# Neuroimmune guidance cues for vascular health

Zhang, H.

# Chapter 6

## Prediction power on cardiovascular disease of neuroimmune guidance cues expression by peripheral blood monocytes determined by machine learning methods

Zhang, H., Bredewold, E. O. W., Vreeken, D., Duijs, J., de Boer, H. C., Kraaijeveld,A. O., Jukema, J. W., Pijls, N. H., Waltenberger, J., Biessen, E. A. L., van der Veer,E. P., van Zonneveld, A. J. and van Gils, J. M.

*International Journal of Molecular Science*, 2020;21.

## Abstract

Atherosclerosis is the underlying pathology in a major part of cardiovascular disease, the leading cause of mortality in developed countries. The infiltration of monocytes into the vessel walls of large arteries is a key denominator of atherogenesis, making monocytes accountable for the development of atherosclerosis. With the development of high-throughput transcriptome profiling platforms and cytometric methods for circulating cells, it is now feasible to study in-depth the predicted functional change of circulating monocytes reflected by changes of gene expression in certain pathways and correlate the changes to disease outcome. Neuroimmune guidance cues comprise a group of circulating- and cell membrane-associated signaling proteins that are progressively involved in monocyte functions. Here, we employed the CIRCULATING CELLS study cohort to classify cardiovascular disease patients and healthy individuals in relation to their expression of neuroimmune guidance cues in circulating monocytes. To cope with the complexity of human datasets featured by noisy data, nonlinearity and multidimensionality, we assessed various machine-learning methods. Of these, the linear discriminant analysis, Naïve Bayesian model and stochastic gradient boost model yielded perfect or near-perfect sensibility and specificity and revealed that expression levels of the neuroimmune guidance cues SEMA6B, SEMA6D and EPHA2 in circulating monocytes were of predictive values for cardiovascular disease outcome.



## Keywords

cardiovascular diseases; monocytes; machine-learning methods; neuroimmune guidance cues

## 6.1 Introduction

Cardiovascular diseases (CVD) remain a leading cause of death in the more economically developed countries, despite improvements in surgical and drug treatments. Much of the CVD-related mortality and morbidity is attributable to atherosclerosis [1]. Atherosclerosis is a systemic chronic inflammatory and immune disease [2,3]. Monocytes and their derived macrophages play a key role in the development of atherosclerosis. Under conditions of dyslipidemia and chronic systemic inflammation, circulating monocytes and the endothelium become activated, resulting in monocyte infiltration and differentiation into macrophages in the vessel wall. Upon the excessive uptake of lipids, these macrophages become foam cells and participate decisively in the development and exacerbation of atherosclerosis, coronary stenosis and its clinical sequela, such as acute myocardial infarctions [3-7]. Neuroimmune guidance cues (NGCs) comprise the netrin, semaphorin, ephrin and slit families of proteins of ligands and receptors, which were originally characterized to direct cell and axon migration during neural development. In the last two decades, it has become increasingly clear that these proteins can also play a major role in (pathological) immune responses by directly regulating leukocyte trafficking and directly impacting the pathogenesis of atherosclerosis [8-10]. Indeed, numerous studies using murine atherosclerosis models have found multifaceted roles of NGCs in the development of atherosclerosis [11-15]. In addition, several observations also support a role for NGCs in human CVD. For instance, three NGC genes are located on human chromosome 1, in the locus that has been identified as the premature myocardial infarction susceptibility locus [16]. In addition, the axonal guidance pathway is found enriched with genetic variants that have significant associations with CVD, and several novel genetic risk loci for CVD contain NGCs genes [17,18]. However, whether the monocytic expression of NGCs is informative for human CVD has not been described yet. Transcriptomics can reveal key alterations in biological processes causing human diseases, thereby present novel instruments that are not only useful for the understanding of the disease mechanisms but, also, for molecular diagnosis and clinical therapy [19]. Since monocytes are among the culprit cells of atherosclerosis development, monocytic expression levels of NGCs could provide insights into the underlying mechanisms in atherosclerosis development and can be used to improve the evidence-based treatment of CVD to reduce the global burden of this disease. The CIRCULATING CELLS study was designed to study the role of several cellular mediators of atherosclerosis as biomarkers of CVD to predict the suscep-

tibility of patients to the progression of CVD [20]. By applying different machine-learning methods (also known as predictive modeling methods) on the gene expression data of peripheral monocytes from the CIRCULATING CELLS study cohort, we investigated whether monocytic NGC expression is informative to distinguish between healthy individuals and CVD patients. As machine-learning methods are developed to explore complex relationships between predictors and outcomes, they are suitable tools to tackle the difficulties due to the complexity of human datasets featured by noisy data, nonlinearity and multidimensionality. Some machine-learning methods take simplistic approaches and work with linear relationships between features and outcomes, while other methods are more complex and are able to capture nonlinear patterns and to tolerate a low information-noise ratio, owing to the difference of their pre-assumptions and learning logics. We compared the performance of multiple modeling methods to explore the best predicting potential of our dataset. In our study, we included commonly used models like logistic regression and linear discriminant analysis, as well as more complex nonlinear models and tree-based models. Linear models, like partial least square, have supervised dimension reduction functionality, which benefits model performances in the case of high between-feature correlations. Nonlinear models, like Naïve Bayesian, make probabilistic calls based on the information provided by the features independently, possibly performing better in situations when between-feature correlations are low. In addition to modeling of the data itself, stochastic gradient boost also models the residuals, thereby increasing the learning ability when the information-noise ratio is low. We compared the results of the different modeling methods to gain insights on the nature of the dataset. Altogether, this allowed us to give a proof of concept that the expression of a small set of functional genes can be a prediction value for complex diseases like CVD.

## 6.2 Materials and Methods

### *6.2.1 Study Population*

The study population consists of a subgroup of 369 patients from the CIRCULATING CELLS study cohort [20] (**Figure 1**). In brief, CIRCULATING CELLS was a prospective multicenter study in which patients scheduled for coronary angiography due to CVD were included. For this subgroup, extensive clinical characteristics were recorded (**Table 1**), and the transcriptomes of purified circulating CD14+ monocytes were profiled. To minimize the potential influence of the presence of profound acute myocardial ischemia on monocytic gene expression profiles, patients with ST-elevation myocardial infarction (STEMI) were excluded.

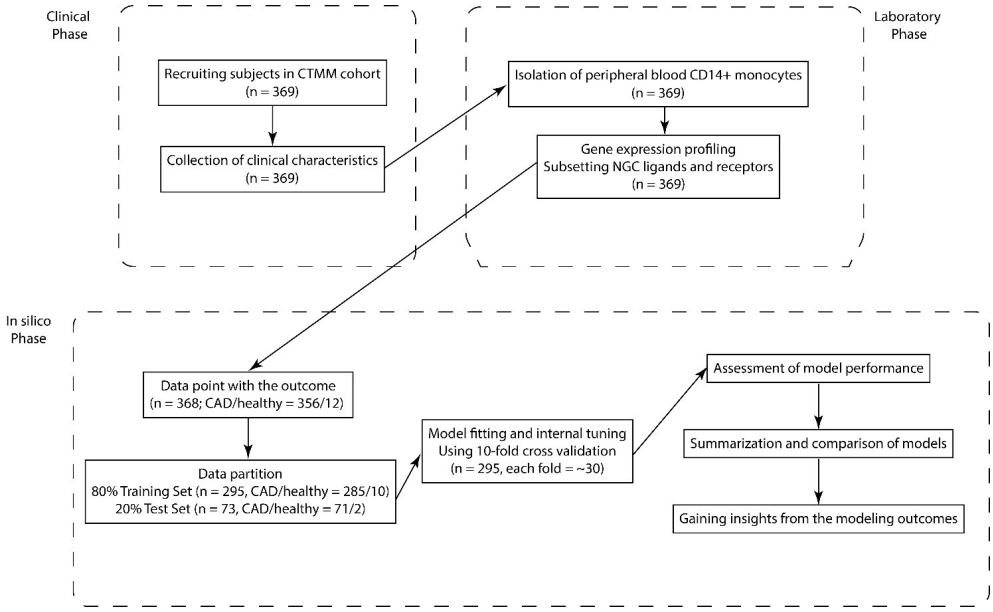



**Figure 1. Flowchart of predictive modeling using neuroimmune guidance cues (NGCs)** Subjects in the CIRCULATING CELLS cohort were recruited based on inclusion criteria. The clinical characteristics were collected, peripheral blooded CD14-positive monocytes were isolated, and their transcriptomes were profiled. The expression of neuronal guidance cues was subset. The individuals were randomly assigned to the Training Set or Test Set for external assessment of the model performance. Classification models that were built on the Training Set data and model performance were internally assessed with cross-validation. Finally, we made comparisons between the models and gained insights on the choice of model and features.

