

Supplementary Material

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

This supplementary file contains the supplementary tables and figures cited in the main manuscript. All development-stage tuning was performed using NHANES 2007–2012 only, and temporal-validation data from NHANES 2013–2018 were not used for model selection, hyperparameter tuning, calibration-method selection, or operating-point selection.

Supplementary Table S1. Candidate hyperparameter search spaces and selected configurations

Model	Development-only search space	Selected configuration
Logistic regression	$C \in \{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100\}$; all 11 regularization strengths evaluated.	$C = 0.03$.
Random forest	$n_estimators \in \{300, 500, 700, 900\}$; $max_depth \in \{\text{None}, 6, 10, 14, 20\}$; $min_samples_split \in \{2, 5, 10\}$; $min_samples_leaf \in \{2, 5, 10, 20\}$; $max_features \in \{\text{sqrt}, 0.5, 0.75\}$. Sixteen randomized configurations were evaluated.	$n_estimators=900$; $max_depth=\text{None}$; $min_samples_split=2$; $min_samples_leaf=20$; $max_features=\text{sqrt}$.
XGBoost	$n_estimators \in \{300, 500, 700, 900\}$; $learning_rate \in \{0.02, 0.03, 0.05, 0.08\}$; $max_depth \in \{2, 3, 4, 5\}$; $min_child_weight \in \{1, 3, 5, 8\}$; $subsample \in \{0.75, 0.85, 1.0\}$; $colsample_bytree \in \{0.75, 0.85, 1.0\}$; $reg_alpha \in \{0, 0.1, 0.5, 1.0\}$; $reg_lambda \in \{1, 2, 5, 10\}$. Twenty randomized configurations were evaluated.	$n_estimators=500$; $learning_rate=0.03$; $max_depth=2$; $min_child_weight=8$; $subsample=0.75$; $colsample_bytree=0.85$; $reg_alpha=1.0$; $reg_lambda=2.0$.
LightGBM	$n_estimators \in \{300, 500, 700, 900\}$; $learning_rate \in \{0.02, 0.03, 0.05, 0.08\}$; $num_leaves \in \{15, 31, 47, 63\}$; $max_depth \in \{-1, 5, 8, 12\}$; $min_child_samples \in \{20, 30, 50, 80\}$; $subsample \in \{0.75, 0.85, 1.0\}$; $colsample_bytree \in \{0.75, 0.85, 1.0\}$; $reg_alpha \in \{0, 0.1, 0.5, 1.0\}$; $reg_lambda \in \{1, 2, 5, 10\}$. Twenty randomized configurations were evaluated.	$n_estimators=300$; $learning_rate=0.02$; $num_leaves=15$; $max_depth=-1$; $min_child_samples=30$; $subsample=0.85$; $colsample_bytree=0.75$; $reg_alpha=0.0$; $reg_lambda=1.0$.

Note. Five-fold stratified development-only cross-validation used the same locked fold assignments across candidate algorithms. Mean ROC-AUC was the prespecified model-ranking criterion. The selected XGBoost configuration achieved mean ROC-AUC 0.8442, average precision 0.4055, and Brier score 0.08015 during tuned development-stage cross-validation. LightGBM achieved mean ROC-AUC 0.8440, average precision 0.4063, and Brier score 0.08010; the very small difference supports interpreting XGBoost selection as criterion-based rather than evidence of substantial algorithmic superiority.

Supplementary Table S2. Predictor-specific missingness in the final eligible analytical cohort

Predictor	Missing, n	Missing, %
Poverty-to-income ratio	3,529	10.20
HDL cholesterol	3,371	9.74
Total cholesterol	3,371	9.74
Waist circumference	3,362	9.72
Serum cotinine	3,301	9.54
HbA1c	3,012	8.70
Systolic blood pressure	2,755	7.96
Diastolic blood pressure	2,755	7.96
BMI	1,818	5.25
Diabetes	828	2.39
Moderate recreational activity	4	0.01
Vigorous recreational activity	2	0.01
Age	0	0.00
Sex	0	0.00
Race/ethnicity	0	0.00

Note. Denominator $N = 34,602$ adults in the final outcome-eligible analytical cohort. Participants were not excluded because of predictor missingness. Data-dependent preprocessing, including imputation and missingness-indicator construction, was fitted only within the relevant development training data and then applied to held-out or temporal-validation observations.

