits such as height and high cholesterol.

While penalized regression models like the lasso and the elastic net impose explicit regularization, statistical boosting represents an alternative approach by introducing an implicit algorithmic regularization when combined with early stopping (Bühlmann and Hothorn, 2007; Mayr and Hofner, 2018). Boosting algorithms iteratively fit pre-defined base-learners to the gradient of the loss function, selecting the most influential base-learner in each step. The main tuning parameter of boosting algorithms is the number of iterations, which enables implicit variable selection and leads to sparse models. Due to its modular structure, boosting allows to combine possible base-learners with any convex loss function. These algorithms hence offer a great flexibility for statistical modelling, including various response types and the estimation of non-linear or other types of effects. A recent work has incorporated boosting into PRS modelling via a three-step approach (Maj et al., 2022): First, a marginal screening approach was applied on all variants to identify potentially informative ones. Then, multivariable algorithms including probing with boosting (Thomas et al., 2017) were applied on blocks of variants in LD to select (“fine-map”) the most informative variants. Finally, a statistical boosting model was fitted on the variant set created by joining the selected variants of all chunks. This approach yielded particularly sparse and interpretable models, whose predictive performance was superior to PRS derived by univariate methods like clumping and

thresholding (Euesden et al., 2014) and was outperformed by the predictive performance achieved by the lasso via the BASIL algorithm. However this approach includes pre-filtering of the variants and is computationally demanding.

In this article we introduce a new framework to boost PRS, starting with a new computational approach to build L_2 -boosting models on large-scale genotype data for quantitative traits. Similar to the snpnet approach for the lasso, our algorithm iteratively works on smaller batches of variants. Yet, in contrast to recent boosting methods (Staerk and Mayr, 2021; Maj et al., 2022), the variants do not need to be pre-filtered in our snpboost approach and the batches are not pre-defined or randomly sampled, but chosen iteratively and deterministically in a data-driven way based on the correlations of the variants to the remaining residuals. By restricting the set of available base-learners in each step to those variants which were most correlated with residuals from a previous iteration, we are able to reduce the search space and decrease the computational time compared to a classical component-wise boosting algorithm.

We conducted a simulation study to examine the performance of our adapted boosting algorithm snpboost compared to the original L_2 -boosting on a reduced but still high-dimensional data set, on which the application of standard L_2 -boosting was still computationally feasible. Furthermore, we simulated data of higher dimensionality and larger sample size to investigate the influence of various hyperparameters (including the batch size) on the prediction accuracy and computational burden of the snpboost approach in a typical large-scale setting. We discuss reasonable default values for the hyperparameters which are incorporated in the provided R implementation (<https://github.com/hklinkhammer/snpboost>). Finally, we constructed multivariable PRS for various traits on data from the UK Biobank via application of snpboost and compared the performance of our approach to the lasso estimates from the BASIL algorithm proposed by Qian et al. as well as to further commonly used methods. On the examined phenotypes we found highly comparable predictive performance while our adapted boosting approach had a tendency to select sparser models compared to the lasso and the other methods. Finally, we illustrate how the framework can be conveniently extended to the classification of binary phenotypes by the incorporation of different loss functions.

2 Methods

For $n \in \mathbb{N}$ individuals, let $\mathbf{y} = (y_1, \dots, y_n)' \in \mathbb{R}^n$ denote a particular continuous phenotype of interest. Furthermore, let X_j correspond to the genetic variant j , for $j = 1, \dots, p$. The observed dosage data of n individuals is given in the genotype matrix $\mathbf{X} = (x_{i,j}) \in [0, 2]^{n \times p}$, where $\mathbf{x}_j \in [0, 2]^n$ corresponds to the j th column of $\mathbf{X}$. We consider a linear regression model

$$\mathbb{E}(y_i | \mathbf{X}) = \beta_0 + \sum_{j=1}^p \beta_j x_{i,j}, \quad i = 1, \dots, n, \quad (1)$$

With coefficients $\beta_0 \in \mathbb{R}$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)' \in \mathbb{R}^p$. The aim is to determine coefficients $\hat{\beta}_0 \in \mathbb{R}$ and $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ such that the estimator $\hat{\mathbf{y}} = \hat{\beta}_0 + \mathbf{X}\hat{\boldsymbol{\beta}}$ minimizes the mean squared error of prediction on an independent test set $\text{MSEP} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\hat{y}_{\text{test},i} - y_{\text{test},i})^2$. Additionally, one is often interested in relatively sparse models in the sense that only a fraction of the coefficient vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is non-zero.

In high-dimensional settings with $p > n$ it is not feasible to apply classical estimation techniques like the ordinary least squares estimator. A commonly-used solution is to consider further constraints on the coefficient vector resulting in penalized regression methods including the lasso (Tibshirani, 1996). The lasso incorporates an L_1 -penalty on the coefficient vector such that the lasso estimate $\hat{\beta}^{\text{lasso}}$ is given by

$$\hat{\beta}^{\text{lasso}} = \underset{(\beta_0, \beta)}{\operatorname{argmin}} \left[\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right] \quad (2)$$

for some $\lambda \geq 0$. The explicit L_1 -penalization of the coefficient vector leads to shrinkage of the coefficient estimates. In contrast to ridge regression (Hoerl and Kennard, 2000), the use of the L_1 -penalty enables to set some parameters exactly to zero corresponding to sparse models. There has been extensive research on the theoretical properties of the lasso including oracle inequalities in high-dimensional settings (e.g., Fu and Knight (2000); Greenshtein and Ritov (2004); Bunea et al. (2007); van de Geer (2008)). Nevertheless, there are situations leading to variable selection problems of the lasso, particularly in the presence of high correlations between signal and noise variables (Hepp et al., 2016). When working with genotype data, high correlations between signal and noise variables might often be present as a result of LD, i.e., genetic variants that have close positions on the DNA strand tend to be highly correlated.

An alternative to explicitly penalized regression methods such as the lasso is statistical gradient boosting (Bühlmann and Hothorn, 2007; Mayr and Hofner, 2018). Gradient boosting requires the specification of a loss function $f(y, \hat{y})$ and so-called base-learners h_j that are iteratively fitted to the response. In detail, the aim is again to fit the linear regression model (1) which is performed in an iterative fashion. Starting at iteration $m = 0$ with a starting value $\hat{y}^{(0)} = 0$, the following steps are repeated until a maximum number m_{stop} of boosting iterations is reached (Bühlmann and Hothorn, 2007):

1. Set $m := m+1$ and compute the negative gradient vector of the loss function:

$$\mathbf{u}^{(m)} = - \frac{\partial f(y, \hat{y})}{\partial \hat{y}} \Big|_{\hat{y} = \hat{y}^{(m-1)}}$$

2. Fit every base-learner h_j separately to the negative gradient vector $\mathbf{u}^{(m)}$ and select the best fitting base-learner $\hat{h}_{j^*}^{(m)}(X_j)$.
3. Update the predictor with the learning rate $0 \leq \nu \leq 1$: $\hat{y}^{(m)} = \hat{y}^{(m-1)} + \nu \hat{h}_{j^*}^{(m)}(X_j)$
4. Stop if $m = m_{\text{stop}}$.

Stopping the algorithm before it converges (early stopping) leads to implicit regularization and shrinkage of effect estimates. The component-wise L_2 -boosting algorithm (Bühlmann and Yu, 2003; Bühlmann and Hothorn, 2007) employs the squared error $f(y, \hat{y}) = \|y - \hat{y}\|_2^2$ as a loss function (Bühlmann and Yu, 2003) and separate univariate linear regression models of the residuals on the j th genetic variant as base-learners (i.e., $h_j(X_j) = \beta_0 + \beta_j X_j$, for $j = 1, \dots, p$). In low-dimensional ($p < n$) settings this set-up mimics a classical Gaussian linear model and converges to the least squares solution for large values of m_{stop} . The general boosting procedure can be interpreted as gradient descent in function space, where the residual vector

represents the gradient of the L_2 loss and the function space is provided by the different base-learner solutions (Friedman, 2001; Bühlmann and Yu, 2003; Mayr and Hofner, 2018). The previously described steps transform therefore into the following procedure (shown in grey in Figure 1): The best fitting base-learner in boosting step $m+1$ corresponds to the variant j^* with the highest Pearson correlation $\rho(x_{j^*}, \mathbf{r}^{(m)})$ to the residuals $\mathbf{r}^{(m)} = \mathbf{y} - \hat{\mathbf{y}}^{(m)}$ resulting from the previous boosting step m . We then fit a linear regression model of the current residuals $\mathbf{r}^{(m)}$ on the variant j^* and update the corresponding coefficient $\hat{\beta}_{j^*}^{(m+1)}$ as well as the intercept $\hat{\beta}_0^{(m+1)}$. This is repeated until a maximum number of boosting iterations is reached or any other early stopping criterion is fulfilled. If additional covariates apart from the genetic variants are included in the model, they are treated as mandatory covariates—similar to the intercept. The additional covariates are included in each single base-learner and are hence updated in each boosting step without competing with the genetic variants.

Hepp et al. (2016) investigated the commonalities and differences between the lasso and statistical boosting: while there are (low-dimensional) settings in which the gradient boosting approximates the lasso coefficient paths arbitrarily close when the learning rate ν is approaching 0, their results generally differ if the coefficient paths are not monotone. The authors note that, in contrast to the lasso which limits the sum of the absolute values of the coefficients for each penalty parameter λ separately, boosting limits the total L_1 -arc-length of all coefficient curves (Hepp et al., 2016). Interpreting this as the total absolute distance “travelled” by all coefficients among the coefficient paths through the iterations $m = 1, \dots, m_{\text{stop}}$, it becomes clear that the solution in a certain iteration depends on all previous solutions of the iterative algorithm. This might lead to more stable pathways particularly in settings with high correlations between independent variables, which is typical for genetic data. Hepp et al. conducted several numerical experiments including high-dimensional settings in which they found similar predictive performance of lasso and boosting. In detail, boosting tended to yield slightly better prediction results while the lasso tended to result in sparser models with faster computations. On the other hand, the boosting algorithm can be easily extended to different response types as well as to different effects, including non-linear and interaction effects. In terms of genetic data, interaction effects can be used to model and identify epistatic effects and gene-environment interactions.

When working on genetic data from large cohort studies we do not only face a high-dimensional setting with $p > n$ but also a large-scale setting with large sample sizes n and large numbers of variants p . Large-scale settings often lead to extended computational times as well as memory issues. To overcome these and apply statistical boosting on genotype data, we implemented an adapted component-wise L_2 -boosting algorithm that is built on the snpnet framework (Qian et al., 2020) and works on batches of variants. To do so, we additionally incorporate a batch-building step before starting the boosting iterations (shown in blue in Figure 1). In this step we extract the p_{batch} variants ($p_{\text{batch}} \ll p$) with the highest correlation $\rho(x_j, \mathbf{r}^{(m)})$ to the current residual vector and include them in the batch B_k . A maximum number of m_{batch} boosting iterations is performed on batch B_k before the next batch is built based on the correlations of all p variants to the updated residuals. In total, we fit a maximum of b_{max} batches or stop early if an early stopping criterion is fulfilled. The algorithm is summarised in Table 1 and Figure 1.

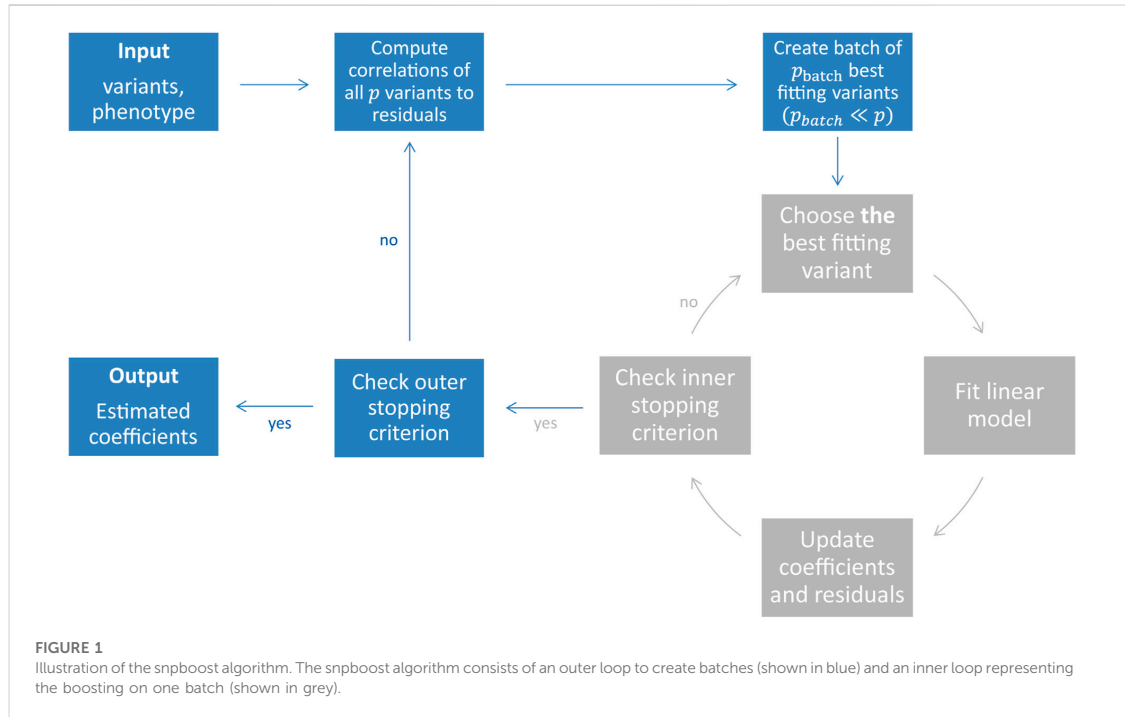


TABLE 1 Definition of the snpboost algorithm without additional covariates. If additional covariates apart from the genetic variants should be included in the model, they are treated as mandatory covariates—similar to the intercept. The additional covariates are included in each single base-learner and are hence updated in each boosting step without competing with the genetic variants.

Algorithm: snpboost	
Input: phenotype data $\mathbf{y} \in \mathbb{R}^n$, genotype data $\mathbf{X} \in \{0, 2\}^{n \times p}$, batch size $p_{\text{batch}} \in \{1, \dots, p\}$, learning rate $\nu > 0$, maximum number of boosting iterations per batch $m_{\text{batch}} \in \mathbb{N}$, maximum number of batches $b_{\text{max}} \in \mathbb{N}$, stopping lag for outer stopping criterion $b_{\text{stop}} \in \mathbb{N}$.	
Algorithm:	
1. Initialization: Set boosting index $m = 0$, residuals $\mathbf{r}^{(0)} = \mathbf{y} - \bar{\mathbf{y}}$, coefficients $\beta_j^{(0)} = \bar{y}$, $\beta_j^{(0)} = 0$, $j = 1, \dots, p$.	
2. Outer loop: Set outer counter $k = 1$.	
a. Screening: Compute correlations $c_j^{(m)} = \rho(\mathbf{r}^{(m)}, \mathbf{x}_j)$, $j = 1, \dots, p$. Create batch B_k of p_{batch} variants with highest absolute correlations $ c_j^{(m)} $. Save the highest absolute correlation outside the batch $c_{\text{stop}} = \max_{j \notin B_k} c_j^{(m)} $.	
b. Inner loop: Set inner counter $l = 1$.	
(1) If $l > 1$, compute correlations inside batch: $c_j^{(m)} = \rho(\mathbf{r}^{(m)}, \mathbf{x}_j)$, $j \in B_k$.	
(2) Choose variant j^* with the highest absolute correlation $ c_{j^*}^{(m)} = \max_{j \in B_k} c_j^{(m)} $. If the current maximum absolute correlation inside the batch is smaller than the highest correlation outside the batch, i.e. if $ c_{j^*}^{(m)} < c_{\text{stop}}$, stop the inner loop; else set $m := m + 1$.	
(3) Fit linear model: $\mathbb{E}(\mathbf{r}^{(m-1)} \mathbf{X}) = \beta_0 + \beta_{j^*} \mathbf{x}_{j^*}$.	
(4) Update coefficients with learning rate ν and β from iii.: $\beta_0^{(m)} = \beta_0^{(m-1)} + \nu \beta_0$, $\beta_{j^*}^{(m)} = \beta_{j^*}^{(m-1)} + \nu \beta_{j^*}$, $\beta_j^{(m)} = \beta_j^{(m-1)}$, $j \in \{1, \dots, p\} \setminus \{j^*\}$ as well as residuals $\mathbf{r}^{(m)} = \mathbf{y} - \beta_0^{(m)} - \beta_{j^*}^{(m)} \mathbf{x}_{j^*}$	
(5) If $l = m_{\text{batch}}$, end inner loop, else increase inner counter $l := l + 1$.	
c. If $k = b_{\text{max}}$ or if the MSE on the validation set has not decreased for b_{stop} batches, stop the outer loop; else increase the outer counter $k := k + 1$.	
3. Final model choice: Find $m_{\text{stop}} \in \{1, \dots, m\}$ corresponding to the lowest MSE on the validation set. The final coefficient estimates are given by $\beta_0^{(m_{\text{stop}})} \in \mathbb{R}$ and $\beta^{(m_{\text{stop}})} \in \mathbb{R}^p$.	

By iteratively working on batches of variants we save computational time and memory because only parts of the variants have to be loaded into memory at once. Additionally, not every step

requires the calculation of all potential base-learner solutions and the updated correlations for all variants. By this, we encourage additional sparsity by restricting the search space in terms of the set of available base-learners (as variants not included in the current batch cannot be selected). To examine when a new set of base-learners should be considered, which corresponds to the question when to stop the inner loop (inside the batches) and create a new batch of variants, we incorporated another step: we monitor the correlations of the variants inside a batch to the residuals and compare them to the correlations of variants outside of the batch. When creating a batch B_k we therefore compute and store the highest outer correlation $c_{\text{stop}} := \max_{j \notin B_k} |\rho(\mathbf{r}^{(m)}, \mathbf{x}_j)|$. After each boosting step m we check if the greatest absolute correlation of the variants inside the batch B_k to the current residual vector $\mathbf{r}^{(m)}$ is smaller than c_{stop} :

$$c_{\text{stop}} > \max_{j \in B_k} |\rho(\mathbf{r}^{(m)}, \mathbf{x}_j)|. \quad (3)$$

If inequality Eq. 3 holds true, we stop the inner loop and create a new batch since a variant outside the batch may provide a better fit to the current residual vector. In the original L_2 -boosting without batches, the variant with the highest correlation to the residuals would be chosen in each boosting step. The incorporation of batches in general limits this choice to the variants inside the batch. However, the proposed stopping criterion provides an indication to consider variants outside the batch which may be higher correlated with the current residuals. Actually, if all variants were independent, the proposed stopping criterion would lead to the same choice of variants in each boosting step in snpboost as in the original L_2 -boosting. Despite LD, our simulation results show that the

proposed stopping criterion yields reasonable variant choices and results in a competitive predictive performance (Section 3.1.1). Additionally, the inner loop is also stopped if the number of updates inside the batch reaches m_{batch} .

Furthermore, we need to determine after how many batches the algorithm should terminate. In classical statistical boosting the number of boosting iterations is often selected by cross-validation or resampling techniques—mimicking an additional data set to validate the predictive performance of the resulting models. However, if the data set is large enough, one can also directly divide the data into training and validation set. As in Qian et al. (2020), we hence simultaneously monitor the predictive performance of our model on an independent validation set while fitting on the training set. As a validation criterion for the predictive performance we use the MSE on the validation set. The outer loop consisting of the batch-building step is stopped if the MSE on the validation set has not decreased for b_{stop} batches or after a maximum number of b_{max} batches have been processed.

The proposed method is implemented as an add-on to the snpnet package by Qian et al. (2020) in the statistical computing environment R (R Core Team (2021), <https://github.com/hklinkhammer/snpboost>). While we are also incorporating PLINK 2.0 (Chang et al., 2015) to compute the correlations and build the batches in the outer loop, we replaced the fitting of the lasso by the adapted component-wise L_2 -boosting algorithm on the resulting batches (Table 1; Figure 1).

3 Empirical results

3.1 Simulation study

We conducted a simulation study to investigate the behaviour of the proposed snpboost algorithm in various controlled data scenarios. The simulation study aims at two main goals: first, to examine potential differences in performance compared to the original component-wise L_2 -boosting (Bühlmann and Yu, 2003) in smaller settings and, second, to gain insights on how to choose the included hyperparameters in practical situations.

Simulations are based on the UK Biobank genotype data (Bycroft et al., 2018) obtained under application number 81202 combined with simulated phenotypes. We restricted the individuals to white British ancestry and used the PLINK 2.0 function `thin-indiv-count` to randomly sample n individuals, of which 50%, 20% and 30% were assigned to the training, validation and test set, respectively (Chang et al., 2015; Purcell and Chang, 2015). Then, p variants with minor allele frequency not less than 1% were randomly sampled using PLINK 2.0's `thin-count`. Missing genotypes were replaced by the reference allele using the R package `bigsnpr` (Privé et al., 2018).

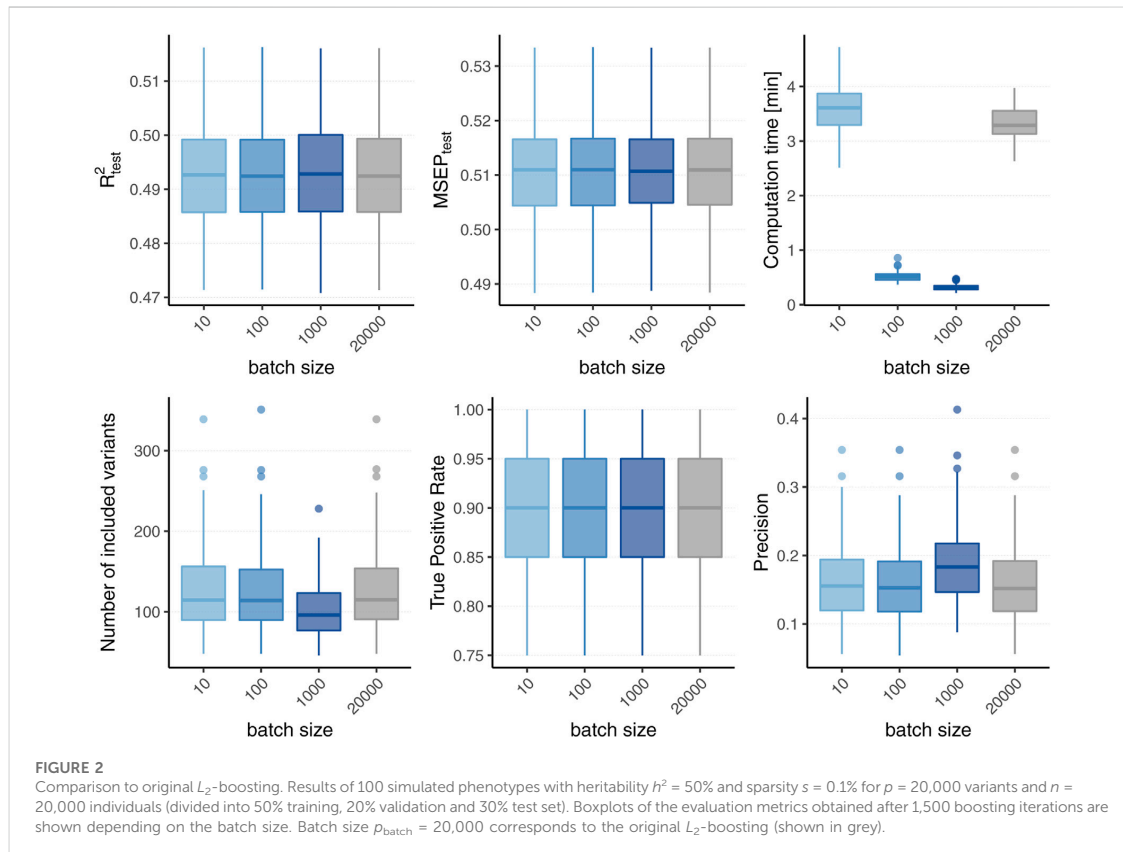
Continuous phenotypes were simulated from a linear model with Gaussian distributed noise and effect sizes using `bigsnpr`. To account for different genetic architectures, we considered varying heritability h^2 and sparsity s , defined as the amount of variance explained by the genetic liability and the proportion of causal variants, respectively. For each setting of h^2 and s , we simulated 100 different datasets. PRS models were derived by snpboost and evaluated by using various metrics regarding the predictive performance and the accuracy of the estimated coefficients. In detail, the predictive performance was measured by the MSE and the R^2 value defined as the squared

correlation between the predicted and the true phenotype on the independent test set. To assess the computational efficiency we measured the computation time of the algorithm. The accuracy of the resulting estimates was evaluated by the number of included variants in the final model and the mean squared error (MSE) of the estimates as well as the true positive (TP) rates and precision regarding variant selection. Additional results for all considered settings as well as comparisons to snpnet can be found in the Supplementary Material (Supplementary Figures S1–S6).

3.1.1 Comparison to original L_2 -boosting in smaller settings

To analyse the performance of snpboost compared to the original component-wise L_2 -boosting algorithm (Bühlmann and Yu, 2003), we used a single large batch with batch size $p_{\text{batch}} = p$ in the snpboost algorithm on simulated data with reduced dimensionality. We then compared the results to the ones derived by using smaller batches in terms of predictive performance, computation time, mean squared errors of the estimated coefficients as well as true positive rates and precision regarding variant selection. The simulations were conducted for $n = 20,000$ observations (10,000 training set, 4,000 validation set, 6,000 test set) and $p = 20,000$ variants as well as for varying degrees of heritability and sparsity. To obtain comparable results we chose a fixed number of boosting iterations independent of the batch size p_{batch} and a fixed learning rate $\gamma = 0.1$. For each simulation, 10 CPUs with 1 GB memory each were used.

Figure 2 displays the boxplots of each metric obtained after 1,500 boosting iterations for heritability $h^2 = 50\%$ and sparsity $s = 0.1\%$, i.e., 20 influential variants. Incorporating batches did not largely affect the predictive performance in terms of R^2 and MSE nor the MSE of the coefficient estimates (MSE results not shown). However, different batch sizes do not always yield the same models as L_2 -boosting as can be observed from the number of variants included in the final models. The models resulting from a batch size of $p_{\text{batch}} = 1,000$ tend to contain less variants than the ones from the original L_2 -boosting (batch size $p_{\text{batch}} = 20,000$). This could be explained by the reduced search space in each boosting step and a trade-off between exploration (genome-wide search) and exploitation (search inside the batch). As a consequence, variants within the batch that are already in the model are more often updated instead of including new variants outside of the batch. The same holds true when comparing the number of chosen variants for batch size $p_{\text{batch}} = 1,000$ to smaller batch sizes (i.e., $p_{\text{batch}} = 10$ and $p_{\text{batch}} = 100$). As all models tend to overestimate the number of influential variants, the lower number of selected variants for batch size $p_{\text{batch}} = 1,000$ corresponds to a higher precision since less false positives are included. The fact that the other metrics remain almost constant suggests that either only variants with very small effects are not included when using a larger batch size or the variants that are updated are highly correlated with the ones not included. Furthermore, incorporating batches in the algorithm has a major effect on the computation time. To interpret the results shown in Figure 2 it is important to understand the two drivers of the computation time. On the one hand, it increases with the number of correlations that have to be calculated in each boosting step which explains the increased computation time of the original L_2 -boosting (i.e., a batch size of $p_{\text{batch}} = 20,000$ and 20,000 computed correlations in each boosting step) compared to smaller batch sizes such as $p_{\text{batch}} = 100$ and $p_{\text{batch}} = 1,000$. On the other hand, reading the genotype data from disk when building the batches also increases the



computation time leading to a higher computation time for smaller batches with $p_{\text{batch}} = 10$ for which more reads-from-disk have to be carried out. The varying computation times therefore reflect a trade-off between the number of correlations computed in each boosting step and the number of created batches.

