

# Benchmarking machine learning for cardiovascular disease classification across 46 countries with geographic and temporal transportability assessment

Wingston Felix Ng'ambi<sup>1✉23</sup>

[wingston.ngambi@gmail.com](mailto:wingston.ngambi@gmail.com)

Kamuzu University of Health Sciences

Cosmas Zyambo<sup>2✉3</sup>

University of Zambia

Lawrence Kazembe<sup>4✉5</sup>

University of Malawi

---

## Research Article

**Keywords:** cardiovascular disease, machine learning, artificial intelligence, WHO STEPS, class imbalance, calibration, transportability, external validation, explainable AI, global health

**Posted Date:** October 14th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11021830/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

## Abstract

## Background

Machine-learning models for cardiovascular disease (CVD) are often evaluated with random participant splits, which may overstate performance when countries differ in prevalence, risk-factor distributions, healthcare access, and survey implementation. We benchmarked statistical and machine-learning classifiers for prevalent self-reported CVD with explicit assessment of class imbalance, calibration, explainability, and geographic and temporal transportability.

## Methods

We analysed harmonised WHO STEPwise approach to noncommunicable disease risk factor surveillance (STEPS) surveys conducted from 2013 to 2023. The cohort included 187,791 adults, including 16,305 CVD cases, from 46 countries. Logistic regression, elastic-net logistic regression, random forest, and XGBoost were evaluated using five-fold country-grouped internal-external cross-validation with survey, country-year, and class weighting. We assessed ROC AUC, precision-recall AUC (PR AUC), Brier score, calibration, threshold-dependent performance, SHAP importance, and earlier-to-later transportability.

## Results

Weighted CVD prevalence was 7.78%. XGBoost achieved ROC AUC 0.671, PR AUC 0.162, and Brier score 0.0692; logistic regression achieved 0.669, 0.154, and 0.0694. Country-level performance varied substantially. Removing hypertension and diabetes reduced ROC AUC by 0.038–0.040. Temporal logistic-regression ROC AUC was 0.634 in Mongolia, 0.579 in Uganda, and 0.670 in Viet Nam.

## Conclusions

Complex models produced only small pooled gains over logistic regression, whereas geographic and temporal heterogeneity was substantial. Responsible public-health AI benchmarking should prioritise minority-class performance, calibration, robustness, explainability, and transportability rather than pooled ROC AUC alone.

## Background

Cardiovascular diseases remain the leading cause of death globally. The World Health Organisation estimated 19.8 million CVD deaths in 2022, accounting for about 32% of global deaths, with more than three-quarters occurring in low- and middle-income countries [1–4]. This burden, together with major cross-country differences in risk-factor exposure, diagnosis, treatment access, and health-system capacity, creates a strong rationale for scalable analytic tools that can support population health surveillance and chronic disease management [1–4].

Machine learning can model nonlinear relationships, interactions, and high-dimensional predictor structures that may be difficult to specify a priori. In CVD research, such flexibility has encouraged extensive comparison of logistic regression with tree-based ensembles and boosting algorithms [5, 6]. However, a model that performs well under a random participant split can still fail when transported to a different country, subregion, survey wave, or healthcare context. Structured data can make ordinary random cross-validation optimistic because dependence associated with geography, time, and data-generating systems is shared between training and test samples [4].

This problem is especially important for population surveys. WHO STEPS was designed as a standardised platform for measuring behavioural and biological noncommunicable disease risk factors across countries [7, 8]. Standardisation improves comparability, but survey populations still differ in age structure, prevalence, healthcare access, measurement completeness, and national implementation. In addition, prevalent self-reported CVD is not a prospectively adjudicated clinical endpoint. A participant can report prior disease only if disease occurred, was recognised, and was known to the participant. Predictors such as hypertension or diabetes may therefore improve classification partly because they encode underlying cardiometabolic risk and partly because they reflect prior contact with health services.

Class imbalance adds another challenge. When CVD prevalence is low, ROC AUC may look acceptable even when precision among positive predictions is poor. Precision-recall analysis is therefore particularly informative when the positive class is uncommon [9, 10]. Calibration is equally important because a classifier can rank participants correctly while systematically overestimating or underestimating absolute probabilities [11, 12]. Current TRIPOD + AI guidance emphasises transparent reporting of model development and evaluation regardless of whether regression or machine-learning methods are used, while PROBAST + AI places explicit emphasis on the quality, bias, and applicability of both model development and evaluation [13, 14].

We therefore developed and applied a multidomain CVD AI benchmarking framework in global STEPS data. Rather than asking which algorithm has the single highest ROC AUC, the framework evaluates harmonisation, preprocessing, class imbalance, model discrimination, calibration, classification performance, dependence on detection-related predictors, interpretability, geographic transportability, and temporal transportability. We hypothesised that ensemble models might offer modest gains in pooled discrimination, but that model rankings and calibration would vary across countries and subregions, and that temporal performance would be weaker than suggested by pooled geographic metrics [15–17].

## Methods

### Study design and data source

We conducted a secondary analysis of harmonised WHO STEPS microdata. STEPS uses standardised questionnaires, physical measurements, and, where available, biochemical measurements to support comparable surveillance of noncommunicable disease risk factors [7, 8]. The source analytic file contained 455341 records from 89 countries or territories. Surveys in the final benchmarking cohort were conducted between 2013 and 2023.

The study was designed as a model benchmarking and transportability evaluation rather than a causal analysis. Because the outcome was prevalent self-reported CVD measured at the survey visit, we use the term classification rather than future risk prediction. Reporting was guided by TRIPOD + AI and the evaluation concepts in PROBAST + AI [13, 14].

### Outcome and cohort eligibility

The binary outcome was self-reported history of CVD based on the harmonised STEPS item: "Have you ever had a heart attack or chest pain from heart disease (angina) or a stroke (cerebrovascular accident or incident)?" Responses coded Yes and No were retained. The outcome represents reported or diagnosed prevalent CVD and is not a clinically adjudicated endpoint.

Of the 89 countries or territories in the harmonised source file, 48 had at least one observation with a valid CVD outcome. For country-specific discrimination, a survey population had to contain both outcome classes. Bahamas and Sudan had CVD-positive observations but no observed CVD-negative controls among records with a valid harmonised outcome and were therefore excluded from binary model benchmarking. The primary analytic cohort comprised 187,791 adults from

46 countries across 15 subregions, including 16,305 CVD cases. A complete country-by-country eligibility audit, including the names of all 46 included countries and their survey years, is provided in Additional file 1 (Supplementary Table S1).

## Predictor sets

The full predictor set contained 16 harmonised variables available across the analytic surveys: age, sex, residence, education level, occupation, marital status, smoking history, harmful alcohol use, physical activity, high salt intake, low fruit intake, low vegetable intake, body mass index (BMI), hypertension status, diabetes status, and a harmonised DASH dietary score. Continuous variables were retained as continuous where available. Categorical variables were encoded consistently within the modelling pipeline.

A prespecified detection-adjusted predictor set removed hypertension and diabetes status. These variables were selected because they can represent both underlying cardiometabolic biology and prior diagnostic or healthcare contact. The contrast between full and adjusted sets was therefore interpreted as dependence on the information carried by detection-related variables, not as proof of detection bias.

## CVD AI benchmarking framework

The proposed framework had nine connected components (Fig. 1). First, outcome, predictors, survey design variables, country, year, and subregion were harmonised. Second, preprocessing was estimated within development data only. Third, class imbalance was addressed through class-weighted training and PR AUC. Fourth, four modelling families were benchmarked. Fifth, probabilities were recalibrated using development data. Sixth, performance was compared after removing detection-related variables. Seventh, XGBoost was interpreted using SHAP values. Eighth, geographic transportability was evaluated with country-grouped internal-external cross-validation and summarised by subregion. Ninth, temporal transportability was assessed from an earlier survey to the immediately available later survey in countries with two eligible waves after 2013.

## Preprocessing and missing data

All preprocessing was fitted within the development data. Age, BMI, and DASH score were treated as numeric predictors and median-imputed. Numeric predictors were standardised for logistic and elastic-net regression. Categorical predictors were imputed to the most frequent category and one-hot encoded; rare one-hot levels with fewer than 10 observations in the development data were not represented as separate encoded columns. Unseen categories in evaluation data were handled without refitting the encoder. No preprocessing parameter was estimated from held-out countries.

## Survey, country-year, and class-imbalance weighting

Survey weights were converted to numeric form, nonpositive or missing weights were replaced with the median positive weight within the corresponding country-year survey, and weights were normalised to mean 1 within each country-year. To prevent large surveys from dominating training, each observation also received an inverse country-year sample-size factor. The product of the survey weight and the country-year balancing factor was then multiplied by a class factor so that the weighted positive and negative classes contributed equally to the training objective.

$$w_i^* = w_i^{(\text{survey})} b_{c(i),t(i)} a_{y_i}, \quad a_k = \frac{W}{2W_k}$$

In Eq. 1,  $b$  denotes the country-year balancing factor,  $W$  is the total weighted mass in the development sample, and  $W_k$  is the weighted mass for class  $k$ . This approach preserves within-survey weighting while addressing both survey-size dominance and CVD class imbalance. Synthetic oversampling was not used in the primary analysis because it can alter the joint distribution of predictors and complicate the interpretation of weighted, clustered survey data [10].

