International Journal of Analysis and Applications

Translational Evaluation of Interpretable Machine Learning for Cardiovascular Risk Prediction: Calibration Decomposition, Subgroup Audits, and Decision-Utility Analysis

Hathaichanok Chompoopong, Warawut Narkbunnum*

Mahasarakham Business School, Mahasarakham University, Mahasarakham, Thailand

**Corresponding author: warawut.n@acc.msu.ac.th*

ABSTRACT. Cardiovascular disease risk prediction models are often evaluated primarily by discrimination, although translational decision making depends on well-calibrated probabilities, subgroup reliability, and demonstrated clinical utility at actionable risk thresholds. This study conducted a translational evaluation of interpretable machine-learning models for heart disease prediction using a deployment-oriented framework integrating discrimination, calibration (including Murphy decomposition), explainability, subgroup stability, and decision-utility analysis via decision curve analysis. Using a large secondary dataset (308,774 observations; 19 predictors; prevalence 8.1%), models were trained with a stratified hold-out design and evaluated on a fixed test set. Histogram-based gradient boosting achieved the strongest discrimination (PR-AUC 0.3177; AUROC 0.8407) and strong probabilistic accuracy (Brier score 0.0633; ECE 0.0045), with Murphy decomposition indicating minimal reliability loss while preserving resolution. Explainability analyses (SHAP with PDP/ALE/ICE diagnostics) enabled transparent assessment of feature contributions and nonlinear effects relevant to plausibility and governance. Subgroup analyses indicated broadly stable discrimination but more variable calibration across age and self-reported general health strata, supporting the need for subgroup-aware monitoring. Decision curve analysis demonstrated positive net benefit relative to treat-all and treat-none strategies across screening-relevant thresholds (0.05–0.15), with workload trade-offs informing threshold selection for practice.

1. Introduction

Cardiovascular disease (CVD) remains a leading cause of preventable death and disability worldwide; consequently, screening and triage programs increasingly rely on risk-prediction models that estimate an individual's absolute probability of future events [1]. In routine practice,

Received Mar. 23, 2026

2020 Mathematics Subject Classification. 62P10, 62H30, 68T05, 62G20.

Key words and phrases. cardiovascular disease prediction; explainable machine learning; probability calibration; Murphy decomposition; decision curve analysis; clinical utility.

https://doi.org/10.28924/2291-8639-24-2026-202

ISSN: 2291-8639

probability estimates, not mere rankings, drive actions because thresholds on predicted risk trigger referrals and preventive therapy, thereby allocating finite resources [1], [2], [3], [4]. When probabilities are biased, even models with excellent discrimination can harm patients; contemporary guidance therefore stresses that reporting AUC without calibration is inadequate [5], [6]. Empirical evaluations of widely used cardiovascular calculators show frequent miscalibration, often resulting in material overestimation or underestimation of absolute risk, which can alter treatment eligibility at guideline thresholds [7], [8], [9]. Decision-analytic work further shows that miscalibration degrades net benefit and can render a high-AUC model net-harmful at clinically relevant thresholds [6], [10]. Hence, rigorous external validation, calibration assessment, and where necessary, model updating or recalibration should precede deployment, in line with TRIPOD+AI reporting and PROBAST risk-of-bias guidance [3], [11], [12], [13], [14].

Despite these concerns, much of the medical machine learning (ML) literature still emphasizes discrimination metrics, while calibration, decision-analytic utility, and subgroup stability are underreported [1], [15], [16]. Models may appear well calibrated in development yet become miscalibrated after transport due to differences in case mix, predictor measurement, or predictor–outcome relationships; such shifts are well documented in external-validation methodology and measurement-heterogeneity studies [17], [18], [19]. Simply adjusting the intercept ("recalibration-in-the-large") may not correct decision-relevant miscalibration and can yield suboptimal choices at clinically used thresholds [4], [20]. Moreover, miscalibration and performance can differ systematically across demographic groups: for instance, evaluations of cardiovascular and stroke risk models show subgroup-specific performance gaps, and a recent fairness analysis highlighted unequal calibration of the Framingham Offspring diabetes risk score underestimating risk in Black individuals and overestimating risk in White individuals [21], [22]. A comprehensive evaluation should therefore verify that predicted probabilities correspond to observed outcomes across relevant subgroups and ensure that any recalibration strategy preserves fairness and clinical usefulness [3], [15], [23].

Beyond probability accuracy, clinicians need to know whether a model will improve decisions relative to existing practice. Decision curve analysis (DCA) addresses this need by plotting net benefit across risk thresholds and comparing model-guided strategies to "treat all" and "treat none" policies [2], [10]. Recent primers emphasize that DCA identifies threshold regions where a model is beneficial and where its use would be detrimental, even when discrimination and calibration appear similar [24], [25]. Usage has expanded markedly, and contemporary reporting guidance encourages authors to assess clinical utility commonly via DCA alongside discrimination and calibration [3], [26]. Incorporating DCA into evaluation frameworks helps align evidence with real clinical decisions.

Interpretability and fairness add further layers to this evaluation. Subgroup performance differences may arise from heterogeneity in risk factors, data quality, or systemic biases.

Calibration fairness requires predicted risks to reflect observed outcomes comparably across demographic groups; unequal calibration constitutes unfairness and can systematically over- or under-estimate risk [22]. In parallel, interpretable models and post-hoc explanation methods, such as SHAP, help clinicians understand which features drive predictions at both global and individual levels, thereby supporting transparency, trust, and post-deployment monitoring [27].

Research objective. This study aims to determine whether interpretable ML models can deliver calibrated, clinically useful, and equitable cardiovascular risk estimates across a realistic window of screening thresholds. By implementing a full evaluation stack combining discrimination metrics, probability calibration diagnostics (Brier score, expected calibration error, reliability curves, and Murphy decomposition), clinical usefulness via decision curve analysis, subgroup fairness/stability assessment, and interpretability, we seek to provide evidence suitable for governance and audit, bridging the gap between statistical performance and real-world deployment.

Research Questions and Hypotheses

RQ1/H1 (Discrimination and calibration): Post-hoc calibration reduces expected calibration error and Brier score, lowering the Murphy Reliability component without materially degrading Resolution.

RQ2/H2 (Clinical usefulness): Models yield higher net benefit than treat-all/none across a prespecified threshold window (e.g., 0.05–0.15) and admit an actionable operating threshold that offers clinically meaningful trade-offs.

RQ3/H3 (Subgroup fairness and stability): Differences in discrimination, calibration, and net benefit between key subgroups (e.g., sex, age) are small, within tolerance bands, and supported by overlapping confidence intervals, indicating fair calibration and performance

2. Literature Review

2.1 From Ranking to Calibrated Probabilities

Many machine learning (ML) studies in medicine still prioritize discrimination metrics such as the area under the receiver operating characteristic curve (ROC AUC) or the precision–recall curve (PR AUC), which reflect only a model's ability to rank individuals relative to each other. Yet clinical decision-making depends on absolute probabilities of whether a predicted risk crosses a threshold that prompts further testing, preventive pharmacotherapy, or lifestyle intervention. Calibration, defined as the agreement between predicted probabilities and observed event frequencies, therefore represents a central property of clinical readiness [1], [28]. Recent methodological reviews highlight that reporting discrimination alone can obscure critical miscalibration: a model with a high AUC may systematically over- or underestimate risk, thereby inducing inappropriate treatment or missed prevention opportunities [28]. Such discrepancies

have been documented in several cardiovascular risk calculators that maintain acceptable discrimination but exhibit calibration drift in contemporary populations, altering treatment eligibility at guideline thresholds [7], [8].

Two summary statistics are widely used to quantify overall calibration: the Brier score, which measures the mean squared deviation between predicted probabilities and observed outcomes, and the Expected Calibration Error (ECE), which aggregates absolute deviations across probability bins [28], [29], [30], [31], [32]. Because ECE is sensitive to bin selection, guidelines recommend reporting it together with graphical diagnostics [28]. To gain a deeper understanding of where calibration errors originate, the Murphy decomposition partitions the Brier score into Uncertainty (irreducible outcome variability), Reliability (systematic miscalibration), and Resolution (the ability to distinguish high- and low-risk individuals). Originally developed in meteorology [30], [31], [33], this decomposition has recently been adopted in medical ML to disentangle whether recalibration procedures improve probability reliability without degrading resolution (Huang et al., 2020). In parallel, reliability curves, also known as calibration plots, visually compare predicted and observed probabilities across the risk spectrum. Modern recommendations advocate displaying confidence bands around these curves, particularly in clinically relevant threshold regions, to facilitate judgment of calibration adequacy [34].

Calibration must also be transportable across populations. A model that is well-calibrated in its derivation cohort may perform poorly elsewhere due to differences in baseline risk, case-mix, or measurement processes [17], [18], [19]. Reflexive recalibration, such as simply updating the intercept, may fail to correct bias and, in some cases, introduce new errors that degrade decision quality at relevant thresholds [20], [35]. Contemporary research thus calls for causal interpretations of miscalibration sources (e.g., shifts in predictor–outcome relations or measurement heterogeneity) and for evidence-based recalibration or model-updating strategies that maintain clinical utility [20], [35]. These insights directly motivate the methodological framework described in Chapter 3 of this study, which combines quantitative calibration metrics (Brier, ECE, Murphy decomposition, and reliability curves with confidence intervals) with decision-analytic and subgroup-stability assessments to ensure transparency and clinical readiness.

