JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY VOL. 76, NO. 14, 2020

Machine Learning Improves Cardiovascular Risk Definition for Young, Asymptomatic Individuals

Fátima Sánchez-Cabo, PhD,[a,*] Xavier Rossello, MD, PhD,[a,b,c,*] Valentín Fuster, MD, PhD,[a,d,†] Fernando Benito, MSc,[a] Jose Pedro Manzano, MSc,[a] Juan Carlos Silla, MSc,[a] Juan Miguel Fernández-Alvira, PhD,[a] Belén Oliva, MSc,[a] Leticia Fernández-Friera, MD, PhD,[a,b,e] Beatriz López-Melgar, MD, PhD,[a,e] José María Mendiguren, MD,[f] Javier Sanz, MD,[a,d] Jose María Ordovás, PhD,[a,g,h] Vicente Andrés, PhD,[a,b] Antonio Fernández-Ortiz, MD, PhD,[a,b,i] Héctor Bueno, MD, PhD,[a,j] Borja Ibáñez, MD, PhD,[a,b,k] José Manuel García-Ruiz, MD,[a,b,l] Enrique Lara-Pezzi, PhD[a,b,†]

ABSTRACT

BACKGROUND Clinical practice guidelines recommend assessment of subclinical atherosclerosis using imaging techniques in individuals with intermediate atherosclerotic cardiovascular risk according to standard risk prediction tools.

OBJECTIVES The purpose of this study was to develop a machine-learning model based on routine, quantitative, and easily measured variables to predict the presence and extent of subclinical atherosclerosis (SA) in young, asymptomatic individuals. The risk of having SA estimated by this model could be used to refine risk estimation and optimize the use of imaging for risk assessment.

METHODS The Elastic Net (EN) model was built to predict SA extent, defined by a combined metric of the coronary artery calcification score and 2-dimensional vascular ultrasound. The performance of the model for the prediction of SA extension and progression was compared with traditional risk scores of cardiovascular disease (CVD). An external independent cohort was used for validation.

RESULTS EN-PESA (Progression of Early Subclinical Atherosclerosis) yielded a c-statistic of 0.88 for the prediction of generalized subclinical atherosclerosis. Moreover, EN-PESA was found to be a predictor of 3-year progression independent of the baseline extension of SA. EN-PESA assigned an intermediate to high cardiovascular risk to 40.1% (n = 1,411) of the PESA individuals, a significantly larger number than atherosclerotic CVD (n = 267) and SCORE (Systematic Coronary Risk Evaluation) (n = 507) risk scores. In total, 86.8% of the individuals with an increased risk based on EN-PESA presented signs of SA at baseline or a significant progression of SA over 3 years.

CONCLUSIONS The EN-PESA model uses age, systolic blood pressure, and 10 commonly used blood/urine tests and dietary intake values to identify young, asymptomatic individuals with an increased risk of CVD based on their extension and progression of SA. These individuals are likely to benefit from imaging tests or pharmacological treatment. (Progression of Early Subclinical Atherosclerosis [PESA]; NCT01410318) (J Am Coll Cardiol 2020;76:1674–85) © 2020 The Authors. Published by Elsevier on behalf of the American College of Cardiology Foundation. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Listen to this manuscript's audio summary by Editor-in-Chief Dr. Valentin Fuster on JACC.org.

From the [a]Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), Madrid, Spain; [b]CIBER de enfermedades CardioVasculares (CIBERCV), Spain; [c]Hospital Universitari Son Espases & Health Research Institute of the Balearic Islands (IdISBa), Mallorca, Spain; [d]The Zena and Michael A. Wiener Cardiovascular Institute/Marie-Josée and Henry R. Kravis Center for Cardiovascular Health, Mount Sinai School of Medicine, New York, New York; [e]HM Hospitales-Centro Integral de Enfermedades Cardiovasculares HM CIEC, Madrid, Spain; [f]Banco de Santander, Madrid, Spain; [g]IMDEA Food Institute, CEI UAM + CSIC, Madrid, Spain; [h]U.S. Department of Agriculture Human Nutrition Research Center on Aging, Tufts University, Boston, Massachusetts; [i]Hospital Clínico San Carlos, Madrid, Spain; [j]Hospital Universitario 12 de Octubre, Madrid, Spain; [k]IIS-Fundación Jiménez Díaz Hospital, Madrid, Spain; and the [l]Hospital Universitario Central de Oviedo, Asturias, Spain. *Drs. Sánchez-Cabo and Rossello contributed equally to this work. †Drs. Fuster and Lara-Pezzi are co-corresponding authors. The PESA study is cofunded equally by the Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), Madrid, Spain, and Banco Santander, Madrid, Spain. The study also receives funding from the Instituto de Salud Carlos III (PI15/02019) and the European Regional Development Fund "Una manera de hacer Europa." The CNIC is supported by the Ministerio de Ciencia, Innovacion y Universidades and the Pro CNIC

ISSN 0735-1097 https://doi.org/10.1016/j.jacc.2020.08.017

Atherosclerosis is a systemic disease that has a long asymptomatic phase before manifesting as acute myocardial infarction, stroke, angina, intermittent claudication, or a fatal event. The detection of subclinical atherosclerosis (SA) is thus critical for improving cardiovascular prevention and outcomes (1). In the subclinical phase, imaging techniques have emerged as crucial tools for redefining traditional risk scores, which tend to underestimate midterm and lifetime cardiovascular risk in asymptomatic individuals (2). In particular, the 2019 American College of Cardiology/American Heart Association (ACC/AHA) guidelines for primary prevention of cardiovascular disease (3) and the 2019 European Society of Cardiology/European Atherosclerosis Society guidelines for the management of dyslipidemias (4) recommend the use of noninvasive imaging techniques to assess SA as a risk modifier in asymptomatic individuals at intermediate risk of atherosclerotic cardiovascular disease (ASCVD). Unfortunately, traditional risk scores often fail to appropriately assess risk in young, asymptomatic individuals (5–7), and hence the number of individuals that should be screened using imaging techniques is underestimated.

SEE PAGE 1686

In recent years, several prospective studies have gathered large amounts of longitudinal phenotypic and molecular ("omics") data that improve our understanding about how and when SA leads to cardiovascular events (8–10). Moreover, it remains unclear how a variety of psychosocial, lifestyle, dietary, and demographic variables affects disease initiation and progression (11). The Santander-PESA (Progression of Early Subclinical Atherosclerosis) study (5,12) was conceived to tackle these issues by providing a thorough characterization of SA through a combination of imaging and deep phenotyping in a large cohort of young, asymptomatic participants.

In parallel with these achievements, machine learning (ML) techniques have blossomed, with successful applications in fields ranging from image analysis to natural language processing (13). The strength of ML lies in its capacity to handle large amounts of interrelated predictors to build complex models. However, these techniques require large amounts of accurately collected data (14), and this has limited extensive use of ML techniques in biomedicine (15,16). Despite this, ML-derived scores have been reported recently to outperform traditional risk-score calculators in the prediction of cardiovascular events (17–21).

Using an unbiased, data-driven approach, we used the data produced by the Santander-PESA project to develop a machine-learning model based solely on routine quantitative variables from standard tests (i.e., blood and urine tests or questionnaires) that could serve as an inexpensive, easy-to-calculate estimate of the extent of SA. This model can help physicians identify individuals who would benefit from further imaging studies or preventative therapy, particularly those at intermediate cardiovascular risk.