## Candidate models

Four model families were selected to represent transparent statistical baselines and commonly used nonlinear ensemble approaches: standard logistic regression, elastic-net logistic regression, random forest, and XGBoost. Logistic regression used L2 regularisation with  $C = 1.0$ . Elastic net used a 0.5 L1/L2 mixing ratio with  $C = 0.5$  [18]. Random forest used 160 trees, maximum depth 14, minimum leaf size 10, and square-root feature sampling [19]. XGBoost used 200 boosting rounds, maximum depth 4, learning rate 0.05, subsampling 0.8, column subsampling 0.8, minimum child weight 5, L2 regularisation 1.0, and a binary logistic objective [20]. Hyperparameters were prespecified and held constant across external folds to avoid selecting configurations on held-out countries.

$$\text{logit} [P(Y_i = 1 | X_i)] = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$$

2

$$\min_{\beta} \left\{ - \sum_i w_i^* [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] + \lambda \left[ \alpha \| \beta \|_1 + \frac{1 - \alpha}{2} \| \beta \|_2^2 \right] \right\}$$

3

### Geographic internal-external cross-validation

The primary transportability analysis used five-fold stratified group cross-validation with country as the grouping unit. All participants from a country were assigned to the same external fold. Thus, every participant in an evaluation fold came from countries that were entirely absent from model fitting. This design is an internal-external cross-validation strategy at the country level and is better aligned with the intended question of geographic transportability than a random participant split [4].

Within each development fold, 15% of participants were reserved as a stratified calibration partition. The model was fitted on the remaining 85% using the combined survey, country-year, and class-balancing weights. No observation from an external country was used for fitting, imputation, encoding, calibration, or threshold selection.

## Probability recalibration and threshold selection

Raw predicted probabilities were recalibrated using logistic, or Platt-type, recalibration in the development calibration partition. The logit of the raw model probability was entered as the single predictor in a high- $C$  logistic model using survey weights. The final external probabilities were obtained by applying the fitted recalibration equation to the untouched external fold.

$$\text{logit} \left( p_i^{(\text{cal})} \right) = a + b \text{logit} \left( p_i^{(\text{raw})} \right)$$

4

Classification thresholds were selected within the calibration partition only. Candidate thresholds from 0.02 to 0.70 were evaluated, and the threshold maximising survey-weighted F1 score was carried forward to the external fold. This avoided optimising classification cut-points on evaluation countries.

## Performance measures

Discrimination was evaluated with survey-weighted ROC AUC and PR AUC. PR AUC was treated as co-primary because the positive CVD class was uncommon [9]. Overall probabilistic accuracy was evaluated with the weighted Brier score. Calibration was described by calibration-in-the-large through the fitted intercept, calibration slope, and weighted calibration plots. Classification measures at the development-selected threshold included sensitivity, specificity, positive predictive value (PPV), negative predictive value, and F1 score. Accuracy was not emphasised because of class imbalance.

$$\text{Brier} = \frac{\sum_i w_i (y_i - p_i)^2}{\sum_i w_i}$$

5

A well-calibrated external model is expected to have a calibration intercept near 0 and slope near 1, while recognising that no single calibration statistic is sufficient [11, 12]. Performance was first summarised across all externally generated predictions, then within each held-out country and within the 15 prespecified subregions.

## Detection-robustness analysis

Each model was refitted and externally evaluated using the detection-adjusted predictor set. We calculated adjusted-minus-full changes in ROC AUC, PR AUC, Brier score, sensitivity, specificity, PPV, F1 score, and calibration slope. A drop in performance was interpreted as evidence that the model relies on information carried by hypertension and diabetes status, which may combine underlying disease risk with differential opportunities for prior detection.

## Temporal benchmarking

Temporal benchmarking was restricted to countries with two eligible STEPS survey years from 2013 onward. For each country, the earlier survey was used for development, and the later survey was kept untouched for evaluation. The resulting pairs were Mongolia 2013 to 2019, Uganda 2014 to 2023, and Viet Nam 2015 to 2021. The same preprocessing, class weighting, calibration, threshold-selection, and performance procedures were applied, with all model development occurring only in the earlier wave.

## Interpretability

For the full-predictor XGBoost model, global feature importance was summarised using mean absolute SHAP values. SHAP values decompose model output into feature contributions using an additive game-theoretic framework [21]. Importance was interpreted as predictive contribution, not as a causal effect or modifiable intervention target.

$$f(x) = \phi_0 + \sum_{j=1}^M \phi_j$$

6

## Subgroup analyses

Externally generated predictions were also summarised by sex, age group, and residence. These analyses assessed heterogeneity in discrimination, calibration, and threshold-dependent error measures. Similarity in ROC AUC alone was not interpreted as evidence of fairness because prevalence, PPV, sensitivity, and calibration may differ across groups.

## Software and reproducibility

The reference analysis was executed in Python 3.13 using pandas, NumPy, scikit-learn, XGBoost, and SHAP-compatible model interpretation routines. A fixed random seed of 20260903 was used. The reproducibility package contains the harmonisation specification, model settings, country-year temporal pair definitions, performance outputs, and figure-generation code. A parallel R implementation can be supplied as an additional reproducibility file for journal submission.

Generative AI tools were used as assistive tools during code review, analytical documentation, manuscript drafting, and formatting [22, 23]. They did not generate the source data, independently execute the statistical analyses, define authorship, or replace author verification. All analytical outputs and interpretations reported in this manuscript were checked against executable scripts and saved results.

## Results

### Study population and CVD prevalence

Among 455,341 source records from 89 countries or territories, 188,032 participants from 48 countries had a valid harmonised CVD outcome. After excluding Bahamas and Sudan from binary benchmarking because only one observed outcome class was present, 187,791 participants from 46 countries remained, including 16,305 CVD cases. Survey-weighted CVD prevalence was 7.78%, mean age was 37.3 years, and 51.3% of participants were female (Table 1). The country names and survey years are reported in Supplementary Table S1.

Subregional prevalence varied substantially. It was highest in Micronesia at 17.66% and Eastern Asia at 13.82%, and lowest in Middle Africa at 3.35%, Northern Africa at 4.39%, and Western Africa at 4.68%. Central Asia included four countries and 12648 participants, while Western Asia included six countries and 22818 participants. The largest analytic grouping was the pooled Sub-Saharan Africa category with 29873 participants across five countries.

Table 1  
Participant distribution and CVD prevalence by subregion

| Subregion                       | Countries | Participants | CVD cases | Weighted CVD prevalence (%) | Mean age, years |
|---------------------------------|-----------|--------------|-----------|-----------------------------|-----------------|
| Overall                         | 46        | 187791       | 16305     | 7.78                        | 37.3            |
| Central Asia                    | 4         | 12648        | 999       | 5.35                        | 36.3            |
| Eastern Asia                    | 1         | 11740        | 1601      | 13.82                       | 38.5            |
| Eastern Europe                  | 3         | 14106        | 2047      | 11.04                       | 41.5            |
| Latin America and the Caribbean | 5         | 17472        | 1585      | 8.82                        | 40.3            |
| Melanesia                       | 1         | 2506         | 187       | 7.30                        | 36.6            |
| Micronesia                      | 3         | 5552         | 936       | 17.66                       | 36.2            |
| Middle Africa                   | 1         | 2418         | 90        | 3.35                        | 36.7            |
| Northern Africa                 | 3         | 16928        | 869       | 4.39                        | 37.4            |
| Polynesia                       | 5         | 8258         | 632       | 7.45                        | 40.0            |
| South-eastern Asia              | 2         | 11982        | 1070      | 7.57                        | 38.3            |
| Southern Africa                 | 2         | 6914         | 341       | 5.43                        | 33.9            |
| Southern Asia                   | 2         | 11247        | 869       | 7.17                        | 35.9            |
| Sub-Saharan Africa              | 5         | 29873        | 2045      | 6.32                        | 34.2            |
| Western Africa                  | 3         | 13329        | 674       | 4.68                        | 35.5            |
| Western Asia                    | 6         | 22818        | 2360      | 8.82                        | 37.9            |

CVD, cardiovascular disease. Weighted prevalence uses normalised survey weights within each country-year. Country names and survey years are provided in Supplementary Table S1; detailed demographic and risk-factor characteristics are provided in the supplementary material.

## Overall external geographic performance

Across externally generated country-held-out predictions, differences between model families were small (Table 2 and Fig. 2). With the full predictor set, XGBoost had the highest ROC AUC at 0.671 and PR AUC at 0.162, with a Brier score of 0.0692. Logistic regression had ROC AUC 0.669, PR AUC 0.154, and Brier score 0.0694. Elastic net was nearly identical to logistic regression, and random forest had ROC AUC 0.665 and PR AUC 0.156.

