CLINICAL RESEARCH

Improved cardiovascular risk prediction using targeted plasma proteomics in primary prevention

Renate M. Hoogeveen [1†], **João P. Belo Pereira**[1†], **Nick S. Nurmohamed** [1,2], **Veronica Zampoleri**[3], **Michiel J. Bom**[2], **Andrea Baragetti** [3], **S. Matthijs Boekholdt**[4], **Paul Knaapen**[2], **Kay-Tee Khaw** [5], **Nicholas J. Wareham** [6], **Albert K. Groen**[1], **Alberico L. Catapano** [3,7], **Wolfgang Koenig**[8,9,10], **Evgeni Levin**[1,11], and **Erik S.G. Stroes** [1]*

[1]Department of Vascular Medicine, Amsterdam University Medical Centers, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands; [2]Department of Cardiology, Amsterdam University Medical Centers, Vrije Universiteit Amsterdam, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands; [3]Department of Pharmacological and Biomolecular Sciences, University of Milan, Via Balzaretti 9, 20133 Milan, Italy; [4]Department of Cardiology, Amsterdam University Medical Centers, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands; [5]Department of Public Health and Primary Care, University of Cambridge, 2 Worts' Causeway, Cambridge, UK; [6]Medical Research Council Epidemiology Unit, University of Cambridge, Cambridge CB2 0QQ, UK; [7] Multimedica IRCCS, Milano, Italy; [8]Klinik für Herz- und Kreislauferkrankungen, Deutsches Herzzentrum München, Technische Universität München, Munich, Germany; [9]DZHK (German Centre for Cardiovascular Research), Partner site Munich Heart Alliance, Munich, Germany; [10]Institute of Epidemiology and Medical Biometry, Ulm University, Ulm, Germany; and [11]HorAlzon BV, Delft, the Netherlands

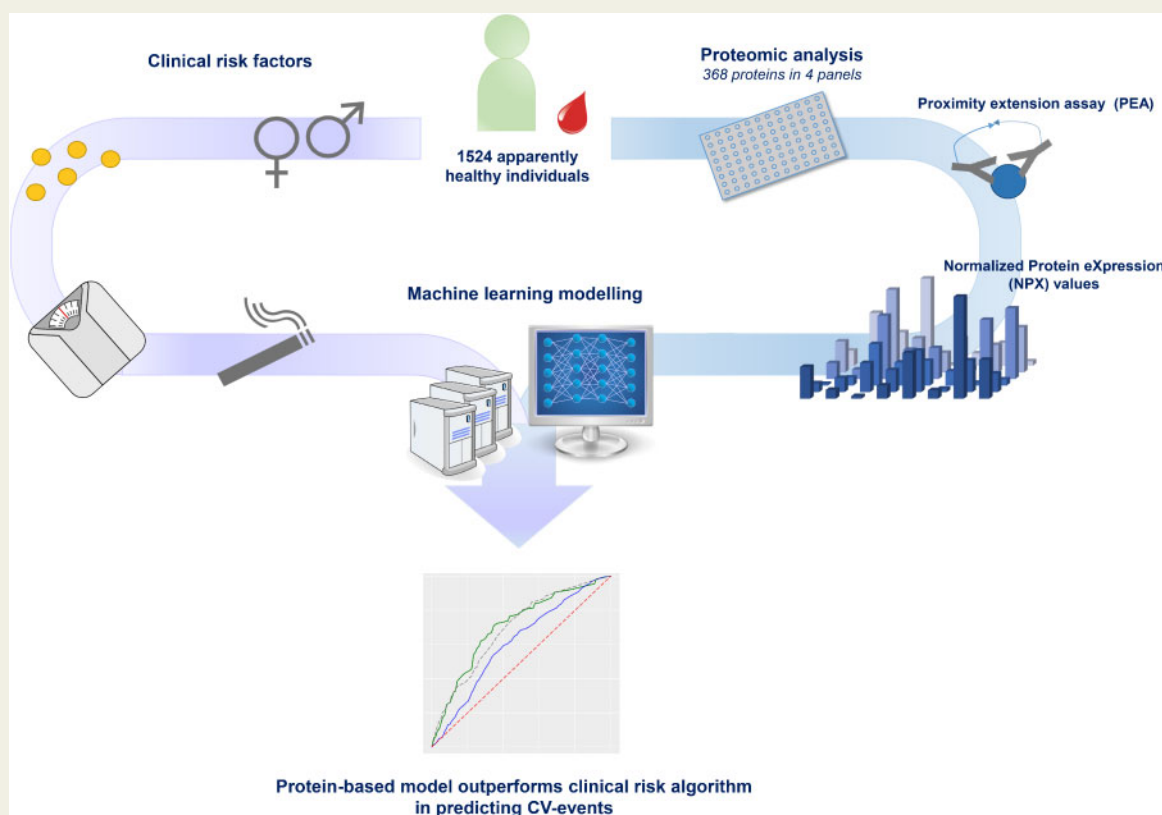|	
Aims	In the era of personalized medicine, it is of utmost importance to be able to identify subjects at the highest cardiovascular (CV) risk. To date, single biomarkers have failed to markedly improve the estimation of CV risk. Using novel technology, simultaneous assessment of large numbers of biomarkers may hold promise to improve prediction. In the present study, we compared a protein-based risk model with a model using traditional risk factors in predicting CV events in the primary prevention setting of the European Prospective Investigation (EPIC)-Norfolk study, followed by validation in the Progressione della Lesione Intimale Carotidea (PLIC) cohort.
Methods and results	Using the proximity extension assay, 368 proteins were measured in a nested case–control sample of 822 individuals from the EPIC-Norfolk prospective cohort study and 702 individuals from the PLIC cohort. Using tree-based ensemble and boosting methods, we constructed a protein-based prediction model, an optimized clinical risk model, and a model combining both. In the derivation cohort (EPIC-Norfolk), we defined a panel of 50 proteins, which outperformed the clinical risk model in the prediction of myocardial infarction [area under the curve (AUC) 0.754 vs. 0.730; $P < 0.001$] during a median follow-up of 20 years. The clinically more relevant prediction of events occurring within 3 years showed an AUC of 0.732 using the clinical risk model and an AUC of 0.803 for the protein model ($P < 0.001$). The predictive value of the protein panel was confirmed to be superior to the clinical risk model in the validation cohort (AUC 0.705 vs. 0.609; $P < 0.001$).
Conclusion	In a primary prevention setting, a proteome-based model outperforms a model comprising clinical risk factors in predicting the risk of CV events. Validation in a large prospective primary prevention cohort is required to address the value for future clinical implementation in CV prevention.

---

* Corresponding author. Tel: +31 20 5665978, Fax: +31 20 6968833, Email: e.s.stroes@amsterdamumc.nl
† The first two authors contributed equally to the study.