ABBREVIATIONS AND ACRONYMS

ASCVD = atherosclerotic cardiovascular disease

AUC = area under the curve

CACS = coronary artery calcium score

CT = computed tomography

EN = elastic net

GD = general disease

ML = machine learning

ND = no disease

SA = subclinical atherosclerosis

SCORE = Systematic Coronary Risk Evaluation

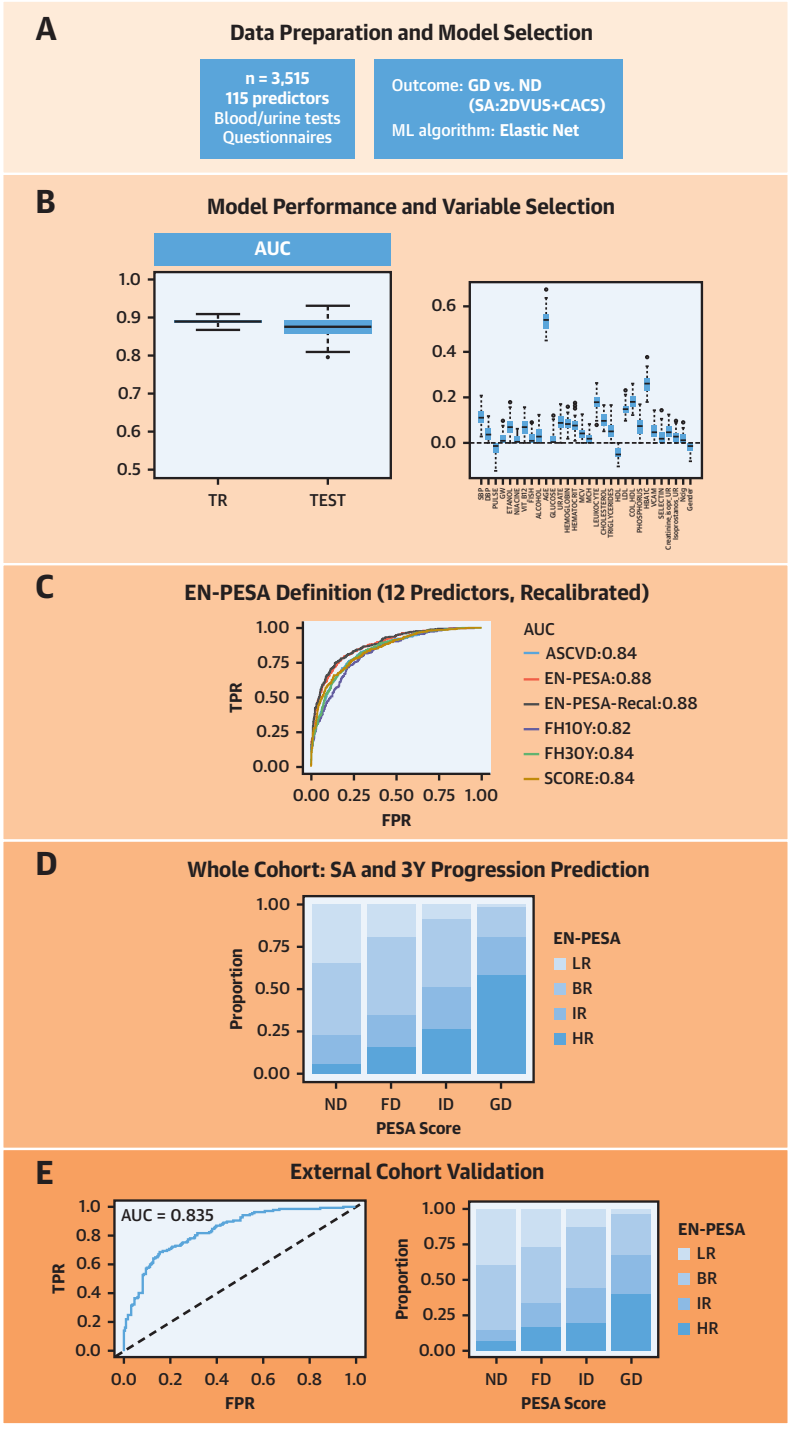
METHODS

STUDY POPULATION AND ASSESSMENT OF THE OUTCOME (SUBCLINICAL ATHEROSCLEROSIS EXTENSION AND PROGRESSION). Since 2010, the CNIC-Santander PESA study prospectively enrolled 4,184 asymptomatic participants between the ages of 40 and 54 years (mean age 45.8 years; 63% men) to evaluate the systemic extent of atherosclerosis. All participants underwent baseline imaging to detect plaques by 2-dimensional vascular ultrasound (2DVUS) in the carotid, abdominal aortic, and femoral territories and by computed tomography (CT) in the coronary arteries (coronary artery calcium score [CACS]) (12). The study protocol was approved by the Instituto de Salud Carlos III Ethics Committee, and all eligible participants provided written informed consent. SA was defined as the presence of ≥1 atherosclerotic plaque in the peripheral territories or

Foundation, and is a Severo Ochoa Center of Excellence (SEV-2015-0505). Dr. Bueno has received research funding from the Instituto de Salud Carlos III, Spain (PIE16/00021 and PI17/01799), AstraZeneca, Bristol-Myers Squibb and Novartis; has received consulting fees from AstraZeneca, Bayer, Bristol-Myers Squibb-Pfizer, and Novartis; and has received speaker fees or support for attending scientific meetings from Amgen, AstraZeneca, Bayer, Bristol-Myers Squibb-Pfizer, Novartis, and MEDSCAPE-the heart.org. All other authors have reported that they have no relationships relevant to the contents of this paper to disclose. Matthew Budoff, MD, served as Guest Associate Editor for this paper. P.K. Shah, MD, served as Guest Editor-in-Chief for this paper. The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the *JACC* author instructions page.

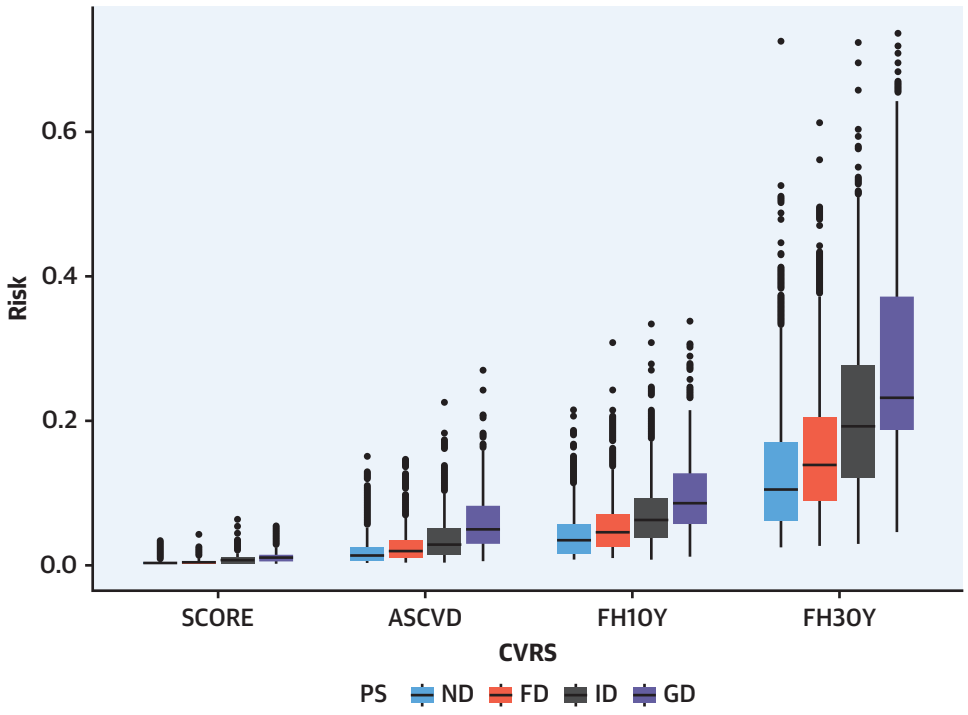
Manuscript received June 14, 2020; revised manuscript received July 27, 2020, accepted August 3, 2020.

