

Explainable Machine Learning for Prevalent Cardiovascular Disease Stratification Using Clinical, Sociodemographic, and Lifestyle Data: Temporal Validation in NHANES

MD ROKI HOSEN

rockyhosen50@gmail.com

Nanjing University of Information Science and Technology

TAWHIB SALAH ABDO MOQBEL

Nanjing University of Information Science and Technology

Research Article

Keywords: cardiovascular disease, machine learning, XGBoost, explainable artificial intelligence, SHAP, probability calibration, temporal validation, NHANES, decision curve analysis, prevalent disease stratification

Posted Date: October 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10984635/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Explainable Machine Learning for Prevalent Cardiovascular Disease Stratification Using Clinical, Sociodemographic, and Lifestyle Data: Temporal Validation in NHANES

HOSEN MD ROKI

*Department of Artificial Intelligence
Nanjing University of Information Science and Technology
Nanjing, China
Email: rockyhosen50@gmail.com*

MOQBEL TAWHIB SALAH ABDO

*Department of Artificial Intelligence
Nanjing University of Information Science and Technology
Nanjing, China
Email: tawhibsalah@gmail.com*

Corresponding author: HOSEN MD ROKI

Email: rockyhosen50@gmail.com

Abstract

Background: Machine-learning methods can capture nonlinear associations in multidomain health data, but discrimination alone does not establish the reliability of probability-based disease stratification. Calibration, temporal transportability, threshold-dependent behavior, subgroup heterogeneity, and model interpretability require separate evaluation.

Objective: To develop and temporally validate an explainable machine-learning framework for stratifying prevalent physician-diagnosed cardiovascular disease (CVD), while evaluating discrimination, probability calibration, decision-analytic behavior, subgroup robustness, and explanation stability.

Methods: National Health and Nutrition Examination Survey (NHANES) data from 2007–2018 were analyzed. Adults aged ≥ 20 years with a known composite CVD outcome were included. Prevalent CVD was defined from self-reported physician diagnoses of congestive heart failure, coronary heart disease, angina, heart attack, or stroke. NHANES 2007–2012 constituted the development period and 2013–2018 the chronologically later temporal-validation period. Logistic regression, random forest, XGBoost, and LightGBM were compared and tuned using development-only five-fold cross-validation. XGBoost was selected according to the predefined development ROC-AUC ranking criterion. A sigmoid calibration mapping and a Youden-derived analytical operating point were established using development data and locked before temporal evaluation. The uncalibrated and isotonic probability outputs were retained as sensitivity analyses. Temporal evaluation included ROC-AUC, average precision, Brier score, log loss, calibration intercept and slope, bootstrap confidence intervals, threshold-dependent metrics, decision-curve analysis, predictor-domain ablation, subgroup analysis, and SHAP-based explainability.

Results: The analytical cohort included 34,602 adults: 17,626 in development and 16,976 in temporal validation. The locked sigmoid-calibrated XGBoost achieved a temporal ROC-AUC of 0.847 (95% CI 0.839–0.856) and average precision of 0.403 (95% CI 0.382–0.427), compared with a prevalence baseline of approximately 0.113. The Brier score was 0.083 (95% CI 0.080–0.086), calibration intercept was -0.067 (95% CI -0.164 to 0.038), and calibration slope was 0.981 (95% CI 0.943 – 1.022). The uncalibrated XGBoost had a lower temporal Brier score (0.0809), indicating that the sigmoid layer did not improve overall probability accuracy. At the development-locked analytical operating point of 0.0760, sensitivity was 0.820, specificity 0.718, positive predictive value 0.270, and negative predictive value 0.969. Operating characteristics varied substantially by age. Age was the dominant SHAP predictor, and original-predictor importance rankings were stable across temporal cycles (pairwise Spearman $\rho = 0.982$ – 0.996).

Conclusions: The development-locked framework maintained useful discrimination and near-ideal aggregate calibration when transported chronologically to later NHANES cycles, although sigmoid calibration did not improve the Brier score relative to the uncalibrated model. Social-context and lifestyle/exposure variables provided modest incremental predictive information beyond the core clinical-demographic predictors. Because the outcome was prevalent physician-diagnosed CVD rather than prospective cardiovascular events, the model should be interpreted as a prevalent-CVD status-stratification framework rather than a longitudinal event-risk calculator. Independent external and prospective validation, including purpose-specific threshold development, is required before clinical use.

Keywords: cardiovascular disease; machine learning; XGBoost; explainable artificial intelligence; SHAP; probability calibration; temporal validation; NHANES; decision curve analysis; prevalent disease stratification.

1 Introduction

Cardiovascular disease (CVD) remains a major global public-health burden and encompasses a heterogeneous group of disorders affecting the heart and vascular system. Its development and clinical expression reflect the combined influence of aging, blood-pressure abnormalities, metabolic dysfunction, adiposity, tobacco exposure, socioeconomic conditions, behavioral factors, and established comorbidities [1, 2]. Because these determinants may interact in nonlinear and context-dependent ways, cardiovascular data provide an important setting for the development and evaluation of multivariable prediction and stratification models.

Cardiovascular risk assessment has traditionally relied on regression-based models constructed from established clinical predictors. The Framingham general cardiovascular risk profile estimates future cardiovascular risk using variables including age, blood pressure, cholesterol, smoking, and diabetes [3]. The American College of Cardiology/American Heart Association Pooled Cohort Equations similarly estimate 10-year risk of a first atherosclerotic cardiovascular disease event [4]. QRISK3 incorporates a broader set of demographic, socioeconomic, comorbidity-related, and clinical characteristics [5], while SCORE2 provides region-specific estimation of future cardiovascular-event risk in European populations [6]. These models demonstrate the continuing value of carefully developed conventional prediction approaches and provide important methodological context for newer machine-learning methods.

Machine-learning (ML) algorithms offer a complementary approach by allowing nonlinear associations and higher-order interactions to be learned directly from data. Tree-based ensemble methods, including random forests, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM), have consequently received substantial attention for prediction using structured clinical and population-health data. XGBoost combines gradient boosting with regularization and efficient tree construction and has become widely used for tabular prediction problems [7].

Greater algorithmic flexibility, however, does not necessarily produce large improvements in cardiovascular prediction. Previous studies have reported that machine-learning approaches can outperform conventional regression in some settings, but the magnitude of improvement is often modest and dependent on study design, outcome definition, predictor availability, and validation strategy [8, 9, 10, 11]. Systematic reviews have also identified important limitations in cardiovascular ML research, including heterogeneous methodological quality, incomplete reporting, limited independent validation, inconsistent calibration assessment, and substantial risk of bias [11, 12, 13]. These observations suggest that reliable cardiovascular ML requires more than selecting the algorithm with the highest discrimination metric.

Discrimination represents only one dimension of predictive performance. A model may achieve a high area under the receiver operating characteristic curve (ROC-AUC) while producing probability estimates that are systematically too high, too low, or excessively extreme. Calibration—the agreement between predicted probabilities and observed outcome frequencies—is therefore important whenever model probabilities are interpreted quantitatively or used to define threshold-dependent classifications [14, 15]. Evaluation based only on ROC-AUC or accuracy may consequently provide an incomplete picture of model reliability.

Temporal transportability is another important consideration. Many machine-learning studies use random training and testing partitions drawn from the same overall observation period. Such designs are useful for internal validation but do not directly evaluate whether a fully developed model continues to perform when applied to individuals sampled during later calendar

periods. Population characteristics, disease prevalence, exposure distributions, laboratory methods, healthcare practices, and measurement processes may change over time. Chronological temporal validation therefore provides a more demanding assessment of model stability because it more closely approximates the situation in which a model developed using historical data is later applied without refitting to newly observed individuals [16, 17].

The National Health and Nutrition Examination Survey (NHANES) provides a useful setting for examining these questions. NHANES combines standardized interviews, physical examinations, laboratory measurements, and health questionnaires within repeated national survey cycles. Its repeated-cycle structure permits chronological separation of earlier and later survey periods, while its multidomain measurements include demographic, socioeconomic, anthropometric, clinical, laboratory, and lifestyle/exposure information. Official NHANES guidance also provides a framework for accounting for the survey’s complex sampling design in population-oriented descriptive analyses [18].

Recent NHANES-based studies have demonstrated the potential of machine-learning methods for cardiovascular classification and risk-factor characterization. Huang and Huang applied XGBoost to NHANES data for coronary artery disease and used model-interpretation methods to examine relevant predictors [19]. Cao et al. evaluated multiple ML algorithms for cardiovascular disease prediction in NHANES and reported strong XGBoost performance while identifying age, hypertension, and poverty-to-income ratio among important predictors [20]. Ahiduzzaman and Hasan compared several machine-learning algorithms using NHANES dietary and health information and incorporated SHAP and LIME to improve model interpretability [21]. These studies demonstrate the usefulness of NHANES for cardiovascular ML, but they also reinforce the need to examine model reliability beyond a single discrimination estimate.

Interpretability is particularly relevant when flexible ensemble models are used with clinical and population-health variables. Unlike a conventional regression coefficient, the internal decision structure of a boosted-tree model is not immediately interpretable. SHapley Additive exPlanations (SHAP) provide a theoretically grounded framework for attributing fitted model output to individual features and can be summarized at both global and participant levels [22]. Cardiovascular explainable-AI studies have shown that SHAP-based analyses can identify clinically recognizable patterns while improving transparency regarding model behavior [23]. Nevertheless, SHAP values explain the fitted model rather than establish causal relationships, and explanation quality should not be reduced to a single importance plot. If a model is transported over time, stability of the predictor-importance hierarchy provides an additional dimension of model auditing.

Threshold-dependent evaluation also requires attention. Conventional performance measures do not directly characterize the consequences of classifying participants as positive or negative at a chosen probability cutoff. Decision-curve analysis addresses this issue by estimating net benefit across threshold probabilities and comparing model-guided decisions with reference strategies such as assessing all participants or assessing none [24]. Such analyses can complement discrimination and calibration by describing decision-analytic trade-offs, although they do not demonstrate that use of the model improves patient outcomes in real clinical practice.

Contemporary guidance increasingly emphasizes these broader dimensions of prediction-model quality. TRIPOD+AI recommends transparent reporting of participants, predictors, outcomes, model development, validation, and performance for prediction models using regression or machine-learning methods [25]. PROBAST+AI similarly provides a framework for evaluating model quality, risk of bias, and applicability [26]. Together, these principles support evaluating a machine-learning system as a complete development-and-validation framework rather than focusing solely on its highest ROC-AUC.

1.1 Research gap and study contributions

Existing evidence supports the feasibility of machine learning for cardiovascular stratification, but systematic reviews continue to identify limited independent validation, inconsistent calibration assessment, heterogeneous methodological quality, and substantial risk of bias [11, 12, 13]. NHANES-based studies have also shown that XGBoost, random forests, SHAP, and multidomain health information can be informative [19, 20, 21]. The remaining methodological question is therefore not simply whether a flexible classifier can achieve a high ROC-AUC, but whether a complete development-locked analytical framework remains reliable when applied unchanged to later calendar periods.

The present study addressed this question through strict chronological separation of model development and temporal evaluation. Algorithm comparison, hyperparameter tuning, preprocessing, calibration, analytical operating-point selection, and all other development-stage choices were performed using NHANES 2007–2012 only. The resulting framework was then applied unchanged to NHANES 2013–2018.

The study had four principal contributions:

1. development-only model comparison and tuning followed by locked chronological temporal validation;
2. integrated evaluation of discrimination, precision–recall performance, proper scoring rules, aggregate calibration, bootstrap uncertainty, and threshold-dependent operating characteristics;
3. secondary decision-curve, subgroup, temporal-cycle, and predictor-domain analyses to characterize decision-analytic behavior, heterogeneity, temporal stability, and incremental information from social-context and lifestyle/exposure variables; and
4. global and local SHAP analysis together with temporal assessment of original-predictor importance-rank stability.

The primary objective was to determine whether a development-selected machine-learning framework maintained useful discrimination and acceptable probability behavior when transported from NHANES 2007–2012 to 2013–2018. Secondary objectives were to characterize incremental predictor-domain value, temporal-cycle stability, subgroup heterogeneity, threshold-dependent behavior, decision-analytic net benefit, and SHAP-derived explanation stability.

