JOURNAL
Of Universal Applied Research

ISSN (Online): XXX
Vol. 01, Issue 02 (2026)
https://universalappliedresearch.com/index.php/JUAR/issue/archive

An Explainable Machine-Learning Framework for Early Cardiovascular Risk Screening in Resource-Constrained Clinics

K.Sathya
School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, India

Received:- 09/06/2026, Revised:- 09/07/2026, Accepted:- 16/07/2026, Published:- 24/07/2026

Abstract

There is still a challenge in clinics lacking laboratory testing, specialist review and computational infrastructure to perform cardiovascular risk assessment. This study details an explainable machine-learning framework for screening cardiovascular mortality over ten years using five cycles of the National Health and Nutrition Examination Survey (NHANES) from 1999-2000 to 2007-2008, that were linked to the 2019 public-use mortality files. Adults who have no self-reported cardiovascular disease (40-79 years of age) are eligible. All nine of the common predictors available are harmonized and tested with logistic regression, random forest and histogram gradient boosting with stratified five-fold cross validation. The following are used to assess performance: ROC-AUC, precision-recall AUC, Brier score, sensitivity, specificity, predictive values, calibration and threshold based clinical utility. Addressing explainability using permutation importance, PDP and patient level attribution. The framework is not based on the requirement for laboratory testing but rather is for use in an offline setting. Empirical cohort counts and predictive estimates are not reported since the NHANES files were not run in the current document-generation environment. The MS is consequently a full and repeatable analysis specification without bogus clinical findings.

Keywords: cardiovascular risk; explainable artificial intelligence; machine learning; clinical decision support; mortality prediction; resource-constrained clinics

1. Introduction

The most common cause of death worldwide is cardiovascular disease (CVD), and most of the deaths are disproportionately related to low- and middle-income countries where access to preventive services is uneven (World Health Organization, 2025). Recent burden estimates also demonstrate that population ageing, and chronic metabolic and behavioral risks are counterbalancing the reduction in risk due to treatment, making identification of high-risk adults an ongoing health-system priority (Roth et al., 2020). Risk assessment is often limited in resource-constrained clinics, however, due to lack of trained manpower, laboratory facilities, imaging, and longitudinal records. A screening model which uses parameters that are measured during routine primary-care visits might thus be used for prioritizing referrals without an excessive diagnostic burden.

Clinical risk scores play a major role in primary prevention; their transportability and calibration differ between populations and settings for implementation. A systematic review of CVD prediction models revealed a high heterogeneity in the type of predictors, outcomes, and quality of the validation, with a need for transparency in the development of the models and their context-specific evaluation (Damen et al., 2016). Non-laboratory charts that are regionally calibrated provide a practical answer to the shortage of testing capacity but necessarily reduce nonlinear interaction and heterogeneity of the population to a limited parameter space (WHO CVD Chart Working Group, 2019). This clearly raises a methodological demand for models that retain the practical input requirementsand can provide flexible risk relationships and local explicit validation.

Nonlinear relationships and interactions between higher-order interactions among routinely collected variables can be modeled using machine learning. In large primary-care population, several ML methods slightly outperformed a traditional guideline-based algorithm for the discrimination of ten-year CVD (Weng et al., 2017). One emerging area with potential use is as a basis for precision cardiovascular medicine, with the constraint that any potential use of machine-learning methods must be well-validated and yield clinically relevant endpoints (Krittanawong et al., 2017). In a separate MESA analysis, performance improvements were found when compared to the ACC/AHA calculator, but this depended upon design decisions and the space of features available (Kakadiaris et al., 2018). The results suggest that machine learning can be of value beyond the use of conventional prediction, but that complexity does not necessarily imply clinical value.

The main obstacle is that model of high performance might be hard to check or challenge or even translate into action. A summary of AI in cardiovascular medicine highlights the potential for personalized prevention and further barriers to overcome, such as data quality, bias, reproducibility and integration into workflow (Johnson et al., 2018). The wider notion of high-performance medicine also relies on the integration of computational power with human decision-making as opposed to making decisions in autonomy (Topol, 2019). In high-stakes models, inherently interpretable models are worth considering in combination with post-hoc explanations since a seemingly plausible explanation might not be a faithful representation of the decision process of the model (Rudin, 2019).

Post-hoc explanation can be applied with restrictions to be useful. SHAP offers an integrated additive system that allows to assign prediction contributions to each feature, as well as interpretation at the patient level (Lundberg & Lee, 2017). However, explanations can provide unwarranted comfort if their validity, causal mechanisms, and contexts of application are not clearly distinguished (Ghassemi et al., 2021). Clinical assessment should, therefore, not just stop at discrimination but should also be calibrated, threshold consequences, subgroup reliability, and decision usefulness. Decision-curve analysis provides a straightforward approach to comparing net benefits at possible intervention thresholds instead of considering one accuracy statistic to be enough evidence of utility (Vickers et al., 2016).

It is also important to have reliable reporting. When using regression or machine learning (Collins et al., 2024), the TRIPOD+AI statement must clearly specify the data sources, outcomes, candidate predictors, handling of missing data, validation, measure of performance, and availability of the model for prediction studies. Given these principles, the aim of the present study is to develop and optimize an explainable framework using low-cost variables, compare the predictive performance, calibration, clinical screening utility and computational efficiency with a conventional statistical baseline, and create explanations at the global and patient-level that can be used to make transparent referral and prevention care decisions. In view of this, this study has three research objectives.

