Development and Validation of a Polygenic Risk Score for Coronary Artery Disease

George Busby PhD, Alessandro Bolli PhD, Paolo Di Domenico, Giordano Botta PhD

INTRODUCTION

Advances in genetic research are leading to improved precision medicine in healthcare

An individual's risk of developing disease results from a complex interaction between their environment and their genes. The collection and analysis of large amounts of data is now allowing us to understand these relationships in more detail than ever before and are aiding the development of data-driven approaches to risk prediction that will transform the way that healthcare is provided [1]. From the electronisation of health records to the introduction of wearable devices, developments in digitisation are allowing the collection of biomedical big data at an unprecedented scale.

At the same time, methods involving advanced analytics like **machine learning** are leading to a greater understanding of these data. The integration of these concepts into healthcare systems is paving the way for a new era of **precision medicine**, the goal of which is to use data to identify those at highest risk of developing disease so that interventions and treatments can be targeted at the groups that need them most. This will ensure that finite healthcare resources are used as efficiently as possible and that disease is caught early enough to improve patient outcomes.

Precision medicine is made increasingly possible because we now have the analytical methods and computational power necessary to understand big, complex datasets. These datasets contain clinical outcomes on a large number of people for a variety of diseases and include additional information on physiological, lifestyle and environmental factors such as age, sex and family history of disease. When these data are combined and analysed, sophisticated statistical models can find patterns in the combinations of risk factors that lead to disease, which can provide an estimate of an individual's risk relative to others in their population with similar values across these factors.

As we understand more about the role that **DNA** has to play in understanding disease risk, it is becoming increasingly important to include genetic or genomic information in these datasets. This is because progress in **genomics** means that this novel and potentially powerful type of data can be added into disease risk prediction models. The use of genomic data in risk prediction is now possible thanks to the confluence of three major scientific advances:

1. Bigger sample sizes in genome-wide association studies (GWAS) which lead to greater statistical power to identify genetic variants involved with disease.
2. Better statistical methods to identify the most predictive set of variants to estimate risk for a given disease.

3. The availability of genome-wide data from hundreds of thousands of people linked with thousands of environmental and physiological measurement in the UK Biobank [2]. This magnificent data resource allowed the validation of the algorithms predictive power at an unprecedented scale.

Developing a robust procedure for estimating a CAD PRS

In its essence, computing an individual's polygenic risk score (PRS) for a given disease is straightforward. By multiplying the number of risk alleles a person carries by the effect size of each variant and then summing these across all risk loci [3] one can estimate an individual's genetic liability for a given disease. However, identifying the alleles at loci that confer disease risk, estimating the size of the allele's effect on disease and incorporating estimates of uncertainty in a PRS remaining challenging. Furthermore, the accuracy of a PRS depends on several conditions [4].

The first is that the GWAS providing the summary statistics for the score – known as the **discovery** set – should involve an independent set of samples to those on which the scores are being calculated. Secondly, the amount of variation in disease or trait liability that can be accounted for by the genetic variants used in the PRS, known as SNP **heritability**, will influence how predictive a PRS will be, which will also be affected by the **genetic architecture** of the disease. Finally, the sample size of the discovery GWAS will affect how well effect sizes are estimated and therefore in turn affect its accuracy. The best performing PRSs use summary statistics from a discovery GWAS involving hundreds of thousands of independent individuals on a trait with high SNP heritability.

Identifying which SNPs have the best predictive power is a central challenge to developing a robust PRS. There are two main objectives to this effort. The first is to understand at what **threshold** of statistical significance SNPs should be removed from the score generation algorithm. Because there are correlations between the effect sizes of variants that are close to each other in the genome, the second objective is to explore how to combine evidence across multiple variants.

Fortunately, procedures have been developed to select subsets of SNPs that rely on looking only at **GWAS summary statistics** [5]. The simplest approach, known as **clumping and thresholding** (C+T), iterates between two methodological steps [6]. First, genetic variants are filtered, or clumped, so that only the variants with the highest effect size and that are not in **linkage disequilibrium** (*LD*) are used. In the second thresholding step, genetic variants with a *P* value larger than a chosen threshold are removed. This process is repeated for different *LD* windows and *P* value thresholds.

A more sophisticated **Bayesian** approach involves modelling *LD* to shrink each variant effect size to an extent that is proportional to the *LD* between SNPs [7]. This approach, implemented in the software *LDPred*, requires the definition of a tuning parameter ρ, which is a statement of the researcher's prior belief on the proportion of genetic variants assumed to be causal.

Recently, a third method involving machine learning that combines C+T and the **LASSO** statistical procedure, called **stacked clumping and thresholding** (SCT) has been developed [8].

In SCT, clumping and thresholding are systematically repeated over a four dimensional grid of parameters (comprising LD squared correlation and p-value thresholds). The algorithm generates over 100,000 alternative C+T variants and combines them through a LASSO-based penalized logistic regression.

Validating and testing PRS is possible with the availability of Biobanks

The algorithms outlined above require a **validation phase** where different PRSs generated with alternative parameter values are validated against an external dataset (Validation dataset). The **test phase** involves computing PRS in a test population (Test dataset) and assessing its **predictive power** in order to confirm its predictive performance and to rule out any possibility of **over-fitting** that may have occurred during the validation step.

Development of PRSs for a number of diseases has been accelerated by the availability of the UK Biobank (UKB) dataset. This is a large **prospective cohort study** that enrolled around 500,000 individuals from across the UK, ranging in age from 40 to 69 years at the time of recruitment and whose genomes have been genotyped and imputed to more than 90 million variants. The astonishing size of the UKB genomic data means that researchers can no build suitably sized Validation and Testing datasets.