FIGURE 1 Analysis Pipeline



(A) Data preparation and selection of the Elastic Net (EN) model. **(B)** Model performance and variable selection using only individuals with GD versus ND based on the PESA Score. **(C)** EN-PESA model definition: recalibration and simplified model based on 12 predictors. Comparison with the performance of traditional risk scores: 10Y and 30Y Framingham (FH10Y, FH30Y), SCORE, and ASCVD risk score. **(D)** Prediction of SA and 3Y progression for PESA participants with any extension of SA. **(E)** Validation in an external independent cohort: AWHS. ASCVD = atherosclerotic cardiovascular disease; AUC = area under the curve; AWHS = Aragon Workers Health Study; BR = borderline risk; CACS = Calcio Score; FD = focal disease; FPR = false positive rate; GD = generalized disease; HR = high risk; ID = intermediate disease; IR = intermediate risk; LR = low risk; ML = machine learning; ND = no disease; SA = subclinical atherosclerosis; TPR = true positive rate; 2DVUS = 2-dimensional vascular ultrasound.

FIGURE 2 Cardiovascular Risk Scores Prediction of SA



Boxplots with the distribution of traditional cardiovascular risk scores across all levels of the PESA Score (PS). For each boxplot, the **horizontal line** within the boxplot represents the median, the **bottom and top lines of the box** represent the first quartile (Q1, 25th percentile) and third quartile (Q3, 75th percentile), respectively. The **upper whisker** extends to the largest value no further than 1.5 × interquartile range (IQR) from Q3 (where IQR is the distance between the first and third quartiles) and the **lower whisker** extends at most 1.5 × IQR below Q1. Data beyond the end of the whiskers are called outliers. CVRS = cardiovascular risk scores; other abbreviations as in Figure 1.

CACS ≥0.5; the extent of SA was defined as the presence of plaque or CACS ≥0.5 for each vascular site (right and left carotid arteries, aorta, right and left femoral arteries, and coronary arteries) (5). This information was used to classify study participants as having no disease (ND) or focal disease (FD) (1 vascular site), intermediate disease (ID) (2 to 3 vascular sites) or generalized disease (GD) (4 to 6 vascular sites). A total of 3,515 individuals remained after filtering participants with no ASCVD risk score (n = 467, missing blood test data, total cholesterol <130 or >320 mg/dl, high-density lipoprotein (HDL) <20 or >100 mg/dl, systolic blood pressure <90 or >200 mm Hg, statin treatment, or LDL >190 mg/dl), no SCORE (Systematic Coronary Risk Evaluation) risk score (n = 74, reported diabetic), no PESA score (n = 74, 2DVUS or CACS data missing), or with missing data in any of the predictors used to build the ML model (n = 54). The baseline characteristics are shown in Supplemental Table 1. Progression was defined as previously published (22). In total, 3,081 of the 3,515 individuals with complete data also had information about their 3-year progression. A logistic mixed model was fit with the R glmer function (R Foundation, Vienna, Austria) to assess if EN-PESA was a predictor of 3-year progression adjusted by the PESA score.

POTENTIAL PREDICTORS AND DATA PREPARATION. In our study, we included 115 quantitative variables measured using routine techniques as potential candidate predictors of SA (Supplemental Table 2). Variables were selected in an unbiased manner from the following categories: physical examination, demographics (age and sex), diet (nutrients or foods), blood and urine tests, and physical activity (objectively quantified by accelerometers). These categories were preferred over others (such as psychosocial variables) because they were considered to be quantitative and accurately measured. Participants' usual energy, nutrient, and food intake was estimated from dietary histories for the preceding 12 months, collected using the ENRICA questionnaire (23); intake for each participant was normalized to body weight. Physical activity data were obtained over 7 days with

TABLE 1 EN-PESA Model

	Estimate	OR	SE	z Value	p Value
Intercept	−2.01729				
Age, yrs	0.97596	2.63	0.07883	12.38	<0.001
HbA1c, %	0.47405	1.61	0.07981	5.94	<0.001
Col/HDL	0.39388	1.48	0.09509	4.142	<0.001
Leukocytes, $\times 10^3$/ul	0.38745	1.47	0.07387	5.245	<0.001
Hemoglobin, mg/dl	0.33174	1.39	0.09663	3.433	<0.001
Vit-B12, g/body kg	0.31686	1.37	0.07288	4.348	<0.001
LDL, mg/dl	0.29515	1.34	0.08956	3.296	<0.001
Phosphorus, g/body kg	0.29505	1.34	0.0722	4.086	<0.001
Systolic blood pressure, mm Hg	0.24012	1.27	0.0802	2.994	0.003
Isoprostanes/creatinine	0.22694	1.25	0.07375	3.077	0.002
Ethanol, g/body kg	0.19059	1.21	0.07162	2.661	0.008
Urate, mg/dl	0.18247	1.19	0.09066	2.013	0.044

General linear model (GLM) to predict generalized subclinical atherosclerosis (SA) using all predictors with an absolute median importance larger than 0.05 across 100 runs of the EN model and with an estimate >0.1 in the general lineal model (Supplemental Table 6). Estimate: Regression coefficient = ln(OR); The description of the variables and their units can be found in Supplemental Table 2. All variables were standardized: 1-U increase in the standardized variable represents the reported OR.

EN-PESA = Elastic Net Progression of Early Subclinical Atherosclerosis; HDL = high-density lipoprotein; LDL = low-density lipoprotein; OR = odds ratio; SE = standard error.

a waist-fitted ActiTrainer accelerometer (Actigraph, Pensacola, Florida). Family history of cardiovascular disease and the number of cigarettes smoked per day for current smokers were also used as predictors. For exploratory purposes, the correlation between predictors was calculated with the Pearson correlation test and plotted using the R ggplot2 package (Supplemental Figure 1). The workflow of the whole analysis pipeline is depicted in **Figure 1**.

COMPARISON OF ML METHODS. A total of 4 ML methods were applied to classify the individuals with GD versus those without any sign of the disease: naive Bayes, Elastic Net (EN), gradient boosting machine, and distributed random forest. We used a naive approach just for exploration splitting the data in 4 groups: 3 were used as training and 1 as test. The mean of the proper (log-loss, Brier Score) and improper rules (area under the curve [AUC], F-scores, balanced accuracy) on the 4 test sets are reported in Supplemental Table 3. EN was the best performer based on all metrics. Further details about the implementation of EN and the definition of the EN-PESA score can be found in the Supplemental Appendix (24-26).