The target was *prevalent physician-diagnosed cardiovascular disease* assessed at NHANES participation. It was not incident CVD observed prospectively. The predicted probability therefore represents membership in the prevalent-CVD group and must not be interpreted as a 5-year or 10-year probability of a future cardiovascular event. Established scores such as the Framingham general cardiovascular risk profile, the ACC/AHA Pooled Cohort Equations, QRISK3, and SCORE2 target future events over defined horizons [3, 4, 5, 6]; direct numerical comparison with the present cross-sectional target would therefore not be like-for-like.

2 Materials and Methods

2.1 Study design

A population-based machine-learning study with chronological temporal validation was conducted using continuous NHANES data from 2007 through 2018. Before model fitting, NHANES 2007–2008, 2009–2010, and 2011–2012 were assigned to the development period, whereas 2013–2014, 2015–2016, and 2017–2018 were assigned to the temporal-validation period.

The temporal-validation cohort was not used to select predictors, choose the algorithm, tune hyperparameters, determine the primary calibration specification, select the analytical operating point, or modify preprocessing. These choices were established using development data and locked before temporal evaluation. The 2013–2018 cohort therefore served as the principal out-of-sample assessment of temporal generalization for the frozen analytical framework.

2.2 Data source

NHANES is conducted by the National Center for Health Statistics and uses interviews, standardized physical examinations, and laboratory measurements to characterize the health and nutritional status of the United States population. The present analysis used harmonized continuous NHANES data from 2007 through 2018.

The analytical dataset integrated demographic, questionnaire, examination, and laboratory components using the NHANES respondent sequence identifier. Variables included demographic characteristics, anthropometric measurements, repeated blood-pressure measurements, total and HDL cholesterol, glycated hemoglobin, physician-reported diabetes, serum cotinine, recreational physical activity, and self-reported physician diagnoses used to construct the CVD outcome.

The full merged source population included participants from all six survey cycles. Adults aged ≥ 20 years with sufficient information to define the composite CVD outcome were eligible for the primary analysis. Final cohort counts are reported in Section 3.1.

Survey-design variables, including MEC examination weights, strata, and primary sampling units, were retained for population-oriented descriptive analysis. Survey weighting was used for descriptive population estimates but not for the primary participant-level machine-learning performance metrics.

2.3 Outcome definition

The primary target was a binary composite indicator of prevalent physician-diagnosed CVD.

Five NHANES medical-condition variables were used:

- MCQ160B: congestive heart failure;
- MCQ160C: coronary heart disease;
- MCQ160D: angina or angina pectoris;
- MCQ160E: heart attack; and
- MCQ160F: stroke.

For each component, a definite “Yes” response was mapped to 1 and a definite “No” response to 0. The composite CVD outcome was assigned as positive if any of the five component conditions was positive.

A participant was classified as CVD-negative only when all five components were definite negative responses. Participants without a positive component but with incomplete or uncertain component information were considered to have an unknown composite outcome and were excluded from primary modeling.

This rule avoided treating incomplete information as confirmed absence of disease. Because the component outcomes represented previous self-reported physician diagnoses, the composite target represented prevalent CVD rather than future incident cardiovascular events.

2.4 Predictor selection

Fifteen original predictors representing four interpretable information domains were included.

2.4.1 Demographic and social-context predictors

Age, sex, race/ethnicity, and poverty-to-income ratio were included. Age was derived from RIDAGEYR, sex from RIAGENDR, race/ethnicity from RIDRETH1, and poverty-to-income ratio from INDFMPIR.

2.4.2 Anthropometric and cardiovascular predictors

Body mass index (BMI), waist circumference, systolic blood pressure, and diastolic blood pressure were included.

BMI was derived from BMXBMI, and waist circumference was derived from BMXWAIST.

Multiple NHANES blood-pressure measurements were available. According to the implemented analysis rule, when only one measurement was available that measurement was retained. When multiple measurements were available, the first reading was excluded and the subsequent available measurements were averaged. For diastolic measurements, zero values were excluded when positive measurements were also available after application of the reading rule.

2.4.3 Laboratory and metabolic predictors

Total cholesterol, HDL cholesterol, glycated hemoglobin (HbA1c), and diabetes status were included.

Total cholesterol was derived from LBXTC, HDL cholesterol from LBDHDD, and HbA1c from LBXGH.

Diabetes was derived from DIQ010. Definite “Yes” and “No” responses were coded as 1 and 0, respectively. Borderline, refused, or unknown responses were retained as missing rather than being assigned to either binary category.

2.4.4 Lifestyle and exposure predictors

Serum cotinine, vigorous recreational physical activity, and moderate recreational physical activity represented the lifestyle/exposure domain.

Serum cotinine (LBXCOT) was used as an objective tobacco/nicotine-exposure biomarker rather than a derived smoking-status variable. Negative concentrations were treated as invalid and converted to missing.

Vigorous and moderate recreational physical activity were derived from PAQ650 and PAQ665, respectively. Definite “Yes” and “No” responses were converted to 1 and 0, whereas refused or unknown responses were treated as missing.

Alcohol consumption, fasting glucose, triglycerides, and a derived categorical smoking-status variable were not included in the final 15-predictor harmonized model.

2.5 Population-oriented descriptive analysis

Because NHANES uses a complex survey design, descriptive characteristics intended to reflect the sampled US population were estimated using survey-domain methods.

The complete merged survey frame was retained for variance estimation, while the analytical cohort was defined as the domain of adults aged ≥ 20 years with a known composite CVD outcome.

MEC examination weights (WTMEC2YR) were rescaled according to the number of combined survey cycles. Survey strata (SDMVSTRA) and primary sampling units (SDMVPSU) were incorporated, and Taylor-series linearization was used for variance estimation. Confidence intervals were based on the survey-design degrees of freedom.

Survey-weighted descriptive estimates are reported in Section 3.1. These survey weights were used to characterize population-level descriptive features of the development and temporal-validation cohorts and were not incorporated into model fitting, model selection, probability calibration, threshold selection, or the primary participant-level predictive-performance metrics.

Accordingly, the primary discrimination, calibration, probability-accuracy, and threshold-dependent performance estimates should be interpreted as participant-level model-performance measures rather than survey-weighted population-performance estimates.

2.6 Missing data and preprocessing

No imputation, scaling, or encoding was performed before chronological separation of the development and temporal-validation cohorts. Preprocessing was incorporated within the machine-learning pipelines so that all data-dependent transformations were estimated exclusively from the corresponding training data.

Continuous numerical predictors—age, poverty-to-income ratio, BMI, waist circumference, systolic blood pressure, diastolic blood pressure, total cholesterol, HDL cholesterol, and HbA1c—underwent median imputation, creation of missingness indicators, and standardization.

Serum cotinine was processed separately because of its strongly right-skewed distribution. A $\log_{1p}$ transformation was applied before median imputation, missingness-indicator creation, and

standardization.

Categorical and binary predictors—sex, race/ethnicity, diabetes, vigorous recreational activity, and moderate recreational activity—were imputed using the most frequent observed category and subsequently one-hot encoded.

All preprocessing transformations were fitted independently within each development cross-validation training fold and subsequently applied to the corresponding validation fold. For final temporal evaluation, preprocessing was fitted using the complete development cohort and applied unchanged to the temporal-validation cohort. Temporal-validation information therefore did not contribute to imputation values, scaling parameters, or category encoding.

The final preprocessing pipeline transformed the 15 original predictors into 32 model features after one-hot encoding and missingness-indicator generation. Predictor-specific missingness in the development and temporal-validation cohorts is reported in Supplementary Table S2. The final machine-learning structure is shown in Figure 1.

2.7 Candidate algorithms

Four candidate algorithms spanning linear and nonlinear modeling paradigms were evaluated:

- logistic regression;
- random forest;
- XGBoost; and
- LightGBM.

Logistic regression was included as a regularized linear reference model. The initial specification used L_2 regularization, $C=1.0$, the `lbfgs` solver, a maximum of 5,000 iterations, and no class weighting.

Random forest was included as a bagged tree-ensemble model [27]. Its initial specification used 500 trees, Gini impurity, unrestricted maximum depth, `min_samples_split=2`, `min_samples_leaf=5`, and `max_features=sqrt`. Bootstrap sampling was enabled, and no class weighting was applied.

XGBoost was included as a regularized gradient-boosted tree model [7]. The initial specification used 500 boosting iterations, learning rate 0.03, maximum depth 4, minimum child weight 3, subsample fraction 0.85, column-sampling fraction 0.85, L_1 regularization of 0.1, and L_2 regularization of 1.0. The binary logistic objective and histogram-based tree-construction method were used.

LightGBM was included as an additional gradient-boosted decision-tree approach [28]. Its initial specification used gradient-boosted decision trees with 500 boosting iterations, learning rate 0.03, 31 leaves, unrestricted maximum depth, minimum child sample size 30, subsample fraction 0.85, column-sampling fraction 0.85, L_1 regularization of 0.1, and L_2 regularization of 1.0.

Class weighting was not used in the primary candidate models because the study evaluated probability estimates and subsequent probability calibration under the observed development-cohort class distribution. Class imbalance was instead characterized using both ROC-AUC and average precision and addressed operationally through development-derived threshold selection.

Final Machine-Learning Model Structure

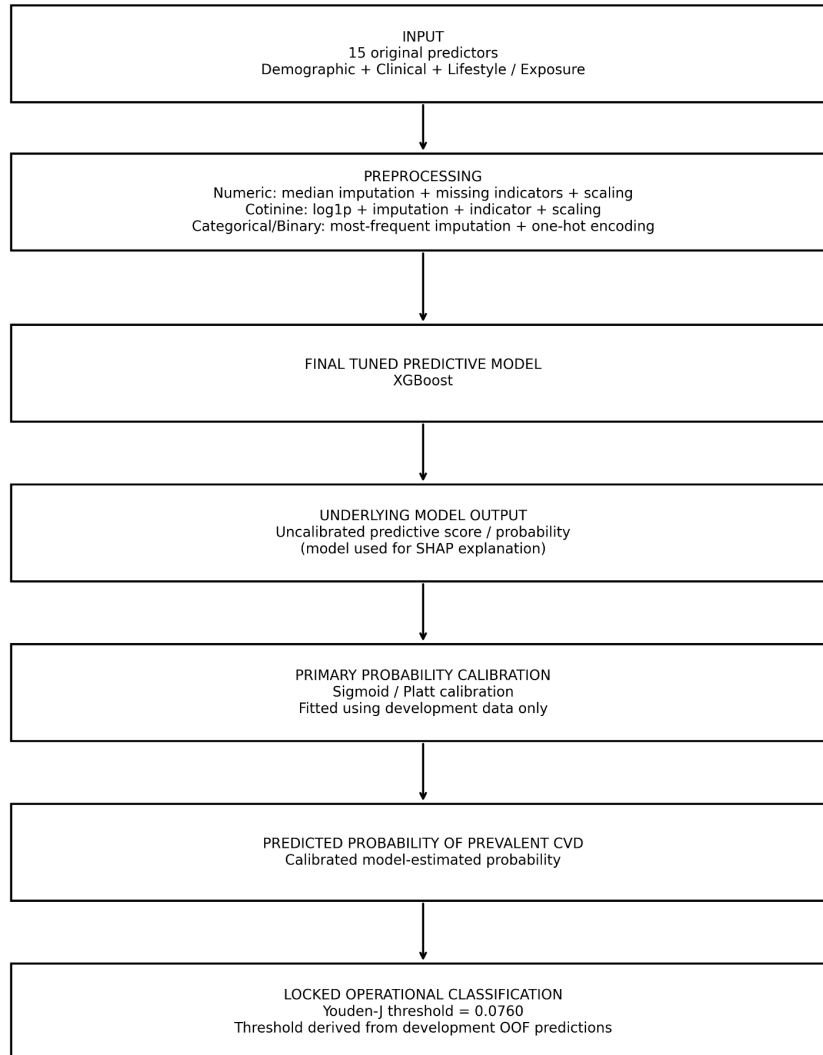


Figure 1: Final machine-learning model structure. Fifteen original demographic, social-context, clinical, laboratory, and lifestyle/exposure predictors were processed through a leakage-controlled preprocessing pipeline and entered into the final tuned XGBoost model. The underlying XGBoost output was used for SHAP explanation, whereas calibrated probabilities were used for final prevalent-CVD stratification. The analytical operating point was determined from development out-of-fold predictions and locked before temporal evaluation.

