Archives available at journals.mriindia.com

**International Journal on Advanced Computer Theory and Engineering**

ISSN: 2319 - 2526

Volume 15 Issue 01s, 2026

# Automated Prediction of Heart Disease based on EMR data using Deep Learning Algorithm: A Review and Evaluation Study for various Feature Extraction Techniques

[1]Mrs. Raghuwanshi Sapana Bhushan, [2]Dr. Suryawanshi Nilesh Ashok

*[1] Research Scholar, Gangamai College of Engineering, Nagoan, Dhule – 424005, Maharashtra, India*

*[2] Assistant Professor, Gangamai College of Engineering, Nagoan, Dhule – 424005, Maharashtra, India*

*Email: [1] sapana.salunkhe99@gmail.com, [2] nileshsuryawanshipatil088@gmail.com*

**Peer Review Information**

*Submission: 08 Dec 2025*

*Revision: 25 Dec 2025*

*Acceptance: 10 Jan 2026*

**Keywords**

*Cardiovascular disease, Machine learning, Electronic Medical Record, Risk Mass screening, Electronic health record.*

**Abstract**

For many years, healthcare providers in the United States have relied on the ACC/AHA Pooled Cohort Equations (PCE) Risk Calculator as their main resource for estimating a person's likelihood of developing atherosclerotic cardiovascular disease, a leading form of heart-related illness. Despite being widely used, the calculator occasionally overestimates or underestimates risk but doesn't always perform equally across different groups. To overcome these challenges, we developed an automated ASCVD risk-prediction model tailored to particular patient groups, using machine-learning techniques applied to real-world electronic medical record data. We then compared its performance with the PCE approach. In our study, we reviewed 101,110 electronic medical records from patients seen between January 1, 2009, and April 30, 2020. The machine-learning models were trained using either cross-sectional clinical data alone or a combination of cross-sectional information and longitudinal patterns drawn from lab results and vital-sign measurements. To understand how each model identified true cases, we introduced a cost-focused metric called the "Screened Cases Percentage at a given Sensitivity," which shows how many patients would require follow-up testing to detect most ASCVD cases. In every analysis we conducted, the machine-learning models outperformed the PCE calculator. The strongest results came from a random forest model that used both cross-sectional and longitudinal data, achieving an AUC of 0.902 (95% CI: 0.895–0.910). To identify 90% of ASCVD cases, this approach required screening only 43% of patients, compared with 69% when relying on the PCE. Overall, combining CS and LT data led to accurate predictions and fewer unnecessary screenings.

## Introduction

Athero-sclerotic cardiovascular disease (ASCVD) continues to be a leading global health factor with tremendous medical and economic consequences. North America, the Middle East and Central Asia have highest cardiovascular disease prevalence, but Eastern Europe and Central Asia see the highest cardiovascular mortality rates [1] globally. A cornerstone to ASCVD prevention is early assessment of risk. For those found at high risk, awareness of their ASCVD score can lead to earlier preventive therapy and protect them from unnecessary diagnostics [2-4].

Figure 1. Workflow for building and testing ML models. Data is extracted from EMRs and

filtered. ML models are built on DataMain, compared with each other for performance, and compared against the PCE risk score for the DataPCE subgroup. Abbreviations: ASCVD, atherosclerotic cardiovascular disease; CS, cross-sectional; EMR, electronic medical record; LT, longitudinal; machine-learning models: RF, random forest; LR, logistic regression; NN, neural networks; NB, naïve Bayes. Improving ASCVD risk predictions is becoming more and more important in routine clinical practice in order to lower the cost of preventive care and lessen reliance on costly or invasive testing [5,6]. To gauge someone's likelihood of experiencing a major ASCVD event within the next ten years—such as a non-fatal heart attack, a non-fatal stroke, or death caused by coronary heart disease or stroke—current clinical guidelines primarily depend on the pooled cohort equations (PCE) [4,7]. Although there are other scoring systems, they are frequently designed to target particular populations or cardiovascular outcomes [8, 9].

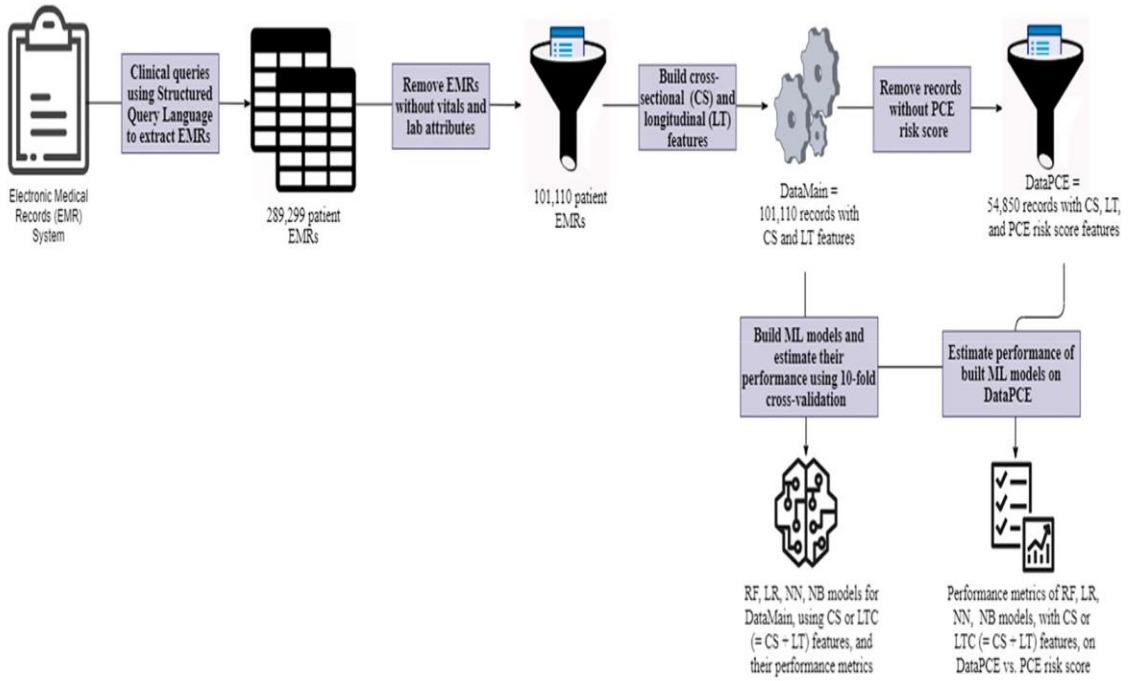



*Figure 1. Workflow for building and testing ML models. Data is extracted from EMRs and filtered. ML models are built on DataMain, compared with each other for performance, and compared against the PCE risk score for the DataPCE subgroup. Abbreviations: ASCVD, atherosclerotic cardiovascular disease; CS, cross-sectional; EMR, electronic medical record; LT, longitudinal; machine-learning models: RF, random forest; LR, logistic regression; NN, neural networks; NB, naïve Bayes.*

Improving ASCVD risk predictions is becoming more and more important in routine clinical practice in order to lower the cost of preventive care and lessen reliance on costly or invasive testing [5,6]. To gauge someone's likelihood of experiencing a major ASCVD event within the next ten years—such as a non-fatal heart attack, a non-fatal stroke, or death caused by coronary heart disease or stroke—current clinical guidelines primarily depend on the pooled cohort equations (PCE) [4,7]. Although there are other scoring systems, they are frequently designed to target particular populations or cardiovascular outcomes [8, 9].