Graphical Abstract

Protein-based model outperforms clinical risk algorithm
in predicting CV-events

Introduction

Identification of asymptomatic people at the greatest cardiovascular (CV) risk remains a major challenge in primary prevention.[1,2] Clinically used risk algorithms, including the Framingham risk score, pooled cohort equations, and Systemic Coronary Risk Evaluation (SCORE) system, are based on traditional risk factors for CV disease and predict future events with limited accuracy.[3,4] Accordingly, a substantial proportion of the general population at risk remains unidentified until their first clinical event.[2] Despite adding individual plasma biomarkers such as pro-brain natriuretic peptide (BNP), high sensitivity troponins, and high sensitivity C-reactive protein (CRP) to clinical risk engines, the overall improvement has been limited.[3] This may be explained by the fact that the vast majority of single markers are selected based on specific pathophysiological concepts, which do not reflect the true complexity of atherosclerosis. In fact, CV risk is the result of an interplay between comorbidities (chronic inflammatory diseases, metabolic derangements) and exogenous risk factors,

propagated by a variety of pathophysiological axes, comprising but not limited to lipids, coagulation, and inflammation.[5]

Simultaneous assessment of a large number of plasma proteins may hold a promise to further refine risk assessment.[6] To this end, either discovery proteomics, aiming to identify new diagnostic markers or therapeutic targets, or targeted proteomics, aimed at quantification of proteins of specific interest, can be applied.[6] Widespread use of proteomics has been precluded by labour intensiveness, high costs, and the complex clinical interpretation of the bulky results. More recently, these limitations have largely been resolved. Technical advances now allow for high-throughput proteomic analysis in a reproducible and cost-effective manner.[7] In parallel, advanced computational modelling has facilitated the interpretation of large data sets for clinical implementation.[8,9] Using these innovations, a targeted protein panel was found to modestly improve the prediction of incident atherosclerotic CV disease in primary prevention,[10] whereas Ganz *et al.*[11] substantiated that targeted proteomics also outperformed refit Framingham in predicting recurrent coronary events. In
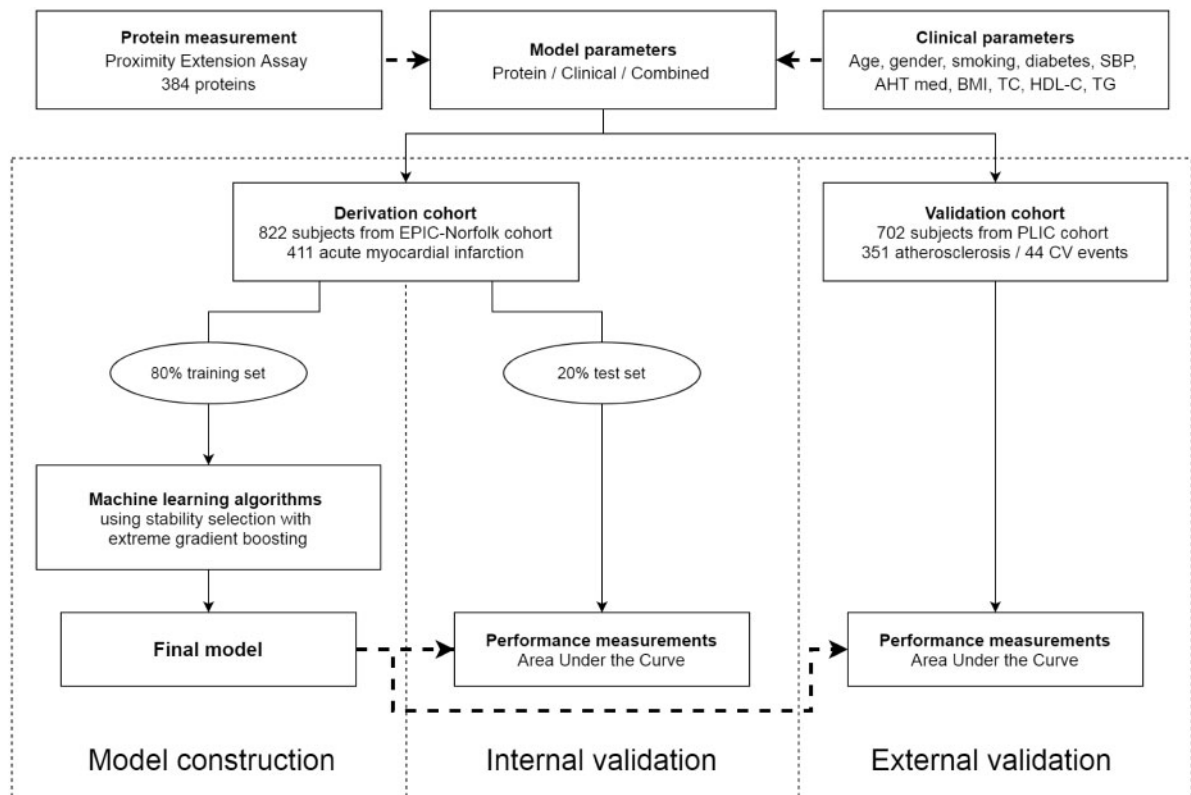
Figure 1 Machine learning workflow of model construction and validation. AHT med, antihypertensive medication; BMI, body mass index; CV, cardiovascular; EPIC, European Prospective Investigation; HDL-C, high-density lipoprotein cholesterol; PLIC, Progressione della Lesione Intimale Carotidea; SBP, systolic blood pressure; TC, total cholesterol; TG, triglycerides.

support, we recently identified two complementary protein signatures predicting the presence of high-risk plaque and the absence of coronary atherosclerosis in subjects referred for the analysis of anginal complaints,[12] clearly outperforming the traditional risk algorithm.[13]

In the present study, we hypothesized that a protein-based risk model can outperform prediction using traditional risk factors in the primary prevention setting. Therefore, we tested the ability of a targeted proteomics panel comprising 368 proteins, related to pathways and/or risk factors involved in atherogenesis, to predict CV event risk in a nested case–control sample of the European Prospective Investigation (EPIC)-Norfolk population study,[14] using advanced machine learning techniques. The findings were subsequently validated in the independent, external primary prevention cohort [Progressione della Lesione Intimale Carotidea (PLIC)].[15]

Methods

Study populations

The derivation cohort was a nested case–control sample derived from the EPIC-Norfolk prospective population study, comprising 25 633 individuals recruited from general practices in the Norfolk area, UK.[14] Study participants aged between 39 and 79 years were enrolled between 1993 and 1997. At baseline, patients completed general health questionnaires and a panel of measurements was performed. During follow-up, all

individuals were flagged for mortality at the UK Office of National Statistics and vital status was ascertained for the entire cohort. Data on all hospital contacts throughout England and Wales were obtained using National Health Service numbers through linkage with the East Norfolk Health Authority (ENCORE) database. Hospital records and death certificates were coded by trained nosologists and categorized according to the International Classification of Disease 10th revision (ICD-10). The study protocol was approved by the Norwich District Health Authority Ethical Committee. All individuals gave written informed consent. For the current study, we selected 822 apparently healthy individuals in a nested case–control sample from the EPIC-Norfolk study. Apparently healthy individuals were defined as study participants who did not report a history of CV disease. A total of 411 individuals who developed an acute myocardial infarction (either hospitalization or death with ICD code I21-22 coded as the underlying cause) between baseline and follow-up through 2016 were selected together with 411 apparently healthy individuals who remained free of any CV disease during follow-up (*Figure 1*).[16,17]

The validation cohort was the PLIC cohort, a single-centre, observational, cross-sectional, and prospective study of subjects enrolled on a voluntary basis in 1998–2000 and followed for 11 years on average in the northern area of Milan.[15] The 2606 Caucasian subjects who were enrolled in the study underwent four periodic visits. Data about clinical, pathological, familial, and pharmacological history and lifestyle habits were collected based on medical records and self-reporting during these visits. Blood samples were withdrawn, and subjects underwent carotid

ultrasound to assess the presence or absence of carotid vascular damage. The presence of documented stenosis or vascular damage on aorta and limb arteries was included in the definition of subclinical atherosclerosis. For the validation cohort, 702 subjects were selected, of whom 351 developed atherosclerosis, comprising subclinical atherosclerosis and 44 subjects who suffered from a CV event, and 351 gender-matched controls during follow-up (*Figure 1*). Cardiovascular events were defined as a combined endpoint of coronary heart disease (myocardial infarction, unstable angina, coronary revascularization, silent ischaemia) and cerebrovascular disease (ischaemic stroke and transient ischaemic attack). This study was approved by the ethics committee and was performed in accordance with the Declaration of Helsinki. All participating subjects signed informed consent.

