Received: 29 May 2019 | Revised: 5 December 2019 | Accepted: 23 December 2019

DOI: 10.1002/gepi.22279

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

RESEARCH ARTICLE

Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status

Damian Gola[1] | **Jeannette Erdmann**[2] | **Bertram Müller-Myhsok**[3] | **Heribert Schunkert**[4] | **Inke R. König**[1]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Lübeck, Germany

[2]Institute for Cardiogenetics, Universität zu Lübeck, Lübeck, Germany

[3]Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry, Munich, Germany

[4]Deutsches Herzzentrum München, Technische Universität München, München, Germany

Correspondence

Inke R. König, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany.
Email: Inke.Koenig@imbs.uni-luebeck.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: KO2250/7; Bundesministerium für Bildung und Forschung, Grant/Award Number: 81Z1700103

Abstract

Coronary artery disease (CAD) is the leading global cause of mortality and has substantial heritability with a polygenic architecture. Recent approaches of risk prediction were based on polygenic risk scores (PRS) not taking possible nonlinear effects into account and restricted in that they focused on genetic loci associated with CAD, only. We benchmarked PRS, (penalized) logistic regression, naïve Bayes (NB), random forests (RF), support vector machines (SVM), and gradient boosting (GB) on a data set of 7,736 CAD cases and 6,774 controls from Germany to identify the algorithms for most accurate classification of CAD status. The final models were tested on an independent data set from Germany (527 CAD cases and 473 controls). We found PRS to be the best algorithm, yielding an area under the receiver operating curve (AUC) of 0.92 (95% CI [0.90, 0.95], 50,633 loci) in the German test data. NB and SVM (AUC ~ 0.81) performed better than RF and GB (AUC ~ 0.75). We conclude that using PRS to predict CAD is superior to machine learning methods.

KEYWORDS

classification, coronary artery disease, machine learning, polygenic risk scores, prediction

1 | INTRODUCTION

An essential part of precision medicine is the development of diagnostic and prognostic models. This can be challenging especially in the analysis of complex diseases like coronary artery disease (CAD), as many environmental and genetic variants simultaneously affect disease risk (Smith et al., 2005). CAD is caused by deposits in the arterial walls of the coronary arteries, which leads to a reduced or incomplete blood flow and thus to a reduced oxygen supply to the heart. Thus, this chronic disease develops over years and leads to concomitant symptoms such as cardiac arrhythmias or myocardial infarction. CAD is currently one of the most common causes of death or disability worldwide (Lopez, Mathers, Ezzati, Jamison, & Murray, 2006). In addition to lifestyle and environmental factors (Yusuf et al., 2004), familial clusters of this disease indicate a significant genetic background (Marenberg, Risch, Berkman, Floderus, & Faire, 1994). Numerous studies reviewed by Khera and Kathiresan (2017) have been published so far, associating a total of approximately 60 individual genetic variants with CAD. Thus, CAD is a polygenic disease with a substantial heritability which makes risk estimation based on the genetic background attractive. Models for risk assessment for CAD have already been proposed and entered clinical routine, but these are mainly based only on clinical variables, such as the *HeartScore* (Thomsen, 2005) and the *Framingham Risk Score* (Wilson et al., 1998).

The extent to which the addition of scores based on individual genetic variants to such clinical scores can improve these existing models has been repeatedly investigated (Abraham et al., 2016; Beaney, Cooper, Drenos, & Humphries, 2017; Krarup et al., 2015; Ripatti et al., 2010; Tada et al., 2016). These polygenic risk scores (PRS) had the limitation to consider only genetic variants for which an association with CAD has previously been established. In contrast, in two recent papers by Khera et al. (2018) and Inouye et al. (2018), genome-wide polygenic risk scores (GPRS) were proposed that use millions of genetic variants to predict the risk of CAD and other complex diseases at a high level of accuracy, regardless of their association with the disease. The authors state that this would enable risk prediction at the time of birth and thus early and effective prevention programs. However, the proposed risk score prediction models are built as simple sums of the single genotypes, weighted by their univariable effect on disease, thus based on the assumption of linear additive effects of the underlying clinical and genetic factors only.

Here, we want to address the question whether accounting for nonlinear effects of these factors can further improve the predictiveness of models. To answer this question, we utilize various methods from the field of machine learning which offer attractive algorithms to model nonlinear effects. Our assumption is that more complex algorithms should provide a better way to model the complex genetic driving structures of a complex disease like CAD. The aim of this study is to assess how good different algorithms can discriminate CAD cases from controls, that is a classification, using solely the genetic information. The results of this study will assist researchers in finding the best approach to incorporate the genetic information into risk modeling approaches by compressing the genetic information in a way that provides the most discriminative value.

2 | METHODS

2.1 | Algorithms

Throughout this study we assume that n_D samples in a data set D are characterized by p predictor variables (predictors) $\mathbf{X} \in \mathcal{X}^{n_D \times p} = \mathcal{X}_1^{n_D} \times \cdots \times \mathcal{X}_p^{n_D}$. The predictors are genotypes of single nucelotide polymorphisms (SNP), thus $\mathcal{X}_j = \{0,1,2\}$. Additionally, the outcome (outcome) $Y \in y = \{-1,1\}$ denotes the control (-1) or case (1) status of a sample. The task for a classification model $M^D_{\mathrm{A};\mathbf{h}}$, based on algorithm A with hyperparameter settings $\mathbf{h}$ and trained on D, is to estimate the case probability ($Y = 1$), given a realization $\mathbf{x}$ of the predictors: $\hat{\mathbb{P}}(Y = 1 \mid \mathbf{x}) = M^D_{\mathrm{A};\mathbf{h}}(\mathbf{x})$. In this study we considered six commonly used algorithms for creating classification models: GPRS, naïve Bayes (NB) classifier, regularized regression, random forest (RF), gradient boosting (GB), and support vector machine (SVM). Detailed information on these prediction algorithms are presented in the Supporting Information, or can be found for example in Hastie, Tibshirani, and Friedman (2009). All algorithms can be used in conjunction with the R package mlr (Bischl et al., 2016), which was used for benchmarking, hyperparameter tuning, variable selection, training, and testing in version 2.12.

2.2 | Data sets

We used six imputed data sets with samples of European descent from the German population. The original data includes a total of 9,314 observations with diagnosed CAD (cases) and 8,160 observations without CAD (controls). An observation has been defined as a case if he/she has myocardial infarction, acute coronary syndrome, angina pectoris, or coronary stenosis greater than 50%. Details on the single data sets including the respective number of observations (columns 6–8) are listed in Table 1. Detailed age information is not available in the data sets at hand.