2.2 Decision-Curve Analysis for Clinical Value

Even a model with excellent discrimination and calibration might not be clinically useful. Decision curve analysis (DCA) assesses whether a prediction model improves decision-making compared to default strategies by estimating net benefit across a range of risk thresholds and comparing model-guided strategies with “treat all” and “treat none” policies [2], [10]. Recent guidance explicitly recommends evaluating clinical utility often, using DCA alongside discrimination and calibration, when developing or validating prediction models [1], [3]. Net benefit formalizes the trade-off between true positives and false positives, weighted by clinical

consequences, ensuring that the evaluation reflects patient preferences and resource constraints, which are limitations not addressed by discrimination and calibration metrics alone [6], [24], [36]. Decision curves display net benefit versus threshold probability; when a model's curve lies above both defaults within relevant thresholds, it indicates greater expected clinical benefit, whereas curves below indicate potential harm [2], [24]. Methodological advances in 2024–2025 further enhance DCA such as Bayesian formulations and approaches for using summary data, while diverse recent applications underscore its practical value across settings, from prehospital severe-trauma triage and emergency department workflows to cardiovascular prognostication and multi-use digital health models [36], [37], [38], [39], [40], [41].

DCA is particularly valuable in screening and triage contexts common in cardiovascular care, where disease prevalence is low and interventions carry nontrivial risks or costs. In such settings, decision curves identify threshold windows (e.g., 0.05–0.15) where model-guided decisions offer positive net benefit and support the selection of an operating threshold by visualizing sensitivity specificity trade-offs at clinically plausible cut-points [10], [25]. Importantly, two models with similar AUC and apparently acceptable calibration can produce markedly different decision curves, revealing divergent clinical utility that conventional performance statistics would miss [4], [6]. Adoption and guidance have expanded rapidly. Recent methodological updates and reporting recommendations encourage assessing clinical utility (often via DCA) alongside discrimination and calibration, and empirical surveys show the method's widespread uptake in the literature [1], [3], [42]. Incorporating DCA within evaluation frameworks, therefore, helps ensure that statistical performance translates into actionable clinical benefit.

2.3 Subgroup Fairness and Stability

Model performance can differ across demographic groups due to variations in underlying risk distributions, predictor prevalence, and data quality. A fairness-aware evaluation, therefore, assesses whether discrimination, calibration, and decision utility are consistent across predefined subgroups (e.g., sex, age, and race/ethnicity) [1], [43]. One specific criterion is calibration fairness; predicted probabilities should align with observed event rates similarly across groups. Uneven calibration indicates that one group consistently receives higher or lower risk estimates than appropriate [22]. Empirical evidence shows such disparities: for instance, the Framingham Offspring score has been found to underestimate type-2 diabetes risk in Black individuals while overestimating it in White individuals, demonstrating ethically significant miscalibration by subgroup [22], [44]. Beyond calibration, fairness summaries should include differences between groups in standard metrics, such as ΔAUC for discrimination, ΔECE for calibration error, and Δ net benefit for decision utility, reported with confidence intervals to reflect statistical uncertainty [43].

Performance instability often results from small subgroup samples or diverse risk factors. Recent guidance recommends using bootstrap methods to measure uncertainty around subgroup metrics and warns that wide confidence intervals may prevent definitive fairness conclusions, encouraging transparent reporting of precision and, if needed, delaying claims of equivalence [1]. When disparities are identified, mitigation strategies include group-specific recalibration or updating, reweighting during training, and applying fairness constraints, chosen with consideration of clinical trade-offs and potential effects on overall utility [22]. Because deployment environments change, fairness should be monitored after implementation: population shifts and changes in data quality can cause fairness drift, requiring surveillance plans alongside ongoing performance evaluation [13], [43].

2.4 Interpretability for Plausibility and Governance

Clinicians and patients are more likely to trust and adopt predictive models when they understand how predictions are made. Early work on interpretable models like generalized additive models (GAMs) showed that transparent functional components allow for clinical plausibility checks and effective communication [45]. Later healthcare applications proved that understandable models could achieve competitive accuracy while remaining auditable, for example, GA2M models used for pneumonia risk and readmission [46]. Post-hoc explanation techniques then emerged to interpret complex models at the individual level, notably LIME and Shapley-value–based methods [47], [48], [49]. Specifically, SHAP provides local and global attributions with a solid foundation in additivity; for tree ensembles, TreeSHAP offers consistent, interaction-aware explanations that support clinical review of feature effects [48]. Complementing SHAP, Accumulated Local Effects (ALE) plots estimate marginal effects while considering feature dependence, avoiding the extrapolation artifacts seen in partial-dependence plots and better reflecting clinically relevant ranges [50], [51].

Interpretability also enhances governance. Visual summaries (e.g., SHAP summary and dependence plots; ALE profiles) help auditors spot misleading associations, potential signs of data leakage or bias, and compare inferred relationships with established domain knowledge [52], [53]. Current views also stress choosing inherently interpretable models whenever possible or combining transparent models (e.g., GAMs with interaction terms, monotonic constraints) with post-hoc explanations to verify behavior, thus boosting plausibility and stakeholder trust [48], [54]. Collectively, these tools aid not just in model development but also in lifecycle governance supporting pre-deployment audits, documentation, and ongoing monitoring of feature–risk relationships as populations and practices change [53].

2.5 Synthesis: Why This Evaluation Stack

The literature indicates that a credible assessment of risk prediction models must go beyond discrimination. Calibration diagnostics, including the Brier score, ECE, reliability (calibration) curves with confidence bands, and the Murphy decomposition measure probability

accuracy and pinpoint sources of error, helping prevent the use of models with biased absolute risks even if their rankings seem strong [1], [28], [30], [31]. Decision curve analysis (DCA) complements these methods by evaluating whether model-guided decisions provide more net benefit than default strategies across relevant thresholds; notably, models with similar AUCs and acceptable calibration can differ in clinical usefulness on DCA [6], [10], [24]. Additionally, subgroup fairness and stability require that discrimination, calibration, and decision utility be consistent across predefined demographic groups; disparities in calibration or net benefit among groups risk reinforcing inequalities [22]. Lastly, interpretability tools such as SHAP and ALE, when used alongside inherently interpretable models, allow for plausibility checks, detection of false associations, and transparent communication with stakeholders [48], [51], [54].

Despite these advantages, the components are seldom integrated into published studies: meta-research consistently reveals a narrow focus on discrimination (and occasionally a single calibration plot) with under-reporting of clinical utility and subgroup analyses [1], [15]. Empirical evaluations in cardiovascular risk further demonstrate that miscalibration can change treatment eligibility at guideline thresholds, while subgroup-specific miscalibration, such as underestimation in Black individuals and overestimation in White individuals, highlights ethical and clinical consequences [7], [8], [22]. Decision curves clearly show that similar AUCs do not equate to similar clinical value, reinforcing the need for utility-focused evaluation [4], [24].

Accordingly, we implement a comprehensive evaluation framework that combines discrimination and calibration diagnostics (including Murphy decomposition), DCA, subgroup fairness and stability audits, and interpretability. This integrated approach aligns statistical performance with real-world decisions, equity concerns, and transparency, and directly influences the methodological choices in Chapter 3 (metrics, graphical diagnostics with uncertainty, subgroup reporting with CIs, and governance-oriented explanations). In short, our goal is to determine whether interpretable ML can provide calibrated, clinically useful, and equitable cardiovascular risk predictions.

3. Methodology

3.1 Study Design and Data Source

This study employed a retrospective secondary data design using a publicly available health survey dataset that included demographic, behavioral, and clinical risk factor variables relevant to cardiovascular outcomes. The analytic objective was to develop and evaluate probabilistic risk prediction models under a deployment-oriented evaluation framework, emphasizing discrimination, calibration, clinical utility, subgroup stability, and interpretability. Because the dataset is de-identified and publicly accessible, the analysis did not involve direct human-subject contact and was conducted in accordance with applicable ethical guidance for secondary analysis of anonymized data.

3.2 Outcome Definition and Eligibility Criteria

The primary outcome was a binary indicator of cardiovascular disease status (CVD), defined according to the dataset's standard coding scheme. Records were included if they contained valid outcome labels and complete values for the prespecified predictors after preprocessing. Exclusion criteria were applied to remove records with non-informative labels, inconsistent codes, or missingness exceeding the prespecified threshold. The final analytic cohort was summarized by sample size, outcome prevalence, and key baseline characteristics.

3.3 Predictors and Preprocessing Pipeline

Predictor candidates were selected based on clinical relevance and prior evidence in CVD risk modeling, including demographic attributes, lifestyle factors, and comorbidity-related measures. Preprocessing comprised: (i) standardization of variable encodings (categorical mapping and ordinal ordering where appropriate), (ii) handling of missing values using a consistent strategy to avoid information leakage, and (iii) removal of duplicates and logically inconsistent entries. All preprocessing steps were implemented as a reproducible pipeline, applied identically across splits, ensuring that transformations learned from training data were applied to validation/test data without contamination.

3.4 Data Splitting and Validation Strategy

To ensure an unbiased estimate of generalization performance, the dataset was partitioned into mutually exclusive training and test subsets using a fixed random seed. Model selection and hyperparameter tuning were conducted only within the training subset, using either a validation split or cross-validation as defined in the pipeline. The held-out test set was used exclusively for final evaluation, including calibration diagnostics and decision-analytic assessment. This separation was maintained throughout to mitigate optimistic bias and inadvertent leakage.

3.5 Model Development and Algorithmic Configuration

Multiple machine-learning algorithms were evaluated under a common framework, including Logistic Regression, EBM, LightGBM, XGBoost, HGB, and a stacked logistic ensemble. These models were selected because they are widely used for structured tabular healthcare data and can capture nonlinear relationships and feature interactions. Hyperparameters were configured to balance predictive capacity and generalizability, using a moderate learning rate, a bounded tree depth, and a fixed number of estimators. To address class imbalance, cost-sensitive learning was incorporated by adjusting class weights (e.g., scale_pos_weight) based on the training-set class ratio. Training employed regularization controls (e.g., subsampling, column sampling, and/or L1/L2 penalties depending on the configured setting) and early stopping when applicable. A baseline benchmark model (e.g., logistic regression or an alternative tree-based model) was also included to contextualize performance under the same evaluation protocol.