Biochemical analyses

In EPIC-Norfolk, non-fasting blood was drawn at baseline from study participants, from which total cholesterol, high-density lipoprotein (HDL) cholesterol, and triglycerides were determined with the RA1000 analyser (Bayer Diagnostics, Basingstoke, UK). The Friedewald formula was used for the calculation of low-density lipoprotein (LDL) cholesterol levels.[18] After blood withdrawal, ethylene diamine tetra acetic acid (EDTA) samples were kept overnight at room temperature before transporting to the EPIC-Norfolk laboratory for centrifugation. Hereafter, the remaining plasma was stored at -80°C for future analyses.

In the PLIC cohort, blood samples were collected after overnight fasting. Samples were kept on ice after blood withdrawal and centrifuged within 1 h at 3000 rounds per minute for 12 min (Eppendorf 5810R centrifuge). Plasma samples were subsequently stored in 200 μL aliquots at -80°C. Since multiple aliquots were stored, multiple freeze/thaw cycles were prevented. Total cholesterol, HDL cholesterol, triglyceride and glucose levels were determined in serum samples with the Cobas Mira Plus Analyser (Horiba, ABX, France). Again, the Friedewald formula was used for the calculation of LDL cholesterol levels.[18]

In 2019, we selected cases and controls from both cohorts, whereupon aliquots were thawed and plasma was transferred, on ice, to 96-well plates. The 96-well plates were shipped to Olink proteomics AB (Uppsala, Sweden) on dry ice for analysis using the proximity extension assay technology. Levels of 368 proteins were measured from the CV II, CV III, Cardiometabolic, and Inflammation panels. These panels were selected for their known associations with CV disease. Cases and controls were randomly distributed across plates and assays were performed in a blinded fashion. Data are Normalized Protein eXpression values. Using an internal extension control and an interpolate control, data quality is controlled and normalized. All assay validation data are available on the manufacturer's website (www.olink.com).

Statistical analysis

Data are presented as mean ± standard deviation for normally distributed variables or median with inter-quartile range for skewed data. Categorical variables are expressed as absolute number and percentages. Independent sample *t*-tests and Mann–Whitney *U* tests were used where appropriate. Two-sided *P*-values ≤0.05 were considered statistically significant. Data were analysed using R version 3.5.1 (R Foundation, Vienna, Austria).

Model construction

A combination of stacking generalization framework,[19,20] tree-based ensemble methods, and multiple gradient boosting classifiers[21] was used to best discriminate between cases and controls. Using these techniques, explained in detail below, different models were constructed. First, a clinical risk model was built. The clinical risk model included parameters of different validated risk scores, the Framingham Risk Score, pooled cohort equations, and SCORE. Parameters included in the clinical risk model were age, gender, body mass index, systolic blood pressure, smoking status, and presence of diabetes, the use of antihypertensive medication, total cholesterol levels, HDL cholesterol levels, and triglyceride levels. Second, a protein-based model was constructed using the measured plasma proteins only. A third model was formed by stacking the clinical risk parameters with the protein parameters. The proteins and clinical parameters were allowed to compete in the formation of this model. All three models were validated in the validation cohort without adjustments.

Next, considering the long-term follow-up of subjects and the fact that proteins are subject to change due to lifestyle and medical interventions, we assessed the optimal time point of prediction of acute myocardial infarction in the derivation cohort using Markov-Chain Monte Carlo techniques. For this optimal time point of prediction, similar to the long-term modelling, a clinical risk model, protein model, and a combined model were formed. Specifically for this time point, we calculated the net reclassification improvement (NRI) as described by Pencina *et al.*[22] for case–control studies. Theretofore, we used the acute myocardial infarction prevalence of the total EPIC-Norfolk cohort in the same period.

In addition, we constructed survival models for both the protein model and the clinical risk model in the derivation cohort, to compare model performance across all possible time points. This time-to-event analysis was performed using identical machine learning techniques as the binary models, with the implementation of a survival loss function. Inverse probability of censoring weighting was used to cope with the right-censored data.[23] Using these survival models, time-dependent area under the curves (AUCs) were calculated with a 2-year interval starting from 3 years up to the median follow-up of 20 years.

Machine learning techniques

All binary models were constructed using the same machine learning techniques (*Figure 1*). First, to avoid overfitting of the models, the derivation data set was split into two sets: a training set of 80% and a test set of 20%. The model was not exposed to data from the 20% test set; this was only used for the performance measurements. Ten percentage of the 80% training set was used for model refinement before the model performance was tested in the test set. In construction of the models and identification of the most reliable biomarker signature in our datasets (both proteomics and clinical), we used stability selection with extreme gradient boosting. Gradient boosting is a statistical learning technique, which produces a non-linear model in the form of an ensemble of weak prediction tree-based models. It builds the model in a stage-wise fashion, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. The extreme gradient boosting classification algorithm optimizes a cost function by iteratively choosing a weak hypothesis that points in the negative gradient direction.[20,21] Using a fivefold cross-validation by random reshuffling of the training set, overfitting was avoided. For increased confidence, this procedure was repeated multiple times on a completely reshuffled dataset. Furthermore, the method was coupled with a rigorous stability selection procedure to ensure the reliability and robustness of the obtained parameters.[24] Finally, we applied a permutation (randomization test) to evaluate statistical validity of the results,[25] since standard univariate significance tests cannot be applied to the used models due to the large number of features. The permutation test comprised 1000 reruns of the model, every time randomly permuting the output variable (presence/absence of the event). By evaluating the distribution of all the results obtained in these simulations and comparing it to the true outcomes, we computed statistical significance associated with the joint panel of the selected markers. We also reported

Table 1 Baseline characteristics

	EPIC—case (n = 411)	EPIC—control (n = 411)	PLIC—case (n = 351)	PLIC—control (n = 351)
Age (years)	66 ± 7.8	62 ± 7.7	55 ± 8.1	54 ± 8.2
Male gender	282 (68.6)	254 (61.8)	117 (33.3)	116 (33.0)
BMI (kg/m^2)	26.8 ± 3.7	26.6 ± 3.6	26.9 ± 4.2	26.4 ± 3.2
Systolic blood pressure (mmHg)	144 ± 19	136 ± 17	134 ± 17	130 ± 16
Diastolic blood pressure (mmHg)	86 ± 12	83 ± 11	84 ± 9	82 ± 9
Current smoker	61 (15)	22 (5.4)	78 (22.2)	56 (16.0)
Total cholesterol (mg/dL)[a]	250 ± 47	243 ± 43	225 ± 39	220 ± 38
HDL cholesterol (mg/dL)[a]	50 ± 14	53 ± 15	55 ± 15	58 ± 15
LDL cholesterol (mg/dL)[a]	164 ± 41	157 ± 39	147 ± 37	142 ± 35
Triglycerides (mg/dL)[b]	168 (115–239)	151 (106–222)	102 (66–143)	86 (61–119)
hsCRP (mg/L)	2.1 (1.1–5.0)	1.3 (0.7–2.9)	—	–
HbA1c (%)	5.77 ± 1.28	5.38 ± 0.79	—	—
Antidiabetic drug use baseline	—	—	3 (0.9)	2 (0.6)
Lipid lowering drug use baseline	9 (2.2)	6 (1.5)	—	—
Antihypertensive drug use baseline	150 (36.5)	75 (18.2)	92 (26.2)	68 (19.4)
Median time of follow-up (years)	15.1 (7.7–19.6)	20.5 (19.6–21.2)	11.1 (10.9–11.3)	11.1 (11.0–11.3)

Values are n (%), mean ± standard deviation, or median (IQR) for skewed data.
BMI, body mass index; EPIC, European Prospective Investigation; HDL, high-density lipoprotein; hsCRP, high sensitivity C-reactive protein; IQR, inter-quartile range; LDL, low-density lipoprotein; PLIC, Progressione della Lesione Intimale Carotidea.
[a]To convert to mmol/L, divide with 38.7.
[b]To convert to mmol/L, divide with 88.6.

importance scores for each of the proteins that demonstrate preferences of model when constructing non-linear prediction function based on the selected biomarkers. Python version 3.7 (www.python.org), with packages Numpy, Scipy, and Scikits-learn, was used for machine learning models and visualizations.