Threshold-dependent performance showed the practical effect of class imbalance. XGBoost sensitivity was 0.337, specificity 0.876, PPV 0.187, and F1 score 0.240. Logistic regression sensitivity was 0.350, specificity 0.870, PPV 0.185, and F1 score 0.242. Thus, the small improvement in XGBoost PR AUC did not translate into a clearly superior F1 score. Calibration slopes were 0.898 for XGBoost and 0.934 for logistic regression, indicating some compression of predicted risk in both models.

Table 2  
Overall country-external performance of full and detection-adjusted models

| Predictor set      | Model        | ROC AUC | PR AUC | Brier  | Sensitivity | Specificity | PPV   | F1    | Calibration slope |
|--------------------|--------------|---------|--------|--------|-------------|-------------|-------|-------|-------------------|
| Full               | Logistic     | 0.669   | 0.154  | 0.0694 | 0.350       | 0.870       | 0.185 | 0.242 | 0.934             |
| Full               | ElasticNet   | 0.670   | 0.154  | 0.0694 | 0.360       | 0.863       | 0.182 | 0.241 | 0.934             |
| Full               | RandomForest | 0.665   | 0.156  | 0.0694 | 0.339       | 0.863       | 0.172 | 0.228 | 0.894             |
| Full               | XGBoost      | 0.671   | 0.162  | 0.0692 | 0.337       | 0.876       | 0.187 | 0.240 | 0.898             |
| Detection-adjusted | Logistic     | 0.630   | 0.123  | 0.0707 | 0.461       | 0.724       | 0.123 | 0.195 | 0.857             |
| Detection-adjusted | ElasticNet   | 0.630   | 0.123  | 0.0707 | 0.460       | 0.724       | 0.123 | 0.195 | 0.858             |
| Detection-adjusted | RandomForest | 0.625   | 0.126  | 0.0707 | 0.370       | 0.794       | 0.131 | 0.194 | 0.807             |
| Detection-adjusted | XGBoost      | 0.632   | 0.132  | 0.0705 | 0.442       | 0.745       | 0.128 | 0.198 | 0.802             |

All metrics are survey-weighted and are calculated from predictions for countries absent from their model-development fold. PPV, positive predictive value; PR AUC, precision-recall area under the curve; ROC AUC, receiver operating characteristic area under the curve.

## Calibration

External calibration plots showed that no model was uniformly calibrated across the full probability range (Fig. 3). Mean predicted probability was close to the observed weighted prevalence overall, but calibration slopes below 1 for all full models were consistent with predictions that were somewhat too extreme or with heterogeneity that was averaged across countries. XGBoost had a calibration intercept of -0.266 and a slope of 0.898, while logistic regression had intercept - 0.176 and slope 0.934. These findings reinforce the distinction between discrimination and probability reliability.

## Geographic and subregional transportability

Country-level results were substantially more heterogeneous than the pooled metrics. For the full predictor set, the median country ROC AUC was 0.646 (IQR 0.609 to 0.702) for XGBoost and 0.644 (IQR 0.606 to 0.697) for logistic regression. Median country PR AUC was 0.150 (IQR 0.095 to 0.219) for XGBoost and 0.149 (IQR 0.093 to 0.201) for logistic regression. This close central tendency masked large between-country and between-subregion variation.

Subregional XGBoost ROC AUC ranged from 0.798 in Central Asia and 0.746 in Western Asia to 0.570 in the pooled Sub-Saharan Africa grouping, 0.559 in Melanesia, and 0.512 in Micronesia (Figs. 4 and 5). Logistic regression showed a similar pattern, ranging from 0.798 in Central Asia to 0.510 in Micronesia. PR AUC was influenced by prevalence as well as discrimination; for example, Micronesia had comparatively high PR AUC near 0.20 despite near-random ROC discrimination because weighted CVD prevalence was 17.66%.

The similarity in subregional ranking across algorithms suggests that data transportability and population context contributed more to performance differences than the choice among the four model families. Importantly, pooling across all countries would have obscured poor discrimination in several Pacific and African subregions.

# Dependence on hypertension and diabetes information

Removing hypertension and diabetes reduced discrimination for every model (Table 3 and Fig. 6). The decrease in ROC AUC ranged from – 0.038 for XGBoost to -0.040 for elastic net and logistic regression. PR AUC decreased by approximately – 0.030 across all model families, while Brier score worsened by 0.0012 to 0.0013. The detection-adjusted XGBoost model retained the highest overall ROC AUC at 0.632 and PR AUC at 0.132, but its calibration slope fell to 0.802.

For logistic regression, removing hypertension and diabetes increased sensitivity from 0.350 to 0.461 but reduced specificity from 0.870 to 0.724, reflecting a lower development-selected threshold and weaker separation. PPV fell from 0.185 to 0.123 and F1 from 0.242 to 0.195. Similar patterns occurred for elastic net and XGBoost. These findings indicate that hypertension and diabetes carry substantial classification information, but they do not by themselves distinguish biological risk information from diagnostic opportunity.

Table 3  
Change after removing hypertension and diabetes

| Model        | Delta ROC AUC | Delta PR AUC | Delta Brier | Delta sensitivity | Delta specificity | Delta PPV | Delta F1 |
|--------------|---------------|--------------|-------------|-------------------|-------------------|-----------|----------|
| Logistic     | -0.040        | -0.031       | + 0.0013    | + 0.111           | -0.146            | -0.062    | -0.047   |
| ElasticNet   | -0.040        | -0.031       | + 0.0013    | + 0.100           | -0.139            | -0.058    | -0.047   |
| RandomForest | -0.040        | -0.030       | + 0.0012    | + 0.032           | -0.069            | -0.041    | -0.034   |
| XGBoost      | -0.038        | -0.030       | + 0.0013    | + 0.106           | -0.131            | -0.059    | -0.042   |

Values are detection-adjusted minus full-predictor performance. Negative discrimination values indicate lower performance after removal. A positive Brier difference indicates worse probabilistic error.

## Temporal transportability

Temporal transportability was weaker and country-specific (Table 4 and Fig. 7). In Mongolia, logistic regression trained on the 2013 survey and evaluated in 2019 achieved ROC AUC 0.634 and PR AUC 0.223; XGBoost achieved ROC AUC 0.600 and PR AUC 0.198. In Uganda, performance from 2014 to 2023 was lower across all algorithms: logistic regression ROC AUC was 0.579, and XGBoost ROC AUC was 0.568. In Viet Nam, logistic regression transported better, with ROC AUC 0.670 and PR AUC 0.185, while XGBoost achieved 0.658 and 0.152.

Calibration changed markedly over time. The logistic calibration slope was 0.636 in Mongolia and 0.427 in Uganda, but 1.396 in Viet Nam. Calibration intercepts were – 0.359, -0.713, and 1.377, respectively. Thus, even where there was moderate discrimination, the level and spread of predicted probabilities were not temporally stable. Removing hypertension and diabetes reduced temporal discrimination further in most country-model combinations.

Table 4  
Earlier-to-later temporal performance with the full predictor set

| Country  | Train year | Test year | Model        | ROC AUC | PR AUC | Brier  | Calibration slope |
|----------|------------|-----------|--------------|---------|--------|--------|-------------------|
| Mongolia | 2013       | 2019      | Logistic     | 0.634   | 0.223  | 0.1216 | 0.636             |
| Mongolia | 2013       | 2019      | ElasticNet   | 0.606   | 0.208  | 0.1217 | 0.614             |
| Mongolia | 2013       | 2019      | RandomForest | 0.627   | 0.211  | 0.1214 | 0.702             |
| Mongolia | 2013       | 2019      | XGBoost      | 0.600   | 0.198  | 0.1220 | 0.965             |
| Uganda   | 2014       | 2023      | Logistic     | 0.579   | 0.205  | 0.1299 | 0.427             |
| Uganda   | 2014       | 2023      | ElasticNet   | 0.580   | 0.205  | 0.1299 | 0.429             |
| Uganda   | 2014       | 2023      | RandomForest | 0.573   | 0.197  | 0.1300 | 0.411             |
| Uganda   | 2014       | 2023      | XGBoost      | 0.568   | 0.193  | 0.1304 | 0.410             |
| Viet Nam | 2015       | 2021      | Logistic     | 0.670   | 0.185  | 0.0809 | 1.396             |
| Viet Nam | 2015       | 2021      | ElasticNet   | 0.671   | 0.185  | 0.0808 | 1.385             |
| Viet Nam | 2015       | 2021      | RandomForest | 0.660   | 0.170  | 0.0812 | 1.497             |
| Viet Nam | 2015       | 2021      | XGBoost      | 0.658   | 0.152  | 0.0815 | 1.280             |

Each later survey was completely withheld from fitting, preprocessing, calibration, and threshold selection. PR AUC, precision-recall area under the curve; ROC AUC, receiver operating characteristic area under the curve.