3.6 Probability Calibration

Because downstream clinical decisions rely on absolute risk estimates, post-hoc calibration was applied to transform raw model scores into well-calibrated probabilities. Calibration methods (e.g., Platt scaling/sigmoid or isotonic regression) were trained on validation data only, then applied to the test set to avoid optimistic calibration. Calibration quality was assessed using both scalar metrics and visual diagnostics, including the Brier score, expected calibration error (ECE), and calibration curves with uncertainty quantification where available. In addition, the Murphy decomposition of the Brier score was computed to separate reliability (systematic miscalibration) from resolution (risk stratification), thereby enabling a more diagnostic interpretation of probabilistic performance.

3.7 Predictive Performance Metrics

Discrimination was assessed using ROC-AUC and PR-AUC to capture ranking performance under potential class imbalance. Threshold-dependent metrics (e.g., sensitivity, specificity, PPV, NPV, and F1-score) were computed at prespecified operating points, including clinically plausible thresholds or thresholds optimized on the training/validation data. Uncertainty was quantified via resampling procedures (e.g., bootstrap confidence intervals) performed on the held-out test set to provide robust interval estimates for key performance measures.

3.8 Subgroup Stability and Fairness Assessment

To evaluate stability and potential inequities, performance was stratified across clinically relevant demographic subgroups (e.g., sex, age bands, and other available attributes). For each subgroup, discrimination, calibration (including calibration curves and ECE/Brier-based summaries), and decision-analytic utility were reported. Differences across groups were interpreted with attention to sample size, shifts in prevalence, and uncertainty intervals, prioritizing calibration within groups as a core criterion for fairness in probabilistic clinical risk estimation.

3.9 Interpretability and Model Behavior Diagnostics

Interpretability analyses were conducted to support plausibility checks and governance-oriented auditing. Global and local explanation methods were applied to characterize feature contributions and heterogeneity of effects. Specifically, SHAP-based summaries were used to quantify global feature importance and to inspect individual-level attributions. Functional diagnostics, including partial dependence plots (PDP), individual conditional expectation (ICE), and accumulated local effects (ALE), were used to evaluate nonlinear patterns, interaction structures, and localized sensitivities. These diagnostics were treated as complementary: SHAP provided contribution-based explanations, whereas PDP/ICE/ALE provided functional assessments of model response across the predictor space.

3.10 Clinical Usefulness: Decision Curve Analysis

Clinical usefulness was evaluated using decision curve analysis (DCA), which estimates net benefit across a range of threshold probabilities corresponding to plausible intervention or referral policies. The model was compared against "treat all" and "treat none" strategies and, when relevant, against baseline comparators. To enhance interpretability for implementation, net benefit was also translated into workload-oriented quantities (e.g., avoidable unnecessary interventions per 1000 patients screened) under representative thresholds.

3.11 Software and Reproducibility

All analyses were conducted in a reproducible computational environment using standard Python machine-learning libraries. Versioning details, fixed random seeds, and deterministic splits were documented to facilitate replication. Code and derived artifacts (e.g., model configurations, calibration objects, prediction outputs, and evaluation tables/figures) were organized as a structured pipeline, with public links provided in the Data Availability Statement.

4. Result

4.1 Data Overview and Outcome Prevalence

We analyzed a large secondary observational dataset comprising 308,854 records and 19 predictors (7 numeric; 12 categorical). The binary outcome was heart disease, with "Yes" as the positive class. The sample included 24,971 positives and 283,883 negatives, corresponding to an outcome prevalence of 0.08085 (95% Wilson CI: 0.07989–0.08182; Table 1). Data integrity checks identified 80 complete duplicate rows, which were removed during preprocessing, yielding a final analytic cohort of 308,774 observations used for model development and evaluation (Table 1). No column-wise missingness was observed. Descriptive statistics for numeric variables and frequency distributions for categorical variables are provided in the Appendix to support transparent reporting.

Table 1. Data overview.

n_rows	308,854
n_cols	19
target_col	Heart_Disease
positive_label	Yes
n_positive	24,971
n_negative	283,883
prevalence	0.08085
prev_ci95_lo	0.079894
prev_ci95_hi	0.081817
n_numeric	7
n_categorical	12
n_fullrow_duplicates	80

4.2 Train/Test Partition

A fixed stratified hold-out split was used to estimate out-of-sample performance, with test_size = 0.20 and seed = 42. Sample sizes and class prevalence by split are summarized in Table 2, and the partition flow is illustrated in Figure 1. The resulting split preserved the low-prevalence structure across subsets, supporting stable estimation of discrimination, calibration, and decision-analytic utility.

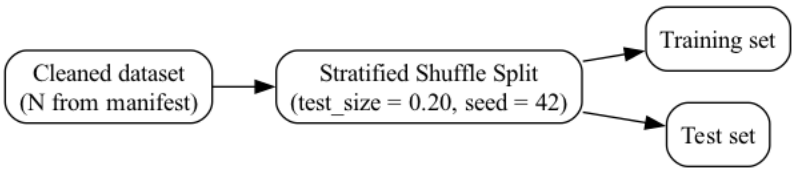

Figure 1. Train and test partition flow.

Table 2. Sample sizes and class prevalence for the stratified hold-out split.

Set	n	Positive (n)	Prevalence	Prev_CI95_L	Prev_CI95_U
All	308,774	24,971	0.08087	0.07992	0.08184
Train	247,019	19,977	0.08087	0.07980	0.08195
Test	61,755	4,994	0.08087	0.07874	0.08304

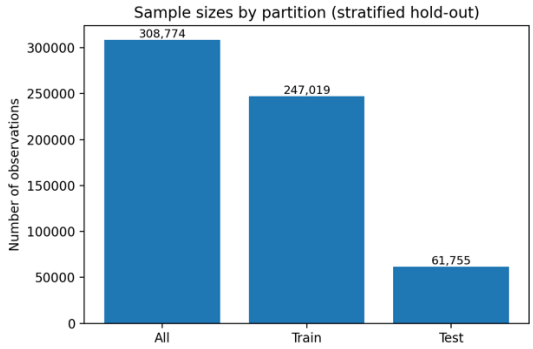

Figure 2. Sample sizes count.

4.3 Model Discrimination

Discrimination was assessed on the fixed held-out test set using PR-AUC as the primary metric (given the low prevalence) and ROC-AUC as a secondary metric. Summary results are reported in Table 3, with uncertainty visualized in Figure 3. Across models, performance clustered tightly. The histogram-based gradient boosting (HGB) model achieved the strongest discrimination with PR-AUC = 0.3177 and ROC-AUC = 0.8407, followed closely by the stacked logistic ensemble and other boosted learners (Table 3). Pairwise comparisons using DeLong tests for AUROC indicated no statistically significant differences among the top-performing models on this split (e.g., HGB vs. logistic: p = 0.9272; HGB vs. stacking-logit: p = 0.9197; Table 4).

Because discrimination can mask probability miscalibration, we additionally report the Brier score in Table 3. Linear baselines exhibited substantially higher Brier values (approximately

0.172) relative to tree/boosting and additive learners (approximately 0.063–0.064), indicating that high-ranking performance alone did not guarantee reliable probability estimates.

Table 3. Test-set discrimination and Brier score.

model	roc_auc	pr_auc	brier
hgb	0.8407	0.3177	0.0633
stack_logit	0.8392	0.3166	0.0634
ebm	0.8391	0.3162	0.0635
xgb	0.839	0.3159	0.0634
logistic	0.8388	0.3149	0.1716
logistic_l1	0.8388	0.3149	0.1716
lgbm	0.8385	0.3144	0.0636

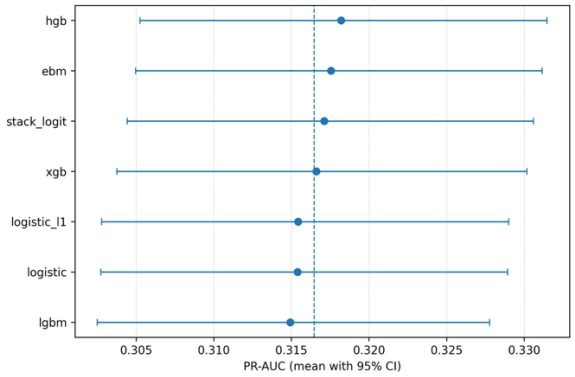

Figure 3. PR-AUC (mean with 95% bootstrap CI) by model.