2.3 | Imputation and data set-specific quality control

Before imputation, the data sets were individually subjected to the same quality control steps, as described in Andlauer et al. (2016). This includes the exclusion of samples whose proportion of missing genotypes exceeds 2% or whose proportion of heterozygous genotypes deviates four standard deviations from the mean value in the respective data set. In addition, samples who were identified as population outliers based on multidimensional scaling (MDS) components were excluded. One sample of pairs with estimated cryptic relationship less than four, that is any closer than first cousins pairs, was excluded. Genetic variants were excluded if their proportion of missing genotypes exceeded 2%, the minor allele frequency (MAF) estimated in the data set was less than 1%, or the p value for the test on deviation from Hardy–Weinberg equilibrium (HWE) was lower than 1×10^{-5}. In addition, palindromic SNPs

TABLE 1 Description of available data sets

Data set	First described	Array	CAD Definition	Controls taken from	Cases[a]	Controls[a]	Total[a]	Cases (% female)[b]	Controls (% female)[b]	Total (% female)[b]
GerMIFSI	Samani et al. (2007)	Affymetrix® GeneChip® Human Mapping 500 K Array Set	Myocardial infarction before the age of 60 and at least one first-degree relative with CAD	Wichmann, Gieger, and Illig (2005)	875	1,644	2,519	622 (33.3)	1,521 (51.3)	2,143 (46.1)
GerMIFSII	Erdmann et al. (2009)	Affymetrix® Genome-Wide Human SNP Array 6.0	Myocardial infarction before the age of 60 or at least one first-degree relative with CAD	Kolz et al. (2009); Krawczak et al. (2006)	1,222	1,298	2,520	1,188 (20.7)	1,238 (47.9)	2,426 (34.6)
GerMIFSIII	Erdmann et al. (2011)	Affymetrix® Genome-Wide Human SNP Array 5.0 (cases), Affymetrix® Genome-Wide Human SNP Array 6.0 (controls)	Myocardial infarction between ages 26 and 74	Kolz et al. (2009); Krawczak et al. (2006)	1,157	1,748	2,905	1,048 (20.0)	1,419 (48.5)	2,467 (36.4)
GerMIFSIV	Deloukas et al. (2012)	Illumina® MetaboChip Array	CAD diagnosed before the age of 65 (men) or 70 (women)	German population	1,254	1,404	2,658	940 (35.2)	1,128 (61.3)	2,068 (49.4)
GerMIFSV	Brænne et al. (2017)	Illumina® Human OmniExpress or Illumina® Human Omni2.5	Unknown	German population	2,459	1,445	3,904	2,392 (24.3)	1,537 (52.4)	3,929 (35.3)
LURIC	Deloukas et al. (2012)	Illumina® Metabochip Array	>50% angiographic confirmation of vascular obstruction in at least one coronary vessel	Winkelmann et al. (2001)	2,347	621	2,968	2,085 (25.1)	591 (43.1)	2,676 (29.1)
Total					9,314	8,160	17,474	8,275 (23.4)	7,434 (51.3)	15,709 (37.6)

[a] These columns refer to the original nonimputed data sets.

[b] These columns refer to the imputed data sets.

with alleles A/T or G/C were removed. The *Sanger Imputation Service* (McCarthy et al., 2016; Wellcome Trust Sanger Institute, n.d.) was used for imputation of the individual data sets. Estimation of the haplotypes was done using SHAPEIT2 (Delaneau, Zagury, & Marchini, 2012) and IMPUTE2 (Howie, Donnelly, & Marchini, 2009) was used as the imputation algorithm. The *Haplotype Reference Consortium* served as reference panel (McCarthy et al., 2016). The variants were subjected to further quality control after the imputation. Variants with an imputation quality less than 0.8, a MAF less than 1%, a *p* value in the test for deviation from HWE less than 1×10^{-20} or a proportion of missing genotypes greater than 2% were removed from the imputed data sets. After imputation and quality control, a total of 8,275 CAD cases and 7,434 controls with 5,539,917 genetic variants present in all data sets are available. Details on the number of observations in each imputed data set are shown in Table 1 columns 9–11.

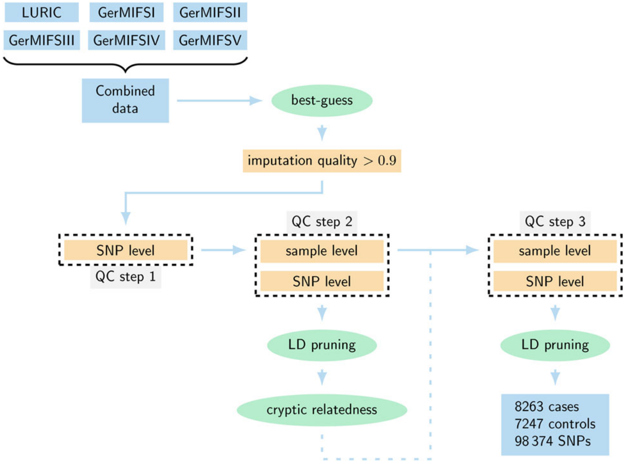

FIGURE 1 Flow chart illustrating the preprocessing and quality control. Solid lines indicate data flow, and the loosely dashed line indicates flow of cryptic relatedness analysis results

2.4 | Preprocessing and quality control on combined data

For the following analyses, the data sets G1, G2, G3, G4, G5, and LURIC were combined to form one data set to achieve more stable model estimates due to the increased number of samples.

Combining the data sets required further quality control which is summarized in the flow chart (Figure 1).

Genotype probabilities were converted into fixed genotypes using PLINK, version 1.9b4.4 (Chang et al., 2015) using the best-guess method, with the highest probability genotype selected as the fixed genotype. Genotypes for which the highest probability was less than 0.9 were set to missing. Furthermore, SNPs with imputation quality <0.9 in any data set were excluded.

Further quality control included three steps and were also performed with PLINK. The first step was quality control at the SNP level. SNPs were removed if the proportion of missing genotypes exceeded 2% (call rate > 98%), if the MAF was less than 5% or if the test for deviation from HWE yielded a *p* value of less than 1×10^{-5}. In the second step, criteria were applied at the sample level. Observations were removed if the proportion of missing genotypes exceeded 2% (call rate > 98%) or if the proportion of heterozygous SNPs differed by more than three standard deviations from the mean in the respective data set. For the last step, the cryptic relationship in the data sets was analyzed. For the estimation of the cryptic relationship SNPs in minimal LD to each other were selected. Regions with 2×10^6 base pairs were considered, in each of which one SNP was removed from SNP pairs with $r^2 \geq 0.2$. The regions were gradually shifted by 2×10^5 base pairs. Based on the remaining SNPs, the cryptic relationship was analyzed. In the third step, the SNPs pruned for cryptic relatedness were added back into the data, and one of two samples was removed for pairs in which the degree of the cryptic relationship was less than two. Finally, the criteria used in the second step were reapplied.

2.5 | Training and test data sets

To allow the testing of the classification models, the combined data set consisting of the data sets G1, G2, G3, G4, G5, and LURIC was divided into a training data set D_{train} and an independent test data set D^*_{test}. For the test data set D^*_{test}, $n^*_{\text{test}} = 1000$ randomly selected samples were removed from the base data set.

To avoid unnecessarily increasing the computational effort due to highly correlated predictors, from SNP pairs in D that are in high LD one SNP was removed using PLINK (LD pruning). A threshold of $r^2 = 0.5$ was chosen. The regions considered included 1×10^4 base pairs. With each iteration, this region was shifted by 1×10^3 base pairs.

For quality control, the imputed data were transformed into unique genotypes using a best guess procedure, as described above. Here, sporadically missing values can occur. However, some of the prediction algorithms used cannot handle missing values. To avoid the exclusion of too many samples or predictors due to isolated missing values, the imputed data of the genetic markers remaining in the data sets after quality control and LD pruning were retransformed using the expected genotype, that is $x_{i,j} = p_1 + 2p_2$, where p_1 and p_2 are the probabilities of sample i to have one or two alternate alleles at SNP j. Thus, the support of the predictors in all data sets is $\mathcal{X}_j = [0,2] \subset \mathbb{R}, j = 1, \ldots, p$.

2.6 | Genome-wide association

As described in the Supporting Information, the PRS needs weights for each SNP. Since the available data sets contributed to most of the published GWAS results for CAD, taking public effect estimates could induce a bias when evaluating the classification performance. Therefore, a GWAS was first performed on the training data set D_{train}. For the estimation of the single SNP effects PLINK was used on the best guess genotypes. Since the phenotype in the present data sets is dichotomous, a logistic regression was performed with sex as an additional covariable. The effect estimates from this GWAS were used as weights for the PRS.