Supplementary Table S3. Original-predictor global SHAP importance

Rank	Original predictor	Mean SHAP
1	Age	1.1439
2	Total cholesterol	0.2329
3	Poverty-to-income ratio	0.1962
4	Waist circumference	0.1777
5	Serum cotinine	0.1656
6	Diabetes	0.1582
7	HDL cholesterol	0.1060
8	Systolic blood pressure	0.0917
9	Race/ethnicity	0.0885
10	Diastolic blood pressure	0.0736
11	Vigorous recreational activity	0.0691
12	HbA1c	0.0687
13	BMI	0.0454
14	Sex	0.0396
15	Moderate recreational activity	0.0263

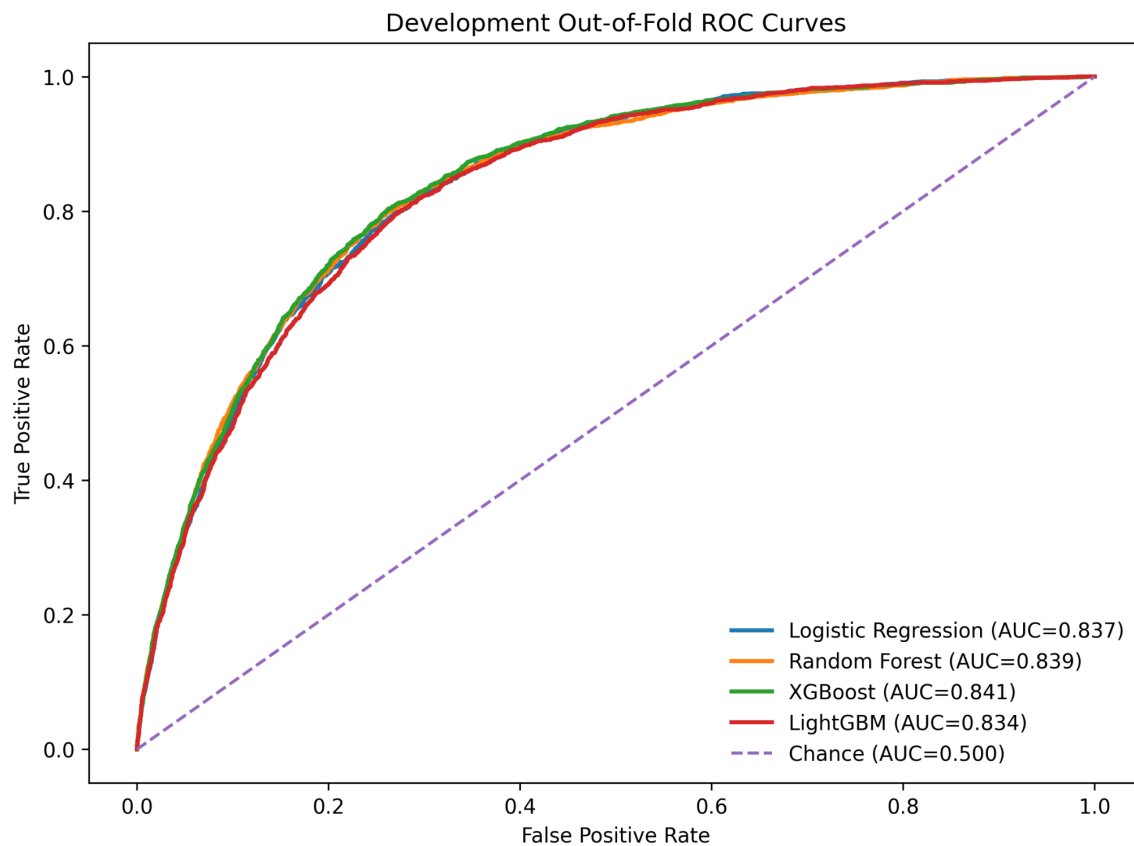
Note. Original-predictor importance was derived from SHAP values for the underlying uncalibrated XGBoost estimator. Contributions from transformed features belonging to the same original predictor were grouped for original-predictor reporting. SHAP values describe fitted-model behavior and should not be interpreted as causal effects.