Table 4. Pairwise AUROC, DeLong tests.

model_a	model_b	p_value	auroc_a	auroc_b
logistic	logistic_l1	0.9999	0.8388	0.8388
logistic	hgb	0.9272	0.8388	0.8407
logistic	xgb	0.9353	0.8388	0.8405
logistic	ebm	0.9612	0.8388	0.8398
logistic	lgbm	0.9823	0.8388	0.8393
logistic	stack_logit	0.9851	0.8388	0.8392
logistic_l1	hgb	0.9271	0.8388	0.8407
logistic_l1	xgb	0.9352	0.8388	0.8405
logistic_l1	ebm	0.9611	0.8388	0.8398
logistic_l1	lgbm	0.9822	0.8388	0.8393
logistic_l1	stack_logit	0.9850	0.8388	0.8392
hgb	xgb	0.9919	0.8407	0.8405
hgb	ebm	0.9659	0.8407	0.8398
hgb	lgbm	0.9449	0.8407	0.8393
hgb	stack_logit	0.9421	0.8407	0.8392
xgb	ebm	0.9740	0.8405	0.8398

xgb	lgbm	0.9530	0.8405	0.8393
xgb	stack_logit	0.9502	0.8405	0.8392
ebm	lgbm	0.9789	0.8398	0.8393
ebm	stack_logit	0.9761	0.8398	0.8392
lgbm	stack_logit	0.9972	0.8393	0.8392

4.4 Calibration Results

Calibration was evaluated on the held-out test set using the Brier score and expected calibration error (ECE; 20 equal-width bins). Summary metrics are shown in Table 5, while reliability diagrams are provided in Figure 4. Tree boosting models exhibited consistently strong probability accuracy. HGB and XGBoost achieved the lowest Brier scores (both 0.0633) with small ECE values (0.0045 and 0.0044, respectively), and LightGBM showed comparable behavior (Table 5). The reliability curves in Figure 4 corroborate these results: boosted learners and the stacked logistic model closely track the 45-degree line across the central probability mass, whereas uncalibrated linear baselines deviate more substantially.

Table 5. Test-set calibration summary.

model	brier	ece_20bins
hgb	0.0633	0.0045
xgb	0.0633	0.0044
ebm	0.0634	0.0065
stack_logit	0.0634	0.0063
lgbm	0.0634	0.0031
logistic_l1	0.1716	0.2768
logistic	0.1716	0.2768

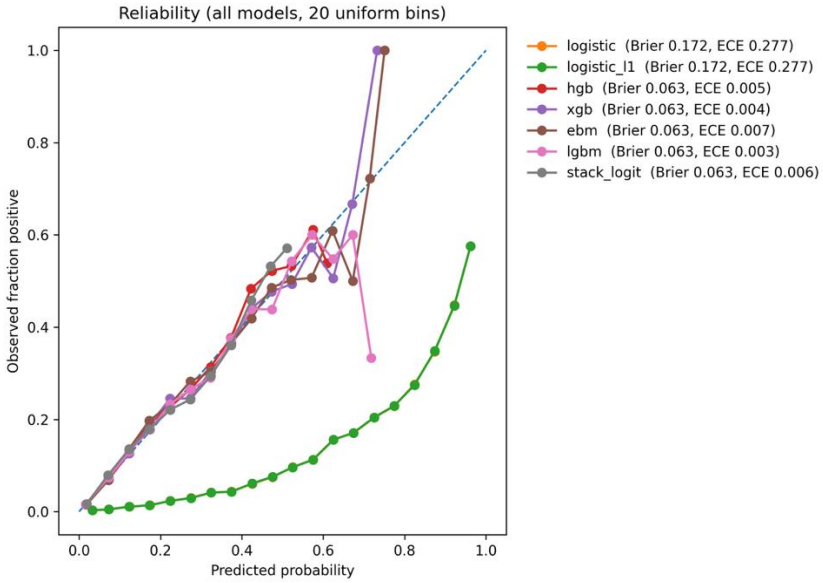

Figure 4. Reliability diagram (20 bins) for all models.

4.5 Murphy Decomposition

To diagnose the source of probabilistic error, we decomposed the test-set Brier score using Murphy's identity (Brier = Reliability − Resolution + Uncertainty) with 20 equal-width probability bins. The decomposition components are reported in Table 6 and visualized in Figure 5. As expected, the Uncertainty term was identical across models (reflecting the fixed outcome prevalence). High-performing non-linear models and the stacking ensemble showed near-zero Reliability, indicating minimal systematic miscalibration, while maintaining meaningful Resolution (risk stratification). The relationship between Murphy Reliability and ECE is shown in Figure 6, supporting consistent identification of models that achieved both low calibration error and low reliability loss.

Table 6. Murphy components on the test set (20 bins).

model	variant	brier	reliability	resolution	uncertainty	ece
logistic	no_cal	0.1716	0.1081	0.0109	0.0743	0.2768
logistic	platt	0.0635	0.0002	0.0108	0.0743	0.0066
logistic	isotonic	0.0634	0.0000	0.0108	0.0743	0.0031
logistic_l1	no_cal	0.1716	0.1081	0.0109	0.0743	0.2768
logistic_l1	platt	0.0635	0.0002	0.0108	0.0743	0.0066
logistic_l1	isotonic	0.0634	0.0001	0.0108	0.0743	0.0033
hgb	no_cal	0.0633	0.0001	0.0109	0.0743	0.0045
hgb	platt	0.0633	0.0001	0.0109	0.0743	0.0034
hgb	isotonic	0.0633	0.0000	0.0109	0.0743	0.0028
xgb	no_cal	0.0633	0.0001	0.0109	0.0743	0.0044
ebm	no_cal	0.0634	0.0001	0.0109	0.0743	0.0065
ebm	platt	0.0634	0.0001	0.0109	0.0743	0.0063
ebm	isotonic	0.0634	0.0001	0.0109	0.0743	0.0039
lgbm	no_cal	0.0634	0.0001	0.0107	0.0743	0.0031
lgbm	platt	0.0661	0.0025	0.0105	0.0743	0.0245
lgbm	isotonic	0.0641	0.0007	0.0107	0.0743	0.0120
stack_logit	no_cal	0.0634	0.0001	0.0108	0.0743	0.0063
stack_logit	platt	0.0634	0.0001	0.0108	0.0743	0.0064
stack_logit	isotonic	0.0634	0.0001	0.0109	0.0743	0.0040

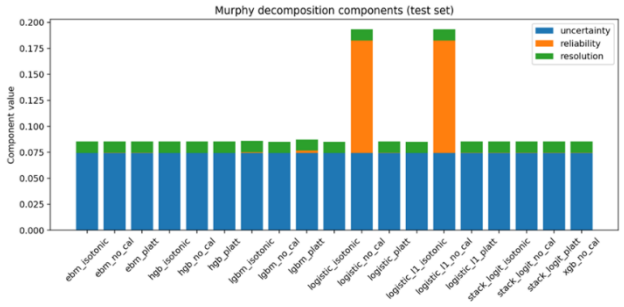

Figure 5. Stacked Murphy components.

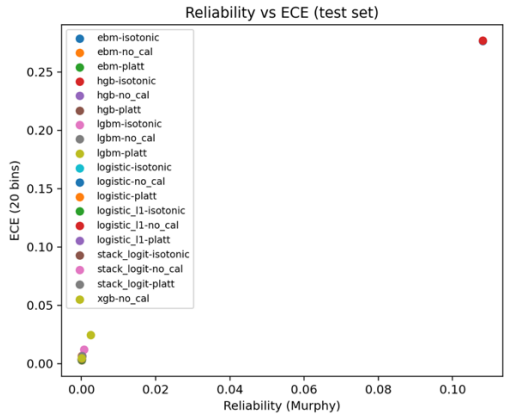

Figure 6. Scatter of Reliability (Murphy) vs ECE.

4.6 Explainability

Explainability analyses focused on the strongest-performing model family to support plausibility checks and governance-oriented interpretation. Global attribution patterns are summarized in Figure 7 (grouped by raw predictors) and Figure 8 (one-hot-encoded levels). Functional diagnostics were then used to examine directionality and non-linearity. Figure 9 presents partial dependence plots (PDPs) for leading raw features, while Figure 10 reports 1D accumulated local effects (ALE) for key numeric predictors. Individual conditional expectation (ICE) plots in Figure 11 highlight local heterogeneity for exemplar numeric predictors. Finally, pairwise interaction structure was assessed using 2D ALE visualizations (Figure 12) to illustrate joint effects in clinically relevant feature combinations.

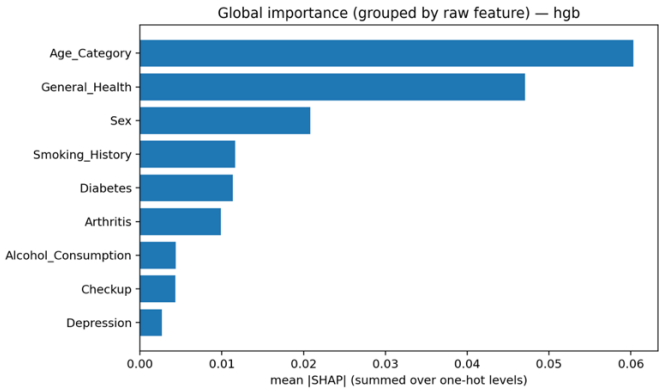

Figure 7. Global importance grouped by raw features (mean |SHAP|), HGB.

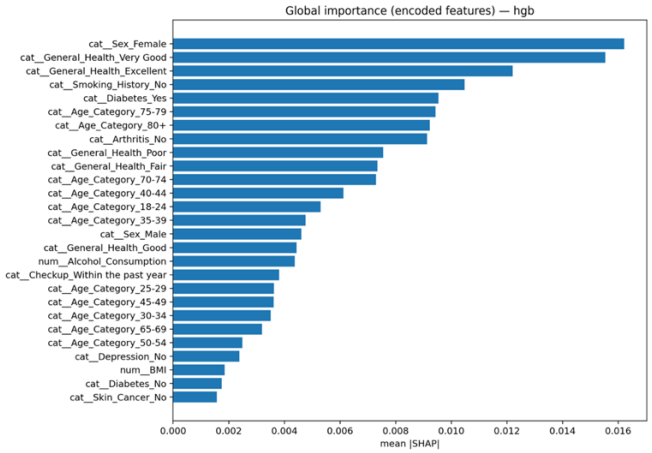

Figure 8. Global importance at the encoded (one-hot) level (mean |SHAP|), HGB.