Clinical utility and implications of CAD PRS use in the European population

Several studies have assessed the ability of PRS to identify individuals at high risk of developing polygenic diseases. For example, Inouye and colleagues [9] showed that men in the top 20% of PRS distribution reached a threshold of 10% cumulative coronary artery disease (CAD) risk by 61 years of age, ten years earlier than men in the bottom 20% distribution.

Additionally it has been shown that CAD PRS has higher predictive performance than a range of the traditional risk factors (e.g. total cholesterol, family history of heart disease) used by physicians to decide primary prevention strategies [10]. In a second study, PRS-based models identified 8.0, 6.1, 3.5, 3.2, and 1.5% of the European population at greater than threefold increased risk for Coronary Artery Disease (CAD), Atrial Fibrillation (AT), Type 2 Diabetes (T2D), Inflammatory Bowel Syndrome (IBD), and Breast Cancer (BC), respectively. Most notably for CAD, the prevalence of been carrier of high PRS was shown to be 20-fold higher than the prevalence of carriers of the familial hypercholesterolemia mutations while conferring the same risk.

Current technological limitations in using PRS

Generating PRS is computationally intensive and so their potential to be used as a tool for precision medicine is currently undervalued. The computers required to generate PRS need hundreds of Gigabytes of RAM and complex computational infrastructures which are extremely difficult to implement and maintain. Additionally, deep bioinformatics expertise is required to run the entire pipeline, from generating genomic data, through quality control to result visualisation. For this reason, analytical laboratories are currently excluded from the possibility to use PRS on routine basis.

PRSs are constructed on the basis of SNPs and their effect sizes discovered through GWASs. One limitation of GWAS is that they are performed with samples genotyped with microarrays that do not cover the entire genome, but only a small portion of it. Therefore the causal SNP associated with a phenotype is rarely genotyped, instead the association is attributed to the genotyped SNP in LD with the causal one. However, different ethnic groups are characterised by specific LD patterns, so we can expect that for different ancestries the causal SNP could have a different SNP in LD. For this reason, the SNPs used in a PRS are highly dependent on the genetic structure (i.e., ancestry) of the initial population of the training GWAS. Since the vast majority of available GWAS is based on population of European descendent (79% of all GWAS partecipants), PRS constructed on these GWAS have the highest predictive power on individuals of the same ancestry. This represents the most critical limitation to genetics in precision medicine and increasing the representation of diverse populations has recently become a higher priority for the research community.

A NEW POLYGENIC RISK SCORE FOR CORONARY ARTERY DISEASE

CAD is a disease caused by the narrowing or blockage of the coronary arteries and is usually caused by atherosclerosis, a hardening of the arteries. Whilst environmental and lifestyle factors can modulate an individual's risk of getting CAD, there is also a genetic component, making a perfect candidate for the investigation of the how well PRS can predict disease.

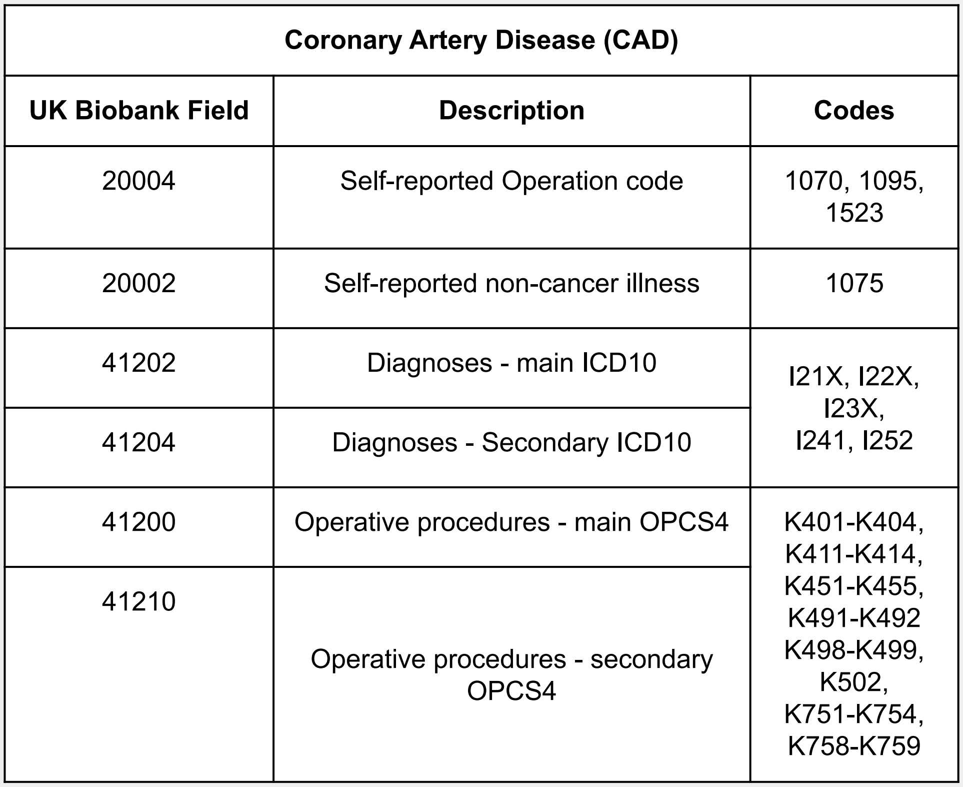
Validation and testing of a new PRS for CAD

To build a new PRS, we used the SCT algorithm of Privè and colleagues [8], implemented in *R* [11]. The algorithm uses summary statistics from a published GWAS [12] and genetic and clinical data from a Validation dataset. For this we used the interim release of UK Biobank (i.e. individuals genotyped through batches from 1 to 22 (**Table 1**)). The UK Biobank **fields** we used to define cases of CAD are reported in **Table 2**. SCT uses per-SNP effect sizes and *P* values to perform repeated clumping and thresholding (C+T) over a four dimensional grid of parameters (comprising *LD* squared correlation, *P* value threshold, clumping window size, and imputation quality).