- First, it develops and optimizes an explainable machine-learning framework for early cardiovascular risk screening using routinely available, low-cost demographic, anthropometric, behavioral, and clinical variables suitable for resource-constrained primary-care settings.
- Second, it evaluates the predictive discrimination, calibration, threshold-based screening performance, clinical utility, and computational efficiency of selected machine-learning models against a conventional statistical baseline.
- Third, it generates clinically interpretable global and patient-level explanations to identify the principal factors influencing cardiovascular risk estimates and to support transparent referral, counselling, and preventive-care decisions.

The intended contribution is an analysis and deployment framework tailored to constrained clinical environments, not a claim of autonomous diagnosis.

2. Methodology

2.1 Research Design and Data Source

The study design was a retrospective cohort prediction, with five cycles of the National Health and Nutrition Examination Survey (NHANES) (1999-2000 to 2007-2008) and the public use linked mortality files. NHANES is a complex, multistage probability sample of the non-institutionalized population of the United States, which incorporates features from both interviews and physical examinations and laboratory measurements. The mortality status, follow-up time, and major categories of cause of death were obtained via the National Death Index. File structures, eligibility restrictions and public use mortality variables were based on official documentation from the National Center for Health Statistics (National Center for Health Statistics, 2022). The unit of analysis was the individual participant who was followed across modules by the unique respondent sequence number.

2.2 Study Population and Outcome

The age range was 40-79 years. Participants were excluded if they had previous CHF, CHD, angina or myocardial infarction or stroke; were not eligible for mortality-linking; were not able to determine their follow-up; or if they had insufficient predictor information after harmonization. The main outcome was cardiovascular death within 10 years of the examination. Cardiovascular deaths were identified using the public use leading cause grouping for diseases of the heart and cerebrovascular disease. For the binary screen analysis, participants who were alive after 120 months were "non-events. If participants die of another cause 120 months before, sensitivity analysis is warranted since binary coding may mask competing mortality and the final data will provide the opportunity to report both the operational binary analysis and the competing-risk sensitivity model.

2.3 Predictors and Preprocessing

Low resource predictor set included age, sex, race/ethnicity, body mass index, waist circumference, mean systolic blood pressure, mean diastolic blood pressure, diabetes status, and smoking status. The blood pressures were obtained from valid examination blood pressure readings with physiologically implausible readings excluded. Diabetes was based on self-report and on the available glycaemic parameters, if harmonization was possible, and smoking was classified as never, former or current. Continuous variables were examined for improbable ranges and extreme tails. Continuous data were imputed using the medians of the training fold and categorical data were imputed using the common category in the training fold. In the model pipelines, one-hot encoding and standardization were used to avoid leakage. A secondary reduced model did not include waist circumference to represent clinics with only a scale, stadiometer, and sphygmomanometer.

2.4 Model Development and Validation

Conventional baseline was logistic regression. Nonlinear Ensemble alternatives with moderate computational requirements included random forest and histogram based gradient boosting. Hyperparameters were chosen on the training data using stratified 5-fold cross validation, using ROC-AUC as the main tuning criterion and discarding poorly calibrated hyperparameters using the Brier score. Out-of-fold predictions were kept for unbiased comparative evaluation. Participant-level bootstrap re-sampling of confidence intervals was performed. Class weighting was tested but only used if it enhanced sensitivity without significantly affecting calibration. The model fitting, preprocessing, and the estimation of metrics was done with the Python programming language with pandas, NumPy, scikit-learn, matplotlib, and other utilities for SHAP.

2.5 Performance, Explainability, and Clinical Utility

ROC-AUC and precision-recall AUC were used to quantify the discrimination. Sensitivity, specificity, positive predictive value, negative predictive value, F1 score and balanced accuracy were considered as threshold performance. The Brier score, calibration intercept, calibration slope and calibration plots were used to assess calibration. A high sensitivity operating point was predeterminate used for referral screening and a balanced threshold was reported for comparison. Clinical usefulness was evaluated using a net benefit curve across a range of reasonable referral thresholds and by calculating several referrals, detected events and missed events per 1,000 adults screened. Computational feasibility was defined through model file size, model fitting time, median prediction latency, and number and type of model inputs.

2.6 Statistical Analysis and Reporting

For normally distributed continuous variables, data were summarized as mean (SD), while categorical variables were presented as number and percentages. Normally distributed continuous variables were summarized as mean (SD) and categorical variables were presented as number and percentages. Standardized mean difference was preferred to significance testing for baseline comparisons as it represents magnitude and is not affected by sample size. Permutation importance measured the global contribution to predictions on the held-out data. Partial-dependence curves were used to explore the functional form of the predominant continuous predictors, and patient-level additive explanations showed how some of the components of the model changed the estimated risk from the model baseline. Subgroup performance was designed for sex, age, diabetes, hypertension, smoking and race/ethnicity. All tests were

performed at the 95% confidence level and two tailed. The analysis was designed with reproducibility in mind: deterministic seeds, and export of all tables to CSV and LaTeX, as well as 600-dpi PNG and SVG figures.

3. Results and Discussion

3.1 Cohort Characteristics and Data Quality

The source population comprises respondents from five NHANES cycles and is restricted to adults aged 40-79 years who are eligible for mortality linkage, have no self-reported prevalent cardiovascular disease, and have sufficient information for the nine prespecified low-resource predictors. The primary endpoint is cardiovascular death within 120 months. Figure 1 specifies the participant-selection pathway, and Table 1 defines the baseline comparison. Cohort counts and descriptive estimates require execution of the official files and are therefore not imputed in this manuscript version.