Supplementary Table S4. Detailed subgroup performance in temporal validation

Dimension	Subgroup	N	Cases	Prev.	ROC-AUC	AP	Brier	Sens.	Spec.	PPV	NPV	F1
Age	20–44	7,010	130	0.0185	0.802	0.121	0.018	0.162	0.991	0.250	0.984	0.196
Age	45–64	5,825	616	0.1058	0.754	0.273	0.088	0.683	0.686	0.204	0.948	0.315
Age	≥ 65	4,141	1,175	0.2837	0.710	0.473	0.187	0.964	0.141	0.308	0.909	0.467
Sex	Female	8,816	828	0.0939	0.833	0.329	0.074	0.768	0.738	0.233	0.968	0.358
Sex	Male	8,160	1,093	0.1339	0.855	0.453	0.093	0.859	0.695	0.303	0.970	0.449
Race/ethnicity	Mexican American	2,483	179	0.0721	0.839	0.300	0.059	0.732	0.766	0.196	0.974	0.309
Race/ethnicity	Non-Hispanic Black	3,660	439	0.1199	0.811	0.370	0.092	0.788	0.675	0.248	0.959	0.378
Race/ethnicity	Non-Hispanic White	6,235	938	0.1504	0.847	0.462	0.103	0.877	0.657	0.312	0.968	0.460
Race/ethnicity	Other Hispanic	1,788	168	0.0940	0.840	0.355	0.072	0.732	0.765	0.245	0.965	0.367
Race/ethnicity	Other race/multiracial	2,810	197	0.0701	0.868	0.330	0.056	0.772	0.821	0.246	0.979	0.373

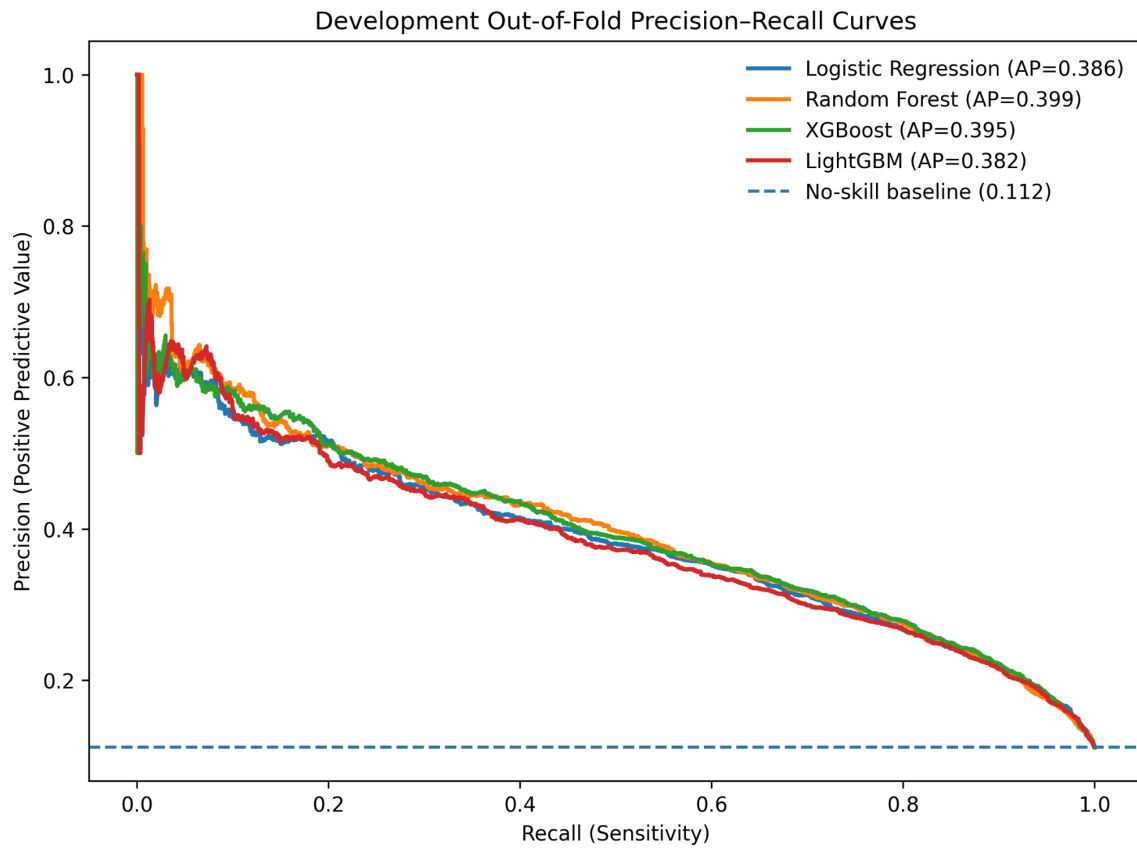
Note. Prev., empirical unweighted subgroup prevalence; AP, average precision; Sens., sensitivity; Spec., specificity; PPV, positive predictive value; NPV, negative predictive value. All threshold-dependent values use the same development-derived analytical operating point of 0.076028. The frozen sigmoid-calibrated model was used without subgroup-specific tuning, recalibration, or threshold optimization. ROC-AUC 95% confidence intervals were: age 20–44, 0.758–0.839; age 45–64, 0.735–0.773; age ≥ 65, 0.694–0.726; women, 0.821–0.846; men, 0.844–0.865; Mexican American, 0.812–0.864; non-Hispanic Black, 0.790–0.829; non-Hispanic White, 0.834–0.856; Other Hispanic, 0.813–0.864; other race/multiracial, 0.842–0.892. Confidence intervals were obtained from 500 participant-level bootstrap resamples per eligible subgroup.