Cohort	N total	N CAD (prevalent | incident)
Validation dataset	129,853	7,912 (4,962 | 3,220)
Testing dataset	130,253	2,268 (0 | 2,268)

Table 1: Samples used to build the new CAD PRS. Here we show the makeup of the Discovery and Validation datasets used to build and test the CAD PRS. Note that individuals with a history of CAD at baseline were removed from the Validation dataset to allow the PRS to identify only those with incident CAD.

Overall, SCT generates 123,200 alternative C+T configurations, each of which is used to compute a corresponding per-individual PRS in the validation population. Per-individual PRS scores were generated as the sum of the **genotype dosage** of each risk allele at each SNP weighted by its respective effect size. These PRS scores are then used as the predictive variables in a LASSO-penalized logistic regression model with disease phenotype as the binary response variable, generating a regression coefficient for each C+T configuration. The stacking phase follows, where effect sizes and regression coefficients of the 123,200 alternative C+T configurations are linearly combined to generate, for each disease, a final optimal panel of SNPs for the PRS. We define a PRS panel as a 3-column table of SNPs, effect alleles, and corresponding effect sizes, which were taken from the GWAS of Nikpay et al [12].



Coronary Artery Disease (CAD)		
UK Biobank Field	**Description**	**Codes**
20004	Self-reported Operation code	1070, 1095, 1523
20002	Self-reported non-cancer illness	1075
41202	Diagnoses - main ICD10	I21X, I22X, I23X, I241, I252
41204	Diagnoses - Secondary ICD10	
41200	Operative procedures - main OPCS4	K401-K404, K411-K414, K451-K455, K491-K492 K498-K499, K502, K751-K754, K758-K759
41210	Operative procedures - secondary OPCS4	

Table 2: Lists of UK Biobank fields and codes to define cases of Coronary Artery Disease.

Finding the PRS with the best predictive performance

We assessed the predictive performance of the new PRSs on an independent Testing dataset comprising the second release of UK Biobank (i.e. individuals genotyped in batches from 23 to 95; **Table 1**). For comparison, we also re-computed the published PRS panels from Khera [6] and Inouye [9] using summary statistics downloaded from the Broad Institute Cardiovascular Disease knowledge portal*.

*http://www.broadcvdi.org/informational/data

Specifically, we used each PRS panel to calculate a per-individual PRS score in the testing population. Per-individual PRS values were used as predictive variables in a logistic regression model, where the response variable (the variable to predict) was the disease phenotype. The logistic regression model comprised additional covariates as control variables such as: age, gender, genotyping array, and the first 4 principal components (PCs) of ancestry. We assessed the predictive performance of each PRS panel by computing the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve.

Using variants with small effect sizes improves the predictive performance of CAD PRS

Two PRS for CAD have recently been published. Khera and colleagues[6] used the *LDpred* algorithm [7] to develop a CAD PRS using 6.6 million SNPs, while Inouye and colleagues[9] aggregated three PRS for CAD to generate a *metaPRS* using around 1.7 million SNPs. With *LDpred*, Khera and colleagues obtained a PRS with the best predictive performance with a parameter value that indicates that 0.1% of the variants in the analysis are causal. This implies that only the 0.1% of the 6.6 million SNPs in the PRS should have an effect on the prediction, while the remaining 99.9% have an effect size close to zero (of about 3x10^{6}).

Given that in this analysis the vast majority of SNPs are estimated to have minimal or no effect on polygenic risk, doubt has been cast on the utility of including this large fraction of low-weight SNPs in PRSs [13]. We addressed the effect of including different numbers of SNPs in a PRS by recreating the Khera PRS and comparing predictive power across progressively smaller subsets of SNPs (**Table 3**): the full PRS (6.6 million SNPs), a PRS made by the top 1% of SNPs with highest effect sizes (66,300 SNPs), a second PRS constituted by the top 0.1% of SNPs with highest effect sizes (6630 SNPs), as well as a PRS generated with genome-wide significant SNPs only (*P* value < 5x10−^{8}, corresponding to 74 SNPs). AUC, or Area Under the Curve, a quantitative measure for the discrimination abilility of a PRS [14], and Positive Predictive Value (PPV) at 3% (i.e the percentage of CAD cases identified at the top 3% of PRS distribution) were calculated for each PRS in the testing dataset. A decrease in the number of SNPs used in a PRS is matched by a corresponding decrease in its discriminatory ability in both AUC and PPV (**Table 4**).

PRS Panel	SNPs in PRS	AUC (95% CI)	PPV (3%)	Cases in top 3%
Khera full 1,031	6,630,150	0.805 (0.798 0.812)	12.35	1,031
Inouye full	1,745,180	0.805 (0.799 0.812)	12.5	1,044
SCT	291,969	0.808 (0.8 0.815)	13.06	1091
SCT+ khera $^{(1)}$	6,699,370	0.808 (0.801 0.814)	12.99	1,085
SCT+ Inouye $^{(1)}$	1,920,136	0.810 (0.803 0.816)	13.36	1,116

Table 3: List of the PRS for CAD assessed in this study. Khera full refers to the whole PRS for cad developed by Khera et al . Inouye full refers to the whole PRS for cad developed by Inouye et al . SCT refers to the PRS for CAD developed in this paper with the SCT algorithm. SCT + Khera refers to the PRSs generated by combining SCT and full Khera PRS $^{(1)}$.

(Table 3 cont.) SCT + Inouye refers to the PRSs generated by combining SCT and full Inouye PRS[(1)]. For each PRS, **Table 3** shows the number of genetic variants composing the PRS (**SNPs in PRS**), the predictive performances quantified as AUC values and 95% confidence intervals (**AUC (95% CI)**), the positive predictive values in the top 3% of PRS distributions (**PPV (3%)**), and the number of CAD cases in the top 3% of PRS distributions (**Cases in top 3%**). [(1)]For SNPs with effect sizes from both SCT and the second panel (khera or Inouye), effect sizes from SCT were taken.