All stochastic candidate models used the common random seed of 42. Each model was embedded within the same leakage-controlled preprocessing framework to ensure a consistent comparison.

2.8 Development-stage model selection and tuning

All candidate-model comparison and hyperparameter optimization were performed exclusively within the NHANES 2007–2012 development cohort. Five-fold stratified cross-validation with shuffling and random seed 42 was used. Fold assignments were generated once and reused across candidate algorithms to ensure paired development-stage comparisons.

Preprocessing remained embedded within each machine-learning pipeline. Imputation, missingness-indicator construction, standardization, transformation, and categorical encoding were therefore estimated from each training fold and applied only to its corresponding validation fold.

Initial fixed-specification comparisons were followed by development-only hyperparameter optimization using `RandomizedSearchCV`. Mean cross-validated ROC-AUC was the predefined model-ranking and tuning criterion; average precision and Brier score were calculated as complementary measures. All 11 regularization strengths were evaluated for logistic regression. Randomized searches evaluated 16 parameter configurations for random forest, 20 for XGBoost, and 20 for LightGBM, using the same five locked development folds and random seed.

Because the same development folds contributed to hyperparameter selection and to the tuned model comparison, the tuned cross-validated values reported in Section 3.2 are interpreted as *development-stage model-selection estimates*, not as an independent post-selection estimate of generalization performance. The chronologically later 2013–2018 cohort was kept untouched throughout this process and provides the principal evaluation of the finalized framework.

2.9 Hyperparameter tuning and final model

Model-specific hyperparameter spaces were defined before temporal evaluation. Full candidate search spaces and selected configurations are reported in Supplementary Table S1.

For XGBoost, the development search space included `n_estimators` $\in \{300, 500, 700, 900\}$, `learning_rate` $\in \{0.02, 0.03, 0.05, 0.08\}$, `max_depth` $\in \{2, 3, 4, 5\}$, `min_child_weight` $\in \{1, 3, 5, 8\}$, `subsample` $\in \{0.75, 0.85, 1.0\}$, `colsample_bytree` $\in \{0.75, 0.85, 1.0\}$, `reg_alpha` $\in \{0, 0.1, 0.5, 1.0\}$, and `reg_lambda` $\in \{1, 2, 5, 10\}$.

After development-only hyperparameter optimization, XGBoost achieved the highest mean cross-validated ROC-AUC and was selected as the final base learner. The selected configuration consisted of 500 boosting iterations, a learning rate of 0.03, maximum tree depth of 2, minimum child weight of 8, subsample fraction of 0.75, column-sampling fraction of 0.85, L_1 regularization parameter of 1.0, and L_2 regularization parameter of 2.0.

Although XGBoost achieved the highest mean ROC-AUC, its performance was nearly identical to that of the tuned LightGBM model. Model selection therefore reflected the predefined primary ranking criterion rather than evidence of substantial algorithmic superiority.

The selected XGBoost architecture and hyperparameters were locked before probability calibration, analytical operating-point selection, and temporal validation. No hyperparameter was modified according to performance observed in NHANES 2013–2018.

2.10 Probability calibration

Calibration was evaluated as a distinct model property rather than assumed to improve probability accuracy. Three probability outputs were examined for the selected XGBoost model:

- uncalibrated XGBoost probabilities;
- sigmoid/Platt-calibrated probabilities; and
- isotonic-calibrated probabilities.

The sigmoid mapping had been designated during development as the primary calibrated specification and was locked before temporal evaluation. The uncalibrated output and isotonic calibration were retained as sensitivity analyses. This designation was not changed after inspection of the 2013–2018 cohort, because doing so would have used the temporal-validation data for retrospective model selection.

Calibration was fitted using development data only. The implementation used five-fold outer development out-of-fold prediction with five-fold inner calibration so that a calibration mapping was not fitted and evaluated on the same observations. Calibration performance was assessed using the Brier score, log loss, calibration intercept, calibration slope, and reliability curves. The calibration analysis was therefore intended to characterize probability behavior, not to presume that a calibration layer necessarily improved every proper scoring rule.

2.11 Decision-threshold selection

A single development-derived analytical operating point was selected from sigmoid-calibrated development out-of-fold probabilities using Youden’s J statistic:

$$J = \text{Sensitivity} + \text{Specificity} - 1.$$

The value maximizing J was $p_t = 0.076028$. This cutoff was treated strictly as an analytical operating point for reproducible threshold-dependent evaluation; it was not treated as a clinically established diagnostic, treatment, or intervention threshold.

The operating point was locked before temporal evaluation and applied unchanged to NHANES 2013–2018. Neither probability recalibration nor operating-point reoptimization was permitted using temporal-validation data. The corresponding development operating characteristics are reported in Section 3.3.

2.12 Primary temporal validation

After all modeling decisions had been fixed, the complete pipeline was applied to NHANES 2013–2018.

The model was not refitted or recalibrated, and the operational classification threshold was not reoptimized.

Primary probability-level performance measures were ROC-AUC, average precision, Brier score, log loss, calibration intercept, and calibration slope.

Threshold-dependent metrics included sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, and accuracy.

2.13 Confidence intervals

Uncertainty in primary temporal-validation metrics was quantified using participant-level non-parametric bootstrap resampling with replacement.

A total of 1,000 bootstrap replicates were completed. The fitted model, calibration transformation, and operational threshold remained fixed within each replicate. No model refitting, recalibration, or threshold reoptimization was performed.

Ninety-five percent percentile bootstrap confidence intervals were calculated.

This procedure estimated uncertainty in performance of the fixed developed model and did not incorporate variability associated with repeating the complete model-development process.

2.14 Exploratory decision-curve analysis

Decision-curve analysis was used as a secondary, exploratory assessment of net benefit across threshold probabilities from 0.01 to 0.50. For a threshold probability p_t , net benefit was calculated as

$$NB = \frac{TP}{N} - \frac{FP}{N} \left(\frac{p_t}{1 - p_t} \right).$$

The model was compared with the reference strategies of assessing all participants and assessing none. In this analysis, “assessment” is a generic model-guided action rather than a tested clinical intervention; no specific downstream diagnostic test, treatment, or workflow was evaluated. The development-locked analytical operating point was shown without temporal optimization. Decision-curve results were therefore interpreted as mathematical decision-analytic behavior, not evidence of clinical effectiveness.

2.15 Subgroup robustness

The unchanged primary model was evaluated in prespecified temporal-validation subgroups defined by age, sex, and race/ethnicity.

Age categories were 20–44 years, 45–64 years, and ≥ 65 years. Sex and race/ethnicity categories were evaluated according to the harmonized NHANES categories used in the primary analytical dataset.

Subgroup-specific model refitting, probability recalibration, and threshold optimization were not performed. The same fitted XGBoost model, sigmoid calibration mapping, and development-locked analytical operating point of 0.076028 were applied unchanged within every subgroup.

Subgroup performance was characterized using ROC-AUC, average precision, Brier score, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Subgroup sample size, number of observed CVD cases, and unweighted CVD prevalence were also reported to aid interpretation of performance differences.

To avoid reporting highly unstable estimates from very small groups, subgroup results were reported only when the subgroup contained at least 100 participants, 20 CVD cases, and 20 non-cases.

Uncertainty was quantified using participant-level nonparametric bootstrap resampling with replacement. A total of 500 bootstrap replicates were performed separately within each eligible subgroup, and 95% percentile bootstrap confidence intervals were calculated for the subgroup performance metrics. The fitted model, calibration mapping, and analytical operating point remained fixed within all bootstrap replicates; no model refitting, recalibration, or subgroup-specific threshold selection was performed.

These analyses were intended to characterize heterogeneity in discrimination, probability accuracy, and threshold-dependent operating characteristics rather than to constitute a comprehensive algorithmic fairness assessment. Detailed subgroup results are reported in Supplementary Table S4.

2.16 Predictor-domain ablation

A secondary analysis evaluated the incremental predictive information provided by nested predictor domains:

- **A: Core clinical-demographic model**—age, sex, BMI, waist circumference, systolic blood pressure, diastolic blood pressure, total cholesterol, HDL cholesterol, HbA1c, and diabetes;
- **B: Core plus social-context variables**—Block A plus race/ethnicity and poverty-to-income ratio; and
- **C: Full model plus lifestyle/exposure information**—Block B plus serum cotinine, vigorous activity, and moderate activity.

The selected XGBoost hyperparameters were held fixed. Reduced models were fitted using development data and evaluated in the unchanged temporal-validation cohort. Paired bootstrap analysis using 1,000 replicates quantified performance differences relative to the full model. These analyses were interpreted as predictive incremental-value analyses and not as evidence of causal effects of any predictor domain.

2.17 Explainability analysis

SHAP was used to characterize the fitted XGBoost base learner. Because the sigmoid calibration layer transforms the base-model probability, SHAP values were calculated for the underlying uncalibrated XGBoost estimator, whereas final participant probabilities used for stratification came from the locked sigmoid-calibrated framework.

TreeExplainer was applied to a fixed simple random sample of 2,000 participants from the temporal-validation cohort. Sampling was independent of outcome status, predicted probability, and SHAP values.

SHAP values were first obtained in the transformed model-feature space, which contained one-hot encoded levels and missingness indicators. For original-predictor importance ranking and temporal-stability analysis, the mean absolute contributions of all derived features belonging

to the same original predictor were aggregated to their parent variable. The main beeswarm figure displays transformed model-feature contributions because this preserves the direction and distribution of individual one-hot levels; original-predictor rankings are reported separately. Mean absolute SHAP value was used to rank global importance. SHAP results were interpreted as explanations of the fitted model, not as causal effects.

2.18 Local explainability

Participant-level SHAP explanations were generated to complement the global feature-importance analysis and to illustrate how individual predictor contributions combined for specific temporal-validation participants.

Three representative participants were selected using predefined mathematical rules that were independent of the resulting SHAP values.

The first example represented the decision boundary. This participant was defined as the temporal-validation participant whose sigmoid-calibrated predicted probability was closest in absolute distance to the development-locked analytical operating point of 0.076028.

The second example represented a typical participant with observed CVD. Among temporal-validation participants with CVD, the participant whose calibrated predicted probability was closest to the median predicted probability of all temporal CVD cases was selected.

The third example represented a typical participant without observed CVD. Among temporal-validation non-cases, the participant whose calibrated predicted probability was closest to the median predicted probability of all temporal non-cases was selected.

Participant selection was therefore based exclusively on predefined probability-based criteria and was not influenced by the appearance, magnitude, or direction of the resulting SHAP explanations.

For each selected participant, SHAP values were calculated from the underlying uncalibrated XGBoost estimator, consistent with the global explainability analysis. These local explanations illustrate the contribution of individual predictors to the fitted model output and should not be interpreted as causal explanations.

The corresponding participant-level SHAP visualizations are provided in Supplementary Figures S5–S7.

2.19 Temporal explanation stability

SHAP importance was calculated separately within the 2013–2014, 2015–2016, and 2017–2018 temporal-validation cycles.

Original-predictor importance rankings were compared across cycles using Spearman rank correlation.

This analysis assessed whether the relative explanation structure remained stable over time without modifying the primary model.

2.20 Laboratory harmonization

Laboratory variables were harmonized using released NHANES variables and common units across the 2007–2018 analytical period.

Official cycle documentation was reviewed for changes in assays, equipment, and laboratory sites. Documented transitions included HbA1c analytic considerations during 2007–2010, lipid-analyzer changes during 2013–2014, and serum-cotinine method/equipment changes during 2013–2014.

No post-hoc bridging equation was fitted after examination of temporal-validation performance. These transitions were treated as potential sources of temporal measurement heterogeneity.

2.21 Software and reproducibility

The final analysis was executed using Python 3.13.5.

Major package versions included pandas 2.3.3, NumPy 2.2.6, SciPy 1.16.3, scikit-learn 1.7.2, XGBoost 3.4.1, LightGBM 4.7.0, SHAP 0.52.0, and joblib 1.5.2.