| Clinical Characteristics | All | CVD | Healthy |
|---|---|---|---|
| *Demographic data* | | | |
| Number (male/female) | 368 (273/95) | 356 (264/92) | 12 (9/3) |
| Age | 61.8 (±10.4) | 62.2 (±10.3) | 49.2 (±6.3) |
| BMI | 27.3 (±4.3) | 27.4 (±4.3) | 23.7 (±2.3) |
| *Coronary risk factors* | | | |
| Hypertension | 231 (63%) | 231 (0.65) | 0 (0) |
| Hypercholesterolemia | 70 (19%) | 65 (0.18) | 5 (0.42) |
| Diabetes | 77 (21%) | 77 (0.22) | 0 (0) |
| Current smoker | 73 (20%) | 73 (0.21) | 0 (0) |
| Family MI history | 141 (39%) | 137 (0.39) | 4 (0.33) |
| Previous MI | 112 (30%) | 112 (0.31) | 0 (0) |
| Positive family history | 157 (43%) | 152 (0.43) | 5 (0.42) |
| *Therapeutic decision* | | | |
| PTCA | 130 (35%) | 130 (0.37) | 0 (0) |
| CABG | 32 (9%) | 32 (0.09) | 0 (0) |
| *NYHA Classification* | | | |
| NYHA Class I | 248 (67%) | 236 (0.66) | 12 (1) |
| NYHA Class II | 78 (21%) | 78 (0.22) | 0 (0) |
| NYHA Class III | 26 (7%) | 26 (0.07) | 0 (0) |
| NYHA Class IV | 16 (4%) | 16 (0.04) | 0 (0) |
| *Current medication* | | | |
| β-blocker | 228 (69%) | 228 (0.72) | 0 (0) |
| Ca-antagonist | 95 (29%) | 95 (0.30) | 0 (0) |
| Aspirin | 260 (79%) | 260 (0.82) | 0 (0) |
| Vitamin K antagonist | 29 (9%) | 29 (0.09) | 0 (0) |
| Low molecular weight heparin | 10 (3%) | 10 (0.03) | 0 (0) |
| ADP receptor blocker | 168 (51%) | 168 (0.53) | 0 (0) |
| ACE inhibitor | 116 (35%) | 116 (0.36) | 0 (0) |
| ATII receptor blocker | 71 (22%) | 71 (0.22) | 0 (0) |
| Diuretic | 76 (23%) | 76 (0.24) | 0 (0) |
| Statins | 252 (77%) | 252 (0.79) | 0 (0) |

**Table 1. Clinical characteristics of the CTMM patient cohort**

Values are N ± SD or N (%). Abbreviations: BMI—body mass index, MI—myocardial infarction, PTCA—percutaneous transluminal coronary angioplasty, CABG—coronary artery bypass graft, NYHA—New York Heart Association, and CVD—cardiovascular disease.

Additional exclusion criteria were age ¡ 18 years, inability to give informed consent, suspected drug or alcohol abuse, serious concomitant disease, serious recent infectious disease in the last 6 weeks or suspected elevated state of the immune system and noncooperativeness. The study was approved by the medical ethical committees of the participating centers and conformed to the Declaration of Helsinki. All patients received oral and written information about the objectives of the study and provided written informed consent.

### *6.2.2 Isolation of Peripheral Blood CD14-Positive Monocytes*

The full procedures for the isolation of peripheral blood CD14-positive monocytes were described previously [20]. Briefly, 60 mLof EDTA blood was collected from patients via the arterial sheath catheter. Peripheral blood mononuclear cells (PBMCs) were isolated by density gradient centrifugation over Ficollpaque Plus (GE Healthcare, Diegem, the Netherlands). For further purification of monocytes, the PBMC fraction was incubated with magnetic beads coated with anti-CD14 antibodies (BD Biosciences, Breda, the Netherlands), and monocytes were purified with a MACS separation system according to the manufacturer's instructions (BD Biosciences, Breda, the Netherlands). Cells in CD14-positive fraction were resuspended, lysed in Trizol and aliquoted. The aliquots were stored at -80 °C for RNA isolation.

### *6.2.3 RNA Isolation and Microarray Analysis*

Monocyte samples were shipped to Eurofins Genomics for semiautomated extraction of RNA using RNeasy 96-well plates (Qiagen). RNA samples were quantified using a Beckman Coulter DTX880 system, and only samples that displayed RIN values $> 9$ (Agilent Bioanalyzer) were included. Labeled RNA was prepared and used on the array for hybridization. Hybridized chips were scanned by Illumina BeadStation (Illumina, Inc., San Diego, CA, USA). Raw image analysis and signal extraction was performed with Illumina Beadstudio Gene Expression software with default settings (no background subtraction). Data were exported as text files. The gene expression profiling data were integrated and archived using the self-developed software "Circucel" [20]. The expressions of NGC ligands and receptors were extracted along with the phenotypic profiles of the patients. We excluded patient records without the required outcome parameter—in this case, a "confirmed diagnosis". For NGC expression profiles, there were no missing values. Therefore, data imputation was not necessary.

### *6.2.4 Statistical Analysis*

Univariate correlation of NGC expressions (or other continuous variables) with a categorical variable was tested by 2 mean Student's t-tests. p-values were obtained from t-statistics. Univariate correlation of NGC expressions (or other continuous variables) with a continuous variable was tested with linear regression. p-values were obtained from the t-statistic of the coefficient of the variable. Correlation of 2 categorical variables was tested using Pearson's chi-squared test of the cross-tabulation. p-values were obtained from the chi-square statistics. For all the tests, a p-value of less than 0.05 was considered significant.

### *6.2.5 Model Fitting and Assessment of Model Performance*

We used R package "caret" and its multiple dependencies (summarized in **Table 2**) for modeling the predicting power of NGCs to the disease status of patients [33]. Performance statistics for binary classification models were calculated, including accuracy, Cohen's kappa ($\varkappa$), sensitivity and specificity. The Cohen's kappa ($\varkappa$) value is given by formula

$$\frac{P_{observed} - P_{expected}}{1 - P_{expected}} \tag{6.1}$$

which indicates the performance gain from the modeling over random guessing (the higher, the better). Since this data set features a smaller number of healthy individuals, we set the models to aim for picking up healthy individuals as events. The sensitivity (true positive prediction) and specificity (true negative prediction) values were also calculated in accord to this principle. For the external assessment of model performance, data partitions were created to have a training set (90% of the dataset) and a test set (10% of the dataset). This was done using the "createDataPartition" function in the "caret" package to ensure proportional and representative coverage of individuals in both the training and test sets. Data in the training set were used for model building, and the data in the test set were held out in the training process and were used to determine the model performance pseudo-externally by comparing the prediction on the test set with the actual outcome of the test set. The distributions of NGC expressions in the training set and test set were illustrated in **Figure S3**. For the internal assessment of model performance and stability, 10-fold cross-validations were done, which means 10% of the training data were kept out of each iteration to evaluate the model generated by the other 90% training data over 10 iterations. Similar performance statistics were calculated at each iteration. The average values and standard deviations of the parameters were summarized to assess the model performance and

stability, respectively. Some models were tuned in ranges of tuning parameters to control their complexity and adaptivity (**Table 2**). Tuning parameters giving the best performance statistics in the cross-validation were chosen as optimal tuning of the model, and a final model was built using these parameters on the complete training data.