CARDIOVASCULAR RISK SCORES. To test the EN-PESA score, we compared its performance with that of well-established cardiovascular risk scores designed for the prediction of cardiovascular events (27). The cardiovascular risk scores used included the European Society of Cardiology SCORE (Systematic Coronary Risk Evaluation), which calculates 10-year risk of fatal cardiovascular disease (28), and the atherosclerotic cardiovascular disease algorithm for 10-year risk based on Pooled Cohort Equations (ASCVD) (29). For the ASCVD risk score, ACC/AHA guidelines suggest the following risk categories: low risk (LR) (<5%); borderline risk (BR) (5% to 7.4%); intermediate risk (IR) (7.5% to 20%); and high risk (HR) (≥20%). For the SCORE, we used European Heart Association categories: LR (<1%); medium risk (1% to 5%); and HR (≥5%). We also considered the 10- and 30-year risk of coronary heart disease (FH10Y, FH30Y) from the Framingham Heart Study (2). For each cardiovascular risk score, a logistic regression model was built to predict the extent of SA. The distribution of the predicted probabilities for generalized SA were plotted in individuals with GD or ND according to the image-based PESA Score (Supplemental Figure 2).

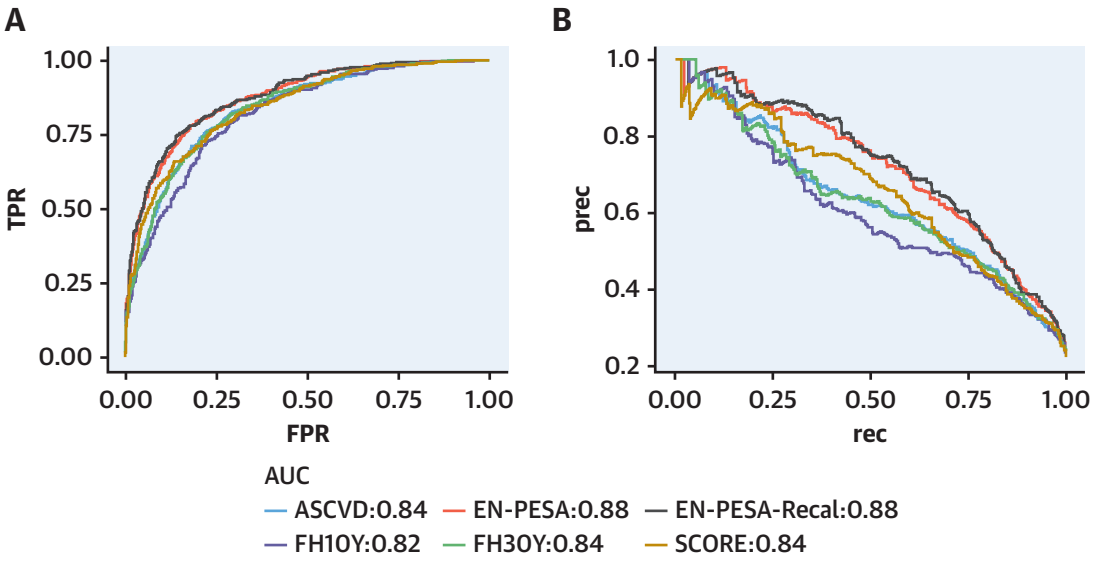
EXTERNAL VALIDATION USING AWHS COHORT. The AWHS (Aragon Workers' Health Study) was designed to assess cardiovascular risk and subclinical atherosclerosis in a cohort of middle-aged men from 40 to 59 years of age (30). Carotid and femoral ultrasound together with noncontrast coronary CT were performed on all participants. Subclinical atherosclerosis was defined similarly to the definition in the PESA study as the presence of any plaque in the aorta, carotid, and femoral arteries and positive CACS. PESA score was calculated using the same definition as in the PESA study. Further details about the data preparation can be found in the Supplemental Appendix.

RESULTS

TRADITIONAL RISK SCORES FAIL TO IDENTIFY INDIVIDUALS AT RISK IN THE PESA COHORT. We first explored the ability of 4 widely used cardiovascular risk scores (FH10Y, FH30Y, ASCVD, and SCORE) to predict SA. **Figure 2** shows the distribution of these scores in PESA participants according to SA extent, from ND to GD. The detailed distribution of cardiovascular risk scores and traditional risk factors across PESA participants is shown in Supplemental Table 1. As expected, all tested scores predicted an increased cardiovascular risk in individuals with a larger extent of SA. However, all 4 scores missed a large proportion of individuals with generalized SA (Supplemental Table 4) (false negatives: 58% to 66%).

We further assessed the capacity of the 2 scores specific for young asymptomatic individuals (i.e., ASCVD and SCORE) to predict the extension of SA (Supplemental Table 5). ASCVD set an intermediate-to-high risk to only 267 individuals, and it missed 88.5% of the individuals with signs of SA in at least 1 territory and 84.9% of those with SA in 2 or more

FIGURE 3 Performance Metrics for the EN-PESA Ensemble Model and the Recalibrated EN-PESA Model

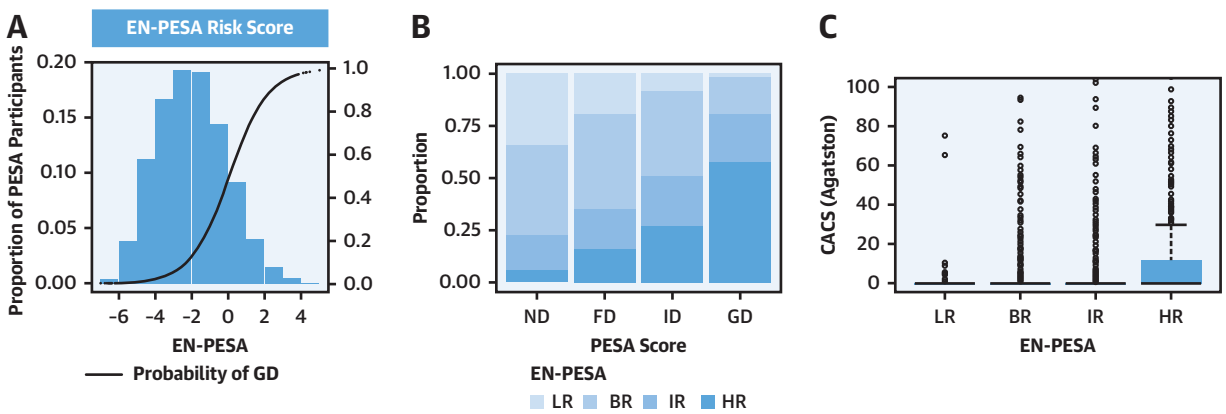


The discriminant ability to distinguish GD versus ND PESA individuals using the ensemble EN-PESA model and the EN-PESA recalibrated model (EN-PESA-recal) were assessed by means of **(A)** AUC and **(B)** Precision/Recall and compared against traditional cardiovascular risk factors: ASCVD, SCORE, FH10Y, and FH30Y. prec = precision; rec = recall; other abbreviations as in Figure 1.

territories. Similar results were found for SCORE. Only 507 individuals were identified as having medium-to-high risk, missing 79.2% of the individuals with signs of SA in at least 1 territory and 72.3% of those with SA in 2 or more territories.