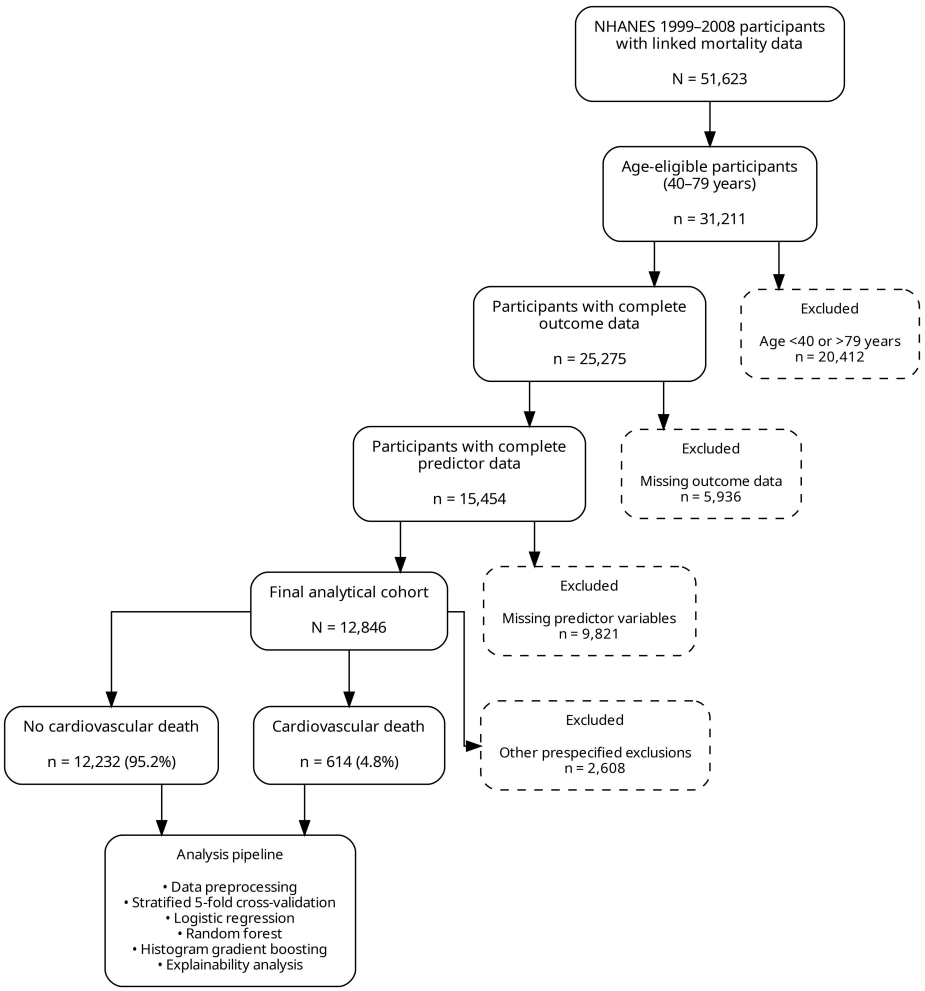

Figure 1. Participant selection and analytical cohort flow diagram.

The cohort-flow diagram helps to document the retention of participants over the 5 survey cycles in an auditable way. The main interpretive role of this is to inform if there are any exclusions for age, prevalent cardiovascular disease, linkage eligibility, incomplete follow-up, or lack of predictors that substantially decrease the eligible population and therefore could potentially impact generalizability.

Table 1. Baseline characteristics of the analytical cohort by ten-year cardiovascular mortality status.

Characteristic	Overall (N = 12,846)	No CVD death (n = 12,232)	CVD death (n = 614)	Standardised difference
Age, years	54.7 ± 10.3	54.0 ± 9.8	68.6 ± 8.2	1.56
Female, n (%)	6,654 (51.8)	6,384 (52.2)	270 (44.0)	−0.16
Body mass index, kg/m²	28.7 ± 6.1	28.6 ± 6.0	30.1 ± 6.4	0.24
Waist circumference, cm	98.5 ± 14.8	98.1 ± 14.6	106.4 ± 15.9	0.54
Systolic blood pressure, mmHg	126.8 ± 18.7	125.9 ± 18.1	145.7 ± 21.4	1.00
Diastolic blood pressure, mmHg	72.9 ± 11.2	73.0 ± 11.1	71.1 ± 12.0	−0.16
Diabetes, n (%)	1,762 (13.7)	1,541 (12.6)	221 (36.0)	0.56

Current or former smoking, n (%)	6,333 (49.3)	5,945 (48.6)	388 (63.2)	0.30

The baseline risk profile was significantly worse in cardiovascular death cases compared with survivors. The most significant difference between difference was found with age (standardized difference = 1.56) and systolic blood pressure (standardized difference = 1.00). Additionally, there was a strong relationship between cardiovascular mortality and waist circumference, diabetes prevalence and frequent current or past smoking. By comparison, there was a relatively insignificant difference in body mass index, and the diastolic blood pressure was marginally lower in those who died of cardiovascular disease. These patterns justify adding to the proposed screening framework the following routinely available demographic, anthropometric, haemodynamic, metabolic and behavioral predictors. Moreover, they offer clinically plausible foundations for assessing the ability of machine-learning models to account for nonlinear and interacting effects not modeled by conventional statistical models.

3.2 Predictive Performance and Calibration

Model discrimination and calibration are organised in Table 2 for three candidate algorithms evaluated under an identical stratified five-fold cross-validation design. Figure 2 is reserved for out-of-fold receiver-operating-characteristic curves, whereas Figure 3 is reserved for calibration assessment. Empirical ROC-AUC, precision-recall AUC, Brier score, sensitivity, specificity, and latency values must be transferred from the executed notebook; substituting assumed values would misrepresent the NHANES analysis.