## 6.3 Results

### *6.3.1 NGC Expressions in Monocytes and Feature Selections*

We made use of the CIRCULATING CELLS study cohort [20] to address the question of whether NGC expression profiles of circulating blood cells can be related to cardiovascular health. From 368 subjects out of this cohort (CVD patients and healthy controls), the transcriptomes of their CD14-positive monocytes were profiled (**Figure 1**). The individuals received different treatments and medications, and some of them suffered from other diseases, resembling the reality of complexity of most human cohorts (**Table 1**). Next, we aimed to classify CVD patients and healthy individuals using the differential expression of the NGC transcripts. To that end, we first sought to include NGCs with high expression levels and good univariate correlation with the outcome. **Figure 2A** shows the expression of NGCs in the cohort. Based on the detection threshold of the profiling platform, NGCs with signals higher than 6.75 (log2 scale unless specified otherwise) were unconditionally included in the modeling as potential features. To validate the microarray analysis, we compared the monocytic NGC expression profile obtained by microarray to that obtained by real-time PCR. Both methods showed a similar expression profile, with the exception of SEMA3E (**Figure S1**). To gain understanding of univariate correlation of the features to the outcome, violin plots of the NGC expressions were created to compare the distribution of NGC expression levels in both the CVD group and healthy group (**Figure 2B and Table S1**). The ranges of expressions showed overlaps in both groups, suggesting that the univariate prediction power will be minimal. In addition, with the ranges being widespread, the information to noise ratio is relatively low in this dataset. To quantify the univariate correlation of NGC expressions to the disease status, we calculated the p-value with two mean t-tests between the CVD group and healthy group. A volcano plot was created to observe the t-test p-value in relation to the fold change (**Figure 3A**). We identified several NGC ligands and receptors to be significantly different between CVD patients and healthy individuals, although with small fold changes. Among the significantly different genes, 10 had expression signals lower than 6.75. These 10 were added to the modeling procedure despite

| **Model Name** | **Abbr.** | **Type** | **Best tuning parameter (Tuning Range)** | **R Package Dependency** |
|---|---|---|---|---|
| Boosted Logistic Regression | Logit | Linear | $n\_iter = 41(11, 101)$ | "caTools" |
| Linear Discriminant Analysis | Lda | Linear | $NA$ | "MASS" |
| Partial Least Squares | Pls | Linear | $n\_comp = 1(1, 10)$ | "pls" |
| Support Vector Machines | Svm | Nonlinear | $Cost = 0.25(2^{-2}, 2^{-7})$ | "kernlab" |
| Nearest Shrunken Centroids | Pam | Linear | $Threshold = 0(0, 25)$ | "pamr" |
| Mixture Discriminant Analysis | Mda | Nonlinear | $Subclasses = 11(2, 16)$ | "mda" |
| Flexible Discriminant Analysis | Fda | Nonlinear | $Degree = 4(1, 5)$<br>$n\_pruning = 5(2, 5)$ | "earth", "mda" |
| k-Nearest Neighbors | Knn | Nonlinear | $n\_neighbor = 5(5, 23)$ | "class" |
| Naive Bayesian | Nb | Nonlinear | $Laplacecorrection = 1(1, 3)$<br>$Kernalfunction = F(F, T)$<br>$BandwidthAdj = 1(1, 3)$ | "naivebayes" |
| Bagged CART | BagCart | Tree/Rule-based | $NA$ | "ipred", "plyr" "e1071" |
| Random Forest | Rf | Tree/Rule-based | $n\_param = 2(2, 37)$ | "randomForest" |
| Stochastic Gradient Boosting | Gbm | Tree/Rule-based | $Interactiondepth = 1(1, 7)$<br>$n\_trees = 450(100, 1000)$<br>$Shrinkage = 0.1(0.01, 0.1)$<br>$Minnodesize = 5(5, 7)$ | "gbm", "plyr" |
| C5.0 Tree | C5 | Tree/Rule-based | $NA$ | "C50", "plyr" |

**Table 2. Summary of model names, types and tuning parameters**

their low expression levels. In total, 35 NGC genes were further to be used as features in subsequent modeling. Finally, to avoid instability caused by between-feature collinearity in some models, we calculated between-feature correlations of the expressions of selected NGCs (**Figure 3B**). No pairwise correlations of NGC expressions exceeded the threshold of 0.75, suggesting that there would be a minimal influence of collinearity. Therefore, none of the selected NGCs were eliminated based on between-feature correlations.

### *6.3.2 Gender and Age are Unlikely to be Confounding Factors in the Current Study*

Gender and age are conventional confounding factors in clinical situations when revealing relationships between measurements of phenotypes and diseases. For machine learning, if age or gender affect both the features and the outcome, they would be confounding factors by definition. Firstly, we examined the relationship of age to NGC expressions and disease outcome. NGC expressions plotted against age in scatterplots with linear fittings showed no significant correlation between age and the NGC expressions (**Figure 4A**). The age ranges of both groups overlap, although the younger age dominates in the healthy group (**Figure 4B**). These observations suggested that age would add an additional prediction power in our modeling but would not be a confounding factor, as it does not link directly to NGC expressions. Next, we examined the relationship of sex to NGC expressions and disease outcomes. Using violin plots, we compared the distribution of NGC expressions in both sexes (**Figure 5A**). The distribution of NGC expressions was barely affected by sex, including the X-linked PLXNA3, PLXNB3 and EFNB1 genes (**Figure 5A**). There were six NGCs with significantly different expressions comparing males to females, albeit with a very small fold change relative to the variations (**Figure 5B**). Cross-tabulations of sex and disease outcomes were made, and the frequency distribution of both genders was similar among CVD patients and healthy individuals (**Figure 5C**). Therefore, sex was also unlikely to be a confounding factor in this study. Regardless, sex and age were included in our modeling process, as it is common practice to control for these conventional confounding factors.

### *6.3.3 Performance of Different Models*

Different machine-learning methods were applied, and optimal tuning was obtained for each model listed in **Table 2**. The performance statistics for each of the models was calculated and summarized in **Figure 6**. As measurements for model stability, we examined standard deviations of the cross-validations for accuracies and Cohen's Kappa. Model per-

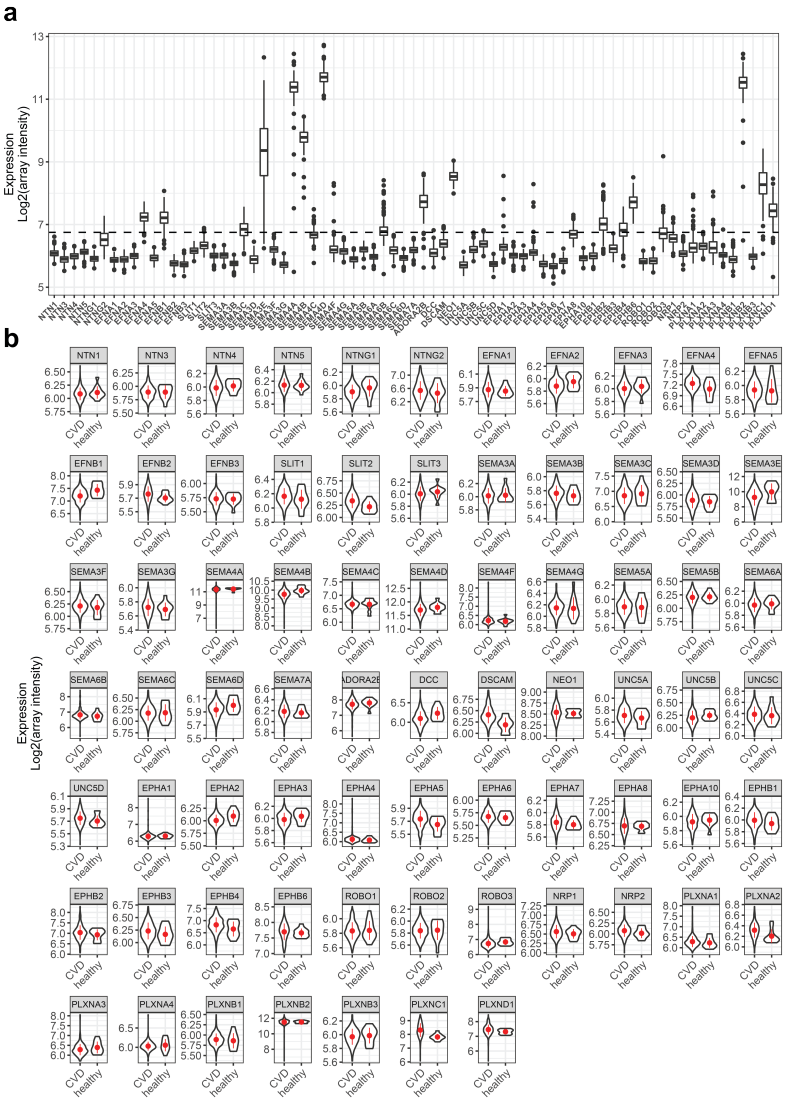



**Figure 2. NGC expressions in patients and healthy subjects**

(a) Box plots with quartiles were created using NGC expressions from all individuals. Baseline signal of the platform (6.75) was indicated with the dashed line. (b) Violin plots of all NGC expressions were created for cardiovascular disease (CVD) patients and healthy individuals. The violin shapes represent the density distribution of NGC expressions in the groups. The NGC expressions of CVD patients overlap with those of healthy individuals. Due to the small number of healthy individuals, their NGC expressions were sometimes not normally distributed.