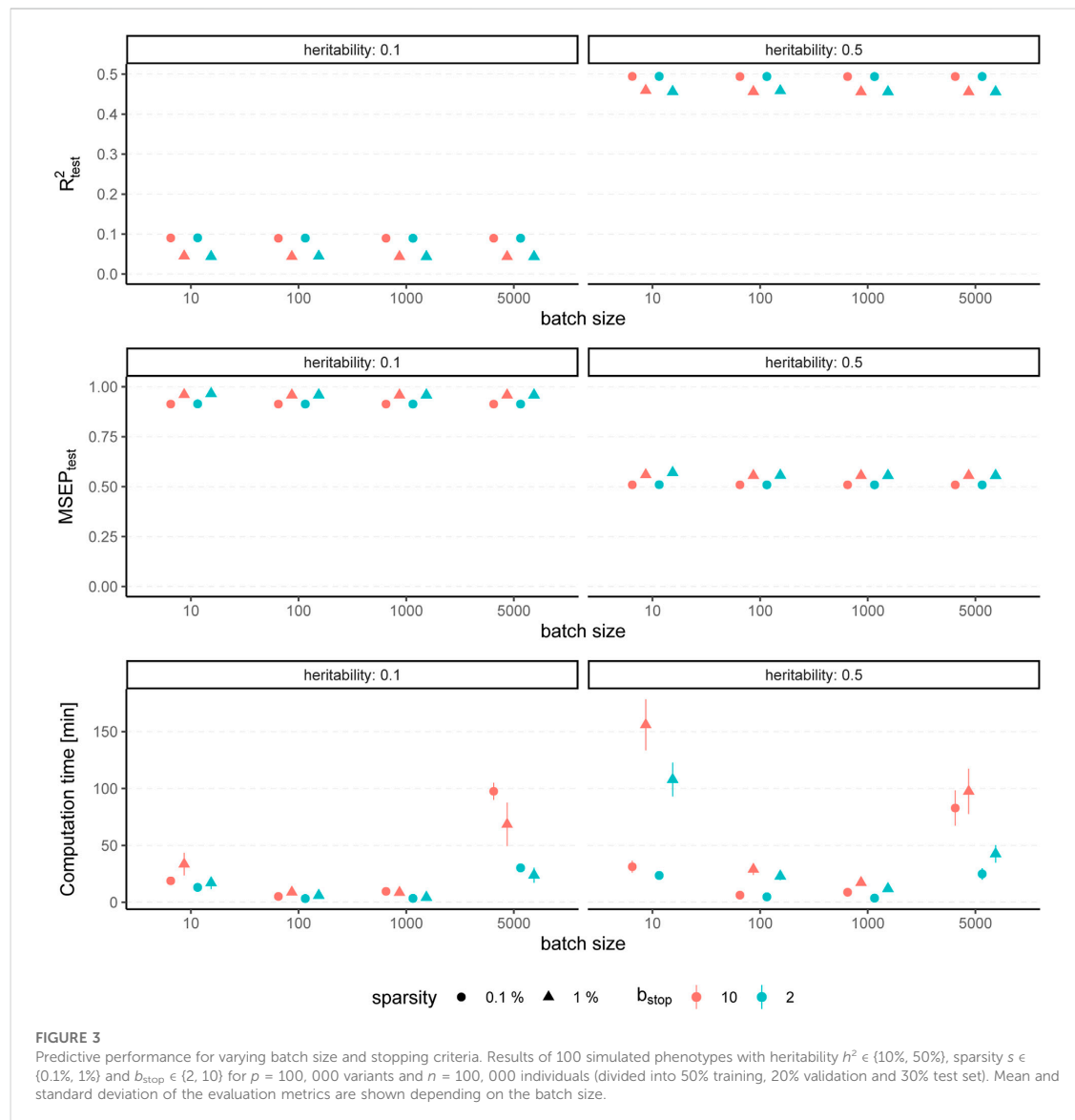
In summary, the incorporation of batches in the boosting algorithm did not affect the predictive performance of the model in our scenarios, while computation time was substantially reduced. However, snpboost does not always lead to the same models as the original L_2 -boosting algorithm, in particular in terms of the included variants and sparsity. The results for further settings with different heritability and sparsity were comparable and can be found in the [Supplementary Material](#).

3.1.2 Choice of hyperparameters for large-scale applications

The proposed snpboost algorithm includes various hyperparameters, namely the batch size p_{batch} , the learning rate ν , the maximum number of boosting iterations per batch m_{batch} , the maximum number of processed batches b_{max} and the stopping lag for the outer early stopping criterion b_{stop} . In this section we discuss default values for the hyperparameters to facilitate the applicability of the algorithm in practice. The majority of these parameters do not need to be tuned but can be specified with reasonable default values,

e.g., based on results from the literature and experience with the original boosting algorithm. For the remaining ones (p_{batch} and b_{stop}) we examine how they influence the computational and predictive performance of snpboost in a simulation study.

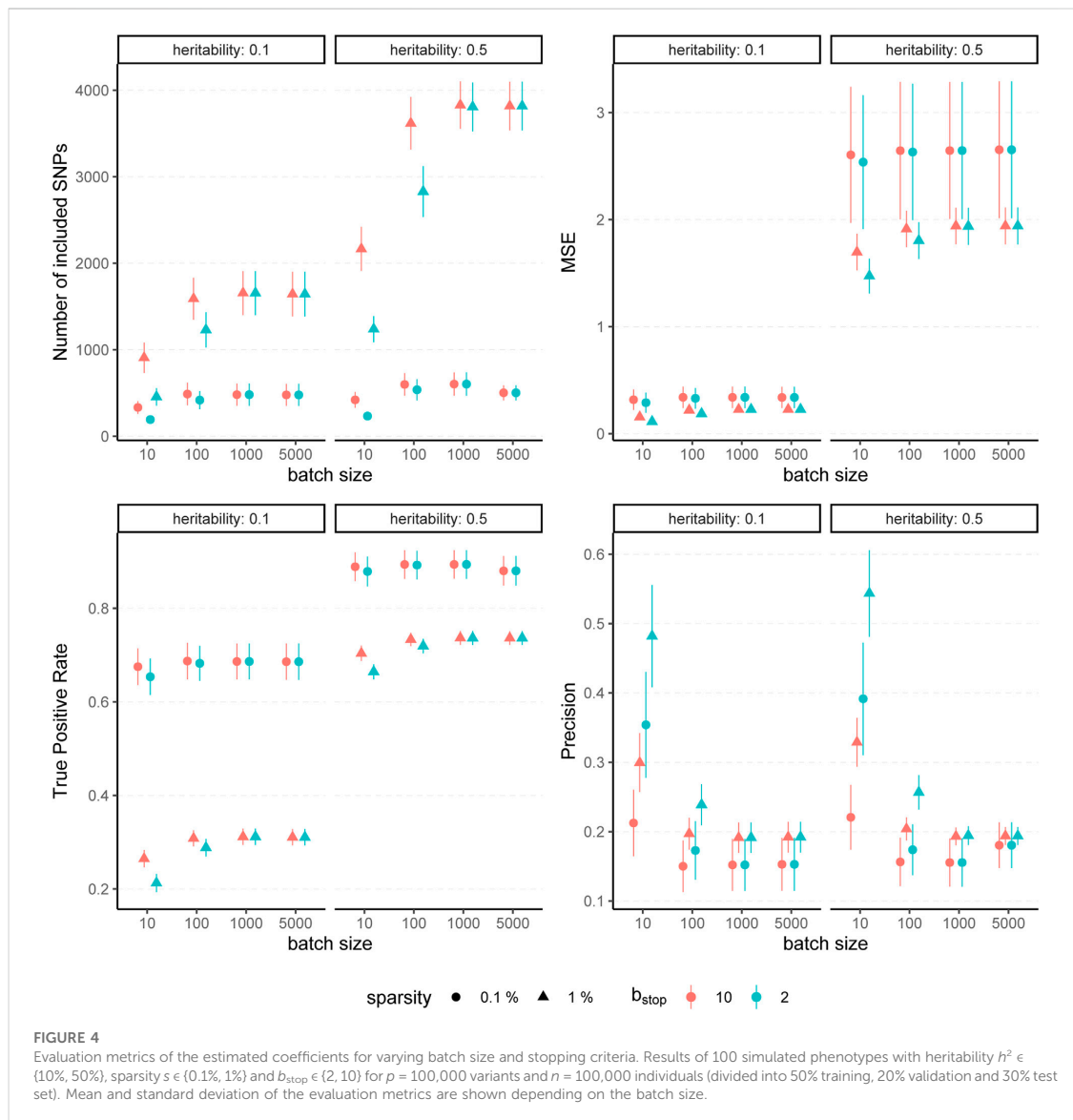
The choice of the learning rate ν can be leaned on widely-used boosting algorithms. A rather small learning rate prevents boosting algorithms from overfitting on single base-learners and is therefore favorable regarding predictive performance. Nevertheless, a smaller learning rate will increase the number of needed boosting iterations to fit the full effect of the base-learners and simultaneously increase the algorithm's computation time. Widely used R packages such as mboost (Bühlmann and Hothorn, 2007; Hothorn et al., 2010) and xgboost (Chen and Guestrin, 2016) use default learning rates of 0.1 and 0.3, respectively. As the effect of the learning rate will be comparable in the proposed adapted boosting algorithm, we decided to specify a fixed default value of $\nu = 0.1$ in all our simulations. For the batch-related hyperparameters we varied the batch size p_{batch} over a range of possible values namely $p_{\text{batch}} \in \{10, 100, 1,000, 5,000\}$ to analyse its effect. For each batch we allow a maximum number of boosting iterations m_{batch} equivalent to the batch size p_{batch} . Since we specified the learning rate with a rather small fixed value and due to the correlation-based early stopping criterion, this choice should prevent the algorithm from overfitting on one batch. If one or more variants inside the batch are still among the most influential ones out of all



variants they will also be included in the next batch. For the outer stopping criterion we specified a large maximum number of batches $b_{\text{max}} = 20,000$ to ensure that the algorithm terminates even in case the MSEP on the validation set has not decreased for b_{stop} batches. Since we do not want the algorithm to stop too early and simultaneously minimize the computation time, in our simulations we consider the choices $b_{\text{stop}} = 2$ and $b_{\text{stop}} = 10$. We then fitted PRS models using snpboost with the previously described hyperparameters. For the computations we used 10 CPUs with 2 GB RAM each.

The results for simulated phenotypes with 10% and 50% heritability are shown in Figure 3 and Figure 4. Results for further degrees of heritability can be found in the supplement. Independently

of the heritability and the sparsity of the simulated data, the predictive performance was not affected in our settings by varying batch sizes in terms of R^2 and MSEP. However, the computation time differed crucially, resulting in considerably higher values for rather small ($p_{\text{batch}} = 10$) or rather large ($p_{\text{batch}} = 5,000$) batches. Furthermore, larger batches led to a higher number of included variants in the final model. This effect was stronger for phenotypes which have a less sparse genetic architecture and associated with a later stopping of the algorithm, i.e., more boosting steps were required to derive the final model. A higher number of variants in the final model was associated with a slightly higher MSE of the coefficients as well as higher true positive rates on the one hand but also smaller precision on the other



hand. As expected, a higher m_{stop} increased the computation time of the fitting process for all batch sizes. In contrast, there was no considerable effect on the predictive performance. However, $b_{\text{stop}} = 2$ and $b_{\text{stop}} = 10$ had an impact on the coefficient estimates as can be seen in Figure 4, e.g., by a tendency to include more variants in the model when choosing $b_{\text{stop}} = 10$. This tendency was only apparent for batch sizes $p_{\text{batch}} < 1,000$, suggesting that for larger batches the choice of b_{stop} is only of minor importance for both, prediction performance and coefficient estimates. The results clearly indicate that a more favorable signal-to-noise ratio (i.e., a higher heritability) and less influential variants (i.e., a higher sparsity) are in general beneficial for the performance of our approach. For

phenotypes with a sparser genetic architecture, the considered evaluation metrics tended to show less variability.

In summary, the choice of the hyperparameters had no major influence on the predictive performance measures R^2 and MSE but on the computation time, which was lowest for medium size batches ($100 \leq p_{\text{batch}} \leq 1,000$). The accuracy of the coefficient estimates measured via MSE, TP and TN rate varied with the batch size, as larger batches tended to lead to more (true positive) variants included in the final model, but also to a slightly higher MSE and a smaller TN rate. While the differences in MSE, TP, and TN rate were only small, smaller batches yielded sparser models in particular for phenotypes with a high heritability.

To conclude, batch sizes of $100 \leq p_{\text{batch}} \leq 1,000$ seem to be the most favorable regarding the computation time and the other evaluation metrics. We propose a batch size of $p_{\text{batch}} = 1,000$ as the default value because the results suggest less dependency on the b_{stop} parameter than for a batch size of 100 variants. Accordingly, we recommend a default value of $b_{\text{stop}} = 2$ to keep the computation time as low as possible. In practice, genotype data most often contain more than 100,000 variants, which further supports the choice of $p_{\text{batch}} = 1,000$ with regard to the computation time. Although our simulation study suggests that those default values should provide reasonable results in most cases, it is recommendable to take the genetic architecture of the examined phenotype as well as the main aim of the analysis into account. Phenotypes with a high expected heritability might be better fitted by using smaller batches, while for phenotypes with many causal variants larger batches might be favorable to increase the TP rate. If one is interested in extremely sparse models identifying only the most-informative variants one could also try to use smaller batches to avoid an overestimation of the number of causal variants.

3.2 Application to the UK Biobank

We applied our proposed method on data from the UK Biobank resource under Application Number 81202. Besides the validation of the results from the previous section, we compared our boosting models fitted *via* the proposed snpboost approach to the ones derived by fitting the lasso *via* the BASIL algorithm implemented in the snpnet package, which have already been shown to outperform commonly-used PRS models for various traits (Qian et al., 2020). Furthermore, we compared our results to PRSs (Ge et al., 2019), LDpred2 (Privé et al., 2021) and SBayesR (Lloyd-Jones et al., 2019), which are based on summary statistics, as well as to multivariable methods *via* LDAK (Zhang et al., 2021) based on Bolt-LMM (Loh et al., 2015), Ridge Regression (Henderson, 1950) and BayesR (Moser et al., 2015). The UK Biobank (UKBB, Bycroft et al., 2018) is a large-scale prospective cohort study including more than half a million participants from the United Kingdom aged between 40 and 69 years when recruited. The database comprises genome-wide genotype data of each individual as well as various in-depth phenotypic information such as biological measurements as well as blood and urine biomarkers. The data have been collected since 2006 and are continually updated with follow-up data.

Our aim is to estimate PRS for various phenotypes, covering several heritability and sparsity levels. The heritability of a trait is an upper bound for the predictive performance based on genotype information. Thus, we used the analyses of Tanigawa et al. (2022) as a proxy and specifically considered five appropriate continuous phenotypes: standing height in cm (UKBB field 50), LDL-cholesterol in mmol/l (UKBB field 30780), blood glucose level in mmol/L (UKBB field 30740), lipoprotein A in nmol/L (UKBB field 30790) and BMI in kg/m² (UKBB field 21001).

Height and BMI are quantitative traits with a relatively high heritability and a rather polygenic structure. Twin-studies estimated a heritability of approximately 69% for height and 42% for BMI after adjusting for covariates (Hemani et al., 2013). For a long time, genetic models could not explain this estimated heritability, a phenomenon known as “missing heritability” (Maher, 2008; Gibson, 2010). More recent studies have indicated that this may be primarily due to many

influential common variants with small effect sizes (Yang et al., 2010; Wood et al., 2014; Yang et al., 2015) underlining the high polygenicity of those traits. In contrast to this, the distribution of the biomarker lipoprotein A, which is a strong risk factor for coronary heart disease, is mainly explained by variants in the LPA gene on chromosome 6 (Kronenberg and Utermann, 2013). Thus, we expect a sparse PRS with a relatively high prediction accuracy for this trait. For LDL-cholesterol it is known that it is associated with several genes such as LDLR and PCSK9 (Sanna et al., 2011; Sabatine, 2019). Therefore, we expect signal in several genomic regions. Recent studies compared different approaches including the lasso to derive PRS, and found that multivariable methods can reach a predictive performance of up to 20% (Maj et al., 2022; Tanigawa et al., 2022). As in previous works (Sinnott-Armstrong et al., 2021), we adjusted the measured LDL-cholesterol value by a factor of 0.684 for individuals taking statins lowering the blood lipid. For blood glucose we are not aware of a genetic impact and also Tanigawa et al. (2022) found the genetic background only explaining a small fraction (less than 2%) of the biomarker's variance.

Out of the over 500,000 individuals from UK Biobank we filtered for unrelated (based on UKBB resource 668) individuals with self-reported white British ancestry (UKBB field 21000) and available data for all chosen phenotypes, resulting in $n = 262,171$ observations. Additionally, the covariates age and sex as well as the first ten principal components of the genotype matrix are available. We randomly divided the data set into training ($n_{\text{train}} = 157,204$), validation ($n_{\text{val}} = 52,416$) and test set ($n_{\text{test}} = 52,551$). We used genome-wide genotype data and filtered for variants with a genotyped rate of at least 90% and a minor allele frequency of at least 0.1%, resulting in $p = 562,723$ genetic variants. Missing genotypes are imputed by the corresponding mean of the complete observations.

For both the boosting and lasso approaches, we first estimated a PRS using only the genotyped variants as predictors. We used the training set to fit the model and the validation set to simultaneously monitor the predictive performance for choosing the main tuning parameters of the algorithms (i.e., the number of iterations for boosting and the penalty parameter for the lasso). To fit the lasso we used the R package snpnet (Qian et al., 2020) with the provided default hyperparameters. Following the results of our simulation study, for the snpboost algorithm we chose a batch size of $p_{\text{batch}} = 1,000$ variants, a learning rate of $\nu = 0.1$ and an outer stopping lag of $b_{\text{stop}} = 2$ batches. Using the resulting estimated PRS we fitted two linear models on the combined training and validation set, namely the first one (M_{PRS}) incorporating only the PRS as a single predictor variable:

$$M_{\text{PRS}}: Y = \gamma_0 + \gamma_{\text{PRS}} \widehat{\text{PRS}} \quad (4)$$

and the second one (M_f) including the first ten principal components, sex and age as additional covariates:

$$M_f: Y = \gamma_0 + \gamma_{\text{PRS}} \widehat{\text{PRS}} + \gamma_1 \text{PC}_1 + \dots + \gamma_{10} \text{PC}_{10} + \gamma_{\text{sex}} \text{sex} + \gamma_{\text{age}} \text{age}. \quad (5)$$

To measure the actual benefit in accuracy of including a PRS in the prediction model, we also fitted a model including only covariates (M_c):

$$M_c: Y = \gamma_0 + \gamma_1 \text{PC}_1 + \dots + \gamma_{10} \text{PC}_{10} + \gamma_{\text{sex}} \text{sex} + \gamma_{\text{age}} \text{age}. \quad (6)$$

Finally, we also included the covariates in the fitting process to derive the PRS, corresponding to the model $M_{\text{PRS},c}$:



TABLE 2 Comparison of computational efficiency of snpnet and snpboost on eight phenotypes from the UKBB. Computational times of the algorithms snpnet and snpboost for multivariable polygenic models with and without inclusion of the covariates for the prediction of eight phenotypes from the UKBB. M_{PRS} corresponds to the application of the algorithms without including covariates and $M_{PRS,c}$ to the inclusion of the covariates sex, age and the first ten principal components. The experiments were run on 16 CPUs with 2 GB RAM each.

Phenotype	Model	Computation time in minutes	
		snpnet	snpboost
Height	M_{PRS}	132.44	116.65
Height	$M_{PRS,c}$	97.49	299.98
BMI	M_{PRS}	54.36	94.34
BMI	$M_{PRS,c}$	54.81	156.49
LDL	M_{PRS}	37.92	50.64
LDL	$M_{PRS,c}$	45.61	64.27
glucose	M_{PRS}	14.86	11.38
glucose	$M_{PRS,c}$	14.71	16.33
lipoprotein A	M_{PRS}	28.99	25.08
lipoprotein A	$M_{PRS,c}$	33.67	30.14
asthma	M_{PRS}	3.97	5.31
asthma	$M_{PRS,c}$	4.21	6.63
coeliac	M_{PRS}	3.11	5.46
coeliac	$M_{PRS,c}$	5.00	6.69
HBP	M_{PRS}	46.63	90.27
HBP	$M_{PRS,c}$	30.07	181.23

$$M_{PRS,c}: Y = \beta_0 + \beta_{PC1}PC_1 + \dots + \beta_{PC10}PC_{10} + \beta_{sex}sex + \beta_{age}age + \sum_{j=1}^p \beta_j X_j. \quad (7)$$

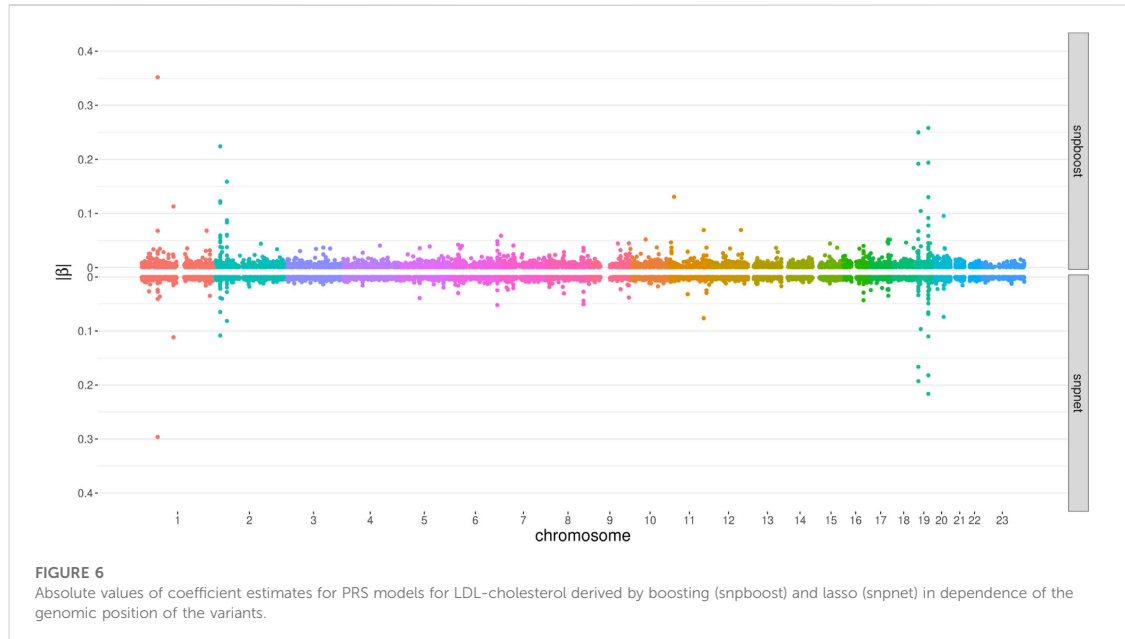
All models were evaluated on the independent test set and compared with respect to their predictive performance, computational efficiency and sparsity. To measure the predictive performance we used the R^2 value on the test set given by the squared correlation between the observed and predicted phenotypes as well as the root mean squared error of prediction (RMSEP). The computational efficiency was measured as the computation time in minutes of the respective algorithm and the sparsity is given by the number of included variants in the final PRS. All computations were conducted on a computer cluster with 16 CPUs and 2 GB RAM each. The derivation of the PRS by the use of further methods (namely PRScs, LDpred2, SBayesR, Bolt-LMM, Ridge Regression and BayesR) was based on the same training and validation data and is described in the [Supplementary Material](#). All models were tested on the same independent test set.

The results of snpboost as well as of snpnet for all phenotypes are given in [Figure 5](#) and [Table 2](#). The resulting RMSEP is shown in [Supplementary Figure S7](#). Overall, snpnet and snpboost yield comparable results regarding the predictive performance, without one approach being consistently superior to the other. Both the

resulting R^2 and RMSEP are very close. Furthermore, the shown R^2 values are in line with previously reported R^2 resulting from snpnet, which has been shown to be highly competitive to various other (univariate) PRS methods ([Qian et al., 2020](#); [Li et al., 2022](#); [Tanigawa et al., 2022](#)). The PRS estimated via snpnet and snpboost both clearly increase the predictive performance compared to the covariates-only model M_c for all shown phenotypes. With respect to sparsity, our boosting approach tends to select less variants (on average 26% less variants compared to the lasso). The computation time of both approaches is highly dependent on the genetic architecture, i.e., the heritability and sparsity of the phenotype. In particular, a higher and more polygenic signal tends to lead to longer computation times. In case of fitting the PRS based solely on the genotype data and including the covariates in a subsequent linear model, snpboost tends to be faster than snpnet; however, the computation times for snpboost increase substantially when including covariates in the fitting process for LDL-cholesterol and height. This is partly due to more coefficients being fitted and updated in each boosting step and partly due to larger PRS models resulting from more boosting steps. Nevertheless, the models are still fitted in reasonable time using our batch-based approach. As described in [Hepp et al. \(2016\)](#), boosting is generally expected to be slower than the lasso, which can only be observed for less sparse models in the examined phenotypes. In general, the model $M_{PRS,c}$ outperforms the model M_f regarding the predictive performance, implying that including the covariates already in the fitting of the PRS is favorable regarding the detection of the genetic signal. However, the effect is only substantial for phenotypes with a high association to covariates (i.e., height). Furthermore, the model $M_{PRS,c}$ tends to select more variables than estimating the PRS based only on the genotypes (M_{PRS}) and the computation time is considerably increased when using the snpboost approach. Therefore, it might be advisable to only consider the covariates in the fitting process if there is a large association already in the covariates-only model.

[Figure 6](#) displays the absolute values of the resulting estimated non-zero coefficients for LDL-cholesterol for the boosting and lasso approaches. Both tend to detect variants with higher effect sizes in the same genetic regions, e.g., at chromosome 2 and chromosome 19. In total, there are 3,030 genetic variants that are present in both PRS, out of 7,163 variants selected by snpboost and 8,924 variants selected by snpnet. While snpboost results in less variants, the included variants have larger effect sizes and less variants with very small effect sizes close to zero are included in the model. [Supplementary Figure S8](#) displays the coefficients again with shared variants marked in black. All SNPs with comparably high effect sizes in the snpnet PRS are included in both models but the snpboost PRS incorporates further SNPs with stronger effects. The results are similar for the other phenotypes and included in the [Supplementary Material](#). In conclusion, the snpboost PRS tends to include less variants in total, but more variants with comparably high effect sizes corresponding to less shrinkage for the variants included in the model compared to the lasso.

The [Supplementary Material](#) comprises results for comparisons to further commonly used methods to derive PRS ([Supplementary Tables S1, S2](#)). Our proposed algorithm yielded consistently higher prediction performance compared to the summary statistics based PRScs and LDpred2 methods; furthermore, it yielded competitive results compared to summary statistics based SBayesR and four different multivariable approaches, while tending to select the sparsest models.



4 Extension to binary data

While traits like blood biomarkers or physical measurements are often quantitative, it is, for example, also of interest to predict the probability of the occurrence of a disease for a particular patient. In this case we deal with binary data $y_i \in \{0, 1\}$ and proceed as in a logistic regression by modelling the logit of the expected value as a linear model

$$\text{logit}(P(y_i = 1|\mathbf{X})) = \ln\left(\frac{P(y_i = 1|\mathbf{X})}{1 - P(y_i = 1|\mathbf{X})}\right) = \beta_0 + \sum_{j=1}^p \beta_j x_{i,j}, \quad i = 1, \dots, n. \quad (8)$$

The estimated probability $\hat{p}_i(\mathbf{X}) = P(y_i = 1|\mathbf{X})$ is then given by

$$\hat{p}_i(\mathbf{X}) = P(y_i = 1|\mathbf{X}) = \frac{\exp(\hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{i,j})}{1 + \exp(\hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{i,j})}. \quad (9)$$

To fit binary outcomes *via* boosting we replace the L_2 loss by the log loss

$$f_{\ln}(\mathbf{y}, \hat{\mathbf{p}}) = -\frac{1}{n} \sum_{i=1}^n y_i \ln(\hat{p}_i) + (1 - y_i) \ln(1 - \hat{p}_i). \quad (10)$$

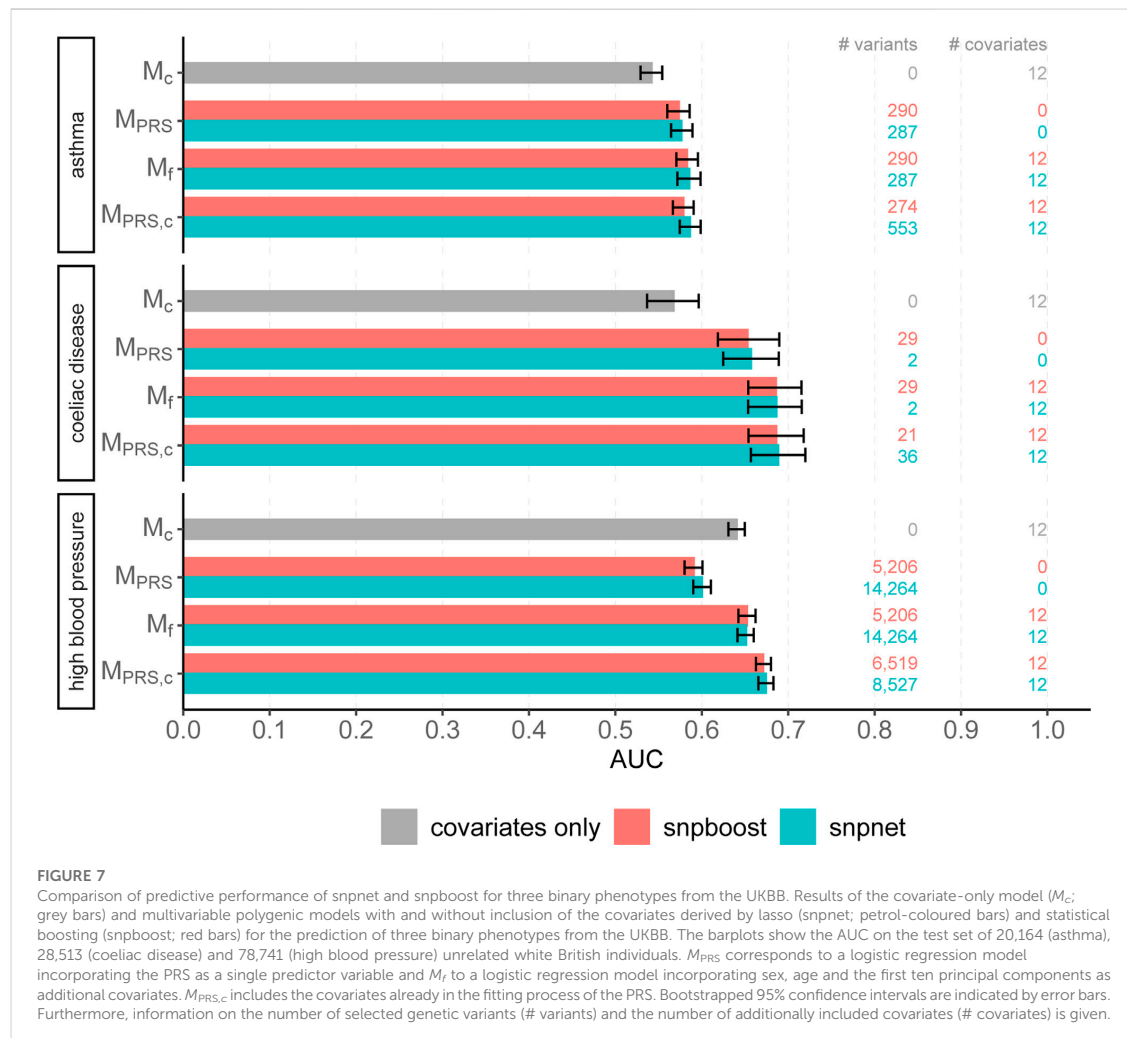
Note that following this new loss function, the gradient is no longer represented by the residuals. The base-learners are hence fitted now to the first derivative of the loss function in Eq. 10. Consequently, batches are built out of the p_{batch} variants with the highest absolute correlation to the first derivative of the loss in Eq. 10 instead of the residual. However, the other components of the algorithm including the base learners remain unchanged. We also keep the hyperparameters derived in Section 3.1.2 fixed. We applied the extended algorithm on data of the UKBB for three binary

phenotypes: the occurrence of asthma (UKBB field 22127), coeliac disease (UKBB field 21068) and high blood pressure (UKBB field 6150). All three traits are associated to many environmental factors but also have a genetic component (Arora and Newton-Cheh, 2010; Trynka et al., 2010; Yang et al., 2017; El-Husseini et al., 2020). Tanigawa et al. (2022) estimated high blood pressure to be a rather polygenic trait while the genetic component of asthma and coeliac disease is determined by fewer common variants.