2.7 | Variable selection

The training data set D_{train} comprises a large number of predictors. From these, RFs, regularized regression and GB use internal mechanisms to identify those that are relevant for the prediction. The predictive power of PRS, NB classifier, and SVM, on the other hand, depends on how many noninformative predictors are present in the data set. Nevertheless, the prediction quality can also be improved for the algorithms with internal variable selection, but above all the duration of the training of the corresponding algorithms can be reduced if an external variable selection is carried out beforehand. One possibility of variable selection is the prioritization of the predictors on the basis of an importance measure.

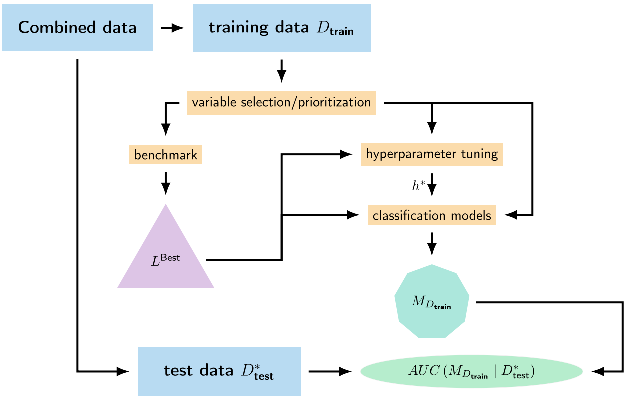
Combined data
training data D_{train}
variable selection/prioritization
benchmark
hyperparameter tuning
h^*
classification models
L^{Best}
$M_{D_{\text{train}}}$
test data D^*_{test}
$AUC\,(M_{D_{\text{train}}} \mid D^*_{\text{test}})$

FIGURE 2 Flow chart illustrating the different steps of building the classification models

During training of each of the classification algorithms, the best $\tilde{p}$ predictors regarding the importance measure were to be selected (see Figure 2). This value was added as an additional hyperparameter to be optimized to the hyperparameter search spaces defined in Table 2. Thus, the optimal number of SNPs as predictors was chosen free of any further hypotheses automatically during the optimization process for each algorithm.

As Wright, Ziegler, and König (2016) report, RFs are generally able to account for interactions between predictors for predictions, although it is not possible to identify interactions as such. Therefore, the corrected Gini importance, an unbiased measure of the total decrease in node impurity (Nembrini, König, & Wright, 2018), was used as the measure of variable importance. The corrected Gini importance is unbiased like the permutation importance and is almost as computationally fast as the classic Gini importance. Specifically, this is achieved by adding a randomized version of each predictor variable during training of a RF and correcting the Gini importance of each original predictor variable by the Gini importance of the corresponding randomized predictor variable. For this purpose, a RF with 50,000 trees and mtry (the number of variables available for splitting at each tree node) set to 10% of all available predictors was grown on the training data set with the R package ranger, version 0.8.1-300 (Wright & Ziegler, 2017). Only variables with a corrected Gini import greater than 0 were used to train the algorithms.

2.8 | Benchmark

To select the classification algorithms best suited for the classification of CAD, they were compared by nested cross-validation (CV) with 10-fold outer CV by the area under the receiver operating characteristic (ROC) curve (AUC). During this performance comparison, each outer CV training data set was randomly reduced to 20% of the samples to limit the computational overhead. The hyperparameters of the individual algorithms were optimized by means of a sequential model based optimization (SMO) on each reduced outer CV training data set using a fivefold inner CV. For this, the R package mlrMBO, version 1.1.0 (Bischl et al., 2017), was used. As surrogate model in the SMO a RF with 500 trees and mtry $= \lfloor \sqrt{d_{\mathcal{H}}} \rfloor$ was used, where $d_{\mathcal{H}}$ is the number of hyperparameters of the respective algorithm to be optimized. As implementation for the surrogate model, the R package ranger was used. The remaining

TABLE 2 Hyperparameter search spaces and optimal hyperparameter settings of the classification algorithms

Algorithm	Hyperparameter[a]			Search space	Optimal hyperparameter[b]	SNPs[c]	AUC[d]
Polygenic risk score	Weight			{TRUE, FALSE}	TRUE	50,633	0.9106
Support vector machine	Type			{C-svc, nu-svc}	C-svc	8469	0.8149
		C-svc	C	$10^{[-5,5]} \subset \mathbb{R}$	0.0011		
		nu-svc	nu	$[0,1] \subset \mathbb{R}$			
	Kernel			{vanilladot, rbfdot, polydot, laplacedot, besseldot}	besseldot		
		rbfdot	sigma	$10^{[-5,2]} \subset \mathbb{R}$			
		polydot	degree	$[1,5] \subset \mathbb{N}$			
			scale	$10^{[-5,5]} \subset \mathbb{R}$			
			offset	$2^{[-3,3]} \subset \mathbb{R}$			
		laplacedot	sigma	$10^{[-5,2]} \subset \mathbb{R}$			
		besseldot	order	$[0,6] \subset \mathbb{N}$	1		
			degree	$[1,5] \subset \mathbb{N}$	5		
			sigma	$10^{[-5,2]} \subset \mathbb{R}$	3.52×10^{-5}		
	Shrinking			{TRUE, FALSE}	TRUE		
Naïve Bayes classifier	Laplace			$[0,10] \subset \mathbb{R}$	0.0738	10,508	0.8137
Random forest	num.trees			$\{100, 200, ..., 5000\} \subset \mathbb{N}$	3,400	1,357	0.7649
	mtry			$[0.001, 0.1] \subset \mathbb{R}$	0.00206		
	min.node.size			$[10, 100] \subset \mathbb{N}$	79		
	replace			{TRUE, FALSE}	TRUE		
Gradient boosting	eta			$[0,1] \subset \mathbb{R}$	9.69×10^{-5}	3,120	0.7646
	booster			{gbtree, gblinear}	gbtree		
		gbtree	gamma	$[0,10] \subset \mathbb{R}$	4.35		
			max_depth	$[1,14] \subset \mathbb{N}$	13		
			min_child_weight	$2^{[0,7]} \subset \mathbb{R}$	2.16		
			subsample	$[0,1] \subset \mathbb{R}$	0.0644		
			colsample_bytree	$[0,1] \subset \mathbb{R}$	0.703		
			colsample_bylevel	$[0,1] \subset \mathbb{R}$	0.369		
		gblinear	lambda	$2^{[-10,10]} \subset \mathbb{R}$			
			lambda_bias	$[0,10] \subset \mathbb{R}$			
			alpha	$[0,1] \subset \mathbb{R}$			
	base_score			$[0,1] \subset \mathbb{R}$	0.587		
	nrounds			$[1,5000] \subset \mathbb{N}$	3,887		
Logistic regression	link			{logit, probit, cloglog}			
Regularized regression	alpha			$[0,1] \subset \mathbb{R}$			
	s			{lambda.1se, lambda.min}			
	standardize			{TRUE, FALSE}			

Note: Details about the individual hyperparameters can be found in the descriptions of the individual classification algorithms and in the documentation of the respective R packages.

Abbreviations: AUC, area under the receiver operating curve; SNP, single nucleotide polymorphism; SVC, support vector machine.

[a] Hyperparameters in this column should be read as nested hierarchical list. The left most values denote level 1 hyperparameters, the value below and right denote a selected value for the level 1 hyperparameter, and finally, the right most values denote level 2 hyperparameters dependent on the selected value of the respective level 1 hyperparameter. For example, the Support Vector Machine algorithm supports two different values for the level 1 hyperparameter type, namely C-svc, and nu-svc. If during hyperparameter tuning, C-svc is selected, there is a dependent level 2 hyperparameter C with its own search space, and if nu-svc is selected, another level 2 hyperparameter nu which has to be tuned over its own search space.