Results

Study populations

Baseline characteristics of individuals from the derivation and validation cohort are provided in *Table 1*. In short, cases in the derivation cohort were more likely to have traditional risk factors for CV disease (older, more likely male, smokers and had higher blood pressure and cholesterol levels). In the validation cohort, cases were more likely to be older, to smoke, and have high blood pressure. In all participants, 368 preselected proteins were measured; proteins were excluded if ≥90% of the values were below lower limit of detection. Due to the latter and overlap between the panels, the final analysis included 333 unique proteins (Supplementary material online, *Table S1*).

Prediction of acute myocardial infarction

Prediction of myocardial infarction using a machine learning model consisting of 50 plasma proteins over a median follow-up of 20 years resulted in an receiver operating characteristic (ROC) AUC of 0.754 ± 0.011 (permutation test $P = 0.0099$; *Figures 2* and *3A* and *Table 2*). In comparison, the use of the clinical risk model resulted in an ROC AUC of 0.730 ± 0.015 (permutation test $P = 0.0099$). Combining the protein panel with clinical risk model resulted in an ROC AUC of 0.764 ± 0.015 (permutation test $P = 0.0099$). The

biomarker model was superior to the clinical risk model ($P < 0.001$). The combined protein and clinical risk model showed a small incremental AUC of 0.01 in comparison with the protein model alone. Using Markov-Chain Monte Carlo techniques, the optimal time point for prediction was found at 1132 days (~3 years), which included 66 acute myocardial infarctions and all 411 non-myocardial infarction controls. Of the 50 proteins that were selected for the ~3 years prediction model, 33 overlapped with the original 20 years model (Supplementary material online, *Figure S1*). Focusing on the events occurring within ~3 years after baseline blood withdrawal, the ROC AUC increased to 0.803 ± 0.093 (permutation test $P = 0.0145$; *Figure 3B*), as opposed to 0.732 ± 0.164 (permutation test $P = 0.0099$) using the clinical risk model. The combination of the protein and clinical risk parameters resulted in an ROC AUC of 0.808 ± 0.085 (permutation test $P = 0.0178$). Now, the biomarker model was superior to the clinical risk model with an incremental AUC of 0.07 ($P = 0.025$) but not to the combination of the protein and clinical risk model ($P = 0.721$). For the short-term prediction, the NRI of the protein model in comparison to the clinical risk model was 6.6%. In the survival analysis, the protein model resulted in a mean time-dependent AUC of 0.717 ± 0.027, which was superior across all time points compared to the clinical risk model mean AUC of 0.653 ± 0.031 ($P < 0.001$; Supplementary material online, *Figure S2*).

Validation of the predictive value

We validated the discriminatory ability of the 50 proteins from the derivation cohort in the validation cohort. First, we investigated the ability of the proteins to predict subclinical atherosclerosis. The prediction was relatively poor with an ROC AUC of 0.648 ± 0.056 (permutation test $P = 0.0297$; Supplementary material online, *Figure S3*).
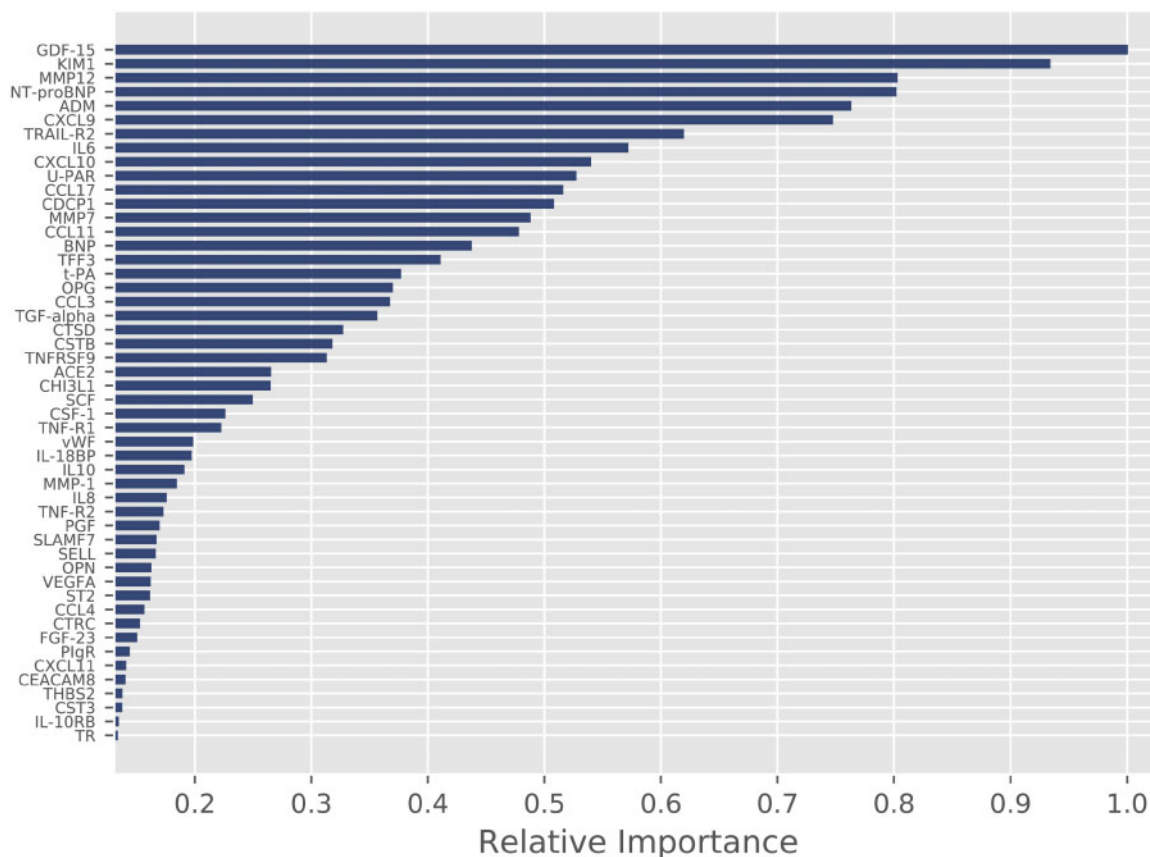
Figure 2 Importance plot of proteins. Relative importance of 50 proteins predictive in derivation cohort.
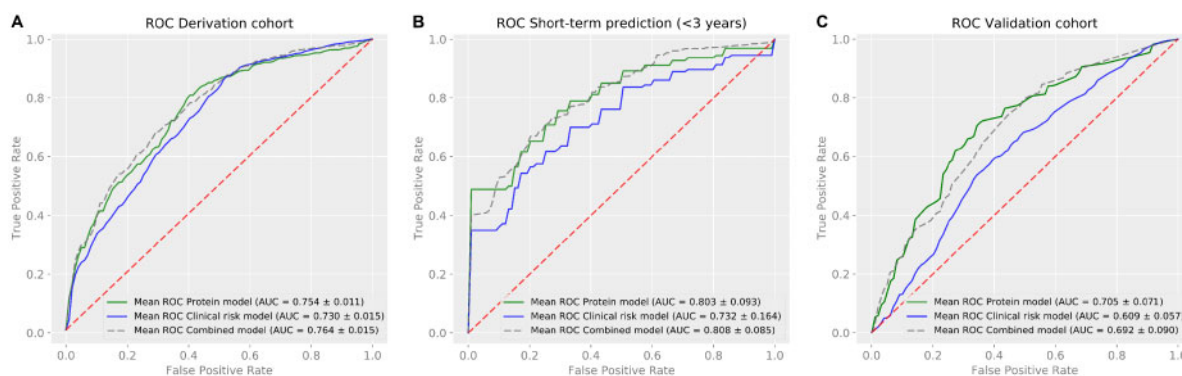
Figure 3 Receiver operating characteristics of prediction models. (*A*) Prediction of events with protein, clinical risk, and combined model in derivation cohort. (*B*) Short-term prediction (<3 years) of events with protein, clinical risk, and combined model in derivation cohort. (*C*) Prediction of events with protein, clinical risk, and combined model in validation cohort. AUC, area under the curve; ROC, receiver operating characteristic.