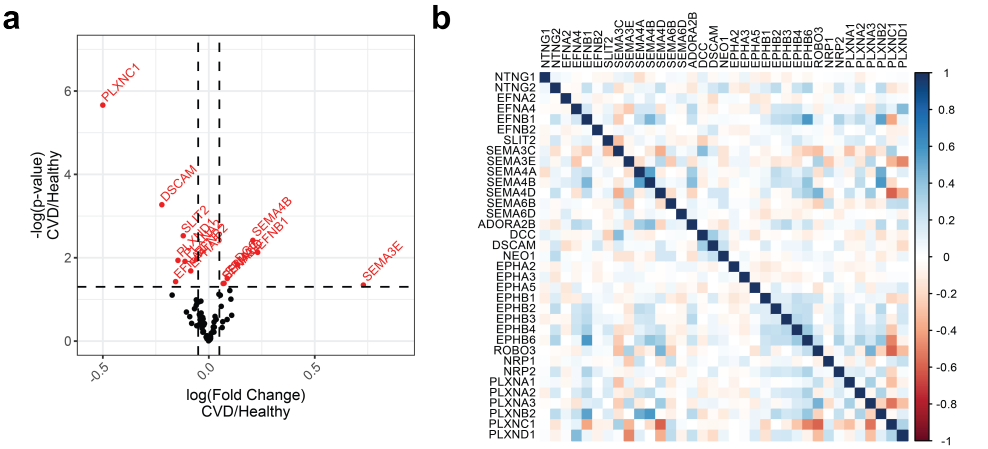



**Figure 3. Univariate correlation of NGC expressions with disease phenotypes and pairwise correlations of NGC expressions**

(a) Volcano plot showing the fold changes and the significance of difference in the 2 mean Student's t-tests between NGC expressions in CVD patients and healthy individuals. NGCs with significant differences ($p < 0.05$) are labeled with gene names and red color. For most NGCs, the univariate correlation determined by the 2 mean Student's t-tests with the outcomes is minimal. (b) Pairwise covariances of NGC expressions were illustrated in the heatmap to examine the between-feature correlations. Blue color indicates a positive correlation, whereas red color indicates a negative correlation. There are no absolute pairwise covariances higher than 0.75, the common cutoff for highly correlated features.

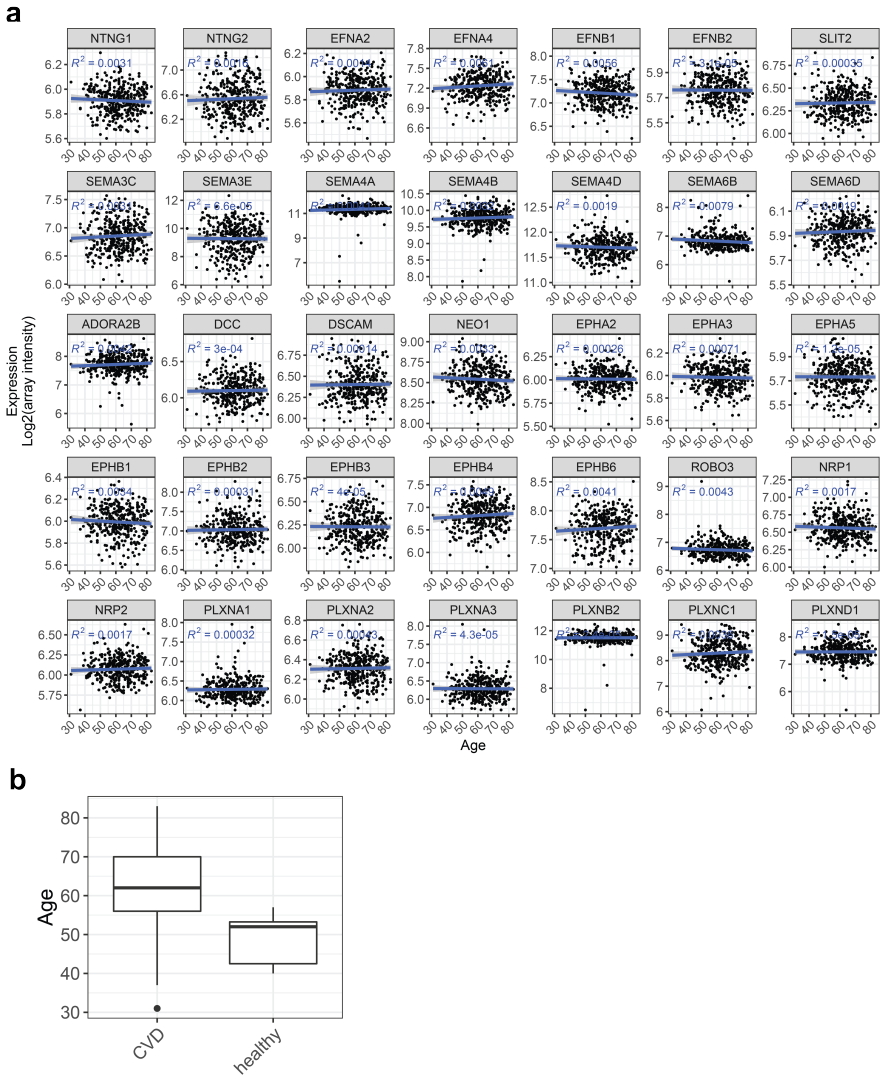
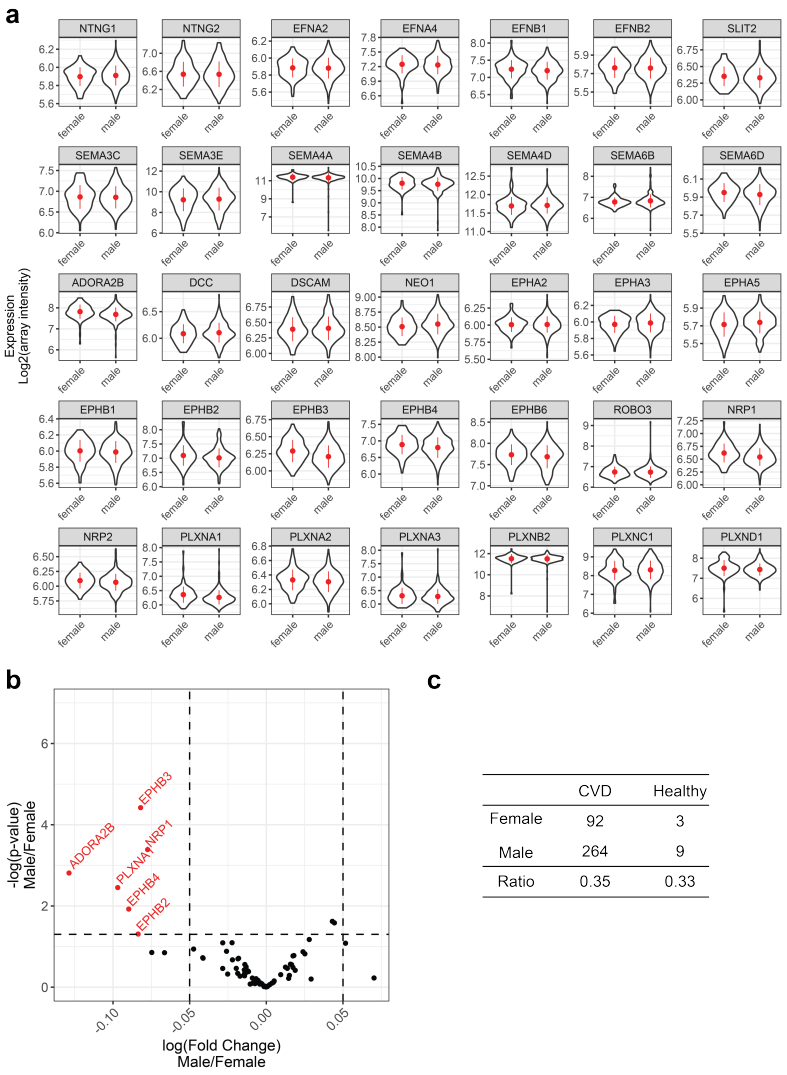



**Figure 4. Influence of age as a potential confounding factor**

(a) Scatter plots of expressions of selected NGCs in relation to age were made to show the influence of age on NGC expressions. Linear regressions were done with age being the dependent variable providing the fitted trend line (blue line), the 95% confidence interval of the trend (gray area) and the regression R-squared. Correlations of NGC expressions and age are minimal, as indicated by the R-squared values. (b) Boxplot of age distribution in CVD patients and healthy individuals. Young age dominates in the healthy individuals.




| | CVD | Healthy |
|---|---|---|
| Female | 92 | 3 |
| Male | 264 | 9 |
| Ratio | 0.35 | 0.33 |