[b] Optimized hyperparameters determined by 10-fold cross-validation.

[c] Optimal number of SNPs with the highest corrected Gini importance which enter the classification models as predictors.

[d] Average AUC over 10 cross-validation test data sets with the optimal hyperparameter settings after training.

hyperparameters of the surrogate model remained in the default settings. The internal hyperparameter optimization was limited to a maximum of 100 iterations and a maximum runtime of 4 weeks. The following R packages were used as implementations of each of the classification algorithms described above:

1. *Polygenic risk score* `riskScoreR`, version 0.6 (own development, available at https://github.com/imbs-hl/riskScoreR);
2. *Naïve Bayes classifier* `e1071`, Version 1.6-7 (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2017)*;*
3. *Regularized regression* `glmnet`, Version 2.0-5 (Friedman, Hastie, & Tibshirani, 2000);
4. *Random forest* `ranger`, Version 0.8.1-300 (Wright & Ziegler, 2017);
5. *Gradient boosting* `xgboost`, version 0.6-4 (Chen et al., 2018);
6. *Support vector machine* `kernlab`, version 0.9-25 (Karatzoglou, Smola, Hornik, & Zeileis, 2004).

The hyperparameter search spaces of each algorithm were defined as shown in Table 2.

2.9 | Hyperparameter tuning

Running the benchmark allowed for the selection of the best algorithms. However, from the benchmark no optimal hyperparameter settings for any algorithm can be derived, as in each fold of the outer CV, other training data is taken. Thus, for each outer fold one optimal model for each algorithm has been created by hyperparameter optimization in the inner CV. Therefore, for the best algorithms selected by the benchmark, a hyperparameter optimization by means of SMO was subsequently performed with respect to the AUC as a measure of quality on the complete training data set D_{train}. Again, the R package `mlrMBO` and a forest with 500 trees and $\texttt{mtry} = \lfloor\sqrt{d_{\mathcal{H}}}\rfloor$ were used as a surrogate model, where $d_{\mathcal{H}}$ is the number of hyperparameters of the respective algorithm to be optimized. As an implementation for the surrogate model, the R package ranger was used. The remaining hyperparameters of the surrogate model remained in the default settings. Hyperparameter optimization was limited to a maximum of 100 iterations and a maximum runtime of 4 weeks with 10-fold CV. We used the same hyperparameter spaces as for the benchmark, listed in Table 2.

Finally, a final classification model was compiled for each of the best algorithms with the optimal hyperparameter settings on the entire training data set.

For an overview of the steps to build the final classification models see Figure 2.

2.10 | Testing of models

To check the classification quality, the final models were primarily evaluated on the test data set D^*_{test} with respect to the AUC and their 95% confidence intervals according to DeLong, DeLong, and Clarke-Pearson (1988). Performance in terms of mean misclassification error, balanced accuracy, true positive and negative rates, and false positive and negative rates was also evaluated for different thresholds converting the probability estimation of the respective models into hard classifications. In addition, a graphical evaluation is carried out on the basis of the ROC curves, and the rank correlation of the predicted class probabilities of the individual classification models is examined on the basis of the Kendall correlation (Kendall, 1938).

3 | RESULTS

3.1 | Preprocessing and quality control

After merging data sets G1, G2, G3, G4, G5, and LURIC, the combined data set comprised a total of 15,709 observations, of which 8,275 were cases of CAD and 7,434 were controls. A total of 4,243,908 SNPs with an imputation quality greater than 0.9 were present in all data sets.

After quality control, the combined data set comprised 15,510, observations and 2,777,815 SNPs. The exclusion of highly correlated SNPs provided 98,374 SNPs for further steps. The proportion of CAD cases was 52.96%, and the proportion of women was 37.51%.

Details on the single quality control steps can be found in the Supporting Information.

3.2 | Training and test data sets

After preprocessing and quality control, the combined data set included 15,510 observations with 7,247 (46.7%) controls and 8,263 (53.3%) CAD cases. Of these, $n^*_{\text{test}} = 1000$ observations (473 controls and 527 CAD cases) were set apart for the test data set D^*_{test} for later testing of the classification models. Thus, the training data set D_{train} comprised a total of $n = 14,510$ observations with 6,774 (46.7%) controls and 7,736 (53.3%) CAD cases.

3.3 | Variable importance

The calculation of the corrected Gini importance required about 22 hr with 16 parallel threads on an Intel® Xeon® E5-2680 2.70 GHz CPU. Overall, 50,646

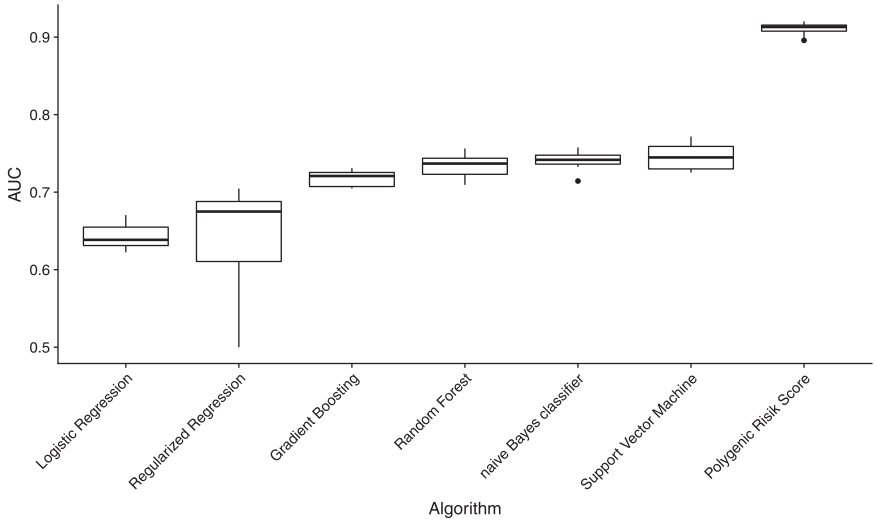

FIGURE 3 Result of the benchmark. Shown are the values of the quality measure area under the receiver operating characteristic curve per algorithm from the outer cross-validation as box plots

SNPs had a positive corrected Gini importance and were included in the further steps.

3.4 | Benchmark

All algorithms reached the maximum number of iterations for the SMO within the respective 4-week runtime. Figure 3 shows the performance in terms of AUC of the different classification algorithms in the benchmark. The PRS leads this benchmark with a median AUC of 0.9131 well ahead of the other classification algorithms. On the second to fifth place follow the SVM, NB classifier, RF, and GB. These four algorithms achieve a median AUC > 0.7 in the benchmark. In particular, the SVM, the NB classifier and the RF are very close together with median AUCs of 0.7448, 0.7417, and 0.7369. Regularized regression and logistic regression are some way behind GB and below a median AUC of 0.7. These classification algorithms were therefore excluded from the further steps.