This finding demonstrates that even a set of low weight SNPs can play a crucial role for a more accurate reclassification of individuals into the higher-risk CAD category.

PRS Panel	SNPs in PRS	AUC (95% CI)	PPV (3%)	Cases in top 3%
Khera full	6,630,150	0.805 (0.798 0.812)	12.35	1,031
Khera 1%	66,300	0.798 (0.792 0.805)	11.31	945
Khera 0.1%	6,630	0.794 (0.788 0.801)	10.88	909
Khera 74	74	0.789 (0.797 0.784)	9.63	804

Table 4: List of the PRS for CAD assessed in this study. Khera full refers to the whole PRS for CAD developed by Khera et al[6]. Khera 1% refers to the PRS generated with the 1% of genetic variants with highest effect sizes from Khera et al . Khera 0.1% refers to the PRS generated with the 0.1% of genetic variants with highest effect sizes from Khera et al . Khera 74 refers to the PRS generated with genome-wide significant SNPs only as described in Khera et al. For each PRS, **Table 2** shows the number of genetic variants composing the PRS (**SNPs in PRS**), the predictive performances quantified as AUC values and 95% confidence intervals (**AUC (95% CI)**), the positive predictive values in the top 3% of PRS distributions (**PPV (3%)**), and the number of CAD cases in the top 3% of PRS distributions (**Cases in top 3%**).

Development of a new CAD PRS with improved predictive performance

We next assessed the predictive performance of our new PRS in the test dataset. The new SCT PRS had higher predictive performance (AUC: 0.808, PPV at 3%: 13.06%) than the PRSs from Khera et al [6] (AUC: 0.805, PPV at 3%: 12.35%) and Inouye et al[9] (AUC: 0.805, PPV at 3%: 12.5%; **Table 3**). Notably, the final CAD PRS we developed using SCT was composed of only c.300,000 genetic variants, a number that corresponds to only the 5% and 17% of the SNPs of CAD PRS from Khera and Inouye, respectively.

In light of our finding outlined above, that a larger number of SNPs - even if with low effect sizes - improve the predictive performance of a PRS, we asked whether integrating the large SNP sets from the Khera or Inouye studies to our CAD PRS could further improve its predictive performance.

We found that the addition of SNPs from Inouye to the SCT PRS led to a new CAD PRS (denoted as SCT + Inouye or SCT-I) with further improved predictive performance, as quantified by an increased value of the AUC (0.81) and the PPV at 3% (13.36%; **Table 3**). This finding highlights the highly polygenic nature of this common disease.

Predictive performance of the newly developed CAD PRS: SCT-I

For the remaining analyses we used the new SCT-I CAD PRS which we showed to have the highest predictive performance (**Table 3**). We computed the PRS of the individuals in the Testing dataset and plotted the distributions of the scores for CAD cases and controls (**Figure 1A**). The distributions are both gaussian, with CAD cases showing a greater median PRS than controls (median: 0.52 and -0.03, respectively) and an AUC of 0.81.

We next evaluated the ability of the SCT-I PRS to stratify CAD risk separately for sub-populations of men and women in the testing dataset. We divided the two PRS distributions into percentiles and computed the prevalence of CAD in each percentile. Here we use disease prevalence in the test dataset as a measure of the risk of developing CAD. Risk stratification for men and women in the test dataset are shown in **Figure 1C** and **D**, respectively. CAD risk rises sharply as PRS percentile increases, ranging from 1.34% to 25.67% (for men) and from 0.26% to 8.62% (for women), for the lowest and highest percentiles, respectively. As previously shown, men have a higher CAD risk than women[9].

For each sex we estimated the relative increased risk, which is the ratio between the prevalence at the top 5% of the PRS distribution and the prevalence in the average of the distribution (defined as between the 40% and the 60% percentiles, dashed lines in **Figure 1C** and **1D**). For men, the relative risk in the top 5% is 3 times higher than the average while for women this value rises to 4. This means that the CAD SCT-I PRS is able to detect individuals with a three fold relative risk of developing CAD which is comparable to that conferred by rare highly penetrant familial hypercholesterolemia mutations [15].

Above, we showed how the SCT-I PRS can stratify the empirical risk of CAD in a test population with known disease prevalence. However, the clinical value of a PRS is in its ability to *predict the risk of disease*. To assess the ability of the SCT-I PRS to predict CAD risk in the Testing dataset we compared the predicted prevalence values with observed ones. For each individual within the Testing dataset, the probability of having the disease was calculated using a logistic regression model with the per-individual PRS score as predictive variable. The predicted prevalence of CAD within each percentile of the PRS distribution was calculated as the average probability in each percentile. For all percentiles, predicted CAD prevalence was plotted against the corresponding values of observed prevalence (**Figure 1B**). Values of observed and predicted CAD prevalence are in excellent agreement as demonstrated by the localization of the points of the bisector in **Figure 1B**. We also tested the level of agreement between the predicted and observed prevalence through the Hosmer-Lemeshow (HL) test. This is a goodness of fit test for logistic regression, especially for risk prediction models. Specifically, the HL test calculates if the observed prevalence matches the predicted prevalence in population subgroups represented by PRS percentiles. The non-signicant p-value generated by the HL test (**Figure 1B**) implies that there is no statistical evidence of a deviation between observed and predicted prevalence values, thus confirming the good fit of the calibration that can be observed in (**Figure 1B**).