Many prediction models are already being used in clinics, but they can sometimes misrepresent an individual's true risk. This happens when models underestimate or overestimate risk in people with differing medical backgrounds, demographics, or socioeconomic circumstances [2, 5, 10-12]. When risk is inaccurately calculated, patients [2] may receive too much treatment, too little treatment, or experience delays in care—issues that contribute to ongoing clinical inertia. Although modern risk calculators are increasingly built into electronic medical record (EMR) systems that offer decision support, there is still a strong need for tools that provide a more complete understanding of both long-term and short-term ASCVD risk—something current PCE methods [2, 5, 8] often fail to capture. ML involving mathematics, statistics and computer science had become a promising tool for predicting medical data [13]. It is used in multiple health system decision support tools [14-17] today and has been

reported to have performance as good as or better than human developed risk models for cardiology [5,10].

Machine learning offers a way to make better use of the long-term information stored in electronic medical records, helping improve how clinicians assess a patient's risk of developing atherosclerotic cardiovascular disease. Prior research has examined machine learning models to identify ASCVD-related events from cross-sectional clinical variables [10], sometimes with the addition of CAC scores [10, 18]. Our study aimed to tackle these gaps by developing a clinically oriented ASCVD prediction model that avoids many of the limitations found in current methods.

**Materials And Methods**

This study was carried out as a retrospective, records-based analysis using longitudinal electronic medical record data from living patients treated within a regional healthcare system in the United States. Data extraction was performed using a clinical decision-support interface embedded in the EMR, which applied structured query language to identify relevant records within the St. Elizabeth Healthcare System in Kentucky. Patients were eligible if they had at least one clinical encounter between January 1, 2009, and April 30, 2020, during which low-density lipoprotein cholesterol (LDL-C) was recorded. All living patients with at least one LDL-C measurement were included. Because statins reduce LDL-C [4, 19-25] in predictable ways, pretreatment LDL-C levels for active statin users were estimated by multiplying their most recent recorded value by a validated adjustment factor of 1.43.

**Table 1.** Diagnostic Criteria for Comorbidities in the Study Population.

| Diagnosis | Diagnostic Criteria | Reference |
|---|---|---|
| Atherosclerotic cardiovascular disease (ASCVD) | Having either coronary artery disease (CAD), cerebrovascular stroke (CVS), or peripheral artery disease (PAD) | |
| Atherosclerotic coronary artery disease (CAD) | Active CAD diagnosis or ICD-10: I20, I21, I22, I23, I24, or I25 on the EMR problem list or having at least 3 instances of CAD appearing as an encounter diagnosis in the last 2 years or at least 3 CAD claim diagnoses in the last 2 years | [33] |
| Premature coronary artery disease (Premature CAD) | CAD occurring before age 55 years in males or 60 years in females | [34] |
| Ischemic cerebrovascular stroke (CVS) | Active CVS diagnosis or ICD-10: I63, I74, or I75 on the EMR problem list | [33] |
| Peripheral artery disease (PAD) | Active PAD diagnosis or ICD-10: I63, I74, or I75 on the EMR problem list | [33] |
| Diabetes mellitus (DM) | Active DM diagnosis on the EMR problem list or HbA1c ≥ 6.5% more than once or random peripheral blood glucose > 200 mg/dl plus HbA1c ≥ 6.5% and no gestational diabetes type 1 or type 2 | [35] |
| Obesity (OB) | Active obesity diagnosis on the EMR problem list or most recent BMI ≥ 30 kg/m2 | [36] |
| Essential hypertension (HTN) | Active essential HTN diagnosis on the EMR problem list | [37] |
| Congestive heart failure (CHF) | Active CHF diagnosis on the EMR problem list | [38] |

All extracted data were anonymized according to HIPAA regulations. De-identified datasets can be shared upon reasonable request with institutional approval. The study received review board authorization, including a waiver of informed consent due to its retrospective nature. From 289,299 screened encounters, 101,110 records with complete laboratory and vital sign

data formed the primary machine-learning dataset (DataMain), while 54,850 records with PCE scores comprised DataPCE. The study design enabled direct comparison of machine-learning models with PCE calculations [7] and assessment of the influence of 10-year longitudinal features.
The diagnostic definitions used to identify comorbid conditions are outlined in Table 1. In this study, ASCVD includes patients diagnosed with coronary artery disease (CAD), cerebrovascular stroke (CVS), or peripheral artery disease (PAD). Once a patient meets the criteria for CAD or CVS, their medical record is marked with a corresponding diagnosis date and remains classified under that condition for the duration of the study. Some patients may be diagnosed with more than one ASCVD condition, such as both CAD and CVS, and may also have additional comorbidities.

**Cross-Sectional Features and Longitudinal Features**

In developing our machine-learning models, we incorporated both cross-sectional (CS) and longitudinal (LT) features. The CS feature set, consisting of 31 variables, included patient demographics, aggregated clinical risk scores, family history, assigned clinical care categories, laboratory measurements, vital signs, and documented comorbid conditions **(Table 2a).**

**Table 2a.** Cross-Sectional (CS) Features used in the ML Models.

| Feature | Description(s) |
|---|---|
| Demographics | Age |
| | Gender |
| | Age categories: <30, [30,40), [40,50), [50,55), [55,60), [60,65), [65,70), [70,75), [75,80), >=80 |
| Aggregate risk scores | ASCVD 10-year risk score (PCE) |
| | ASCVD 10-year risk score (PCE) categorical, discretized to 3 categories: null value, <5, and ≥ 5 |
| | Numerical score for the family history group of the Dutch Lipid Clinic Network (DLCN) (0,1) |
| | Hierarchical Condition Category Risk Score (Risk Score) |
| | Numeric score for the LDL-C group of the DLCN (0,1,3,5,8) |
| Family history (FHX) | Family history of any coronary artery disease (FHX-++) |
| | Family history of premature coronary artery disease (FHX Premature) |
| | Family history of non-premature coronary artery disease (FHX Non-premature) |
| Clinical care group | Current insurance carrier (Carrier) |
| | Current primary care provider is an employee of the healthcare system where the study is conducted or not (SEP Affiliation) |
| | Have seen endocrinologist in the past or not (Saw Endo) |
| | Patient has account with the MyChart personal health record or not (MyChart) |
| Laboratory values | Maximum LDL-C (whether EHR-documented or last estimated pretreatment) ≥ 190 mg/dL at least twice (LDL-C > 190 x2) |
| | Maximum LDL-C (whether EHR-documented or last estimated pretreatment) ≥ 190 mg/dL at least once (LDL-C > 190) |
| | The last LDL-C reading before a CAD diagnosis, or the last LDL-C reading in absence of CAD LDL (Num Before CAD Avg) |
| | The last Non-HDL-C reading before a CAD diagnosis, or the last Non-HDL-C reading in absence of CAD (Non-HDL-C Num Before CAD Avg) |
| | The last VLDL-C reading before CAD diagnosis, or the last VLDL-C-reading in absence of CAD (VLDL-C-Num-Before-CAD-Avg) |
| | Maximum Lp(a) (MAX LPA) |
| | Maximum Lp(a) group (whether MAX LPA < 29 or > 50 or null value) (MAX LPA cat) |
| Vital signs | Last mean arterial blood pressure (MAP) reading before a CAD diagnosis, or the last MAP reading in absence of CAD (MAP BEFORE CAD Avg) |
| | Last systolic arterial blood pressure (SYS) reading before a CAD diagnosis, or the last SYS reading in absence of CAD (SYS BP BEFORE CAD Avg) |
| | Last diastolic arterial blood pressure (DIA) reading before a CAD diagnosis, or the last DIA reading in absence of CAD (DIA BP BEFORE CAD Avg) |