Finally, we assessed the accuracy of traditional risk scores to identify individuals in whom there is a substantial change in SA (known as SA progression) over a 3-year period (22). Only 11.4% of the progressors were assigned an intermediate-to-high risk

FIGURE 4 EN-PESA Score Generalization to PESA Individuals With Any Extension of SA (n = 3,515)



(A) (Right axis, solid line) Probability of generalized disease for all PESA individuals. **(Left axis, histogram)** Proportion of PESA individuals with different levels of the EN-PESA risk score. **(B)** EN-PESA risk distribution for individuals with different SA extension according to the PESA score. **(C)** Distribution of the CACS in Agatston units across different categories of the EN-PESA risk score. For each boxplot, the **horizontal line** within the boxplot represents the median (Q2/50th percentile), the **bottom and top lines of the box** represent the first quartile (Q1/25th percentile) and third quartile (Q3/75th percentile), respectively. The **upper whisker** extends to the largest value no further than 1.5 × IQR from Q3 (where IQR is the interquartile range, or distance between the first and third quartiles) and the **lower whisker** extends at most 1.5 × IQR below Q1. Data beyond the end of the whiskers are called outliers. Abbreviations as in Figures 1 and 2.

TABLE 2 EN-PESA Predicts 3-Year SA Progression Independently of Basal SA Extent

	Estimate	OR	SE	z Value	p Value
Intercept	−0.87				
BR vs. LR	0.33	1.39	0.11	2.90	0.004
IR vs. LR	0.77	2.16	0.13	5.85	<0.001
HR vs. LR	1.12	3.06	0.14	7.94	<0.001

Mixed effects logistic regression model adjusted by PESA score. Estimate: regression coefficient of the mixed effects logistic regression.

BR = EN-PESA borderline risk; HR = EN-PESA high risk; IR = EN-PESA intermediate risk; LR = EN-PESA low risk; OR = odds ratio; SE = standard error.

according to the ASCVD score and only 20.1% based on the SCORE.

These results show that, unlike the proven evidence about the accuracy of traditional risk factors for predicting events, they failed to adequately predict SA extension and progression.

AGE, HBA1C, TOTAL CHOLESTEROL TO HDL RATIO, LEUKOCYTE VOLUME, AND HEMOGLOBIN ARE THE TOP 5 PREDICTORS OF GENERALIZED SA IN THE PESA STUDY. To overcome the limitations of traditional scores, we applied ML techniques to the data gathered in the PESA study (12) to improve the identification of individuals with increased risk of CVD. For this purpose, we built a model to predict the presence of generalized disease versus no disease using the set of quantitative variables in the study (Supplemental Table 2). Mean AUC was 0.89 ± 0.01 for the training sets and 0.88 ± 0.02 for the test sets (Supplemental Figure 3A). Precision and recall were also extremely good and were reproducible across random sets (Supplemental Figures 3B and 3C). The optimum value for the model alpha hyperparameter was 0.2, meaning that there are many variables with similar weight that contribute to the outcome prediction. These can be seen in the homogeneous importance variable of the top variables (Supplemental Figure 3D, Supplemental Table 6). The EN-PESA ensemble model correctly classified 80% of individuals with GD and 79% of those with ND. A general linear model based on these predictors and without collinearities (**Table 1**, Supplemental Table 6) performed similarly to the EN-PESA ensemble model in terms of AUC, precision, and recall (**Figure 3**), but was better calibrated (Supplemental Figure 4) and was hence considered the final model for practical use. The EN-PESA risk score can be calculated online (31).

TABLE 3 EN-PESA Validation in the AWHS Cohort

	Estimate	OR	SE	z Value	p Value
(Intercept)	−0.15677				
Age, yrs	1.03618	2.80	0.17178	6.032	<0.001
Hba1c, %	0.76109	2.14	0.26333	2.89	0.004
Col/HDL	0.59065	1.80	0.16038	3.683	<0.001
Alcohol, g/body kg	0.41333	1.51	0.16338	2.53	0.011
Leukocytes, $\times 10^3$/ul	0.35036	1.41	0.14479	2.42	0.015
Systolic blood pressure, mm Hg	0.28315	1.32	0.14194	1.995	0.046
Hemoglobin, mg/dl	0.23203	1.25	0.14364	1.615	0.106
Vit-B12, g/body kg	0.18751	1.20	0.14402	1.302	0.192
LDL, mg/dl	-0.03491	0.97	0.15365	-0.227	0.820

GLM to predict generalized SA using the 9 predictors from the EN-PESA available at AWHS cohort. Estimate: regression coefficient = ln(OR); The description of the variables can be found in Supplemental Table 2. All variables were standardized: 1-U increase in the standardized variable represents the reported OR.

Abbreviations as in **Table 1**.

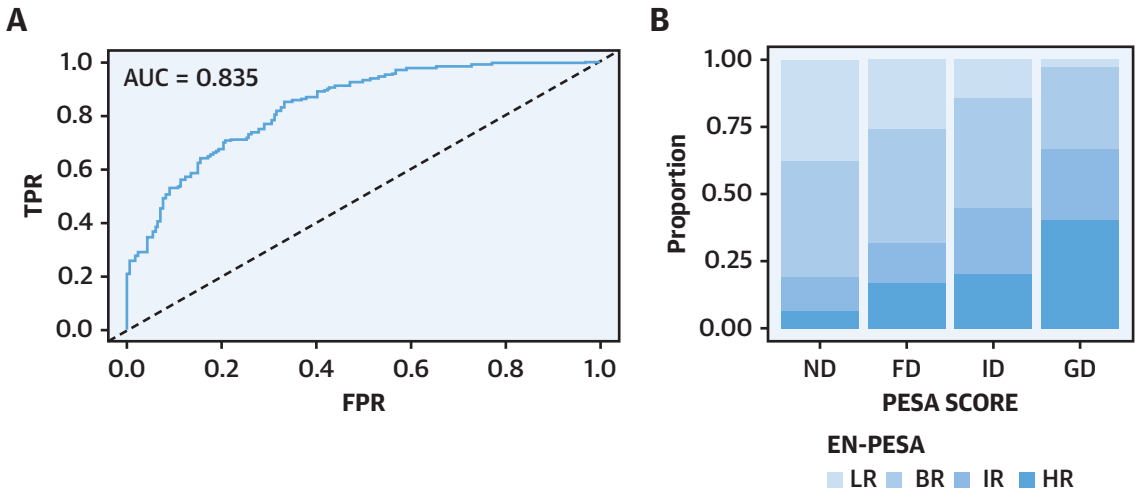
THE EN-PESA ALGORITHM PREDICTS THE EXTENSION OF SA FOR ALL PESA PARTICIPANTS. After the training phase (**Figures 1B and 1C**) with data from individuals with extremes of disease extent (ND vs. GD), we used the EN-PESA score to predict SA extent in all PESA participants (**Figure 1D**). Comparison of observed versus predicted outcomes (**Figure 4A**) showed an adequate goodness-of-fit between the EN-PESA score and the probability of generalized disease. EN-PESA correctly assigned a higher probability of generalized disease to individuals with imaging evidence of SA in multiple vascular territories (**Figure 4B**, Supplemental Table 5).