PRS is more effective at predicting CAD risk than family history

Family history of CAD is a well-recognized risk factor and prospective studies demonstrate a consistent association with the disease [16]. Family history can be easily and systematically queried in the clinical setting. As such, current prevention guidelines recommend that family history to be incorporated into the risk estimation process that guides treatment decisions [17]. In this section, we consider the relationship between two risk factors for CAD: family history and PRS. In particular we want to answer the following questions:

1. Can SCT-I PRS stratify risk in people with family history?
2. Is SCT-I PRS a better predictor than family history?
3. Does prediction performance increase if a combination of family history and PRS is Used?

We computed scaled per-individual distributions for CAD cases and controls for those individuals in the test dataset with at least one first-degree relative with a history of CAD. Risk distributions for cases and controls are shown in **Figure 2A**. As before, both distributions are gaussian with cases having a higher median value than controls (median: 0.60 and 0.08, respectively). This suggests that the good discriminatory ability of the SCT-I PRS is maintained even in individuals already considered at CAD risk based on family history.

We then evaluated the ability of the SCT-I PRS to stratify CAD risk in the sub-populations of men and women with at least one first-degree relative with a history of heart disease. As before, we calculated for both men and women the per-individual PRS distributions and CAD prevalence for each percentile of the PRS distributions. CAD risk stratification for men and women with family history of heart disease are shown in **Figure 2C** and **D**, respectively. Even with individuals considered at higher risk based on family history, the SCT-I PRS is able to further stratify CAD risk over a range of values comprised between 2.10% and 33% (for men) and between 0.56% and 10% (for women), for the lowest and highest percentiles, respectively. For both men and women, observed prevalence is higher in individuals with family history than in the general population for any percentile considered. For men with family history, the relative risk in the top 5% is 3 fold higher than the average while for women this value rises to 4.

Lastly, we assessed the predictive performance of family history, PRS, and the combination of the two, by computing AUC. **Figure 2B** shows that PRS displays a higher AUC value than family history and it is therefore a better predictor of CAD disease. When both risk factors are combined, the predictive performances further improves.

These findings demonstrate that family history and PRS capture different components of the risk of CAD and family history cannot be considered in isolation without further PRS risk stratification.

SCT-I for CAD is a genetic risk factor that is independent and orthogonal to other clinical risk factor for CAD such as LDL-cholesterol

Blood levels of LDL-cholesterol are commonly used by clinicians to assess risk of CAD. PRS and LDL-cholesterol levels are both continuous risk factors that can be used to stratify the risk of CAD in a sample population. Here, we compare the performance of the SCT-I PRS and LDL-cholesterol blood levels in risk stratification and prediction (**Figure 3A**). We calculated the two risk gradients as follows. PRS-based CAD stratification has been obtained by determining the per-individual PRS distribution and CAD prevalence for each percentile of the PRS distribution. LDL-cholesterol-based risk stratification has been calculated by using the UK Biobank LDL-cholesterol levels for each individual in the test population and dividing the corresponding distribution in percentiles. For each percentile of each distribution, the prevalence of CAD has been calculated and taken as a proxy of CAD risk. Of note, LDL-cholesterol levels for individuals reported to use cholesterol-lowering medications have been corrected by multiplying LDL-cholesterol values by a correction factor of 1.56 [18].

With increasing values of the percentile, the empirical risk of CAD (green circles in **Figure 3A**) rises more sharply for PRS than for LDL-cholesterol (yellow circles in **Figure 3A**), from 0.79% to 16.31% (PRS) and from 4.3% to 9.27% (for LDL-cholesterol), for the lowest and highest percentiles, respectively. Moreover, the relative risk of CAD, calculated as the ratio between the prevalence at the top 5% and at the average of the distribution (defined by the 40% and the 60% percentiles) is much higher for PRS than for LDL-cholesterol: 3.1 and 1.8, respectively. This finding demonstrates that PRS has a CAD risk stratification power that is 1.7 times higher than that of LDL-cholesterol.

We then assessed whether PRS and LDL-cholesterol are independent risk factors. **Figure 3B** shows, for each individual of the testing population, the correspondence between its percentile value in the PRS and in the LDL-cholesterol distributions (yellow dots). It is not possible to distinguish any clear pattern of correlation, except for a slight increase in point density in the upper right and bottom left corners. The absence of a clear correlation between the risk factors is confirmed by the roughly zero slope of the linear model (yellow line in **Figure 3B**), as well as by the very small correlation (Pearson's correlation coefficient ρ= 0.127).

This finding demonstrates that PRS and LDL-cholesterol levels are orthogonal risk factors that capture different clinical and genetic components of the risk for CAD.

Lastly, we assessed the predictive performances of LDL-cholesterol (LDL-C), PRS (PRS), and the combination of the two (LDL-C+PRS), by means of AUC. **Figure 3C** shows that PRS displays an higher AUC value than LDL-cholesterol and it is therefore a better predictor of CAD disease. When both risk factors are combined, the predictive performances further improves only slightly.