| | |
|---|---|
| Comorbidities | Diabetes (T1 or T2) (Yes or No, Number of months of having diabetes before being CAD diagnosis) |
| | Hypertension (Yes or No, Number of months of having HTN before CAD diagnosis) |
| | Obesity (Yes or No, Number of months of having OB before CAD diagnosis) |

To better capture how a patient's health changed over time, we developed an additional set of 63 longitudinal (LT) features. These variables reflected evolving patterns in lipid levels, HbA1c, blood pressure, and other routinely measured clinical markers. For every patient and each time-varying parameter, we calculated descriptive statistics such as the minimum, maximum, mean, overall time span, value range, standard deviation, average interval between measurements, and coefficient of variation. Collectively, these measures enabled the LT feature set to represent both actual values and individual variability across time.

For patients without an ASCVD diagnosis, all recorded measurements were used to generate longitudinal statistics. For those who developed ASCVD, only data collected before diagnosis were included, ensuring the LT features reflected true pre-diagnostic information and avoided future data influencing model training.

**Table 2b.** Longitudinal Features (LT) for Vital Signs* and Laboratory Values.

| Feature | Description |
|---|---|
| Minimum (MIN) | Lowest value of all patient recorded values for a feature |
| Maximum (MAX) | Highest value of all patient recorded values for a feature |
| Average (MEAN) | Average value for all recorded values of the feature |
| Reading Number (COUNT) | Number of recorded readings for each measure |
| Reading-time range (TRANGE)‡ | Time difference, in days, between the first and last recorded values for the feature |
| Reading-value range (VRANGE)§ | Difference between the smallest and largest recorded value for the feature |
| Standard Deviation (STDEV) | Amount of variation between the recorded values for the feature |
| Average reading days (Avg-Test-Day)\|\| | Average time, in days, between consecutive recorded values for the feature |
| Coefficient of variation (CV)# | Standard measure of dispersion of a probability distribution or frequency distribution |

*Vital signs: diastolic BP, systolic BP, mean arterial pressure (MAP).

laboratory values: LDL-C, total cholesterol, HDL-C, non-HDL-C, triglycerides, HbA1c.

‡Reading-time range: to determine if the length of the patient's care (reflected by the history of vital and laboratory records) has had any impact on the risk for developing an ASCVD.

§ Reading-value range: might be significant for the cases with large reading differences.

|| Average reading days: to determine if the frequency of patient care, as reflected by patients' vital signs (e.g., BP), being checked in a professional environment and more frequent laboratory tests (e.g., lipid profile, LDL-C) have any impact on the risk for developing ASCVD.

# Coefficient of variation (also known as the relative standard deviation): to study whether the fluctuation of laboratory values and vital sign readings contribute to a patient's risk for developing ASCVD.

**ML Models**

To assess each patient's risk of developing ASCVD, we built automated prediction models using four different machine-learning techniques: logistic regression, naïve Bayes, neural networks, and random forest. Each model was developed twice—first using only cross-sectional (CS) clinical features, and again using a combined set of CS and longitudinal (LT) features (LTC). Using a standard two-by-two contingency structure, we assessed model performance through multiple evaluation measures, applying 10-fold cross-validation on two datasets, DataMain and DataPCE, to ensure reliable comparison.

**Screened Cases Percentage @ Sensitivity Level**

For a prediction model to be practical, it must achieve high sensitivity without requiring

excessive patient screening. To address this, we introduced the Screened Cases Percentage at a given Sensitivity (SCP@Sensitivity), a metric that indicates how much of the population must be evaluated to reach a chosen sensitivity while limiting unnecessary tests. The corresponding formula is provided below:

$$SCP\ @Sensitivity(S) = \frac{|Subpopulation\ of\ patients\ who\ must\ be\ screened\ to\ achieve\ the\ target\ sensitivity\ level\ of\ S|}{|Overall\ Patient\ Population|}$$

where |X| is a notation for the cardinality of a set X.

Practically speaking, this quantity also has clinical value: it provides a way to calculate the resources required for screening a population at sensitivity target. Therefore, it supports efforts to reduce unnecessary diagnostic tests. In theory, the proportion of patients who would need to be screened to identify every ASCVD-positive case—represented as SCP@Sensitivity(1)—could drop to as low as 11.56% in the DataPCE cohort. This value is for the 6,339 people with confirmed ASCVD among the possible 54,850 patients in the analysis in DataPCE.

## Results

Model Performance presents a comparison of all developed models and their AUC results to highlight which combinations of methods and feature sets performed most effectively. Impact of Features used to Build the Models takes a closer look at the top-performing model and examines which features played the most important roles. ML comparison with the PCE features and without the PCE features assesses model performance with and without PCE variables to determine whether the PCE score provides meaningful additional value. ML comparison with the PCE Calculator summarizes statistical comparisons, Screened cases percentage at sensitivities 50% and 90% [SCP@Sensitivity(0.5) and SCP@Sensitivity(0.9)] introduces the SCP@Sensitivity metric, and Probability Threshold for DataMain Model Performance evaluates threshold-based performance of the top model.

### Model Performance

Among all models evaluated, the RF-LTC approach demonstrated the strongest overall performance for ASCVD prediction, achieving an AUC of 0.902 (95% CI: 0.895–0.910), clearly outperforming RF-CS (AUC 0.82), NN-LTC (AUC 0.896), LR-LTC (AUC 0.888), and NB-LTC (AUC 0.817). The AUC values shown in Fig. 2 represent an aggregated assessment derived from probability estimates corresponding to each sensitivity–specificity pair on the ROC curve.

The neural network model was structured with an input layer, a single hidden layer of 150 nodes, and an output layer, and was trained using backpropagation with a learning rate of 0.01 and momentum of 0.9. The random forest model employed bootstrap sampling and the Gini criterion without restrictions on tree depth. Logistic regression was implemented using an L2 penalty, C = 1.0, and a maximum of 100 iterations. The naive Bayes model followed a Gaussian framework without predefined priors. All algorithms were executed using the Scikit-learn library [26].