Because the PESA score combines 2DVUS and coronary artery calcium quantification, we also assessed the ability of the EN-PESA score to detect individuals with a positive CACS. The EN-PESA score performed well in the prediction of which individuals would have high coronary artery calcium levels (**Figure 4C**), classifying 70% of individuals with CACS >0.5 as having an intermediate-to-high risk.

THE EN-PESA SCORE IS A PREDICTOR OF SA PROGRESSION INDEPENDENT OF BASELINE SA EXTENSION. Based on these results, we explored the ability of the EN-PESA score to predict not only SA extension but also disease progression. Individuals with an intermediate EN-PESA risk (IR; n = 604) were 2.16 times more likely to progress in 3 years than individuals with LR EN-PESA (n = 542), independently of their basal extension of SA (**Table 2**). More importantly, individuals with HR EN-PESA (n = 593) were 3.06 times more likely to progress than those individuals with LR. Even individuals with BR (n = 1,342) progressed significantly more than LR individuals.

Overall, EN-PESA assigned an intermediate-to-high risk to 1,411 PESA individuals (40.1%). Of these, 86.8% either progressed in a 3-year period or already had

FIGURE 5 Validation of the EN-PESA in the AWHS Cohort



(A) Receiver-operating curve for the prediction of GD versus ND in the AWHS cohort. **(B)** EN-PESA risk distribution for individuals with different SA extension according to the PESA score in the AWHS cohort. Abbreviations as in Figures 1 to 3.

signs of SA at the basal visit. In particular, 50.4% of the 3-year progressors in the PESA study were identified as having intermediate-to-high risk using the EN-PESA score as opposed to only 11.4% or 20.1% when using the ASCVD or the SCORE (Supplemental Table 7).

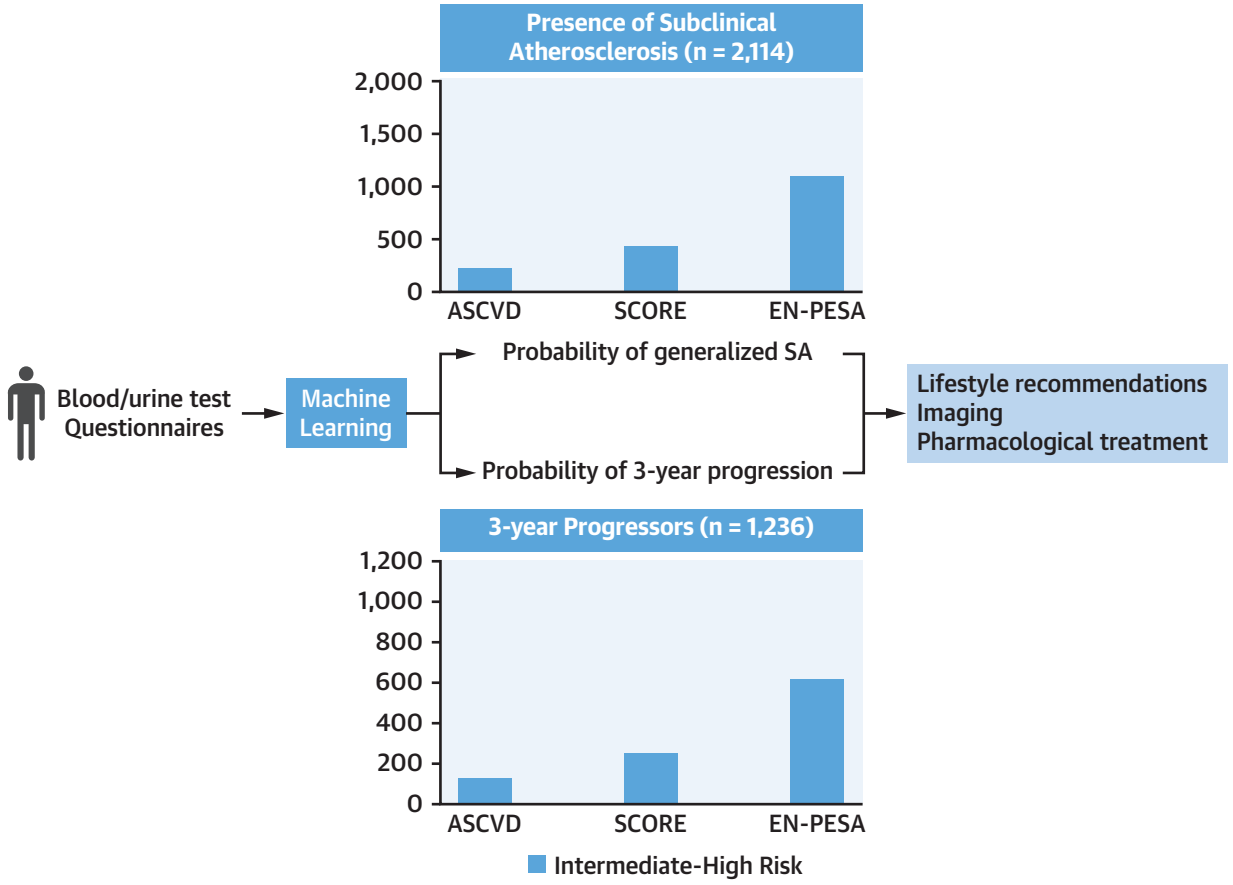
EN-PESA VALIDATION IN THE AWHS. We used the AWHS cohort as an external validation for the EN-PESA score. Nine out of the 12 EN-PESA predictors were available to estimate participants' risk of CVD. Despite the small sample size, all predictors except LDL were independent predictors of generalized SA (**Table 3**), with similar estimates to those observed in the PESA cohort (**Table 1**). In fact, using the same coefficients that were adjusted for the PESA cohort, an AUC of 0.83 was achieved to predict generalized disease (n = 187) versus no disease (n = 159) (**Figure 5A**). Similar to what was observed for the PESA cohort (**Figure 4B**), the EN-PESA score correlated with the extension of SA measured by imaging also for individuals with focal and intermediate extension of SA (**Figure 5B**).

DISCUSSION

There is abundant evidence demonstrating that the detection and quantification of SA extension (32–34) and progression (35,36) with imaging techniques improves risk prediction and classification compared with conventional risk factors. Recent studies have shown that the quantification of CACS by CT improves ASCVD risk stratification (37) and that ultrasound assessment of carotid or femoral plaque burden is similarly effective as CACS at predicting cardiovascular events (33,34). CACS quantification by CT has been suggested as a cost-effective strategy for selecting individuals who would benefit from long-term statin therapy (38). However, current guidelines recommend performing imaging only on those individuals with intermediate risk of cardiovascular events according to traditional scores that might underestimate the risk of young, asymptomatic individuals (**Central Illustration**).