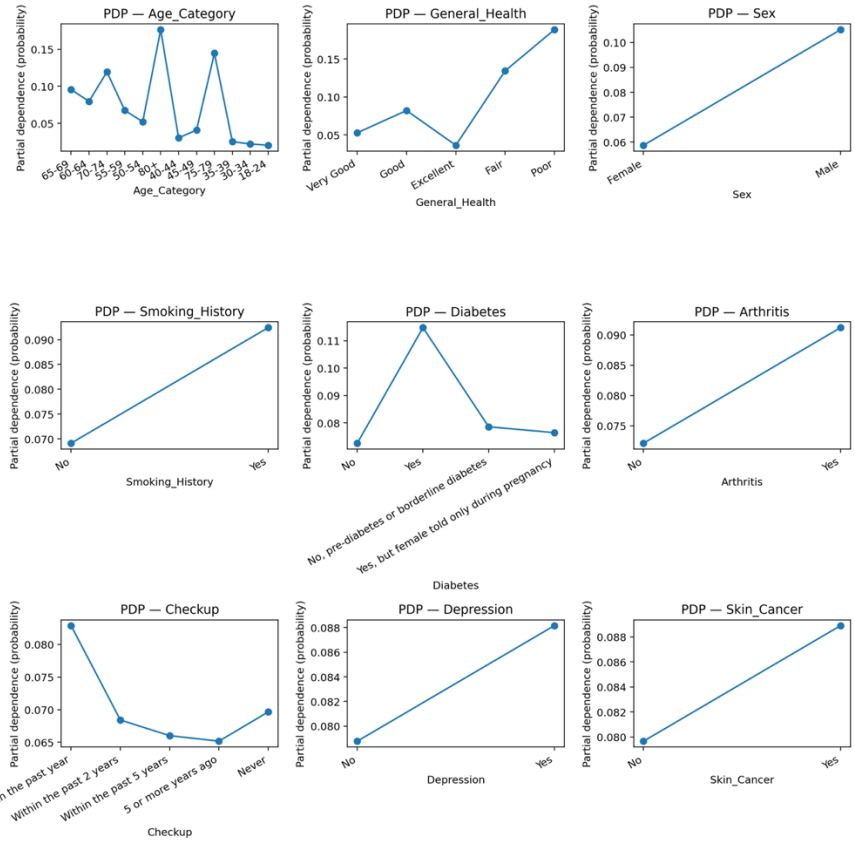

Figure 9. PDPs for leading categorical and numeric raw features.

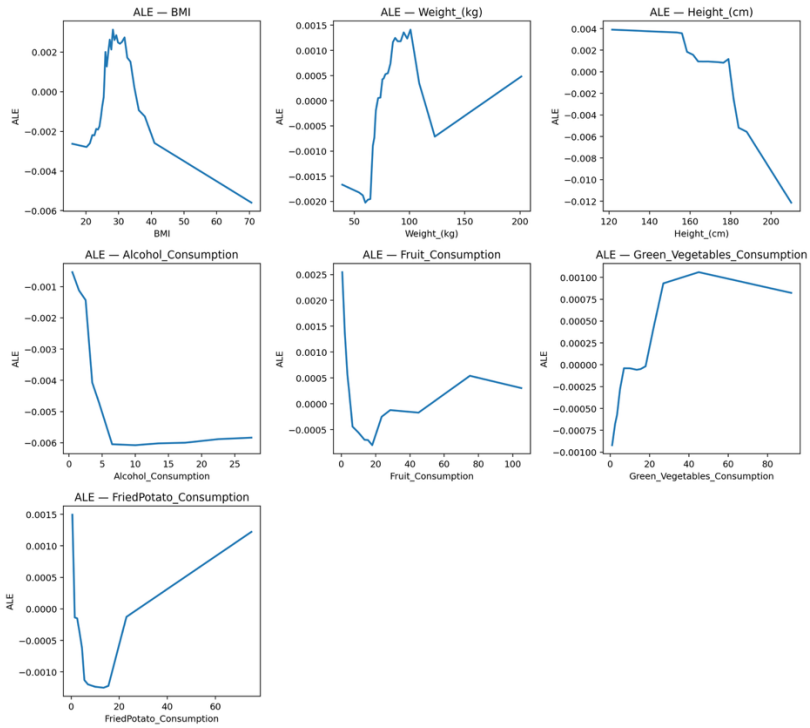

Figure 10. 1D ALE for numeric features (Alcohol, BMI, Weight, Height, FriedPotato, Green_Vegetables, Fruit).

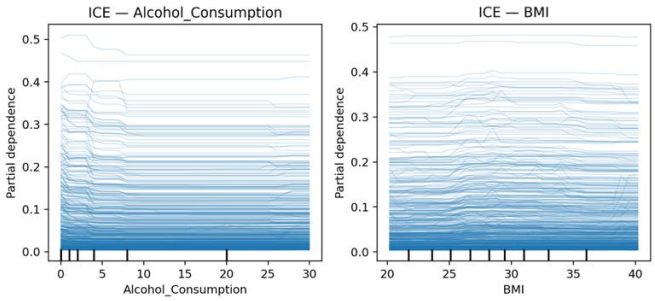

Figure 11. ICE plots for Alcohol_Consumption and BMI.

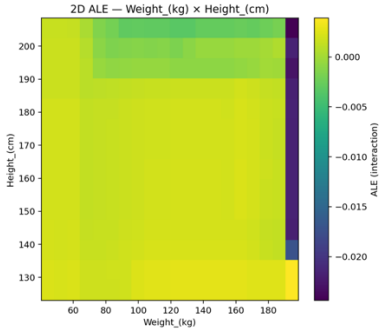

Figure 12. 2D ALE for BMI × Alcohol_Consumption and Weight_(kg) × Height_(cm).

4.7 Subgroup Fairness and Stability

Subgroup analyses were conducted on the fixed test set across routinely available strata, reporting discrimination (ROC-AUC; PR-AUC) and probabilistic accuracy (Brier; ECE with 20 bins). Summary patterns are visualized using facet grids for key subgroups (Figures 13–18). By sex, discrimination was broadly similar between females and males (near-identical ROC-AUC), while PR-AUC was higher among males, consistent with higher event prevalence. More pronounced heterogeneity emerged by age group and general health. ROC-AUC peaked in midlife (approximately 40–49 years) and declined in the oldest group (80+), whereas PR-AUC increased with age as prevalence rose. A clinically coherent gradient was observed across general health strata: as self-reported health worsened, PR-AUC increased (reflecting risk concentration), while ROC-AUC decreased and both Brier/ECE tended to worsen. For deployment, these results indicate that while overall discrimination is stable, calibration and clinical reliability vary meaningfully across subgroups, particularly among older and less healthy groups, motivating subgroup-aware monitoring and potential recalibration strategies.

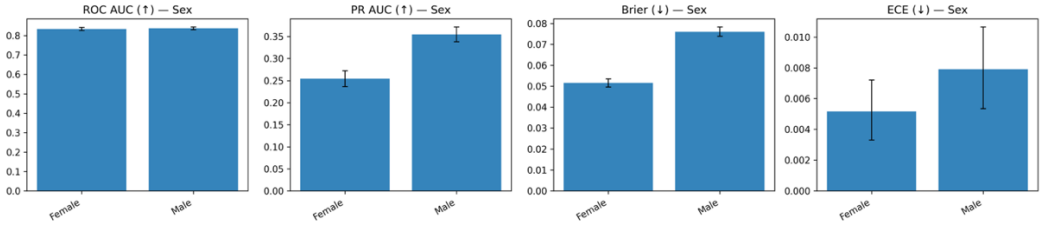

Figure 13. Metrics by Sex.

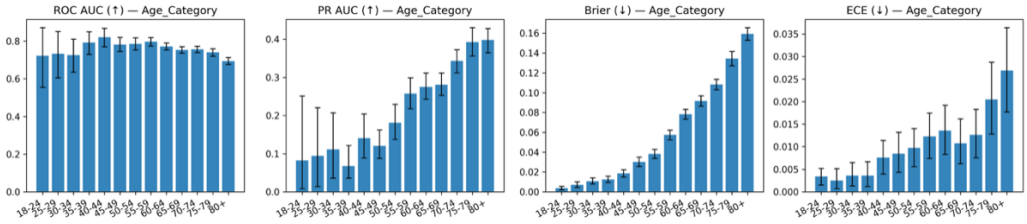

Figure 14. Metrics by Age Category.

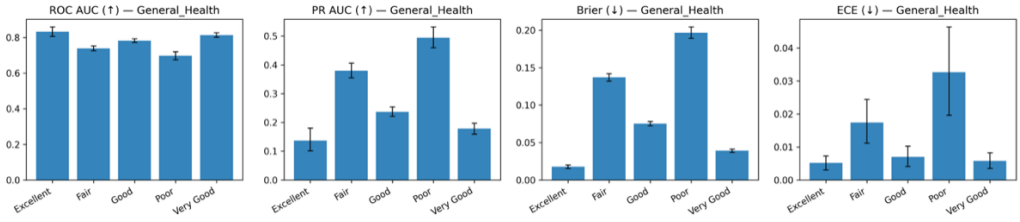

Figure 15. Metrics by General Health.

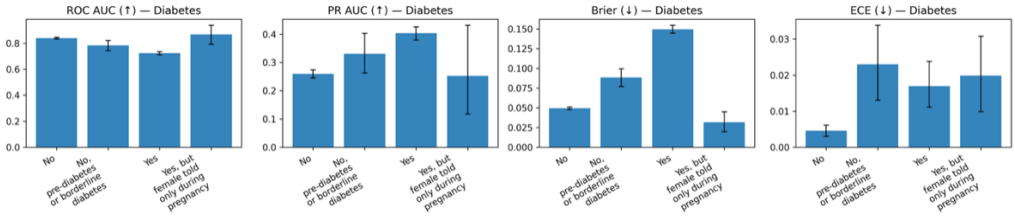

Figure 16. Metrics by Diabetes.