Supplementary Figure S1. Development out-of-fold ROC curves for fixed candidate models



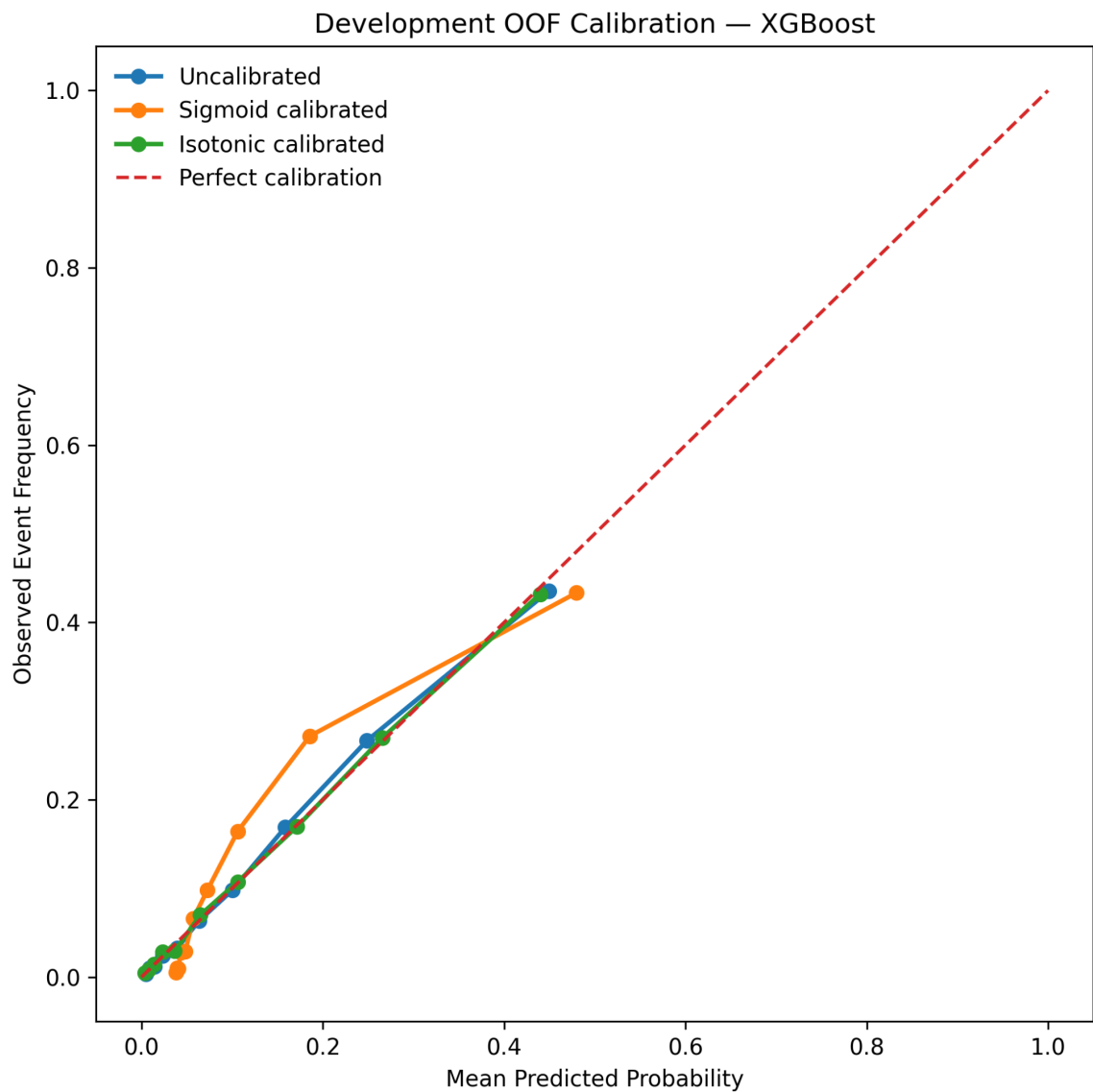
Supplementary Figure S1. Development out-of-fold receiver operating characteristic curves for the fixed logistic regression, random forest, XGBoost, and LightGBM candidate models. These curves summarize development-stage performance only and were not generated from the temporal-validation cohort.