In this study, we present the first unbiased ML tool able to predict the presence and extent of SA from quantitative variables obtained in routine tests, in our analysis collected from the 4,184 asymptomatic participants in the PESA study (12). The advantage of ML methods over other approaches lies in its power to handle hundreds of inter related variables. However, these methods are often referred to as "data-hungry" (14), and this likely explains why, despite impressive results in other disciplines, ML methods have yet to be used extensively in large epidemiological cohort studies. In recent years, several proof-of-concept studies have showed how ML approaches can provide useful insight into CVD (19). The most recent advance showed how coupling of Support Vector

CENTRAL ILLUSTRATION **EN-PESA Is a Machine-Learning Model That Improves Definition of Cardiovascular Risk in Young, Asymptomatic Individuals**



Sánchez-Cabo, F. et al. J Am Coll Cardiol. 2020;76(14):1674–85.

A total of 115 predictors based on urine and blood tests and in diet and demographics questionnaires were used to train a machine-learning algorithm (Elastic Net [EN]) to predict the probability of generalized subclinical atherosclerosis (SA) in the individuals of the PESA (Progression of Early Subclinical Atherosclerosis) cohort. EN-PESA was also an independent predictor of 3-year SA progression, being able to accurately identify a larger number of individuals at risk than traditional risk scores such as atherosclerotic cardiovascular disease (ASCVD) or the SCORE (Systematic Coronary Risk Evaluation) risk scores. Based on these improved predictions more objective decisions might be undertaken about follow-up or intervention plans.

Machines to oversampling techniques improved the prediction of cardiovascular events, outperforming the consensus ACC/AHA CVD Risk Calculator (20). Also, recently, an ML algorithm was able to predict all-cause CVD with relative success based on more than 400,000 individuals (21). The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) consortium is currently preparing guidelines for the reporting results obtained with ML methods (39).

Based on simulated data (14) and on our own experience, EN appears to be a useful method to apply in cohort studies such as PESA. EN not only is able to handle intercorrelated variables, but also provides an understandable model and allows weighting of variables to prioritize predictors in terms of potential interventions. The EN-PESA model ranked the top several variables linked to well-known cardiovascular risk factors such as age, diabetes (hemoglobin A1c), and dyslipidemia (total cholesterol to HDL ratio and LDL cholesterol). Interestingly, hemoglobin A1c is not considered in its whole range in any of the known cardiovascular risk scores, but EN-PESA identified it as the second strongest predictor of SA, a result that was confirmed in the AWHS cohort. This finding illustrates the value of modeling risk factors in their whole range beyond thresholds that dichotomize their meaning, as has already been proven for

LDL cholesterol levels (40). Interestingly, sex was not one of the most predictive variables of the model. Because leukocyte volume differs between men and women, this parameter might be partially capturing the sex-related difference in risk, perhaps adding further information about inflammatory status beyond traditional risk factors.

The main purpose of risk prediction is to support clinical decision-making about the initiation, discontinuation, or intensification of preventive interventions. In general, "high-risk" patients are considered to benefit the most from aggressive risk-factor treatment in terms of absolute risk reduction; however, in meta-analyses of lipid-lowering and blood-pressure-lowering therapy trials, subgroup analyses have shown that relative risk reduction is roughly stable across risk categories (27). This is of particular interest in primary prevention, because there are many more patients at low or intermediate risk than at high risk. Of the 3,515 PESA participants with complete data used in the present study, 1,411 (40.1%) were assigned an intermediate to high risk using the EN-PESA versus only 267 (7.5%) using the ASCVD risk score and 507 (14.4%) using the SCORE. From those selected by the EN-PESA, 86.8% present either SA in at least 1 territory or progression of SA in a 3-year period. It would be interesting to see how the remaining 13.2% progressed in a further 6-year follow-up period. In practical terms, we would suggest performing imaging studies for individuals with an intermediate-to-high risk according to the EN-PESA score, because the large majority of them are expected to already have SA or to develop SA in the near future and are hence more at risk of cardiovascular events.

Finally, we validated the EN-PESA model in an independent external cohort. EN-PESA was also able to identify individuals at higher risk of SA (AUC 0.83) and the key predictors remained independent predictors of generalized SA, confirming the ability of the EN-PESA to identify young, asymptomatic individuals at risk of CVD.

STUDY LIMITATIONS. The PESA study population consists of young (age 40 to 54 years at recruitment), asymptomatic individuals with relatively homogeneous socioeconomic, lifestyle, and ethnic characteristics. Although the EN-PESA was validated in an external cohort, further exploration should be provided in other less-controlled settings to define general risk-category thresholds. In this study, SA is defined as the presence of plaque by 2DVUS in any of the 5 territories studied, plaque calcification in the coronaries, or both. Although most individuals with a generalized extent of disease have coronary calcification, the clinical impact of having plaque in one territory only is still a matter of intense debate. Finally, some of the variables in the EN-PESA model could be expensive to obtain. Further studies on the use of other more standard variables with similar predictive value should be undertaken to allow the applicability of EN-PESA in clinical practice.

CONCLUSIONS

We have built an ML model based on a handful of minimally-invasive, routine, quantitative variables from standard tests (blood test and questionnaires) that could serve as an inexpensive, easy-to-calculate estimate of the extent of subclinical atherosclerosis. Our results show that the EN-PESA model is a useful resource for refining cardiovascular risk estimates, especially for individuals with an inconclusive risk score according to traditional cardiovascular risk scales. It will contribute to further personalize cardiovascular risk, which will translate into more tailored treatments and follow-up plans.

ACKOWLEDGMENTS The authors thank all PESA technical staff: Alberto Ávila, Aurora del Barrio, Iris García, Marta Gavilán, Ángel Macías, Rosario Pérez, Braulio Pérez, Virginia Mass, Inés Gutierrez, Ana Álvarez, Chawar Hayoun, Susana Linares, Vicente Martínez de Vega, Diana Mollinedo and Verónica Muñoz, and cardiologists: Inés García-Lunar, Sandra Gómez-Talavera and Andrea Moreno. The authors also thank the IT team (Sergio Cárdenas and Jesús Molina), PESA manager Antonio Quesada, Alberto Sanz (CNIC General Manager), the medical and technical team at Banco Santander Health Services, Dr. Martin Laclaustra for his help with the AWHS cohort database, and Simon Bartlett (CNIC editing services) for English revision. Finally, the authors are particularly grateful to all of the PESA study participants.

ADDRESS FOR CORRESPONDENCE: Dr. Valentin Fuster, The Zena and Michael A. Wiener Cardiovascular Institute, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy. Place, Box 1030 New York, New York 10029 OR Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), Melchor Fernandez Almagro, 3. 28029 Madrid, Spain. E-mail: vfuster@cnic.es OR valentin.fuster@mountsinai.org. OR Dr. Enrique Lara-Pezzi, Centro Nacional de Investigaciones Cardiovasculares (CNIC), Melchor Fernández Almagro, 3. 28029 Madrid, Spain. E-mail: elara@cnic.es. Twitter: @ELaraPezzi, @fsanchezcabo, @RosselloXavier.

PERSPECTIVES

COMPETENCY IN SYSTEMS-BASED PRACTICE: Automated collection of quantitative data coupled with ML algorithms promise to enhance cardiovascular risk assessment by providing highly accurate estimates of the probability that an apparently healthy individual will develop subclinical or progressive atherosclerosis.

TRANSLATIONAL OUTLOOK: Further research is needed to enhance the analytic algorithms employed for cardiovascular risk assessment and translate prognostic projections into individualized management plans for individuals classified by conventional risk scores as at intermediate cardiovascular risk.