When validating the proteins in the 44 participants who suffered from CV events vs. the 351 participants with no signs of atherosclerosis, the protein model resulted in an ROC AUC of 0.705 ± 0.071 (permutation test *P* = 0.0099; *Figure 3C*), compared to the clinical risk model ROC AUC 0.609 ± 0.057 (permutation test *P* = 0.0700;

Table 2). The protein model was significantly better than the clinical risk model in the validation cohort (*P* < 0.001). The combined protein and clinical risk model resulted in an ROC AUC of 0.692 ± 0.090 (permutation test *P* = 0.0099), which was not better than the protein model alone (*P* = 0.618).

Table 2　Receiver operating characteristic area under the curve of prediction

	Derivation cohort	Derivation (<3 years)	Validation cohort
Protein model	0.754 ± 0.011	0.803 ± 0.093	0.705 ± 0.071
Clinical risk model	0.730 ± 0.015	0.732 ± 0.164	0.609 ± 0.057
Combined clinical and protein model	0.764 ± 0.015	0.808 ± 0.085	0.692 ± 0.090

Average receiver operating characteristics area under the curve of the prediction models.

Take-home figure　　Derivation and validation of a plasma proteomic model improves cardiovascular risk prediction in a primary prevention setting, demonstrating the potential of a proteomics panel to further refine risk assessment. CV, cardiovascular; NPX, Normalized Protein eXpression; PEA, proximity extension assay.
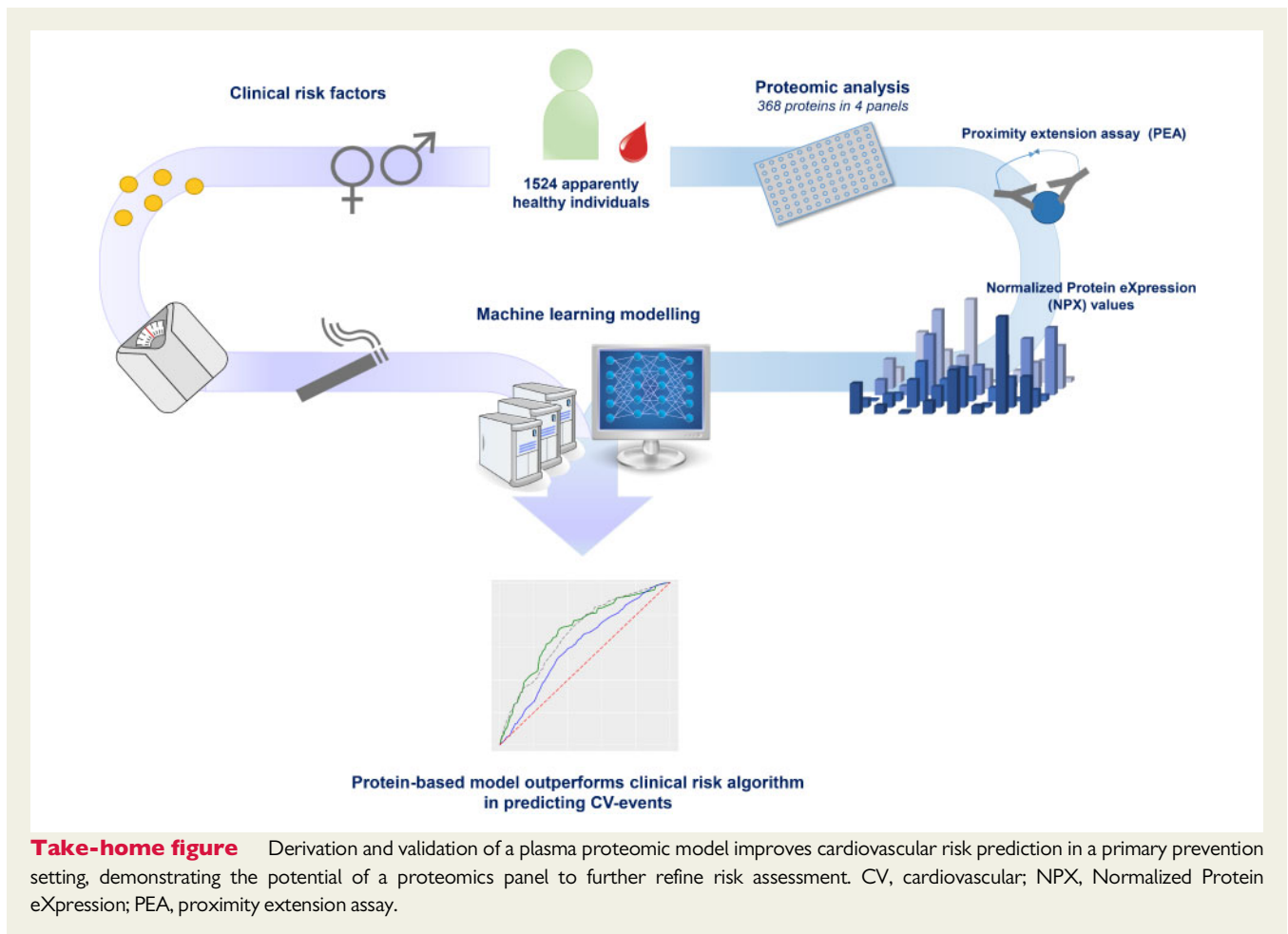
Discussion

Using targeted proteomics, we show that a panel of 50 proteins outperforms the clinical risk model in predicting the risk of myocardial infarction (<3 years) in a primary prevention setting with an AUC increase in the ROC curve of 0.07. Improvement in predicting CV events during the entire (median) 20-year follow-up period was significant, albeit modest. In an external independent validation cohort, the predictive value of the protein panel for CV events was confirmed and superior to the clinical risk model (incremental AUC 0.10). Survival analysis showed superiority of the protein model to the clinical risk model at all tested time points ($P < 0.001$). Collectively, these data show that a novel proteomic panel offers a significant improvement in CV risk discrimination compared to a clinical risk model based on traditional risk factors (*Take-home figure*).