Supplementary Figure S2. Development out-of-fold precision–recall curves for fixed candidate models



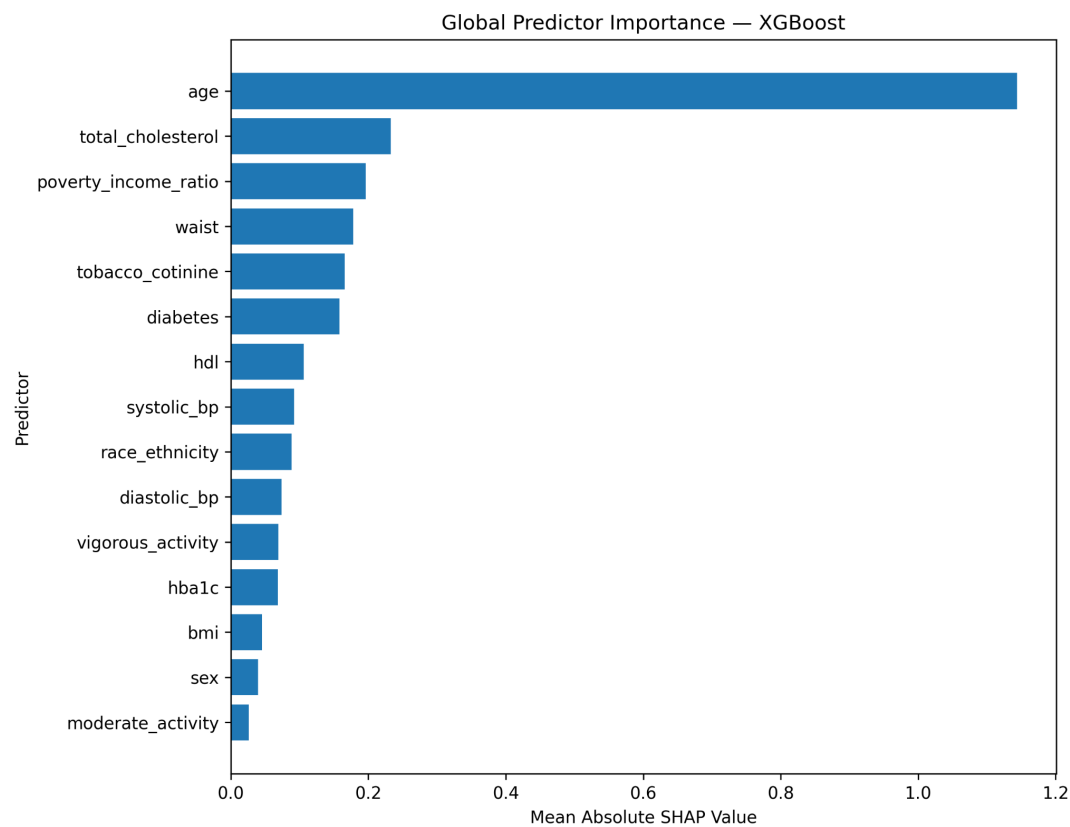
Supplementary Figure S2. Development out-of-fold precision–recall curves for the fixed candidate models. Average precision is interpreted relative to the development-cohort prevalence-based no-skill reference.