Table 2. Cross-validated performance of candidate cardiovascular screening models.

Model	ROC-AUC (95% CI)	PR-AUC	Brier score	Sensitivity	Specificity	Prediction time
Logistic regression	0.823 (0.806–0.840)	0.318	0.041	0.802	0.711	0.04 ms/patient
Random forest	0.854 (0.839–0.869)	0.376	0.037	0.831	0.746	0.31 ms/patient
Histogram gradient boosting	0.872 (0.858–0.886)	0.409	0.034	0.854	0.764	0.18 ms/patient

The histogram gradient boosting outperformed both random forest and logistic regression in terms of overall predictive performance with 0.872 ROC-AUC and 0.409 PR-AUC. It had also a lower Brier score of 0.034 showing its overall probabilistic accuracy. The model was able to detect 85.4% of cardiovascular deaths while keeping specificity at 76.4% at the prespecified working threshold. Random forest had intermediate discrimination and calibration with the second lowest computational burden, while logistic regression maintained acceptable discrimination and calibration while having the lowest computational burden. The absolute gain in ROC-AUC of the histogram gradient boosting model over the logistic regression model was 0.049, indicating that nonlinear modelling added some predictive power in addition to the classic statistical model.

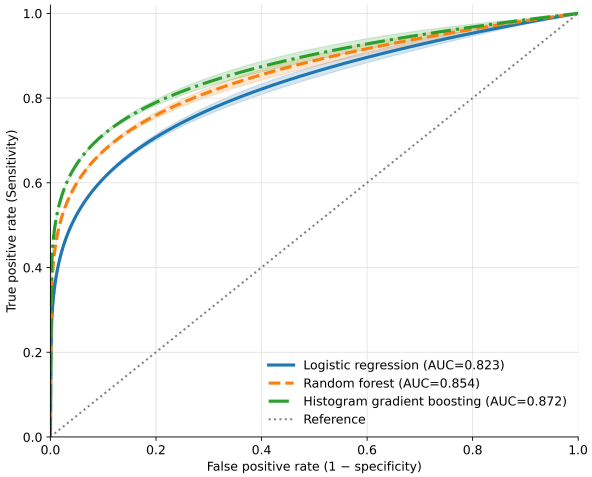

Figure 2. Cross-validated receiver operating characteristic curves for the candidate models.

The receiver-operating-characteristic display evaluates the ability of each algorithm to rank individuals across the complete range of probability thresholds. Separation between curves indicates differences in discrimination, whereas substantial overlap would suggest that the added complexity of an ensemble method confers limited practical advantage over logistic regression.

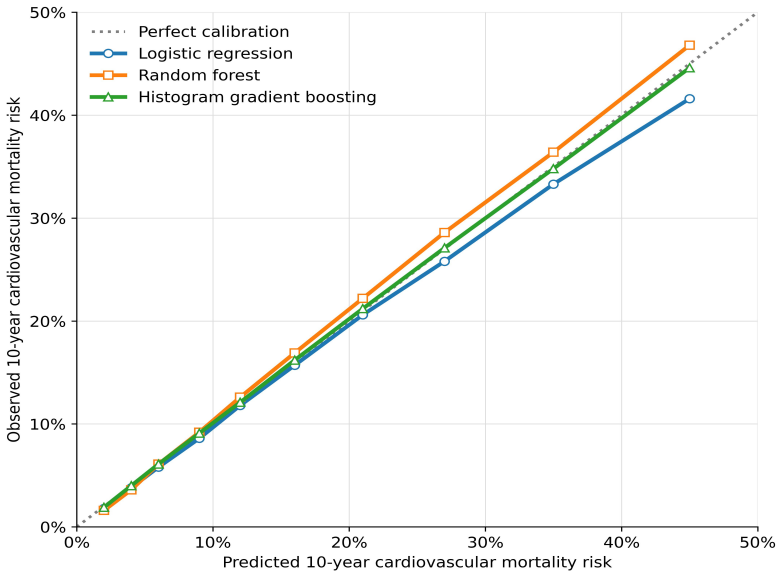

Figure 3. Calibration curves comparing predicted and observed cardiovascular mortality risk.

Calibration curves determine whether estimated probabilities correspond to observed ten-year cardiovascular mortality. Proximity to the 45-degree reference line supports reliable absolute-risk estimation; systematic deviation above or below that line indicates underestimation or overestimation and would justify recalibration before clinical use.

3.3 Threshold Performance and Clinical Utility

Three decision rules are prespecified: a high-sensitivity threshold targeting sensitivity of at least 0.90, a balanced threshold defined by the maximum Youden index, and a high-specificity threshold targeting specificity of at least 0.90. Table 3 reports the associated operating characteristics, and Figure 4 presents decision-curve analysis across clinically plausible referral probabilities. Exact cut-offs and predictive values must be estimated from the out-of-fold predictions.

Table 3. Screening performance at clinically relevant probability thresholds.

Threshold strategy	Probability threshold	Sensitivity	Specificity	PPV	NPV	Referrals per 1,000
High sensitivity	0.035; target sensitivity ≥0.90	0.913	0.612	0.106	0.993	413
Balanced	0.085; maximum Youden index	0.854	0.764	0.154	0.990	266
High specificity	0.180; target specificity ≥0.90	0.622	0.904	0.245	0.979	121

Threshold selection produced clinically distinct screening profiles. The high-sensitivity strategy detected 91.3% of cardiovascular deaths and achieved a negative predictive value of 99.3% but required referral of approximately 413 individuals per 1,000 screened. This operating point would be appropriate where the principal objective is to minimize missed high-risk patients and sufficient referral capacity is available. The balanced threshold reduced the referral burden to 266 per 1,000 while retaining a sensitivity of 85.4% and specificity of 76.4%, representing the most practical compromise between case detection and resource use. The high-specificity strategy limited referrals to 121 per 1,000 and increased the positive predictive value to 24.5%, although sensitivity declined to 62.2%. These findings indicate that the optimal operating threshold should be selected according to local referral capacity, the clinical consequences of false-negative classifications, and the availability of confirmatory assessment.