We incorporated unrelated individuals of white British ancestry in our analysis and divided the samples randomly into training, validation and test sets. In total we used 8,397 cases ($n_{\text{train}} = 4,266$, $n_{\text{val}} = 1,709$ and $n_{\text{test}} = 2,522$) and 58,428 controls ($n_{\text{train}} = 29,079$, $n_{\text{val}} = 11,707$, $n_{\text{test}} = 17,642$) for asthma, 1,793 cases ($n_{\text{train}} = 882$, $n_{\text{val}} = 361$ and $n_{\text{test}} = 550$) and 92,646 controls ($n_{\text{train}} = 46,234$, $n_{\text{val}} = 18,449$, $n_{\text{test}} = 27,963$) for coeliac disease and 71,235 cases ($n_{\text{train}} = 35,720$, $n_{\text{val}} = 14,210$ and $n_{\text{test}} = 21,305$) and 190,422 controls ($n_{\text{train}} = 94,740$, $n_{\text{val}} = 38,246$, $n_{\text{test}} = 57,436$) for high blood pressure.

The applicability to binary traits was also one of the first extension of snpnet and Qian et al. (2020) showed impressive results for a number of binary traits. Due to that, we again also apply snpnet to the same data to evaluate the quality of our results.

We evaluated the accuracy of the resulting predictions on the test set using both the log loss as well as the AUC. Results are shown in Figure 7 and in Supplementary Figure S17. The overall predictive performance is comparable for all three phenotypes. For high blood pressure with a polygenic genetic component snpboost yields a sparse model with a high predictive performance. For sparse binary phenotypes as asthma and coeliac disease, snpboost and snpnet yield similar sparse models. The result for coeliac disease, which appears to be rather oligogenic than polygenic, for snpnet is outstanding, but in line with the results of Tanigawa et al. (2022). Nevertheless, also snpboost also estimates a very sparse PRS with a



high predictive performance. Table 2 illustrates the computation time for binary data of snpnet and snpboost on a computer cluster with 16 CPUs and 2 GB RAM each. Both, snpboost and snpnet yield very limited computation times, with snpnet being faster.

In summary, this illustrates how easily and conveniently the snpboost framework can be extended to different data types by incorporating different loss functions. Even though we simply plugged in the log loss and did not optimize the hyperparameters such as the batch size or the learning rate of our algorithm for binary data, snpboost yields a competitive predictive performance compared to the BASIL algorithm.

5 Discussion

In this work we have proposed a new methodological framework to derive multivariable PRS models *via* applying a statistical boosting

approach directly on genotype data. Currently, PRS are most often built based on summary statistics from GWAS that were estimated by simple and univariate linear regression models (Choi et al., 2020). This methodologically simple approach is mainly justified by the computational hurdle resulting from the ultra-high-dimensionality of the genotype data. For example, in the past it had been unfeasible to fit a lasso model on the complete genotype data due to the high computational complexity. To overcome this, Mak et al. (2017) developed lassosum, an approach to approximate the lasso path by only using summary statistics and LD reference panels. However, recently published works provided methods to enable statistical modelling by penalized multivariable regression approaches on genotype data (Privé et al., 2018; Qian et al., 2020). Qian et al. demonstrated that lasso-based PRS were able to outperform several PRS derived by methods based on univariate summary statistics (Qian et al., 2020). First approaches to apply statistical boosting on genotype data employed a three-step-approach to fit multivariable PRS (Maj

et al., 2022): first, variants are pre-filtered based on their univariate associations with the examined phenotype. Second, statistical modelling and variable selection approaches such as AdaSub (Staerk et al., 2021) and boosting with probing (Thomas et al., 2017) are used to identify the informative variants on blocks of variants in LD. Finally, a multivariable PRS based on the selected variants is constructed *via* component-wise L_2 -boosting (Bühlmann and Hothorn, 2007). While this approach yielded particularly sparse models and could compete with common methods like clumping and thresholding (Euesden et al., 2014), lasso *via* snpnet yielded more accurate results regarding the predictive performance which is usually the main objective of PRS modelling.

In the present article we introduced the boosting algorithm snpboost that works on smaller batches of variants similar to the BASIL algorithm. Our framework enables the application of statistical boosting directly on the complete original genotype data. In a smaller but still high-dimensional simulation setting, we were able to show that the adapted boosting algorithm yields similar performance to the original component-wise L_2 -boosting, indicating that we do not lose predictive performance due to the incorporation of batches. In a further setting with more realistic dimensionality we have derived appropriate default values for the application of snpboost on large-scale data. We were able to show that the specified default values resulted in models with good performance in most cases but also gave advice on how to adapt them based on the genetic architecture of the examined phenotype and the specific research questions.

We applied the newly proposed snpboost algorithm on large-scale genotype data of the UKBB. In particular, we have compared the performance of snpboost to the one achieved by the lasso *via* snpnet, which has been shown to outperform many classical PRS (Qian et al., 2020). Our results indicate that the snpboost algorithm leads to PRS models that are highly competitive to lasso-based PRS models in both predictive performance and computation time. Although it might have been expected that the computation time would be higher for statistical boosting than for the lasso (Hepp et al., 2016), our approach had a tendency to result in sparser models. These sparser models correspond to an earlier stopping of the algorithm which reduces the computation time of boosting. The incorporation of further covariates such as age, sex and principal components in the fitting process of the PRS resulted in increased computation times for some phenotypes, particularly for height. However, in such cases, the boosting algorithm yielded an improved predictive performance with larger numbers of included variants. This illustrates that sparsity is not always favorable in regards of predictive performance. Additionally, we compared the performance of snpboost to further predictive PRS tools, which are either summary statistics based as PRSs, LDpred2 and SBayesR or multivariable approaches *via* the LDAK implementation of BayesR, Ridge Regression and Bolt-LMM (Zhang et al., 2021). While these methods do not apply variable selection, the predictive performance of snpboost was still highly competitive.

Our analyses show that there is a large overlap of the chosen variants by lasso and boosting, in particular regarding the variants with high estimated effect sizes. However, boosting has been found to include less variants in the final model and to induce less shrinkage on the effect estimates compared to the lasso. In clinical practice, a sparser PRS model might be of particular interest if the aim is not only prediction but also the identification of risk loci in the genome. In fact, functional annotations of the selected variants can better elucidate the underlying biological mechanisms influencing the analyzed trait.

Thus, statistical boosting might be one way to yield more biologically interpretable PRS models.

Despite the presented promising results, the proposed method also inherited some limitations from statistical boosting. In contrast to classical regression methods, boosting does not provide closed formulas for standard errors of effect estimates or confidence intervals that could be used for inference. Furthermore, as mentioned before, statistical boosting is in general associated with a slightly higher computational complexity compared to methods such as the lasso (Hepp et al., 2016) and has a known tendency to include too many variables in low-dimensional settings (Staerk and Mayr, 2021; Strömer et al., 2022). Our results suggest that the incorporation of batches substantially reduced the computational time. Additionally, the reduction of the search space in each boosting step might partially prevent the algorithm from selecting too many variables. However, the implementation of the batch-based statistical boosting in snpboost is currently limited to linear base-learners, each corresponding to one genetic variant.

Apart from those technical limitations, using individual-level data raises ethical and logistical questions: While summary statistics are easily shared and do not allow for identification of unique individuals, individual-level data involve the risk of identification. It is therefore crucial, that providers as well as researchers using individual-level data follow ethical standards. Furthermore, the storage and transfer of individual-level data require more capacities which might not be at everyone's disposal in the complete research community. However, the resulting PRS can be published by sharing only the included variants, alleles and coefficients—exactly like summary statistics (Lambert et al., 2021). To make use of available summary statistics and to avoid the limitations associated with individual-level data, it might be of interest to develop an approximation of a component-wise boosting algorithm based on summary statistics and LD panels, analogously as lassosum for the lasso. From a computational perspective, this is not necessary as snpboost only requires limited resources (e.g., our analysis of the UKBB data was run with only 32 GB RAM in total).

By incorporating the log loss we made our framework applicable also to binary traits and demonstrated the convenience of further extensions of the snpboost framework beyond the case of continuous phenotypes. Without re-specifying our hyperparameters we were able to yield similar results as the snpnet framework.

In future research we want to further exploit the modular structure of boosting to model more complex biological phenomena. We will incorporate different loss functions to extend the snpboost framework to be applicable also to count and time-to-event data. To account for the uncertainty in the prediction, one could also construct subject-specific prediction intervals based on quantile regression (Mayr et al., 2012). Besides extending the approach *via* new loss functions, one could also change the base-learners in various ways. For example, base-learners could be adapted to take into account different models of inheritance beside the classical additive component typically used in the polygenic models, such as dominant and recessive hereditary schemes. Further possibilities for future research include the extension of the set of possible base-learners, e.g., to model gene-environment interactions as well as epistatic effects across variants which can play a relevant role in biological phenotypes (Li and Lehner, 2020). To do so, base-learners including interactions between variants and variant-covariate interactions could be incorporated. Apart from that, biological knowledge can also be used a priori. Márquez-Luna et al. (2021) have shown that the incorporation of functional annotations of the genetic variants contribute to a rise in

prediction accuracy. Previous works in the field of penalized regression and boosting have proposed to make use of biologically meaningful groups of genomic variants such as genes or pathways as described by Luan and Li (2008), Wei and Li (2007) as well as Liu et al. (2013). While those previous methods were computationally limited to smaller datasets our framework opens the possibility to include those ideas in the multivariable modelling of PRS. Besides those methodological extensions of our proposed snpboost framework, future research will also focus on the practical application of PRS derived by our framework. An important aspect of PRS research is the transferability of PRS models to different ethnicities, as PRS are often derived on cohorts of European ancestry and a substantial loss of predictive performance is observed when applied on further cohorts with different ethnicities (Landry et al., 2018; Evans et al., 2022). Previous studies have indicated that sparser models may contribute to overcome this issue (Maj et al., 2022) and it is of particular interest to examine the transferability of PRS derived by statistical boosting.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: This research has been conducted using the UK Biobank resource under application number 81202 (<http://www.ukbiobank.ac.uk>). Requests to access these datasets should be directed to UK Biobank, <http://www.ukbiobank.ac.uk>.

Ethics statement

The studies involving human participants were reviewed and approved by the North West Multi-centre Research Ethics Committee (MREC) as a Research Tissue Bank (RTB) approval. (UK Biobank, <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics>). The patients/participants provided their written informed consent to participate in this study.

References

- Arora, P., and Newton-Cheh, C. (2010). Blood pressure and human genetic variation in the general population. *Curr. Opin. Cardiol.* 25, 229–237. doi:10.1097/hco.0b013e3283383e2c
- Beesley, L. J., Salvatore, M., Fritsche, L. G., Pandit, A., Rao, A., Brummett, C., et al. (2020). The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. *Statistics Med.* 39, 773–800. doi:10.1002/sim.8445
- Bühlmann, P., and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Stat. Sci.* 22, 477–505. doi:10.1214/07-STS242
- Bühlmann, P., and Yu, B. (2003). Boosting with the l2 loss. *J. Am. Stat. Assoc.* 98, 324–339. doi:10.1198/016214503000125
- Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Statistics* 1, 169–194. doi:10.1214/07-EJS008
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209. doi:10.1038/s41586-018-0579-z
- Chang, C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* 4, 7. doi:10.1186/s13742-015-0047-8
- Chen, T., and Guestrin, C. (2016). “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (New York, NY, USA: Association for Computing Machinery), KDD ’16, 785. doi:10.1145/2939672.2939785
- Choi, S. W., Mak, T. S.-H., and O'Reilly, P. F. (2020). Tutorial: A guide to performing polygenic risk score analyses. *Nat. Protoc.* 15, 2759–2772. doi:10.1038/s41596-020-0353-1
- El-Husseini, Z. W., Gossens, R., Dekker, F., and Koppelman, G. H. (2020). The genetics of asthma and the promise of genomics-guided drug target discovery. *Lancet Respir. Med.* 8, 1045–1056. doi:10.1016/s2213-2600(20)30363-5
- Euesden, J., Lewis, C. M., and O'Reilly, P. F. (2014). PRSice: Polygenic risk score software. *Bioinformatics* 31, 1466–1468. doi:10.1093/bioinformatics/btu848
- Evans, D. G., van Veen, E. M., Byers, H., Roberts, E., Howell, A., Howell, S. J., et al. (2022). The importance of ethnicity: Are breast cancer polygenic risk scores ready for women who are not of white European origin? *Int. J. Cancer* 150, 73–79. doi:10.1002/ijc.33782
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statistics* 29, 1189–1232. doi:10.1214/aos/1013203451
- Fu, W., and Knight, K. (2000). Asymptotics for lasso-type estimators. *Ann. Statistics* 28, 1356–1378. doi:10.1214/aos/1015957397
- Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019). Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nat. Commun.* 10, 1776. doi:10.1038/s41467-019-09718-5
- Gibson, G. (2010). Hints of hidden heritability in GWAS. *Nat. Genet.* 42, 558–560. doi:10.1038/ng0710-558
- Greenshtein, E., and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* 10, 971–988. doi:10.3150/bj/1106314846

Author contributions

HK, AM, CS, CM, and PK contributed to conception and design of the method. HK wrote the code, performed the experiments and wrote the first draft of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

Funding

The work on this article was supported by the Deutsche Forschungsgemeinschaft (DFG, grant number 428239776, MA7304/1-1).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2022.1076440/full#supplementary-material>

- Hassanin, E., May, P., Aldisi, R., Spier, I., Forstner, A. J., Nöthen, M. M., et al. (2021). Breast and prostate cancer risk: The interplay of polygenic risk, rare pathogenic germline variants, and family history. *Genet. Med.* 24, 576–585. doi:10.1016/j.gim.2021.11.009
- Hemani, G., Yang, J., Vinkhuyzen, A., Powell, J., Willemsen, G., Hottenga, J.-J., et al. (2013). Inference of the genetic architecture underlying bmi and height with the use of 20, 240 sibling pairs. *Am. J. Hum. Genet.* 93, 865–875. doi:10.1016/j.ajhg.2013.10.005
- Henderson, C. R. (1950). Estimation of genetic parameters. *Ann. Math. Stat.* 21, 309.
- Hepp, T., Schmid, M., Gefeller, O., Waldmann, E., and Mayr, A. (2016). Approaches to regularized regression – a comparison between gradient boosting and the lasso. *Methods Inf. Med.* 55, 422–430. doi:10.3414/ME16-01-0033
- Hoerl, A. E., and Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 42, 80–86. doi:10.1080/00401706.2000.10485983
- Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010). Model-based Boosting 2.0. *J. Mach. Learn. Res.* 11, 2109–2113.
- National Human Genome Research Institute (2021). The cost of sequencing a human genome. Available at: <https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost> (Accessed 11 04, 2021).
- Kronenberg, F., and Utermann, G. (2013). Lipoprotein(a): Resurrected by genetics. *J. Intern. Med.* 273, 6–30. doi:10.1111/j.1365-2796.2012.02592.x
- Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., Phillips, K.-A., Mooij, T. M., Roos-Blom, M.-J., et al. (2017). Risks of breast, ovarian, and contralateral breast cancer for *brca1* and *brca2* mutation carriers. *JAMA* 317, 2402–2416. doi:10.1001/jama.2017.7112
- Lambert, S. A., Gil, L., Jupp, S., Ritchie, S. C., Xu, Y., Buniello, A., et al. (2021). The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* 53, 420–425. doi:10.1038/s41588-021-00783-5
- Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L., and Bonham, V. L. (2018). Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Aff.* 37, 780–785. doi:10.1377/hlthaff.2017.1595
- Li, R., Chang, C., Justesen, J. M., Tanigawa, Y., Qian, J., Hastie, T., et al. (2022). Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank. *Biostatistics* 23, 522–540. doi:10.1093/biostatistics/kxab038
- Li, X., and Lehner, B. (2020). Biophysical ambiguities prevent accurate genetic prediction. *Nat. Commun.* 11, 4923. doi:10.1038/s41467-020-18694-0
- Liu, J., Huang, J., Ma, S., and Wang, K. (2013). Incorporating group correlations in genome-wide association studies using smoothed group lasso. *Biostatistics* 14, 205–219. doi:10.1093/biostatistics/kxs034
- Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., et al. (2019). Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nat. Commun.* 10, 5086. doi:10.1038/s41467-019-12653-0
- Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., et al. (2015). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* 47, 284–290. doi:10.1038/ng.3190
- Luan, Y., and Li, H. (2008). Group additive regression models for genomic data analysis. *Biostatistics* 9, 100–113. doi:10.1093/biostatistics/kxm015
- Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature* 456, 18–21. doi:10.1038/456018a
- Maj, C., Staerk, C., Borisov, O., Klinkhammer, H., Yeung, M. W., Krawitz, P., et al. (2022). Statistical learning for sparser fine-mapped polygenic models: The prediction of LDL-cholesterol. *Genet. Epidemiol.* 46, 589–603. doi:10.1002/gepi.22495
- Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* 41, 469–480. doi:10.1002/gepi.22050
- Márquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S. S., Furlotte, N., Auton, A., et al. (2021). Incorporating functional priors improves polygenic prediction accuracy in UK biobank and 23andme data sets. *Nat. Commun.* 12, 6052. doi:10.1038/s41467-021-25171-9
- Mayr, A., and Hofner, B. (2018). Boosting for statistical modelling—a non-technical introduction. *Stat. Model.* 18, 365–384. doi:10.1177/1471082X17748086
- Mayr, A., Hothorn, T., and Fenske, N. (2012). Prediction intervals for future BMI values of individual children – a non-parametric approach by quantile boosting. *BMC Med. Res. Methodol.* 12, 6. doi:10.1186/1471-2288-12-6
- Meinshausen, N. (2007). Relaxed lasso. *Comput. Statistics Data Analysis* 52, 374–393. doi:10.1016/j.csda.2006.12.019
- Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet.* 11, e1004969. doi:10.1371/journal.pgen.1004969
- Privé, F., Arbel, J., and Vilhjálmsson, B. J. (2021). Ldpred2: Better, faster, stronger. *Bioinformatics* 36, 5424–5431. doi:10.1093/bioinformatics/btaa1029
- Privé, F., Aschard, H., Ziyatdinov, A., and Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: Bigstatsr and bigsnpr. *Bioinformatics* 34, 2781–2787. doi:10.1093/bioinformatics/bty185
- Purcell, S., and Chang, C. (2015). Plink 2.0. Available at: www.cog-genomics.org/plink/2.0/.
- Qian, J., Tanigawa, Y., Du, W., Aguirre, M., Chang, C., Tibshirani, R., et al. (2020). A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet.* 16, e1009141. doi:10.1371/journal.pgen.1009141
- R Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Sabatine, M. S. (2019). PCSK9 inhibitors: Clinical evidence and implementation. *Nat. Rev. Cardiol.* 16, 155–165. doi:10.1038/s41569-018-0107-8
- Sanna, S., Li, B., Mulas, A., Sidore, C., Kang, H. M., Jackson, A. U., et al. (2011). Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet.* 7, e1002198. doi:10.1371/journal.pgen.1002198
- Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., et al. (2021). Genetics of 35 blood and urine biomarkers in the UK biobank. *Nat. Genet.* 53, 185–194. doi:10.1038/s41588-020-00757-z
- Staerk, C., Kateri, M., and Ntzoufras, I. (2021). High-dimensional variable selection via low-dimensional adaptive learning. *Electron. J. Statistics* 15, 1797. doi:10.1214/21-ejs1797
- Staerk, C., and Mayr, A. (2021). Randomized boosting with multivariable base-learners for high-dimensional variable selection and prediction. *BMC Bioinforma.* 22, 441. doi:10.1186/s12859-021-04340-z
- Strömer, A., Staerk, C., Klein, N., Weinhold, L., Titze, S., and Mayr, A. (2022). Deselection of base-learners for statistical boosting—With an application to distributional regression. *Stat. Methods Med. Res.* 31, 207–224. doi:10.1177/09622802211051088
- Tanigawa, Y., Qian, J., Venkataraman, G., Justesen, J. M., Li, R., Tibshirani, R., et al. (2022). Significant sparse polygenic risk scores across 813 traits in UK biobank. *PLoS Genet.* 18, e1010105–e1010121. doi:10.1371/journal.pgen.1010105
- Thomas, J., Hepp, T., Mayr, A., and Bischl, B. (2017). Probing for sparse and fast variable selection with model-based boosting. *Comput. Math. Methods Med.* 2017, 1421409–1421418. doi:10.1155/2017/1421409
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x
- Trynka, G., Wijmenga, C., and van Heel, D. A. (2010). A genetic perspective on coeliac disease. *Trends Mol. Med.* 16, 537–550. doi:10.1016/j.molmed.2010.09.003
- van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statistics* 36, 614–645. doi:10.1214/009053607000000929
- Vilhjálmsson, B., Yang, J., Finucane, H., Gusev, A., Lindström, S., Ripke, S., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97, 576–592. doi:10.1016/j.ajhg.2015.09.001
- Wei, Z., and Li, H. (2007). Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics* 8, 265–284. doi:10.1093/biostatistics/kxl007
- Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 46, 1173–1186. doi:10.1038/ng.3097
- Yang, I. V., Lozupone, C. A., and Schwartz, D. A. (2017). The environment, epigenome, and asthma. *J. Allergy Clin. Immunol.* 140, 14–23. doi:10.1016/j.jaci.2017.05.011
- Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Lee, S. H., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* 47, 1114–1120. doi:10.1038/ng.3390
- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common snps explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569. doi:10.1038/ng.608
- Zhang, Q., Privé, F., Vilhjálmsson, B., and Speed, D. (2021). Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.* 12, 4192. doi:10.1038/s41467-021-24485-y
- Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67, 301–320. doi:10.1111/j.1467-9868.2005.00503.x

3.3 Publication 3: Genetic prediction modeling in large cohort studies via boosting targeted loss functions

Klinkhammer H, Staerk C, Maj C, Krawitz PM, Mayr A. Genetic prediction modeling in large cohort studies via boosting targeted loss functions. In: Statistics in Medicine 2024; 43: 5412-5430

<https://doi.org/10.1002/sim.10249>

Supplementary data can be found at:

<https://doi.org/10.1002/sim.10249>

Implementations are available on GitHub:

<https://github.com/hklinkhammer/snpboost>

RESEARCH ARTICLE **OPEN ACCESS**

Genetic Prediction Modeling in Large Cohort Studies via Boosting Targeted Loss Functions

 Hannah Klinkhammer^{1,2}  | Christian Staerk^{1,3,4} | Carlo Maj^{2,5} | Peter M. Krawitz² | Andreas Mayr¹ 

¹Institute of Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Bonn, Germany | ²Institute of Genomic Statistics and Bioinformatics, Medical Faculty, University of Bonn, Bonn, Germany | ³IUF – Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany | ⁴Department of Statistics, TU Dortmund University, Dortmund, Germany | ⁵Center for Human Genetics, Philipps-University Marburg, Marburg, Germany

Correspondence: Hannah Klinkhammer (klinkhammer@imbie.uni-bonn.de)

Received: 22 March 2024 | **Revised:** 11 September 2024 | **Accepted:** 1 October 2024

Funding: The authors received no specific funding for this work.

Keywords: boosting | polygenic scores | prediction | quantile regression | time-to-event data

ABSTRACT

Polygenic risk scores (PRS) aim to predict a trait from genetic information, relying on common genetic variants with low to medium effect sizes. As genotype data are high-dimensional in nature, it is crucial to develop methods that can be applied to large-scale data (large n and large p). Many PRS tools aggregate univariate summary statistics from genome-wide association studies into a single score. Recent advancements allow simultaneous modeling of variant effects from individual-level genotype data. In this context, we introduced snpboost, an algorithm that applies statistical boosting on individual-level genotype data to estimate PRS via multivariable regression models. By processing variants iteratively in batches, snpboost can deal with large-scale cohort data. Having solved the technical obstacles due to data dimensionality, the methodological scope can now be broadened—focusing on key objectives for the clinical application of PRS. Similar to most methods in this context, snpboost has, so far, been restricted to quantitative and binary traits. Now, we incorporate more advanced alternatives—targeted to the particular aim and outcome. Adapting the loss function extends the snpboost framework to further data situations such as time-to-event and count data. Furthermore, alternative loss functions for continuous outcomes allow us to focus not only on the mean of the conditional distribution but also on other aspects that may be more helpful in the risk stratification of individual patients and can quantify prediction uncertainty, for example, median or quantile regression. This work enhances PRS fitting across multiple model classes previously unfeasible for this data type.

1 | Introduction

Polygenic risk scores (PRS) aim to capture the individual's predisposition to various traits (diseases or clinical outcomes) influenced by genetics. Often, those traits are biologically complex diseases or phenotypes and present specific problems such as polygenicity and heterogeneity that require tailored statistical methods [1, 2]. In particular, PRS are typically estimated based on large-scale genotype data with both large

sample sizes n (hundreds of thousands of participants) and large numbers of variables p (millions of single nucleotide polymorphisms, SNPs), comprising common genetic variants with low to medium effect sizes. In contrast, single variants with a very strong signal are rare. Most traits have a limited genetic component, that is, a low signal-to-noise ratio [3]. Another statistical challenge is that closely located genetic variants tend to show high correlations (linkage disequilibrium, LD).

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *Statistics in Medicine* published by John Wiley & Sons Ltd.

Due to the high-dimensionality of the data, many common statistical methods to derive PRS such as LDpred [4], PRSCs [5] and PRSice [6] are based on summary statistics corresponding to simple univariate association tests for each variant in genome-wide association studies (GWAS). Additional methods such as Bayesian approaches can be applied to account for the correlation among the SNPs and to combine these univariate effects. From a statistical perspective, to build polygenic prediction models, it is desirable to use multiple regression modeling to jointly estimate multivariable-adjusted effects for the variants, accounting directly for LD. To do so, we introduced the statistical boosting framework snpboost [7], which works on batches of variants to overcome the computational hurdle of large p and yields sparse linear prediction models while automatically selecting the most informative variants. Previous results suggest that such multivariable methods often yield higher prediction accuracy than methods based on summary statistics [7, 8].

Until now, the majority of PRS methods, including snpboost, have been limited to continuous and binary outcomes, utilizing classical linear and logistic regression, respectively. However, in biostatistical practice, outcomes of interest might not always be continuous or binary but can also represent time-to-event data, such as the age of onset of a disease (for an example in the context of PRS, see Li et al. [9]), or count data, such as the number of acute episodes of a chronic illness. In particular, time-to-event data usually contain censored observations that demand adequate modeling, while classical linear and logistic regression models do not account for this. Furthermore, outcome data might be skewed or compromised by outliers and measurement errors. This raises important issues, especially as researchers are interested in building individual prediction intervals based on the genetic predisposition. There is hence the need to go one step further in developing PRS as flexible statistical prediction models for different types of outcomes, rather than simple biomarkers constructed as sums of weighted effect alleles.