Supplementary Figure S3. Development out-of-fold calibration of the selected XGBoost model



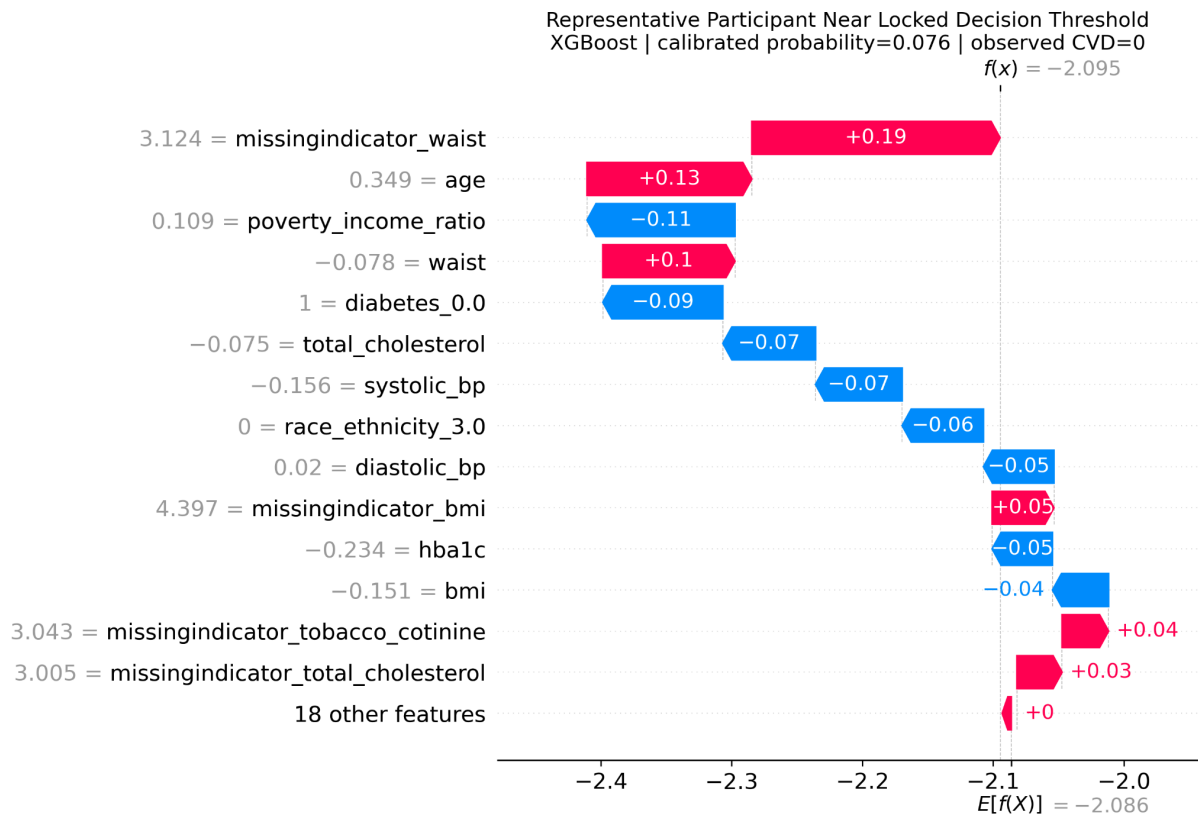
Supplementary Figure S3. Development out-of-fold calibration curves for the selected XGBoost model, comparing uncalibrated probabilities with sigmoid/Platt calibration and isotonic calibration. Sigmoid calibration was designated as the primary calibrated specification during development; alternative probability outputs were retained as sensitivity analyses.