The analytical workflow generated cohort-membership hashes, model artifacts, source-file check-sums, environment-version records, and a results-artifact manifest to support computational traceability.

3 Results

3.1 Cohort characteristics

Among 59,842 participants in the merged 2007–2018 source, 34,770 were adults aged ≥ 20 years. After requiring sufficient information to define the composite CVD outcome, 34,602 participants remained in the primary analytical cohort.

Chronological partitioning resulted in 17,626 development participants and 16,976 temporal-validation participants.

The development cohort contained 1,966 CVD cases and 15,660 non-cases, corresponding to an unweighted prevalence of 11.15%. The temporal-validation cohort contained 1,921 CVD cases and 15,055 non-cases, corresponding to an unweighted prevalence of 11.32%. The chronological cohort composition by NHANES cycle is summarized in Table 1.

Table 1: Chronological cohort composition by NHANES cycle.

NHANES cycle	Role	Participants	CVD cases	Unweighted prevalence
2007–2008	Development	5,902	717	12.15%
2009–2010	Development	6,184	668	10.80%
2011–2012	Development	5,540	581	10.49%
2013–2014	Temporal validation	5,752	595	10.34%

NHANES cycle	Role	Participants	CVD cases	Unweighted prevalence
2015–2016	Temporal validation	5,691	635	11.16%
2017–2018	Temporal validation	5,533	691	12.49%

Survey-weighted mean age was 47.01 years (95% CI 46.29–47.73) during development and 47.88 years (95% CI 47.29–48.46) during temporal validation.

Survey-weighted CVD prevalence was 8.4% (95% CI 7.7–9.1) during development and 9.0% (95% CI 8.3–9.7) during temporal validation.

These weighted estimates describe population-oriented cohort characteristics and should not be conflated with the unweighted participant-level prevalence used for predictive performance calculations. Survey-weighted characteristics of the development and temporal-validation cohorts are presented in Table 2.

Table 2: Survey-weighted characteristics of the development and temporal-validation cohorts.

Characteristic	Development 2007–2012	Temporal validation 2013–2018
Age, years	47.01 (46.29–47.73)	47.88 (47.29–48.46)
Male	48.1% (47.4–48.8)	48.1% (47.2–48.9)
Female	51.9% (51.2–52.6)	51.9% (51.1–52.8)
Poverty-to-income ratio	2.96 (2.87–3.06)	2.99 (2.89–3.10)
BMI, kg/m ²	28.67 (28.47–28.87)	29.46 (29.18–29.73)
Waist circumference, cm	98.25 (97.68–98.82)	100.23 (99.53–100.94)
Systolic BP, mmHg	121.02 (120.42–121.63)	122.92 (122.40–123.43)
Diastolic BP, mmHg	70.33 (69.68–70.98)	70.85 (70.28–71.42)
Total cholesterol, mg/dL	196.32 (195.27–197.36)	190.36 (188.72–192.00)
HDL cholesterol, mg/dL	52.70 (52.14–53.27)	54.16 (53.50–54.81)
HbA1c, %	5.6 (5.6–5.6)	5.7 (5.6–5.7)
Diabetes, yes	8.9% (8.2–9.6)	11.2% (10.4–11.9)
Serum cotinine, ng/mL, median [IQR]	0.04 [0.01–9.80]	0.03 [0.01–4.15]
Vigorous recreational activity, yes	24.2% (21.9–26.5)	27.2% (25.5–28.9)
Moderate recreational activity, yes	44.7% (42.6–46.8)	45.9% (44.0–47.9)
Composite CVD, yes	8.4% (7.7–9.1)	9.0% (8.3–9.7)

Several characteristics shifted between periods. Survey-weighted mean BMI increased from 28.67 to 29.46 kg/m², waist circumference from 98.25 to 100.23 cm, and systolic blood pressure from 121.02 to 122.92 mmHg.

Weighted diabetes prevalence increased from 8.9% to 11.2%, whereas total cholesterol decreased from approximately 196.3 to 190.4 mg/dL.

3.2 Candidate-model discrimination

All four candidate algorithms demonstrated useful discrimination in the development data. During the initial fixed-model comparison, XGBoost achieved the highest out-of-fold ROC-AUC at 0.8412, followed by random forest at 0.8385, logistic regression at 0.8372, and LightGBM at 0.8337.

Following development-only hyperparameter optimization, discrimination improved modestly across the leading models. XGBoost and LightGBM achieved nearly identical mean cross-validated ROC-AUC values, whereas random forest and logistic regression showed slightly lower discrimination. The tuned development-stage model comparison is summarized in Table 3.

Table 3: Development-only tuned model comparison.

Model	Mean ROC-AUC	SD	Mean PR-AUC	SD	Mean Brier
XGBoost	0.8442	0.0097	0.4055	0.0229	0.0801
LightGBM	0.8440	0.0097	0.4063	0.0237	0.0801
Random forest	0.8399	0.0097	0.4014	0.0203	0.0810
Logistic regression	0.8375	0.0071	0.3895	0.0166	0.0813

XGBoost achieved a mean cross-validated ROC-AUC of 0.84420 compared with 0.84399 for LightGBM, corresponding to an absolute difference of approximately 0.00022. LightGBM achieved a marginally greater mean PR-AUC and a numerically similar Brier score.

XGBoost was selected because mean ROC-AUC was the primary development-stage model-ranking criterion established before temporal evaluation. The extremely small difference between XGBoost and LightGBM indicates that the selected model should not be interpreted as demonstrating substantial algorithmic superiority. Rather, several candidate algorithms achieved broadly comparable predictive performance within the development cohort.

Development out-of-fold ROC and precision–recall curves for the fixed candidate models are provided in Supplementary Figures S1 and S2, respectively.

3.3 Development calibration and locked threshold

Development out-of-fold probability performance was examined for the uncalibrated XGBoost model and the two calibration approaches. The uncalibrated model achieved ROC-AUC 0.8443, average precision 0.4015, Brier score 0.0801, calibration intercept approximately 0.002, and calibration slope approximately 0.999. Sigmoid calibration yielded ROC-AUC 0.8438, average precision 0.4017, Brier score 0.0824, calibration intercept 0.025, and calibration slope 1.012. Isotonic calibration yielded ROC-AUC 0.8433, average precision 0.3923, Brier score 0.0803, calibration intercept -0.028 , and calibration slope 0.980.

Thus, sigmoid calibration did not improve development Brier score relative to the uncalibrated probabilities. It was nevertheless preserved as the development-designated primary calibrated specification because changing the probability specification after inspection of the later cohort would compromise temporal-validation independence. The alternative probability outputs are

therefore reported as sensitivity analyses rather than used for retrospective model replacement. Development calibration curves are provided in Supplementary Figure S3.

Using sigmoid-calibrated development out-of-fold probabilities, the Youden-derived analytical operating point was $p_t = 0.076028$. At this point, development sensitivity was 0.8159, specificity 0.7285, PPV 0.2740, NPV 0.9692, F1 score 0.4102, and accuracy 0.7383. The operating point was then locked and applied unchanged to temporal validation.

3.4 Temporal-validation discrimination

In NHANES 2013–2018, the frozen sigmoid-calibrated XGBoost achieved a ROC-AUC of 0.8474 (95% bootstrap CI 0.8395–0.8558). The temporal-validation ROC curve is shown in Figure 2.

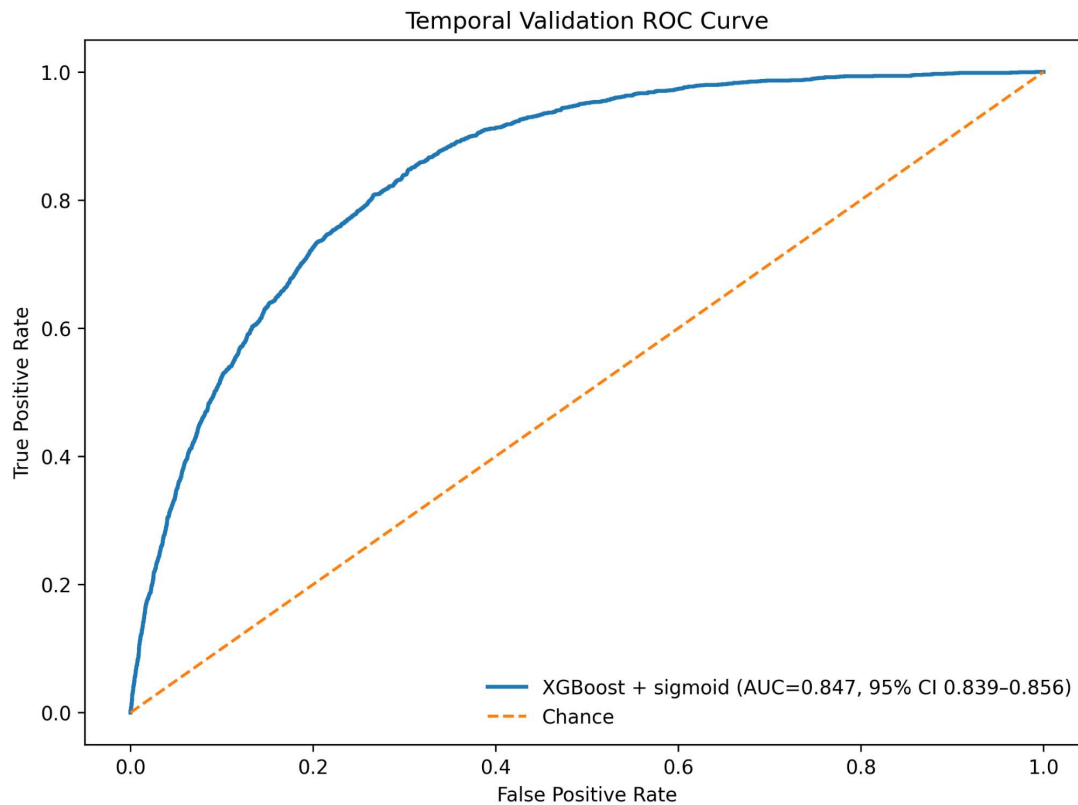


Figure 2: Temporal-validation ROC curve. The locked sigmoid-calibrated XGBoost achieved ROC-AUC 0.847 (95% CI 0.839–0.856) when transported from NHANES 2007–2012 to the later NHANES 2013–2018 cohort.

3.5 Temporal precision–recall performance

Temporal average precision was 0.4034 (95% CI 0.3817–0.4271), compared with an unweighted prevalence-based no-skill reference of approximately 0.113. The corresponding precision–recall curve is shown in Figure 3.

The model therefore achieved substantially greater precision–recall performance than the prevalence-based reference.

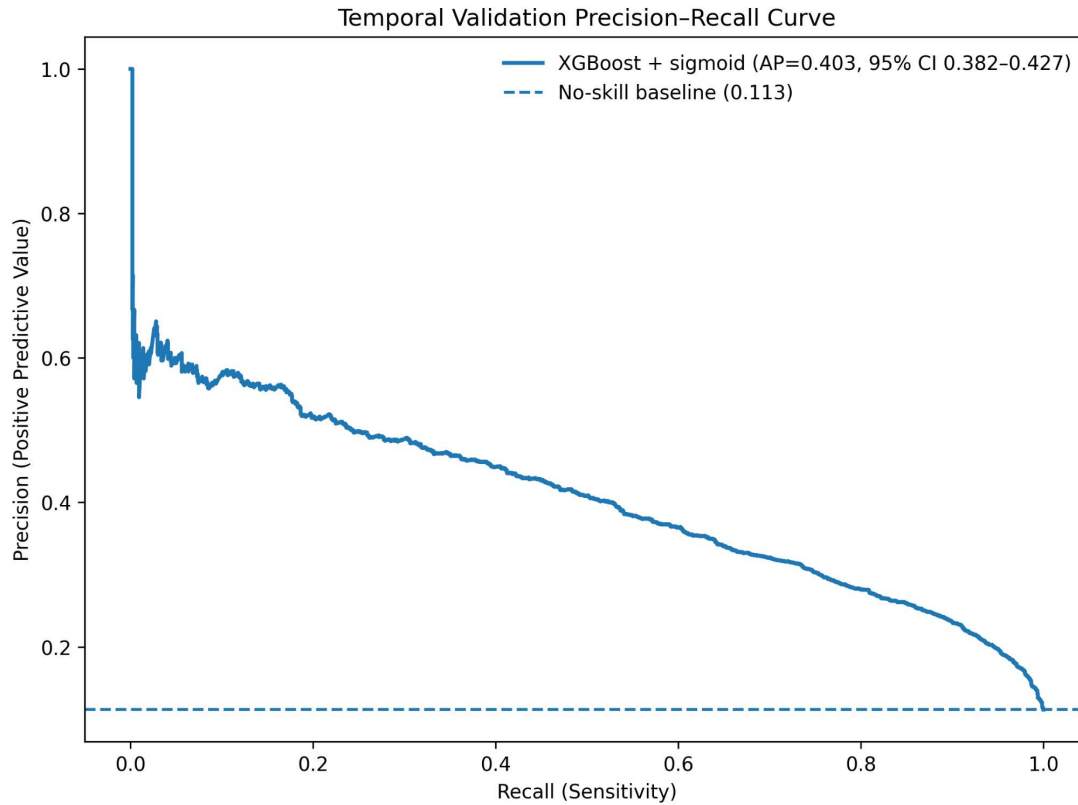


Figure 3: Temporal-validation precision–recall curve. Average precision was 0.403 (95% CI 0.382–0.427), compared with a no-skill prevalence baseline of approximately 0.113.