## Interpretability

XGBoost feature importance was concentrated in a small number of predictors (Fig. 8). Age had the largest mean absolute SHAP value (0.215), followed by hypertension (0.191), DASH score (0.090), BMI (0.077), sex (0.060), high salt intake (0.056), smoking history (0.049), and physical activity (0.048). Diabetes had a much smaller global mean absolute SHAP value (0.0019) than hypertension, even though the joint removal of hypertension and diabetes materially reduced discrimination. SHAP ranking was interpreted as global predictive contribution and not as causal importance.

## Subgroup performance

Subgroup analyses also showed meaningful heterogeneity. For the full logistic regression, ROC AUC was 0.660 in females and 0.677 in males. By age, ROC AUC ranged from 0.595 among participants aged 18 to 29 years to 0.648 among those aged 50 to 59 years, while PR AUC increased from 0.073 in the youngest group to 0.230 among those aged 60 to 69 years, partly reflecting rising CVD prevalence with age. Rural participants had a logistic ROC AUC of 0.687 compared with 0.652 in urban participants. These differences were not interpreted as a formal fairness verdict because calibration, prevalence, and error rates must also be considered.

## Discussion

### Principal findings

In this 46-country benchmark of 187791 adults, no machine-learning model clearly dominated a transparent logistic baseline across all evaluation domains. XGBoost achieved the highest pooled ROC AUC and PR AUC, but the absolute gain over logistic regression was small: approximately 0.001 in ROC AUC and 0.008 in PR AUC. Logistic regression had a

slightly higher F1 score and a calibration slope closer to 1. These results challenge the common practice of presenting the highest pooled AUC as sufficient evidence that a more complex model is preferable.

These findings align with several recent systematic reviews that have examined machine learning for CVD prediction. Liu and colleagues conducted a systematic review and meta-analysis of ML models using electronic health records and found that while random forest and deep learning achieved higher pooled AUCs (0.865 and 0.847, respectively) than conventional risk scores (0.765), significant heterogeneity ( $I^2 > 99\%$ ) and methodological concerns limited clinical applicability [5]. Similarly, a systematic review by Damen and colleagues identified substantial variability in CVD prediction model performance across populations, underscoring the importance of external validation [6]. Our finding that XGBoost achieved only modest gains over logistic regression, approximately 0.001 in ROC AUC, contrasts with some earlier studies that reported larger improvements. This discrepancy likely reflects differences in study design: many previous comparisons used random participant splits, which may overstate performance by allowing models to exploit country-specific patterns during training. When we enforced country-grouped cross-validation, the performance gap narrowed considerably, reinforcing the argument that evaluation design substantially influences model rankings [4].

This pattern is also visible within the STEPS literature itself. In a related analysis restricted to twelve African countries drawing on the same survey programme, XGBoost achieved a considerably higher ROC AUC of 0.769 than the 0.671 observed here [15]. That analysis used a conventional 80:20 random split with synthetic minority oversampling rather than country-grouped external validation, and it evaluated a smaller and more geographically homogeneous set of countries. The gap between the two AUC estimates, despite both studies drawing on WHO STEPS data with broadly comparable predictors, illustrates how strongly evaluation design and geographic scope shape apparent model performance, and reinforces the caution urged throughout this benchmark against comparing pooled discrimination figures across studies that used different validation strategies.

The more consequential result was heterogeneity in transportability. Country-level and subregional performance varied far more than the pooled algorithm differences. Central Asia and Western Asia showed substantially stronger discrimination than several Pacific and African subregions. Temporal validation in Mongolia, Uganda, and Viet Nam further showed that a model developed in one survey wave cannot be assumed to retain the same discrimination or calibration in a later wave. These findings support model evaluation strategies that preserve geographic and temporal structure [4] and align with broader recommendations to characterise performance across settings rather than relying on a single aggregate estimate [24].

## Why complex models did not clearly outperform logistic regression

Several factors can explain the small gap between model classes. First, the predictor set consisted largely of conventional demographic, behavioural, and cardiometabolic variables. These may have strong monotonic or approximately additive associations with prevalent reported CVD, limiting the incremental value of nonlinear interactions. Second, survey harmonisation reduces some forms of measurement heterogeneity but also constrains the feature space to variables that are common enough to compare internationally. Third, the geographic evaluation design is intentionally harder than a random participant split because it tests unseen national contexts. Improvements that emerge from exploiting country-specific patterns during random cross-validation may disappear when the entire country is held out.

Our results are consistent with those of Li and colleagues, who found that the choice of ML algorithm had less influence on predictive performance than the underlying data structure and population characteristics [25]. Weng and colleagues similarly reported that machine learning offered only modest improvements over logistic regression for cardiovascular risk prediction using routine clinical data [26]. In the African context, Ng'ambi and colleagues recently applied ML models to WHO STEPS data from 12 African countries and found that XGBoost achieved an AUC of 0.769, a larger improvement over baselines than observed in our multicountry sample [15]. The difference may be attributable to their use of random

train-test splits rather than country-grouped external validation, as well as differences in country composition and outcome prevalence [16, 17].

The result does not imply that machine learning is unnecessary. XGBoost had the best pooled PR AUC and performed slightly better in several subregions. The implication is narrower: complexity should be justified by material gains in the performance domain relevant to the intended use. For low-prevalence CVD classification, a modest improvement in PR AUC may be useful if it improves case finding without degrading calibration or equity, but it should be weighed against interpretability, recalibration burden, computational cost, and the need for external validation.

## **Class imbalance and the importance of PR AUC**

Weighted CVD prevalence was 7.78%, meaning that accuracy and ROC AUC alone could provide an incomplete picture. The best full-model PPV was only about 0.19 at thresholds optimised for weighted F1, despite specificity near 0.87. This is not a contradiction; it is a direct consequence of low outcome prevalence and overlapping predictor distributions. Precision-recall evaluation is particularly informative under such imbalance because precision responds directly to false positive burden [9]. Class weighting improved the influence of CVD cases during model fitting without manufacturing synthetic survey participants, while threshold optimisation was isolated to development data.

The importance of precision-recall evaluation in imbalanced settings has been emphasised across multiple prediction modelling frameworks [10, 27]. Our findings extend this principle to multicountry survey data, where low prevalence combined with geographic heterogeneity makes ROC AUC particularly misleading. Other researchers have advocated for similar approaches in low- and middle-income country settings, where CVD prevalence is often low and false positives carry substantial resource implications [27].

## **Calibration and local updating**

Calibration was less stable than pooled discrimination. Full-model calibration slopes below 1 overall and marked temporal changes in intercept and slope suggest that transported probabilities should not be assumed to retain their original meaning. This is consistent with the wider prediction-modelling literature, which treats calibration as essential when probabilities are intended to support decisions [11, 12]. The heterogeneity observed here suggests that local recalibration should be considered before deployment, especially when CVD prevalence, diagnosis patterns, or population risk distributions differ from the development settings.

The calibration instability we observed echoes findings from Van Calster and colleagues, who identified calibration as the "Achilles heel" of predictive analytics [12]. In our temporal benchmarks, calibration slopes varied from 0.427 in Uganda to 1.396 in Viet Nam, far exceeding the variation in discrimination. This suggests that even when ranking performance is preserved, probability estimates may require local recalibration before deployment. Hippisley-Cox and colleagues recently demonstrated that recalibration can substantially improve the transportability of CVD risk algorithms across populations [28].

The current benchmark used a calibration partition drawn from development countries and deliberately did not recalibrate on the final test country. Future implementation studies could evaluate staged transport strategies such as intercept-only updating, intercept-and-slope recalibration, or limited local refitting using a prospectively defined local calibration sample. Such updating should be reported separately from untouched external performance because it answers a different question: adaptability rather than zero-shot transportability.

## **Detection-related predictors and reported CVD**

Hypertension and diabetes improved performance across all model families. Their removal reduced ROC AUC by about 0.04 and PR AUC by about 0.03. This is epidemiologically plausible because both are established cardiometabolic risk markers. However, for a self-reported prevalent outcome, these variables may also correlate with prior clinical contact, screening, treatment, and awareness. A participant whose hypertension or diabetes has been identified has had an opportunity to interact with the health system, which may also increase the probability that prior heart attack, angina, or stroke was recognised and reported.

The dependence on hypertension and diabetes status has implications for AI deployment in settings with limited healthcare access. In many African countries, hypertension and diabetes are underdiagnosed, meaning that models relying on these variables may systematically underestimate risk in populations with poor diagnostic coverage [29]. This is consistent with broader concerns about algorithmic fairness and the risk that AI tools may perpetuate rather than reduce health disparities. Ng'ambi and colleagues recently documented substantial heterogeneity in CVD prevalence, prevention, and treatment across African countries, highlighting the need for context-specific approaches to risk assessment [16]. Read alongside the continental hypertension surveillance reported by Nambiema and colleagues, who pooled 61 surveys from 37 WHO African Region countries and found wide country- and sex-level variation in hypertension awareness, treatment, and control between 2003 and 2022 [29], our subregional findings suggest that heterogeneity in CVD-related classification and heterogeneity in care-cascade performance may share common upstream drivers: health-system contact, screening coverage, and data infrastructure, rather than being independent statistical artefacts.