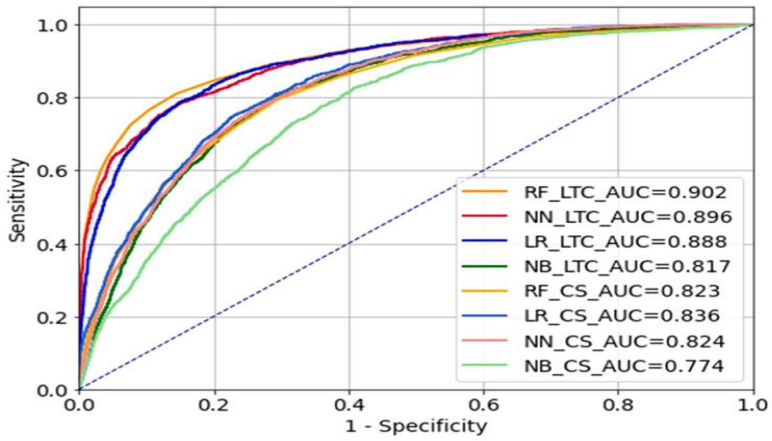
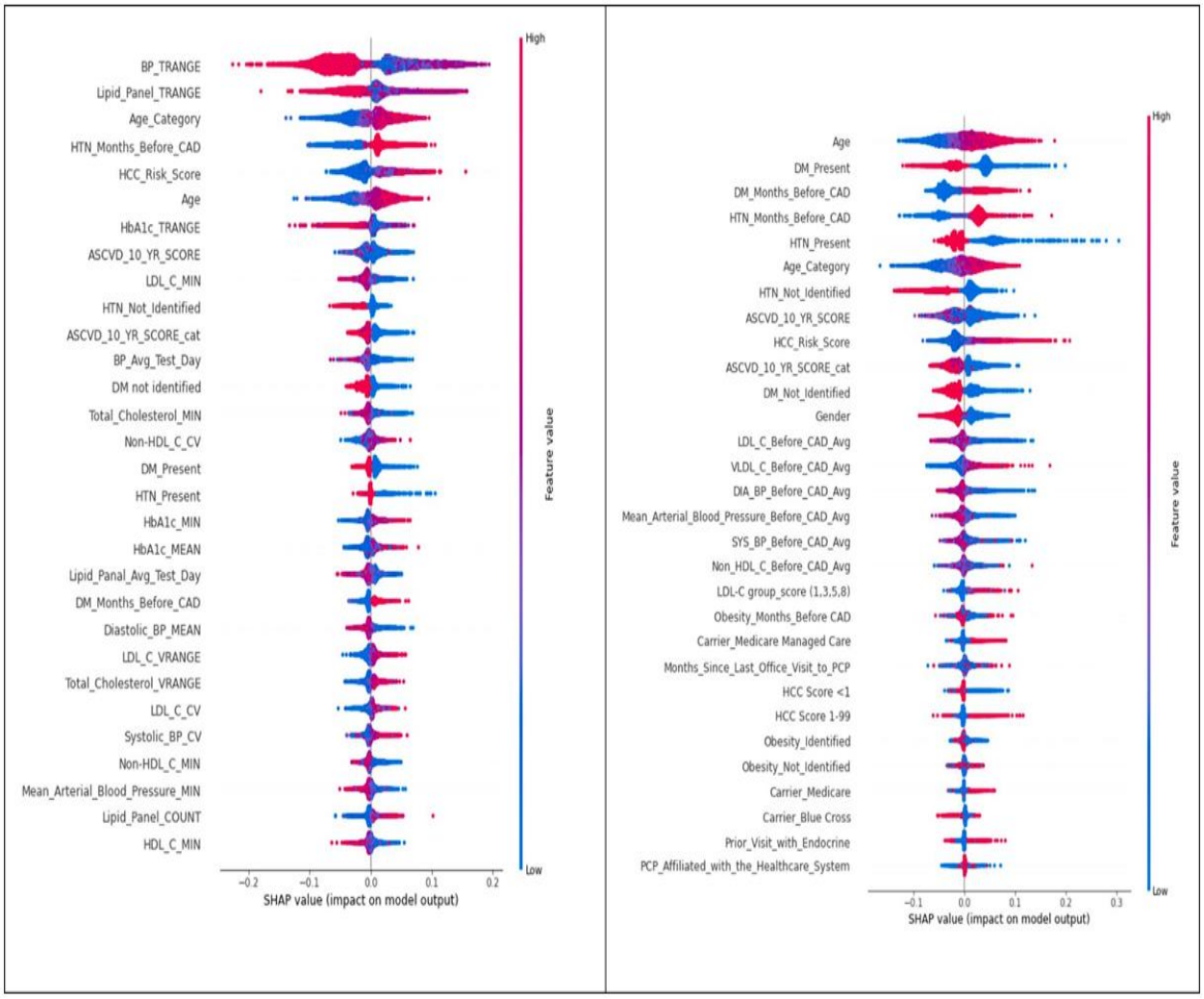



*Figure 2. Area Under Curve (AUC) for DataMain. As a measure for individual model performance for predicting ASCVD in the DataMain cohort, RF-LTC produced the best AUC for ASCVD prediction.*

**Impact of Features used to build the Models**

To better understand how different variables influenced the models, we used Shapley Additive Explanations (SHAP) on the DataMain dataset (Fig. 3). In the RF-CS model, age stood out as the most powerful predictor, followed by existing health conditions, overall risk scores, and LDL-C levels. In contrast, the LTC model placed greater weight on longitudinal information—especially trends in blood pressure, lipid levels, and HbA1c. Even so, age remained one of the strongest contributors in both the RF-LTC and RF-CS models, underscoring its well-known importance in ASCVD risk. Interestingly, the PCE-derived features did not rank among the top drivers of prediction. The ASCVD_10_YR_SCORE_score appeared eighth in both models, while the categorical version of this score ranked even lower—tenth in the RF-LTC model and eleventh in the RF-CS model.




**a) RF-LTC** **b) RF-CS**

*Figure 3. Shapley Additive Explanations (SHAP)[39] Diagram. This illustrates the relative importance of features for: (a) longitudinal features plus cross-sectional features (LTC), and (b) cross-sectional features (CS) only on the RF models, since RF-LTC showed the best performance according to the AUC measure.*

The blue and red points in each row represent data cases having low to high values of the specific variable: blue for low and red for high. The X-axis represents the SHAP value, which quantifies the variable's impact on the model [i.e., tendency to drive the predictions toward an event (positive SHAP value, i.e., ASCVD) or non-event (negative SHAP value, i.e., non-ASCVD)]. The top 20 variables contributing most to the class separability of the model are shown in the figure. Age, comorbid- ities, and aggregate risk scores were the most predictive features in the RF-CS model, followed by LDL-C features. The BP TRANGE measure, lipid TRANGE measure, and HbA1c TRANGE measure were the most predictive LT features. In both models, age was one of the most important features. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**ML Comparison with the PCE Features and without the PCE Features**

To better understand the impact of PCE-derived variables on the models, we re-evaluated all models in the DataPCE cohort, running them both with and without the PCE scores and their categorical forms. What stood out was that the NN-LTC model performed exactly the same, holding an AUC of 0.896 even after the PCE inputs were removed. The RF-LTC model experienced only a slight decline in performance, with its AUC decreasing to 0.894—just a small drop of 0.006. Overall, these results indicate that the core longitudinal and cross-sectional features used by the models carry most of the predictive power, even when the PCE variables are not included.