Supplementary Figure S4. Original-predictor SHAP importance



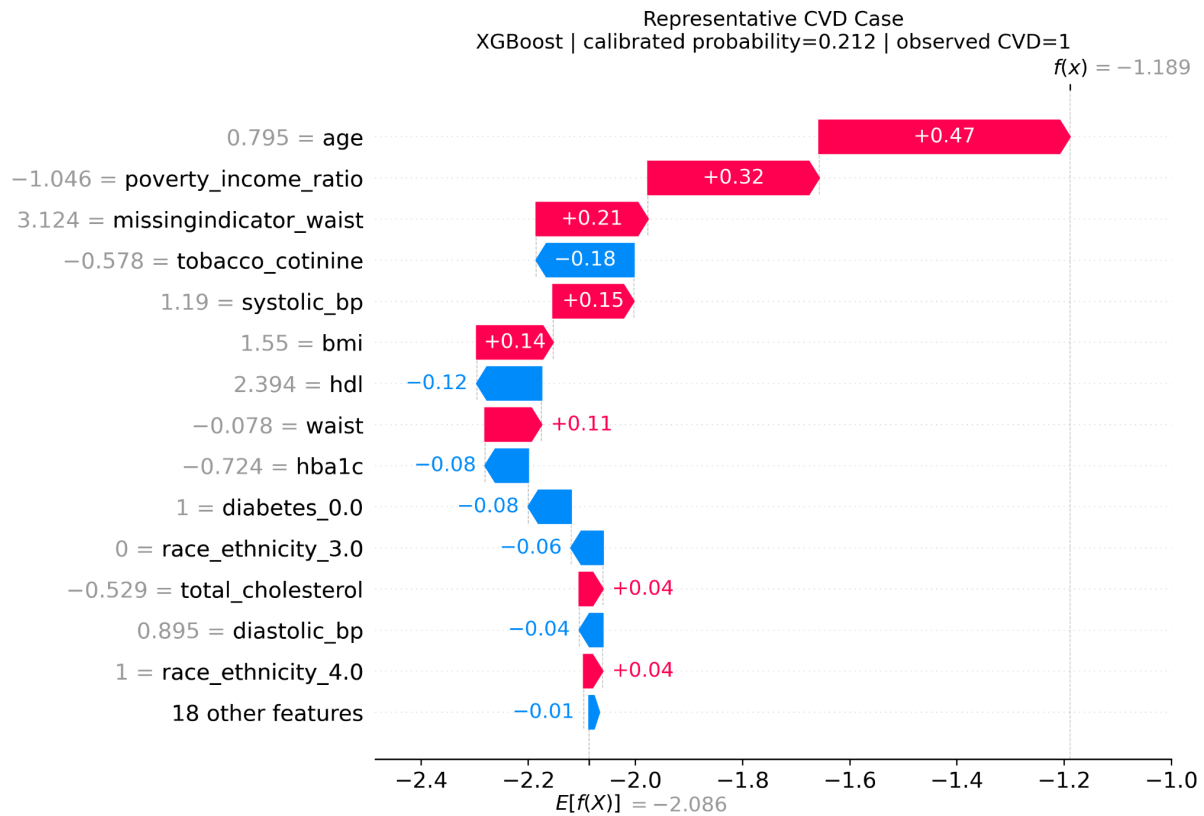
Supplementary Figure S4. Global original-predictor SHAP importance for the final XGBoost base learner. Age was the dominant original predictor, followed by total cholesterol, poverty-to-income ratio, waist circumference, serum cotinine, diabetes, HDL cholesterol, and systolic blood pressure. SHAP importance describes fitted-model behavior rather than causal effects.