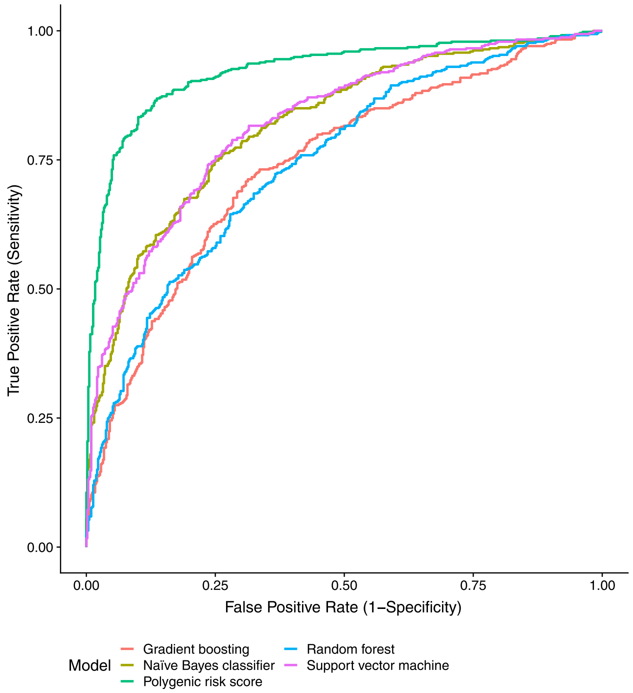

FIGURE 4 Receiver operating characteristic curves of the final models on the test data set D^*_{test}

3.5 | Hyperparameter tuning

The hyperparameter tuning was performed for the classification algorithms PRS, SVM, NB classifier, RF, and GB. For all algorithms, 100 optimization iterations could be performed within the maximum runtime of 4 weeks. The optimal hyperparameter settings of the individual classification algorithms are summarized in the last columns of Table 2 together with the optimal number of SNPs entering the respective model and the respective performance after hyperparameter optimization as mean *AUC*. Particularly noteworthy here are the very large differences in the optimal number of SNPs, which enter into the individual models as predictors. Thus, for the PRS with 50,633 SNPs almost all available SNPs are included in the final classification model, while in the final RF model only 1,357 SNPs are considered.

3.6 | Testing of models

The testing of the final classification models on the test data set D^*_{test} yielded the following AUC values and 95% confidence intervals:

1. Polygenic risk score: 0.9222, [0.9045, 0.94]

2. Support vector machine: 0.8228, [0.7972, 0.8484]
3. Naïve Bayes classifier: 0.8189, [0.7929, 0.8449]
4. Random forest: 0.7453, [0.7151, 0.7754]
5. Gradient boosting: 0.7399, [0.7092, 0.7707].

The corresponding ROC curves of the classification models on D^*_{test} are shown in Figure 4. Clearly, the classification model of the PRS dominates all other classification models. Also, the AUC confidence interval does not overlap with any confidence interval of the other models. The SVM model and the NB classifier model are nearly equivalent in terms of specificity and sensitivity for different thresholds. Accordingly, the AUC confidence intervals are almost completely superimposed. The same applies to the weakest models, the GB model and the RF model. Here, however, it can be observed that the model of the GB in the range of a specificity between approximately 0.45 and 0.75 has a slightly higher sensitivity than the RF model. At a specificity lower than 0.45, this reverses. The confidence intervals of the models of the SVM and the NB classifier show no overlap with the confidence intervals of the GB and RF models. These relationships between the individual models can also be seen by considering the Kendall correlation of the predicted case probabilities (Figure 5). Thus, there is a strong correlation between the ranks of the predicted values of the SVM model and the NB model of $\tau = 0.7$. The correlation between the predictive values of the RF and GB model is somewhat weaker with $\tau = 0.6$, but just as strong as between the PRS model and NB and SVM models. All other correlations are $\tau = 0.6$.

Figure 6 shows the distributions of the predicted probabilities for the class "CAD case," stratified according to the true CAD status for the final classification models. Noticeable is the distribution of the probabilities for the model of the NB classifier, which are strongly pushed to the edges 0 and 1 for both CAD status groups, which is due to the NB classifier attempting to maximize the a posteriori class probabilities. In contrast, the probability distributions of the other classification models tend to be more symmetrical. In addition, for the models of GB, SVM, and RF, the predicted class probabilities take values from extremely short intervals. Thus, the probabilities predicted by the GB model are in the range [0.5575, 0.587], those by the SVM model in the range [0.533149, 0.53315] and those by the RF model in the range [0.5058, 0.5546]. Only the models of the NB classifier and the PRS draw on the complete range [0, 1]. The distributions also show that the PRS model is best suited to discriminate observations between the two CAD status classes.

The limited ranges of predicted probabilities affect the search for an optimal threshold for dichotomizing the predictions. Figure 7 shows different quality measures depending on different thresholds for the final classification models. It can easily be seen that the PRS model is also better in other performance measures, such as mean misclassification and balanced accuracy, than the other models.

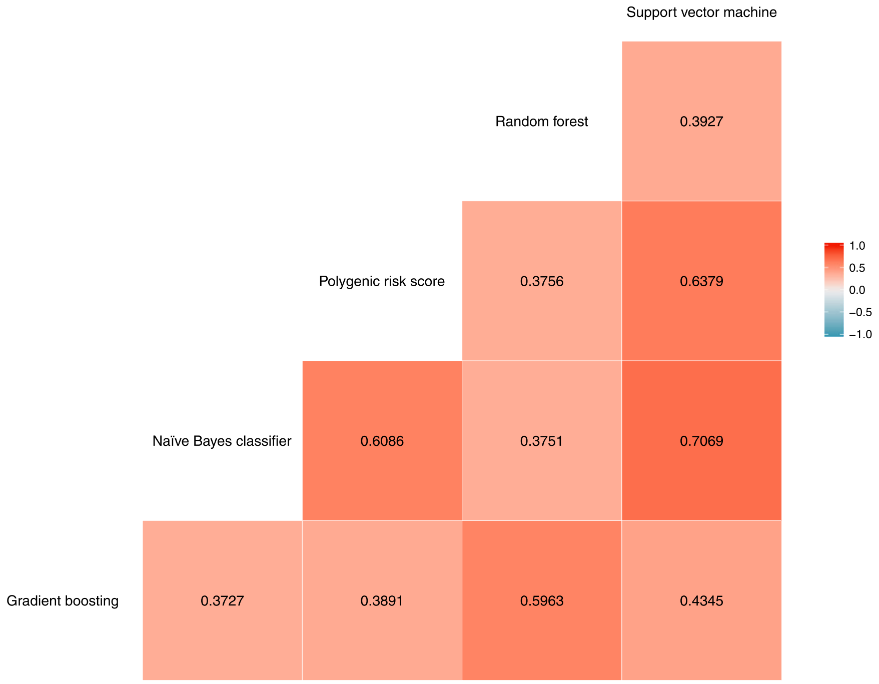

FIGURE 5 Kendall correlation of class probabilities between the classification models on the test data set D^*_{test}