**ML Comparison with the PCE Calculator**

To evaluate performance, we compared our automated machine-learning models with the PCE 10-year risk score commonly used in clinical practice. Although the ML models were developed using the DataMain cohort, the comparison relied on the DataPCE dataset, as PCE scores were available only there.

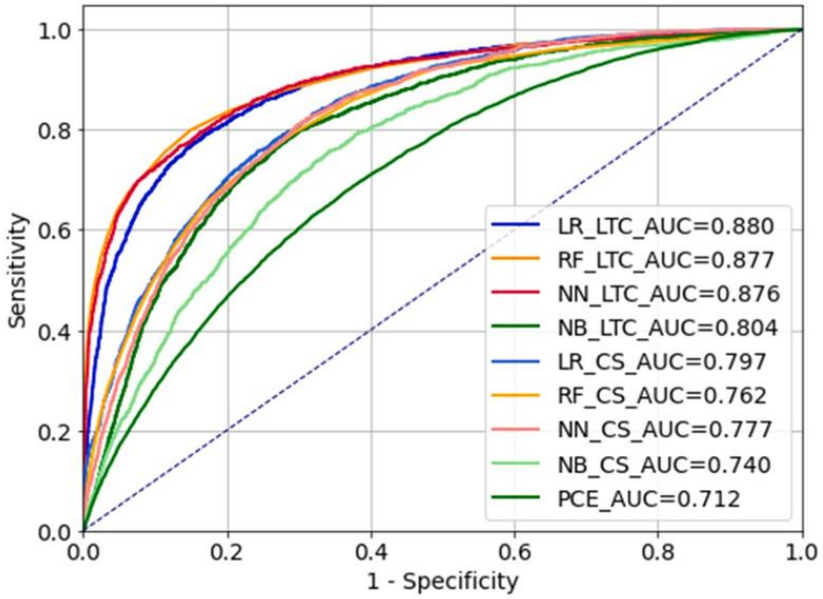



*Figure 4. Area under the curve (AUC) for DataPCE. As a measure of individual model performance for predicting ASCVD in the DataPCE group, LR-LTC pro- duced the best AUC.*

The AUC values derived from 10-fold cross-validation are shown in Fig. 4. When predicting ASCVD risk, all ML methods performed better than the conventional PCE calculator (AUC 0.712; 95 percent CI 0.700–0.730). Among them, the LR-LTC model showed the strongest performance, achieving an AUC of 0.880 (95% CI 0.867–0.894) in the DataPCE dataset.

**Screened Cases Percentage at Sensitivities 50% and 90% [SCP@Sensitivity(0.5) and** SCP@Sensitivity(0.9)**]**

In clinical practice, risk-prediction models are most useful when they help identify individuals who are more likely to have a positive diagnosis while limiting the number of people who must undergo additional laboratory or confirmatory testing. This balance is critical because increasing sensitivity—although desirable for

capturing more true positive cases—also increases the proportion of the population requiring further diagnostic evaluation. Therefore, a model that achieves higher sensitivity with fewer individuals needing follow-up testing (i.e., lower SCP@Sensitivity) is considered more efficient and clinically meaningful. At any chosen sensitivity level (S), an effective machine-learning model should demonstrate a higher positive predictive value (PPV) and a reduced SCP@Sensitivity(S). In current clinical workflows, PCE thresholds of 5% and 20% are frequently used to distinguish low-risk from high-risk patients [4, 30]. Accordingly, we assessed the SCP@Sensitivity of all models at sensitivity levels of 0.90—aligned with the 5% PCE cutoff—and 0.50—aligned with the PCE score of 18.5%. Within the DataPCE cohort, the theoretical minimum SCP@Sensitivity(0.90) is 10.40%. In contrast, the PCE method requires screening all individuals with a risk score of 5% or higher, which in our dataset accounted for 68.8% of the population (S7 Table). The NN-LTC model demonstrated a substantial improvement, identifying the same proportion of true positives while requiring screening of only 43.4% of the population at a probability threshold of 3.2%. All ML models outperformed the PCE method at this sensitivity level, consistently reducing the proportion of individuals requiring further evaluation.

Similarly, to capture 50% of true positive cases, the theoretical minimum SCP@Sensitivity(0.50) is 5.78% in the DataPCE cohort. As shown in S8 Table, the PCE calculator requires screening 25.6% of the population at a cutoff of 18.6%. In comparison, the RF-LTC model achieves the same sensitivity by screening only 7.1% of individuals at the same probability threshold. Again, all ML models showed superior performance relative to the PCE score, requiring substantially fewer individuals to be screened while maintaining equivalent sensitivity.

**Probability Threshold for DataMain Model Performance**

In clinical settings, decision-making is typically binary, with outcomes classified as either positive (1) or negative (0). In contrast, machine-learning models generate continuous risk scores ranging between 0 and 1, which must be interpreted using a threshold probability (t). A predicted score equal to or exceeding t is considered a positive prediction, while a score below t is interpreted as negative. Given that the RF-LTC model demonstrated the highest overall AUC values (Fig. 2), we present its performance along with corresponding threshold probabilities (Table 3) to guide clinicians in selecting optimal thresholds based on clinical priorities. The AUC values reported represent point-level estimates associated with a single sensitivity–specificity pairing at the chosen threshold. Performance metrics for the other top models, NN-LTC and LR-LTC, along with their thresholds. All remaining models underperformed relative to these three.

**Table 5.** RF-LTC Model Performance on DataMain for Various Probability Threshold Values*ψ

| Cut-off Probability | AUC | NPV | Specificity | $F_0$ | PPV | Sensitivity | $F_1$ | SCP@Sensitivity |
|---|---|---|---|---|---|---|---|---|
| 0.05 | 0.673 | 0.982 | 0.378 | 0.546 | 0.247 | 0.967 | 0.393 | 68.2% |
| 0.1 | 0.759 | 0.975 | 0.590 | 0.735 | 0.323 | 0.928 | 0.479 | 50.0% |
| 0.15 | 0.803 | 0.968 | 0.718 | 0.824 | 0.398 | 0.888 | 0.550 | 38.8% |
| 0.2 | 0.820 | 0.960 | 0.797 | 0.871 | 0.466 | 0.842 | 0.600 | 31.4% |
| 0.25 | 0.826 | 0.953 | 0.853 | 0.900 | 0.533 | 0.799 | 0.640 | 26.1% |
| 0.3 | 0.824 | 0.945 | 0.892 | 0.918 | 0.596 | 0.756 | 0.666 | 22.1% |
| 0.35 | 0.816 | 0.938 | 0.920 | 0.929 | 0.651 | 0.712 | 0.680 | 19.0% |
| 0.4 | 0.806 | 0.932 | 0.941 | 0.936 | 0.704 | 0.672 | 0.687 | 16.6% |
| 0.45 | 0.793 | 0.925 | 0.956 | 0.940 | 0.751 | 0.629 | 0.685 | 14.6% |
| 0.5 | 0.776 | 0.917 | 0.968 | 0.942 | 0.795 | 0.584 | 0.673 | 12.8% |
| 0.55 | 0.759 | 0.910 | 0.977 | 0.942 | 0.832 | 0.541 | 0.655 | 11.3% |
| 0.6 | 0.740 | 0.903 | 0.983 | 0.941 | 0.861 | 0.497 | 0.630 | 10.0% |
| 0.65 | 0.720 | 0.895 | 0.988 | 0.940 | 0.892 | 0.451 | 0.599 | 8.8% |
| 0.7 | 0.699 | 0.888 | 0.992 | 0.937 | 0.916 | 0.406 | 0.563 | 7.7% |
| 0.75 | 0.676 | 0.880 | 0.995 | 0.934 | 0934 | 0.358 | 0.517 | 6.7% |
| 0.8 | 0.652 | 0.872 | 0.997 | 0.930 | 0.952 | 0.307 | 0.465 | 5.6% |
| 0.85 | 0.626 | 0.864 | 0.998 | 0.926 | 0.964 | 0.254 | 0.402 | 4.6% |
| 0.9 | 0.598 | 0.855 | 0.999 | 0.922 | 0.975 | 0.197 | 0.328 | 3.5% |
| 0.95 | 0.568 | 0.846 | 1.000 | 0.916 | 0.987 | 0.136 | 0.239 | 2.4% |