Accordingly, the sensitivity analysis should not be read as proving detection bias. It demonstrates that the model depends materially on variables that can carry both biological and healthcare-contact information. This distinction matters for responsible AI: a classifier can appear highly accurate because it reconstructs who has previously been diagnosed rather than identifying disease burden equivalently across populations with different access to detection. Future work could separate measured blood pressure and glucose from diagnosis and treatment variables and explicitly model recent healthcare contact.

## Geographic benchmarking through subregions

Interpreting the global analysis by 15 subregions made heterogeneity easier to see than a single worldwide average. For example, Central Asia showed ROC AUC near 0.80 for both logistic regression and XGBoost, whereas Micronesia showed ROC AUC near 0.51. These differences could arise from varying prevalence, risk-factor distributions, coding completeness, outcome reporting, sample size, or unmeasured health-system factors. They also demonstrate why algorithm rankings should not be assumed to generalise across world regions.

The subregional heterogeneity we observed, with ROC AUC ranging from 0.512 in Micronesia to 0.798 in Central Asia, is consistent with the findings of Debray and colleagues, who emphasised that prediction model performance should be summarised across settings rather than pooled [24]. Riley and colleagues have similarly argued that sample size and population heterogeneity are critical determinants of model transportability [30]. Our results suggest that global benchmarks should be disaggregated by region and that models intended for deployment in specific settings should be validated using locally representative data.

Subregional summaries should not be interpreted as fixed biological properties of regions. Some subregions contained only one or a few contributing countries, and the region labels were used as an interpretation layer rather than as model-development strata. Country-level external predictions remain the primary unit of transportability evidence. The subregional layer is useful because it reveals coherent performance patterns without allowing the largest countries to dominate the narrative.

## Temporal transportability

The three temporal benchmarks provide a distinct test that geographic cross-validation cannot replace. Uganda showed weak discrimination in 2023 when models were trained on 2014 data, while Viet Nam retained moderate discrimination from 2015 to 2021. Mongolia showed intermediate performance and a marked decline after detection-related variables were removed. These contrasts illustrate temporal dataset shift: even within a country, population composition, risk-factor patterns, healthcare access, measurement practice, and outcome awareness can change between survey waves.

Temporal dataset shift is increasingly recognised as a major challenge for clinical AI deployment. Our finding that models trained on earlier survey waves performed poorly in later waves, particularly in Uganda, mirrors observations from the COVID-19 prediction modelling literature, where model performance deteriorated rapidly over time [31]. This underscores the need for ongoing model monitoring and periodic recalibration in real-world deployment settings [32].

Temporal benchmarking was intentionally defined as earlier-to-later evaluation rather than randomly mixing both survey years. This preserves the direction of deployment and creates a more realistic test of whether a model developed from an earlier surveillance round can support future surveillance. The approach can be extended when additional countries accumulate repeated STEPS waves.

## **Interpretability and responsible use**

Age and hypertension dominated global XGBoost SHAP importance, followed by dietary score, BMI, sex, salt intake, smoking, and physical activity. The ranking is clinically plausible, but SHAP values explain the fitted model rather than the data-generating causal system [21]. High feature importance can result from correlation, measurement availability, coding patterns, and the model architecture. Therefore, SHAP should be used to audit what the model relies on, not to claim that changing the highest-ranked feature will cause a corresponding change in CVD.

The prominence of age and hypertension in SHAP importance rankings is broadly consistent with contemporary reviews of electronic-health-record-based cardiovascular AI research that highlight model interpretability, including SHAP-based approaches, as a priority for clinical translation [33]. As Lundberg and Lee emphasised, SHAP values explain the fitted model rather than the underlying causal structure [21]. This distinction is particularly important in the context of population health surveillance, where feature importance may be conflated with measurement availability and healthcare access.

This interpretability role is particularly valuable for cross-country AI. If a model's most influential variables are also those with large between-country differences in measurement or diagnosis, reviewers and implementers can identify potential mechanisms of transport failure. In this study, the prominence of hypertension together with the performance decline after removing hypertension and diabetes provides a coherent example of how interpretability and robustness analyses can be integrated rather than reported as separate decorative exercises.

## **Implications for digital public health and chronic disease surveillance**

For chronic disease management, population-health AI should be evaluated as part of an implementation pathway rather than as a leaderboard. A model intended for surveillance triage, case-finding support, or population segmentation should demonstrate acceptable minority-class performance, reliable probabilities, robustness to detection-related predictors, and external performance in the settings where it will be used. The present framework provides a practical sequence for doing so with multicountry survey data [22, 23, 34].

The findings also support maintaining transparent statistical baselines. Logistic regression was highly competitive and easier to inspect and recalibrate. A complex model should therefore be selected only when its incremental value is reproducible in external countries or later survey waves, or when it offers a specific operational advantage that justifies

the added complexity. This is consistent with TRIPOD + AI's emphasis on transparent model reporting and with PROBAST + AI's emphasis on evaluation quality and applicability [13, 14].

## Strengths and limitations

This study has several strengths. It used a large harmonised multicountry dataset, treated countries as external grouping units, incorporated survey weighting and country-year balancing, addressed class imbalance without primary synthetic oversampling, evaluated discrimination and calibration jointly, prespecified a detection-robustness analysis, added temporal evaluation, and linked interpretability to transportability. The same externally generated predictions were used for overall, country, subregional, and subgroup summaries, reducing opportunities for inconsistent evaluation.

Several limitations require emphasis. First, CVD was self-reported and not clinically adjudicated, so the models classify reported prevalent disease rather than incident future CVD. Second, the available harmonised predictor set did not contain all clinically relevant biomarkers, medication details, or healthcare-contact measures consistently across countries. Third, survey weights address population representation within their design assumptions but cannot remove measurement error, nonresponse bias, or cross-country differences in case ascertainment. Fourth, the country-grouped five-fold design evaluates unseen countries but is not identical to prospective external validation in an independently collected clinical cohort. Fifth, some subregions were represented by few countries, so regional summaries should not be overgeneralized. Sixth, only Mongolia, Uganda, and Viet Nam provided two eligible survey years for the requested temporal benchmark. Seventh, SHAP importance is model-specific and noncausal.

Finally, the primary analysis used prespecified rather than extensively tuned hyperparameters. This design reduced the risk of tuning to the external test structure and made model-family comparisons reproducible, but it may not represent the maximum attainable performance of every algorithm. Future work could perform fully nested tuning within development countries, provided that all tuning remains isolated from external countries and later survey waves.

## Conclusions

In global WHO STEPS data, XGBoost produced only modest pooled discrimination gains over logistic regression for prevalent self-reported CVD, while country, subregional, and temporal heterogeneity was much larger than the difference between algorithms. Removing hypertension and diabetes reduced performance across every model family, demonstrating meaningful dependence on detection-related information. Calibration was imperfect and changed over time. These findings argue against selecting public-health AI models by pooled ROC AUC alone.

A stronger benchmark evaluates class imbalance, PR AUC, calibration, threshold-dependent errors, predictor robustness, explainability, country-external performance, and earlier-to-later temporal transportability together. For chronic disease management and global surveillance, the preferred model should be the one that offers the most defensible balance of performance, calibration, interpretability, and transportability for its intended setting, not simply the most complex algorithm.

## Declarations

## Ethics approval and consent to participate

This study used de-identified secondary WHO STEPS survey data. Original STEPS surveys obtained country-specific ethics approval and participant informed consent under national procedures and WHO STEPS protocols. No new participant recruitment or contact was undertaken for this secondary analysis, and no identifiable participant information is reported.

## Consent for publication

Not applicable. No identifiable individual participant information is reported.

## Availability of data and materials

WHO STEPS survey data are available through the WHO noncommunicable disease microdata and country data mechanisms subject to the access and use conditions attached to each survey. The data and reproducibility materials generated during and/or analysed during the current study are available in the Zenodo repository at <https://doi.org/10.5281/zenodo.22860310>.

## Competing interests

The authors declare that they have no competing interests.

## Funding

No specific funding was received for this analysis.

Clinical trial number: not applicable.

## Authors' contributions

WFN conceptualised the study, developed the analytical framework, curated and harmonised the data, designed the statistical and machine-learning analyses, interpreted the results, and drafted the manuscript. CZ and LK contributed to methodological review, interpretation of the findings, and critical revision of the manuscript. All authors reviewed and approved the final manuscript.