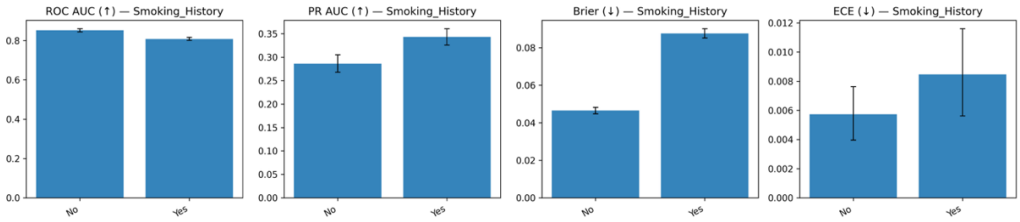

Figure 17. Metrics by Smoking History.

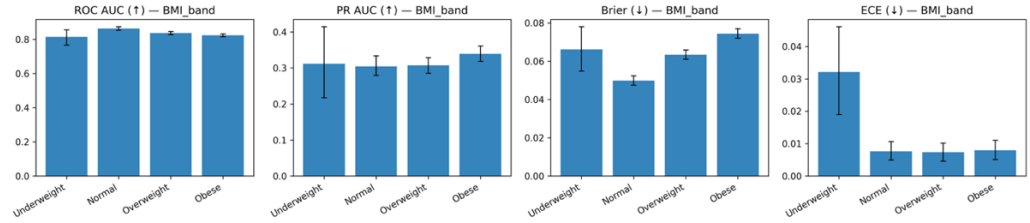

Figure 18. Metrics by BMI band.

4.8 Clinical Usefulness

Clinical utility was assessed on the held-out test set using decision curve analysis (DCA), anchored against treat-none (NB = 0) and treat-all (NB computed from test-set prevalence). Net benefit curves were traced over thresholds 0.01–0.50, with emphasis on the predefined screening-relevant policy window. Figure 19 shows that within the screening window, non-linear learners (HGB, EBM, XGBoost, LightGBM) and the stacked logistic model maintained positive net benefit and dominated both reference strategies. To improve separability among high-performing models, Figure 20 displays incremental net benefit (ΔNB) relative to treat-none within the policy window $\tau \in [0.05, 0.15]$. Within this range, HGB formed a tight upper envelope and provided the highest average incremental clinical value across screening-relevant thresholds.

At prespecified operating points, workload and accuracy were compatible with low-prevalence triage. At $\tau = 0.08$, HGB achieved NB ≈ 0.044, sensitivity ≈ 0.822, and PPV ≈ 0.202, corresponding to approximately 330 alerts per 1,000 screened (approximately 263 false positives per 1,000). Tightening to $\tau = 0.12$ reduced workload to approximately 213 alerts per 1,000, while maintaining NB ≈ 0.032–0.033 and PPV ≈ 0.24–0.26 across the top models. Collectively, these findings indicate that model-guided screening confers clinically meaningful benefit over default strategies within the policy window, with HGB offering the most favorable average trade-off.

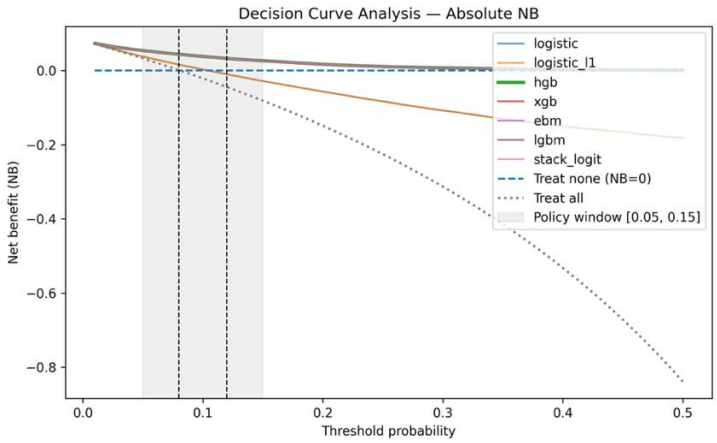

Figure 19. Decision Curve Analysis—absolute net benefit on the test set.

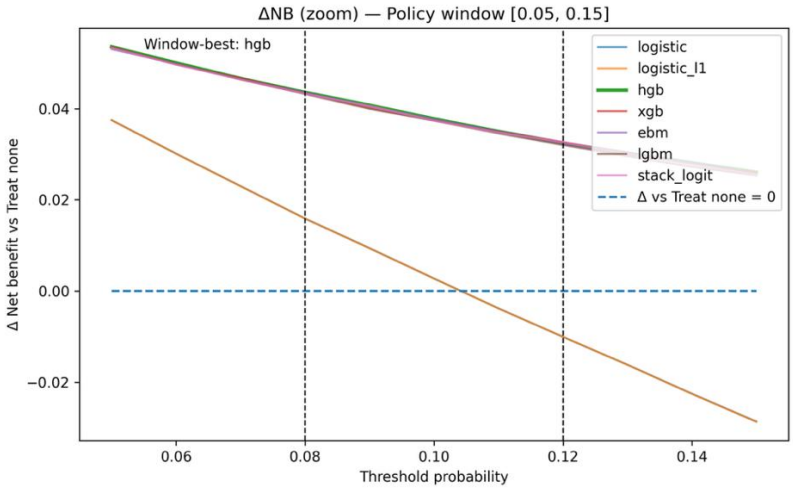

Figure 20. Decision Curve Analysis—ΔNB (zoom) within the policy window (0.05–0.15).

5. Discussion

5.1 Principal Findings

Across a fixed, leakage-safe holdout, the candidate models achieved consistent out-of-sample discrimination while materially improving probability accuracy after post-hoc calibration. The histogram-based gradient boosting (HGB) attained the highest PR-AUC (0.3177; ROC-AUC 0.8407), closely followed by the stacking logistic ensemble and EBM. Raw probability calibration differed markedly by family: tree/ensemble models already exhibited low Brier and ECE (e.g., HGB Brier 0.0633; ECE 0.0045; LightGBM ECE 0.0031), whereas uncalibrated logistic variants were poorly calibrated (Brier ≈ 0.1716; ECE ≈ 0.2768). Murphy decomposition confirmed that the observed Brier reductions stemmed from large decreases in the Reliability component with Resolution preserved: for logistic, Reliability dropped from 0.1081 (no_cal) to ~0.0000 under isotonic, while Resolution remained ~0.0108; HGB showed near-zero Reliability with a small

additional gain after isotonic. These findings support H1: post-hoc calibration reduces ECE and Brier by lowering Reliability without materially degrading Resolution.

Decision-curve analysis (DCA) demonstrated clinically meaningful net benefit within the screening-relevant threshold window. Non-linear models (HGB, EBM, LightGBM, XGBoost) formed the upper envelope with net benefit above both treat-none and treat-all for thresholds ≈ 0.05–0.15. At the prespecified operating points, performance/workload trade-offs were compatible with low-prevalence triage: at $\tau = 0.08$, HGB achieved NB ≈ 0.044 with PPV ≈ 0.202 and sensitivity ≈ 0.822 (≈330 alerts/1,000 screened); at $\tau = 0.12$, models clustered with NB ≈ 0.032–0.033 and PPV ≈ 0.24–0.26. These results substantiate H2: model-guided decisions provide positive net benefit across the intended policy window. Subgroup audits revealed small, clinically acceptable differences with overlapping confidence intervals across sex and age strata. Specifically, discrimination was similar between females and males; PR-AUC tracked subgroup prevalence; and calibration was modestly worse in older ages and among individuals with poorer self-reported health. Taken together, the models are interpretable, calibrated, and decision-useful in the targeted threshold range while maintaining subgroup stability consistent with H3.

5.2 Interpretation in Relation to Model Evaluation and Deployment

A key implication of the results is that discrimination alone is insufficient to characterize deployment readiness in a low-prevalence screening context. Although discrimination clustered tightly among top models, calibration metrics, and Murphy reliability provided additional evidence that probability estimates were sufficiently aligned with observed frequencies for clinically meaningful thresholding. This distinction is practically important because policy actions (e.g., referrals or preventive interventions) are triggered by predicted risk thresholds rather than by rank ordering. The Murphy decomposition adds interpretive value by clarifying whether improvements in Brier score arise from improved reliability (reduced systematic miscalibration) while preserving resolution (risk stratification), thereby avoiding an overly optimistic interpretation of global probabilistic accuracy.

These findings align with prior evidence and guidance emphasizing that legacy cardiovascular risk scores (e.g., Framingham-derived models and QRISK) often require recalibration in contemporary populations, and that miscalibration can materially shift treatment eligibility at fixed thresholds [55], [56], [57]. Published assessments of traditional scores commonly emphasize discrimination, whereas comprehensive calibration diagnostics, decision-analytic evaluation, and subgroup-oriented audits remain less consistently reported relative to clinical need [3], [57]. In contrast, the present study operationalizes a governance-ready evaluation that couples probability diagnostics (Brier score, ECE, reliability curves, Murphy decomposition) with decision curve analysis (DCA) and predefined subgroup audits. This integrated design demonstrates that post hoc calibration can reduce the Murphy Reliability

component to zero without degrading Resolution, while threshold-specific net benefit clarifies when model-guided actions improve upon treat-all or treat-none strategies [10], [24]. Although our dataset does not permit a head-to-head comparison with legacy calculators, the framework is directly applicable: intercept–slope updating or isotonic recalibration followed by DCA within guideline-relevant thresholds and subgroup-specific calibration checks can yield deployment-grade evidence for traditional scores in current populations [3], [10], [24].