* Various methods may be used to calculate the probability threshold, *t*: i) assigned as 0.5 (halfway within the 0–1 range); ii) based on model performance [select the value from the set (0.05, 0.1, 0.15, …, 0.95) that achieves the best performance metric (AUC, PPV or sensitivity); or iii) based on the SCP, PPV, and sensitivity of the model for that threshold value (the higher *t*, the higher the PPV, but the lower the sensitivity and SCP of the model).
ψ Color variation indicates low (lightest) to high (darkest) AUC.

For the RF-LTC model, the highest AUC (0.826) was achieved at a threshold of 0.25. In terms of the F1 score, optimal performance (0.687) occurred at a threshold of 0.40, with additional metrics detailed in Table 3. Clinicians can select thresholds according to specific needs; for example, a threshold of 0.25 achieves approximately 80% sensitivity, while a threshold of 0.50 corresponds to 80% positive predictive value (PPV). This flexibility enables careful balancing between minimizing misclassification of high-risk patients as low risk and maximizing opportunities for timely interventions and potentially life-saving management strategies.

## Discussion

Research exploring machine-learning approaches for estimating atherosclerotic cardiovascular disease (ASCVD) risk has yielded mixed but generally encouraging insights. Early studies [10, 18] indicated that ML-based models can complement standard cardiac evaluations, either by improving interpretation of imaging results or [27] by predicting a broad range of ASCVD-related outcomes [5,10,18,28-30]. Notable work by Motwani et al. [28], van Rosendael et al. [30], and Nakanishi et al. [10]—across cohorts ranging from 8,844 to over 66,000 participants—used coronary calcium scores and imaging data to estimate short- and long-term mortality and coronary heart disease risk. Ward et al. [18] and colleagues examined five-year ASCVD risk using structured clinical information from electronic medical records.

In comparison, our study observed similar mean ASCVD risk scores in the DataMain cohort (7.19 ± 11.313), while the DataPCE cohort showed substantially higher averages (13.26 ± 12.47). Although mortality could not be assessed because deceased individuals were not included, we focused on current ASCVD risk by incorporating both clinical symptom (CS) data and longitudinal trajectory (LT) patterns preceding diagnosis [30]. Earlier models paired clinical factors with imaging, but many were limited by selection bias [30]. Consistent with previous findings, our ML models surpassed pooled cohort equation (PCE) estimates [18]. Using 94 variables (31 CS, 63 LT), we found that LT data added important predictive value [10, 18, 28]. Across 101,110 individuals, models integrating both data types performed strongly, with RF-LTC achieving an AUC of 0.902. These improvements suggest ML tools may reduce unnecessary testing, and SCP@Sensitivity offers a practical metric for guiding clinical deployment.

### Study Limitations

While earlier studies have emphasized how useful CAC and CCTA scores can be for evaluating ASCVD risk, only a small portion of our population—[2,8,32] around 2%—actually had those imaging results. Because of that, we left those measures out of our feature set. Even so, the model performed strongly, reaching an AUC of 0.92. Our work was based entirely on structured EMR data. We recognize, however, that symptoms reported by both patients and clinicians often provide important clues when diagnosing ASCVD. Incorporating symptom information from unstructured clinical notes is something we plan to explore in the future, and it may further improve the model's performance. One of the key limitations of our study is that we did not conduct external validation, which means our findings may not fully generalize to other settings. Another concern is that including the PCE score might introduce some information bias, given that knowledge of the score can influence how clinicians manage a patient and how the disease progresses. That said, our models also capture a wide range of other cardiovascular risk factors, and the predictive accuracy remained essentially unchanged even when we removed the PCE feature.

### Study Strengths

To our knowledge, this is the first study to examine machine-learning models that can estimate population-specific ASCVD risk—aside from mortality—without depending on CAC scores, while also drawing on both cross-sectional and longitudinal EMR data from an entire regional health network. Because our dataset includes people with and without symptoms, it mirrors real-world clinical practice rather than a narrowly defined research sample. The ability of these models to provide short-term risk estimates may help clinicians explain risks more clearly to their patients and motivate timely lifestyle changes, even for individuals who have no symptoms. Unlike earlier [5, 28, 30] ML work that focused on highly selected cohorts, our approach uses routinely captured EMR data that

contain substantial clinical detail and are largely unaffected by referral patterns. We also introduce the SCP@Sensitivity measure as a practical way to guide cost-effective ASCVD screening and to complement existing tools such as net benefit analyses.

In predicting ASCVD, our results show that machine-learning models that use both cross-sectional data and longer-term clinical trends outperform the conventional PCE calculator. The predictions became even more accurate when changes in vital signs and lab results over time were taken into account. The majority of EMR systems already have this kind of data, so it is plausible that routine clinical practice will incorporate the use of ML to estimate ASCVD risk. Clinicians may be able to intervene earlier, customize treatments more accurately, and possibly eliminate the need for tests like stress imaging, CCTA, or CAC scoring thanks to these models. These methods are now being expanded in current research to forecast more precise ASCVD-related outcomes.

## Summary

### Existing Knowledge

Machine-learning (ML) models are being used more and more to estimate a person's risk of developing atherosclerotic cardiovascular disease (ASCVD), providing data-driven insights that can enhance and support traditional clinical risk assessment methods.

### Contributions of This Study

This study offers a detailed comparison between machine-learning models for ASCVD prediction and the commonly used ACC/AHA Pooled Cohort Equations (PCE). Our findings show that ML-based approaches can give clinicians a more accurate and personalized assessment of ASCVD risk. When we combined both longitudinal (LT) data—such as changes in lab results and vital signs over time—with cross-sectional (CS) information from electronic medical records, the models became noticeably more accurate. These longitudinal patterns added valuable context that single-point clinical measurements alone cannot capture. We also identified which features had the strongest influence on ASCVD risk, offering clinicians clearer guidance on which factors matter most. In addition, we introduced a new metric—Screened Cases Percentage at a given Sensitivity (SCP@Sensitivity)—to help estimate how much of the population would need further screening while still keeping sensitivity high. This metric supports more efficient use of healthcare resources without compromising the quality of risk detection.