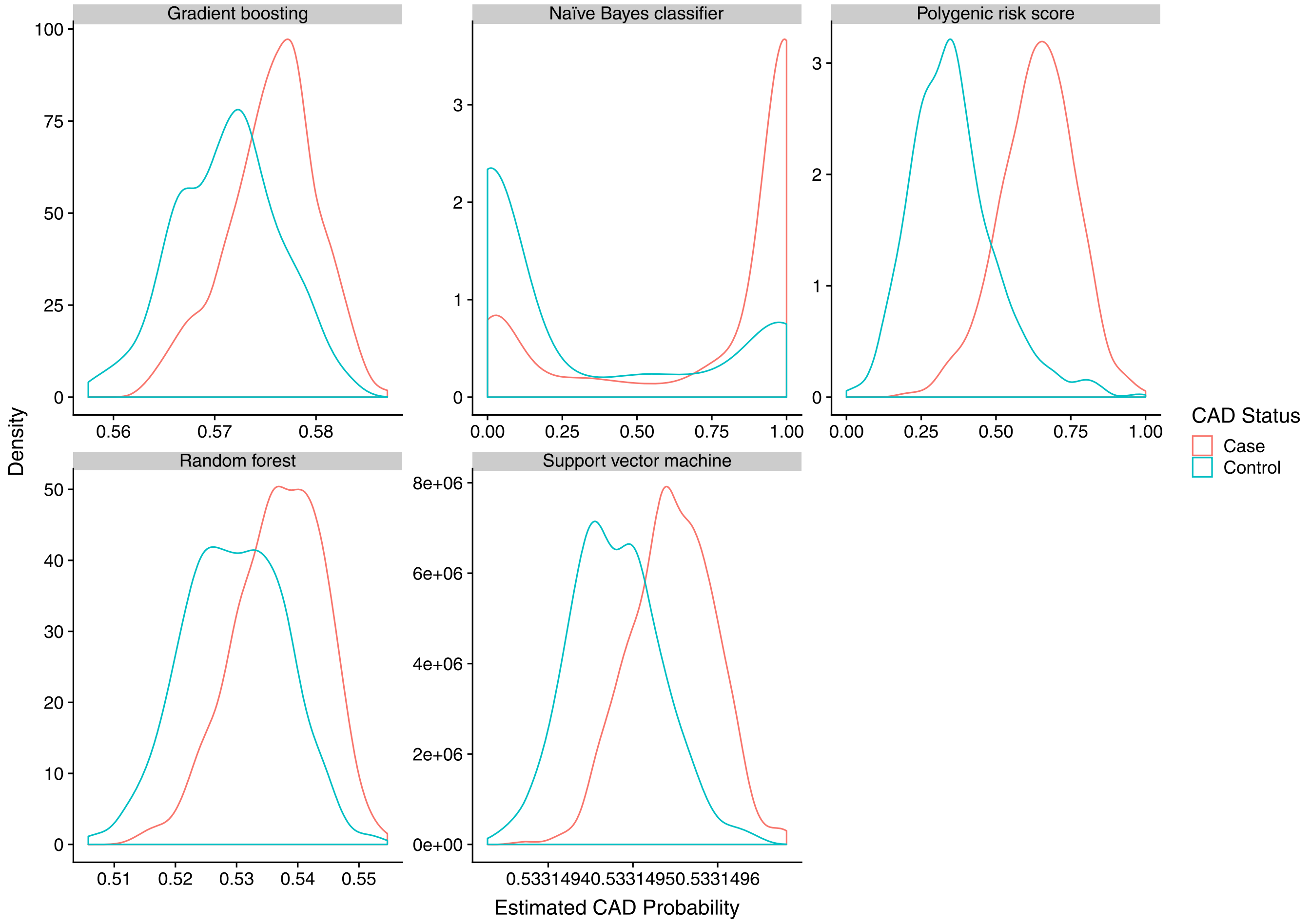

FIGURE 6 Distributions of the estimated case probabilities by true disease status in the test data set D^*_{test}

4 | DISCUSSION

In this study, classification models have been developed with the aim to discriminate individuals with CAD from healthy individuals by using genetic information only. For the application of the PRS, a GWAS was performed on the training data set to obtain weights for the individual SNPs. This was necessary because the available data sets contributed to most of the published GWAS results for CAD, and thus taking public effect estimates would have induced a bias when evaluating the classification performance.

To allow a fair comparison of the different classification algorithms, a benchmark was performed by means of nested CV on the training data. The clear winner was the simple PRS with a mean AUC of 0.9131. The more complex classification algorithms SVM, NB classifier, RF, and GB all achieved at least an AUC > 0.7. These five algorithms qualified for hyperparameter optimization and final modeling. The algorithms regularized regression and logistic regression achieved poor classification performance and were not considered further. For logistic regression, this is due to the poor convergence property of the algorithm: If more than approximately 850 predictors were included in the training, convergence of the algorithm was no longer possible. However, it is unclear why regularized regression fared similarly poorly.

After the optimization of the hyperparameters, the different numbers of SNPs used in the final models was striking. With 50,633 SNPs, almost all available predictors entered the final model of the PRS, while only 1,357 SNPs were used for the final RF model. One reason for this could be that the importance of SNPs was also determined by a RF. However, by design, other classification algorithms may use other variables to generate good predictions, so they will need to look at more SNPs for modeling until the variables specific to the algorithm are available. This reasoning is also supported by the fact that the second-lowest number of SNPs was used for the gradient-boosting model in which the basic models were also decision trees. Overall, far more SNPs are used for the prediction in all models than are significant in a GWAS. This supports an altered view on the underlying genetic mechanisms in complex diseases currently discussed in the literature. For example, Boyle, Li, and Pritchard (2017) suggest the “omnigenetic model” in

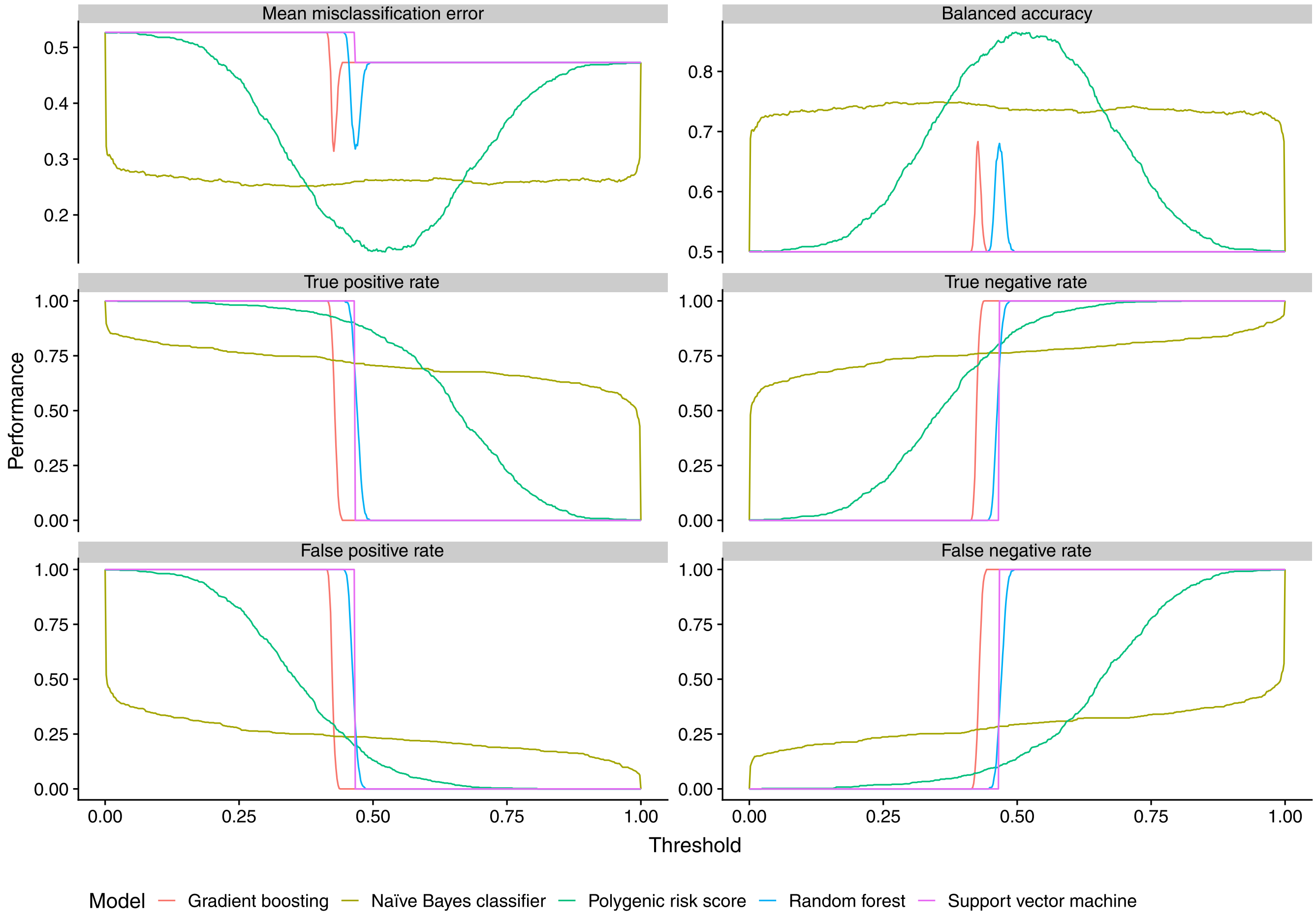

FIGURE 7 Performance measures depending on varying thresholds for transforming the predicted class probability into hard class predictions in the test data set D^*_{test}

which all genes that are expressed in disease-relevant cells ultimately influence the etiology of complex diseases through their influence on disease-relevant genes.