More broadly, our results support the established view that probability accuracy and decision utility must be evaluated explicitly for clinical readiness [1], [3]. By integrating Brier score, ECE, and Murphy decomposition with DCA and subgroup audits, we align with recommendations that emphasize calibrated risk, threshold-specific utility, and equity considerations during external validation [18], [19]. Our decision-analytic findings further accord with reports that models with similar AUC values can differ in clinical usefulness once threshold preferences and class imbalance are considered [2], [6], [24]. Finally, pairing explainability analyses (e.g., SHAP/ALE) with subgroup auditing reflects the shift toward governance-ready evaluation, in which plausibility checks and fairness monitoring are treated as first-class requirements rather than optional add-ons [49], [51], [53], [54].

5.3 Clinical Utility and Workload Implications

Decision curve analysis indicated that model-guided strategies yield positive net benefit across plausible threshold probabilities, supporting clinical usefulness in screening or triage. The additional workload translation provides an operational interpretation: at lower thresholds, the model can capture a larger proportion of true positives but incurs higher alert volume and false positives; at higher thresholds, workload decreases at the cost of reduced sensitivity. Importantly, the net benefit curves indicate that, within the policy window considered, the top-performing models deliver benefits beyond default strategies, and the incremental net benefit analysis shows that HGB offers a consistently strong trade-off across this range. From an implementation perspective, these results can inform the selection of thresholds aligned with capacity constraints (e.g., available clinic slots) and the harm–benefit preferences of the healthcare setting.

5.4 Explainability and Plausibility of Model Behavior

Explainability analyses (SHAP-based summaries and functional diagnostics such as PDP/ALE/ICE) served two roles: (i) plausibility checking to ensure that the model's learned associations are clinically coherent, and (ii) governance support by revealing non-linearities, local heterogeneity, and interaction structures that could influence threshold-based decision rules. The combined use of attribution-based explanations (SHAP) and response-function diagnostics (PDP/ALE/ICE) is methodologically complementary: SHAP elucidates which predictors contribute most to the prediction function overall and at the individual level, while PDP/ALE/ICE characterizes how predicted risk changes as features vary. This dual perspective

is particularly relevant when models are used as decision support tools, because clinicians and governance bodies need both "what drives the model" and "how the model responds" under realistic variation in patient profiles.

5.5 Subgroup Stability, Fairness Considerations, and Monitoring

Subgroup analyses indicated that global performance does not necessarily translate into uniform reliability across all strata. While discrimination was broadly similar across sex groups, greater heterogeneity emerged across age groups and self-reported general health. The observed pattern improved PR-AUC with rising prevalence alongside degraded ROC-AUC and worsening calibration error in older or less healthy strata—highlights that subgroup shifts in case mix and baseline risk can alter both ranking and calibration behavior. For practical deployment, these findings motivate subgroup-aware monitoring, including periodic calibration checks within key strata and recalibration strategies when drift is detected. From a fairness perspective, calibration within groups is especially important in clinical risk estimation because systematic over- or underestimation in particular subpopulations can translate into unequal access to preventive care or follow-up evaluation.

5.6 Strengths

This study has several strengths. First, it uses a large-scale dataset with a clear preprocessing pipeline and leakage-safe evaluation design, enabling stable estimates of discrimination and calibration in a low-prevalence context. Second, the evaluation is comprehensive and aligned with deployment needs, integrating discrimination, calibration, diagnostic decomposition, decision-analytic utility, subgroup stability, and interpretability. Third, uncertainty quantification via bootstrap intervals and the use of DeLong tests provide transparent evidence for comparative claims, avoiding overinterpretation of small metric differences between top-performing models. Finally, the inclusion of workload translation complements decision curves by connecting statistical performance to operational decision-making.

5.7 Limitations

Several limitations should be considered when interpreting these results. First, the dataset is derived from a public health survey and may rely on self-reported variables and outcome status, which can introduce measurement error and limit direct comparability to models trained on electronic health records. Second, external validity is not established in this analysis; performance may vary across different populations, clinical workflows, or measurement practices for predictors. Third, subgroup analyses can be sensitive to sample size and prevalence differences within strata; therefore, subgroup-specific conclusions should be interpreted with uncertainty intervals and complemented by prospective monitoring in deployment settings. Fourth, although explainability diagnostics help check plausibility, they do not establish causal

relationships and should not be used to infer causal effects of predictors. Future work should focus on external validation, prospective impact evaluation, and subgroup-aware recalibration strategies to support safe clinical decision support.

5.8 Future Work

Future research should explore this evaluation in three ways. First, external validation using independent datasets or sites should be performed to test transportability and measure calibration drift. Second, strategies for model updating and recalibration, possibly tailored to key subgroups, should be examined to ensure reliability over time. Third, impact-focused studies are necessary to determine whether model-guided decisions improve clinical outcomes, resource efficiency, and equity in real-world settings.

6. Conclusions

This study shows that predicting heart disease risk in a low-prevalence setting benefits from an evaluation framework that goes beyond discrimination to include probability reliability, clinical utility, and deployment safeguards. Using a large secondary dataset, multiple non-linear models achieved strong discrimination, with histogram-based gradient boosting (HGB) providing the highest PR-AUC and showing similar AUROC to other top models. Importantly, calibration tests revealed that the best models produced well-aligned probability estimates, indicated by low Brier scores and small ECE values, confirmed by reliability diagrams.

Murphy decomposition further clarified the sources of probabilistic error and showed that the best models achieved almost zero reliability loss while maintaining meaningful resolution. This indicates that probabilistic accuracy was achieved not only by moving predictions toward the base rate but also by effective risk stratification, an essential feature of threshold-based clinical decision support. Subgroup analyses provided additional insights relevant to deployment: while discrimination remained generally stable, calibration and error patterns varied across age groups and self-reported general health levels, underscoring the importance of subgroup-aware monitoring and, when necessary, recalibration to ensure fair and reliable performance.

Decision curve analysis confirmed that model-guided strategies can yield a positive net benefit relative to treat-all and treat-none policies across a screening-relevant threshold range. When translated into operational terms, the findings highlight a practical trade-off between sensitivity and workload, enabling threshold selection aligned with local capacity constraints and policy preferences. Taken together, the results support the use of a unified, governance-oriented evaluation stack that integrates discrimination, calibration with diagnostic decomposition, subgroup stability, interpretability, and decision-analytic utility as a pragmatic pathway for assessing the readiness of machine-learning models for clinical screening and triage.

Future work should prioritize external validation in independent populations, systematic assessment of temporal and site-specific drift, and evaluation of subgroup-aware updating and

recalibration strategies. Prospective impact studies are also needed to determine whether model-guided decision support improves clinical outcomes, resource efficiency, and equity when embedded in real-world care pathways.

Data and Code Availability: A fully reproducible package containing all analysis code and research artifacts will be deposited on Zenodo (DOI: https://doi.org/10.5281/zenodo.17421151). The package comprises: (i) configuration files and the data schema; (ii) deterministic train/validation/test indices; (iii) fitted pipelines, post-hoc calibrators, and trained model binaries; (iv) out-of-fold and held-out predictions with cross-validation summaries; (v) evaluation outputs (tables and figures for discrimination, calibration with Murphy decomposition, decision-curve analysis, subgroup audits, and explainability SHAP, PDP/ICE, ALE); and (vi) run logs and audit metadata to enable exact re-execution. This section is not mandatory but may be added if there are patents resulting from the work reported in this manuscript. The raw data are available from Kaggle (dataset: "Cardiovascular Diseases Risk Prediction Dataset" by Alphiree). Due to the dataset's licensing on Kaggle, we do not redistribute the raw files. Our Zenodo package includes scripts and exact commands to retrieve the dataset and reproduce all derived artifacts (processed tables, figures, and model outputs).

Reproducibility and Seeds: We provide a manifest with random seeds, software versions, and CLI commands for each step. All preprocessing is performed on the training data only; train/test indices and per-record predictions are provided to ensure like-for-like recomputation without resplitting.

Data Availability and Third-Party Terms: The raw data are available from Kaggle (dataset: "Cardiovascular Diseases Risk Prediction Dataset" by Alphiree). Due to the dataset's licensing on Kaggle, we do not redistribute the raw files. Our Zenodo package includes scripts and exact commands to retrieve the dataset and reproduce all derived artifacts (processed tables, figures, and model outputs).

Licensing: Code is released under a permissive open-source license (to be specified in the Zenodo record). Model binaries and derived artifacts are redistributed under the same license unless otherwise stated; any third-party data remains under their original licenses/terms.

Ethics Approval and Consent to Participate: This study used publicly available, de-identified secondary data. According to institutional policy, it does not constitute human-subjects research and does not require IRB approval or informed consent.

Acknowledgments: This paper was financially supported by Mahasarakham Business School, Mahasarakham University, Thailand.

Author Contributions: Hathaichanok Chompoopong (H.C.): Conceptualization; Methodology; Software; Formal analysis; Data curation; Visualization; Writing-original draft.

Warawut Narkbunnum (W.N.): Conceptualization; Methodology; Validation; Supervision; Project administration; Funding acquisition; Resources; Writing-review and editing.

Conflicts of Interest: The author(s) declare that there are no conflicts of interest regarding the publication of this paper.