## References


[1] S.S. Virani, A. Alonso, H.J. Aparicio, E.J. Benjamin, M.S. Bittencourt, C. W. Callaway, A.P. Carson, A.M. Chamberlain, S. Cheng, F.N. Delling, M.S.V. Elkind, K.R. Evenson, J.F. Ferguson, D.K. Gupta, S.S. Khan, B.M. Kissela, K.L. Knutson, C. D. Lee, T.T. Lewis, J. Liu, M.S. Loop, P.L. Lutsey, J. Ma, J. Mackey, S.S. Martin, D. B. Matchar, M.E. Mussolino, S.D. Navaneethan, A.M. Perak, G.A. Roth, Z. Samad, G.M. Satou, E.B. Schroeder, S.H. Shah, C.M. Shay, A. Stokes, L.B. VanWagner, N.-Y. Wang, C.W. Tsao, "Heart Disease and Stroke Statistics"—2021 Update: A Report From the American Heart Association, Circulation 143 (8) (2021), https://doi.org/10.1161/CIR.0000000000000950.

[2] D.M. Lloyd-Jones, L.T. Braun, C.E. Ndumele, S.C. Smith, L.S. Sperling, S.S. Virani, R.S. Blumenthal, "Use of Risk Assessment Tools to Guide Decision-Making in the Primary Prevention of Atherosclerotic Cardiovascular Disease: A Special Report From the American Heart Association and American College of Cardiology", Circulation 139 (25) (2019), https://doi.org/10.1161/CIR.0000000000000638.

[3] Karmali KN, Persell SD, Perel P, Lloyd-Jones DM, Berendsen MA, Huffman MD., " Risk scoring for the primary prevention of cardiovascular disease," Cochrane Database Syst Rev. 2017;3:CD006887. Epub 2017/03/16. doi: 10.1002/14651858. CD006887.pub4. PubMed PMID: 28290160; PubMed Central PMCID: PMCPMC6464686.

[4] Grundy SM, Stone NJ, Bailey AL, Beam C, Birtcher KK, Blumenthal RS, et al. 2018 HA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary. Circulation. 2018:CIR0000000000000624. Epub 2018/12/20. doi: 10.1161/CIR.0000000000000624. PubMed PMID: 30565953.

[5] Kavuri, S. (2023). Machine learning approaches for security vulnerability detection in software testing. Computer Fraud & Security, 2023(10). https://doi.org/10.52710/cfs.837.

[6] K.M. Chinnaiyan, P. Peyser, T. Goraya, K. Ananthasubramaniam, M. Gallagher, A. DePetris, J.A. Boura, E. Kazerooni, C. Poopat, M. Al-Mallah, S. Saba, S. Patel, S. Girard, T. Song, D. Share, G. Raff, "Impact of a Continuous Quality Improvement

Initiative on Appropriate Use of Coronary Computed Tomography Angiography," J. Am. Coll. Cardiol. 60 (13) (2012) 1185–1191.

[7] Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Sr., Gibbons R, et al. 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. J Am Coll Cardiol. 2014;63(25 Pt B):2935-59. Epub 2013/11/19. doi: 10.1016/j.jacc.2013.11.005. PubMed PMID: 24239921; PubMed Central PMCID: PMCPMC4700825.

[8] M.O. Gore, C.R. Ayers, A. Khera, C.R. deFilippi, T.J. Wang, S.L. Seliger, V. Nambi, E. Selvin, J.D. Berry, W.G. Hundley, M. Budoff, P. Greenland, M.H. Drazner, C. Ballantyne, B.D. Levine, J.A. de Lemos, "Combining Biomarkers and Imaging for Short-Term Assessment of Cardiovascular Disease Risk in Apparently Healthy Adults," JAHA 9 (15) (2020), https://doi.org/10.1161/JAHA.119.015410.

[9] P.M. Ridker, J.E. Buring, N. Rifai, N.R. Cook, "Development and Validation of Improved Algorithms for the Assessment of Global Cardiovascular Risk in Women: The Reynolds Risk Score," JAMA 297 (6) (2007) 611, https://doi.org/10.1001/jama.297.6.611.

[10] R. Nakanishi, P.J. Slomka, R. Rios, J. Betancur, M.J. Blaha, K. Nasir, M. D. Miedema, J.A. Rumberger, H. Gransar, L.J. Shaw, A. Rozanski, M.J. Budoff, D. S. Berman, "Machine Learning Adds to Clinical and CAC Assessments in Predicting 10-Year CHD and CVD Deaths," JACC Cardiovasc Imaging. 14 (3) (2021) 615–625, https://doi.org/10.1016/j.jcmg.2020.08.024.

[11] M. Kavousi, M.J.G. Leening, D. Nanchen, P. Greenland, I.M. Graham, E. W. Steyerberg, M.A. Ikram, B.H. Stricker, A. Hofman, O.H. Franco, "Comparison of Application of the ACC/AHA Guidelines, Adult Treatment Panel III Guidelines, and European Society of Cardiology Guidelines for Cardiovascular Disease Prevention in a European Cohort," JAMA 311 (14) (2014) 1416, https://doi.org/10.1001/jama.2014.2632.

[12] Rana JS, Tabada GH, Solomon MD, Lo JC, Jaffe MG, Sung SH, et al. "Accuracy of the Atherosclerotic Cardiovascular Risk Equation in a Large Contemporary," Multiethnic Population. J Am Coll Cardiol. 2016;67(18):2118-30. Epub 2016/05/07. doi: 10.1016/j.jacc.2016.02.055. PubMed PMID: 27151343; PubMed Central PMCID: PMCPMC5097466.

[13] J.A.M. Sidey-Gibbons, C.J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," BMC Med. Res. Method. 19 (1) (2019), https://doi.org/10.1186/s12874-019-0681-4.

[14] Y. Ye, F. Tsui, M. Wagner, J.U. Espino, Q. Li, "Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers," J. Am. Med. Inform. Assoc. 21 (5) (2014) 815–823.

[15] H. Zhai, P. Brady, Q.i. Li, T. Lingren, Y. Ni, D.S. Wheeler, I. Solti, "Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children," Resuscitation 85 (8) (2014) 1065–1071.

[16] F. Doshi-Velez, R.H. Perlis, "Evaluating Machine Learning Articles," JAMA 322 (18) (2019) 1777, https://doi.org/10.1001/jama.2019.17304.

[17] N. Hong, H. Park, Y. Rhee, "Machine Learning Applications in Endocrinology and Metabolism Research: An Overview," Endocrinol Metab 35 (1) (2020) 71, https://doi.org/10.3803/EnM.2020.35.1.71.

[18] A. Ward, A. Sarraju, S. Chung, J. Li, R. Harrington, P. Heidenreich, L. Palaniappan, D. Scheinker, F. Rodriguez, "Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population," npj Digit. Med. 3 (1) (2020), https://doi.org/10.1038/s41746-020-00331-1.