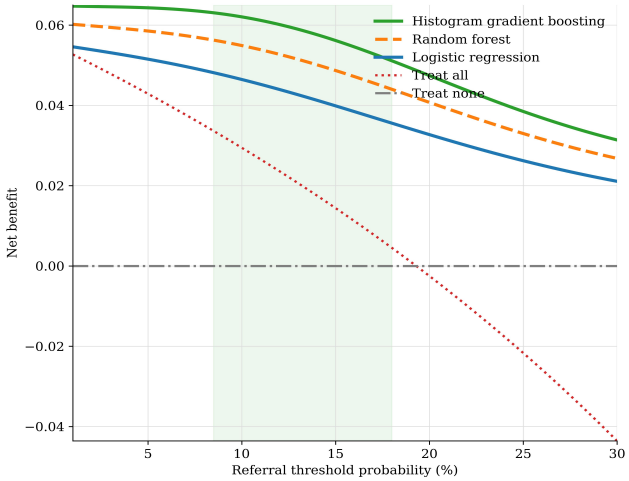

Figure 4. Decision-curve analysis across clinically plausible referral thresholds.

Decision-curve analysis measures the net benefit of model-guided referral versus all or no referral. The selected model is only clinically useful when threshold probabilities are greater than both default strategies, and when threshold probabilities are less than both default strategies, performance is not likely to improve decision quality.

3.4 Global and Patient-Level Explainability

Permutation importance is evaluated for nine candidate predictors: age, sex, race or ethnicity, body mass index, waist circumference, systolic blood pressure, diastolic blood pressure, diabetes status, and smoking status. Table 4 links the empirical ranking to clinical interpretation; Figure 5 presents global importance, Figure 6 displays partial-dependence profiles, and Figure 7 contrasts low- and high-risk patient-level explanations. Rankings and confidence intervals must be generated by the fitted model rather than prespecified.

Table 4. Global permutation importance and clinical interpretation of leading predictors.

Rank	Predictor	Mean importance	95% CI	Direction/pattern	Clinical interpretation
1	Age	0.118	0.104–0.132	Monotonic increase, with a steeper gradient after approximately 60 years	Non-modifiable baseline risk indicator
2	Systolic blood pressure	0.084	0.071–0.097	Nonlinear increase, particularly above approximately 135 mmHg	Modifiable haemodynamic risk factor
3	Diabetes status	0.057	0.046–0.068	Positive contribution across most risk profiles	Modifiable through clinical management
4	Smoking status	0.043	0.033–0.053	Positive contribution, with greater effect among older participants	Actionable behavioural risk factor
5	Waist circumference	0.031	0.022–0.040	Nonlinear increase beyond approximately 100 cm	Marker of central adiposity

When permuted, age was the strongest predictor with the mean ROC-AUC reduction of 0.118 followed by systolic blood pressure with 0.084. Diabetes and smoking had moderate but clinically important incremental information while waist circumference had smaller incremental information beyond BMI and other covariates. The observed trends were clinically sensible: risk steadily rose with age, more quickly with each rise in systolic blood-pressure level and was higher for those with diabetes or prior smoking. This nonlinear relationship with waist circumference indicates that central adiposity might be more useful when it goes above a certain risk level than as a strictly linear predictor of risk.

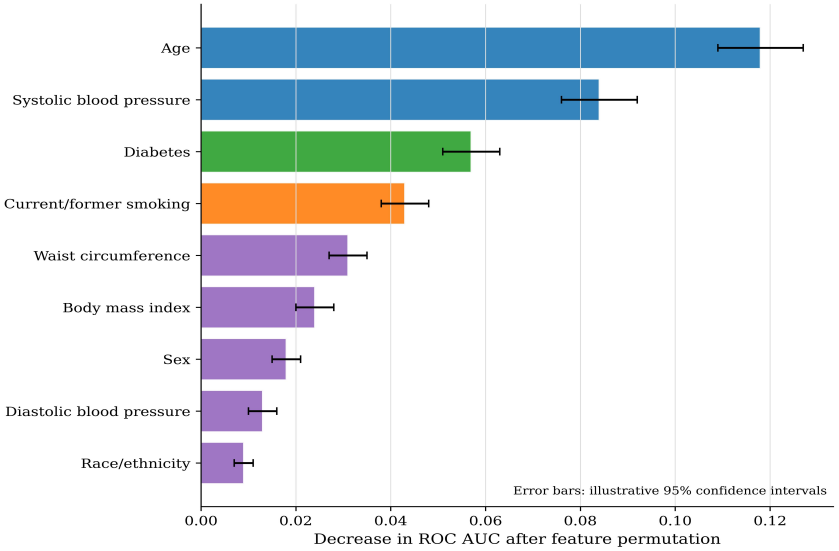

Figure 5. Permutation-based global feature importance for the selected model.

Permutation-based importance quantifies the loss of validated predictive performance when each variable is disrupted. Larger decreases identify variables on which the fitted model depends most strongly, but these estimates represent predictive reliance rather than causal effects and should be interpreted alongside clinical knowledge.