3.6 Temporal probability accuracy and calibration

The sigmoid-calibrated temporal predictions had a mean predicted probability of 0.1159, compared with an observed unweighted CVD prevalence of 0.1132. The temporal Brier score was 0.0831 (95% CI 0.0799–0.0864), and log loss was 0.2802 (95% CI 0.2712–0.2893). The calibration intercept was -0.0669 (95% CI -0.1641 to 0.0379), and the calibration slope was 0.9813 (95% CI 0.9429–1.0223).

Ideal aggregate calibration corresponds to an intercept of 0 and a slope of 1. The bootstrap confidence intervals included these reference values, supporting limited aggregate miscalibration, although these summary parameters do not by themselves establish perfect calibration. The temporal reliability curves are shown in Figure 4.

Sigmoid calibration did not improve temporal Brier score. The uncalibrated XGBoost achieved a Brier score of 0.08085 and the isotonic sensitivity specification 0.08110, compared with 0.08315 for the primary sigmoid-calibrated specification. Consequently, the sigmoid layer should not be interpreted as improving overall probability accuracy. It was preserved only because it had been designated and locked during development; replacing it after observing temporal performance would constitute retrospective model selection on the evaluation cohort.

3.7 Performance at the locked analytical operating point

At the development-derived analytical operating point of 0.076028, the temporal-validation confusion matrix contained 10,806 true negatives, 4,249 false positives, 346 false negatives, and

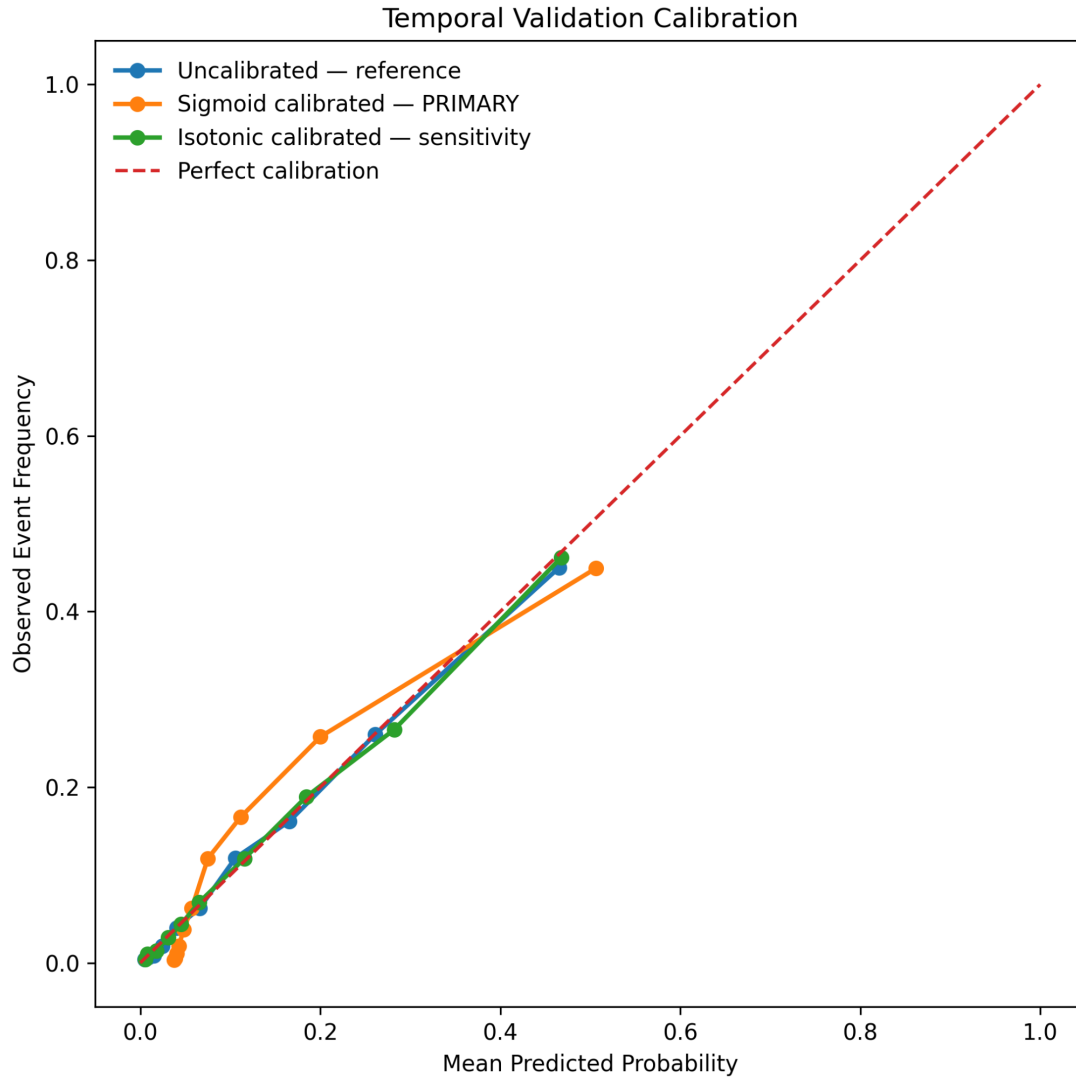


Figure 4: Temporal-validation reliability curves for the uncalibrated XGBoost model, the primary sigmoid-calibrated specification, and the isotonic-calibrated sensitivity analysis. The diagonal line represents perfect calibration. The figure should be interpreted together with proper scoring rules and calibration intercept/slope rather than as a stand-alone measure of probability quality.

1,575 true positives.

Primary temporal-validation performance of the locked sigmoid-calibrated XGBoost model is summarized in Table 4.

Table 4: Primary temporal-validation performance of the locked sigmoid-calibrated XGBoost model.

Metric	Point estimate	95% CI
ROC-AUC	0.847	0.839–0.856
Average precision	0.403	0.382–0.427
Brier score	0.083	0.080–0.086
Log loss	0.280	0.271–0.289
Calibration intercept	−0.067	−0.164–0.038
Calibration slope	0.981	0.943–1.022
Sensitivity	0.820	0.803–0.836
Specificity	0.718	0.711–0.726
Positive predictive value	0.270	0.259–0.282
Negative predictive value	0.969	0.966–0.972
F1 score	0.407	0.392–0.420
Accuracy	0.729	0.723–0.737

At the locked analytical operating point, approximately 82% of participants with observed CVD were classified positive, while approximately 72% of participants without observed CVD were classified negative. Positive predictive value was approximately 27%, whereas negative predictive value was approximately 97%. These operating characteristics describe prevalent-CVD status stratification and should not be interpreted as the performance of a confirmatory diagnostic test.

To characterize the dependence of classification performance on the chosen probability cutoff, a secondary threshold-sensitivity analysis was performed using several fixed probability thresholds. These alternative thresholds were evaluated only to illustrate changes in operating characteristics and were not used to replace or reoptimize the development-derived primary operating point. The threshold-sensitivity analysis is summarized in Table 5.

Table 5: Threshold-sensitivity analysis in temporal validation.

Threshold	Sensitivity	Specificity	PPV	NPV	F1	Accuracy	Predicted positive (%)
0.050	0.940	0.531	0.204	0.986	0.335	0.578	52.19
0.076028	0.820	0.718	0.270	0.969	0.407	0.729	34.31
0.100	0.741	0.788	0.308	0.960	0.436	0.783	27.20
0.150	0.602	0.867	0.365	0.945	0.455	0.837	18.65
0.500	0.203	0.976	0.517	0.906	0.291	0.888	4.45

The sensitivity analysis demonstrated the expected trade-off between case detection and false-positive classifications. Lower thresholds increased sensitivity and negative predictive value but classified a larger proportion of participants as positive. For example, lowering the cutoff to

0.05 increased sensitivity to 0.940 but reduced specificity to 0.531, with approximately 52.2% of temporal participants classified positive.

Conversely, increasing the threshold progressively improved specificity and PPV while reducing sensitivity. At a threshold of 0.15, sensitivity decreased to 0.602 and specificity increased to 0.867. At the conventional probability cutoff of 0.50, specificity reached 0.976 and PPV 0.517, but sensitivity fell to 0.203, meaning that approximately four-fifths of participants with observed CVD were not classified positive.

These comparisons demonstrate that the 0.076028 value should be interpreted as a development-derived analytical operating point rather than as an intrinsically optimal or clinically established cutoff. The alternative temporal thresholds were descriptive sensitivity analyses only; none was used to retrospectively modify the locked primary analysis.

3.8 Exploratory decision-curve analysis

The locked model maintained positive net benefit relative to assessing none and exceeded the assess-all strategy across an estimated threshold interval of approximately 0.040–0.500. At the development-locked threshold of 0.076028, net benefit was 0.07218 for the model, 0.04019 for assess-all, and 0 for assess-none. The corresponding net-benefit difference was approximately 0.0320 versus assess-all and 0.0722 versus assess-none. The temporal decision-curve analysis is shown in Figure 5.

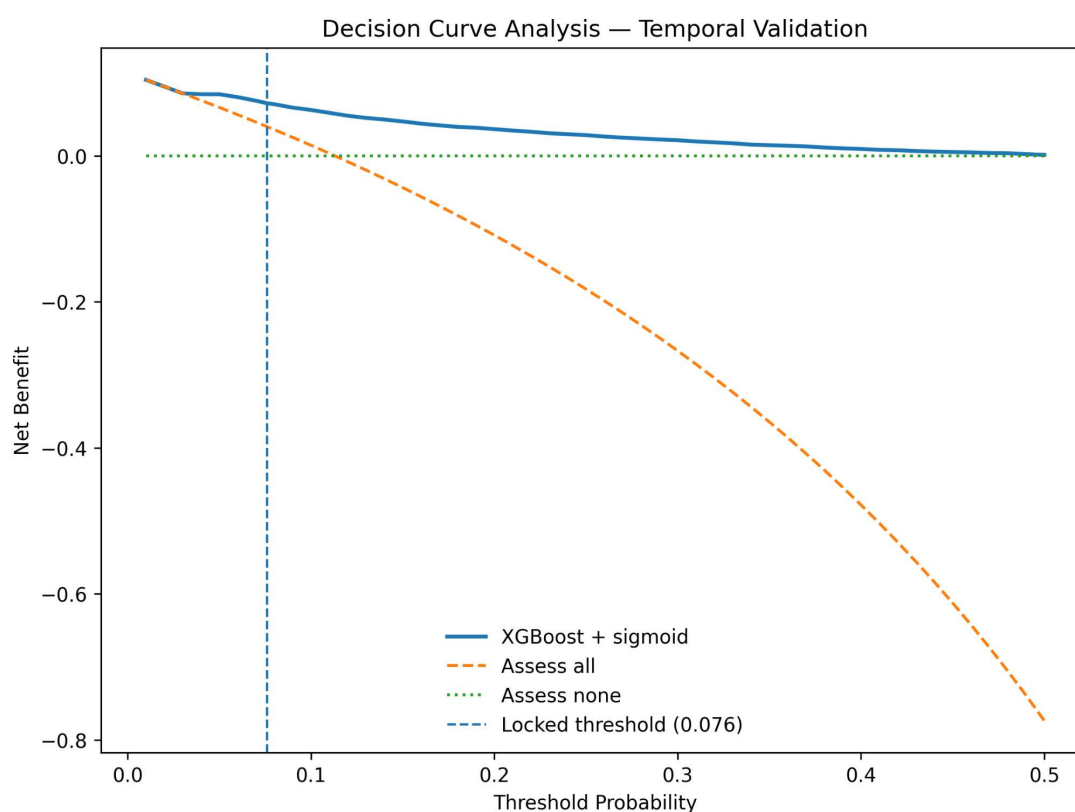


Figure 5: Exploratory decision-curve analysis during temporal validation. The XGBoost-plus-sigmoid specification is compared with assess-all and assess-none strategies. The vertical dashed line represents the locked development-derived analytical operating point of approximately 0.076. No clinical intervention was evaluated; net benefit therefore reflects the mathematical decision assumptions of the analysis.