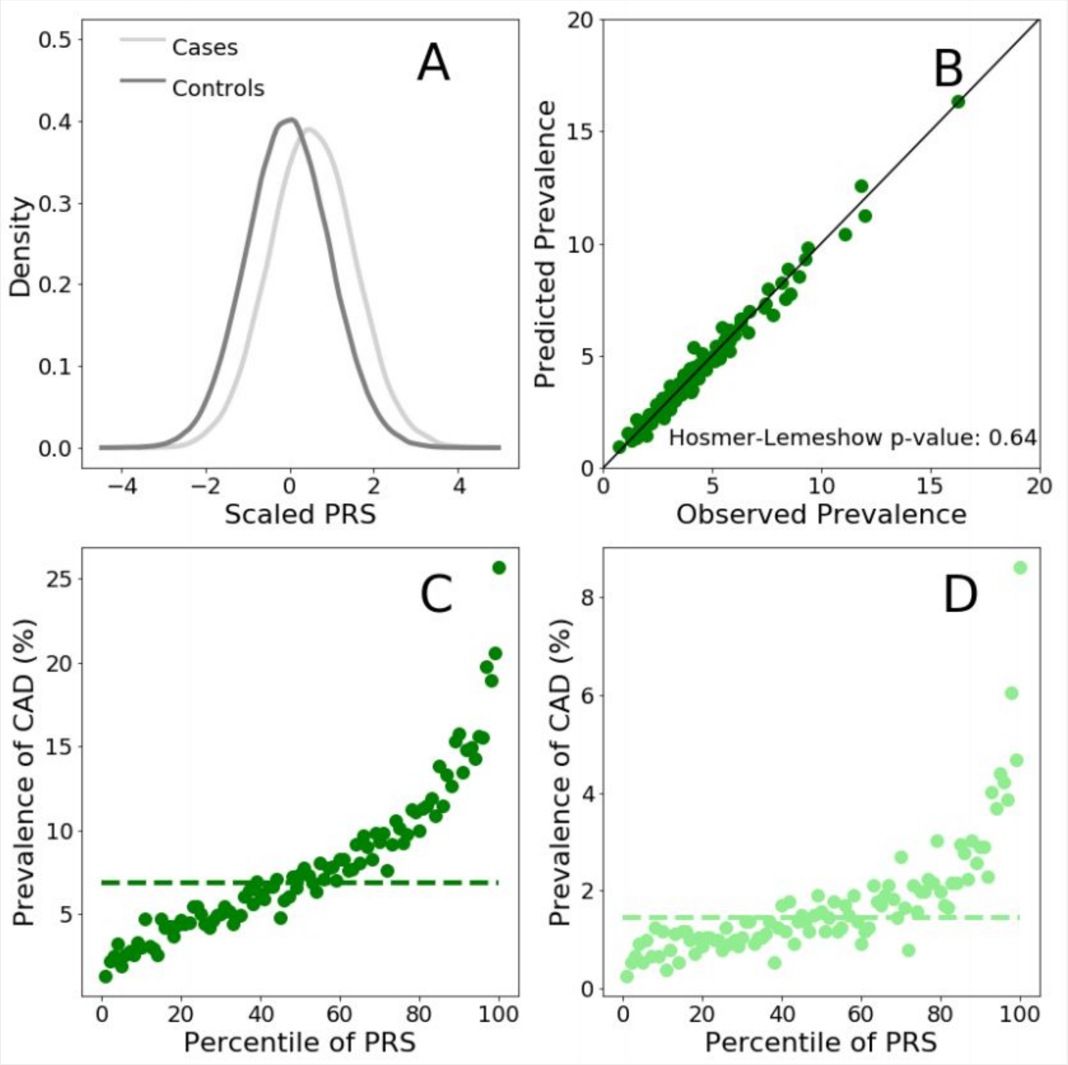


Figure 1: Risk for CAD using the SCT-I PRS panel. A) Distributions (scaled to a mean of 0 and a standard deviation of 1) of the per-individual PRS in the men and women testing populations.

B) Comparison between the observed and predicted CAD prevalence. Observed prevalence was calculated as the per-percentile prevalence of CAD in the per-individual PRS distribution. Predicted CAD prevalence was calculated for each individual using a logistic regression model with per-individual PRS as predictive variables. Within each percentile of the PRS distribution, CAD probability was averaged and this returned the predicted prevalence of CAD. **C)** Prevalence of CAD per percentile of the per-individual PRS distribution calculated in men from the Testing population.

D) Prevalence of CAD per percentile of the per-individual PRS distribution calculated in women from the Testing population. Dashed horizontal lines: CAD prevalence of the average of the PRS distributions (defined as between the 40% and the 60% percentiles) for men (dark green) and women (light green)