**Figure 5. Inspection of the influence of sex as a potential confounding factor**

(a) Violin plots to show the distribution of expressions of selected NGCs in males and females. To be noticed is that distributions of selected NGC expressions are similar between sexes. (b) A volcano plot was made showing the fold change and the significance of differences in a 2 mean Student's t-tests between NGC expressions in males and females. NGCs with significant differences ($p < 0.05$) were labeled with gene names and red color. (c) Contingency table showing that the ratio of sexes is not biased in relation to the outcomes.

formance could be categorized in four groups (**Figure 6A-D**). (1) Partial least squares, support vector machine and nearest shrunken centroids models failed entirely to model the informative part of the gene expression data, as was revealed by having bottom-line accuracy (not better than predicting all the individuals to be CVD) and zero Cohen's Kappa value in the cross-validation, training set and test set. (2) Most models—namely, logistic regression, k-nearest neighbors, mixture discriminant analysis, flexible discriminant analysis, bagged CART (classification and regression tree), random forest and single C5.0 tree—suffered from overfitting, as was characterized by far better performances of models in the training set than in the test set. This can be explained by the modeling process trying to polish these models to perfectly predict the outcome based on the information in the training dataset. However, the modeling for the training set could not be generalized to the test set data. (3) K-nearest neighbors and flexible discriminant models had overall better prediction powers over random guesses but were still not up to standard. (4) Linear discriminant analysis, Naïve Bayesian and stochastic gradient boosting models performed best compared with the other models, within both the training set and test set at an accuracy of more than 0.98 and Cohen's Kappa more than 0.75 in the test set. These results indicate that the linear discriminant analysis, Naïve Bayesian and stochastic gradient boosting models were able to translate the informative part of NGC expression data into disease outcome. The three best-performing models all reached a sensitivity of 1 in the test set, meaning that they were able to discriminate healthy individuals from CVD patients (**Figure 6C**). Cohen's kappa ($\varkappa$) values were 0.79, 0.79 and 1 for the linear discriminant analysis, Naïve Bayesian model and stochastic gradient boost model, respectively (**Figure 6B**). The lower Cohen's kappa ($\varkappa$) for the former two models were due to the misclassification of one healthy individual as a CVD patient (**Table 3**). In the prediction of the training set, the linear discriminant analysis had lower sensitivity in the training set due to the misclassification of three healthy individuals to the CVD group (**Table 3**). Interestingly, the three misclassified healthy individuals still had higher modeled probability to be healthy than all but one misclassified CVD patient, suggesting that the sensitivity problem can be solved by an alternative cutoff value of the classification probability. When the cutoff value was altered from the original 0.5 to 0.28, the linear discriminant analysis achieved the same ideal sensitivity as the Naïve Bayesian model (**Figure S2A,B**). However, to prove the external efficiency of the alternative cutoff, another independent test set would be necessary, which is not feasible considering the size of this study.

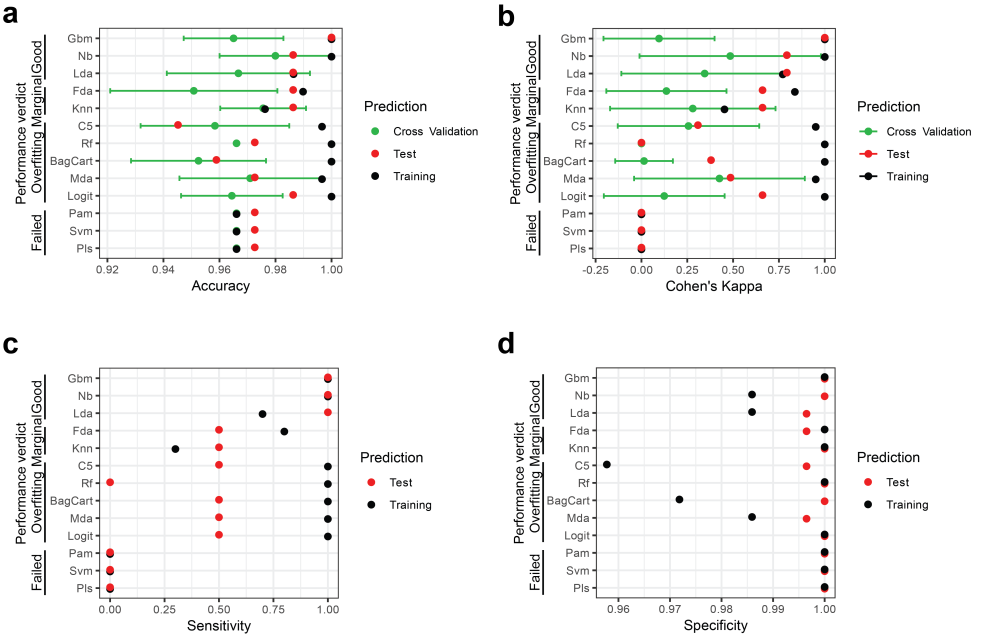



**Figure 6. Model performance**

(a-d) Model performance metrics—namely, accuracy (a), Cohen's Kappa (b), sensitivity (c) and specificity (d)—were illustrated in cross-validations (green dots and error bars), Training Set (black dots) and Test Set (red dots) for all models. Models are categorized into 4 groups based on their performances.

| | **Training** | **Reference** | | **Test** | **Reference** | |
|---|---|---|---|---|---|---|
| Model | Prediction | CVD | Healthy | Prediction | CVD | Healthy |
| Linear Discriminant | CVD | 284 | 3 | CVD | 70 | 0 |
| Analysis | Healthy | 1 | 7 | Healthy | 1 | 2 |
| Naive Bayesian | CVD | 285 | 0 | CVD | 70 | 0 |
| | Healthy | 0 | 10 | Healthy | 1 | 2 |
| Stochastic Gradient | CVD | 285 | 0 | CVD | 71 | 0 |
| Boosting | Healthy | 0 | 10 | Healthy | 0 | 2 |

**Table 3. Confusion matrices of the linear discriminant analysis, Naive Bayesian and stochastic gradient boosting models**

### *6.3.4 Features with the Most Importance in the Models*

Apart from age, the most important three features determined by the linear discriminant analysis and Naïve Bayesian model were PLXNC1, DSCAM and DCC, while the most important three features determined by stochastic gradient boost were SEMA6B, SEMA6D and EPHA2 (**Figure 7A-C**). As we noted before, age was determined to be important contributor to the model but was hardly a confounding factor, considering the weak correlation of age with NGC expressions (**Figure 7A-C**). The functional relevance of the top features will be discussed in the Discussion section.

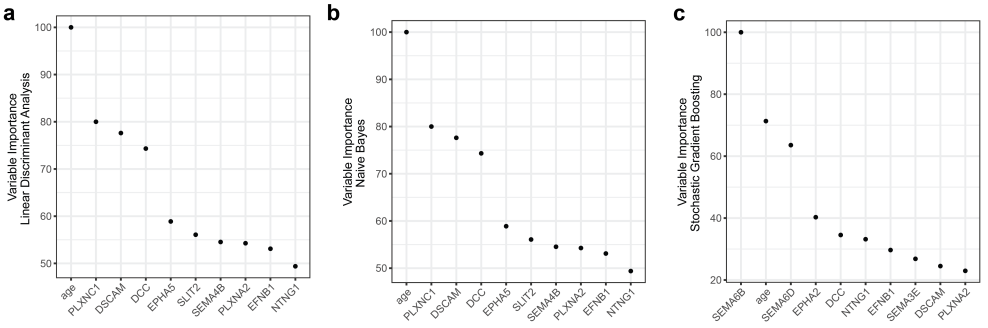



**Figure 7. Variable importance of the models**

(a-c) Variable importance measured in the linear discriminant analysis (a), Naïve Bayesian model (b) and stochastic gradient boost model (c). The importance of the most important feature was scaled to 100.

## 6.4 Discussion

In this study, we used NGC expressions of peripheral blood monocytes for the prediction of CVD. To reveal the true prediction power of monocytic NGC expression profiles, we performed cross-validation and validation using a pseudo-external test set with conventional confounding factors controlled. Of the models, Naïve Bayesian model and stochastic gradient model had satisfactory discrimination in both the training test and test set. The stochastic gradient model with a residual modeling mechanism was able to achieve 100% accuracy. Therefore, we have established the proof of concept that a small set of functional genes, NGCs, is of sufficient prediction power for the classification of CVD patients and healthy individuals. There are several challenging factors in the prediction of CVDs using monocytic NGC expressions. Firstly, nonlinearity is common in the biological effect of proteins.