These results indicate favorable decision-analytic behavior under the evaluated assumptions, but they do not establish clinical effectiveness or patient benefit.

3.9 Stability across temporal cycles

The frozen model was evaluated separately in each of the three temporal-validation cycles. Cycle-specific temporal performance is summarized in Table 6.

Table 6: Cycle-specific temporal performance.

Validation cycle	N	Cases	ROC-AUC	AP	Brier	Cal. slope	Sensitivity	Specificity
2013–2014	5,752	595	0.849	0.405	0.076	1.004	0.800	0.731
2015–2016	5,691	635	0.856	0.422	0.080	1.034	0.819	0.733
2017–2018	5,533	691	0.835	0.388	0.094	0.913	0.838	0.688

Performance was similar in 2013–2014 and 2015–2016.

Some attenuation was observed in 2017–2018, when ROC-AUC decreased to 0.835, Brier score increased to approximately 0.094, and specificity decreased to 0.688.

The 2017–2018 cycle also had the highest unweighted CVD prevalence, at 12.49%.

3.10 Predictor-domain ablation

The core clinical-demographic model achieved temporal ROC-AUC 0.8348 and average precision 0.3849. Adding the social-context variables (race/ethnicity and poverty-to-income ratio) increased ROC-AUC to 0.8445 and average precision to 0.4016. Adding the lifestyle/exposure variables produced the full-model ROC-AUC of 0.8474 and average precision of 0.4034. Incremental predictor-domain performance is summarized in Table 7.

Table 7: Incremental predictor-domain value in temporal validation.

Predictor block	No. predictors	Temporal ROC-AUC	Temporal AP	Brier
Core clinical + demographic	10	0.8348	0.3849	0.08444
+ social context	12	0.8445	0.4016	0.08329
+ lifestyle/exposure	15	0.8474	0.4034	0.08315

Relative to the full model, the core clinical-demographic model had a ROC-AUC difference of -0.01260 (paired-bootstrap 95% CI -0.01542 to -0.00973), an average-precision difference of -0.01855 (95% CI -0.02541 to -0.01098), and a Brier-score difference of 0.00130 (95%

CI 0.00065–0.00191). The social-context model differed from the full model by -0.00288 in ROC-AUC (95% CI -0.00452 to -0.00117); its average-precision difference of -0.00187 had a confidence interval crossing zero. Thus, most of the observed incremental discrimination arose after addition of the social-context variables, whereas the further gain from lifestyle/exposure variables was comparatively small.

3.11 Global explainability

After aggregation to the 15 original predictors, age had the largest mean absolute SHAP importance. The next most influential original predictors were total cholesterol, poverty-to-income ratio, waist circumference, serum cotinine, diabetes, HDL cholesterol, and systolic blood pressure.

Figure 6 displays SHAP contributions in the transformed model-feature space. Consequently, categorical levels and missingness indicators may appear separately in the beeswarm even though they were aggregated to their parent variables for the original-predictor ranking.

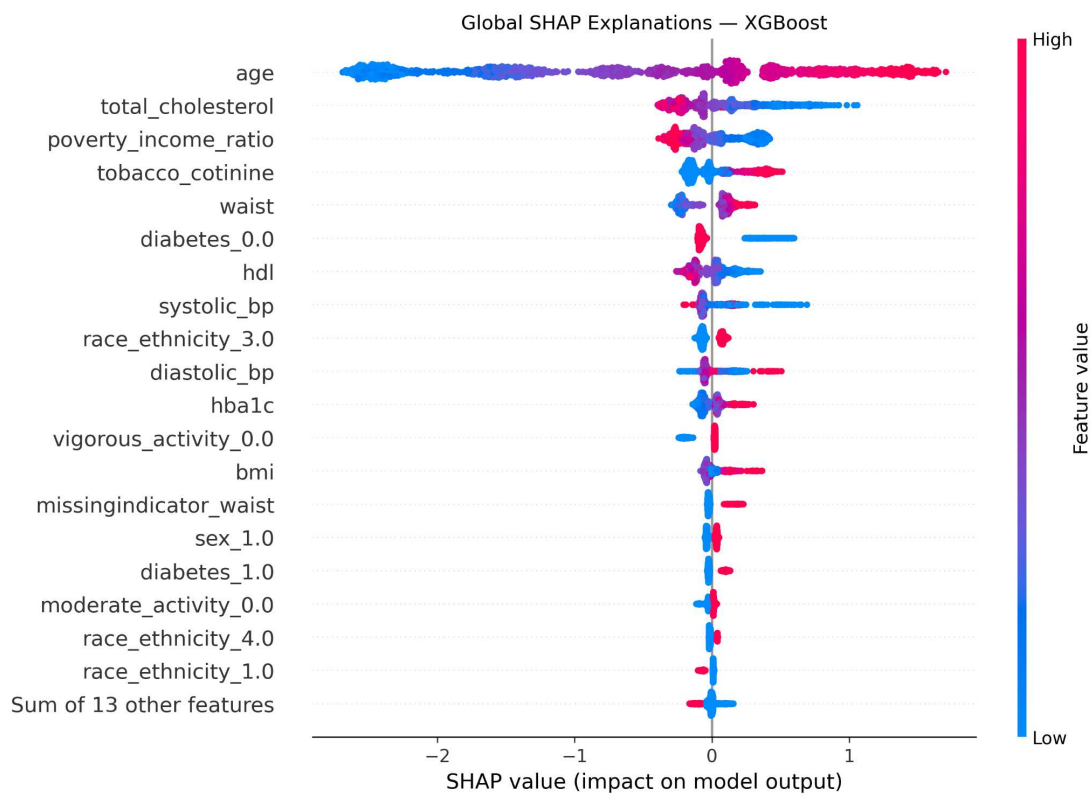


Figure 6: Global SHAP beeswarm for the final XGBoost base learner in the transformed model-feature space. Each point represents a participant-feature contribution to the underlying model output. Positive SHAP values increase the model output and negative values decrease it; color represents the relative feature value. One-hot encoded levels and missingness indicators may appear separately in this display. For original-predictor ranking, derived features were aggregated to their parent predictors. SHAP values explain fitted-model behavior and do not represent causal effects.

Higher age values generally shifted the underlying model output toward the prevalent-CVD class, whereas younger values shifted it in the opposite direction. Other predictors showed more heterogeneous patterns. The dominance of age should be interpreted together with the subgroup results because the unchanged global operating point produced markedly different sensitivity and specificity across age groups.

Exact original-predictor SHAP importance values are reported in Supplementary Table S3, and the corresponding aggregated mean absolute SHAP bar plot is shown in Supplementary Figure S4.

3.12 Temporal explanation stability

Original-predictor importance rankings were highly stable across the three temporal-validation cycles. Pairwise Spearman correlations were 0.982 for 2013–2014 versus 2015–2016, 0.986 for 2013–2014 versus 2017–2018, and 0.996 for 2015–2016 versus 2017–2018. Age ranked first in all three cycles, total cholesterol second, poverty-to-income ratio third, and waist circumference fourth. Most predictors changed by no more than one rank position; vigorous activity showed the largest observed range of two positions.

3.13 Subgroup performance

Subgroup performance was evaluated using the unchanged sigmoid-calibrated XGBoost model and the development-derived analytical operating point of 0.076028. Detailed subgroup sample sizes, prevalence estimates, discrimination, probability accuracy, and threshold-dependent operating characteristics are reported in Supplementary Table S4.

Performance varied substantially across age groups. Among participants aged 20–44 years ($n = 7,010$; 130 CVD cases; prevalence 1.85%), ROC-AUC was 0.802 (95% CI 0.758–0.839), whereas sensitivity at the common analytical operating point was only 0.162 (95% CI 0.098–0.227). Specificity was 0.991 (95% CI 0.988–0.993), PPV was 0.250, and NPV was 0.984.

Among participants aged 45–64 years ($n = 5,825$; 616 CVD cases; prevalence 10.58%), ROC-AUC was 0.754 (95% CI 0.735–0.773). Sensitivity was 0.683 (95% CI 0.646–0.719), specificity was 0.686 (95% CI 0.672–0.699), PPV was 0.204, and NPV was 0.948.

Among adults aged ≥ 65 years ($n = 4,141$; 1,175 CVD cases; prevalence 28.37%), ROC-AUC was 0.710 (95% CI 0.694–0.726). Sensitivity was 0.964 (95% CI 0.954–0.973), whereas specificity was only 0.141 (95% CI 0.129–0.154). PPV was 0.308 and NPV was 0.909.

Sex-specific discrimination was comparatively stable. Among women, ROC-AUC was 0.833 (95% CI 0.821–0.846), sensitivity was 0.768, specificity was 0.738, PPV was 0.233, and NPV was 0.968. Among men, ROC-AUC was 0.855 (95% CI 0.844–0.865), sensitivity was 0.859, specificity was 0.695, PPV was 0.303, and NPV was 0.970.

Race/ethnicity-specific ROC-AUC ranged from 0.811 (95% CI 0.790–0.829) among non-Hispanic Black participants to 0.868 (95% CI 0.842–0.892) among participants categorized as other race/multiracial. Full threshold-dependent results for all race/ethnicity categories are provided in Supplementary Table S4.

These findings indicate substantial heterogeneity in operating characteristics, particularly across age categories. The contrast between low sensitivity in younger adults and low specificity in older adults demonstrates that favorable overall temporal performance does not imply uniform behavior across population subgroups.

No subgroup-specific model refitting, probability recalibration, or analytical operating-point optimization was performed using the temporal-validation data.

4 Discussion

4.1 Principal findings

This population-based study temporally evaluated a development-locked, explainable machine-learning framework for prevalent physician-diagnosed CVD stratification. The principal temporal result was stable discrimination: the frozen XGBoost specification achieved ROC-AUC 0.847 and average precision 0.403 in NHANES 2013–2018. Aggregate calibration of the locked sigmoid specification was close to the ideal intercept and slope, but its Brier score was worse than that of the uncalibrated XGBoost. The calibration layer should therefore be viewed as a development-designated probability mapping that was preserved to protect temporal-validation independence, not as evidence of improved overall probability accuracy.

The development-derived analytical operating point yielded sensitivity of approximately 82%, specificity of approximately 72%, and NPV of approximately 97% in the pooled temporal cohort. These pooled values, however, concealed substantial age-related heterogeneity: sensitivity was low in younger adults and specificity was very low in adults aged ≥ 65 years. Predictor-domain ablation suggested modest incremental information from social-context and lifestyle/exposure variables, while SHAP-derived original-predictor rankings remained highly stable across the three temporal cycles. Taken together, the study's contribution is a multidimensional reliability assessment under chronological transport, rather than evidence that a single algorithm or operating threshold is universally superior.

4.2 Machine learning versus conventional regression

The improvement of XGBoost over logistic regression during development was modest: mean cross-validated ROC-AUC increased from approximately 0.837 to 0.844, an absolute difference of about 0.007. These tuned cross-validation values are development-stage model-selection estimates, because the same development folds contributed to hyperparameter selection, and should not be interpreted as an independent post-selection validation comparison.