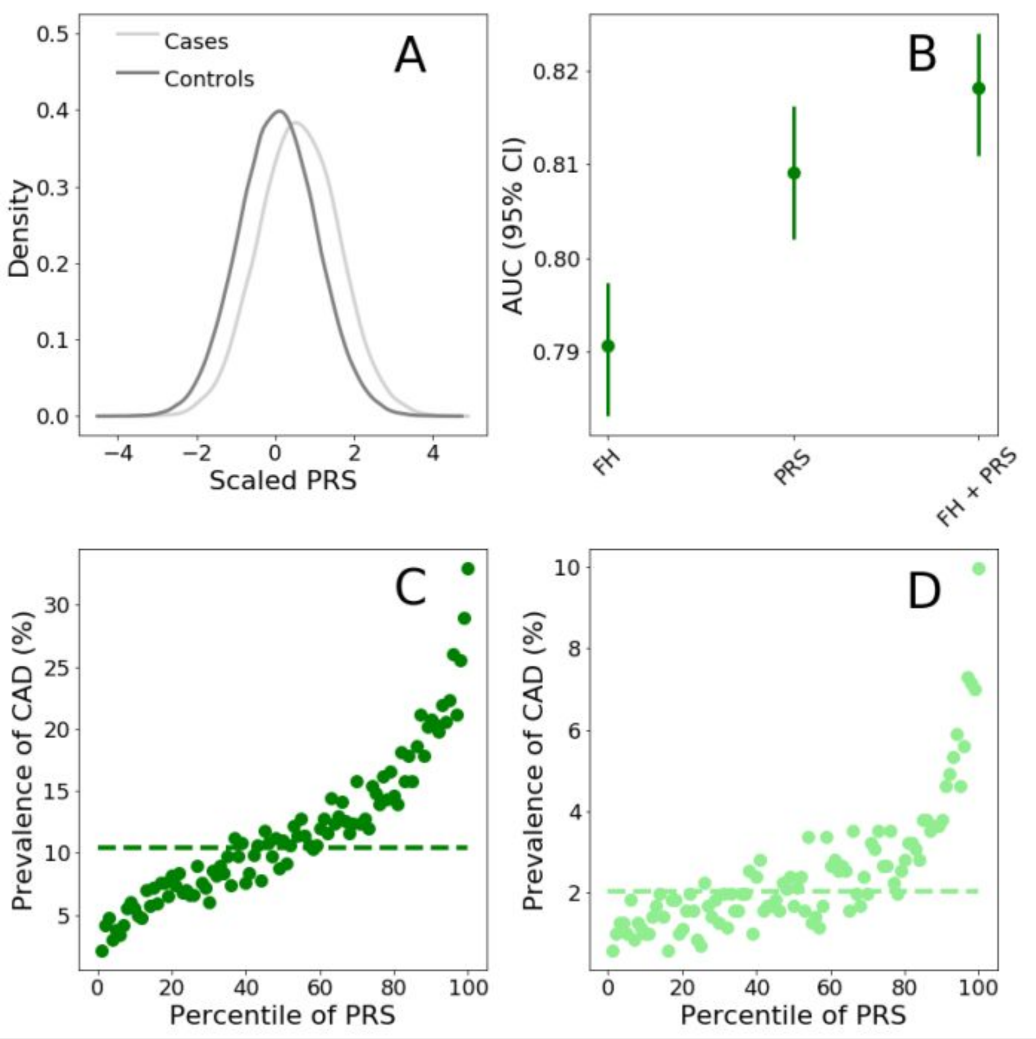


Figure 2: Risk for CAD using the SCT-I PRS panel together with Family history of CAD.
A) Distributions (scaled to a mean of 0 and a standard deviation of 1) of the per-individual PRS score for CAD cases and control individuals with at least one first-degree relative with a history of heart disease. **B)** AUC values on the Testing set computed using a logistic regression model including family history of CAD (FH), per-individual PRS calculated with the SCT-I PRS panel (PRS), or both (FH + PRS) as explanatory variables and presence/absence of CAD as the response. The model comprised additional covariates as control variables such as: age, gender, genotyping array, and the first 4 principal components of ancestry. **C)** Prevalence of CAD per percentile of the per-individual PRS score distribution calculated in the Testing population for the men with at least one first-degree relative with history of heart disease. **D)** As **C** but for women. Dashed horizontal lines: CAD prevalence of the average of the men's and women's PRS distributions (defined as between the 40% and the 60% percentiles).