Supplementary Figure S5. Local SHAP explanation for the temporal participant nearest the locked operating point



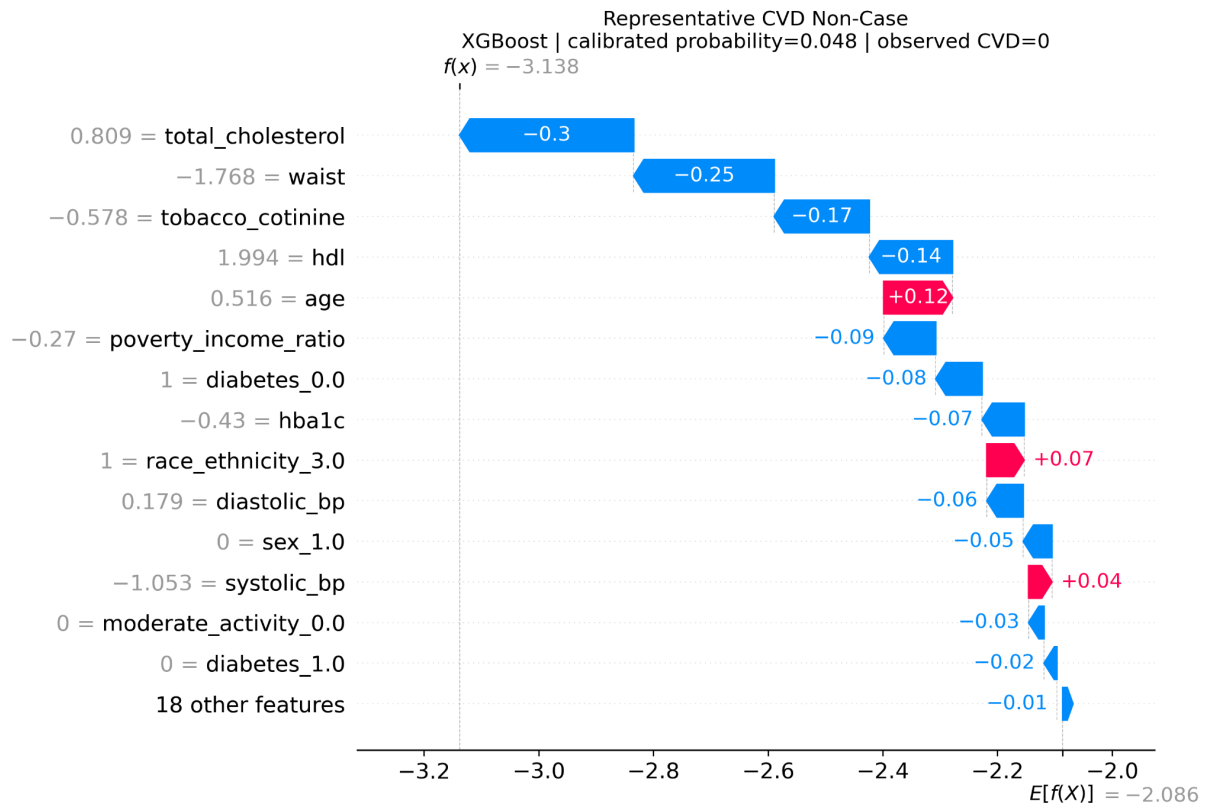
Supplementary Figure S5. Participant-level SHAP explanation for the temporal-validation participant whose sigmoid-calibrated predicted probability was closest to the development-locked analytical operating point of 0.076028. Participant selection was based on predicted-probability proximity to the threshold and was independent of the resulting SHAP values. SHAP contributions refer to the underlying uncalibrated XGBoost estimator.

Supplementary Figure S6. Local SHAP explanation for a representative temporal CVD case



Supplementary Figure S6. Participant-level SHAP explanation for the observed CVD case whose sigmoid-calibrated predicted probability was closest to the median predicted probability among temporal-validation CVD cases. Selection did not depend on SHAP values. SHAP contributions refer to the underlying uncalibrated XGBoost estimator and should not be interpreted causally.

Supplementary Figure S7. Local SHAP explanation for a representative temporal non-case



Supplementary Figure S7. Participant-level SHAP explanation for the temporal-validation non-case whose sigmoid-calibrated predicted probability was closest to the median predicted probability among temporal-validation non-cases. Selection did not depend on SHAP values. SHAP contributions refer to the underlying uncalibrated XGBoost estimator and should not be interpreted causally.