The small numerical advantage is consistent with evidence that greater algorithmic flexibility does not necessarily produce large gains when a relatively small set of structured predictors already captures substantial signal [8, 9, 10, 11]. XGBoost can nevertheless represent nonlinear associations and interactions without prespecifying their functional form, while SHAP provides a complementary description of fitted-model behavior. Conversely, the close performance of logistic regression suggests that simpler models remain relevant when implementation simplicity and direct parameter interpretation are priorities.

The present target differs fundamentally from established prospective risk equations. Framingham, the ACC/AHA Pooled Cohort Equations, QRISK3, and SCORE2 estimate future cardiovascular events over defined prediction horizons [3, 4, 5, 6]. This study modeled prevalent physician-diagnosed CVD status at the NHANES examination. Consequently, direct numerical comparison of AUC or other performance measures with prospective cardiovascular risk scores would not be a like-for-like comparison. The principal contribution is the temporally evaluated reliability framework under a distinct cross-sectional target, not evidence that XGBoost outperforms established prospective scores.

4.3 Importance of chronological validation

A major strength of the study was chronological separation of model development and later evaluation.

The model was developed using NHANES 2007–2012 and subsequently applied to 2013–2018 without refitting, recalibration, or threshold reoptimization.

The later cohort also differed in several measured characteristics, including greater mean BMI and waist circumference, higher systolic blood pressure and diabetes prevalence, and lower mean total cholesterol.

Despite these differences, pooled temporal discrimination remained stable.

Cycle-specific analysis nevertheless identified some attenuation in 2017–2018, including lower ROC-AUC, higher Brier score, and lower specificity than in the two preceding validation cycles.

This pattern illustrates why pooled temporal performance alone may not fully characterize longitudinal model stability. If a related model were implemented in practice, continued performance and calibration monitoring would be necessary.

4.4 Calibration as a core model property

Calibration was evaluated separately from discrimination because a high ROC-AUC does not guarantee reliable predicted probabilities [15]. In temporal validation, the locked sigmoid specification had calibration intercept -0.067 , calibration slope 0.981 , and mean predicted probability 11.59% , close to the observed unweighted CVD prevalence of 11.32% . These findings indicate limited aggregate miscalibration.

The proper-scoring-rule comparison was less favorable. The uncalibrated XGBoost had a Brier score of 0.08085 , compared with 0.08315 for the sigmoid specification and 0.08110 for the isotonic sensitivity specification. The sigmoid layer therefore did not improve overall probability accuracy. Because it had been designated during development and locked before temporal evaluation, it was not replaced after inspection of the later cohort. Reporting the alternatives as sensitivity analyses preserves the independence of the temporal assessment and illustrates why calibration should be judged using multiple complementary measures rather than a single calibration summary.

4.5 Interpretation of the analytical operating point

The development-derived probability cutoff of approximately 0.076 was used as an analytical operating point rather than as a clinically established diagnostic, treatment, or intervention threshold.

This value was selected from development out-of-fold predictions using Youden's J statistic and was locked before temporal validation. Its purpose was to provide a reproducible operating point for evaluating threshold-dependent model behavior, not to define a universally optimal clinical cutoff.

The analytical operating point of approximately 0.076 was substantially lower than the commonly used generic probability cutoff of 0.50 . However, a threshold of 0.50 is not intrinsically optimal for a binary probability model. Threshold-dependent operating characteristics vary according to

outcome prevalence and the relative consequences assigned to false-positive and false-negative classifications.

At the 0.50 cutoff, temporal sensitivity was approximately 20%, whereas sensitivity at the development-derived analytical operating point was approximately 82%. The lower operating point therefore produced a markedly different classification profile, with specificity of approximately 72%, negative predictive value of approximately 97%, and positive predictive value of approximately 27%.

The threshold-sensitivity analysis further demonstrated that these operating characteristics changed substantially across alternative probability cutoffs. Lower thresholds increased sensitivity and negative predictive value at the cost of reduced specificity and a larger proportion of participants being classified positive, whereas higher thresholds produced the opposite pattern.

Accordingly, the value of 0.076028 should not be interpreted as a clinically validated intervention threshold, a diagnostic cutoff, or an intrinsically optimal probability value. It should instead be viewed as a development-derived analytical operating point used to characterize the behavior of the prevalent-CVD status-stratification model under a prespecified decision rule.

These threshold-dependent results therefore describe model operating characteristics rather than evidence that the model can independently diagnose cardiovascular disease or determine clinical management. Clinical use would require independent validation, purpose-specific threshold development, and prospective evaluation of the consequences of acting on model classifications.

4.6 Exploratory decision-analytic evaluation

Decision-curve analysis complemented conventional metrics by combining true-positive and false-positive classifications under explicit threshold-dependent trade-offs [24]. At the locked analytical operating point, the model had greater calculated net benefit than assess-all and assess-none. However, no specific clinical intervention, diagnostic pathway, workflow burden, cost, or patient outcome was evaluated. The decision-curve findings should therefore be treated as exploratory mathematical evidence of potential decision-analytic value rather than proof of clinical utility.

4.7 Incremental value of social-context and lifestyle data

Predictor-domain ablation demonstrated that the ten-variable core clinical-demographic model achieved a temporal ROC-AUC of 0.835. Adding social-context variables increased ROC-AUC to 0.844, and adding lifestyle/exposure information increased it further to 0.847.

The paired bootstrap comparison showed that excluding both social-context and lifestyle/exposure information reduced ROC-AUC by approximately 0.0126 relative to the full model. Poverty-to-income ratio was also among the most influential predictors in the global SHAP analysis.

These findings are consistent with the broader cardiovascular literature showing that socioeconomic circumstances, adiposity, tobacco exposure, and physical activity are associated with cardiovascular health and disease burden [29, 2, 30, 31, 32]. Socioeconomic disadvantage may capture differences in access to care, material resources, health behaviors, environmental exposures, and other contextual determinants relevant to observed cardiovascular disease status [29].

Likewise, central adiposity and obesity have established associations with cardiovascular risk [30], while tobacco exposure represents a major cardiovascular determinant. In the present analysis,

serum cotinine provided an objective biomarker of tobacco/nicotine exposure rather than relying exclusively on self-reported smoking status [31]. Physical activity is also inversely associated with cardiovascular disease risk, providing biological and epidemiological context for inclusion of the recreational-activity variables [32].

The present results therefore suggest that socioeconomic and lifestyle/exposure measurements contain predictive information beyond the selected clinical-demographic variables. However, the magnitude of the incremental gain should not be overstated. Most of the observed improvement occurred after addition of the social-context variables, whereas the further increase associated with lifestyle/exposure variables was comparatively small.

Their use nevertheless raises important questions concerning interpretation and implementation. Variables such as race/ethnicity and socioeconomic status may reflect structural, access-related, and contextual processes rather than isolated biological mechanisms. Consequently, predictive importance does not establish causal influence and does not automatically justify inclusion of such variables in every clinical or population-health decision system.

These predictors should therefore be interpreted primarily as components of the fitted stratification model rather than as causal determinants identified by the present analysis. Any future implementation would require careful consideration of subgroup performance, fairness, potential differential consequences, and whether the incremental predictive information justifies use of socially sensitive variables.

4.8 Interpretation and stability of SHAP explanations

Age was the dominant global predictor, followed by total cholesterol, poverty-to-income ratio, waist circumference, serum cotinine, diabetes, HDL cholesterol, and systolic blood pressure.

These predictors represent several clinically and socially relevant domains, including aging, lipid status, socioeconomic circumstances, central adiposity, tobacco exposure, glycemic disease, and blood pressure. The use of explainable artificial intelligence in cardiovascular prediction has increasingly emphasized the importance of examining how complex machine-learning models use such interacting clinical and contextual information rather than relying on discrimination alone [23].

SHAP values provide participant-level attributions describing how predictor values contribute to the output of the fitted model [22]. They should therefore be interpreted as explanations of model behavior rather than estimates of biological or causal effects. Their direction and magnitude may also depend on nonlinear relationships, interactions, correlations among predictors, and the distribution of other variables in the fitted model. Similar caution is warranted in cardiovascular explainability studies, where feature importance may improve transparency without establishing causal mechanisms [23].

The temporal explanation-stability analysis provided an additional model-auditing dimension. Spearman rank correlations greater than 0.98 across all temporal-cycle pairs indicated that the hierarchy of original-predictor importance changed little between 2013 and 2018. Age, total cholesterol, poverty-to-income ratio, and waist circumference retained the four leading positions across all three cycles.

This stability suggests that the broad hierarchy of predictor use by the fitted model remained consistent during temporal transport. However, explanation stability and predictive stability are conceptually distinct. A stable feature-importance hierarchy does not guarantee that discrimination,

probability accuracy, calibration, or threshold-dependent operating characteristics will remain unchanged. This distinction was illustrated by the modest attenuation in predictive performance observed during the 2017–2018 validation cycle.

4.9 Subgroup heterogeneity

The subgroup analysis demonstrated that favorable overall temporal performance did not translate into uniform operating characteristics across population groups. Detailed subgroup results are reported in Supplementary Table S4.

The largest heterogeneity was observed across age categories. Among adults aged 20–44 years, sensitivity at the unchanged development-derived analytical operating point was only 0.162, despite an age-specific ROC-AUC of 0.802. In contrast, among adults aged ≥ 65 years, sensitivity was 0.964 but specificity was only 0.141, with an age-specific ROC-AUC of 0.710.

These findings illustrate an important distinction between overall discrimination and threshold-dependent performance. The favorable pooled temporal ROC-AUC of approximately 0.847 did not reveal the large differences in sensitivity and specificity that emerged when the same analytical operating point was applied within individual age groups. Thus, a single overall AUC can mask substantial threshold-dependent heterogeneity across clinically relevant subpopulations.

The age-related pattern is consistent with both the dominant contribution of age in the SHAP analysis and the application of one global operating point across the entire adult age range. Predicted probabilities were much more likely to exceed the common operating point among older participants, producing high sensitivity but poor specificity, whereas the opposite pattern occurred among younger participants.

A single global analytical operating point may therefore be inappropriate across all age groups if the model were considered for a future applied setting. Importantly, age-specific operating points were not optimized using the present temporal-validation cohort, because doing so would have used evaluation data for a new development decision and would have compromised the independence of the temporal validation.

Sex-specific discrimination was comparatively more consistent, whereas race/ethnicity subgroup performance showed some variation. The complete age-, sex-, and race/ethnicity-specific results, including sample sizes, prevalence, ROC-AUC, probability accuracy, sensitivity, specificity, PPV, and NPV, are reported in Supplementary Table S4.

These subgroup findings should be interpreted as an assessment of performance heterogeneity rather than as a comprehensive algorithmic fairness evaluation. A dedicated fairness analysis would additionally require subgroup-specific calibration assessment, systematic error-rate comparisons, intersectional analyses, evaluation of differential decision consequences, and investigation of whether any resulting disparities would persist in independent populations.

4.10 Potential interpretation and implementation considerations

The present model should currently be regarded as a research framework for stratifying prevalent physician-diagnosed CVD status.

It has not demonstrated prospective cardiovascular-event prediction, detection of previously undiagnosed disease, or improvement in clinical outcomes.

Accordingly, the current results should not be used to justify clinical diagnosis or treatment decisions.

The predictor-domain results nevertheless provide useful information for future model design. Some predictors, including cholesterol, HbA1c, and serum cotinine, require biological measurements, meaning the complete model is not a questionnaire-only system.

Future studies could therefore investigate laboratory-enabled, routine-clinical, and lower-resource predictor configurations using independent development and validation datasets.

4.11 Strengths

The principal methodological strength was strict chronological separation between development and evaluation. Temporal-validation participants were not used for algorithm selection, hyperparameter tuning, probability calibration, preprocessing, or operating-point selection. The primary temporal results therefore reflect performance of a frozen framework transported to later calendar periods rather than a contemporaneous random split.