Protein-based risk prediction outperforms traditional risk factors

We substantiate the predictive value of a panel comprising 50 proteins for a first MI with an ROC AUC of 0.754 ± 0.011, using targeted proteomics. Although outperforming the prediction by the clinical risk model ($P < 0.001$), the AUC increase of 0.02 is very modest. Interestingly, the prediction of earlier MI (within 3 years after baseline blood sampling) using the plasma protein panel performed better, with an incremental ROC of 0.07. Where genetic prediction models

are advocated to predict lifelong risk, the ability of our protein model particularly in shorter-term risk prediction most likely highlights the property of plasma proteomics to reflect a more proximate time-frame. Confronted with continuous changes in lifestyle as well as medical interventions during the course of a life, repeated proteome-based risk estimation as a 'liquid health check' may help to further improve lifetime risk estimation.[26] The prediction of predominantly short-term MI substantiates our previous findings that proteomics also predicts the presence of high-risk plaques in patients, which are closely associated with an increased risk for ensuing MI.[27] Previous cohort studies have also reported benefit of proteins in risk prediction. The Framingham Heart Study investigators evaluated a panel of 85 plasma proteins in relation to CV events in primary care setting.[28] Using a multi-marker analysis, they reported eight biomarkers predictive for incident CVD, which on top of clinical parameters achieved an ROC of 0.758. These data are in line with data reported for the prediction of recurrent coronary events, where a panel of 9 out of 1130 proteins modestly improved risk prediction (AUC 0.70) compared to the clinical risk algorithm (AUC 0.64).[11]

Proteins predictive of cardiovascular events

Based on previous findings,[13] we used targeted proteomics using proteins relating to cardiometabolic disease, CV disease, and inflammation/immune responses. The majority of proteins in our model were related to immune system response; particularly proteins involved in chemotaxis, migration, apoptosis, and angiogenesis. Most of the proteins found to predict early vs. all events overlapped (33/50). Several proteins merit further attention considering their marked contribution to the final model. Growth Differentiation Factor 15 (GDF-15) was the protein with the largest contribution. In chronic diseases, GDF-15 produced by leukocytes has been shown to enhance inflammation.[29] Other prominent candidates involve the N-terminal pro-B-type natriuretic peptides and BNP, which are established markers for heart failure and predictors of CV events.[30] There is also a preponderance of inflammatory proteins, comprising metallo proteinase-12 (MMP-12), TRAIL receptor 2 and interleukin-6. These proteins, involved in matrix degradation, apoptosis and inflammation induction, reflect major pathways contributing to atherosclerotic lesion formation and destabilization. Interestingly, there is clear overlap in proteins and pathways when comparing our data to previous CVD-proteomic studies. Thus, GDF-15 was also identified as a predictive candidate in previous studies.[13,28,31] Similarly, the relevance of plasma MMP12[11,13,32,33] and various chemo/cyto-kines[28,31,33] underscore consistency between these studies.

Validation in the Progressione della Lesione Intimale Carotidea cohort

Validation of our findings was performed in the primary prevention PLIC cohort, in which both repetitive non-invasive measures for atherosclerosis and CV events were collected during an 11-year follow-up.[34] The 50-protein model from the derivation cohort showed reasonable prediction of CV events with an ROC AUC of 0.705 ± 0.071, with an incremental AUC of 0.10 compared to the clinical risk model in the PLIC cohort (ROC AUC of 0.609 ± 0.057). We also assessed the value of the proteomic model to predict the presence of

subclinical atherosclerotic lesions assessed using ultrasound, revealing an ROC AUC of 0.648 ± 0.056. The failure of plasma proteomics to accurately predict the presence of subclinical atherosclerosis is in line with the findings in the derivation cohort, where the protein signature performed better for early/mid-term CV events than for long-term events.

Clinical perspective

In previous studies, adding single plasma markers to clinical risk algorithms resulted in only a modest improvements of risk prediction.[35,36] Here, we report a marked improvement in CVD risk prediction using a targeted proteomics approach. The hurdles for using proteomic panels in clinical practice have been largely removed with the advent of affordable high-throughput technology requiring only minimal amounts of plasma. More importantly, machine learning technology further facilitates the use of complex, massive data (such as proteomics) in clinical decision making.[8,9] The need for better discrimination of subjects at highest CV event risk is underscored by the advent of expensive medication in CVD preventive therapy beyond generic statins, among which PCSK9-antibodies,[37,38] low-dose Xa inhibition,[39] SGLT2 inhibitors,[40,41] and GLP1 agonists.[42,43] Whereas a high-risk proteomic panel holds a promise to help identify higher-risk subjects, it is tempting to speculate that pathway analysis of the proteomic signature may also allow for the guidance of what medication to use in specific patient categories.[6] This concept is underscored by the CANTOS study, where predominantly CRP responders demonstrated CV benefit of interleukin 1 beta-antibody administration.[44] However, this concept needs further validation with special emphasis on relationships between biomarkers and protein network analysis.[6,45] Hypothetically, the development of a targeted-proteomic based risk score might enable a more patient-tailored approach for the primary prevention of CV events.

Strengths and limitations

The combination of proteomics with machine learning technology is highly promising.[8,9] Machine learning technology can process data that surpasses the capacity of traditional statistics and the human brain to comprehend.[9] One of the most important differences is that our predictive machine learning model is based on multiple proteins in a panel, which collectively leads to a reliable prediction. Using machine learning, non-linear relationships and interactions among proteins are taken into account, in contrast to univariate models that only address up- and/or down-regulation of individual proteins. In the current analyses, we refitted the clinical risk factors from the Framingham risk score and SCORE to best fit our cohort data, aiming to improve the performance of traditional risk factors. By applying analogous machine learning methods for the traditional risk factors, the observed superiority of our protein model over the clinical risk model is distinct.

Several potential limitations deserve closer attention. First, the cohorts used in this study were collected over a decade ago. Over the years, risk factor management has improved, plaque characteristics have altered, and patient characteristics have changed.[46] Second, our validation cohort had a limited number of CVD events. However, validation of our protein model on these events was reasonable and the model outperformed the clinical risk model in the validation cohort, in an even stronger manner than in the derivation

cohort. Third, we used targeted rather than untargeted proteomics in our study. Proteins were preselected as potential biomarkers for CV disease, since clinical verification, rather than protein discovery was the goal in our study. Despite analysing a broad range of proteins, we may have missed other predictors of CV event risk due to the use of targeted proteomics only. As a result, we may have underestimated the true potential of proteomics in CV risk estimation. Fourth, in contrary to other primary prevention risk scores such as the Pooled Cohort Equation,[4] our constructed models do not predict lifetime risk, which could be useful in primary prevention patients characterized by a relatively low short-term CV risk, such as in subjects below 50 years of age. However, in the present study, we preferred shorter-term prediction for several reasons. Most importantly, the mean age of both our derivation and validation cohorts was well above the age of 50 years, resulting in a higher short-term risk even in primary prevention. Furthermore, diagnostic improvement in detecting high-risk patients is currently needed to make decisions on initiating novel medication on top of routine regimens, and for these decisions, relatively short-time horizons are routinely used. Fifth, the samples in the derivation cohort were non-fasting, while the samples in the validation cohort were collected after an overnight fast. Despite this difference, the protein model performed comparable in the validation cohort. Finally, the current analyses were performed in subjects primarily from European ancestry. Hence, the predictive power remains to be validated in different ethnicities.

Conclusions

In primary prevention, proteome-based risk prediction significantly outperforms prediction using clinical risk factors in predicting the risk of acute myocardial infarction and CV events, especially in the first 3 years. In the midst of novel, expensive drugs, prediction of individual CVD risk and treatment benefit is increasingly important. Further large prospective studies will have to determine the true value of proteome-based risk scores in primary prevention.

Supplementary material

Supplementary material is available at *European Heart Journal* online.