In this work, we propose using statistical boosting to fit advanced regression models on large-scale genotype data, enabling flexible modeling of different types of outcomes. In this context, our approach allows, for the first time, the application of multiple model classes for this type of large-scale genetic data. The modular nature of statistical boosting provides the opportunity to extend the snpboost framework based on targeted loss functions. In particular, we incorporate loss functions for continuous, binary, time-to-event, and count data. This allows us to effectively use the most suitable statistical model classes in the context of polygenic risk modeling for the most common types of outcome data. Furthermore, we implement quantile regression in our framework; this includes median regression as a special case, serving as a robust alternative to classical mean regression in the presence of skewed or corrupted data. As an additional feature, quantile regression facilitates the construction of individual prediction intervals to report the uncertainty of point predictions.

In the following, we will describe the snpboost framework and introduce the new loss functions which are provided in an R implementation (<https://github.com/hklinkhammer/snpboost>). We offer various alternative modeling choices covering accelerated failure time (AFT) models and weighted L_2 boosting for time-to-event data, as well as Poisson and negative binomial

regression for count data. While these model classes and objective functions are well-known in biostatistics, they have generally not been feasible for handling these extreme data situations with large n and p . We investigate the performance of the proposed loss functions on simulated phenotypes based on genotype data and compare it to a simplified modeling approach by treating the outcomes as continuous or binary. Then, using high-dimensional data from the UK Biobank [10], we apply the new modeling opportunities by building prediction models for the age of onset of asthma and gout, as well as for the number of sunburn occasions during childhood. Finally, we investigate individual prediction intervals for the body mass index (BMI) and for lipoprotein A.

2 | Methods

For each individual, $i = 1, \dots, n$ we observe the phenotype y_i and measure p genetic variants $x_{i,j}$, for $j = 1, \dots, p$. Let $\mathbf{y} = (y_1, \dots, y_n)^T$ denote the observed phenotype vector and $\mathbf{X} = (x_{i,j}) \in [0, 2]^{n \times p}$ the genotype matrix with the j th column $\mathbf{x}_j$ corresponding to the j th variant. Our aim is to minimize the expected risk $E(\rho(Y\eta)|\mathbf{X})$ defined via a pre-specified loss function ρ with respect to an additive predictor η .

2.1 | Statistical Boosting for Huge Genotype Data

The concept of boosting emerged from machine learning [11, 12] but was later adapted to fit statistical models for potentially high-dimensional data while incorporating automated data-driven variable selection [13, 14].

Statistical boosting estimates the additive predictor $\hat{\eta}$ in an iterative fashion using so-called base-learners. In the genetic context, we currently consider simple linear models as base-learners for each variant given by $h_j(X_j) = \beta_0 + \beta_j X_j$, $j = 1, \dots, p$, leading to a linear predictor $\eta = \beta_0 + \mathbf{X}\boldsymbol{\beta}$. After initializing $\hat{\eta}^{(0)}$ and setting $m = 0$, the following steps are performed until a maximum of m_{stop} boosting iterations is reached: [14]

1. Set $m := m + 1$ and calculate the negative gradient of the loss function

$$\mathbf{u}^{(m)} = -\frac{\partial}{\partial \eta} \rho(\mathbf{y}, \eta) \Big|_{\eta = \hat{\eta}^{(m-1)}}$$

2. Fit each base-learner h_j separately and choose the base-learner $h_{j^*}^{(m)}$ that approximates $\mathbf{u}^{(m)}$ best.
3. Update the linear predictor with step length $\nu > 0$:

$$\hat{\eta}^{(m)} = \hat{\eta}^{(m-1)} + \nu \hat{h}_{j^*}^{(m)}(\mathbf{x}_{j^*})$$

4. Stop if $m = m_{\text{stop}}$.

To enable the application of statistical boosting directly on huge and high-dimensional genotype data, the snpboost algorithm incorporates an additional screening and batch-building step [7]. To build a batch, the $p_{\text{batch}} \ll p$ variants showing the highest absolute Pearson correlations with the current negative gradient

of the loss are selected. On this subset of variants, a maximum of m_{batch} boosting iterations is performed with the corresponding base-learners. This procedure of batch building and boosting on the batch is sequentially repeated for a maximum number of b_{max} (possibly overlapping) batches. To prevent over-fitting, two stopping criteria were incorporated inside and outside of these batches. To avoid over-fitting within a batch, the largest absolute correlation between a variant outside the batch and the current negative gradient vector is saved when entering a new batch. Boosting on the current batch is then stopped, if all correlations between variants inside the batch and the current gradient are smaller than the saved correlation. Furthermore, as an “outer” stopping criterion for building new batches, simultaneously to the fitting process, the predictive risk (evaluated loss) on an independent validation set is monitored and the snpboost algorithm stops if the risk on the validation set has not decreased for two consecutive batches.

2.2 | Time-to-Event Data

Particularly in biostatistical applications, the interest is often not only on an individual’s risk for a disease but also on the time until its occurrence or its course after initial treatment. In oncology, for example, studies typically focus also on the age of onset [15] and on recurrence-free or metastasis-free survival times of patients after chemotherapy [16]. In genetic epidemiology, PRS derived from GWAS for disease risk (occurrence as binary outcome) are often directly plugged into multiple regression models to assess genetic contributions regarding time-to-event data. From a methodological and statistical perspective, optimizing PRS with respect to the time-to-event character of the data (e.g., the hazard rate) rather than solely focusing on the binary outcome of occurrence might be more favorable [17].

To directly estimate PRS models for time-to-event data via boosting, we extended the snpboost algorithm by incorporating various new loss functions suitable for survival times. This illustrates the flexibility of our framework which results from the modular nature of the boosting algorithm: new loss functions can be implemented and directly combined with the overall framework, without altering the algorithm itself. For time-to-event and count phenotypes, one can now easily try different loss functions and pick the most suitable one depending on the phenotype’s characteristics. For time-related outcomes, the variable of interest is the time T_i until the event occurs (for subject i), which can, for example, be the survival time if the event is death, or the age of onset if the event is the first occurrence of a disease. Furthermore, censoring at time C_i might occur due to the end of the study, competing events or further reasons. The observed time for observation i is then given by $\tilde{T}_i = \min(C_i, T_i)$. Furthermore, the variable $\Delta_i = \mathbb{I}(T_i \leq C_i)$ indicates whether the event has been observed.

2.2.1 | Weighted L_2 Boosting

From an algorithmic point of view, an efficient way to develop PRS for time-to-event data is to adapt the classical squared error loss (L_2 loss) to account for censoring. To use the L_2 loss for time-to-event data, we consider the logarithmic survival times

$y_i = \log(T_i)$ and the corresponding logarithmic observed times $\tilde{y}_i = \log(\tilde{T}_i)$. The complete data $y_1, \dots, y_n$ could then be modeled using the squared error loss; however, due to censoring, we only have access to the observed incomplete data $\tilde{y}_1, \dots, \tilde{y}_n$. An effective way to deal with censoring is to incorporate weights [14, 18]. One popular choice is given by inverse probability censoring weights (IPC weights) defined as

$$w_i = \Delta_i \frac{1}{\hat{G}(\tilde{T}_i)} \quad (1)$$

where $\hat{G}(c)$ is the Kaplan–Meier estimator of the probability of censoring $G(c) = P(C > c)$. Based on the IPC weights, the loss function for weighted L_2 -boosting for time-to-event data is then given by

$$\rho_{\text{wL}_2}(\tilde{\mathbf{y}}, \boldsymbol{\eta}) = \frac{1}{n} \sum_{i=1}^n w_i (\tilde{y}_i - \eta_i)^2 \quad (2)$$

2.2.2 | Accelerated Failure Time Models

An attractive parametric alternative approach that avoids ordering and summing over all individuals (and is therefore computationally much less challenging than, for example, a Cox proportional hazard model), is based on accelerated failure time (AFT) models [19]. As an additional benefit, AFT models have a more intuitive interpretation and do not rely on assumptions like proportional hazards [20]. In these models the logarithmic observed survival times $\tilde{y}_i = \log(\tilde{T}_i)$ are modeled via

$$\log(\tilde{T}_i) = \eta_i + \sigma Z_i \quad (3)$$

where $\sigma > 0$ denotes a scaling parameter and Z_i is a random variable with known statistical distribution, for example, Gaussian (“AFT-Gaussian”), logistic (“AFT-logistic”) or extreme value distribution (“AFT-Weibull”). The corresponding loss function for boosting corresponds to the negative log-likelihood of this model and is given by

$$\begin{aligned} \rho_{\text{AFT}}(\tilde{\mathbf{y}}, \boldsymbol{\eta}, \sigma) = & - \sum_{i=1}^n \Delta_i \left(\log f\left(\frac{\tilde{y}_i - \eta_i}{\sigma}\right) - \log(\sigma) \right) \\ & - \sum_{i=1}^n (1 - \Delta_i) \left(\log S\left(\frac{\tilde{y}_i - \eta_i}{\sigma}\right) \right) \end{aligned} \quad (4)$$

where $f(\cdot)$ and $S(\cdot)$ are the probability density function and survival function of Z_i , respectively [21]. During the boosting algorithm, the additional unknown nuisance parameter σ is initialized (e.g., $\hat{\sigma}^{(0)} = 1$) and then updated after each boosting iteration by optimizing

$$\hat{\sigma}^{(m)} = \min_{\sigma > 0} \rho_{\text{AFT}}(\tilde{\mathbf{y}}, \hat{\boldsymbol{\eta}}^{(m)}, \sigma) \quad (5)$$

with $\hat{\boldsymbol{\eta}}^{(m)}$ being the estimated predictor after m boosting steps.

2.2.3 | Cox Proportional Hazard Model and Boosting the C-Index

A commonly used approach to model time-to-event data is the Cox proportional hazard model. [22] Here, the hazard

function is assumed to be the product of a baseline hazard and a factor depending on the linear predictor. The model relies on the assumption of proportional hazards. Furthermore, time-to-event models are often evaluated using the concordance index (C-index) $C = P(\eta_i > \eta_j | T_i < T_j)$ corresponding to the probability that a person with a shorter survival time is assigned with a higher risk score and vice versa [23]. The interpretation of the C-index is similar to that of the area under the receiver operating curve (AUC) for binary data: $C = 1$ indicates perfect discrimination via the risk score, while $C = 0.5$ corresponds to a random assignment of η . In earlier works focusing on low-dimensional data, statistical boosting has been proposed to directly optimize the linear predictor with respect to the C-index [24, 25].

We implemented loss functions for both the Cox proportional hazard model and the C-index in the snpboost framework. As both loss functions and corresponding gradients require ordering and summing over all n observations and predictions in each boosting step, these two approaches are computationally more demanding and limited in practice to small to medium sample sizes (with $n \leq 50\,000$). As our work is focused on the application of statistical boosting for genetic prediction models based on large cohort data, in this study we focus on the proposed alternatives, that is, AFT models and weighted L_2 boosting. Further details and results of snpboost based on the Cox proportional hazard model and the C-index for a reduced subsample with $n = 50\,000$ individuals can be found in the Appendix A. For a recent lasso-based approach to estimate PRS via a Cox model on UK Biobank data, we refer to Li et al. [9].

2.3 | Count Data

In many biomedical applications, the response variable represents count data (e.g., number of cells or number of episodes). In the current literature on genetic risk modeling, count data are typically treated as continuous outcomes, that is, standard linear models are fitted including a PRS that was selected and optimized based on univariate effects. However, it might be more favorable to actually use a model and to optimize the PRS based on a suitable discrete distribution for the count data, particularly if the range of observed values is small or the data is skewed.

2.3.1 | Poisson Regression

The most popular regression approach for count data is Poisson regression, which assumes that the outcome y_i follows a Poisson distribution with parameter $\lambda(\mathbf{X})$. We are then trying to predict $\eta_i = \log(\lambda(\mathbf{X}))$ by minimizing the negative log-likelihood of a Poisson regression, that is, the loss function is given by

$$\rho_{\text{Poisson}}(\mathbf{y}, \boldsymbol{\eta}) = \sum_{i=1}^n (-y_i \eta_i + \exp(\eta_i)). \quad (6)$$

In practice, when dealing with count data, the Poisson distribution is often too restrictive as it assumes that both expected value and variance are equal to $\lambda(\mathbf{X})$.

2.3.2 | Negative Binomial Distribution

In cases of overdispersed count data, a more flexible negative binomial regression model might be more suitable than Poisson regression [26]. The negative binomial distribution includes an additional parameter $\sigma > 0$ that allows for skewness in the data. Figuratively speaking, it describes the number of unsuccessful tries needed to reach a given number of successes in a Bernoulli process. The loss function used in statistical boosting again equals the negative log-likelihood and is given by

$$\rho_{\text{NB}}(\mathbf{y}, \boldsymbol{\eta}, \sigma) = - \sum_{i=1}^n \left(\log \left(\frac{\Gamma(y_i + \sigma)}{\Gamma(y_i + 1)\Gamma(\sigma)} \right) + \sigma \log \left(\frac{\sigma}{\exp(\eta_i) + \sigma} \right) + y_i \log \left(\frac{\exp(\eta_i)}{\exp(\eta_i) + \sigma} \right) \right) \quad (7)$$

where Γ denotes the Gamma function.

Similar as in AFT models, the additional unknown nuisance parameter $\sigma > 0$ is first initialized (e.g., $\hat{\sigma}^{(0)} = 1$) and then updated after each boosting iteration by minimizing

$$\hat{\sigma}^{(m)} = \min_{\sigma > 0} \rho_{\text{NB}}(\mathbf{y}, \hat{\boldsymbol{\eta}}^{(m)}, \sigma) \quad (8)$$

with $\hat{\boldsymbol{\eta}}^{(m)}$ being the estimated linear predictor after m boosting iterations.

2.4 | Quantile Regression

For continuous traits, most often classical mean regression is used to construct PRS, which corresponds to optimizing the L_2 loss (or a Gaussian likelihood). However, in some cases the actual interest might not be the center of the distribution but upper and lower quantiles, for example, in the context of over- or underweight [27]. Furthermore, for outcomes that are susceptible to measurement errors and outliers, estimation of the conditional median (corresponding to the L_1 loss) can be considered as a robust alternative to the conditional mean. Quantile regression, however, can also be used to derive individual conditional prediction intervals for a continuous outcome to quantify the uncertainty of a prediction [28, 29]. Classical prediction intervals, based on standard regression, assume homoscedasticity and therefore yield equidistant interval bounds whose size does not rely on the individuals but on the overall population. To overcome this, a very recent work by Hou et al. [30] propose to calibrate prediction intervals for diverse contexts (subgroups) specified by different covariates. We go one step further, and directly model the bounds of prediction intervals based on the individual genotype information.

In our context, quantile regression aims to predict the individual quantile $\tau \in (0, 1)$ conditioned on the individual's genetic information. To do so, we use the check function [31] as a loss function which is defined as

$$\rho_{\text{quantile}}(\mathbf{y}, \boldsymbol{\eta}) = \sum_{i=1}^n (\tau(y_i - \eta_i) \mathbb{I}(\eta_i \leq y_i) + (1 - \tau)(\eta_i - y_i) \mathbb{I}(\eta_i > y_i)) \quad (9)$$

where $\mathbf{y}$ is the vector of observed phenotypes and $\boldsymbol{\eta}$ the vector of corresponding predictions for the quantile τ . To construct prediction intervals for a coverage of $(1 - \alpha) \cdot 100\%$, we hence follow the general approach by Meinshausen and Ridgeway [28] and model the $\tau = \frac{\alpha}{2}$ and $\tau = 1 - \frac{\alpha}{2}$ quantiles as bounds of the individual intervals. As quantile regression yields point estimates and does not consider uncertainty in the estimated bounds, those can be refined using conformalized quantile regression [32]. To do so, on a validation set with n_{val} individuals, the conformity scores E_i are calculated as

$$E_i = \max\left(\hat{q}_{\frac{\alpha}{2},i} - y_i, y_i - \hat{q}_{1-\frac{\alpha}{2},i}\right) \quad (10)$$

where $\hat{q}_{\tau,i}$ is the estimated quantile τ of individual i . Then, the lower and upper estimated bounds are corrected by the $(1 - \alpha)\left(1 + \frac{1}{n_{\text{val}}}\right)$ th quantile of $\{E_1, \dots, E_{n_{\text{val}}}\}$.

3 | Simulation Study Based On UK Biobank Genotype Data

To assess the performance of our framework in a controlled setting, we conducted a simulation study based on UK Biobank (UKBB) genotype data [10], obtained under application number 81202. For the simulation study, we focused on the imputed genotypes of chromosome 22 and filtered for a minor allele frequency of at least 1% and a genotype rate of at least 90%, resulting in $p = 124\,302$ genetic variants. By using UK Biobank genotype data from chromosome 22, we ensure to base our simulations on realistic correlation structures among the genetic variants (LD). We randomly sampled $n = 100\,000$ individuals of which 50%, 20%, and 30% were assigned to training, validation and test sets, respectively. The genotype processing and sampling was conducted in PLINK 2.0 [33, 34]. In total, we sampled 100 different sets of individuals. For all phenotypes, we randomly chose 1200 influential variants, that is, 1% of the total number of variants. Based on simulated effect sizes $\boldsymbol{\beta}$ and a suitable intercept β_0 , the linear predictor $\boldsymbol{\eta}$ was defined as $\boldsymbol{\eta} = \beta_0 + \mathbf{X}\boldsymbol{\beta}$. Further details on the simulation of the phenotypes as well as the choices of effect sizes $\boldsymbol{\beta}$ and intercepts β_0 are given in the corresponding subsections. All simulations were run on a small HPC cluster using 30 CPUs with 2 GB RAM each.

3.1 | Time-To-Event Data

Our aim was to simulate realistic time-to-event data for the context of PRS, for example, the age of onset of a certain disease that might not occur in every individual. Therefore, we split the model into two components. For both components, the non-zero effect sizes β_j were drawn from a Gaussian distribution with mean $\mu = 0.02$ and standard deviation $s = 0.02$, reflecting the small effect sizes of common variants. In the first component, we simulated a binary variable, reflecting if the individual was at risk of developing the disease. While the linear predictor is positively correlated with the age-of-onset, it should have the opposite effect on the disease risk (i.e., a longer disease-free survival is associated with a lower disease risk). Therefore, in the first component, we used an adapted linear predictor $\boldsymbol{\eta}^{\text{bin}} = \beta_0^{\text{bin}} - \mathbf{X}\boldsymbol{\beta}$. The intercept was set to $\beta_0^{\text{bin}} = 13$. The binary observations $\mathbf{y}_{\text{bin}}$ were drawn from a

Bernoulli distribution with probabilities $\mathbf{p}_{\text{bin}} = \frac{\exp(\boldsymbol{\eta}^{\text{bin}})}{1 + \exp(\boldsymbol{\eta}^{\text{bin}})}$. Then, in the second component, survival and censoring times were simulated following an AFT model with logistic distribution. Here, the intercept was set to $\beta_0 = -10$. The logarithmic survival times were then drawn from $\log(t_{\text{surv}}) = \boldsymbol{\eta} + 0.02\mathbf{z}$, where $\mathbf{z}$ was simulated via a logistic distribution with location parameter of 0 and scale parameter of 1. Finally, the censoring times t_{cens} were drawn from a uniform distribution between 0 and $\max(t_{\text{surv}})$. In our simulations, an individual had an observed event if the binary indicator $y_{\text{bin},i}$ was 1, reflecting that the individual experienced the event, and the simulated survival time was smaller than the random censoring time. This reflects that the event already was observed, for example, the age of onset was smaller than the current age. The *observed* survival times are hence defined as

$$\tilde{t}_i = \begin{cases} t_{\text{surv},i} & \text{if } y_{\text{bin},i} = 1 \text{ and } t_{\text{surv},i} \leq t_{\text{cens},i} \\ t_{\text{cens},i} & \text{otherwise} \end{cases} \quad (11)$$

and the event indicator as

$$\Delta_i = \begin{cases} 1 & \text{if } y_{\text{bin},i} = 1 \text{ and } t_{\text{surv},i} \leq t_{\text{cens},i} \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

The resulting phenotypes had a mean censoring rate of 32.8% and a mean observed survival time of 19.5 years. On average, 25.8% of the individuals were modeled to not be at risk for the disease (binary component). Then, we derived PRS via snpboost using AFT loss functions with Gaussian, logistic and Weibull distributions, as well as weighted L_2 and binary loss functions, respectively. The results are shown in Figure 1. The weighted L_2 boosting showed the best prediction performance in terms of the C-index, followed by the AFT models with logistic and normal distributions. The AFT model with Weibull distribution performed slightly better than modeling the disease risk as a binary outcome. Additionally, we computed Kendall's τ to assess the agreement between the rankings of model predictions and true survival times. All time-to-event loss functions could predict the true ranking better than the binary loss function. Lastly, the weighted L_2 model showed an increased number of included variants compared with the other models which also led to longer runtimes. All models included more variants than the truly informative ones, which might be due to the high correlations of variants in the same region (LD).

3.2 | Count Data

We simulated two types of count phenotypes: one based on the Poisson distribution and the other based on the negative binomial distribution. For the Poisson distributed phenotype, effect sizes $\beta_1, \dots, \beta_{1200}$ were drawn independently from a normal distribution with mean $\mu = 0.04$ and standard deviation $s = 0.01$, reflecting again small effect sizes from common variants. Subsequently, the intercept was adaptively chosen via centering the linear predictor to ensure comparable expected values across simulation runs. Finally, observations were drawn from Poisson distributions with parameters $\lambda = \exp(\boldsymbol{\eta})$ resulting in phenotypes with an average mean of 1.45 and a median of 1. For a more skewed phenotype following a negative binomial distribution, effect sizes

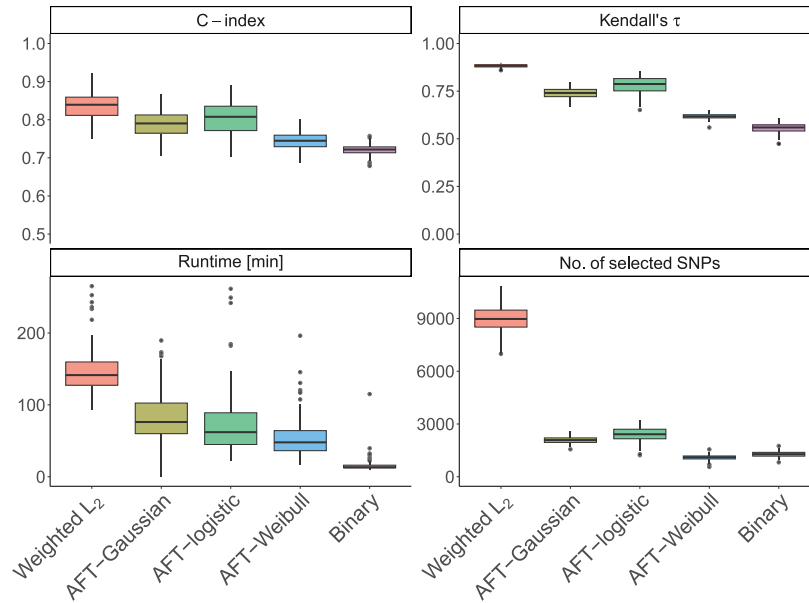


FIGURE 1 | Results of the simulation study for time-to-event phenotypes. PRS were fitted via snpboost using weighted L_2 boosting, AFT loss functions with Gaussian, logistic or Weibull distributions, and binary modeling of event occurrence, respectively. C-index, Kendall's τ with respect to the true event-free survival times, runtime in minutes, and number of selected SNPs in the PRS were evaluated.

$\beta_1, \dots, \beta_{1200}$ were Gaussian distributed with mean $\mu = 0.03$ and standard deviation $s = 0.01$. The intercept was again adaptively chosen and shifted by $\log(3)$. Then, observations were drawn from negative binomial distributions with parameters $\mu = \exp(\eta)$ and $\sigma = 2$ (see Equation (7)). The resulting phenotypes had an average mean of 4.97 and a median of 3. We trained and validated three PRS models on both types of count phenotypes using snpboost, employing (i) the L_2 loss, (ii) the Poisson loss, and (iii) the negative binomial loss. For each model, we assessed the mean absolute error of prediction (MAEP) and the mean squared error of prediction (MSEP) on the independent test set of $n = 30000$ individuals. Additionally, we also assessed the runtime in minutes and the number of selected SNPs in the PRS of the different models.

The results are illustrated in Figure 2 (negative binomial distributed outcome) and Figure B1 (Poisson distributed outcome). Additionally, as a benchmark reference, we calculated the MAEP and MSEP for the intercept model, using the mean of the training data as the intercept. Our results indicate that for phenotypes following Poisson and negative binomial distributions, using loss functions tailored for count data enhances the prediction performance compared to using the L_2 loss, which treats the data as continuous. For Poisson distributed phenotypes, all three PRS models yield a comparable number of selected variants. Regarding the predictive accuracy, the Poisson loss and the negative binomial loss lead to very similar predictive performance. However, as expected, PRS models with the negative binomial loss lead to larger runtimes due to the extra step of optimizing the additional parameter σ . Results are similar for the phenotypes following a negative binomial distribution, but considering the MAEP it is apparent, that the negative binomial loss deals better with an overdispersed phenotype and yields more accurate

predictions. In this case, the Poisson model includes considerably less SNPs than the L_2 , and negative binomial models.

3.3 | Quantile Regression and Prediction Intervals

For quantile regression, we simulated continuous phenotypes from a linear model with Gaussian distributed noise with standard deviation $s_{\text{noise}} = 2$ and effect sizes drawn from a Gaussian distribution with mean $\mu = 0.1$ and standard deviation $s = 0.01$. The intercept was set to $\beta_0 = 0$. The resulting phenotypes had an average genetic heritability of 52.5% (standard deviation 1.6%). We then applied quantile regression via snpboost to estimate 2.5%, 5%, 10%, 90%, 95%, and 97.5% quantiles. Then, prediction intervals with varying desired coverage rates were constructed; namely 80% (using 10% and 90% quantiles as bounds), 90% (using 5% and 95% quantiles as bounds), and 95% prediction intervals (using 2.5% and 97.5% quantiles as bounds). Finally, the achieved coverage rate defined by the proportion of observed phenotype values that lie within the prediction interval was assessed on the test set consisting of $n = 30000$ individuals.

In 100 simulated phenotypes an average coverage rate of 76.51% (standard deviation 0.29%), 86.99% (standard deviation 0.26%), and 92.80% (standard deviation 0.22%) were achieved for the 80%, 90%, and 95% prediction intervals, respectively. For all three prediction intervals, the achieved coverage is close to, but slightly below, the desired level. These results are expected since the estimated quantiles are point estimates that do not account for estimation uncertainty. However, the estimated prediction intervals can be further refined by using conformalized quantile regression [32] (see Section 2.4). These refined prediction intervals exhibit

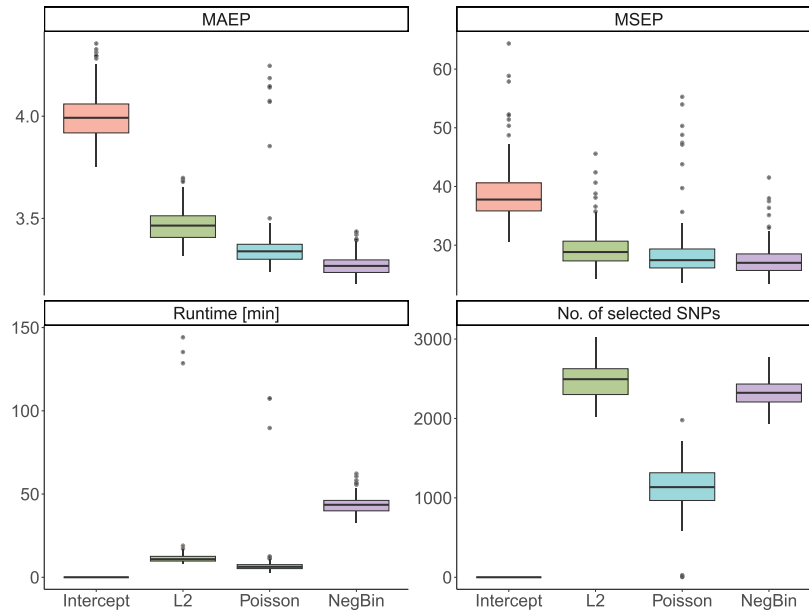


FIGURE 2 | Results of a simulation study for count phenotypes following a negative binomial distribution. PRS were fitted via snpboost using L_2 , Poisson and negative binomial (NegBin) loss functions, respectively. Mean absolute error of prediction (MAEP), mean squared error of prediction (MSEP), runtime in minutes, and number of selected variants in the PRS were evaluated.

average coverage rates of 80.04% (standard deviation 0.36%), 90.00% (standard deviation 0.26%), and 94.99% (standard deviation 0.19%), respectively.