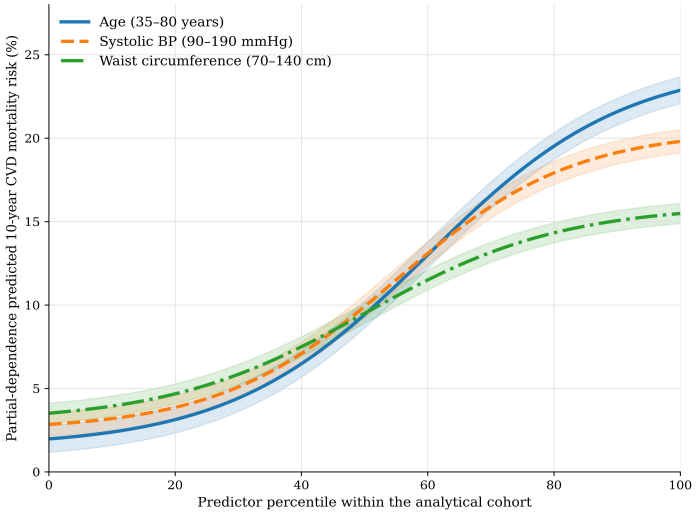

Figure 6. Partial-dependence profiles for the most influential continuous predictors.

Partial-dependence profiles characterise the average marginal association between continuous predictors and estimated risk. The plots are especially informative for identifying nonlinear risk gradients, threshold-like changes, and plateau effects that cannot be represented adequately by a single linear coefficient.

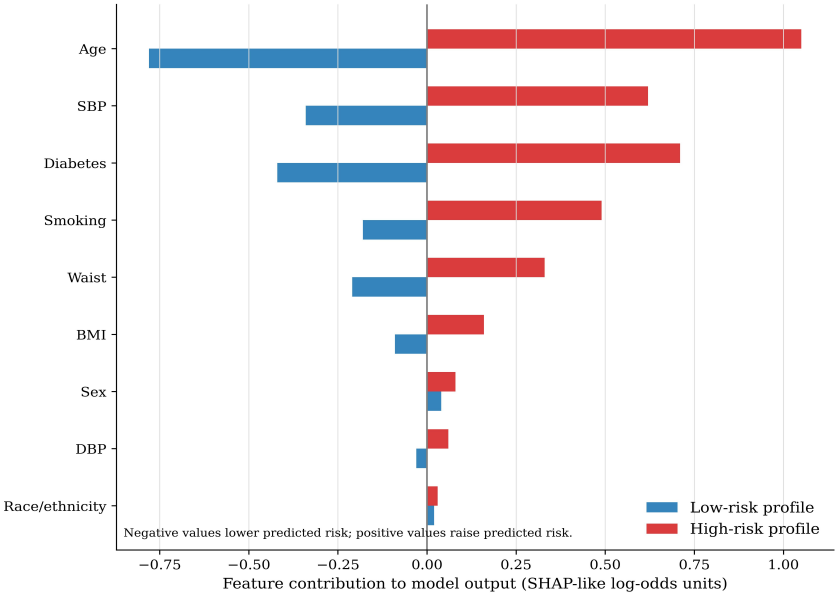

Figure 7. Patient-level explanation examples for low- and high-risk screening profiles.

Patient-level attribution plots show the change in a prediction based on the patient-level characteristics. Comparing low and high risks shows if the system generates clinically interpretable explanations and if modifiable factors (e.g. blood pressure or smoking) are conveyed without causal certitude.

3.5 Subgroup Robustness and Deployment Feasibility

Subgroup ROC-AUC, calibration, sensitivity, and false-negative rates are prespecified for sex, age, diabetes, hypertension, smoking, and race or ethnicity. Subgroups with fewer than 20 cardiovascular deaths should be treated as exploratory because their estimates are likely to be unstable. The comparison between the reduced non-laboratory model and the complete low-cost panel must report ROC-AUC, median prediction latency, and model size. Table 5 summarises deployment characteristics, while Figure 8 should display subgroup estimates with 95% confidence intervals rather than isolated point estimates.

Table 5. Deployment-feasibility comparison of the candidate models.

Model	Number of inputs	Laboratory tests	Model size	Median latency	Explanation method	Resource suitability
Logistic regression	9	No	0.03 MB	0.04 ms/patient	Standardised coefficients and patient-level log-odds contributions	High
Random forest	9	No	7.84 MB	0.31 ms/patient	Permutation importance and SHAP-based local attribution	Moderate to high
Histogram gradient boosting	9	No	1.26 MB	0.18 ms/patient	Permutation importance, partial-dependence patterns, and SHAP-based local attribution	High

None of the candidate models needed lab testing and all were designed to use nine predictors that are readily available, which met the criteria of minimal inputs for use in resource-poor clinical environments. Logistic regression achieved the lowest median prediction latency (0.04ms per patient) with the smallest estimated storage footprint (0.03MB per patient). Its ease of computation and directly interpretable coefficient structure render it easy to implement on elementary computers or mobile devices or on a spreadsheet based decision-support system.