References

1. Yusuf S, Rangarajan S, Teo K, Islam S, Li W, Liu L, Bo J, Lou Q, Lu F, Liu T, Yu L, Zhang S, Mony P, Swaminathan S, Mohan V, Gupta R, Kumar R, Vijayakumar K, Lear S, Anand S, Wielgosz A, Diaz R, Avezum A, Lopez-Jaramillo P, Lanas F, Yusoff K, Ismail N, Iqbal R, Rahman O, Rosengren A, Yusufali A, Kelishadi R, Kruger A, Puoane T, Szuba A, Chifamba J, Oguz A, McQueen M, McKee M, Dagenais G. Cardiovascular risk and events in 17 low-, middle-, and high-income countries. *N Engl J Med* 2014;**371**:818–827.
2. Fernández-Friera L, Fuster V, López-Melgar B, Oliva B, García-Ruiz JM, Mendiguren J, Bueno H, Pocock S, Ibáñez B, Fernández-Ortiz A, Sanz J. Normal LDL-cholesterol levels are associated with subclinical atherosclerosis in the absence of risk factors. *J Am Coll Cardiol* 2017;**70**:2979–2991.
3. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, Cooney M-T, Corrà U, Cosyns B, Deaton C, Graham I, Hall MS, Hobbs FDR, Løchen M-L, Löllgen H, Marques-Vidal P, Perk J, Prescott E, Redon J, Richter DJ, Sattar N, Smulders Y, Tiberi M, Worp HVD, Dis IV, Verschuren WMM; Authors/Task Force Members. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: the Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts. Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation). *Eur Heart J* 2016;**37**:2315–2381.
4. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC, Sorlie P, Stone NJ, Wilson PWF. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American college of cardiology/American heart association task force on practice guidelines. *Circulation* 2014;**129**:S49–S73.
5. Hoogeveen RM, Nahrendorf M, Riksen NP, Netea MG, Winther M. D, Lutgens E, Nordestgaard BG, Neidhart M, Stroes ESG, Catapano AL, Bekkering S. Monocyte and haematopoietic progenitor reprogramming as common mechanism underlying chronic inflammatory and cardiovascular diseases. *Eur Heart J* 2018;**39**:3521–3527.
6. Lindsey ML, Mayr M, Gomes AV, Delles C, Arrell DK, Murphy AM, Lange RA, Costello CE, Jin YF, Laskowitz DT, Sam F, Terzic A, Eyk JV, Srinivas PR. Transformative impact of proteomics on cardiovascular health and disease: a scientific statement from the American Heart Association. *Circulation* 2015;**132**: 852–872.
7. Assarsson E, Lundberg M, Holmquist G, Björkesten J, Bucht Thorsen S, Ekman D, Eriksson A, Rennel Dickens E, Ohlsson S, Edfeldt G, Andersson A-C, Lindstedt P, Stenvang J, Gullberg M, Fredriksson S. Homogenous 96-Plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS One* 2014;**9**:e95192.
8. Deo RC. Machine learning in medicine. *Circulation* 2015;**132**:1920–1930.
9. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019; **380**:1347–1358.
10. Yin X, Subramanian S, Hwang S-J, O'Donnell CJ, Fox CS, Courchesne P, Muntendam P, Gordon N, Adourian A, Juhasz P, Larson MG, Levy D. Prospective study from the systems approach to biomarker. 2014;**34**:939–945.
11. Ganz P, Heidecker B, Hveem K, Jonasson C, Kato S, Segal MR, Sterling DG, Williams SA. Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease. *JAMA* 2016;**315**:2532–2541.
12. Danad I, Raijmakers PG, Driessen RS, Leipsic J, Raju R, Naoum C, Knuuti J, Maki M, Underwood RS, Min JK, Elmore K, Stuijfzand WJ, Royen N, van Tulevski II, Somsen AG, Huisman MC, Lingen AA, van Heymans MW, Ven P. V D, Kuijk C. V, Lammertsma AA, Rossum A. V, Knaapen P. Comparison of coronary CT angiography, SPECT, PET, and hybrid imaging for diagnosis of ischemic heart disease determined by fractional flow reserve. *JAMA Cardiol* 2017;**2**:1100–1107.
13. Bom MJ, Levin E, Driessen RS, Danad I, Kuijk CCV, Rossum A. V, Narula J, Min JK, Leipsic JA, Belo Pereira JP, Taylor CA, Nieuwdorp M, Raijmakers PG, Koenig W, Groen AK, Stroes ESG, Knaapen P. Predictive value of targeted proteomics

for coronary plaque morphology in patients with suspected coronary artery disease. *EBioMedicine* 2019;**39**:109–117.

14. Day N, Oakes S, Luben R, Khaw KT, Bingham S, Welch A, Wareham N. EPIC-Norfolk: study design and characteristics of the cohort. *Br J Cancer* 1999;**80**: 95–103.

15. Elena O, Andrea B, Manuela C, Liliana G, Fabio P, Angela P, Elena T, Luigi CA. Multilevel models to estimate carotid intima-media thickness curves for individual cardiovascular risk evaluation. *Stroke* 2019;**50**:1758–1765.

16. Boekholdt SM, Kuivenhoven JA, Wareham NJ, Peters RJG, Jukema JW, Luben R, Bingham SA, Day NE, Kastelein JJP, Khaw KT. Plasma levels of cholesteryl ester transfer protein and the risk of future coronary artery disease in apparently healthy men and women: the prospective EPIC (European Prospective Investigation into Cancer and nutrition)-Norfolk population study. *Circulation* 2004;**110**:1418–1423.

17. Saleheen D, Scott R, Javad S, Zhao W, Rodrigues A, Picataggi A, Lukmanova D, Mucksavage ML, Luben R, Billheimer J, Kastelein JJP, Boekholdt SM, Khaw KT, Wareham N, Rader DJ. Association of HDL cholesterol efflux capacity with incident coronary heart disease events: a prospective case-control study. *Lancet Diabetes Endocrinol* 2015;**3**:507–513.

18. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem* 1972;**18**:499–502.

19. Wolpert DH. Stacked generalization. *Neural Netw* 1992;**5**:241–259.

20. Caruana R, Niculescu-Mizil A, Crew G, Ksikes A. Ensemble selection from libraries of models. In: *Proceedings of the Twenty-First International Conference on Machine Learning*. Banff, AB, Canada: ACM; 2004. p18.

21. Chen T, Guestrin C, XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA: ACM; 2016. p785–794.

22. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011;**30**:11–21.

23. Vock DM, Wolfson J, Bandyopadhyay S, Adomavicius G, Johnson PE, Vazquez-Benitez G, O'Connor PJ. Adapting machine learning techniques to censored time-to-event health record data: a general-purpose approach using inverse probability of censoring weighting. *J Biomed Inform* 2016;**61**:119–131.

24. Meinshausen N, Bühlmann P. Stability selection. *J R Stat Soc Ser B* 2010;**72**: 417–473.

25. Marques T, Buckland S, Borchers D, Rexstad E, Thomas L. International Encyclopedia of Statistical Science. Berlin, Heidelberg: Springer, 2010. p398–400.

26. Williams SA, Kivimaki M, Langenberg C, Hingorani AD, Casas JP, Bouchard C, Jonasson C, Sarzynski MA, Shipley MJ, Alexander L, Ash J, Bauer T, Chadwick J, Datta G, DeLisle RK, Hagar Y, Hinterberg M, Ostroff R, Weiss S, Ganz P, Wareham NJ. Plasma protein patterns as comprehensive indicators of health. *Nat Med* 2019;**25**:1851–1857.

27. Nerlekar N, Ha FJ, Cheshire C, Rashid H, Cameron JD, Wong DT, Seneviratne S, Brown AJ. Computed tomographic coronary angiography-derived plaque characteristics predict major adverse cardiovascular events: a systematic review and meta-analysis. *Circ Cardiovasc Imaging* 2018;**11**:e006973.