Evaluation extended beyond ROC-AUC to precision–recall performance, Brier score, log loss, calibration intercept and slope, bootstrap uncertainty, threshold-dependent metrics, exploratory decision-curve analysis, cycle-specific stability, predictor-domain ablation, subgroup heterogeneity, and SHAP-based explanation stability. Preprocessing was embedded within the development pipelines to reduce information leakage, and logistic regression was retained as a conventional reference model.

For population-oriented descriptive analyses, NHANES examination weights, strata, primary sampling units, and survey-domain estimation were incorporated. Computational traceability was supported through preservation of fitted models, cohort-membership hashes, source-file checksums, software-version records, and an artifact manifest. These design and reporting choices are consistent with the transparency and validation principles emphasized by TRIPOD+AI and PROBAST+AI [25, 26].

4.12 Limitations

Several limitations constrain interpretation. First, the outcome was cross-sectional and based on self-reported previous physician diagnosis. The model estimates prevalent physician-diagnosed CVD status, not incident cardiovascular events, and participants with undiagnosed disease could have been misclassified as non-cases.

Second, although chronological validation is stronger than a contemporaneous random holdout for assessing temporal transportability, both development and validation cohorts originated from NHANES. The study therefore does not constitute geographic, healthcare-system, or fully independent external validation.

Third, laboratory methods and equipment changed during some NHANES cycles, including measurements relevant to HbA1c, lipids, and serum cotinine. Such measurement heterogeneity may have contributed to cycle-specific performance differences.

Fourth, primary predictive performance was evaluated at the participant level without incorporating NHANES sampling weights into model fitting or performance estimation. Survey weights were used for population-oriented descriptive statistics; the predictive estimates should therefore not be

interpreted automatically as design-weighted national performance.

Fifth, the tuned development cross-validation metrics were obtained after hyperparameter selection using the same development-fold structure. They are therefore model-selection estimates rather than an independent post-selection internal validation. The chronologically later cohort is the principal generalization assessment.

Sixth, a single global analytical operating point produced markedly different operating characteristics across age groups. Sensitivity was low among younger adults and specificity was low among adults aged ≥ 65 years. Pooled metrics therefore do not imply uniform threshold-dependent behavior across clinically relevant subgroups.

Seventh, SHAP describes how the fitted model uses predictors but does not establish causal relationships. Correlated predictors, nonlinear effects, model interactions, one-hot encoding, and missingness indicators can influence feature attribution.

Eighth, potentially informative variables such as medication use, family history, detailed dietary characteristics, kidney-function measurements, and additional cardiovascular biomarkers were not included. Their omission may have limited predictive performance and clinical characterization.

Ninth, calibration transportability was assessed only within the same overarching survey infrastructure. Independent populations with different disease prevalence, measurement systems, and healthcare contexts may require recalibration.

Finally, decision-curve analysis quantified mathematical net benefit under assumed thresholds but did not evaluate patient acceptability, workflow burden, cost-effectiveness, diagnostic consequences, or clinical outcomes. The present framework remains a research model and should not be used for diagnosis, treatment, or other clinical decision-making without independent validation, purpose-specific operating-point development, and prospective impact assessment.

4.13 Future validation and research

Future work should prioritize independent external validation in datasets collected outside NHANES and prospective modeling of incident cardiovascular outcomes with a clearly specified prediction horizon. Such designs would permit like-for-like comparison with established prospective cardiovascular risk equations.

Because age was the dominant predictor and age-stratified operating characteristics differed substantially, future model refinement should also quantify incremental value over simple age-only and age-plus-sex benchmarks. Additional priorities include subgroup-specific calibration assessment, age- or purpose-specific thresholds developed in independent data, intersectional fairness analyses, recalibration in populations with different prevalence, prospective drift monitoring, and comparison with established clinical prediction approaches.

Ultimately, prospective impact studies are required to determine whether model-guided decisions improve clinical or population-health outcomes.

5 Conclusion

A development-locked XGBoost framework for prevalent physician-diagnosed CVD stratification maintained useful discrimination when transported from NHANES 2007–2012 to 2013–2018,

with temporal ROC-AUC 0.847 and average precision 0.403. Aggregate calibration intercept and slope were close to their ideal values, although the sigmoid calibration layer did not improve the Brier score relative to the uncalibrated XGBoost. Social-context and lifestyle/exposure variables provided modest incremental predictive information, and SHAP-derived original-predictor rankings were stable across temporal cycles.

The pooled analytical operating point yielded sensitivity 0.820, specificity 0.718, and NPV 0.969, but threshold-dependent behavior was not uniform across age groups. The framework should therefore be interpreted as a temporally evaluated research model for prevalent-CVD status stratification, not as a prospective cardiovascular risk calculator or a clinically validated diagnostic tool. Independent external validation, prospective outcome evaluation, and purpose-specific clinical-impact assessment are required before implementation.

Declarations

Ethics Approval and Consent to Participate

This study was a secondary analysis of publicly available, de-identified data from the National Health and Nutrition Examination Survey (NHANES). NHANES protocols were reviewed and approved by the National Center for Health Statistics Research Ethics Review Board, and written informed consent was obtained from participants under the original survey procedures. The present analysis involved no direct participant recruitment, intervention, or access to personally identifiable information. Therefore, no additional institutional ethical approval or participant consent was required for this secondary analysis of publicly available de-identified data.

Consent for Publication

Not applicable. The manuscript does not contain identifiable individual-level participant information.

Availability of Data and Materials

The datasets analyzed in this study are publicly available from the National Health and Nutrition Examination Survey (NHANES), conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention. NHANES datasets, documentation, questionnaires, laboratory files, and analytic guidelines are publicly accessible through the official NHANES website. The present analysis used NHANES survey cycles from 2007–2008 through 2017–2018.

Code Availability

The analysis code used for data preprocessing, model development, hyperparameter optimization, probability calibration, temporal validation, bootstrap analysis, decision-curve analysis, predictor-domain ablation, subgroup evaluation, and SHAP-based explainability will be made available in a public repository upon publication.

Funding

This research received no external funding.

Conflict of Interest

The authors declare that they have no competing interests.

Author Contributions

H.M.R.: Conceptualization, methodology, software, data curation, formal analysis, investigation, visualization, validation, writing—original draft, and writing—review and editing.

M.T.S.A.: Methodology, validation, interpretation of results, and writing—review and editing.

Both authors reviewed and approved the final version of the manuscript.

Acknowledgments

The authors acknowledge the National Center for Health Statistics and the participants of the National Health and Nutrition Examination Survey for making the data used in this study publicly available.

References

- [1] Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *Journal of the American College of Cardiology*. 2020;76(25):2982–3021. doi:10.1016/j.jacc.2020.11.010.
- [2] Yusuf S, Hawken S, Ounpuu S, Dans T, Avezum A, Lanas F, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *The Lancet*. 2004;364(9438):937–952. doi:10.1016/S0140-6736(04)17018-9.
- [3] D’Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. 2008;117(6):743–753. doi:10.1161/CIRCULATIONAHA.107.699579.
- [4] Goff DC Jr, Lloyd-Jones DM, Bennett G, Coady S, D’Agostino RB Sr, Gibbons R, Greenland P, 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. *Circulation*. 2014;129(25 Suppl 2):S49–S73. doi:10.1161/01.cir.0000437741.48606.98.
- [5] Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357:j2099. doi:10.1136/bmj.j2099.
- [6] SCORE2 Working Group and ESC Cardiovascular Risk Collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *European Heart Journal*. 2021;42(25):2439–2454. doi:10.1093/eurheartj/ehab309.
- [7] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785–794. doi:10.1145/2939672.2939785.
- [8] Weng SF, Reips J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE*. 2017;12(4):e0174944. doi:10.1371/journal.pone.0174944.
- [9] Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular event prediction by machine learning: the Multi-Ethnic Study of Atherosclerosis. *Circulation Research*. 2017;121(9):1092–1101. doi:10.1161/CIRCRESAHA.117.311312.
- [10] Kakadiaris IA, Vrigkas M, Yen AA, Kuznetsova T, Budoff M, Naghavi M. Machine learning outperforms ACC/AHA CVD risk calculator in MESA. *Journal of the American Heart Association*. 2018;7(22):e009476. doi:10.1161/JAHA.118.009476.
- [11] Liu W, Laranjo L, Klimis H, Chiang J, Yue J, Marschner S, et al. Machine-learning versus traditional approaches for atherosclerotic cardiovascular risk prognostication in primary prevention cohorts: a systematic review and meta-analysis. *European Heart Journal – Quality of Care and Clinical Outcomes*. 2023;9(4):310–322. doi:10.1093/ehjqcc/qcad017.
- [12] Cai Y, Cai YQ, Tang LY, Wang YH, Gong M, Jing TC, et al. Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation

screening tool: a systematic review. *BMC Medicine*. 2024;22(1):56. doi:10.1186/s12916-024-03273-7.

- [13] Liu T, Krentz A, Lu L, Curcin V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *European Heart Journal – Digital Health*. 2025;6(1):7–22. doi:10.1093/ehjdh/ztae080.
- [14] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–138. doi:10.1097/EDE.0b013e3181c30fb2.
- [15] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*. 2019;17:230. doi:10.1186/s12916-019-1466-7.
- [16] Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*. 2014;14:40. doi:10.1186/1471-2288-14-40.
- [17] Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, Snell KIE. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine*. 2021;40(19):4230–4251. doi:10.1002/sim.9025.
- [18] National Center for Health Statistics. National Health and Nutrition Examination Survey: survey methods and analytic guidelines. Centers for Disease Control and Prevention. Available from: <https://wwwn.cdc.gov/nchs/nhanes/analyticguidelines.aspx>. Accessed 2026.
- [19] Huang AA, Huang SY. Use of machine learning to identify risk factors for coronary artery disease. *PLoS ONE*. 2023;18(4):e0284103. doi:10.1371/journal.pone.0284103.
- [20] Cao X, Ma J, He X, Liu Y, Yang Y, Wang Y, Zhang C. Unlocking the link: predicting cardiovascular disease risk with a focus on airflow obstruction using machine learning. *BMC Medical Informatics and Decision Making*. 2025;25(1):50. doi:10.1186/s12911-025-02885-0.
- [21] Ahiduzzaman M, Hasan MN. Interpretable machine learning for cardiovascular risk prediction: insights from NHANES dietary and health data. *PLoS ONE*. 2025;20(11):e0335915. doi:10.1371/journal.pone.0335915.
- [22] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*. 2017;30.
- [23] Kırboğa KK, Küçüksille EU. Identifying cardiovascular disease risk factors in adults with explainable artificial intelligence. *Anatolian Journal of Cardiology*. 2023;27(11):657–663. doi:10.14744/AnatolJCardiol.2023.3214.
- [24] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*. 2006;26(6):565–574. doi:10.1177/0272989X06295361.
- [25] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378.

- [26] Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. doi:10.1136/bmj-2024-082505.
- [27] Breiman L. Random forests. *Machine Learning*. 2001;45:5–32. doi:10.1023/A:1010933404324.
- [28] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. LightGBM: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*. 2017;30:3146–3154.
- [29] Schultz WM, Kelli HM, Lisko JC, Varghese T, Shen J, Sandesara P, et al. Socioeconomic status and cardiovascular outcomes: challenges and interventions. *Circulation*. 2018;137(20):2166–2178. doi:10.1161/CIRCULATIONAHA.117.029652.
- [30] Yusuf S, Hawken S, Ounpuu S, Bautista L, Franzosi MG, Commerford P, et al. Obesity and the risk of myocardial infarction in 27,000 participants from 52 countries: a case-control study. *The Lancet*. 2005;366(9497):1640–1649. doi:10.1016/S0140-6736(05)67663-5.
- [31] Whincup PH, Gilg JA, Emberson JR, Jarvis MJ, Feyerabend C, Bryant A, Walker M, Cook DG. Passive smoking and risk of coronary heart disease and stroke: prospective study with cotinine measurement. *BMJ*. 2004;329(7459):200–205. doi:10.1136/bmj.38146.427188.55.
- [32] Sattelmair J, Pertman J, Ding EL, Kohl HW III, Haskell W, Lee IM. Dose response between physical activity and risk of coronary heart disease: a meta-analysis. *Circulation*. 2011;124(7):789–795. doi:10.1161/CIRCULATIONAHA.110.010710.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterial.pdf](#)