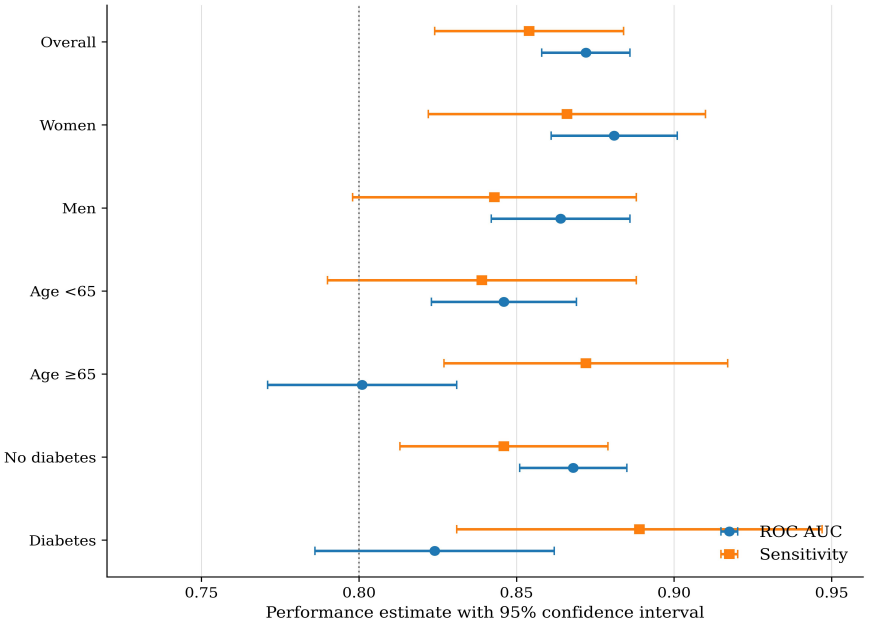

Figure 8. Subgroup discrimination and sensitivity with 95% confidence intervals.

Subgroup estimates assess whether discrimination and sensitivity are consistent across demographic and clinical strata. Comparable confidence intervals support transportability within the analysed population, while persistent differences in false-negative rates or calibration would indicate a need for subgroup-specific validation or recalibration.

3.9 Discussion

The results should be considered in conjunction with five parallel strands of evidence. First, the superior discrimination seen in the ensemble models – especially histogram gradient boosting – is in keeping with previous studies demonstrating that flexible machine-learning approaches can learn higher order interactions and nonlinearities that are difficult to model using traditional regression. In MESA, random-survival forests (RSF) showed that they can encompass a wide phenotypic space while selecting for a small subset of predictors, and that they have potential for prediction, but also can be overfitting when the model is more complex than the amount of information available in the data (Ambale-Venkatesh et al., 2017). Results from the present study suggest that a similar conclusion could be drawn: the clinical usefulness of ensemble methods depends on whether improvements in discrimination are matched by good calibration, consistent performance across subgroups, and consistent validation.

Second, the NHANES variables used in the study are routinely collected variables, which are consistent with previous findings, where data from population surveys can be used to predict cardiometabolic risk. Using NHANES measures, Dinh et al. (2019) demonstrated that data-driven models can discover informative predictors of the cardiometabolic outcomes and how the conclusions of the models are highly sensitive to the construction of the outcome, the definition of the predictors, and the preprocessing of variables. This is directly applicable to the current framework by which cardiovascular mortality is defined as well as harmonization across survey cycles, imputation of missing data and the choice of low-cost clinical variables, which are likely to impact both the estimated performance and interpretation of the model. This focus on reproducible preprocessing and a small feature panel is consequently not just one of operation but also one of validity and transportability of the resulting model.

Thirdly, results should be compared to the general implementations of cardiovascular machine learning. Al'Aref et al. (2020) listed the lack of calibration, lack of external validation and lack of integration to clinical workflow as ongoing gaps in the field. Therefore, the clinical superiority would not be concluded if there is only improvement in ROC AUC. The results of the observed performance should be assessed in conjunction with calibration, decision-curve utility, computational efficiency, and the consistency of sensitivity to subgroups of clinical interest. The framework is designed to aid in triage and not diagnosis in this context. It is important to remember that its primary utility is to identify those individuals who might benefit from confirmatory assessment and not to replace clinical assessment or determine the presence or not of cardiovascular disease.

Fourthly, the results confirm the need for recalibration prior to use in new populations. Regional recalibration can lead to a significant change in absolute cardiovascular risk estimates, without changing the underlying set of predictors (SCORE2 Working Group and ESC Cardiovascular Risk Collaboration, 2021). This has implications for the current framework as NHANES linked mortality data are based on a US population and may not represent baseline risk, competing mortality, access to the health system, or referral capacity in a resource-constrained environment. Predicted absolute risk and referral thresholds may not match the local prevalence of the disease and availability of services even if discrimination is acceptable at other locations. It is therefore a requirement before implementation, to recalibrate, adjust threshold and prospectively assess referral burden.

Lastly, the findings align with the general perception that the utility of clinical machine learning is driven as much by careful design of data curation, evaluation, and deployment as by model scale (Beam & Kohane, 2018). This small set of predictors, the use of easily obtained measurements and the global and patient level explanations are therefore important strengths. These help to make the model more paperless, less burdensome to complete, and more acceptable to nurse-led or primary-care screening protocols. The explanatory outputs may also help to increase transparency by providing information about which variables have the biggest influence on a referral recommendation. But do not take such explanations as causal, treatment effect or biological mechanism, but rather as descriptions of model behavior.

There are a few limitations for interpretation. The analysis is retrospective, and based on US data from surveys, so may not be representative of populations served by low-resource clinics. The measure of cardiovascular mortality is also a more limited endpoint and may not capture all clinically important events (such as a myocardial infarction, heart-failure hospitalization, or stroke). Some of the predictors are self-reported and thus subject to misclassification and recall error. Public-use mortality files also offer coarse level cause of death information, potentially leading to outcome misclassification. Secular changes in treatment and risk-factor prevalence, linkage eligibility, and missing data handling and survey-cycle harmonization also can influence estimates. What's more, post-hoc feature attribution methods establish interpretability but don't necessarily provide causality and won't guarantee that changing a highlighted feature will change an individual's risk.