4 | UK Biobank Application

We applied our adapted boosting approach, utilizing the new targeted loss functions, to analyze various types of outcome data from the UKBB (application number 81202). This dataset also served as the motivation for our methodological developments. The UKBB is a large-scale database that comprises extensive phenotype and genotype data of more than 500 000 individuals. Starting in 2006, individuals aged 40 and older were enrolled and data is added and updated consecutively in follow-up visits and questionnaires. Besides self-reported traits through questionnaires and data from study visits, the UKBB is linked to electronic health records and death registries through the National Health Service. Additionally, blood samples of all participants were sequenced resulting in large-scale genotype data. In our applications, we excluded individuals of self-reported non-European ancestry as well as related individuals. Furthermore, we restricted the genotype data to genotyped variants, and excluded both imputed variants and those on the sex chromosomes, resulting in $p = 604\,967$ genetic variants.

4.1 | Time-to-Event Data

Asthma is a highly prevalent respiratory disease influenced by various risk factors including genetic and environmental factors. Genetically, it shows a polygenic architecture and in particular a low age of onset is characteristic for a genetic predisposition

[35]. In the UKBB, we identified $n = 56\,436$ individuals diagnosed with asthma by using the corresponding ICD-9 (UKBB Field 41271, codes 493, 493[0, 1, 09, 19, 9, 90, 99]) and ICD-10 codes (UKBB Field 41270, codes J450, J451, J548, J459) as well as self-reported cases (UKBB Field 22127). Time of diagnosis varied between the year of birth and 84 years of age, with a mean age at time of first diagnosis of 38 years. Additionally, we identified $n = 339\,208$ controls not diagnosed with asthma. As a second time-to-event outcome variable, we considered age at first diagnosis of gout. Gout, a common disease in Western populations, may be influenced by genetic factors, particularly in cases of early onset. Recently, Li et al. [9] found an association of PRS and age at diagnosis in gout. Using again ICD-10 codes (codes M10[0, 2-4, 9], M100[0-9], M102[0-9], M103[0-9], M104[0-9], M109[0-9], respectively), we found $n = 7\,409$ individuals diagnosed with gout in the UKBB and $n = 339\,202$ controls. Age at first diagnosis ranged from 32 years to 84 years, with a mean age of 67 years.

For both of the time-to-event outcomes, 50% of the individuals were assigned to the training set and 25% to the validation and test set, respectively. The AFT loss functions as well as the weighted L_2 boosting were applied on the complete dataset fitting one PRS with each loss function for both phenotypes using the default values of the snpboost algorithm (batch size $p_{\text{batch}} = 1000$, step length $\nu = 0.1$). Additionally, to compare the new loss functions to the most common approach in the PRS field, we fitted a “binary” PRS via snpboost using the occurrence of the disease as an outcome variable. The application of the Cox proportional hazard model and the C-index on a reduced subsample of $n = 50\,000$ individuals (also compared to all other time-to-event loss functions) is included in Appendix A (Figure A1).

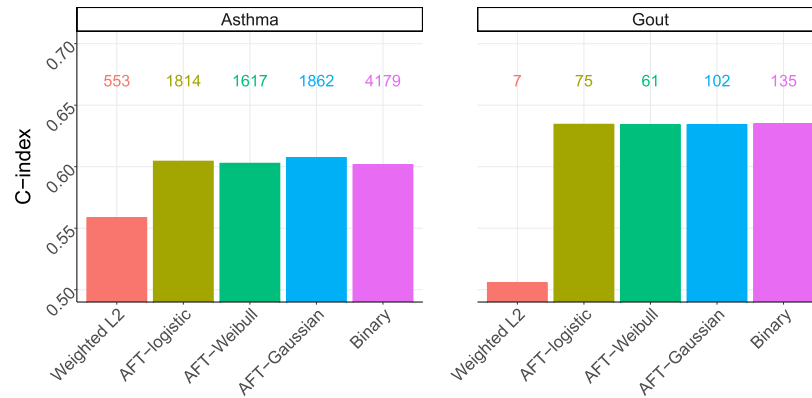


FIGURE 3 | C-indices on the independent test set for the PRS models estimated using different loss functions. The numbers above the bars indicate the number of SNPs with non-zero effect in each model.

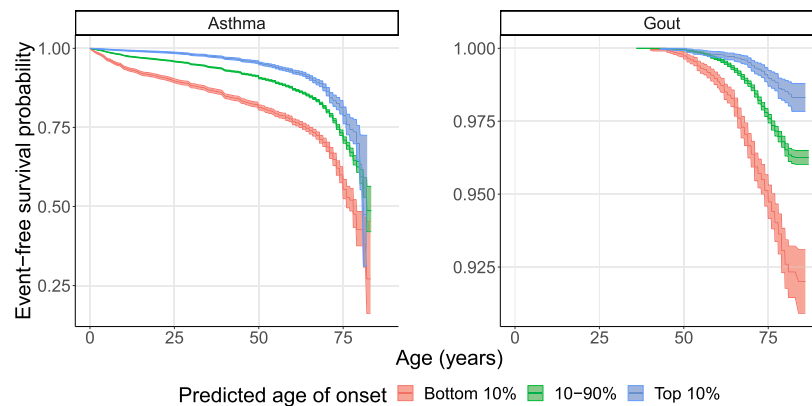


FIGURE 4 | Kaplan–Meier estimates on the independent test set for the AFT-Gaussian models. The individuals were stratified into risk groups according to the percentiles of the predicted age of onset derived from the training and validation sets. Since AFT models regress the logarithmic survival times, the PRS corresponds to the predicted age of onset, that is, a higher PRS is associated with a higher probability of event-free survival.

To compare the performance of the various PRS models, we used the C-index implemented in the function `rcorr.cens` from the R package `Hmisc` on the independent test set that was not used for deriving the PRS. The results are illustrated in Figure 3. While all fitted models except for the weighted L_2 model for gout were able to detect genetic signal in both outcomes, we found the AFT models to be superior to the other loss functions with respect to the prediction performance. In comparison to the “binary” PRS, the time-to-event loss functions yielded sparser models with the same or superior prediction performance. However, the improvement was not as clear as in the simulation study.

In addition to the predictive discriminatory performance in terms of the C-index, we compared the ranking of the individuals according to each model. Figure B2 shows Kendall’s τ coefficient of the PRS for asthma on the test set resulting from different loss functions. The AFT models show a highly similar ranking ($\tau \geq 0.8$), in particular when assuming a Weibull or logistic distribution. Interestingly, the ranking of the AFT models differ moderately from the binary approach ($\tau \in (0.68 - 0.77)$) suggesting that the time-to-event loss functions capture different

information than the binary loss function. As expected due to the reduced performance of the weighted L_2 model, the ranking according to this approach differs highly from the other four models.

We then utilized the best performing model, that is, the AFT-Gaussian model, to stratify individuals into risk groups based on their predicted age of onset. For each outcome separately, we calculated the 10% and 90% quantiles of the combined training and validation set. These quantiles were then used as thresholds to divide all individuals into three risk groups: “bottom 10%,” “10%–90%,” “top 90%.” Figure 4 displays the resulting Kaplan–Meier curves on the independent test set. For both traits, individuals in the tails of the predicted age of onset (PRS) distribution show a decreased or increased disease-free survival probability compared to individuals in the center of the predicted age of onset (PRS) distribution. Notably, since AFT models yield a PRS predicting the survival times, in this context, a higher PRS is associated with a longer survival time and a smaller lifetime risk. Comparing the rankings of the PRS models, 13.7% of the individuals in the test set would have been assigned to different risk groups based on the binary and the AFT-Gaussian models

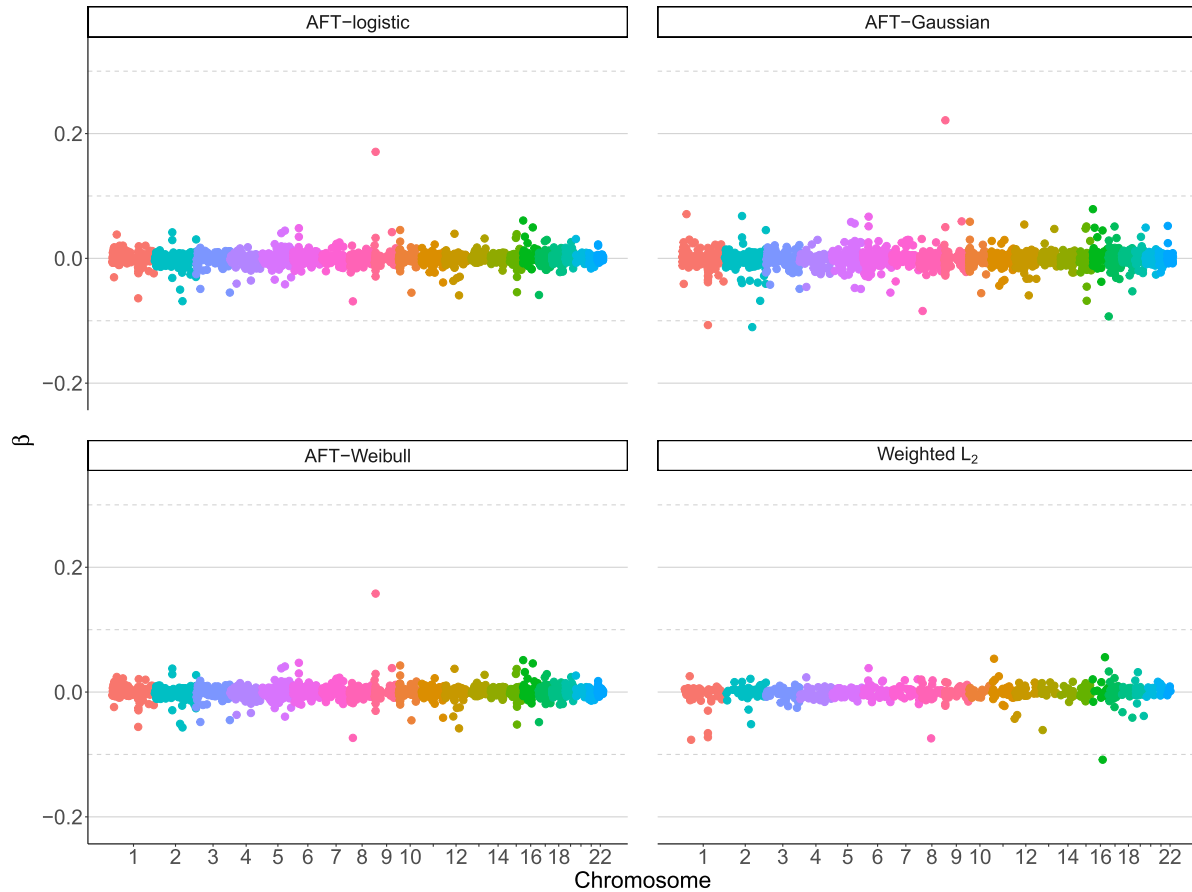


FIGURE 5 | Estimated coefficients for age of onset of asthma using different loss functions. Only non-zero effects are shown.

for asthma and 4% of the individuals would have been differently assigned for gout.

Figure 5 displays the estimated coefficients for asthma using the different loss functions. While the models differ in the number of included variants, the selected variants are spread throughout the genome for each method. As a sensitivity analysis, we included sex and the first 10 principle components as covariates, which resulted in highly correlated effect estimates (Pearson correlation $r > 0.99$, Figure B3). To assess the potential biological association between the variants included in the PRS models and the predicted phenotype, we performed a variant-to-gene mapping using genomic coordinates. Variants were mapped to colocalizing genes via bedtools if located in exons or introns, or to the closest gene within 500 KB if in an intergenic region, to include regulatory variants [36]. In fact, the selected variants might have true mechanistic associations or just be in linkage disequilibrium (LD) with truly associated variants and have no biological effect. Therefore, the selected variants, rather than detecting causal variants, act as proxies for signals in the locus, which could be biologically linked to the mapped genes. The enrichment analysis of the mapped genes using Fisher's Exact Test in Enrichr identified asthma as the top association in the GWAS catalog and showed an odds ratio of 6.34 (adjusted p -value of $9.43\text{E}-48$) [37]. This corroborates that the variants included in the PRS capture the

genetic signal associated with asthma susceptibility. In line with the enrichment analysis, also the SNP with the highest effect size in all three AFT models, namely rs146597587 in chromosome 9, has been previously associated with asthma [38, 39]. Also SNP rs34210653 in chromosome 17, which is included in all models, has been linked to asthma before [40]. The corresponding figure for gout is included in Appendix B (Figure B4).

Table 1 provides the computational times to derive the PRS when running the algorithm on a small HPC cluster with 30 CPUs and 2 GB RAM per CPU. While the computational times for the weighted L_2 boosting and the AFT loss functions are comparable it is notable that they largely depend on the number of included variants.

4.2 | Count Data

To demonstrate the application of our boosting approach using targeted loss functions for count data, we considered the number of sunburn occasions during childhood (UKBB Field 1737). We found $n = 300073$ individuals (randomly assigned to 50% training data, 25% validation data, and 25% test data) with a median number of 0 sunburn occasions during childhood (interquartile range (IQR) 0 – 2 sunburn occasions, maximum 999 sunburn

TABLE 1 | Results on the independent test sets for the age of onset of asthma and gout using five different loss functions.

Trait	PRS	C	No. of selected SNPs	Runtime (h)
Asthma	Binary	0.60	4179	2.04
Asthma	Weighted L_2	0.56	553	1.22
Asthma	AFT-Weibull	0.60	1617	2.64
Asthma	AFT-logistic	0.60	1814	3.03
Asthma	AFT-Gaussian	0.61	1862	5.87
Gout	Binary	0.63	135	0.59
Gout	Weighted L_2	0.51	7	1.09
Gout	AFT-Weibull	0.63	61	0.53
Gout	AFT-logistic	0.63	75	0.51
Gout	AFT-Gaussian	0.63	102	0.73

Note: Reported are the C-index resulting from the PRS models (C), the number of included SNPs in the PRS and the computational time in hours. All calculations were run on an HPC cluster with 30 CPUs and 2 GB RAM each.

TABLE 2 | Results on the independent test sets for the number of sunburn occasions during childhood using L_2 , Poisson and negative binomial (NegBin) loss functions.

Model	MSEP	MAEP	No. of selected SNPs	Runtime (min)
Intercept	22.84	1.92	0	0.0
L_2	22.76	1.90	58	9.14
Poisson	22.64	1.88	1073	16.65
NegBin	22.61	1.85	1297	99.19

Note: For the intercept model, the mean on the training set was used as an intercept. Reported are the mean squared error of prediction (MSEP) as well as the mean absolute error of prediction (MAEP), the number of included SNPs in the PRS, and the computational time in minutes. All calculations were run on an HPC cluster with 30 CPUs and 2 GB RAM each.

occasions). We fitted three PRS models via snpboost on the training data: (1) using the L_2 -loss for quantitative traits, (2) using the Poisson loss and (3) using the negative binomial distribution loss function. The validation set was used for the early stopping of the algorithm. Then, the three models were evaluated on the test set via both the mean squared error of prediction (MSEP) and the mean absolute error of prediction (MAEP). Additionally, we considered a simple intercept model with the median on the training data as the intercept.

The results are shown in Table 2. As expected, due to the additional estimation of σ in each boosting step for the negative binomial model, the fitting of the Poisson model and the L_2 model is substantially faster. Even though there is a trend of improved predictions performance from the simplest intercept model toward the more complex negative binomial model, it is only a modest improvement in terms of MSEP and MAEP. This is probably due to a low genetic signal in the considered traits, poor data quality due to retrospective self-reporting, and other influencing factors such as sun exposure and sun protection.

4.3 | Quantile Regression

Finally, we applied quantile regression via snpboost on a set of $n = 262171$ unrelated individuals from the UK Biobank (divided into training, validation and test set) to construct prediction intervals for BMI (UKBB Field 21001) and lipoprotein A (UKBB Field 30790). For 78% of the individuals in the test set ($n = 52551$),

the observed BMI was covered by individual prediction intervals between the predicted 10% and 90% quantiles that were based on genetic information solely. This could be increased to 80% coverage rate for the conformalized prediction interval. As shown in Figure 6, we found the length of the prediction interval for BMI to be positively correlated with the predicted median as well as the observed BMI. In the observed data, an increased standard deviation is present in both, lower and upper deciles of the BMI (Figure B5). Therefore, the linear increase in the length of the prediction interval indicates that the genetic predisposition might not only have an effect on the level but also on the dispersion parameters, for example, the variance.

With respect to lipoprotein A, 79% of the observed values on the independent test set were covered by the prediction interval (80% by the conformalized prediction interval). For lipoprotein A, one SNP, namely rs10455872, has a strong modulation effect on the blood marker's level. [7] Figure 7 shows the densities of the observed values as well as the predicted quantiles, all on the independent test set. The individuals were subgrouped into those without any effect allele in SNP rs10455872 and those with at least one effect allele in this SNP. In the upper left panel, it is evident that individuals with at least one effect allele in this SNP have higher lipoprotein A values and the variance of the trait is also increased. As seen in the remaining three panels, snpboost is able to capture this shift in all of the three predicted quantiles. Additionally, the density of the 90% quantiles of those individuals with at least one variant in rs10455872 is further shifted to the right compared to the density of the 10% quantiles than for those

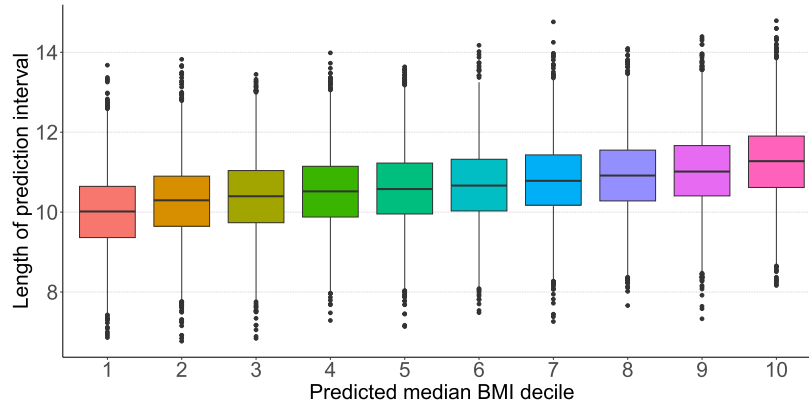


FIGURE 6 | Length of the prediction interval constructed by the predicted 10% and 90% quantiles in relation to the decile of the predicted median BMI on an independent test set.

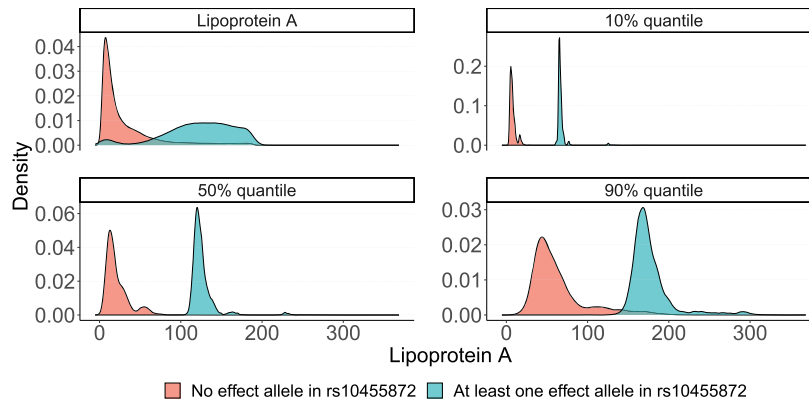


FIGURE 7 | Observed lipoprotein A values on an independent test set and the predicted 10%, 50%, and 90% quantiles. Individuals are subgrouped into those with no effect allele in SNP rs10455872 and those with at least one effect allele in this SNP, since this SNP was found to have a highly modulating effect on lipoprotein A.

individuals without an effect allele in rs10455872. This means, that the difference between the 10% and the 90% quantiles is increased for those individuals with at least one effect allele in rs10455872. As for BMI, this corresponds to a wider prediction interval capturing the higher variance in those individuals.

5 | Discussion

In this work, we address the challenges presented by large-scale and high-dimensional genotype data by proposing solutions that enable for the first time the application of advanced statistical modeling approaches. Specifically, we transfer biostatistical methods to this genetic setting with very large numbers of observations and predictor variables by extending the statistical boosting framework `snptest`. Our approach involves the incorporation of targeted loss functions to construct PRS tailored for diverse outcome types, including time-to-event data, count data as well as skewed and potentially corrupted data. So far, most GWAS and PRS studies are limited to linear and logistic regression focusing on continuous and binary outcomes, respectively [41, 42]. More recent works considered modeling

time-to-event data by using Cox proportional hazard models [17, 43, 44]. However, most of these approaches are based on univariate associations. Li et al. [9] proposed a Lasso algorithm, `snptest`, to apply penalized Cox regression directly on the individual-level genotype data. In this work, we introduce the possibility to use statistical boosting to fit different alternative time-to-event models, such as AFT models, on individual genotype data. These parametric models are computationally less demanding and do not rely on the proportional hazard assumption. We have shown that such models can be used to effectively stratify the cumulative lifetime risk.

For count data, our approach is able to incorporate loss functions that correspond to Poisson and negative binomial regression. While for RNA sequencing data, the use of a negative binomial regression is a standard approach [45], count data in GWAS is so far mostly modeled using classical Gaussian models. The possibility of actually considering count data models is particularly beneficial for data with limited counts and skewed distributions. This might also be of interest for transcriptome-wide association studies (TWAS [46]), which consider the expression data of genes as an outcome regressed on genotype data.

Finally, we incorporate quantile regression in our framework as a robust alternative to mean regression models, particularly in the presence of skewed and corrupted outcome data. Quantile regression enables the construction of individual prediction intervals that are often lacking in standard PRS models. Ongoing research focused on the construction of context-specific prediction intervals which are individualized with respect to covariates like age, sex, and ancestry [30]. Quantile regression provides the possibility to construct individual and possibly asymmetric prediction intervals based on genetic predisposition. To also correct for the uncertainty inherited from estimating the bounds of the intervals via quantile regression, the estimated bounds can be further refined via conformalized quantile regression [32]. Furthermore, quantile regression can help to detect heteroscedasticity.

Interestingly, in our application, we observe only a modest increase in prediction performance when using targeted loss functions for time-to-event and count data. Since we have demonstrated a superior performance of the proposed loss functions in simulated data, we hypothesize that this might be due to the nature of the analyzed UK Biobank data. As this is a large population-based cohort, the number of controls most often greatly exceeds the number of cases, in particular for less prevalent traits. Furthermore, time-to-event data are usually noisier than the binary response if a disease was diagnosed or not. From a genetic point of view, we have only investigated the genetic component that can be measured via common variants, that is, we have neglected rare variants that can have potentially highly relevant effects. To improve prediction performance, it might be of interest to integrate rare variants and the derived PRS in one model as has been done in other works [47, 48]. We have also not considered any additional covariates in our prediction models, such as age, sex, or the first principal components of the genotype matrix. There are several approaches to include covariates in the final prediction model, for example, integrating PRS and covariates in a multiple regression model. While the inclusion of those covariates in the final model would most likely improve the prediction performance of our models, we have focused here on detecting the genetic signal based on the individual-level genotype data. In snpboost, it is also possible to include covariates whose coefficients are then updated in each boosting iteration [7]. In further research, it is of interest to analyze and compare different approaches for covariate correction which could also directly consider covariate-SNP interactions. Noteworthy, our analysis is only based on individuals with European ancestry. It is known that PRS show decreased performances in out-of-target populations, that is, populations of ancestries that are underrepresented in the training data [49]. There is a strong need for more diverse biobanks as well as for the development of methods that effectively include different ancestries to achieve more robust predictions across out-of-target populations and admixed individuals. While the proposed loss functions target specific data types, they do not explicitly aim to increase the transferability across populations of different ancestries. In particular, the PRS derived in our UK Biobank application will be biased toward European ancestry since they were only trained on this population. Further research is warranted to extend the snpboost framework by targeted loss functions that can bridge the gap of transferability.

On the methodological side, our approach inherits some limitations from the original statistical boosting framework. Boosting

does not provide closed formulas for standard errors for coefficient estimates, that is, statistical inference is not directly possible. Confidence intervals can be constructed via permutation based approaches or bootstrapping, however, the computational demand is not negligible. Additionally, the algorithm includes several potential tuning parameters, including the step length, the batch size, and the stopping iteration. While the stopping iteration is tuned via the independent validation set, the step length and the batch size are typically pre-specified and kept fixed. While in previous work [7] we found the algorithm to be quite robust with respect to varying batch sizes and step lengths, we have observed that a much larger step length is beneficial for boosting the C-index. Lastly, we have also incorporated loss functions to derive Cox-like models and optimize the C-index. However, those loss functions and their gradients are computationally highly demanding and are therefore limited to small to medium sized cohorts with $n \leq 50000$ individuals. In our application, we observed a limited prediction performance in a reduced subsample.

Further research is warranted on how snpboost could provide means to identify SNPs that are associated with the variability of a trait rather than the general level. Such SNPs, which are also referred to as vQTLs [50, 51], could be indicators for gene-environment interactions. A future approach to identify them could be to extend our framework to location-scale models in the spirit of generalized additive models for location, scale and shape (GAMLSS [52, 53]), which directly model potential effects of the variants on the variance of a trait. Other model classes that could extend the framework are conditional transformation models [54], as well as multivariate distributional regression models, that allow modeling multiple response variables and their association [55]. However, this requires the simultaneous fitting of several linear predictors which is computationally far more expensive. Additionally, determining the optimal approach for defining different batches of variants for the multiple distribution parameters remains a challenge. Further research could also investigate not only extending the framework to new model classes but also to different types of genetic effects. Currently our framework is focused only on linear base-learners, following the assumption of most polygenic models in the genetics field. A recent work of Heyne et al. [56] investigates the potential of non-linear genotype effects. Hence, in future work, it could be of interest to incorporate non-linear base-learners to take the dominance structure of the SNPs into account, that is, to allow also for a recessive or dominant rather than an additive effect. The inclusion of additional base-learners also represents the potential to model gene-gene and gene-environment interactions [57]. With the inclusion of those interactions, new methodological and computational challenges arise, for example, which interactions should be included to ensure biological meaningfulness and avoid computational explosion. For example, gene-environment interactions could be included based on the detection of vQTLs [58] and gene-gene interactions could be considered pathway-specific [59]. Lastly, snpboost is based on individual-level genotype data. Despite those becoming increasingly available to the scientific community, many PRS approaches are based on summary statistics from genome-wide association studies. From a data protection view, summary statistics are generally easier to share compared to individual-level data and can be combined in a meta-analysis. To overcome this limitation

of individual-level based approaches, federated learning can be deployed in further research to fit individual-level based PRS without violation of data sharing agreements.

In summary, we extended the boosting framework for huge genetic data by incorporating various loss functions enabling targeted PRS modeling. This is in the spirit of developing PRS as actual statistical prediction models, rather than summing up and weighting effect alleles as combined biomarkers. In this context, statistical boosting can be applied to individual-level data to derive outcome-specific prediction models, which may have the potential to improve the clinical usability of PRS in the future. To the best of our knowledge, many of the proposed model classes are currently not feasible to estimate with these kinds of data. For instance, we introduce the first algorithm to fit quantile regression for PRS on large-scale individual-level genotype data, enabling the construction of prediction intervals. Additionally, we demonstrate the modular nature of statistical boosting, which facilitates further extensions of the framework.

Author Contributions

H.K., A.M., C.S., C.M., and P.M.K. contributed to the conception and design of the method. H.K. wrote the code, performed the experiments, and wrote the first draft of the manuscript. All authors contributed to the manuscript revision, read and approved the submitted version.

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 81202.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

This research has been conducted using the UK Biobank resource under application number 81202 (<http://www.ukbiobank.ac.uk>). Our method is available as an R implementation on GitHub (<https://github.com/hklinkhammer/snpboost>). R scripts that were used in the simulation study can be found in the [Supporting Information](#).