## Acknowledgements

The authors acknowledge the World Health Organization, national ministries of health, country STEPS teams, field workers, and survey participants whose work made the multicountry surveillance data available for secondary analysis.

## References

1. World Health Organization. Cardiovascular diseases (CVDs) [Internet]. Geneva: World Health Organization; 2025 [cited 2026 Sep 9]. Cardiovascular diseases (CVDs). Available from: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).
2. Mensah GA, Fuster V, Murray CJL, Roth GA, Mensah GA, Abate YH, et al. Global Burden of Cardiovascular Diseases and Risks, 1990–2022. *J Am Coll Cardiol*. 2023;82(25):2350–473. 10.1016/j.jacc.2023.11.007.
3. World Health Organization. Noncommunicable Diseases Progress Monitor 2022 [Internet]. Geneva: World Health Organization. 2022. Available from: <https://iris.who.int/server/api/core/bitstreams/db06a907-0d4b-43d7-a49a-a7eb3623e489/content>
4. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Aroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*. 2017;40(8):913–29. 10.1111/ecog.02881.

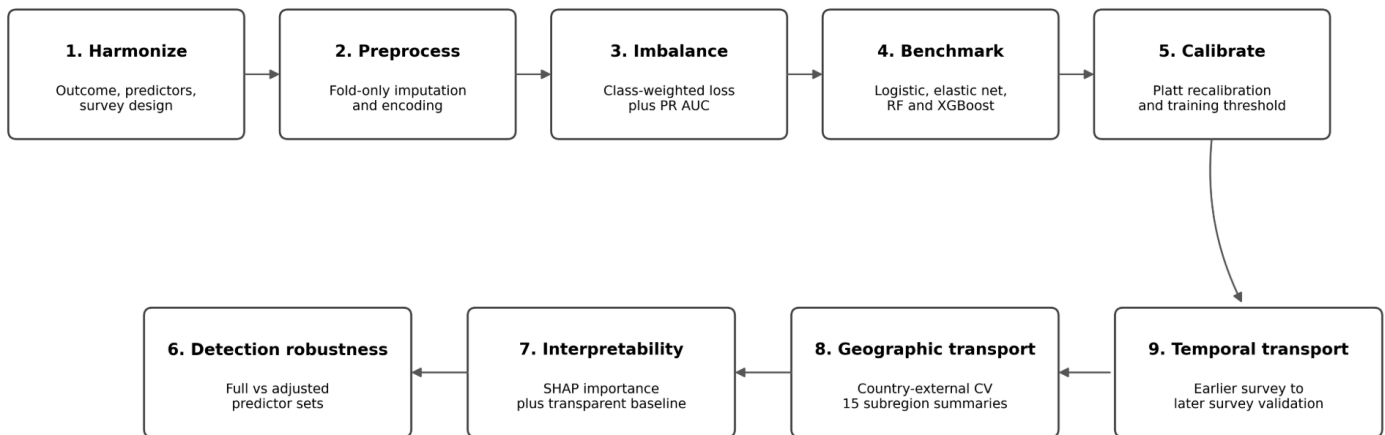
5. Liu T, Krentz A, Lu L, Curcin V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *Eur Heart J - Digit Health*. 2025;6(1):7–22. 10.1093/ehjdh/ztae080.
6. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;i2416. 10.1136/bmj.i2416.
7. World Health Organization. STEPwise approach to NCD risk factor surveillance (STEPS) [Internet]. Geneva: World Health Organization; [cited 2026 Sep 9]. STEPwise approach to NCD risk factor surveillance (STEPS). Available from: <https://www.who.int/teams/noncommunicable-diseases/surveillance/systems-tools/steps>
8. World Health Organization. STEPS Instrument [Internet]. Geneva: World Health Organization; [cited 2026 Sep 9]. STEPS Instrument. Available from: <https://www.who.int/teams/noncommunicable-diseases/surveillance/systems-tools/steps/instrument>
9. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. Brock G, editor. *PLoS ONE*. 2015;10(3):e0118432. 10.1371/journal.pone.0118432
10. Haibo He, Garcia EA. Learning from Imbalanced Data. *IEEE Trans Knowl Data Eng*. 2009;21(9):1263–84. 10.1109/TKDE.2008.239.
11. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology*. 2010;21(1):128–38. 10.1097/EDE.0b013e3181c30fb2.
12. ‘Evaluating diagnostic tests and prediction models’ of the STRATOS initiative, On behalf of Topic Group, Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. 10.1186/s12916-019-1466-7
13. 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. 10.1136/bmj-2023-078378.
14. 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. 10.1136/bmj-2024-082505.
15. Ng’ambi WF, Merzouki FA, Estill J, Keiser O, Orel E. Machine learning-based prediction of cardiovascular disease risk in Africa using WHO Stepwise Surveys: 2014–2019 [Internet]. *Epidemiology*; 2026 [cited 2026 Sep 9]. Available from: <http://medrxiv.org/lookup/doi/10.64898/2026.02.23.26346870> doi:10.64898/2026.02.23.26346870.
16. Ng’ambi WF, Estill J, Merzouki FA, Zyambo C, Banda JC, Beran D, et al. The prevalence, prevention, and treatment of cardiovascular diseases in Twelve African Countries (2014–2019): An analysis of the World Health Organisation STEPwise approach to chronic disease risk factor surveillance. Umer MF. editor *PLoS One*. 2026;21(3):e0320276. 10.1371/journal.pone.0320276.
17. Ng’ambi WF, Estill J, Muula AS, Keiser O. Use of mathematical modelling to estimate the impact of interventions and risk factors on cardiovascular diseases in Africa: A scoping review. *Mal Med J*. 2026;38(1):34–51. 10.4314/mmj.v38i1.6.
18. Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. *J Royal Stat Soc Ser B: Stat Methodol*. 2005;67(2):301–20. 10.1111/j.1467-9868.2005.00503.x.
19. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5–32. 10.1023/A:1010933404324.
20. Chen T, Guestrin C, XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]*. San Francisco California USA: ACM;

2016 [cited 2026 Sep 9]. pp. 785–94. Available from: <https://dl.acm.org/doi/10.1145/2939672.2939785>  
doi:10.1145/2939672.2939785.

21. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions [Internet]. arXiv; 2017 [cited 2026 Sep 9]. Available from: <https://arxiv.org/abs/1705.07874> 10.48550/ARXIV.1705.07874
22. Deo RC. Machine Learning in Medicine. *Circulation*. 2015;132(20):1920–30. 10.1161/CIRCULATIONAHA.115.001593.
23. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med*. 2019;380(14):1347–58. 10.1056/NEJMr1814259.
24. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;i6460. 10.1136/bmj.i6460.
25. Li Y, Sperrin M, Ashcroft DM, Van Staa TP. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. *BMJ*. 2020. m3919.
26. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? Liu B, editor. *PLoS ONE*. 2017;12(4):e0174944. 10.1371/journal.pone.0174944
27. Bisong E, Jibril N, Premnath P, Buligwa E, Oboh G, Chukwuma A. Predicting high blood pressure using machine learning models in low- and middle-income countries. *BMC Med Inf Decis Mak*. 2024;24(1):234. 10.1186/s12911-024-02634-9.
28. Hippisley-Cox J, Coupland CAC, Bafadhel M, Russell REK, Sheikh A, Brindle P, et al. Development and validation of a new algorithm for improved cardiovascular risk prediction. *Nat Med*. 2024;30(5):1440–7. 10.1038/s41591-024-02905-y.
29. Nambiema A, Agboyibor KM, Dangou JM, Antignac M, Diallo CB, Okeibunor JC, et al. Prevalence, Awareness, and Treatment of Hypertension in 37 African Countries. *JACC*. 2025;86(23):2263–88. 10.1016/j.jacc.2025.09.1600.
30. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;m441. 10.1136/bmj.m441.
31. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. 2020. m1328.
32. World Health Organization. HEARTS: Technical package for cardiovascular disease management in primary health care: Risk-based CVD management [Internet]. Geneva: World Health Organization. 2020. Available from: <https://iris.who.int/server/api/core/bitstreams/aef55845-a638-45d0-b816-70dd5bf8885a/content>
33. Dhingra LS, Shen M, Mangla A, Khera R. Cardiovascular Care Innovation through Data-Driven Discoveries in the Electronic Health Record. *Am J Cardiol*. 2023;203:136–48. 10.1016/j.amjcard.2023.06.104.
34. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA*. 2018;319(13):1317. 10.1001/jama.2017.18391.