[19] S.D. de Ferranti, A.M. Rodday, M.M. Mendelson, J.B. Wong, L.K. Leslie, "R. C. Sheldrick, Prevalence of Familial Hypercholesterolemia in the 1999 to 2012 United States National Health and Nutrition Examination Surveys (NHANES)," Circulation 133 (11) (2016) 1067–1072, https://doi.org/10.1161/CIRCULATIONAHA.115.018791. PubMed PMID: 26976914.

[20] M. Benn, G.F. Watts, A. Tybjaerg-Hansen, B.G. Nordestgaard, "Familial hypercholesterolemia in the danish

general population: prevalence, coronary artery disease, and cholesterol-lowering medication," J. Clin. Endocrinol. Metab. 97 (11) (2012) 3956–3964, https://doi.org/10.1210/jc.2012-1563. PubMed PMID: 22893714.

[21] Myocardial Infarction Genetics Consortium I, Stitziel NO, Won HH, Morrison AC, Peloso GM, Do R, et al. Inactivating mutations in NPC1L1 and protection from coronary heart disease. N Engl J Med. 2014;371(22):2072-82. Epub 2014/11/13. doi: 10.1056/NEJMoa1405386. . PubMed PMID: 25390462; PubMed Central PMCID: PMCPMC4335708.

[22] M. Benn, G.F. Watts, A. Tybjærg-Hansen, B.G. Nordestgaard, Mutations causative of familial hypercholesterolaemia: screening of 98 098 individuals from the Copenhagen General Population Study estimated a prevalence of 1 in 217, Eur. Heart J. 37 (17) (2016) 1384–1394.

[23] Eid WE, Sapp EH, Flerlage E, Nolan JR. Lower-Intensity Statins Contributing to Gaps in Care for Patients With Primary Severe Hypercholesterolemia. J Am Heart Assoc. 2021;10(17):e020800. Epub 2021/09/02. doi: 10.1161/JAHA.121.020800. PubMed PMID: 34465130.

[24] W.E. Eid, E.H. Sapp, A. Wendt, A. Lumpp, C. Miller, "Improving Familial Hypercholesterolemia Diagnosis Using an EMR-based Hybrid Diagnostic Model," J. Clin. Endocrinol. Metab. 107 (4) (2022) 1078–1090, https://doi.org/10.1210/clinem/dgab873.

[25] W.E. Eid, E.H. Sapp, T. McCreless, J.R. Nolan, E. Flerlage, "Prevalence and Characteristics of Patients With Primary Severe Hypercholesterolemia in a Multidisciplinary Healthcare System," Am. J. Cardiol. 132 (2020) 59–65, https://doi.org/10.1016/j.amjcard.2020.07.008.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn Res. 12 (null) (2011) 2825–2830.

[27] V. Brandt, T. Emrich, U.J. Schoepf, D.M. Dargis, R.R. Bayer, C.N. De Cecco, C. Tesche, "Ischemia and outcome prediction by cardiac CT based machine learning," Int. J. Cardiovasc. Imaging 36 (12) (2020) 2429–2439.

[28] Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, 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(7):500-7. Epub 2016/06/03. doi: 10.1093/eurheartj/ehw188. PubMed PMID: 27252451; PubMed Central PMCID: PMCPMC5897836.

[29] M. van Assen, A. Varga-Szemes, U.J. Schoepf, T.M. Duguay, H.T. Hudson,S. Egorova, K. Johnson, S. St. Pierre, B. Zaki, M. Oudkerk, R. Vliegenthart, A. J. Buckler, "Automated plaque analysis for the prognostication of major adverse cardiac events," Eur. J. Radiol. 116 (2019) 76–83.

[30] A.R. van Rosendael, G. Maliakal, K.K. Kolli, A. Beecy, S.J. Al'Aref, A. Dwivedi, G. Singh, M. Panday, A. Kumar, X. Ma, S. Achenbach, M.H. Al-Mallah, D. Andreini, J.J. Bax, D.S. Berman, M.J. Budoff, F. Cademartiri, T.Q. Callister, H.-J. Chang, K. Chinnaiyan, B.J.W. Chow, R.C. Cury, A. DeLago, G. Feuchtner, M. Hadamitzky, J. Hausleiter, P.A. Kaufmann, Y.-J. Kim, J.A. Leipsic, E. Maffei, H. Marques, G. Pontone, G.L. Raff, R. Rubinshtein, L.J. Shaw, T.C. Villines, H. Gransar, Y. Lu, E. C. Jones, J.M. Pen˜a, F.Y. Lin, J.K. Min, "Maximization of the usage of coronary CTA derived plaque information using a machine learning based algorithm to improve risk stratification; insights from the CONFIRM registry," J. Cardiovasc. Comput. Tomogr. 12 (3) (2018) 204–209.

[31] K.M. Johnson, H.E. Johnson, Y. Zhao, D.A. Dowe, L.H. Staib, "Scoring of Coronary Artery Disease Characteristics on Coronary CT Angiograms by Using Machine Learning," Radiology 292 (2) (2019) 354–362.

[32] A. Khera, M.J. Budoff, C.J. O'Donnell, C.A. Ayers, J. Locke, J.A. de Lemos, J. M. Massaro, R.L. McClelland, A. Taylor, B.D. Levine, "Astronaut Cardiovascular Health and Risk Modification (Astro-CHARM) Coronary Calcium Atherosclerotic Cardiovascular Disease Risk Calculator," Circulation 138 (17) (2018) 1819–1827.

[33] National Center for Health Statistics. Center For Disease Control and Prevntion. https://icd10cmtool.cdc.gov/?fy=FY202 1. Accessed May 29, 2021. Available from: https://icd10cmtool.cdc.gov/?fy=FY2021.

[34] B.G. Nordestgaard, M.J. Chapman, S.E. Humphries, H.N. Ginsberg, L. Masana, O. S. Descamps, et al., Familial hypercholesterolaemia is

underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: consensus statement of the European Atherosclerosis Society, Eur. Heart J. 34 (45) (2013) 3478–3490, https://doi.org/10.1093/eurheartj/eht273. PubMed PMID: 23956253; PubMed Central PMCID: PMCPMC3844152.

[35] American Diabetes A. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes- 2020. Diabetes Care. 2020;43(Suppl 1):S14-S31. Epub 2019/12/22. doi: 10.2337/dc20-S002. PubMed PMID: 31862745.

[36] Defining Adult Overweight and Obesity. Center For Disease Control and Prevntion. https://www.cdc.gov/obesity/adult/defining.html. Accessed May 29, 2021. Available from: https://www.cdc.gov/obesity/adult/defining.html.

[37] J.J. Boisvenue, C.U. Oliva, D.P. Manca, J.A. Johnson, R.O. Yeung, "Feasibility of identifying and describing the burden of early-onset metabolic syndrome in primary care electronic medical record data: a cross-sectional analysis," cmajo 8 (4) (2020) E779–E787.

[38] Y. Xu, S. Lee, E. Martin, A.G. D'souza, C.T.A. Doktorchik, J. Jiang, S. Lee, C. A. Eastwood, N. Fine, B. Hemmelgarn, K. Todd, H. Quan, "Enhancing ICD-Code- Based Case Definition for Heart Failure Using Electronic Medical Record Data," J. Cardiac Fail. 26 (7) (2020) 610–617.