Taking NTN1 as an example, the repellent effect of NTN1 on monocyte and macrophage migration has an optimal concentration of around 250 ng/ml; higher or lower concentrations are both less effective [3,14]. The biological effect also depends on its target, as NTN1 also has opposite roles acting on smooth muscle cells or macrophages [15]. Secondly, concomitant physiological processes in certain disease conditions can systemically cancel out the change of average gene expressions if they alter the gene expression in the opposite direction. In our case, CVD patients experience changes in the monocyte subpopulations, with increased lipopolysaccharide receptors and the low-affinity FC γ receptor-positive monocytes, referred to as intermediate monocytes [21,22]. These intermediate monocytes have been shown to predict cardiovascular events in subjects referred for elective coronary angiography [23]. At the same time, there is increased mobilization of lipopolysaccharide receptor-positive and low-affinity FC γ receptor-negative naïve classical monocytes from bone marrow, a process termed monocytosis. For any changes induced by monocyte activation/differentiation, monocytosis will cancel out the change because of the added naïve classical monocyte population. Moreover, human measurements in general can be very heterogenic by nature. Even proved biomarkers suffer from false positives and false negatives because of large variations in human measurements. In this study, we sought to tackle these problems by applying multiple modeling methods, each of which incorporates special features in the aspects of the linearity requirements or the learning mechanisms. Machine-learning methods are different from each other in various ways, including learning mechanisms and the assumptions made on the features. For a given dataset, choosing a model with suitable learning mechanisms and proper assumptions of the features can benefit the performance of modeling. In the current dataset, the linear discriminant analysis and the Naïve Bayesian model both adopted the same features, with identical weights on each feature, but the Naïve Bayesian model had better performance (**Figure 6 and 7A,B**). As both models are based on multivariate probability densities, the difference of the model performance should result from the different intrinsic assumptions and learning mechanisms of the models. The linear discriminant analysis assumes a multidimensional Gaussian distribution of the feature data, while the Naïve Bayesian model works with a more flexible distribution. Instead, the Naïve Bayesian model makes a strong assumption that all features are independent, so that the conditional probability of one class will be simply the product of the probability densities of all features. In addition, Naïve Bayesian could model nonlinear relationships between features and the outcome. Due to smaller numbers of individuals in the healthy group, distributions were

sometimes non-normal (**Figure 2B**). As previously described, the features in this dataset have relatively low pairwise covariance (**Figure 3B**), so that the independent-feature assumption required by the Naïve Bayesian model is very likely to be acceptable. Taken together, the structure in the current dataset favors the Naïve Bayesian model, which explains its better performance. The top 10 important NGCs chosen in the linear discriminant analysis and Naïve Bayesian models were all significantly differently expressed in the two mean t-tests, except NTNG1 (**Figure 3A and 7A,B**). This suggests that these two models favored features that have primarily different values in CVD and healthy groups, while the inclusion of NTNG1 served as a supplement to address the remaining variations that had not been explained by the other factors. Stochastic gradient boosting got the ideal classification of CVD and healthy individuals in both the training set and test set with a superior discrimination of CVD and healthy probability than the linear discriminant analysis and Naïve Bayesian model (**Figure 6 and Table 2**). The stochastic gradient boosting model is a classification tree-based model that incorporates functionalities that allows iterations over different choices of features and the modeling of residuals at costs until the residual is smaller than a certain threshold. From our modeling practice, the complexity of the stochastic gradient boosting paid off in comparison to the other two tree-based models—namely, the bagged CART model and the random forest model (**Figure 6** and **Figure S2C-E**). The bagged CART model, which does not iterate among different sets of features, suffered from a larger variation of predicted probabilities (**Figure S2C,D**), suggesting that randomized feature selection did benefit the modeling stability. On the other hand, the random forest model had the ability to produce probability predictions with less variation but failed to deliver a probability above the cutoff in the test set, confirming the extra learning ability made possible by the gradient boosting process (**Figure S2C,E**). Notably, the stochastic gradient boosting model picked considerably different NGCs as features. The most important three features in this model, SEMA6B, SEMA6D and EPHA2, are not included in the top 10 features in the linear discriminant analysis and Naïve Bayesian models (**Figure 7**). The most important feature, SEMA6B, was not even significant in the two mean t-tests (**Figure 3A**). Variable importance in the stochastic gradient boosting model took both the importance of a variable in the building of a decision tree and the subsequent modeling of residuals into account. It is likely that SEMA6B performed well in the modeling of the residual, since the univariate prediction power of SEMA6B should be minimal. The modeling of residuals often improves datasets with lower information-to-noise ratios, which is the case in the current

dataset. The importance of a feature in certain models could sometimes inform us with the relevance of the feature to the outcome. In the current setup, the importance of an NGC in the prediction models might suggest the functional importance of NGC in the development of CVD via monocytes. The linear discriminant analysis and Naïve Bayesian model picked up PLXNC1 as the most important feature. PLXNC1 mediates monocyte migration, adhesion and differentiation in response to its ligand SEMA7A [24]. At regions experiencing disturbed blood flow (atheroprone), an increase of SEMA7A in endothelial cells exacerbates inflammation and atherosclerotic plaque size [12]. SEMA6D and EPHA2 were chosen as the second and third most important features in the stochastic gradient boost model. SEMA6D has a role in immunology as being costimulatory molecule-expressed by dendritic cells [25]. The function of SEMA6D in monocytes is not known yet. EPHA2 promotes the adhesion and differentiation of monocytes [26,27]. In an apolipoprotein E knockout murine atherosclerosis model, the knockdown of EPHA2 using adenovirus-carrying short hairpin RNA resulted in the attenuation of atherosclerotic lesion development [28]. However, the amount of contribution from EPHA2 knockdown on monocytes could not be distinguished from that of endothelial cells. A limitation of this study is that the number of individuals is small in the healthy group, meaning the models were less trained by features of individuals with a healthy phenotype. This also led to some degree of instability of the models in the cross-validation. Another limitation is that the healthy individuals were younger than patients in the CVD group. Although, the ranges of ages in the two groups overlapped, it is to be determined whether the model could discriminate between the two phenotypes when the age range in the healthy group is extended. Future studies should recruit more age-matched individuals. To fully reveal the prediction power of NGC expressions, future works should focus on whether NGC expressions could discriminate between classes with more subtle differences, e.g. between stable and unstable angina. It should also be noted that the pathogenesis of CVD is rather complex. In this article, we focused on the association of NGC expressions to CVD, although multiple pathways are involved, leading to the identification of an incomplete risk gene set. The complex pathogenesis also dampens our ability to draw conclusions on causality, since it is possible that mechanisms that are implicated in CVD also alter the functional states of monocytes, reflected by monocytic gene expressions. In general, the development of clinical risk prediction models often faces certain hurdles, which could lead to less-defined results. Some studies only examined the univariate prediction value and ignored the combined prediction value of all features [29], while other studies did examine the

multivariate prediction value of features but used models that made strong assumptions on the data structure and covariance between features [30,31]. It is also common that studies have accessed the models in a descriptive way, but the external prediction potential is not determined [29,31,32]. In our study, we have incorporated multiple NGCs as features based on both statistical examinations and biological insights. We assessed the performances with cross-validation in the training data and independent prediction in the held-out data, thereby controlling the models on overfitting. Moreover, the models were shown to be not only a descriptive tool to confirm the correlations between NGC expression and CVD outcome but, also, a prediction method that can be applied to new datasets. In addition, we applied one of the complex models, the stochastic gradient boosting, which makes little assumptions on the characteristics of the data structure and returned a better performance. Taken together, this study gave the proof of principle on how machine-learning methods could be applied to the prediction of disease outcomes using the expression of a set of functional genes in circulating cells and allowed us to identify SEMA6B, SEMA6D and EPHA2 as predictive genes for CVD in the current cohort.

## Acknowledgement

This research was funded by the Netherlands Heart Foundation, grant number 2013T127, European Research Area Network on Cardiovascular Diseases, grant number 038 MISsCVD and Centre for Translational Molecular Medicine, grant number 01C-102.

## Conflicts of Interest

The authors declare no conflicts of interest.