## Figures

## CVD AI benchmarking framework



Output: multidomain evidence on discrimination, calibration, minority-class performance, robustness, interpretability and transportability

Primary data: WHO STEPS surveys | 48 countries with observed outcome data | 46 countries eligible for country-specific binary discrimination

Figure 1

CVD AI benchmarking framework. The framework separates harmonisation, fold-specific preprocessing, class-imbalance handling, statistical and machine-learning benchmarking, probability recalibration, detection-robustness analysis, interpretability, country-external geographic evaluation, and earlier-to-later temporal evaluation. RF, random forest; PR AUC, precision-recall area under the curve; CV, cross-validation

### External geographic performance with the full predictor set

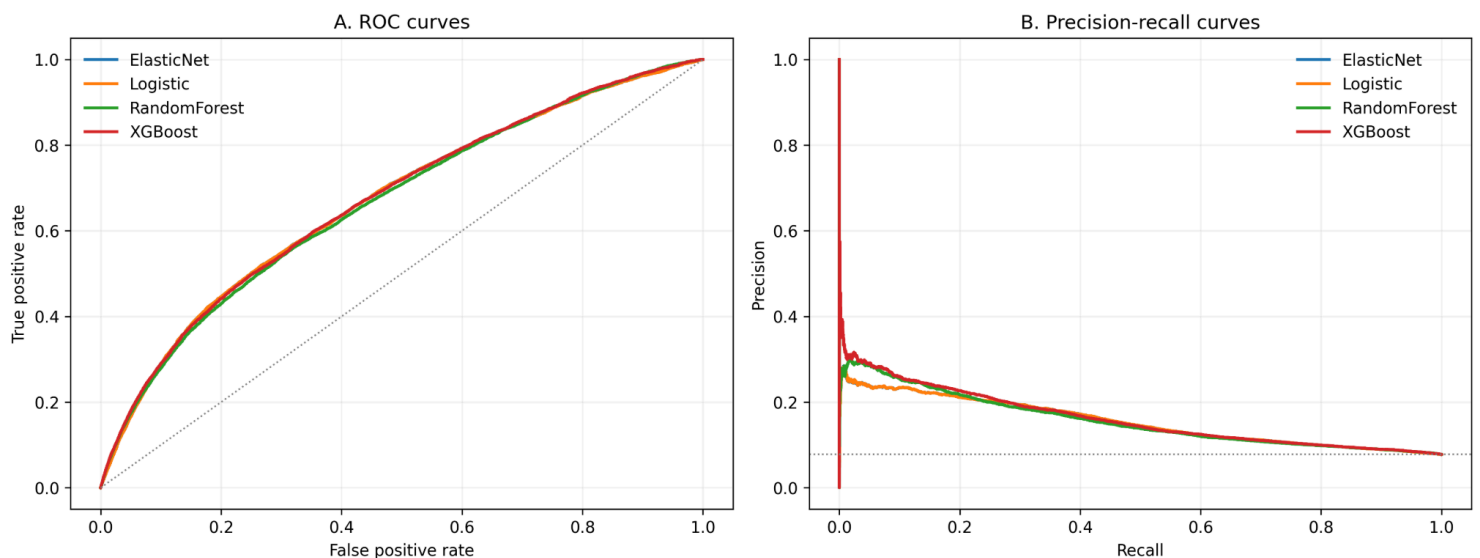


Figure 2

External geographic discrimination with the full predictor set. ROC and precision-recall curves combine predictions generated only when each country was absent from the corresponding development fold. The dotted horizontal line in the precision-recall panel represents weighted CVD prevalence

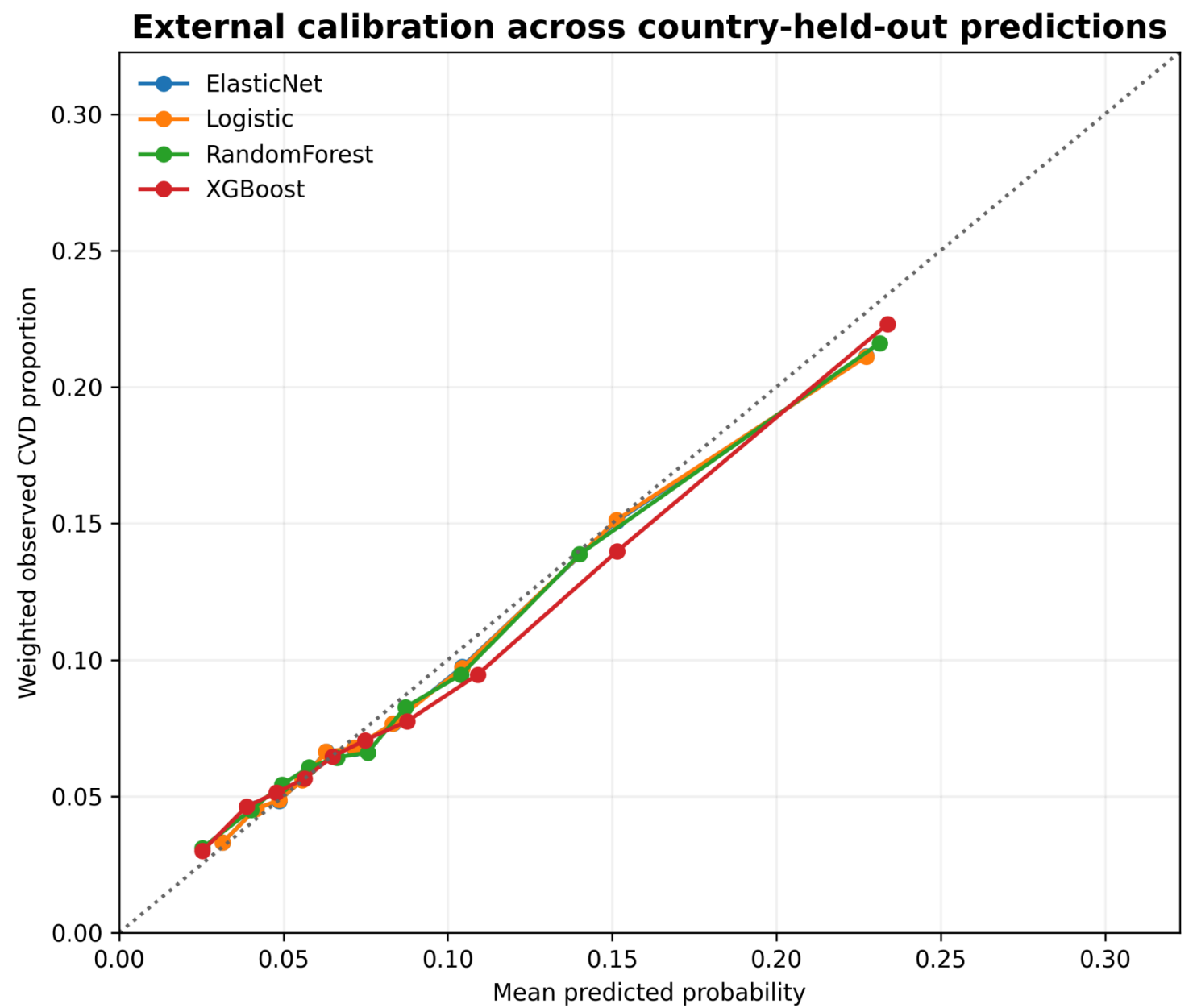


Figure 3

External calibration across country-held-out predictions. Points show survey-weighted observed CVD proportions against mean predicted probabilities within prediction deciles. The diagonal represents ideal calibration. Calibration is summarised across all held-out countries and should be interpreted alongside country-specific results

### A. ROC AUC



**Figure 4**

Subregional transportability of externally evaluated models: survey-weighted ROC AUC after country-grouped internal-external cross-validation. Values were calculated only from predictions generated when the contributing countries were absent from development data



**Figure 5**

Subregional transportability of externally evaluated models: survey-weighted precision-recall AUC after country-grouped internal-external cross-validation. PR AUC depends on outcome prevalence as well as ranking performance