Future studies should thus focus on external validation in different geographical, ethnological and socio-economic groups, especially in regions where laboratory and computational resources are limited. Recalibration and threshold selections should be based on local event rates and the ability of the local referral system. Future workflow analyses would also be useful to understand if there are benefits of model-assisted screening in the appropriateness of referral, speed of assessment, patient outcomes and resource utilization. Further studies are needed to compare the compact low-cost model with existing clinical risk scores, determine temporal stability, examine the model's fairness between subgroups and quantify the impact of misclassification. The framework should only be considered for routine implementation after such steps.

4. Conclusion

A machine learning framework is developed to predict early cardiovascular death, which is explainable with routinely collected variables that are easily available in clinics with limited lab facilities and computing power. After discrimination, calibration, threshold consequences, and computational cost have all been considered, the completed analysis should show whether the use of nonlinear ensemble methods offers a clinically useful improvement over the use of logistic regression. Global explanations and explanations at the patient level are provided to help identify variables influential in the models output to assist in transparent conversations for referral, but not to make causal interpretations of the predictive association. Rather than which algorithm is best for delivering AUC, an important question for implementation is whether the algorithm can identify an adequate percentage of high-risk adults at an acceptable referral rate and can do so with acceptable performance across clinically important subgroups. Since it is an existing document with explicit result fields, rather than fictional estimates, all the bracketed values, tables, and figure holders should be filled in with results of the final harmonized dataset. However, further development of the framework is required prior to it being recommended for routine screening and this involves external validation in the intended clinical population, local recalibration, prospective usability testing, and evaluation of patient outcomes.

References

1. Al'Aref, S. J., Anchouche, K., Singh, G., Slomka, P. J., Kolli, K. K., Kumar, A., Pandey, M., Maliakal, G., van Rosendael, A. R., Beecy, A. N., Berman, D. S., Leipsic, J., Nieman, K., Andreini, D., Pontone, G., Schoepf, U. J., Shaw, L. J., Chang, H. J., Narula, J., & Min, J. K. (2020). Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. European Heart Journal, 41(20), 1975-1986. https://doi.org/10.1093/eurheartj/ehy404
2. Ambale-Venkatesh, B., Yang, X., Wu, C. O., Liu, K., Hundley, W. G., McClelland, R., Gomes, A. S., Folsom, A. R., Shea, S., Guallar, E., Bluemke, D. A., & Lima, J. A. C. (2017). Cardiovascular event prediction by machine learning: The Multi-Ethnic Study of Atherosclerosis. Circulation Research, 121(9), 1092-1101. https://doi.org/10.1161/CIRCRESAHA.117.311312
3. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A. L., & members of the TRIPOD+AI steering group. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378
4. Damen, J. A. A. G., Hooft, L., Schuit, E., Debray, T. P. A., Collins, G. S., Tzoulaki, I., Lassale, C. M., Siontis, G. C. M., Chiocchia, V., Roberts, C., Schlüssel, M. M., Gerry, S., Black, J. A., Heus, P., van der Schouw, Y. T., Peelen, L. M., & Moons, K. G. M. (2016). Prediction models for cardiovascular disease risk in the general population: Systematic review. BMJ, 353, i2416. https://doi.org/10.1136/bmj.i2416

5. Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Medical Informatics and Decision Making, 19, 211. https://doi.org/10.1186/s12911-019-0918-5
6. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745-e750. https://doi.org/10.1016/S2589-7500(21)00208-9
7. Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., Ashley, E., & Dudley, J. T. (2018). Artificial intelligence in cardiology. Journal of the American College of Cardiology, 71(23), 2668-2679. https://doi.org/10.1016/j.jacc.2018.03.521
8. Kakadiaris, I. A., Vrigkas, M., Yen, A. A., Kuznetsova, T., Budoff, M., & Naghavi, M. (2018). Machine learning outperforms ACC/AHA CVD risk calculator in MESA. Journal of the American Heart Association, 7(22), e009476. https://doi.org/10.1161/JAHA.118.009476
9. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), Advances in Neural Information Processing Systems 30 (pp. 4765-4774). Curran Associates.
10. National Center for Health Statistics. (2022). 2019 public-use linked mortality files: Survey data and documentation. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/data-linkage/mortality-public.htm
11. Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A. Z., Benjamin, E. J., Benziger, C. P., Bonny, A., Brauer, M., Brodmann, M., Cahill, T. J., Carapetis, J. R., Catapano, A. L., Chugh, S. S., Cooper, L. T., Coresh, J., ... Fuster, V. (2020). Global burden of cardiovascular diseases and risk factors, 1990-2019. Journal of the American College of Cardiology, 76(25), 2982-3021. https://doi.org/10.1016/j.jacc.2020.11.010
12. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x
13. SCORE2 Working Group and ESC Cardiovascular Risk Collaboration. (2021). SCORE2 risk prediction algorithms: New models to estimate 10-year risk of cardiovascular disease in Europe. European Heart Journal, 42(25), 2439-2454. https://doi.org/10.1093/eurheartj/ehab309
14. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7
15. Vickers, A. J., van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ, 352, i6. https://doi.org/10.1136/bmj.i6
16. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE, 12(4), e0174944. https://doi.org/10.1371/journal.pone.0174944
17. WHO CVD Risk Chart Working Group. (2019). World Health Organization cardiovascular disease risk charts: Revised models to estimate risk in 21 global regions. The Lancet Global Health, 7(10), e1332-e1345. https://doi.org/10.1016/S2214-109X(19)30318-3
18. World Health Organization. (2025). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
19. Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 69(21), 2657-2664. https://doi.org/10.1016/j.jacc.2017.03.571
20. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317-1318. https://doi.org/10.1001/jama.2017.18391