## References

1. *Frostegard, J., Immunity, atherosclerosis and cardiovascular disease. BMC Med, 2013. 11: p. 117.*
2. *Ross, R., Atherosclerosis–an inflammatory disease. N Engl J Med, 1999. 340(2): p. 115-126.*
3. *Hansson, G.K. and P. Libby, The immune response in atherosclerosis: a double-edged sword. Nat Rev Immunol, 2006. 6(7): p. 508-519.*
4. *Lessner, S.M., H.L. Prado, E.K. Waller, and Z.S. Galis, Atherosclerotic lesions grow through recruitment and proliferation of circulating monocytes in a murine model. Am*

*J Pathol, 2002. 160(6): p. 2145-2155.*

5. *Osterud, B. and E. Bjorklid, Role of monocytes in atherogenesis. Physiol Rev, 2003. 83(4): p. 1069-1112.*
6. *Hansson, G.K., Inflammation, atherosclerosis, and coronary artery disease. N Engl J Med, 2005. 352(16): p. 1685-1695.*
7. *Ley, K., Y.I. Miller, and C.C. Hedrick, Monocyte and macrophage dynamics during atherogenesis. Arterioscler Thromb Vasc Biol, 2011. 31(7): p. 1506-1516.*
8. *Funk, S.D. and A.W. Orr, Ephs and ephrins resurface in inflammation, immunity, and atherosclerosis. Pharmacol Res, 2013. 67(1): p. 42-52.*
9. *Mirakaj, V. and P. Rosenberger, Immunomodulatory Functions of Neuronal Guidance Proteins. Trends Immunol, 2017. 38(6): p. 444-456.*
10. *Zhang, H., D. Vreeken, C.S. Bruikman, A.J. van Zonneveld, and J.M. van Gils, Understanding netrins and semaphorins in mature endothelial cell biology. Pharmacol Res, 2018. 137: p. 1-10.*
11. *Ramkhelawon, B., Y. Yang, J.M. van Gils, B. Hewing, K.J. Rayner, S. Parathath, L. Guo, S. Oldebeken, J.L. Feig, E.A. Fisher, and K.J. Moore, Hypoxia induces netrin-1 and Unc5b in atherosclerotic plaques: mechanism for macrophage retention and survival. Arterioscler Thromb Vasc Biol, 2013. 33(6): p. 1180-1188.*
12. *Hu, S., Y. Liu, T. You, J. Heath, L. Xu, X. Zheng, A. Wang, Y. Wang, F. Li, F. Yang, Y. Cao, H. Zhang, J.M. van Gils, A.J. van Zonneveld, H. Jo, Q. Wu, Y. Zhang, C. Tang, and L. Zhu, Vascular Semaphorin 7A Upregulation by Disturbed Flow Promotes Atherosclerosis Through Endothelial beta1 Integrin. Arterioscler Thromb Vasc Biol, 2018. 38(2): p. 335-343.*
13. *Zhu, L., T.J. Stalker, K.P. Fong, H. Jiang, A. Tran, I. Crichton, E.K. Lee, K.B. Neeves, S.F. Maloney, H. Kikutani, A. Kumanogoh, E. Pure, S.L. Diamond, and L.F. Brass, Disruption of SEMA4D ameliorates platelet hypersensitivity in dyslipidemia and confers protection against the development of atherosclerosis. Arterioscler Thromb Vasc Biol, 2009. 29(7): p. 1039-1045.*
14. *Wanschel, A., T. Seibert, B. Hewing, B. Ramkhelawon, T.D. Ray, J.M. van Gils, K.J. Rayner, J.E. Feig, E.R. O'Brien, E.A. Fisher, and K.J. Moore, Neuroimmune*

*guidance cue Semaphorin 3E is expressed in atherosclerotic plaques and regulates macrophage retention. Arterioscler Thromb Vasc Biol, 2013. 33(5): p. 886-893.*

15. *van Gils, J.M., M.C. Derby, L.R. Fernandes, B. Ramkhelawon, T.D. Ray, K.J. Rayner, S. Parathath, E. Distel, J.L. Feig, J.I. Alvarez-Leite, A.J. Rayner, T.O. McDonald, K.D. O'Brien, L.M. Stuart, E.A. Fisher, A. Lacy-Hulbert, and K.J. Moore, The neuroimmune guidance cue netrin-1 promotes atherosclerosis by inhibiting the emigration of macrophages from plaques. Nat Immunol, 2012. 13(2): p. 136-143.*
16. *Wang, Q., S. Rao, G.Q. Shen, L. Li, D.J. Moliterno, L.K. Newby, W.J. Rogers, R. Cannata, E. Zirzow, R.C. Elston, and E.J. Topol, Premature myocardial infarction novel susceptibility locus on chromosome 1P34-36 identified by genomewide linkage analysis. Am J Hum Genet, 2004. 74(2): p. 262-271.*
17. *Ghosh, S., J. Vivar, C.P. Nelson, C. Willenborg, A.V. Segre, V.P. Makinen, M. Nikpay, J. Erdmann, S. Blankenberg, C. O'Donnell, W. Marz, R. Laaksonen, A.F. Stewart, S.E. Epstein, S.H. Shah, C.B. Granger, S.L. Hazen, S. Kathiresan, M.P. Reilly, X. Yang, T. Quertermous, N.J. Samani, H. Schunkert, T.L. Assimes, and R. McPherson, Systems Genetics Analysis of Genome-Wide Association Study Reveals Novel Associations Between Key Biological Processes and Coronary Artery Disease. Arterioscler Thromb Vasc Biol, 2015. 35(7): p. 1712-1722.*
18. *van der Harst, P. and N. Verweij, Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circ Res, 2018. 122(3): p. 433-443.*
19. *Casamassimi, A., A. Federico, M. Rienzo, S. Esposito, and A. Ciccodicola, Transcriptome Profiling in Human Diseases: New Advances and Perspectives. Int J Mol Sci, 2017. 18(8).*
20. *Hoefer, I.E., J.W. Sels, J.W. Jukema, S. Bergheanu, E. Biessen, E. McClellan, M. Daemen, P. Doevendans, P. de Groot, M. Hillaert, S. Horsman, M. Ilhan, J. Kuiper, N. Pijls, K. Redekop, P. van der Spek, A. Stubbs, E. van de Veer, J. Waltenberger, A.J. van Zonneveld, and G. Pasterkamp, Circulating cells as predictors of secondary manifestations of cardiovascular disease: design of the CIRCULATING CELLS study. Clin Res Cardiol, 2013. 102(11): p. 847-856.*
21. *Thomas, G.D., A.A.J. Hamers, C. Nakao, P. Marcovecchio, A.M. Taylor, C. McSkimming, A.T. Nguyen, C.A. McNamara, and C.C. Hedrick, Human Blood Mono-*

*cyte Subsets: A New Gating Strategy Defined Using Cell Surface Markers Identified by Mass Cytometry. Arterioscler Thromb Vasc Biol, 2017. 37(8): p. 1548-1558.*

22. *Moroni, F., E. Ammirati, G.D. Norata, M. Magnoni, and P.G. Camici, The Role of Monocytes and Macrophages in Human Atherosclerosis, Plaque Neoangiogenesis, and Atherothrombosis. Mediators Inflamm, 2019. 2019: p. 7434376.*

23. *Rogacev, K.S., B. Cremers, A.M. Zawada, S. Seiler, N. Binder, P. Ege, G. Grosse-Dunker, I. Heisel, F. Hornof, J. Jeken, N.M. Rebling, C. Ulrich, B. Scheller, M. Bohm, D. Fliser, and G.H. Heine, CD14++CD16+ monocytes independently predict cardiovascular events: a cohort study of 951 patients referred for elective coronary angiography. J Am Coll Cardiol, 2012. 60(16): p. 1512-20.*

24. *Holmes, S., A.M. Downs, A. Fosberry, P.D. Hayes, D. Michalovich, P. Murdoch, K. Moores, J. Fox, K. Deen, G. Pettman, T. Wattam, and C. Lewis, Sema7A is a potent monocyte stimulator. Scand J Immunol, 2002. 56(3): p. 270-275.*

25. *O'Connor, B.P., S.Y. Eun, Z. Ye, A.L. Zozulya, J.D. Lich, C.B. Moore, H.A. Iocca, K.E. Roney, E.K. Holl, Q.P. Wu, H.W. van Deventer, Z. Fabry, and J.P. Ting, Semaphorin 6D regulates the late phase of CD4+ T cell primary immune responses. Proc Natl Acad Sci U S A, 2008. 105(35): p. 13015-13020.*

26. *Mukai, M., N. Suruga, N. Saeki, and K. Ogawa, EphA receptors and ephrin-A ligands are upregulated by monocytic differentiation/maturation and promote cell adhesion and protrusion formation in HL60 monocytes. BMC Cell Biol, 2017. 18(1): p. 28.*

27. *Saeki, N., S. Nishino, T. Shimizu, and K. Ogawa, EphA2 promotes cell adhesion and spreading of monocyte and monocyte/macrophage cell lines on integrin ligand-coated surfaces. Cell Adh Migr, 2015. 9(6): p. 469-482.*

28. *Jiang, H., X. Li, X. Zhang, Y. Liu, S. Huang, and X. Wang, EphA2 knockdown attenuates atherosclerotic lesion development in ApoE(-/-) mice. Cardiovasc Pathol, 2014. 23(3): p. 169-174.*