The testing of the five final models on a part of the combined data set as a test data set showed the same ranking of the algorithms as in the benchmark. The classification performance from testing corresponded approximately to the respective values from the benchmark. Thus, although we are lacking a sensitivity analysis in the sense of performing repeated splits of the entire data set into distinct D_{train} and D^*_{test} and re-evaluating the entire process, the benchmark performance values, as based on 10-fold CV, can be seen as an indicator of the performance in the testing data set. Furthermore, the highly restricted ranges of the predicted probabilities by the SVM, GB, and RF models indicate that these models are very poorly calibrated, that is the proportion of actual CAD cases does not match the predicted probabilities. In contrast, the predicted probabilities of the PRS model and the model of the NB classifier are on the entire [0, 1] interval. One reason for this is the behavior of the individual classification algorithms (Niculescu-Mizil & Caruana, 2005), especially in the use of the AUC as a criterion during hyperparameter optimization, which focuses exclusively on a good discrimination of the CAD disease status. However, a good calibration of predictive models alongside good discrimination is much more important for prognostic models than for diagnostic models (Cook, 2007). As the available data sets are retrospective case/control, only diagnostic models can be developed, so that this aspect is rather negligible for the classification models developed here.

Carrying out the calculation of the variable importance and GWAS outside of the CVs in the performance comparison and during the hyperparameter optimization can be viewed critically. Usually, all steps to determine variable weights should be performed within a CV to avoid over-optimistic estimates of model qualities (Bischl, Mersmann, Trautmann, & Weihs, 2012). However, the additional time required for this would have been enormous because the one-time calculation of the importance of the variables alone required 22 hr. Also, we assume that calculation of the variable importance outside the CVs leads only to slight distortions, since the

RF algorithm already uses internal bootstrapping. Since the final classification models were also tested on independent data sets, little overall distortion of the classification quality is to be expected. In addition, it was found that the AUC did not vary much for all classification models between performance comparison, hyperparameter optimization and testing on D^*_{test}.

More generally, it might be questioned whether the data used for the GWAS and training was, in the strictest sense, independent from the test data set, given that these samples were drawn from the same mix of study groups. Therefore, a truly external validation in data generated with a different technology, at another time point, in a different population is required to evaluate the absolute classification performance of the models. Interestingly, this is the case also for other published GPRS by Khera et al. (2018) and Inouye et al. (2018).

Other procedures might have been applied for the variable selection. As described above, the rank order determined by a RF model may have given higher ranks to variables that are important only for RF models, but less important for others and vice versa. One solution would have been to use methods that form any subset of the available variables. However, this would also entail a dramatically increased computing time.

Several different genotyping chips were used for the six original data sets. One of these studies used even different chips for their cases and controls. This has a tendency to produce false associations. However, the latter is true only for a small subset of the data set, thus this effect should be small.

The age distribution of the samples at hand would have been of interest, but unfortunately, detailed age information was not available. Thus, a bias could have specifically occurred if the controls were younger than the cases, thereby possibly being affected by CAD in later life. However, in this case we assume that the genetic effects would be underestimated, thus resulting in an underestimation of the discrimination performance.

It should be noted that the proportion of females is different between cases and controls. When building risk estimation models one might therefore want to include sex as an important predictor for the CAD phenotype. However, in this study we wanted to show how well the discrimination between cases and controls can be discriminated by just using genetic information, that is how to compress the genetic information in the most discriminative way and thus make the genetic information easily usable in risk estimation models in addition to other clinical and/or demographic variables, and to clarify the importance (or not) of using more complex statistical approaches.

Our study substantiates the findings by Inouye et al. (2018) and Khera et al. (2018) showing that GPRS can boost individual risk prediction of common diseases. Here, we come to the conclusion, that at least for prediction of CAD status there is no need to use a sledgehammer to crack the nut. All machine learning models considered in our work were substantially worse than a simple GPRS in compressing the genetic information in an information preserving way. It is possible that machine learning models might improve in performance if more samples would be available. However, up to this point it seems that the assumption of linear additive effects influencing the CAD disease status is sufficient for creating powerful risk prediction models based on GPRS and that other methods modeling nonlinear effects are not necessary.

ACKNOWLEDGMENTS

I. R. K. was supported by a grant from the Cluster of Excellence Inflammation at Interfaces (funded by the Deutsche Forschungsgemeinschaft), by the Deutsche Forschungsgemeinschaft (grant KO2250/7) and the German Center for Cardiovascular Research (funded by the Bundesministerium für Bildung und Forschung, grant 81Z1700103).

DATA AVAILABILITY STATEMENT

The genetic and phenotypic data that support the findings of this study are available from the respective authors given in Table 1. Restrictions apply to the availability of these data, which were used under license for this study. The result data that support the findings of this study are available from the corresponding author upon reasonable request. The code used to produce these results is available at https://github.com/imbs-hl/cad_classification.

ORCID

Damian Gola http://orcid.org/0000-0002-4980-4582
Inke R. König http://orcid.org/0000-0003-0504-6465

REFERENCES

Abraham, G., Havulinna, A. S., Bhalala, O. G., Byars, S. G., De Livera, A. M., Yetukuri, L., … Inouye, M. (2016). Genomic prediction of coronary heart disease. *European Heart Journal*, *37*(43), 3267–3278. https://doi.org/10.1093/eurheartj/ehw450

Andlauer, T. F. M., Buck, D., Antony, G., Bayas, A., Bechmann, L., Berthele, A., … Muller-Myhsok, B. (2016). Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation.

Science Advances, *2*(6), e1501678. https://doi.org/10.1126/sciadv.1501678

Beaney, K. E., Cooper, J. A., Drenos, F., & Humphries, S. E. (2017). Assessment of the clinical utility of adding common single nucleotide polymorphism genetic scores to classical risk factor algorithms in coronary heart disease risk prediction in UK men. *Clinical Chemistry and Laboratory Medicine*, *55*(10), 1605–1613. https://doi.org/10.1515/cclm-2016-0984

Bischl, B., Lang, M., Richter, J., Bossek, J., Judt, L., Kuehn, T., … Jones, Z. M. (2016). mlr: Machine learning in R. *Journal of Machine Learning Research*, *17*(170), 1–5. Retrieved from http://jmlr.org/papers/v17/15-066.html; http://cran.r-project.org/package=mlr

Bischl, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation*, *20*(2), 249–275. https://doi.org/10.1162/EVCO_a_00069

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., & Lang, M. (2017). mlrMBO: A modular framework for model-based optimization of expensive black-box functions. Retrieved from http://arxiv.org/abs/1703.03373

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell*, *169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Brænne, I., Willenborg, C., Tragante, V., Kessler, T., Zeng, L., Reiz, B., … Schunkert, H. (2017). A genomic exploration identifies mechanisms that may explain adverse cardiovascular effects of COX-2 inhibitors. *Scientific Reports*, *7*(1), 10252. https://doi.org/10.1038/s41598-017-10928-4