References

1. A. Majumdar and B. Pasaniuc, "A Bayesian Method for Estimating Gene-Level Polygenicity Under the Framework of Transcriptome-Wide Association Study," *Statistics in Medicine* 42, no. 26 (2023): 4867–4885, <https://doi.org/10.1002/sim.9892>.
2. X. Shen, Y. Wen, Y. Cui, and Q. Lu, "A Conditional Autoregressive Model for Genetic Association Analysis Accounting for Genetic Heterogeneity," *Statistics in Medicine* 41, no. 3 (2021): 517–542, <https://doi.org/10.1002/sim.9257>.
3. T. Ge, C. Y. Chen, B. M. Neale, M. R. Sabuncu, and J. W. Smoller, "Phenome-Wide Heritability Analysis of the UK Biobank," *PLoS Genetics* 13, no. 4 (2017): e1006711, <https://doi.org/10.1371/journal.pgen.1006711>.
4. F. Privé, J. Arbel, and B. J. Vilhjálmsson, "LDpred2: Better, Faster, Stronger," *Bioinformatics* 36, no. 22–23 (2020): 5424–5431, <https://doi.org/10.1093/bioinformatics/btaa1029>.
5. T. Ge, C. Y. Chen, Y. Ni, Y. C. A. Feng, and J. W. Smoller, "Polygenic Prediction via Bayesian Regression and Continuous Shrinkage Priors," *Nature Communications* 10, no. 1 (2019): 1–10, <https://doi.org/10.1038/s41467-019-09718-5>.
6. S. W. Choi and P. F. O'Reilly, "PRSice-2: Polygenic Risk Score Software for Biobank-Scale Data," *GigaScience* 8, no. 7 (2019): 1–6, <https://doi.org/10.1093/gigascience/giz082>.
7. H. Klinkhammer, C. Staerk, C. Maj, P. M. Krawitz, and A. Mayr, "A Statistical Boosting Framework for Polygenic Risk Scores Based on Large-Scale Genotype Data," *Frontiers in Genetics* 13 (2023): 1076440, <https://doi.org/10.3389/fgene.2022.1076440>.
8. J. Qian, Y. Tanigawa, W. Du, et al., "A Fast and Scalable Framework for Large-Scale and Ultrahigh-Dimensional Sparse Regression With Application to the UK Biobank," *PLoS Genetics* 16, no. 10 (2020): e1009141, <https://doi.org/10.1371/journal.pgen.1009141>.
9. R. Li, C. Chang, J. M. Justesen, et al., "Fast Lasso Method for Large-Scale and Ultrahigh-Dimensional Cox Model With Applications to UK Biobank," *Biostatistics* 23, no. 2 (2022): 522–540, <https://doi.org/10.1093/biostatistics/kxaa038>.
10. C. Bycroft, C. Freeman, D. Petkova, et al., "The UK Biobank Resource With Deep Phenotyping and Genomic Data," *Nature* 562, no. 7726 (2018): 203–209, <https://doi.org/10.1038/s41586-018-0579-z>.
11. R. E. Schapire, "The Strength of Weak Learnability," *Machine Learning* 5, no. 2 (1990): 197–227, <https://doi.org/10.1007/bf00116037>.
12. Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences* 55, no. 1 (1997): 119–139, <https://doi.org/10.1006/jcss.1997.1504>.
13. J. H. Friedman, "Stochastic Gradient Boosting," *Computational Statistics and Data Analysis* 38, no. 4 (2002): 367–378, [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2).
14. P. Bühlmann and T. Hothorn, "Boosting Algorithms: Regularization, Prediction and Model Fitting," *Statistical Science* 22, no. 4 (2007): 477–505, <https://doi.org/10.1214/07-sts242>.
15. N. Dueñas, H. Klinkhammer, N. Bonifaci, et al., "Ability of a Polygenic Risk Score to Refine Colorectal Cancer Risk in Lynch Syndrome," *Journal of Medical Genetics* 60, no. 11 (2023): 1044–1051, <https://doi.org/10.1136/jmg-2023-109344>.
16. G. Saldanha, K. Joshi, K. Lawes, et al., "5-Hydroxymethylcytosine Is an Independent Predictor of Survival in Malignant Melanoma," *Modern Pathology* 30, no. 1 (2017): 60–68, <https://doi.org/10.1038/modpathol.2016.159>.
17. H. Syed, A. L. Jorgensen, and A. P. Morris, "Evaluation of Methodology for the Analysis of 'Time-To-Event' Data in Pharmacogenomic Genome-Wide Association Studies," *Pharmacogenomics* 17, no. 8 (2016): 907–915, <https://doi.org/10.2217/pgs.16.19>.
18. T. Hothorn, "Survival ensembles," *Biostatistics* 7, no. 3 (2005): 355–373, <https://doi.org/10.1093/biostatistics/kxj011>.
19. J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data* (New York, NY: Springer, 2003).
20. P. Hougaard, "Fundamentals of Survival Data," *Biometrics* 55, no. 1 (1999): 13–22.
21. M. Schmid and T. Hothorn, "Flexible Boosting of Accelerated Failure Time Models," *BMC Bioinformatics* 9, no. 1 (2008): 269, <https://doi.org/10.1186/1471-2105-9-269>.
22. D. R. Cox, "Regression Models and Life-Tables," *Journal of the Royal Statistical Society: Series B: Methodological* 34, no. 2 (1972): 187–220.
23. M. Schmid and S. Potapov, "A Comparison of Estimators to Evaluate the Discriminatory Power of Time-To-Event Models," *Statistics in Medicine* 31, no. 23 (2012): 2588–2609.
24. A. Mayr and M. Schmid, "Boosting the Concordance Index for Survival Data – A Unified Framework To Derive and Evaluate Biomarker Combinations," *PLoS One* 9, no. 1 (2014): e84483, <https://doi.org/10.1371/journal.pone.0084483>.

25. A. Mayr, B. Hofner, and M. Schmid, "Boosting the Discriminatory Power of Sparse Survival Models via Optimization of the Concordance Index and Stability Selection," *BMC Bioinformatics* 17, no. 1 (2016): 1–12, <https://doi.org/10.1186/s12859-016-1149-8>.
26. C. I. Bliss and R. A. Fisher, "Fitting the Negative Binomial Distribution to Biological Data," *Biometrics* 9, no. 2 (1953): 176–200, <https://doi.org/10.2307/3001850>.
27. N. Fenske, T. Kneib, and T. Hothorn, "Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression," *Journal of the American Statistical Association* 106, no. 494 (2011): 494–510.
28. N. Meinshausen and G. Ridgeway, "Quantile regression forests," *Journal of Machine Learning Research* 7, no. 6 (2006): 983–999.
29. A. Mayr, T. Hothorn, and N. Fenske, "Prediction Intervals for Future BMI Values of Individual Children—a Non-parametric Approach by Quantile Boosting," *BMC Medical Research Methodology* 12 (2012): 1–13.
30. K. Hou, Z. Xu, Y. Ding, et al., "Calibrated Prediction Intervals for Polygenic Scores Across Diverse Contexts," *Nature Genetics* 56 (2024): 1386–1396.
31. R. Koenker and K. F. Hallock, "Quantile Regression," *Journal of Economic Perspectives* 15, no. 4 (2001): 143–156, <https://doi.org/10.1257/jep.15.4.143>.
32. Y. Romano, E. Patterson, and E. Candes, "Conformalized Quantile Regression," in *Advances in Neural Information Processing Systems*. 32. 33rd Conference on Neural Information Processing Systems, eds. H. Wallach, H. Larochelle, A. Beygelzimer, d F. Alché-Buc, E. Fox, and R. Garnett (Vancouver, Canada: Curran Associates, Inc, 2019).
33. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets," *GigaScience* 4, no. 1 (2015): 1–16, <https://doi.org/10.1186/s13742-015-0047-8>.
34. S. Purcell, "Chang C. PLINK 2.0," 2015, www.cog-genomics.org/plink/2.0/.
35. S. F. Thomsen, "Genetics of Asthma: An Introduction for the Clinician," *European Clinical Respiratory Journal* 2, no. 1 (2015): 24643, <https://doi.org/10.3402/ecrj.v2.24643>.
36. A. R. Quinlan and I. M. Hall, "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features," *Bioinformatics* 26, no. 6 (2010): 841–842.
37. Z. Xie, A. Bailey, M. V. Kuleshov, et al., "Gene Set Knowledge Discovery With Enrichr," *Current Protocols* 1, no. 3 (2021): e90.
38. Z. Zhu, X. Zhu, C. L. Liu, et al., "Shared Genetics of Asthma and Mental Health Disorders: A Large-Scale Genome-Wide Cross-Trait Analysis," *European Respiratory Journal* 54, no. 6 (2019): 1901507, <https://doi.org/10.1183/13993003.01507-2019>.
39. Z. Zhu, Y. Guo, H. Shi, et al., "Shared Genetic and Experimental Links Between Obesity-Related Traits and Asthma Subtypes in UK Biobank," *Journal of Allergy and Clinical Immunology* 145, no. 2 (2020): 537–549, <https://doi.org/10.1016/j.jaci.2019.09.035>.
40. Å. Johansson, M. Rask-Andersen, T. Karlsson, and W. E. Ek, "Genome-Wide Association Analysis of 350000 Caucasians From the UK Biobank Identifies Novel Loci for Asthma, Hay Fever and Eczema," *Human Molecular Genetics* 28, no. 23 (2019): 4022–4041, <https://doi.org/10.1093/hmg/ddz175>.
41. N. Mars, E. Widén, S. Kerminen, et al., "The Role of Polygenic Risk and Susceptibility Genes in Breast Cancer Over the Course of Life," *Nature Communications* 11, no. 11 (2020): 6383, <https://doi.org/10.1038/s41467-020-19966-5>.
42. N. Tesi, v. d S J. Lee, M. Hulsman, et al., "Polygenic Risk Score of Longevity Predicts Longer Survival Across an Age Continuum," *Journals of Gerontology: Series A* 76, no. 5 (2021): 750–759, <https://doi.org/10.1093/gerona/glaa289>.
43. S. E. Ojavee, A. Kousathanas, D. Trejo Banos, et al., "Genomic Architecture and Prediction of Censored Time-To-Event Phenotypes With a Bayesian Genome-Wide Analysis," *Nature Communications* 12, no. 11 (2021): 2337, <https://doi.org/10.1038/s41467-021-22538-w>.
44. E. M. Pedersen, E. Agerbo, O. Plana-Ripoll, et al., "ADuLT: An Efficient and Robust Time-To-Event GWAS," *Nature Communications* 14, no. 1 (2023): 1–12, <https://doi.org/10.1038/s41467-023-41210-z>.
45. T. S. Andrews, V. Y. Kiselev, D. McCarthy, and M. Hemberg, "Tutorial: Guidelines for the Computational Analysis of Single-Cell RNA Sequencing Data," *Nature Protocols* 16, no. 1 (2020): 1–9, <https://doi.org/10.1038/s41596-020-00409-w>.
46. M. Wainberg, N. Sinnott-Armstrong, N. Mancuso, et al., "Opportunities and Challenges for Transcriptome-Wide Association Studies," *Nature Genetics* 51, no. 4 (2019): 592–599, <https://doi.org/10.1038/s41588-019-0385-z>.
47. R. Aldisi, E. Hassanin, S. Sivalingam, et al., "GenRisk: A Tool for Comprehensive Genetic Risk Modeling," *Bioinformatics* 38, no. 9 (2022): 2651–2653, <https://doi.org/10.1093/bioinformatics/btac152>.
48. R. Aldisi, E. Hassanin, S. Sivalingam, et al., "Gene-Based Burden Scores Identify Rare Variant Associations for 28 Blood Biomarkers," *BMC Genomic Data* 24, no. 1 (2023): 1–11, <https://doi.org/10.1186/s12863-023-01155-0>.
49. B. D. Bitarello and I. Mathieson, "Polygenic Scores for Height in Admixed Populations," *G3: Genes, Genomes, Genetics* 10, no. 11 (2020): 4027–4036, <https://doi.org/10.1534/g3.120.401658>.
50. A. I. Young, F. L. Wauthier, and P. Donnelly, "Identifying Loci Affecting Trait Variability and Detecting Interactions in Genome-Wide Association Studies," *Nature Genetics* 50, no. 11 (2018): 1608–1614, <https://doi.org/10.1038/s41588-018-0225-6>.
51. K. E. Westerman, T. D. Majarian, F. Giulianini, et al., "Variance-Quantitative Trait Loci Enable Systematic Discovery of Gene-Environment Interactions for Cardiometabolic Serum Biomarkers," *Nature Communications* 13, no. 1 (2022): 1–11, <https://doi.org/10.1038/s41467-022-31625-5>.
52. A. Mayr, N. Fenske, B. Hofner, T. Kneib, and M. Schmid, "Generalized Additive Models for Location, Scale and Shape for High Dimensional Data—A Flexible Approach Based on Boosting," *Journal of the Royal Statistical Society: Series C: Applied Statistics* 61, no. 3 (2012): 403–427, <https://doi.org/10.1111/j.1467-9876.2011.01033.x>.
53. R. A. Rigby and D. M. Stasinopoulos, "Generalized Additive Models for Location, Scale and Shape," *Journal of the Royal Statistical Society: Series C: Applied Statistics* 54, no. 3 (2005): 507–554, <https://doi.org/10.1111/j.1467-9876.2005.00510.x>.
54. T. Hothorn, T. Kneib, and P. Bühlmann, "Conditional Transformation Models," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 76, no. 1 (2013): 3–27, <https://doi.org/10.1111/rssb.12017>.
55. A. Strömer, N. Klein, C. Staerk, H. Klinkhammer, and A. Mayr, "Boosting Multivariate Structured Additive Distributional Regression Models," *Statistics in Medicine* 42, no. 11 (2023): 1779–1801, <https://doi.org/10.1002/sim.9699>.
56. H. O. Heyne, J. Karjalainen, K. J. Karczewski, et al., "Mono- and Biallelic Variant Effects on Disease at Biobank Scale," *Nature* 613, no. 7944 (2023): 519–525, <https://doi.org/10.1038/s41586-022-05420-7>.
57. J. Werme, v. d S. Sluis, D. Posthuma, and d C A. Leeuw, "Genome-Wide Gene-Environment Interactions in Neuroticism: An Exploratory Study Across 25 Environments," *Translational Psychiatry* 11, no. 1 (2021): 180, <https://doi.org/10.1038/s41398-021-01288-9>.
58. J. Miao, Y. Lin, Y. Wu, et al., "A Quantile Integral Linear Model to Quantify Genetic Effects on Phenotypic Variability," *Proceedings of the*

National Academy of Sciences 119, no. 39 (2022): 1–12, <https://doi.org/10.1073/pnas.2212959119>.

59. B. Sheppard, N. Rappoport, P. R. Loh, S. J. Sanders, N. Zaitlen, and A. Dahl, “A Model and Test for Coordinated Polygenic Epistasis in Complex Traits,” *Proceedings of the National Academy of Sciences* 118, no. 15 (2021): 1–11, <https://doi.org/10.1073/pnas.1922305118>.

60. H. Uno, T. Cai, M. J. Pencina, R. B. D’Agostino, and L. J. Wei, “On the C-Statistics for Evaluating Overall Adequacy of Risk Prediction Procedures With Censored Survival Data,” *Statistics in Medicine* 30, no. 10 (2011): 1105–1117.

61. M. Schmid and T. Hothorn, “Boosting Additive Models Using Component-Wise P-Splines,” *Computational Statistics and Data Analysis* 53, no. 2 (2008): 298–311.

62. N. Fenske, L. Fahrmeier, T. Hothorn, P. Rzehak, and M. Höhle, “Boosting Structured Additive Quantile Regression for Longitudinal Childhood Obesity Data,” *International Journal of Biostatistics* 9, no. 1 (2013): 1–18.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Appendix A

A Further Time-to-Event Loss Functions

Besides the loss functions presented in the main manuscript, we implemented two additional loss functions for time-to-event data in the snpboost framework, corresponding to a Cox proportional hazard model and optimizing the C-index. As they are currently limited to cohorts of small to medium sample sizes, the loss functions and their application on data from the UKBB are presented here.

A.1. | Cox Proportional Hazard Model

In biostatistical practice, the most commonly used approach to model time-to-event data is the Cox proportional hazard model [22]. The hazard function $h(t, \eta_i)$ is assumed to be of the form

$$h(t, \eta_i) = h_0(t) \exp(\eta_i) \quad (\text{A1})$$

where $h_0(t)$ is a baseline hazard that is treated as nuisance. To apply boosting for fitting Cox models, the loss function is then specified as the negative Cox partial log-likelihood defined as

$$\rho_{\text{Cox}}(\mathbf{y}, \boldsymbol{\eta}) = \sum_{i=1}^n \left(\eta_i - \log \left(\sum_{j=1}^n \mathbb{I}(T_j \geq T_i) \exp(\eta_j) \right) \right) \quad (\text{A2})$$

While in low-dimensional settings Cox regression can be considered as the classical approach for time-to-event analysis, in high-dimensional (large p) and large-scale (large n) settings its applicability is limited. Within snpboost, building batches enables us to deal with large numbers of variants p and the general concept allows us to apply it also for large sample sizes n . However, the loss function (A2) of the Cox model requires ordering and summing over all individuals. For very large sample sizes n as in cohort studies like the UK Biobank with hundreds of thousands of individuals, those operations are computationally extremely demanding.

A.2. | Boosting the C-index

A well-known measure to evaluate time-to-event models is the C-index. However, the C-index is usually not optimized by commonly used regression methods such as Cox regression or AFT models. To directly optimize combinations of biomarkers for time-to-event data with respect to the C-index, statistical boosting has been proposed [24, 25]. The approach incorporates an estimator for the C-index on censored survival data which had been proposed by [60]:

$$\hat{C}_{\text{Uno}} = \frac{\sum_{j,k} \left(\hat{G}(\bar{T}_j) \right)^{-2} \mathbb{I}(\bar{T}_j < \bar{T}_k) \mathbb{I}(\eta_j > \eta_k) \Delta_j}{\sum_{j,k} \left(\hat{G}(\bar{T}_j) \right)^{-2} \mathbb{I}(\bar{T}_j < \bar{T}_k) \Delta_j} \quad (\text{A3})$$

Here, $\hat{G}(\cdot)$ is again the Kaplan–Meier estimator for the probability of censoring, that is, the same IPC weights w_i as in the weighted L_2 boosting are used. As our aim is to optimize $\hat{C}_{\text{Uno}}$, it appears to be straightforward to choose the negative C-index as a loss function. However, to apply statistical boosting, the gradient of the loss needs to be computed, that is, the loss function has to be differentiable in $\boldsymbol{\eta}$. Therefore, we replace the indicator function $\mathbb{I}(u > 0)$ by a sigmoid function $K(u) = (1 + \exp(-u/\sigma))^{-1}$, with $\sigma > 0$ controlling the smoothness. The smooth loss is then defined as

$$\begin{aligned} f_C(\mathbf{y}, \boldsymbol{\eta}) &= -\hat{C}_{\text{smooth}}(\mathbf{T}, \boldsymbol{\eta}) \\ &= -\frac{\sum_{j,k} w_j^2 \mathbb{I}(\bar{T}_j < \bar{T}_k) (1 + \exp(-(\eta_j - \eta_k)/\sigma))^{-1}}{\sum_{j,k} w_j^2 \mathbb{I}(\bar{T}_j < \bar{T}_k)} \end{aligned} \quad (\text{A4})$$

The smoothness parameter σ can be set to a fixed small value, for example, $\sigma = 0.1$, and does not have a large impact on the performance of the approach [24].

Comparable to the Cox proportional hazard model, boosting the C-index requires ordering and summing over all n observations and predictions in each boosting step. Again, for very large sample sizes n in cohort studies, this is computationally demanding and therefore the use of the negative C-index as a loss function is limited in practice to small to medium sample sizes (with a few thousand observations).

A.3. | Application on a reduced subset of the UK Biobank

Due to the limitations of the Cox and C-index models, we created reduced subsets of the asthma and gout datasets described in Section 4.1. To do so, we randomly sampled $n = 50000$ individuals from the complete data sets, again divided into 50% training, 20% validation, and 30% test sets. We then applied all six proposed loss functions via snpboost on the training and validation set and evaluated the C-index on the independent test set. Additionally, we derived a PRS for the occurrence of the diseases by using the log loss as a loss function [7]. When using the C-index loss function, we found that a larger step length reduced the computational time and also improved the performance of the algorithm. Therefore, we used an increased step length of $\nu = 10$ for this particular case. The association between the necessary number of iterations and the step length is well-known in the literature [61], as well as the possibility to increase the fixed step length for some particular loss functions [62] or even letting the algorithm converge in case of the concordance index [24, 25].

The results are shown in Figure A1. Compared to Figure 3, we observe a general decline of prediction performance, most likely due to the reduced sample size. The varying results suggest that the “optimal” loss function highly depends on the phenotype and the available data, for example, for gout, the C-index model was able to detect more genetic signals than the AFT models, while it was the opposite for asthma. The runtimes for the different models can be found in Table A1.

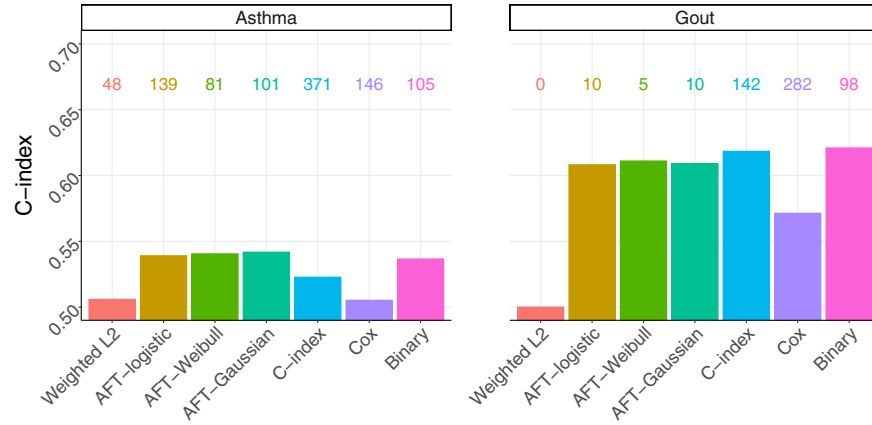


FIGURE A1 | C-indices on the independent test set of the reduced subsamples for asthma and gout for the PRS models estimated using different loss functions. The numbers above the bars indicate the number of SNPs with non-zero effect in each model.

TABLE A1 | Runtime in hours on the reduced subsample ($n = 50\,000$) for the age of onset of asthma and gout using seven different loss functions.

Trait	PRS	Runtime (h)
Asthma	Binary	0.85
Asthma	Weighted L_2	0.12
Asthma	AFT-Weibull	0.17
Asthma	AFT-logistic	0.17
Asthma	AFT-Gaussian	0.18
Asthma	C-index	181.63
Asthma	Cox	47.45
Gout	Binary	0.15
Gout	Weighted L_2	0.14
Gout	AFT-Weibull	0.13
Gout	AFT-logistic	0.13
Gout	AFT-Gaussian	0.14
Gout	C-index	161.74
Gout	Cox	49.25

Note: All calculations were run on an HPC cluster with 30 CPUs and 2 GB RAM each.

Appendix B Additional Results

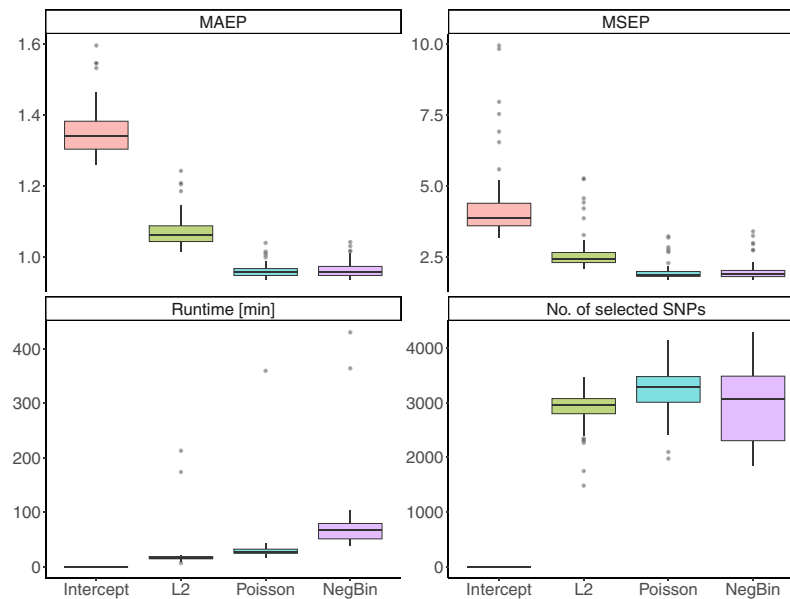


FIGURE B1 | Results of 100 simulated phenotypes following a Poisson distribution. PRS were fitted using L_2 , Poisson and negative binomial (Neg-Bin) loss functions, respectively. Mean absolute error of prediction (MAEP), mean squared error of prediction (MSEP), runtime in minutes and number of selected variants in the PRS were evaluated. Simulations were run using 30 CPUs with 2 GB RAM each.

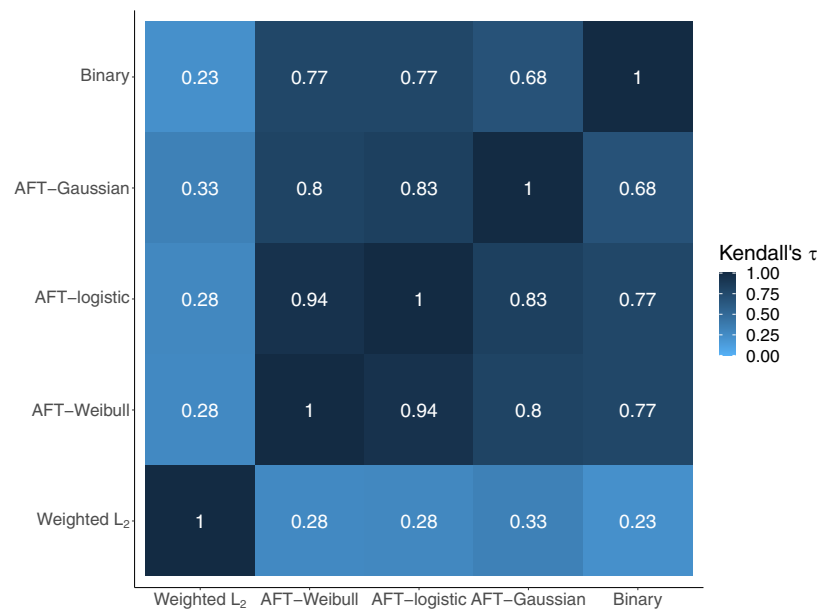


FIGURE B2 | Kendall's τ between different PRS models for age of onset of asthma evaluated on the independent test set.

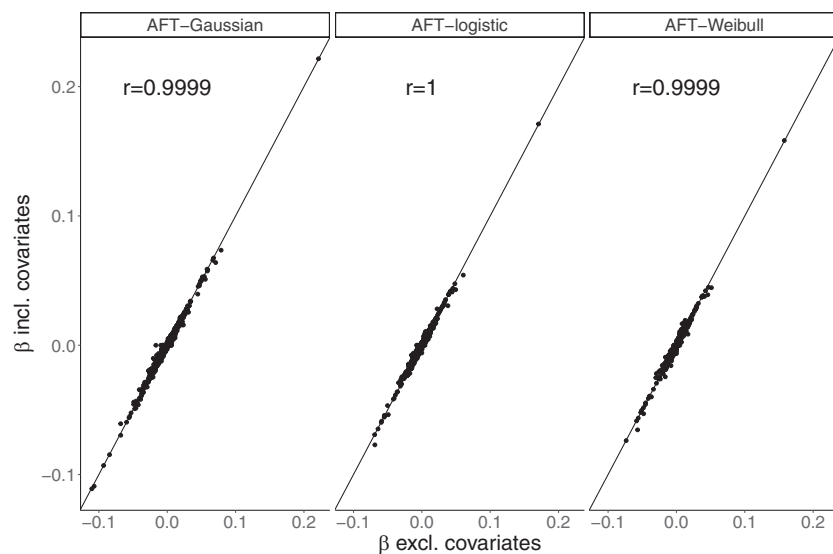


FIGURE B3 | Estimated SNP effects for age of onset of asthma for three AFT models with covariate correction (sex, first ten PCs, y-axis) and without (x-axis).

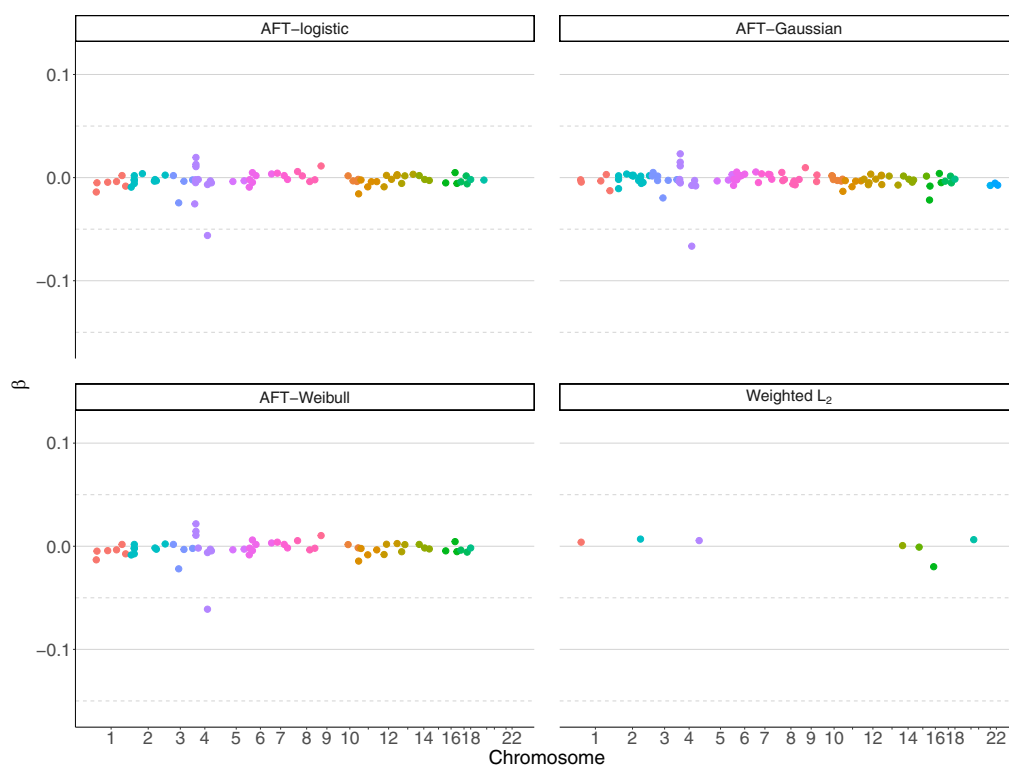


FIGURE B4 | Estimated coefficients for age of onset of gout using different loss functions. Only non-zero effects are shown.