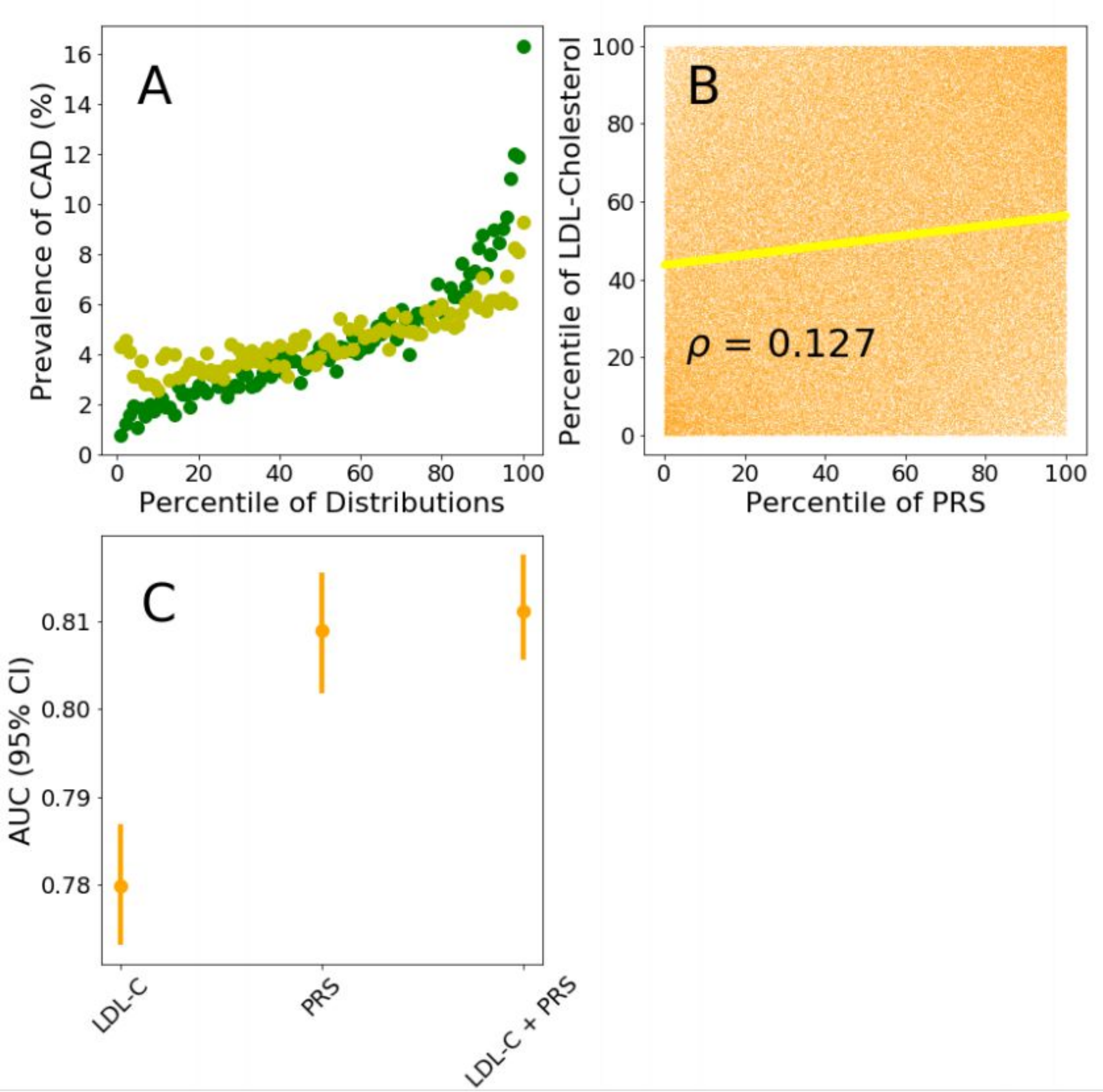


Figure 3: Predictive performance of the SCT-I PRS and LDL-cholesterol. A) Prevalence of CAD per percentile of the per-individual PRS (green circles) and LDL-cholesterol (yellow circles) distributions calculated in the testing population. **B)** Scatter plot of the per-individual percentiles of the PRS distribution plotted against the per-individual percentiles of the LDL-cholesterol distribution. Yellow continuous line: linear regression of the scatter plot. **C)** AUC values on the testing set of UKBB calculated with logistic regression models with LDL-cholesterol levels (LDL-C), per-individual PRS calculated with the SCT-I PRS panel (PRS), or both (LDL-C + PRS) as explanatory variables. The response variable of the logistic regression model was absence/presence of CAD. The logistic regression model comprised additional covariates as control variables such as: age, gender, genotyping array, and the first 4 principal components (PCs) of ancestry.

CONCLUSIONS

In this article we have described the development of a PRS for Coronary Artery Disease. We tested its prediction and risk stratification performances in the UK Biobank, which is the largest population dataset currently available. When compared with previously published PRS, the PRS showed highest predictive performance (Table 3). Additionally, we have demonstrated that PRS for CAD is able to identify a notable fraction of the UK Biobank population (5%) with a 3 fold or higher increased risk of developing CAD compared to the population average. Of note, a similar relative risk of CAD is observed in individuals carrying rare, highly penetrant familial hypercholesterolemia mutations. This large fraction of the population with a high polygenic risk for the above diseases highlights the deep impact that PRS based-screening can have to improve targeted prevention strategies.

The risk stratification ability of our newly developed PRS is maintained even in individuals already considered at risk, based on positive family history or high plasma lipid levels. This suggests that integrated models of PRS together with other lifestyle and clinical factors can enable clinicians to more accurately quantify the risk of the diseases and to consequently adjust prevention and screening strategies.

We have developed an easy to use, intuitive software suite to perform PRS analysis, which is now available to clinical laboratories and research groups as a fully automated, HIPPA and GDPR compliant and certified as a CE marked medical device. The software calculates Polygenic Risk Scores for a number of complex diseases and can analyse thousands of samples in parallel having the potential to improve health care prevention through its large scale implementation into public health practice.

REFERENCES

[1] T. A. o. M. Sciences, “Our data-driven future in healthcare”, no. November, 2018. [Online]. Available: https://acmedsci.ac.uk/file-download/74634438.

[2] C. Bycroft, L. T. Elliott, A. Young, et al., “The UK Biobank resource with deep phenotyping and genomic data”, Nature, vol. 562, no. 7726, pp. 203–209, 2018, ISSN: 0028-0836. DOI: 10.1038/s41586-018-0579-z. [Online]. Available: https://www.nature.com/articles/s41586- 018- 0579-zhttp://www.nature.com/articles/s41586- 018- 0579-z.

[3] N. R. Wray, S. H. Lee, D. Mehta, et al., “Research Review: Polygenic methods and their application to psychiatric traits”, en, Journal of Child Psychology and Psychiatry and Allied Disciplines, vol. 55, no. 10, pp. 1068–1087, 2014, ISSN: 14697610. DOI: 10.1111/jcpp.12295. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jcpp.12295.

[4] A. R. Martin, M. J. Daly, E. B. Robinson, S. E. Hyman, and B. M. Neale, Predicting Polygenic Risk ofù Psychiatric Disorders, Jul. 2019. DOI: 10.1016/j.biopsych.2018.12.015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S000632231832119X.

[5] S. M. Purcell, N. R. Wray, J. L. Stone, et al., “Common polygenic variation contributes to risk of schizophrenia and bipolar disorder”, Nature, vol. 460, no. 7256, pp. 748–752, 2009, ISSN: 00280836. DOI: 10.1038/nature08185.

[6] A. V. Khera, M. Chaffin, K. G. Aragam, et al., “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations”, eng, Nature Genetics, vol. 50, no. 9, pp. 1219–1224, 2018, ISSN: 15461718. DOI: 10.1038/ s41588-018-0183-z.

[7] B. J. Vilhjálmsson, J. Yang, H. K. Finucane, et al., “Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores”, eng, American Journal of Human Genetics, vol. 97, no. 4, pp. 576–592, Oct. 2015, ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2015.09.001.

[8] F. Privé, B. J. Vilhjálmsson, H. Aschard, and M. G. B. Blum, “Making the most of Clumping and Thresholding for polygenic scores”, bioRxiv, p. 653 204, Jan. 2019. DOI: 10.1101/653204. [Online]. Available: http://biorxiv.org/content/early/2019/06/28/653204.abstract.

[9] N. Mavaddat, K. Michailidou, J. Dennis, et al., “Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes”, Am. J. Hum. Genet., vol. 104, no. 1, pp. 21–34, 2019.

[10] H. R. Brewer, M. E. Jones, M. J. Schoemaker, A. Ashworth, and A. J. Swerdlow, “Family history and risk of breast cancer: an analysis accounting for family structure”, Breast Cancer Res. Treat., vol. 165, no. 1, pp. 193–200, 2017.

[11] D. G. Evans, J. Graham, S. O’Connell, S. Arnold, and D. Fitzsimmons, “Familial breast cancer: summary of updated NICE guidance”, BMJ, vol. 346, f3829, 2013.