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., … Li, Y. (2018). *xgboost: Extreme gradient boosting*. Retrieved from https://cran.r-project.org/package=xgboost

Cook, N. R. (2007). Statistical evaluation of prognostic versus diagnostic models: Beyond the ROC curve. *Clinical Chemistry*, *54*(1), 17–23. https://doi.org/10.1373/clinchem.2007.096529

Delaneau, O., Zagury, J.-F., & Marchini, J. (2012). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, *10*(1), 5–6. https://doi.org/10.1038/nmeth.2307

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*(3), 837. https://doi.org/10.2307/2531595

Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T. L., Thompson, J. R., … Samani, N. J. (2012). Large-scale association analysis identifies new risk loci for coronary artery disease. *Nature Genetics*, *45*(1), 25–33. https://doi.org/10.1038/ng.2480

Erdmann, J., Großhennig, A., Braund, P. S., König, I. R., Hengstenberg, C., Hall, A. S., … Schunkert, H. (2009). New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nature Genetics*, *41*(3), 280–282. https://doi.org/10.1038/ng.307

Erdmann, J., Willenborg, C., Nahrstaedt, J., Preuss, M., Konig, I. R., Baumert, J., … Schunkert, H. (2011). Genome-wide association study identifies a new locus for coronary artery disease on chromosome 10p11.23. *European Heart Journal*, *32*(2), 158–168. https://doi.org/10.1093/eurheartj/ehq405

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, *28*(2), 337–407. https://doi.org/10.1214/aos/1016218223

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York, NY: Springer. https://doi.org/10.1007/978-0-387-84858-7

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genetics*, *5*(6), e1000529. https://doi.org/10.1371/journal.pgen.1000529

Inouye, M., Abraham, G., Nelson, C. P., Wood, A. M., Sweeting, M. J., Dudbridge, F., … Samani, N. J. (2018). Genomic risk prediction of coronary artery disease in 480,000 adults. *Journal of the American College of Cardiology*, *72*(16), 1883–1893. https://doi.org/10.1016/j.jacc.2018.07.079

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab—An S4 package for Kernel methods in R. *Journal of Statistical Software*, *11*(9), https://doi.org/10.18637/jss.v011.i09

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1–2), 81–93. https://doi.org/10.1093/biomet/30.1-2.81

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., … Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, *1*, https://doi.org/10.1038/s41588-018-0183-z

Khera, A. V., & Kathiresan, S. (2017). Genetics of coronary artery disease: Discovery, biology and clinical translation. *Nature Reviews Genetics*, *18*(6), 331–344. https://doi.org/10.1038/nrg.2016.160

Kolz, M., Baumert, J., Gohlke, H., Grallert, H., Döring, A., Peters, A., … Illig, T. (2009). Association study between variants in the fibrinogen gene cluster, fibrinogen levels and hypertension: Results from the MONICA/KORA study. *Thrombosis and Haemostasis*, *101*(02), 317–324. https://doi.org/10.1160/TH08-06-0411

Krarup, N., Borglykke, A., Allin, K., Sandholt, C., Justesen, J., Andersson, E., … Hansen, T. (2015). A genetic risk score of 45 coronary artery disease risk variants associates with increased risk of myocardial infarction in 6041 Danish individuals. *Atherosclerosis*, *240*(2), 305–310. https://doi.org/10.1016/j.atherosclerosis.2015.03.022

Krawczak, M., Nikolaus, S., Eberstein, H., von, Croucher, P. J., El Mokhtari, N. E., & Schreiber, S. (2006). PopGen: Population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Public Health Genomics*, *9*(1), 55–61. https://doi.org/10.1159/000090694

Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., & Murray, C. J. (2006). Global and regional burden of disease and risk factors, 2001: Systematic analysis of population health data. *Lancet*, *367*(9524), 1747–1757. https://doi.org/10.1016/S0140-6736(06)68770-9

Marenberg, M. E., Risch, N., Berkman, L. F., Floderus, B., & Faire, U. d. (1994). Genetic susceptibility to death from coronary heart disease in a study of twins. *New England Journal of Medicine*, *330*(15), 1041–1046. https://doi.org/10.1056/NEJM199404143301503

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., … Marchini, J. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2017). *e1071:* Misc functions of the department of statistics, probability theory group (Formerly: E1071), TU Wien. Retrieved from https://cran.r-project.org/package=e1071

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? *BMC Bioinformatics*, *34*(21), 3711–3718. https://doi.org/10.1093/bioinformatics/bty373

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of 22nd international conference of machine learning—ICML '05* (pp. 625–632). New York, NY: ACM Press. Retrieved from https://doi.org/10.1145/1102351.1102430

Ripatti, S., Tikkanen, E., Orho-Melander, M., Havulinna, A. S., Silander, K., Sharma, A., … Kathiresan, S. (2010). A multilocus genetic risk score for coronary heart disease: Case-control and prospective cohort analyses. *Lancet*, *376*(9750), 1393–1400. https://doi.org/10.1016/S0140-6736(10)61267-6

Samani, N. J., Erdmann, J., Hall, A. S., Hengstenberg, C., Mangino, M., Mayer, B., … Schunkert, H. (2007). Genomewide association analysis of coronary artery disease. *New England Journal of Medicine*, *357*(5), 443–453. https://doi.org/10.1056/NEJMoa072366

Smith, G. D., Ebrahim, S., Lewis, S., Hansell, A. L., Palmer, L. J., & Burton, P. R. (2005). Genetic epidemiology and public health: Hope, hype, and future prospects. *Lancet*, *366*(9495), 1484–1498. https://doi.org/10.1016/S0140-6736(05)67601-5

Tada, H., Melander, O., Louie, J. Z., Catanese, J. J., Rowland, C. M., Devlin, J. J., … Shiffman, D. (2016). Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history. *European Heart Journal*, *37*(6), 561–567. https://doi.org/10.1093/eurheartj/ehv462

Thomsen, T. (2005). HeartScore: A new web-based approach to European cardiovascular disease risk management. *European Journal of Cardiovascular Prevention And Rehabilitation*, *12*(5), 424–426. https://doi.org/10.1097/01.hjr.0000186617.29992.11

Wellcome Trust Sanger Institute. (n.d.). Sanger Imputation Service. Retrieved from https://imputation.sanger.ac.uk

Wichmann, H.-E., Gieger, C., & Illig, T. (2005). KORA-gen—Resource for population genetics, controls and a broad spectrum of disease phenotypes. *Das Gesundheitswes*, *67*(S 01), 26–30. https://doi.org/10.1055/s-2005-858226

Wilson, P. W. F., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, *97*(18), 1837–1847. https://doi.org/10.1161/01.CIR.97.18.1837

Winkelmann, B. R., März, W., Boehm, B. O., Zotz, R., Hager, J., Hellstern, P., & Senges, J. (2001). Rationale and design of the LURIC study—A resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease. *Pharmacogenomics*, *2*(1s1), S1–S73. https://doi.org/10.1517/14622416.2.1.S1

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), https://doi.org/10.18637/jss.v077.i01

Wright, M. N., Ziegler, A., & König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics*, *17*(1), 145. https://doi.org/10.1186/s12859-016-0995-8

Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., … Lisheng, L. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART Study): Case-control study. *Lancet*, *364*(9438), 937–952. https://doi.org/10.1016/S0140-6736(04)17018-9

SUPPORTING INFORMATION

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

How to cite this article: Gola D, Erdmann J, Müller-Myhsok B, Schunkert H, König IR. Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. *Genet. Epidemiol.* 2020;44:125–138. https://doi.org/10.1002/gepi.22279