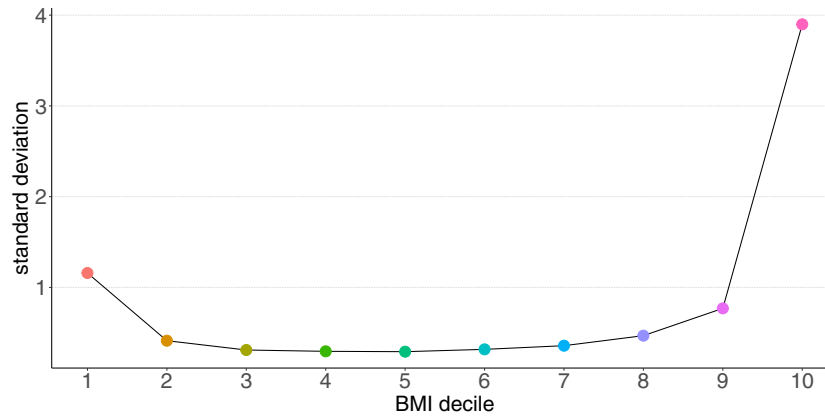


FIGURE B5 | Standard deviation of BMI values within deciles of the original BMI values. The standard deviation is increasing from the middle deciles to the outer deciles, respectively.

3.4 Publication 4: Generalizability of polygenic prediction models: how is the R^2 defined on test data?

Staerk C, Klinkhammer H, Wistuba T, Maj C, Mayr A. Generalizability of polygenic prediction models: how is the R^2 defined on test data? In: BMC Medical Genomics 2024; 17: 132
<https://doi.org/10.1186/s12920-024-01905-8>

Supplementary information can be found at:

<https://doi.org/10.1186/s12920-024-01905-8>

RESEARCH

Open Access

Generalizability of polygenic prediction models: how is the R^2 defined on test data?



Christian Staerk^{1,2*}, Hannah Klinkhammer^{1,3}, Tobias Wistuba¹, Carlo Maj⁴ and Andreas Mayr¹

Abstract

Background Polygenic risk scores (PRS) quantify an individual's genetic predisposition for different traits and are expected to play an increasingly important role in personalized medicine. A crucial challenge in clinical practice is the generalizability and transferability of PRS models to populations with different ancestries. When assessing the generalizability of PRS models for continuous traits, the R^2 is a commonly used measure to evaluate prediction accuracy. While the R^2 is a well-defined goodness-of-fit measure for statistical linear models, there exist different definitions for its application on test data, which complicates interpretation and comparison of results.

Methods Based on large-scale genotype data from the UK Biobank, we compare three definitions of the R^2 on test data for evaluating the generalizability of PRS models to different populations. Polygenic models for several phenotypes, including height, BMI and lipoprotein A, are derived based on training data with European ancestry using state-of-the-art regression methods and are evaluated on various test populations with different ancestries.

Results Our analysis shows that the choice of the R^2 definition can lead to considerably different results on test data, making the comparison of R^2 values from the literature problematic. While the definition as the squared correlation between predicted and observed phenotypes solely addresses the discriminative performance and always yields values between 0 and 1, definitions of the R^2 based on the mean squared prediction error (MSPE) with reference to intercept-only models assess both discrimination and calibration. These MSPE-based definitions can yield negative values indicating miscalibrated predictions for out-of-target populations. We argue that the choice of the most appropriate definition depends on the aim of PRS analysis — whether it primarily serves for risk stratification or also for individual phenotype prediction. Moreover, both correlation-based and MSPE-based definitions of R^2 can provide valuable complementary information.

Conclusions Awareness of the different definitions of the R^2 on test data is necessary to facilitate the reporting and interpretation of results on PRS generalizability. It is recommended to explicitly state which definition was used when reporting R^2 values on test data. Further research is warranted to develop and evaluate well-calibrated polygenic models for diverse populations.

Keywords Polygenic risk scores, Prediction, Coefficient of determination, R squared, Calibration

*Correspondence:

Christian Staerk
christian.staerk@uni-bonn.de

¹ Department of Medical Biometry, Informatics and Epidemiology,
Medical Faculty, University of Bonn, Bonn, Germany

² Institute of Statistics, RWTH Aachen University, Aachen, Germany

³ Institute for Genomic Statistics and Bioinformatics, Medical Faculty,
University of Bonn, Bonn, Germany

⁴ Center for Human Genetics, University of Marburg, Marburg, Germany



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Genetics plays an increasingly important role in predicting disease risk and treatment responses. The development of polygenic risk scores (PRS) and general genetic prediction models for the individual liability to a particular clinical trait or phenotype comes with the promise to improve targeted early prevention and individual treatment [1–3]. However, an important limiting factor in current clinical practice is the issue of transferability of a developed PRS model to out-of-target populations [4–7]. This especially concerns the application of PRS in populations that have been underrepresented in the cohorts used to estimate and train the models, such as those with genetic ancestry different from the majority of participants in current cohorts. As many current PRS models are predominantly trained and optimized on individuals with European ancestry, this disparity also raises important ethical issues [8].

Amid these challenges, PRS work by quantifying the genetic predisposition for a particular trait and are typically computed using weighted sums of risk alleles, where the weights are based on effect sizes that represent associations between variants and the target phenotype [9]. The effect sizes of individual variants (single nucleotide polymorphisms, SNPs) are often small for common traits characterized by a polygenic architecture. As a consequence of differences in allele frequencies and linkage disequilibrium patterns (i.e., variant correlations) across populations, absolute PRS values may be influenced more by population structure than by genuine genetic risk for the phenotype [10]. Such issues should be taken into account when evaluating the transferability of PRS models to out-of-target populations. This is particularly important when the ultimate aim is the development of individual prediction models for personalized medicine, integrating the PRS with further demographic and environmental variables.

The evaluation of polygenic prediction models on test data plays an essential role to examine the generalizability and transferability to different populations. However, even though it may sound straightforward, there is no general consensus on the choice of the evaluation metric [11–14]. The methodological distinction between discrimination and calibration is a well-known issue in the context of prediction models for binary outcomes [15, 16]. While the AUC is the most widely used measure for evaluating the prediction accuracy of polygenic models for binary outcomes, it solely reflects discriminative accuracy. To assess also calibration, additional measures such as the Brier score and graphical tools such as calibration plots are recommended [17, 18]. When we focus on the prediction of a continuous trait, it seems obvious that the predicted values should be as close as possible to the true (observed) values.

But how should this closeness be measured, particularly when there is a test sample of n_{test} observations? Researchers working in statistics might answer with the mean squared prediction error (MSPE) or its root (RMSPE), which take both bias and variance of predictions into account. If there are outliers in the data, other researchers might want to focus on the more robust median absolute error (MAE), while researchers working with the actual predictions might argue that both MSPE and MAE do not provide results that are easily interpretable or comparable across different outcomes and might focus on percentage errors (e.g., the median absolute percentage error).

In the context of polygenic risk modelling (but not only there [14]), a popular measure of prediction accuracy for continuous traits is the R^2 [19–22]. The R^2 in statistics is typically a measure for the goodness-of-fit of a model on the training data and is also called the coefficient of determination. In the context of linear models, it is a well-defined measure on training data and can be interpreted as the proportion of variability of the outcome explained by the model. However, in case of using the R^2 for predictions on test data the situation is generally less clear and the R^2 measure is not unambiguously defined. In particular, one potential definition of the R^2 is simply based on the squared correlation between predicted and observed values, which is always bounded between 0 and 1. This definition makes the R^2 relatively easy to interpret and to compare between different phenotypes (similar to the AUC for binary outcomes). Two alternative definitions of the R^2 are closer to the original definition on training data (cf., [23]), comparing the MSPE of the prediction model in relation to an intercept model (one based on training data, one based on test data). In contrast to the squared correlation which only focuses on the discriminative performance, the MSPE-based definitions of the R^2 take both discrimination and calibration of the prediction model into account.

In this study, we investigate the impact of these three different R^2 definitions when analysing the accuracy of PRS models on test data and illustrate their differences in the context of the generalizability to different populations. Our UK Biobank data analysis shows that phenotype predictions of standard PRS models on out-of-target populations are often considerably biased, which can even lead to negative R^2 values based on the two MSPE-based definitions, strongly indicating miscalibrated predictions. On the other hand, the (squared) correlation with the phenotype may partially be preserved in some cases even when the PRS is systematically biased and poorly calibrated, as it only addresses the discriminative performance. As a consequence, resulting R^2 values from the different definitions lead to considerably different results — not only regarding their absolute values,

but also when using the R^2 to rank different PRS models developed by competing methods. We highlight the strengths and limitations of the different R^2 definitions and discuss recommendations regarding their reporting.

Methods

Different R^2 definitions on test data

The coefficient of determination R^2 was originally introduced as a goodness-of-fit measure for linear models [24] and can be calculated in various but equivalent ways due to the variance decomposition [25]. When applied to training data in linear models, the R^2 can be interpreted as the proportion of variance that is explained by the model. However, the variance decomposition does not generally hold true on test data or for non-linear models, leading to different results depending on the choice of R^2 formula. While the different formulas for the R^2 on training can lead to various ways to define the R^2 on test data [25], here we focus on the three most common definitions.

The first definition is given by the squared correlation r^2 between the observed phenotypes $y = (y_1, \dots, y_{n_{\text{test}}})$ and the predictions $\hat{y} = (\hat{y}_1, \dots, \hat{y}_{n_{\text{test}}})$, i.e.,

$$\begin{aligned} r^2 &= \text{cor}(y, \hat{y})^2 \\ &= \frac{\left(\sum_{i=1}^{n_{\text{test}}} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}) \right)^2}{\sum_{i=1}^{n_{\text{test}}} (y_i - \bar{y})^2 \sum_{i=1}^{n_{\text{test}}} (\hat{y}_i - \bar{\hat{y}})^2}, \end{aligned} \quad (1)$$

where n_{test} is the number of individuals in the test set, and $\bar{y}$ and $\bar{\hat{y}}$ are the means of the observed and predicted phenotypes, respectively. Noteworthy, r^2 based on Eq. (1) is always nonnegative and bounded between 0 and 1. This is the same as fitting a simple linear model with intercept on the test data with the PRS model predictions as the only covariate and computing the classical coefficient of determination for this model. The r^2 value tends to be high in the presence of a strong linear correlation, regardless of whether the predictions are strongly biased and poorly calibrated. In fact, the association could even be negative. Therefore, when considering the squared correlation r^2 , it is recommended to additionally examine scatterplots and the correlation coefficient r itself.

Two alternative definitions of the R^2 on test data, which are closer to the original model-comparison idea of R^2 (cf., [23, 26]), are based on relative comparisons of the out-of-sample residual sum of squares (or, equivalently, the MSPE) for the prediction model in relation to a reference model. In contrast to the squared correlation r^2 , which only evaluates the discriminative performance of the model, the two MSPE-based definitions of the R^2 consider both discrimination and calibration of

the prediction model. In case (A), the intercept model on the training data is used as a reference, leading to the definition

$$R^2_{1-A} = 1 - \frac{\sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{\text{test}}} (y_i - \bar{y}_{\text{train}})^2}, \quad (2)$$

where $\bar{y}_{\text{train}}$ is the mean of the phenotype in the training data (representing the intercept model).

In case (B), the intercept model on the test data is used as a reference, leading to the definition

$$R^2_{1-B} = 1 - \frac{\sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{\text{test}}} (y_i - \bar{y}_{\text{test}})^2}, \quad (3)$$

where $\bar{y}_{\text{test}}$ is the mean of the phenotype in the test data (representing the intercept model on test data).

Both MSPE-based definitions, R^2_{1-A} and R^2_{1-B} , have an upper bound of 1 with equality in case of perfect predictions on test data, i.e., $R^2_{1-A} = R^2_{1-B} = 1$ if and only if

$$\text{MSPE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2 = 0.$$

On the other hand, R^2_{1-A} and R^2_{1-B} are generally not lower bounded and can even be negative, particularly for prediction models that are miscalibrated on the test data (see Results section). If R^2_{1-A} is negative, this implies that a simple intercept-only model on the training data yields a smaller MSPE on the test data than the PRS model. In a similar vein, if R^2_{1-B} is negative, this means that a simple intercept model on the test data performs better than the PRS model.

Note that in situations where the size of the test data is small, considering the mean on the test data leads to increased variability of the R^2_{1-B} measure, and the R^2_{1-A} definition may generally be preferable based on theoretical reasons [23]. Yet, while r^2 and R^2_{1-B} only depend on the predicted and observed phenotypes on the test data, R^2_{1-A} also explicitly depends on the training data, implying that R^2_{1-A} cannot be readily computed when the phenotype mean on the training data is not available.

Data

All analyses were performed on data from the UK Biobank (UKBB) under application number 81202. The UKBB [27] is a large-scale prospective cohort study of approximately 500,000 participants from the United Kingdom, established in 2006. Participants were aged between 40 and 69 years at recruitment. Resources include individual-level genotype data and detailed information on a wide range of phenotypes for each

participant, such as physical measurements as well as blood biomarkers. Baseline data are continuously updated through follow-up studies.

We extracted data for the phenotypes height (UKBB field 50), BMI (UKBB field 21001) and lipoprotein A (UKBB field 30790), as well as for the covariates sex (UKBB field 22001), age (UKBB field 21022) and the first ten principal components (PCs, UKBB field 22009) of the genotype matrix. As performed in other works, individual genetic ancestry was derived considering the closest reference population center (geometric medians) according to the Euclidean distance in the global PC space based on 1K genome data [28]. Incomplete data, related individuals as well as individuals who could not be assigned to European, African, East Asian and South Asian population clusters, were discarded, resulting in $n = 333,587$ individuals comprising $n = 319,456$ of European ancestry, $n = 5,411$ of African ancestry, $n = 6,796$ of South Asian ancestry and $n = 1,924$ of East Asian ancestry. The European subcohort was further split into training ($n = 159,554$), validation ($n = 80,113$) and test set ($n = 79,799$). Genotype data (UKBB category 100315) were filtered for a genotype rate of at least 90% and a minor allele frequency of at least 0.1%, resulting in $p = 549,426$ SNPs.

Computation of PRS

PRS models for each phenotype were derived based on the training and validation sets of the European subcohort via three competing estimation approaches: PRSs [29], snpboost [30] and BayesR [31]. While PRSs applies Bayesian shrinkage on univariate effect estimates from GWAS summary statistics, snpboost and BayesR estimate PRS directly based on individual-level genotype data, using statistical boosting and Bayesian hierarchical modelling, respectively.

To apply PRSs, first a GWAS was conducted in plink2 [32, 33] on the European training set. Afterwards, the European validation set as well as the LD reference based on UKBB data (provided at <https://github.com/getian107/PRSs>) were used to run PRSs and derive the final weights. For BayesR, which also applies Bayesian shrinkage on the effect estimates but works on individual-level data, the implementation in LDAK [34] was used. Per-predictor heritabilities were computed assuming the BLD-LDAK-model and incorporating the GWAS results based on the training set. Using those per-predictor heritabilities, BayesR was run on the European training set and the final model was chosen to optimize the performance on the European validation set. The statistical boosting algorithm snpboost was applied on individual-level data. Effect estimates by snpboost were derived

on the European training set incorporating early stopping based on the predictive performance on the validation set to provide sparse models.

Finally, PRS were computed for all individuals using plink's --score function based on the estimated weights from the three different methods. While snpboost directly provides prediction models for the phenotype of interest, PRSs and BayesR only yield scores that must be rescaled to predict the phenotype. For a fair comparison, for all three methods a multivariable regression was fitted on the combined European training and validation data including the PRS, sex, age and the first ten principal components as predictors. The resulting models were then applied on the test data set comprising individuals of European, African, East Asian and South Asian ancestry.

Additionally, we conducted two sensitivity analyses regarding different variant filtering and adjustment for population stratification. In a first sensitivity analysis, we used more stringent variant filters, i.e., a minor allele frequency filter of 1% and a genotype call rate of at least 99%. In a second sensitivity analysis, we included the first 10 genetic PCs directly in the GWAS and in the training of snpboost.

Results

Figure 1 provides results for the three different definitions of the R^2 measure on European, African, South Asian and East Asian ancestry test populations for the prediction of height, BMI and lipoprotein A on UK Biobank data. Results are based on PRS models derived by PRSs [29], snpboost [30] and BayesR [31] on the European training population (including the covariates sex, age and the first 10 genetic PCs). Sensitivity analyses regarding different variant filtering and adjustment for population stratification lead to overall comparable results and main conclusions (see Supplementary Figure 1 and Supplementary Figure 2).

The three considered phenotypes, height, BMI and lipoprotein A, exhibit a significant genetic component, but present distinct underlying genetic architectures. As such, they are well-suited for evaluating diverse scenarios, specifically regarding different R^2 definitions and competing PRS models with varying degrees of sparsity. In particular, height and BMI are highly polygenic traits [35]; however, while the environmental influence remains minimal for height [36], a substantial portion of the genetic signal in BMI can be attributed to gene-environment interactions [37]. As we do not consider environmental variables in this work, the predictive performance of the PRS models are particularly limited for BMI, with relatively low R^2 values observed for the different models and definitions even on the European test

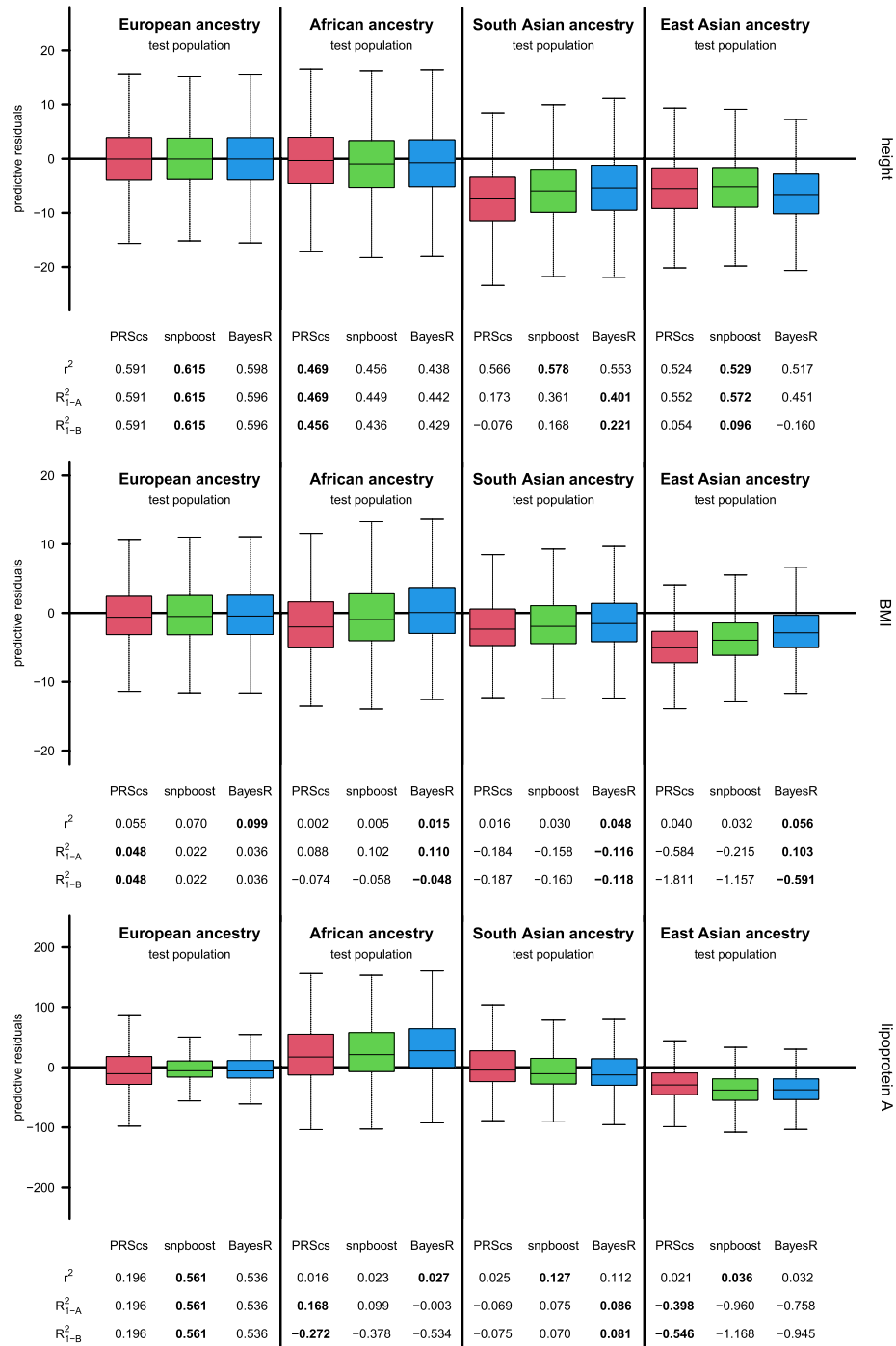


Fig. 1 Boxplots of predictive model residuals for the prediction of height (top), BMI (middle) and lipoprotein A (bottom) on test populations with different ancestries, together with R^2 values based on the three different definitions in Eqs. (1), (2) and (3). Different polygenic prediction models (including sex, age and first 10 genetic PCs) were derived on training data with European ancestry using PRScs, snpboost and BayesR, respectively. Outliers are not shown in the boxplots

population (see Fig. 1). Conversely, lipoprotein A blood levels are predominantly influenced by an oligogenic component, with the majority of the signal originating from variants located within the LPA gene locus [38]. The predictive performance of the competing methods shown in Fig. 1 clearly reflect the sparse genetic architecture of lipoprotein A, as the sparse PRS models derived by snpboost and BayesR yield considerably improved performance compared with the non-sparse model derived by PRSs [29, 30].

On the European test population, the predictive residuals are approximately centered around zero for all PRS methods (Fig. 1), with symmetric residual distributions for height, while for BMI and lipoprotein A the median residuals are negative, reflecting the right-skewed distributions of these phenotypes. On the European test population, the three different definitions of R^2 lead to quite similar values, especially for height and lipoprotein A, where the models derived by snpboost consistently show the best performance across all R^2 definitions. However, already on the European test population, the three definitions of R^2 are not generally equivalent on test data, as can be seen for the case of BMI, where PRSs performs best according to the MSPE-based definitions R^2_{1-A} and R^2_{1-B} (PRSs: $R^2_{1-A} = 0.048$, BayesR: $R^2_{1-A} = 0.036$, snpboost: $R^2_{1-A} = 0.022$), while BayesR performs best according to the squared correlation r^2 between predicted and observed BMI (BayesR: $r^2 = 0.099$, snpboost: $r^2 = 0.070$, PRSs: $r^2 = 0.055$). Note that the rankings of different prediction models developed on the same training data are always the same based on the maximization of R^2_{1-A} or R^2_{1-B} as well as based on the minimization of the MSPE, as both R^2_{1-A} and R^2_{1-B} can just be viewed as different ways of “rescaling” the MSPE.

Figure 1 further reveals that, on test populations with African, South Asian and East Asian ancestry, the distributions of predictive residuals tend to be shifted and not centered around zero, indicating miscalibration of the PRS models. In such situations, the three different definitions of R^2 on test data can lead to considerably different results. For example, for height in the South Asian ancestry population, the distributions are shifted towards negative residuals, reflecting the overestimation on the South Asian population based on the PRS models that were developed on the European training population. Considering the R^2 definition based on the squared correlation between predicted and observed heights still yields relatively large values for the different PRS models on the South Asian population (between $r^2 = 0.578$ for snpboost and $r^2 = 0.553$ for BayesR), while the MSPE-based definitions R^2_{1-A} and R^2_{1-B} yield considerably lower and potentially even negative values (between $R^2_{1-A} = 0.401$, $R^2_{1-B} = 0.221$ for BayesR and

$R^2_{1-A} = 0.173$, $R^2_{1-B} = -0.076$ for PRSs). Therefore, for the prediction of height on the South Asian test population, snpboost performs best according to the definition based on the squared correlation r^2 , while BayesR performs best according to the MSPE-based definitions R^2_{1-A} and R^2_{1-B} .

Figure 2 illustrates the underlying reasons for the conflicting results of the three different definitions of the R^2 measure for the prediction of height, focusing on the PRS model derived by PRSs. The two scatterplots (first row in Fig. 1) indicate relatively strong correlations between the predicted heights from the PRS model and the observed (true) heights on both the European and South Asian test populations. In line with this, the prediction performance according to the R^2 definition based on the squared correlation between predicted and observed heights appears to be only slightly worse on the South Asian ancestry population ($r^2 = 0.566$) compared to the European population ($r^2 = 0.591$). However, while for the European test population the regression line between predicted and observed heights lies on the angle bisector, a clear parallel shift of the regression line is apparent for the test population of South Asian ancestry, indicating a systematic overestimation of height by the PRS model on this population.

The miscalibration of the predictions on the South Asian test population becomes even clearer when inspecting kernel density plots for the distributions of predictive residuals (observed height – predicted height), see second row in Fig. 2. While for the European test population the predictive residuals of the PRS model are centered around zero, the distribution of predictive residuals is shifted towards negative residuals (i.e., smaller observed heights than predicted by the PRS model) for the South Asian test population. In contrast to the definition based on the squared correlation r^2 , the two MSPE-based definitions R^2_{1-A} and R^2_{1-B} are sensitive to the miscalibration of the PRS model on the South Asian population. In particular, while on the European test population the two MSPE-based definitions of R^2 yield similar values to the squared correlation, both MSPE-based definitions yield considerably lower values on the population with South Asian ancestry ($R^2_{1-A} = 0.173$, $R^2_{1-B} = -0.076$ and $r^2 = 0.566$).

Recall that the definition R^2_{1-A} relates the MSPE of the PRS model to the MSPE of the intercept model derived on the training data (see Eq. 2), while the definition R^2_{1-B} relates the MSPE of the PRS model to the variance of observed heights on the test data (see Eq. 3). Therefore, as the distribution of the PRS model residuals on the South Asian population is slightly “closer” (on average in terms of squared distance) to zero compared to the distribution of the predictive residuals from the intercept

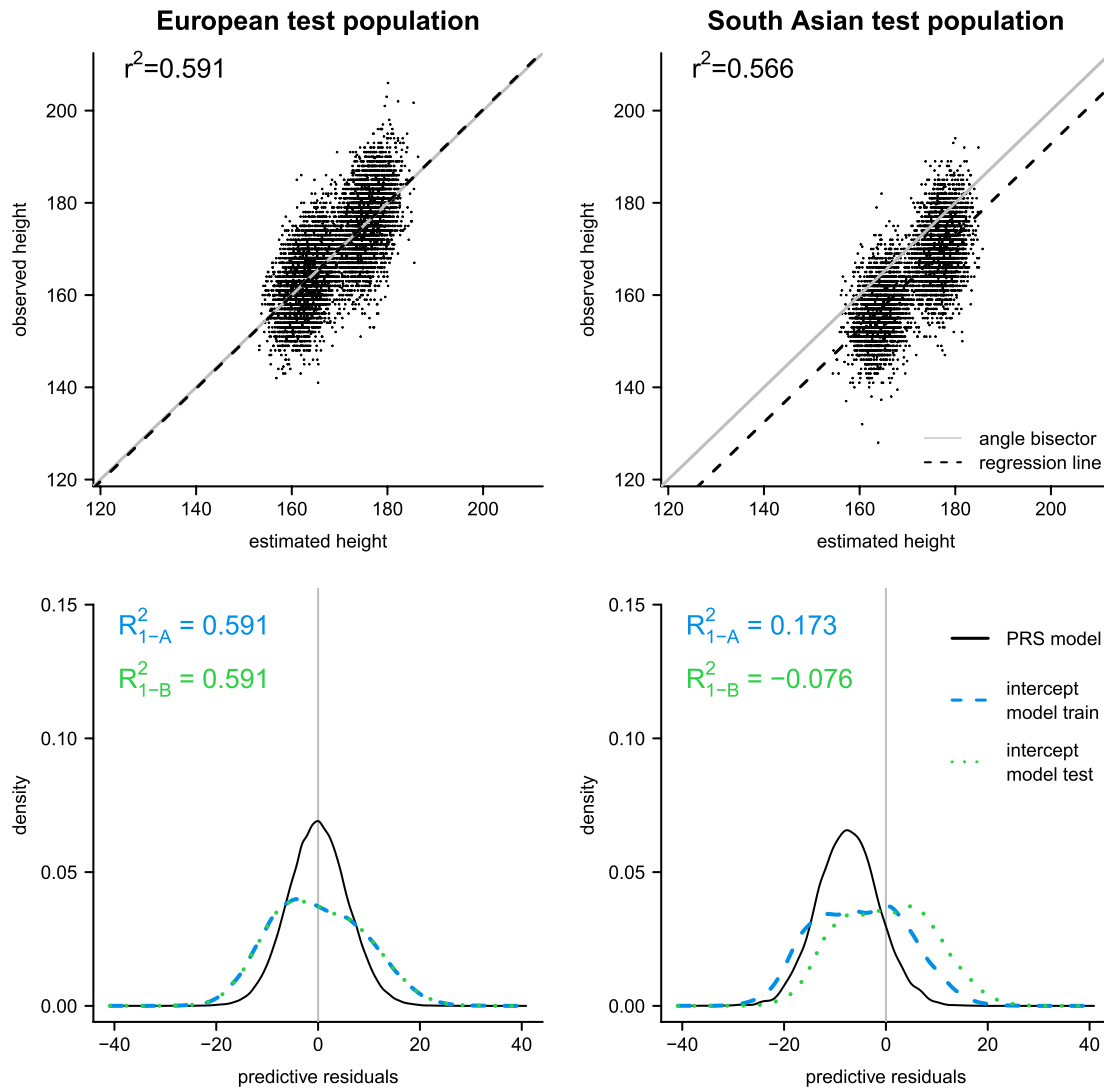


Fig. 2 Illustration of the three different R^2 definitions on test populations with European ancestry (left) and South Asian ancestry (right) for the prediction of height, based on PRSs model derived on training data with European ancestry (including sex, age and first 10 genetic PCs). The scatterplots (top) show the association between predicted and observed heights, illustrating the R^2 definition based on the squared correlation (r^2). The kernel density plots (bottom) show the distributions of predictive residuals, illustrating the MSPE-based definitions of the R^2 with reference to intercept models on training data (R^2_{1-A}) and test data (R^2_{1-B}), respectively

model derived on the training data (shown in blue), the value of $R^2_{1-A} = 0.173$ is still positive. On the other hand, when compared to the intercept model on the South Asian test population (i.e., corresponding to the variance of observed heights, shown in green), the PRS model predictive residuals show a larger variability around zero, so that the value of $R^2_{1-B} = -0.076$ is negative. This example illustrates that the MSPE-based definitions can be positive

or negative on test data depending on the performance of the prediction model relative to the intercept model on training data (R^2_{1-A}) or test data (R^2_{1-B}), whereas the definition based on the squared correlation (r^2) is guaranteed to always yield nonnegative values. Additional scatterplots and distributions of predictive residuals (as in Fig. 2) for the test populations of African and East Asian ancestries are shown in the Supplementary Material (see

Supplementary Figures 3 and 4, respectively). For the East Asian population, it is apparent that a larger difference between phenotype means in the training and test data can lead to considerably different R^2_{1-A} and R^2_{1-B} values, whereas the distributions of the predictive residuals derived from the intercept models almost coincide for the population of African ancestry leading to similar R^2_{1-A} and R^2_{1-B} values.