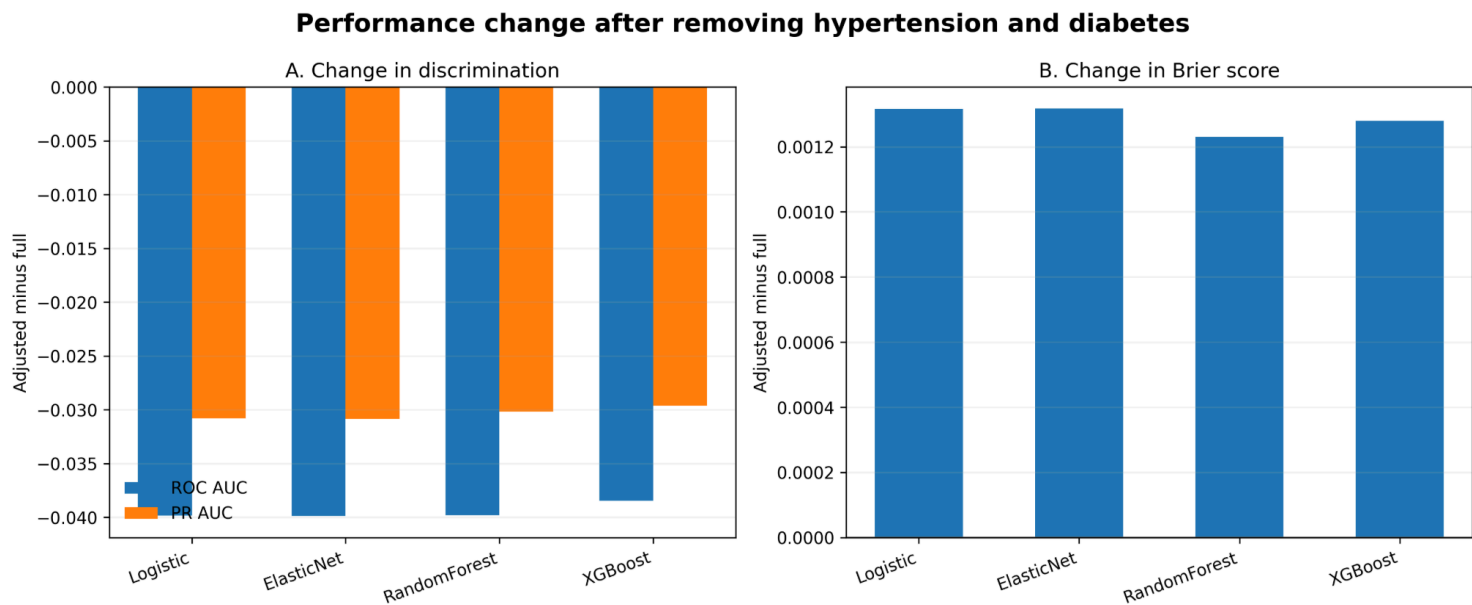


**Figure 6**

Performance change after removing hypertension and diabetes. Bars show detection-adjusted minus full-predictor performance. Negative ROC AUC and PR AUC values indicate loss of discrimination; positive Brier differences indicate increased probability error

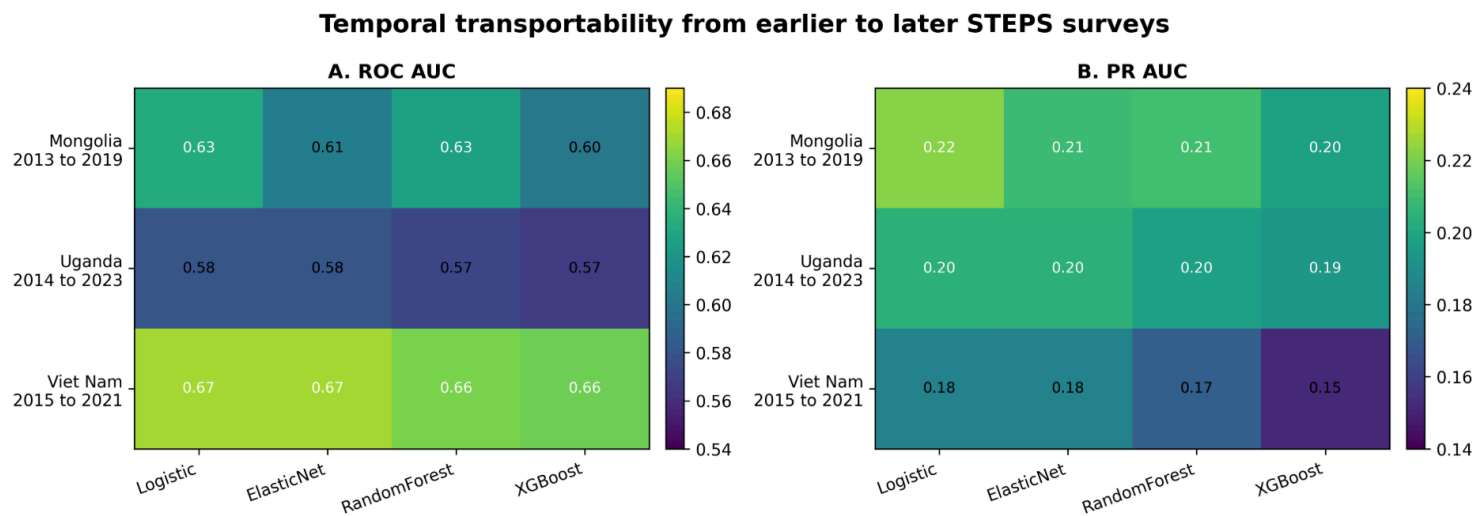


**Figure 7**

Temporal transportability from earlier to later surveys. Models were developed only in the earlier survey wave and evaluated in the later wave. The three pairs were Mongolia 2013 to 2019, Uganda 2014 to 2023, and Viet Nam 2015 to 2021

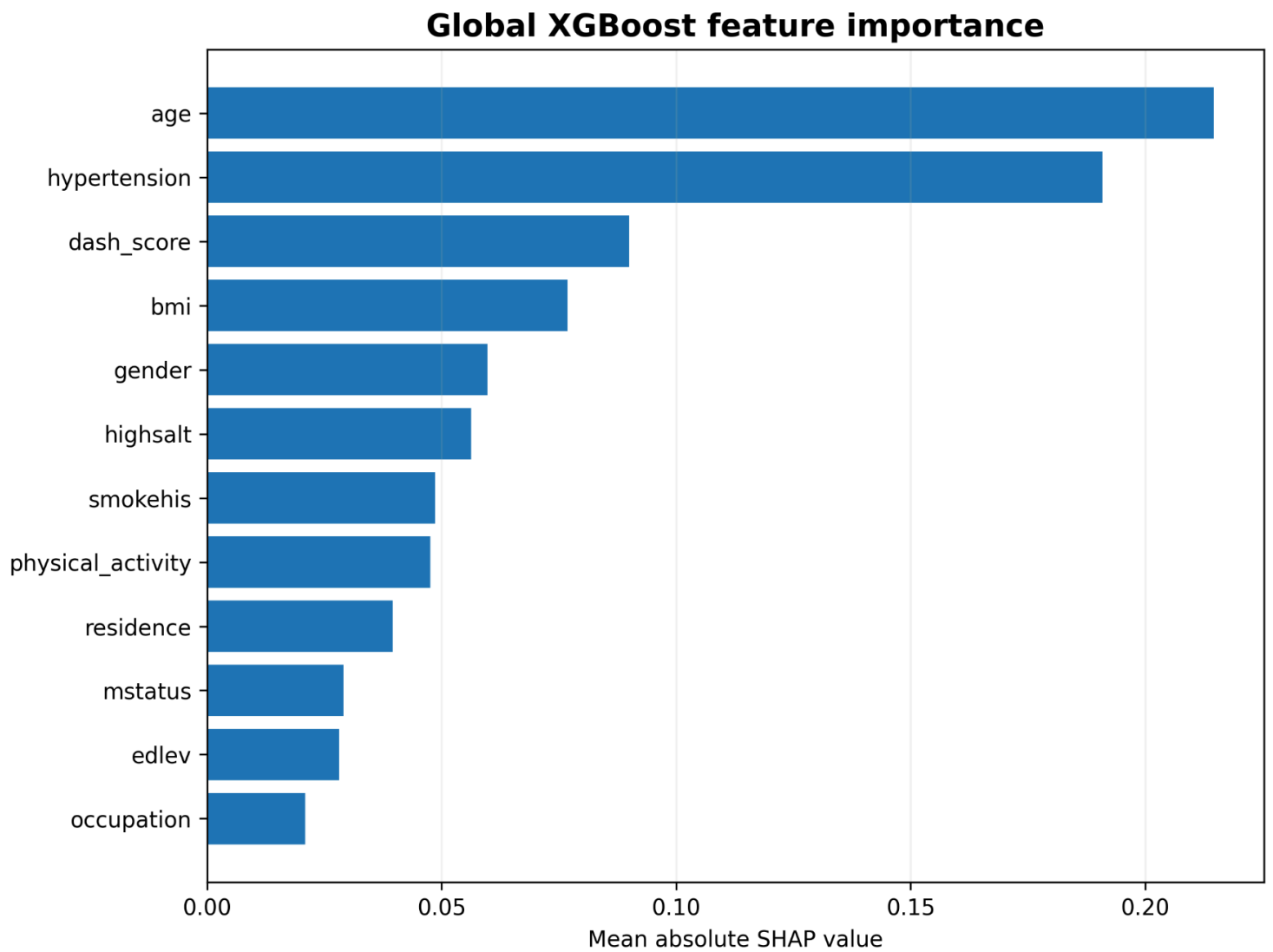


**Figure 8**

Global XGBoost feature importance. Mean absolute SHAP values summarise the magnitude of feature contributions in the full-predictor XGBoost model. Values do not imply causal effects or intervention effects

### Supplementary Files

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

- [AdditionalFile1SupplementaryTables.xlsx](#)
- [AdditionalFile2ReproducibleRanalysis.r](#)
- [AdditionalFile4Reproducibilityinformation.txt](#)
- [READMESUPPLEMENTARYFILES.txt](#)
- [AdditionalFile3CVDanalysisdataset.csv.gz](#)