28. Ho JE, Lyass A, Courchesne P, Chen G, Liu C, Yin X, Hwang SJ, Massaro JM, Larson MG, Levy D. Protein biomarkers of cardiovascular disease and mortality in the community. *J Am Heart Assoc* 2018;**7**:e008108.

29. Kempf T, Zarbock A, Widera C, Butz S, Stadtmann A, Rossaint J, Bolomini-Vittori M, Korf-Klingebiel M, Napp LC, Hansen B, Kanwischer A, Bavendiek U, Beutel G, Hapke M, Sauer MG, Laudanna C, Hogg N, Vestweber D, Wollert KC. GDF-15 is an inhibitor of leukocyte integrin activation required for survival after myocardial infarction in mice. *Nat Med* 2011;**17**:581–588.

30. Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robins SJ, Benjamin EJ, D'Agostino RB, Vasan RS. Multiple biomarkers for the prediction of first major cardiovascular events and death. *N Engl J Med* 2006;**355**:2631–2639.

31. Lind L, Siegbahn A, Lindahl B, Stenemo M, Sundström J, Ärnlöv J. Discovery of new risk markers for ischemic stroke using a novel targeted proteomics chip. *Stroke* 2015;**46**:3340–3347.

32. Stenemo M, Nowak C, Byberg L, Sundström J, Giedraitis V, Lind L, Ingelsson E, Fall T, Ärnlöv J. Circulating proteins as predictors of incident heart failure in the elderly. *Eur J Heart Fail* 2018;**20**:55–62.

33. Nowak C, Carlsson AC, Östgren CJ, Nyström FH, Alam M, Feldreich T, Sundström J, Carrero JJ, Leppert J, Hedberg P, Henriksen E, Cordeiro AC, Giedraitis V, Lind L, Ingelsson E, Fall T, Ärnlöv J. Multiplex proteomics for prediction of major cardiovascular events in type 2 diabetes. *Diabetologia* 2018;**61**: 1748–1757.

34. Olmastroni E, Shlyakhto EV, Konradi AO, Rotar OP, Alieva AS, Boyarinova MA, Baragetti A, Grigore L, Pellegatta F, Tragni E, Catapano AL, Casula M. Epidemiology of cardiovascular risk factors in two population-based studies. *Atheroscler Suppl* 2018;**35**:e14–e20.

35. Curry SJ, Krist AH, Owens DK, Barry MJ, Caughey AB, Davidson KW, Doubeni CA, Epling JW, Kemper AR, Kubik M, Landefeld CS, Mangione CM, Silverstein M, Simon MA, Tseng CW, Wong JB; US Preventive Services Task Force. Risk assessment for cardiovascular disease with nontraditional risk factors: US preventive services task force recommendation statement. *JAMA* 2018;**320**:272–280.

36. Mortensen MB, Falk E, Li D, Nasir K, Blaha MJ, Sandfort V, Rodriguez CJ, Ouyang P, Budoff M. Statin trials, cardiovascular events, and coronary artery calcification. *JACC Cardiovasc Imaging* 2018;**11**:221–230.

37. Schwartz GG, Steg PG, Szarek M, Bhatt DL, Bittner VA, Diaz R, Edelberg JM, Goodman SG, Hanotin C, Harrington RA, Jukema JW, Lecorps G, Mahaffey KW, Moryusef A, Pordy R, Quintero K, Roe MT, Sasiela WJ, Tamby J-F, Tricoci P, White HD, Zeiher AM. Alirocumab and cardiovascular outcomes after acute coronary syndrome. *N Engl J Med* 2018;**379**:2097–2107.

38. Sabatine MS, Giugliano RP, Keech AC, Honarpour N, Wiviott SD, Murphy SA, Kuder JF, Wang H, Liu T, Wasserman SM, Sever PS, Pedersen TR. Evolocumab and clinical outcomes in patients with cardiovascular disease. *N Engl J Med* 2017; **376**:1713–1722.

39. Eikelboom JW, Connolly SJ, Bosch J, Dagenais GR, Hart RG, Shestakovska O, Diaz R, Alings M, Lonn EM, Anand SS, Widimsky P, Hori M, Avezum A, Piegas LS, Branch KRH, Probstfield J, Bhatt DL, Zhu J, Liang Y, Maggioni AP, Lopez-Jaramillo P, O'Donnell M, Kakkar AK, Fox KAA, Parkhomenko AN, Ertl G, Störk S, Keltai M, Ryden L, Pogosova N, Dans AL, Lanas F, Commerford PJ, Torp-Pedersen C, Guzik TJ, Verhamme PB, Vinereanu D, Kim J, Tonkin AM, Lewis BS, Felix C, Yusoff K, Steg PG, Metsarinne KP, Cook Bruns N, Misselwitz F, Chen E, Leong D, Yusuf S. Rivaroxaban with or without aspirin in stable cardiovascular disease. *N Engl J Med* 2017;**377**:1319–1330.

40. Perkovic V, Jardine MJ, Neal B, Bompoint S, Heerspink HJL, Charytan DM, Edwards R, Agarwal R, Bakris G, Bull S, Cannon CP, Capuano G, Chu P-L, Zeeuw D. D, Greene T, Levin A, Pollock C, Wheeler DC, Yavin Y, Zhang H, Zinman B, Meininger G, Brenner BM, Mahaffey KW. Canagliflozin and renal outcomes in type 2 diabetes and nephropathy. *N Engl J Med* 2019;**380**:2295–2306.

41. Wiviott SD, Raz I, Bonaca MP, Mosenzon O, Kato ET, Cahn A, Silverman MG, Zelniker TA, Kuder JF, Murphy SA, Bhatt DL, Leiter LA, McGuire DK, Wilding JPH, Ruff CT, Nilsson GI, Fredriksson M, Johansson PA, Langkilde AM, Sabatine MS. Dapagliflozin and cardiovascular outcomes in type 2 diabetes. *N Engl J Med* 2019;**380**:347–357.

42. Husain M, Birkenfeld AL, Donsmark M, Dungan K, Eliaschewitz FG, Franco DR, Jeppesen OK, Lingvay I, Mosenzon O, Pedersen SD, Tack CJ, Thomsen M, Vilsbøll T, Warren ML, Bain SC. Oral semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *N Engl J Med* 2019;**381**:841–851.

43. Marso SP, Daniels GH, Frandsen KB, Kristensen P, Mann JFE, Nauck MA, Nissen SE, Pocock S, Poulter NR, Ravn LS, Steinberg WM, Stockner M, Zinman B, Bergenstal RM, Buse JB. Liraglutide and cardiovascular outcomes in type 2 diabetes. *N Engl J Med* 2016;**375**:311–322.

44. Ridker PM, Everett BM, Thuren T, MacFadyen JG, Chang WH, Ballantyne C, Fonseca F, Nicolau J, Koenig W, Anker SD, Kastelein JJP, Cornel JH, Pais P, Pella D, Genest J, Cifkova R, Lorenzatti A, Forster T, Kobalava Z, Vida-Simiti L, Flather M, Shimokawa H, Ogawa H, Dellborg M, Rossi PRF, Troquay RPT, Libby P, Glynn RJ. Antiinflammatory therapy with canakinumab for atherosclerotic disease. *N Engl J Med* 2017;**377**:1119–1131.

45. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, Ashley E, Dudley JT. Artificial intelligence in cardiology. *J Am Coll Cardiol* 2018;**71**: 2668–2679.

46. Lammeren GV, Ruijter HM, Den Vrijenhoek JEP, Laan SVD, Velema E, Vries JD, Kleijn DD, Vink A, Borst GD, Moll FL, Bots ML, Pasterkamp G. Time-dependent changes in atherosclerotic plaque composition in patients undergoing carotid surgery. *Circulation* 2014;**129**:2269–2276.