Figure 1 shows substantial differences in the three definitions of R^2 on test data not only for height but also for the other phenotypes on out-of-target populations. While for height the squared correlations r^2 between predicted and observed phenotypes still remain relatively large (cf., Fig. 2), for lipoprotein A the prediction accuracy, even in terms of r^2 , is considerably reduced for populations with non-European ancestry. This indicates that the loss of prediction accuracy of the PRS models for lipoprotein A on out-of-target populations is not only a matter of miscalibration.

Discussion

Although the R^2 is a widely-used and apparently simple measure, it is not clearly defined for the evaluation of prediction models on test data. In our analysis on the UK Biobank, we have illustrated that there are at least three alternative definitions of the R^2 on test data that lead to considerably different results when used for the evaluation of the generalizability of PRS models to out-of-target populations (Fig. 1). In particular, we have shown that these different definitions can also lead to conflicting rankings of competing prediction methods. Therefore, to ensure reproducibility and fair comparisons between methods, we strongly recommend explicitly stating which definition was used when reporting R^2 values.

Our study does not aim to determine which definition of the R^2 should be considered as the *correct* one on test data, since all of them have their advantages and challenges. Yet, we still want to provide some general advice on which might be the most suitable definition depending on the data situation and modelling aims. If the PRS is aimed to be primarily used as a stratification tool, similar to other biomarkers, then the definition based on the squared correlation r^2 between predicted and observed outcomes might be preferred, as it implicitly assumes a re-calibration when used on another population. On the other hand, if the PRS model should be used as a direct prediction model for continuous phenotypes of individuals, then the squared correlation r^2 cannot capture miscalibration and the MSPE-based definitions R^2_{1-A} and R^2_{1-B} might be more suitable. Unlike the squared correlation r^2 , the MSPE-based definitions can also take on negative values. This property should generally not be viewed as a deficiency of the MSPE-based definitions,

since negative values of R^2_{1-A} or R^2_{1-B} are important indicators of poor calibration, showing that the PRS model on the test population is inferior to a basic intercept model.

In this context, one can argue that the R^2_{1-A} definition, which relies on the arithmetic mean on the training data (or the intercept model on the training population), is generally preferable compared to the R^2_{1-B} definition relying on the mean on the test data, particularly in cases where the size of the test data is small leading to increased variability of the R^2_{1-B} measure [23]. On the other hand, the comparison of different PRS models that might have been derived based on different training cohorts is not feasible with the R^2_{1-A} definition in situations where the means on the different training data are not available. Therefore, in such cases, the R^2_{1-B} definition can be considered a pragmatic alternative choice. Furthermore, note that the R^2_{1-B} definition can also be regarded as “more conservative”: to obtain a positive R^2_{1-B} value, the PRS model needs to provide better predictions than the mean on the population it is tested on.

In practice, it can be valuable to consider both the correlation-based and the MSPE-based definitions as they provide complementary information. In particular, when the squared correlation r^2 is high, but the MSPE-based measures R^2_{1-A} or R^2_{1-B} are low, this indicates that a well-calibrated new prediction model on the targeted test population may be achieved after using an appropriate affine transformation of the original model predictions. On the other hand, when both the correlation-based and MSPE-based measures are low for a PRS model on a particular population, the poor discrimination already indicates that the PRS model may generally not be suitable for the targeted population. As the squared correlation r^2 between predicted and observed phenotypes only assesses the discriminatory accuracy of prediction models, to investigate also the calibration of PRS models for continuous phenotypes, we recommend to use additional measures such as the MSPE or the MSPE-based definitions of the R^2 , as well as graphical tools such as scatterplots as in Fig. 2 (cf., [13]). This is especially important when considering out-of-target populations.

In this study, we evaluated PRS accuracy using R^2 population-level metrics on test data based on individuals with different genetic ancestries, for which we employed genetic principal component clustering based on the reference 1K Genome superpopulations [39]. Considering population clusters is an important limitation, as human genetic diversity extends across a continuum, and even populations deemed “homogeneous” exhibit accuracy variations along subcontinental ancestries, influenced by the genetic distance from training data [40]. Our analysis relying on population clusters hence simplifies

the underlying genetic heterogeneity. Nevertheless, this approach is suitable for comparing R^2 performance among categorical groups, identifying reporting biases that would actually emerge within a continuous spectrum. Our results further emphasize the importance of comprehensive global human genome studies to effectively capture genetic heterogeneity, with the goal of offering more universally applicable genetic risk scores. Another limitation of our study is that we focused on state-of-the-art methods to estimate PRS models on a training population with European ancestry; these methods are not specifically designed to provide robust generalizability to out-of-target populations. While more refined methods might be able to provide improved transferability of the resulting PRS models (see [6] for a comprehensive review of recent methods), the issues we discussed regarding the different definitions of the R^2 on test data are still highly pertinent for the evaluation of such models.

The transferability of prediction models remains one of the most important challenges in the practical application of PRS and its implementation in clinical routines and decision rules. For PRS or other genetic scores to play a pivotal role in future public health and disease prevention programs, it is imperative that individuals with non-European ancestry and multi-ethnic backgrounds benefit equally from these tools, just as individuals with European ancestry. Further research is warranted on broader and more diverse cohorts as well as new methods that help to close the substantial gap in accuracy that can be observed in practice. For this purpose, it is essential to have clear and transparent ways to evaluate new methods for the development of PRS also on other cohorts.

Conclusions

In this article, we have investigated three different definitions of the R^2 measure on test data and illustrated how these behave when evaluating the accuracy of European-derived PRS models that were applied to predict continuous phenotypes on individuals with different ancestry. The choice of the R^2 definition on test data might drastically influence the results, particularly when evaluating the prediction performance on out-of-target populations. If the PRS model is primarily aimed for risk stratification, then defining the R^2 as the squared correlation between predicted and observed phenotypes provides the clearest interpretation. If the PRS model is also targeted for individual predictions, model calibration is crucial, which can be assessed via graphical tools and additional measures, including the MSPE-based definitions of the R^2 . To ensure reproducibility and fair comparisons with results from the literature, it is essential to state which definition was used when reporting R^2 measures on test data.

Abbreviations

AUC	Area under the curve
BMI	Body mass index
GWAS	Genome-wide association study
MAE	Mean absolute error
MSPE	Mean squared prediction error
PC	Principal component
PRS	Polygenic risk score
SNP	Single nucleotide polymorphism
UKBB	UK Biobank

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12920-024-01905-8>.

Supplementary Material 1.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 81202.

Authors' contributions

CS, HK, CM and AM conceived the research idea and contributed to the development of the methodology, the analysis and the interpretation of the results. TW contributed to the analysis, the presentation and the interpretation of the results. CS, HK, CM and AM wrote the initial version of the manuscript. All authors provided critical comments and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The data analyzed in this study is subject to the following licenses/restrictions: This research has been conducted using the UK Biobank resource under application number 81202 (<http://www.ukbiobank.ac.uk>). Requests to access these datasets should be directed to UK Biobank, <http://www.ukbiobank.ac.uk>.

Declarations

Ethics approval and consent to participate

The studies involving human participants were reviewed and approved by the North West Multi-centre Research Ethics Committee (MREC) as a Research Tissue Bank (RTB) approval (UK Biobank, <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics>). The patients/participants provided their written informed consent to participate in this study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 1 September 2023 Accepted: 8 May 2024

Published online: 16 May 2024

References

1. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50(9):1219–24.
2. Polygenic Risk Score Task Force of the International Common Disease Alliance. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat Med*. 2021;27(11):1876–84.
3. Kullo IJ, Lewis CM, Inouye M, Martin AR, Ripatti S, Chatterjee N. Polygenic scores in biomedical research. *Nat Rev Genet*. 2022;23(9):524–32.

4. Duncan L, Shen H, Gelaye B, Meijssen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun*. 2019;10(1):3328.
5. D'Amour A, Heller K, Moldovan D, Adlam B, Alipanahi B, Beutel A, et al. Underspecification presents challenges for credibility in modern machine learning. *J Mach Learn Res*. 2022;23(1):10237–97.
6. Kachuri L, Chatterjee N, Hirbo J, et al. Principles and methods for transferring polygenic risk scores across global populations. *Nat Rev Genet*. 2023. <https://doi.org/10.1038/s41576-023-00637-2>.
7. Hassanin E, Maj C, Klinkhammer H, Krawitz P, May P, Bobbili DR. Assessing the performance of European-derived cardiometabolic polygenic risk scores in South-Asians and their interplay with family history. *BMC Med Genomics*. 2023;16(1):1–11.
8. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51(4):584–91.
9. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*. 2020;15(9):2759–72.
10. Curtis D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr Genet*. 2018;28(5):85–9.
11. Alexander DL, Tropsha A, Winkler DA. Beware of R^2 : simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *J Chem Inf Model*. 2015;55(7):1316–22.
12. Li G, Wang X. Prediction accuracy measures for a nonlinear model and for right-censored time-to-event data. *J Am Stat Assoc*. 2019;114(528):1815–25.
13. Archer L, Snell KJ, Ensor J, Hudda MT, Collins GS, Riley RD. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Stat Med*. 2021;40(1):133–46.
14. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R -squared is more informative than SMAPE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput Sci*. 2021;7:e623.
15. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux P, et al. Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *JAMA*. 2017;318(14):1377–84.
16. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):1–7.
17. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc*. 2020;27(4):621–33.
18. Wei J, Shi Z, Na R, Resurreccion WK, Wang CH, Duggan D, et al. Calibration of polygenic risk scores is required prior to clinical implementation: Results of three common cancers in UKB. *J Med Genet*. 2022;59(3):243–7.
19. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med*. 2020;12(1):1–11.
20. Qian J, Tanigawa Y, Du W, Aguirre M, Chang C, Tibshirani R, et al. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet*. 2020;16(10):e1009141.
21. Wand H, Lambert SA, Tamburro C, Iacocca MA, O'Sullivan JW, Sillari C, et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature*. 2021;591(7849):211–9.
22. Momin MM, Lee S, Wray NR, Lee SH. Significance tests for R^2 of out-of-sample prediction using polygenic scores. *Am J Hum Genet*. 2023;110(2):349–58.
23. Hawinkel S, Waegeman W, Maere S. Out-of-sample R^2 : Estimation and inference. *Am Stat*. 2024;78(1):15–25.
24. Wright S. Correlation and Causation. *J Agric Res*. 1921;20(1):557–85.
25. Kvalseth TO. Cautionary note about R^2 . *Am Stat*. 1985;39(4):279–85.
26. Anderson-Sprecher R. Model comparisons and R^2 . *Am Stat*. 1994;48(2):113–7.
27. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562:203–9.
28. Privé F, Aschard H, Carmi S, Folkersen L, Hoggart C, O'Reilly PF, et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am J Hum Genet*. 2022;109(1):12–23.
29. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun*. 2019;10(1):1776.
30. Klinkhammer H, Staerk C, Maj C, Krawitz PM, Mayr A. A statistical boosting framework for polygenic risk scores based on large-scale genotype data. *Front Genet*. 2023;13:1076440.
31. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet*. 2015;11(4):e1004969.
32. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4(1):s13742-015.
33. Purcell S, Chang C. PLINK 2.0. 2023. <https://www.cog-genomics.org/plink/2.0/>. Accessed 28 Oct 2020.
34. 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):4192.
35. Song W, Shi Y, Wang W, Pan W, Qian W, Yu S, et al. A selection pressure landscape for 870 human polygenic traits. *Nat Hum Behav*. 2021;5(12):1731–43.
36. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. *Nature*. 2022;610(7933):704–12.
37. Jung HU, Kim DJ, Baek EJ, Chung JY, Ha TW, Kim HK, et al. Gene-environment interaction explains a part of missing heritability in human body mass index. *Commun Biol*. 2023;6(1):324.
38. Mack S, Coassin S, Rueedi R, Yousri NA, Seppälä I, Gieger C, et al. A genome-wide association meta-analysis on lipoprotein (a) concentrations adjusted for apolipoprotein (a) isoforms. *J Lipid Res*. 2017;58(9):1834–44.
39. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74.
40. Ding Y, Hou K, Xu Z, Pimplaskar A, Petter E, Boulter K, et al. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*. 2023;618(7966):774–81.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4 Discussion with references

The publications in this cumulative dissertation illustrate the potential of polygenic prediction models and introduce a new modeling framework that provides a large flexibility regarding modeling choices. Finally, major challenges of PRS are discussed.

PRS can add value to known clinical and lifestyle risk factors, e.g., *Publication 1* (Dueñas et al., 2023) showed how PRS can provide additional information in the context of incomplete penetrance of monogenic variants. Therefore, accurate polygenic prediction models could enable personalized medicine by successfully stratifying risk and by being used to tailor prevention methods. However, many PRS achieve only limited prediction performance, particularly in out-of-target populations as seen in *Publication 4* (Staerk et al., 2024). But even in the target population, the phenomenon of missing heritability can be observed – partially explained by simplified statistical models that are applied to large-scale high-dimensional genetic data.

While most PRS methods rely on univariate summary statistics from GWAS and an adequate LD reference panel, our framework *snptest* is directly applicable to individual-level genotype data. By incorporating a batch-building step, computational hurdles are overcome and statistical boosting is made feasible for ultrahigh-dimensional genotype data (large n and large p). Thus, *snptest* enables multivariable regression for PRS which takes all SNPs into account simultaneously and implicitly performs variable selection. *Publication 2* (Klinkhammer et al., 2023) describes the *snptest* algorithm in detail and illustrates its potential for polygenic risk modeling. Particularly, *snptest* provides more accurate predictions than GWAS-based approaches and yields competitive performance to other multivariable methods such as *snptest* (Qian et al., 2020). In a next step, the modular nature of boosting has been exploited and the *snptest* framework has been extended. This extension incorporates various loss functions allowing for tailored and adequate statistical modeling of time-to-event and count data (*Publication 3*, Klinkhammer et al., 2024). Additionally, *snptest* features quantile regression, enabling the derivation of individual expected quantiles of the phenotype distribution conditional on an individual's genetic information. By that, individual polygenic prediction intervals can be constructed and refined via conformalized approaches (Romano et al., 2019). This

extension addresses a limitation of most established PRS methods, which only yield point estimates of the mean PRS without accounting for uncertainty. Overall, at its current state, the *snpboost* framework offers a large range of modeling choices, some of which have not been available before for individual-level genotype data. While statistical boosting demonstrates its potential as a powerful tool for advanced PRS modeling, it does not provide closed-form solutions for standard errors of effect estimates, limiting its use for inference. Furthermore, summary statistics from GWAS are easier shareable than individual-level data and involve fewer ethical concerns. A federated learning approach within the *snpboost* framework could enable training on several smaller cohorts without the need to share individual-level data.

Despite the progress made in the field of polygenic risk scores, there are still many challenges remaining. One major hurdle is addressed in *Publication 4* (Staerk et al., 2024), i.e., the transferability gap of PRS based on European ancestry when applied on non-European ancestries. While the latter and minorities are increasingly represented in biobanks, e.g. in the All of Us cohort (Bick et al., 2024), it is of interest to develop methods that can help to bridge the transferability gap and adequately take varying ancestries in the training data into account (Tanigawa and Kellis, 2023). Recently developed regression techniques such as anchor regression (Rothenhäusler et al., 2021) aim for robust prediction performance across different, possibly unseen environments. They can straightforwardly be integrated into the *snpboost* framework by implementing the corresponding loss function. Furthermore, the underlying genetic signal might not be based on linear SNP effects but rather non-linear effects, e.g., in the presence of dominant or recessive variants (Heyne et al., 2023; Ohta et al., 2024) which can be modeled via *snpboost* by replacing the linear base-learners. Last but not least, polygenic predisposition is only one factor in complex traits, and gene-environment interactions might be present. In literature, it has been hypothesized that SNPs that are associated with the variance of a trait can serve as proxies for gene-environment interactions (Wang et al., 2019; Miao et al., 2022). To identify those SNPs, *snpboost* will be extended to additionally support distributional regression, i.e., not only modeling the location parameter of a distribution but also scale and further distribution parameters (Kneib, 2013; Kneib et al., 2023).

4.1 Conclusion

In summary, this cumulative dissertation illustrates the potential and challenges of polygenic risk scores and presents the new statistical boosting framework *snpboost* to derive advanced polygenic prediction models directly from individual-level genotype data of large cohort studies. The modular nature of boosting opens up the opportunity to apply well-studied advanced statistical models to complex ultrahigh-dimensional genetic data. While the publications in this dissertation particularly focused on different outcome distributions, the framework is also conveniently extendable to non-linear effects and distributional regression.

4.2 References

- Bick AG et al. Genomic data in the All of Us Research Program. In: *Nature* 2024; 627 (8003): 340–346
- Dueñas N, Klinkhammer H, Bonifaci N, Spier I, Mayr A, Hassanin E, Diez-Villanueva A, Moreno V, Pineda M, Maj C, Capellà G, Aretz S, Brunet J. Ability of a polygenic risk score to refine colorectal cancer risk in Lynch syndrome. In: *Journal of Medical Genetics* 2023; 60 (11): 1044–1051
- Heyne HO, Karjalainen J, Karczewski KJ, Lemmelä SM, Zhou W, Havulinna AS, Kurki M, Rehm HL, Palotie A, Daly MJ. Mono- and biallelic variant effects on disease at biobank scale. In: *Nature* 2023; 613 (7944): 519–525
- Klinkhammer H, Staerk C, Maj C, Krawitz PM, Mayr A. Genetic Prediction Modeling in Large Cohort Studies via Boosting Targeted Loss Functions. In: *Statistics in Medicine* 2024; 43 (28): 5412–5430
- Klinkhammer H, Staerk C, Maj C, Krawitz PM, Mayr A. A statistical boosting framework for polygenic risk scores based on large-scale genotype data. In: *Frontiers in Genetics* 2023; 13: 1076440
- Kneib T. Beyond mean regression. In: *Statistical Modelling* 2013; 13 (4): 275–303
- Kneib T, Silbersdorff A, Säfken B. Rage Against the Mean – A Review of Distributional Regression Approaches. In: *Econometrics and Statistics* 2023; 26: 99–123

- Miao J, Lin Y, Wu Y, Zheng B, Schmitz LL, Fletcher JM, Lu Q. A quantile integral linear model to quantify genetic effects on phenotypic variability. In: *Proceedings of the National Academy of Sciences* 2022; 119 (39): e2212959119
- Ohta R, Tanigawa Y, Suzuki Y, Kellis M, Morishita S. A polygenic score method boosted by non-additive models. In: *Nature Communications* 2024; 15 (1): 4433
- Qian J, Tanigawa Y, Du W, Aguirre M, Chang C, Tibshirani R, Rivas MA, Hastie T. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. In: *PLOS Genetics* 2020; 16 (10): e1009141
- Romano Y, Patterson E, Candes E. Conformalized Quantile Regression. In: *Advances in Neural Information Processing Systems*. Vol. 32, 33rd Conference on Neural Information Processing Systems, 2019: 449
- Rothenhäusler D, Meinshausen N, Bühlmann P, Peters J. Anchor Regression: Heterogeneous Data Meet Causality. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 2021; 83 (2): 215–246
- Staerk C, Klinkhammer H, Wistuba T, Maj C, Mayr A. Generalizability of polygenic prediction models: how is the R^2 defined on test data? In: *BMC Medical Genomics* 2024; 17 (1): 132
- Tanigawa Y, Kellis M. Power of inclusion: Enhancing polygenic prediction with admixed individuals. In: *The American Journal of Human Genetics* 2023; 110 (11): 1888–1902
- Wang H, Zhang F, Zeng J, Wu Y, Kemper KE, Xue A, Zhang M, Powell JE, Goddard ME, Wray NR, Visscher PM, McRae AF, Yang J. Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. In: *Science Advances* 2019; 5 (8): eaaw3538

Complete publication list

PRS projects

- Dueñas N, Klinkhammer H, Bonifaci N, Spier I, Mayr A, Hassanin E, Diez-Villanueva A, Moreno V, Pineda M, Maj C, Capellà G, Aretz S, Brunet J. Ability of a polygenic risk score to refine colorectal cancer risk in Lynch syndrome. In: *Journal of Medical Genetics* 2023; 60 (11): 1044–1051
- Hassanin E, Maj C, Klinkhammer H, Krawitz P, May P, Bobbili DR. Assessing the performance of European-derived cardiometabolic polygenic risk scores in South-Asians and their interplay with family history. In: *BMC Medical Genomics* 2023; 16 (1): 164
- Hassanin E, Spier I, Bobbili DR, Aldisi R, Klinkhammer H, David F, Dueñas N, Hüneburg R, Perne C, Brunet J, Capella G, Nöthen MM, Forstner AJ, Mayr A, Krawitz P, May P, Aretz S, Maj C. Clinically relevant combined effect of polygenic background, rare pathogenic germline variants, and family history on colorectal cancer incidence. In: *BMC Medical Genomics* 2023; 16 (1): 42
- Klau JH, Maj C, Klinkhammer H, Krawitz PM, Mayr A, Hillmer AM, Schumacher J, Heider D. AI-based multi-PRS models outperform classical single-PRS models. In: *Frontiers in Genetics* 2023; 14: 1217860
- Klinkhammer H, Staerk C, Maj C, Krawitz PM, Mayr A. Genetic Prediction Modeling in Large Cohort Studies via Boosting Targeted Loss Functions. In: *Statistics in Medicine* 2024; 43 (28): 5412–5430
- Klinkhammer H, Staerk C, Maj C, Krawitz PM, Mayr A. A statistical boosting framework for polygenic risk scores based on large-scale genotype data. In: *Frontiers in Genetics* 2023; 13: 1076440
- Maj C, Staerk C, Borisov O, Klinkhammer H, Wai Yeung M, Krawitz P, Mayr A. Statistical learning for sparser fine-mapped polygenic models: The prediction of LDL-cholesterol. In: *Genetic Epidemiology* 2022; 46 (8): 589–603

- Mauschitz MM, Hochbein BJ, Klinkhammer H, Saßmannshausen M, Terheyden JH, Krawitz P, Finger RP. Prevalence and determinants of subretinal drusenoid deposits in patients' first-degree relatives. In: Graefe's Archive for Clinical and Experimental Ophthalmology 2023; 262 (1): 53–60
- Staerk C, Klinkhammer H, Wistuba T, Maj C, Mayr A. Generalizability of polygenic prediction models: how is the R2 defined on test data? In: BMC Medical Genomics 2024; 17 (1): 132

GestaltMatcher

- Arlt A, Knaus A, Hsieh TC, Klinkhammer H, Bhasin MA, Hustinx A, Moosa S, Krawitz P, Ekure E. Next-generation phenotyping in Nigerian children with Cornelia de Lange syndrome. In: American Journal of Medical Genetics Part A 2024; 194 (9): e63641
- Blackburn PR, Ebstein F, Hsieh TC, Motta M, Radio FC, Herkert JC, Rinne T, Thiffault I, Rapp M, Alders M, et al. Loss-of-Function Variants in CUL3 Cause a Syndromic Neurodevelopmental Disorder. In: Annals of Neurology 2024; 97 (1): 76–89
- Hsieh TC, Bar-Haim A, Moosa S, Ehmke N, Gripp KW, Pantel JT, Danyel M, Mensah MA, Horn D, Rosnev S, et al. GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. In: Nature Genetics 2022; 54 (3): 349–357
- Lesmann H, Klinkhammer H, Krawitz PM. The future role of facial image analysis in ACMG classification guidelines. In: Medizinische Genetik 2023; 35 (2): 115–121
- Lyon GJ, Vedaie M, Beisheim T, Park A, Marchi E, Gottlieb L, Hsieh TC, Klinkhammer H, Sandomirsky K, Cheng H, et al. Expanding the phenotypic spectrum of NAA10-related neurodevelopmental syndrome and NAA15-related neurodevelopmental syndrome. In: European Journal of Human Genetics 2023; 31 (7): 824–833
- Oppermann H, Marcos-Grañeda E, Weiss LA, Gurnett CA, Jelsig AM, Vineke SH, Isidor B, Mercier S, Magnussen K, Zacher P, et al. CUX1-related neurodevelopmental disorder: deep insights into phenotype-genotype spectrum and underlying pathology. In: European Journal of Human Genetics 2023; 31 (11): 1251–1260

Other projects

- Aldisi R, Hassanin E, Sivalingam S, Buness A, Klinkhammer H, Mayr A, Fröhlich H, Krawitz P, Maj C. GenRisk: a tool for comprehensive genetic risk modeling. In: *Bioinformatics* 2022; 38 (9): 2651–2653
- Aldisi R, Hassanin E, Sivalingam S, Buness A, Klinkhammer H, Mayr A, Fröhlich H, Krawitz P, Maj C. Gene-based burden scores identify rare variant associations for 28 blood biomarkers. In: *BMC Genomic Data* 2023; 24 (1): 50
- Schmidt A, Danyel M, Grundmann K, Brunet T, Klinkhammer H, Hsieh TC, Engels H, Peters S, Knaus A, Moosa S, et al. Next-generation phenotyping integrated in a national framework for patients with ultrarare disorders improves genetic diagnostics and yields new molecular findings. In: *Nature Genetics* 2024; 56 (8): 1644–1653
- Schmidt A, Röner S, Mai K, Klinkhammer H, Kircher M, Ludwig KU. Predicting the pathogenicity of missense variants using features derived from AlphaFold2. In: *Bioinformatics* 2023; 39 (5): btad280
- Strömer A, Klein N, Staerk C, Klinkhammer H, Mayr A. Boosting multivariate structured additive distributional regression models. In: *Statistics in Medicine* 2023; 42 (11): 1779–1801

5 Acknowledgments

First, I thank my supervisors, Prof. Dr. Andreas Mayr and Prof. Dr. Peter Krawitz, for their invaluable feedback as well as creating an open-minded and supportive research environment. I am equally grateful to Dr. Carlo Maj and Dr. Christian Staerk for the highly enjoyable collaboration within our PRS research group. My sincere gratitude goes to the members of my dissertation committee, PD Dr. Christoph Engel and Prof. Dr. Stefan Aretz, for their guidance and input. I would also like to thank Charlotte, Annika and Christian for their proofreading of my thesis and Charlotte, in particular, for her great support with its formalities.

I deeply appreciate all my colleagues at IGSB and IMBIE for their engaging discussions, lively lunches, and the fun shared during conferences – what a journey it has been.

Above all, I am profoundly grateful to my family and friends for their unconditional support and encouragement, always reminding me of what truly matters in life. Thank you, Jan, for taking me as I am and for everything you have done for me.