REFERENCES

1. Mozaffarian D, Benjamin EJ, Go AS, et al. Executive summary: heart disease and stroke statistics-2016 update: a report from the American Heart Association. Circulation 2016;133:447–54.

2. D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham heart study. Circulation 2008;117:743–53.

3. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease. J Am Coll Cardiol 2019;140:e596–646.

4. Mach F, Baigent C, Catapano AL, et al. 2019 ESC/EAS guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. Eur Heart J 2019;41:111–88.

5. Fernández-Friera L, Peñalvo JL, Fernández-Ortiz A, et al. Prevalence, vascular distribution, and multiterritorial extent of subclinical atherosclerosis in a middle-aged cohort the PESA (Progression of Early Subclinical Atherosclerosis) study. Circulation 2015;131:2104–13.

6. Piepoli MF, Hoes AW, Agewall S, et al. 2016 European guidelines on cardiovascular disease prevention in clinical practice. Eur Heart J 2016;37:2315–81.

7. López-Melgar B, Fernández-Friera L, Oliva B, et al. Subclinical atherosclerosis burden by 3D ultrasound in mid-life: the PESA Study. J Am Coll Cardiol 2017;70:301–13.

8. Bild DE, Bluemke DA, Burke GL, et al. Multi-Ethnic Study of Atherosclerosis: objectives and design. Am J Epidemiol 2002;156:871–81.

9. Schmermund A, Möhlenkamp S, Stang A, et al. Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the Heinz Nixdorf RECALL study. Am Heart J 2002;144:212–8.

10. Falk E, Sillesen H, Muntendam P, Fuster V. The high-risk plaque initiative: primary prevention of atherothrombotic events in the asymptomatic population. Curr Atheroscler Rep 2011;13:359–66.

11. Ahmadi A, Argulian E, Leipsic J, Newby DE, Narula J. From subclinical atherosclerosis to plaque progression and acute coronary events. J Am Coll Cardiol 2019;74:1608–17.

12. Fernández-Ortiz A, Jiménez-Borreguero LJ, Peñalvo JL, et al. The progression and early detection of subclinical atherosclerosis (PESA) study: rationale and design. Am Heart J 2013;166:990–8.

13. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.

14. Van Der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014;14:137.

15. Shameer K, Badgeley MA, Miotto R, Glicksberg BS, Morgan JW, Dudley JT. Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. Brief Bioinform 2017;18:105–24.

16. Johnson KW, Torres Soto J, Glicksberg BS, et al. Artificial intelligence in cardiology. J Am Coll Cardiol 2018;71:2668–79.

17. Motwani M, Dey D, Berman DS, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. Eur Heart J 2017;38:500–7.

18. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can Machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017;12:e0174944.

19. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular event prediction by machine learning: the Multi-Ethnic Study of Atherosclerosis. Circ Res 2017;121:1092–101.

20. Kakadiaris IA, Vrigkas M, Yen AA, Kuznetsova T, Budoff M, Naghavi M. Machine learning outperforms ACC/AHA CVD risk calculator in MESA. J Am Heart Assoc 2018;7:e009476.

21. Alaa AM, Bolton T, Angelantonio E Di, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. PLoS One 2019;14:e0213653.

22. López-Melgar B, Fernández-Friera L, Oliva B, et al. Short-term progression of multiterritorial subclinical atherosclerosis. J Am Coll Cardiol 2020;75:1617–27.

23. Rodríguez-Artalejo F, Graciani A, Guallar-Castillón P, et al. Rationale and methods of the Study on Nutrition and Cardiovascular Risk in Spain (ENRICA). Rev Esp Cardiol 2011;64:876–82.

24. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1–22.

25. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal 2002;6:429–49.

26. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: Visualizing classifier performance in R. Bioinformatics 2005;21:3940–1.

27. Rossello X, Dorresteijn JAN, Janssen A, et al. Risk prediction tools in cardiovascular disease prevention: a report from the ESC Prevention of CVD Programme led by the European Association of Preventive Cardiology (EAPC) in collaboration with the Acute Cardiovascular Care Association (ACCA) and the Association of Cardiovascular Nursing and Allied Professions (ACNAP). Eur J Cardiovasc Nurs 2019;18:534–44.

28. Piepoli MF, Hoes AW, Agewall S, et al. 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 representati. G Ital Cardiol (Rome) 2017;18:547–612.

29. Budoff MJ, Young R, Burke G, et al. Ten-year association of coronary artery calcium with atherosclerotic cardiovascular disease (ASCVD) events: the Multi-Ethnic Study of Atherosclerosis (MESA). Eur Heart J 2018;39:2401–8.

30. Laclaustra M, Casasnovas JA, Fernández-Ortiz A, et al. Femoral and carotid subclinical atherosclerosis association with risk factors and coronary calcium: the AWHS study. J Am Coll Cardiol 2016;67:1263–74.

31. EN-PESA calculator. Available at: http://bioinfo.cnic.es/ENPESA. Accessed September 8, 2020.

32. Van Der Meer IM, Bots ML, Hofman A, Del Sol AI, Van Der Kuip DAM, Witteman JCM. Predictive value of noninvasive measures of atherosclerosis for incident myocardial infarction: the Rotterdam Study. Circulation 2004;109:1089–94.

33. Baber U, Mehran R, Sartori S, et al. Prevalence, impact, and predictive value of detecting

subclinical coronary and carotid atherosclerosis in asymptomatic adults: the bioimage study. J Am Coll Cardiol 2015;65:1065–74.

34. Sillesen H, Sartori S, Sandholt B, Baber U, Mehran R, Fuster V. Carotid plaque thickness and carotid plaque burden predict future cardiovascular events in asymptomatic adult Americans. Eur Heart J Cardiovasc Imaging 2018;19:1042–50.

35. Rodriguez-Granillo GA, Carrascosa P, Bruining N. Progression of coronary artery calcification at the crossroads: sign of progression or stabilization of coronary atherosclerosis? Cardiovasc Diagn Ther 2016;6:250–8.

36. Wannarong T, Parraga G, Buchanan D, et al. Progression of carotid plaque volume predicts cardiovascular events. Stroke 2013;44:1859–65.

37. Lin JS, Evans CV, Johnson E, Redmond N, Coppola EL, Smith N. Nontraditional risk factors in cardiovascular disease risk assessment: updated evidence report and systematic review for the US Preventive Services Task Force. J Am Med Assoc 2018;320:281–97.

38. Hong JC, Blankstein R, Shaw LJ, et al. Implications of coronary artery calcium testing for treatment decisions among statin candidates according to the ACC/AHA Cholesterol Management Guidelines: A Cost-Effectiveness Analysis. J Am Coll Cardiol Img 2017;10:938–52.

39. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577–9.

40. Fernández-Friera L, Fuster V, López-Melgar B, et al. Normal LDL-cholesterol levels are associated with subclinical atherosclerosis in the absence of risk factors. J Am Coll Cardiol 2017;70:2979–91.

KEY WORDS ASCVD, atherosclerosis, cardiovascular risk scores, machine-learning, subclinical

APPENDIX For an expanded Methods section as well as supplemental tables and figures, please see the online version of this paper.