29. *Zawadzki, M., M. Krzystek-Korpacka, A. Gamian, and W. Witkiewicz, Serum cytokines in early prediction of anastomotic leakage following low anterior resection. Wideochir Inne Tech Maloinwazyjne, 2018. 13(1): p. 33-43.*

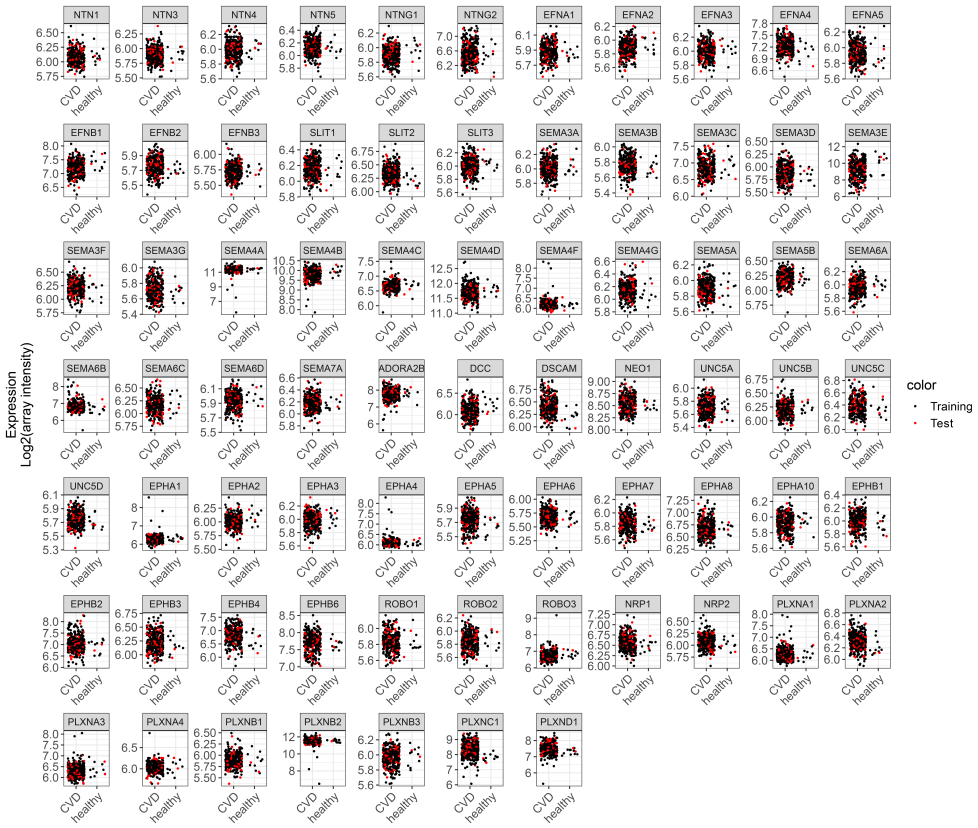
30. *Shimanuki, M., Y. Imanishi, Y. Sato, N. Nakahara, D. Totsuka, E. Sato, S. Iguchi, Y. Sato, K. Soma, Y. Araki, S. Shigetomi, S. Yoshida, K. Uno, Y. Ogawa, T. Tominaga,*

*Y. Ikari, J. Nagayama, A. Endo, K. Miura, T. Tomioka, H. Ozawa, and K. Ogawa, Pretreatment monocyte counts and neutrophil counts predict the risk for febrile neutropenia in patients undergoing TPF chemotherapy for head and neck squamous cell carcinoma. Oncotarget, 2018. 9(27): p. 18970-18984.*

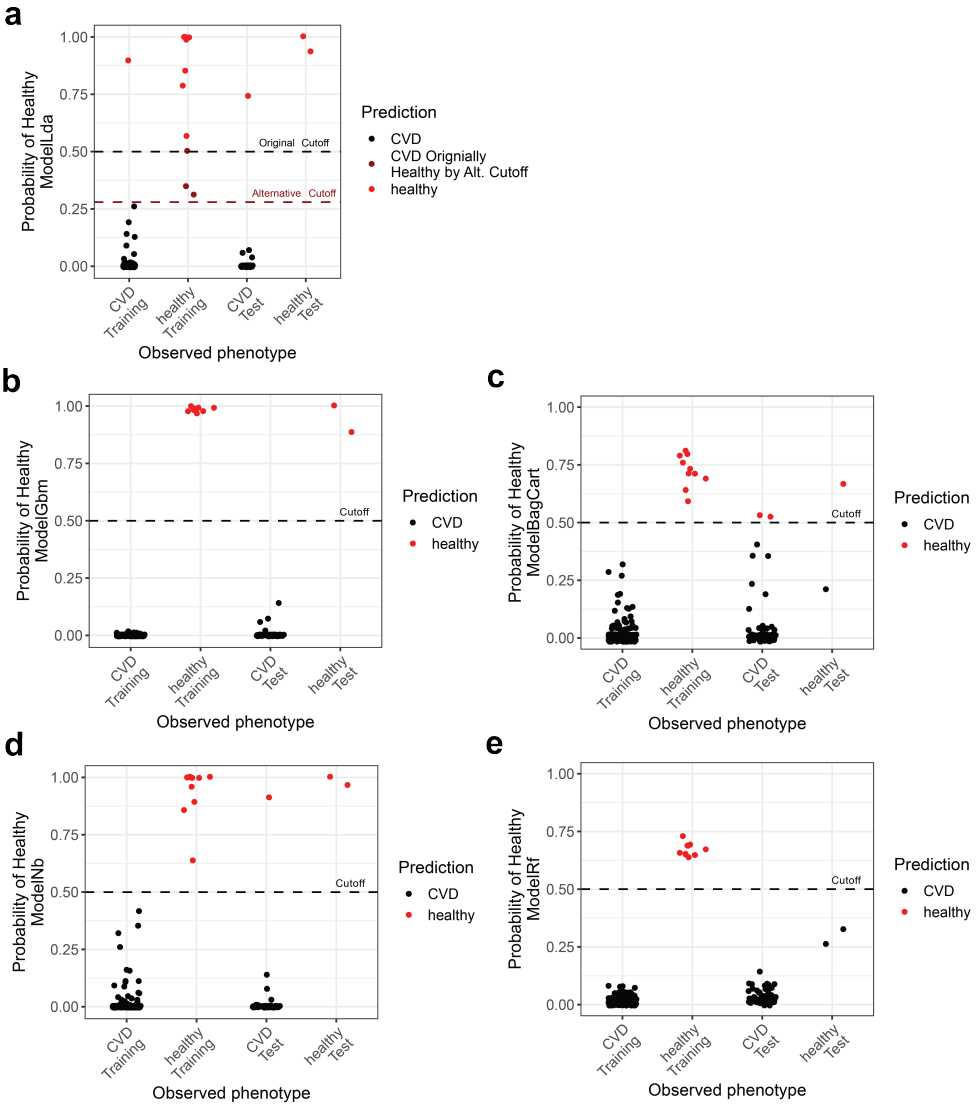
31. *Machado, G.P., G.N. Araujo, C.K. Carpes, M. Lech, S. Mariani, F.H. Valle, L.C.C. Bergoli, S.C. Goncalves, R.V. Wainstein, and M.V. Wainstein, Comparison of neutrophil-to-lymphocyte ratio and mean platelet volume in the prediction of adverse events after primary percutaneous coronary intervention in patients with ST-elevation myocardial infarction. Atherosclerosis, 2018. 274: p. 212-217.*

32. *Guaricci, A.I., V. Lorenzoni, M. Guglielmo, S. Mushtaq, G. Muscogiuri, F. Cademartiri, M. Rabbat, D. Andreini, G. Serviddio, N. Gaibazzi, M. Pepi, and G. Pontone, Prognostic relevance of subclinical coronary and carotid atherosclerosis in a diabetic and nondiabetic asymptomatic population. Clin Cardiol, 2018. 41(6): p. 769-777.*

33. *Kuhn., M., Contributions from Jed Wing and Steve Weston and Andre Williams and Chris Keefer and Allan Engelhardt and Tony Cooper and Zachary Mayer and Brenton Kenkel and the R Core Team and Michael Benesty and Reynald Lescarbeau and Andrew Ziem and Luca Scrucca and Yuan Tang and Can Candan and Tyler Hunt. caret: Classification and Regression Training. R package version 6.0-80, 2018. https://CRAN.R-project.org/package=caret.*




**Supplementary Figure 1. Visualization of data partition**




**Supplementary Figure 2. Visualization of cutoffs for classification**