Abbreviations

The following abbreviations are used in this manuscript:

ALE	Accumulated Local Effects
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BMI	Body Mass Index
CI	Confidence Interval
CVD	Cardiovascular Disease
DCA	Decision Curve Analysis
ECE	Expected Calibration Error
EBM	Explainable Boosting Machine
HGB	Histogram-Based Gradient Boosting
ICE	Individual Conditional Expectation
IRB	Institutional Review Board
LGBM	Light Gradient Boosting Machine (LightGBM)
ML	Machine Learning
NB	Net Benefit
NPV	Negative Predictive Value
PDP	Partial Dependence Plot
PPV	Positive Predictive Value
PR-AUC	Area Under the Precision–Recall Curve
ROC	Receiver Operating Characteristic
SHAP	Shapley Additive Explanations
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

References

[1] R.D. Riley, L. Archer, K.I.E. Snell, J. Ensor, P. Dhiman, et al., Evaluation of Clinical Prediction Models (Part 2): How to Undertake an External Validation Study, BMJ 384 (2024), e074820. https://doi.org/10.1136/BMJ-2023-074820.

[2] A.J. Vickers, E.B. Elkin, Decision Curve Analysis: A Novel Method for Evaluating Prediction Models, Med. Decis. Mak. 26 (2006), 565-574. https://doi.org/10.1177/0272989X06295361.

[3] G.S. Collins, K.G.M. Moons, P. Dhiman, R.D. Riley, A.L. Beam, et al., TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods, BMJ 385 (2024), e078378. https://doi.org/10.1136/BMJ-2023-078378.

[4] B. Van Calster, A.J. Vickers, Calibration of Risk Prediction Models, Med. Decis. Mak. 35 (2014), 162-169. https://doi.org/10.1177/0272989X14547233.

[5] B. Van Calster, D.J. McLernon, M. van Smeden, L. Wynants, et al., Calibration: The Achilles Heel of Predictive Analytics, BMC Med. 17 (2019), 230. https://doi.org/10.1186/s12916-019-1466-7.

[6] A.J. Vickers, B. Van Calster, E.W. Steyerberg, Net Benefit Approaches to the Evaluation of Prediction Models, Molecular Markers, and Diagnostic Tests, BMJ 352 (2016), i6. https://doi.org/10.1136/BMJ.I6.

[7] S. Mora, N.K. Wenger, N.R. Cook, J. Liu, B.V. Howard, et al., Evaluation of the Pooled Cohort Risk Equations for Cardiovascular Risk Prediction in a Multiethnic Cohort from the Women's Health Initiative, JAMA Intern. Med. 178 (2018), 1231-1240. https://doi.org/10.1001/JAMAINTERNMED.2018.2875.

[8] J.S. Rana, G.H. Tabada, M.D. Solomon, J.C. Lo, M.G. Jaffe, et al., Accuracy of the Atherosclerotic Cardiovascular Risk Equation in a Large Contemporary, Multiethnic Population, J. Am. Coll. Cardiol. 67 (2016), 2118-2130. https://doi.org/10.1016/J.JACC.2016.02.055.

[9] C.A. Emdin, A.V. Khera, P. Natarajan, D. Klarin, U. Baber, et al., Evaluation of the Pooled Cohort Equations for Prediction of Cardiovascular Risk in a Contemporary Prospective Cohort, Am. J. Cardiol. 119 (2017), 881-885. https://doi.org/10.1016/j.amjcard.2016.11.042.

[10] B. Van Calster, L. Wynants, J.F. Verbeek, J.Y. Verbakel, E. Christodoulou, et al., Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators, Eur. Urol. 74 (2018), 796-804. https://doi.org/10.1016/J.EURURO.2018.08.038.

[11] K.G. Moons, R.F. Wolff, R.D. Riley, P.F. Whiting, M. Westwood, et al., PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration, Ann. Intern. Med. 170 (2019), W1-W33. https://doi.org/10.7326/M18-1377.

[12] R.F. Wolff, K.G. Moons, R.D. Riley, P.F. Whiting, M. Westwood, et al., PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies, Ann. Intern. Med. 170 (2019), 51-58. https://doi.org/10.7326/M18-1376.

[13] S.E. Davis, R.A. Greevy, C. Fonnesbeck, T.A. Lasko, C.G. Walsh, et al., A Nonparametric Updating Method to Correct Clinical Prediction Model Drift, J. Am. Med. Inform. Assoc. 26 (2019), 1448-1457. https://doi.org/10.1093/JAMIA/OCZ127.

[14] T.L. Su, T. Jaki, G.L. Hickey, I. Buchan, M. Sperrin, A Review of Statistical Updating Methods for Clinical Prediction Models, Stat. Methods Med. Res. 27 (2016), 185-197. https://doi.org/10.1177/0962280215626466.

[15] C. Yang, J.A. Kors, S. Ioannou, L.H. John, A.F. Markus, et al., Trends in the Conduct and Reporting of Clinical Prediction Model Development and Validation: A Systematic Review, J. Am. Med. Inform. Assoc. 29 (2022), 983-989. https://doi.org/10.1093/JAMIA/OCAC002.

[16] Y. Cai, Y.Q. Cai, L.Y. Tang, Y.H. Wang, M. Gong, et al., Artificial Intelligence in the Risk Prediction Models of Cardiovascular Disease and Development of an Independent Validation Screening Tool: A Systematic Review, BMC Med. 22 (2024), 56. https://doi.org/10.1186/s12916-024-03273-7.

[17] T.P. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E.W. Steyerberg, et al., A New Framework to Enhance the Interpretation of External Validation Studies of Clinical Prediction Models, J. Clin. Epidemiol. 68 (2015), 279-289. https://doi.org/10.1016/j.jclinepi.2014.06.018.

[18] K. Luijken, R.H.H. Groenwold, B. Van Calster, E.W. Steyerberg, M. van Smeden, Impact of Predictor Measurement Heterogeneity Across Settings on the Performance of Prediction Models: A Measurement Error Perspective, Stat. Med. 38 (2019), 3444-3459. https://doi.org/10.1002/SIM.8183.

[19] K. Luijken, L. Wynants, M. van Smeden, B. Van Calster, E.W. Steyerberg, et al., Changing Predictor Measurement Procedures Affected the Performance of Prediction Models in Clinical Examples, J. Clin. Epidemiol. 119 (2020), 7-18. https://doi.org/10.1016/J.JCLINEPI.2019.11.001.

[20] A. Mishra, R.L. McClelland, L.Y.T. Inoue, K.F. Kerr, Recalibration Methods for Improved Clinical Utility of Risk Scores, Med. Decis. Mak. 42 (2021), 500-512. https://doi.org/10.1177/0272989X211044697.

[21] C. Hong, M.J. Pencina, D.M. Wojdyla, J.L. Hall, S.E. Judd, et al., Predictive Accuracy of Stroke Risk Prediction Models Across Black and White Race, Sex, and Age Groups, JAMA 329 (2023), 306-317. https://doi.org/10.1001/JAMA.2022.24683.

[22] K. Chakradeo, I. Huynh, S.B. Balaganeshan, O.L. Dollerup, H. Gade-Jørgensen, et al., Navigating Fairness Aspects of Clinical Prediction Models, BMC Med. 23 (2025), 567. https://doi.org/10.1186/S12916-025-04340-3.

[23] C.G. Walsh, K. Sharman, G. Hripcsak, Beyond Discrimination: A Comparison of Calibration Methods and Clinical Usefulness of Predictive Models of Readmission Risk, J. Biomed. Inform. 76 (2017), 9-18. https://doi.org/10.1016/J.JBI.2017.10.008.

[24] A.J. Vickers, B. van Calster, E.W. Steyerberg, A Simple, Step-By-Step Guide to Interpreting Decision Curve Analysis, Diagn. Progn. Res. 3 (2019), 18. https://doi.org/10.1186/s41512-019-0064-7.

[25] M. Sadatsafavi, A. Adibi, M. Puhan, A. Gershon, S.D. Aaron, et al., Moving Beyond AUC: Decision Curve Analysis for Quantifying Net Benefit of Risk Prediction Models, Eur. Respir. J. 58 (2021), 2101186. https://doi.org/10.1183/13993003.01186-2021.

[26] D. Piovani, R. Sokou, A.G. Tsantes, A.S. Vitello, S. Bonovas, Optimizing Clinical Decision Making with Decision Curve Analysis: Insights for Clinical Investigators, Healthcare 11 (2023), 2244. https://doi.org/10.3390/HEALTHCARE11162244.

[27] A.V. Ponce-Bobadilla, V. Schmitt, C.S. Maier, S. Mensing, S. Stodtmann, Practical Guide to SHAP Analysis: Explaining Supervised Machine Learning Model Predictions in Drug Development, Clin. Transl. Sci. 17 (2024), e70056. https://doi.org/10.1111/CTS.70056.

[28] Y. Huang, W. Li, F. Macheret, R.A. Gabriel, L. Ohno-Machado, A Tutorial on Calibration Measurements and Calibration Models for Clinical Prediction Models, J. Am. Med. Inform. Assoc. 27 (2020), 621-633. https://doi.org/10.1093/JAMIA/OCZ228.

[29] G.W. Brier, Verification of Forecasts Expressed in Terms of Probability, Mon. Weather. Rev. 78 (1950), 1-3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[30] A.H. Murphy, A New Vector Partition of the Probability Score, J. Appl. Meteorol. 12 (1973), 595-600. https://doi.org/10.1175/1520-0450(1973)012<0595:anvpot>2.0.co;2.

[31] S. Siegert, Simplifying and Generalising Murphy's Brier Score Decomposition, Q. J. R. Meteorol. Soc. 143 (2017), 1178-1183. https://doi.org/10.1002/QJ.2985.

[32] E.W. Steyerberg, A.J. Vickers, N.R. Cook, T. Gerds, et al., Assessing the Performance of Prediction Models, Epidemiology 21 (2010), 128-138. https://doi.org/10.1097/EDE.0B013E3181C30FB2.

[33] D.B. Stephenson, C.A.S. Coelho, I.T. Jolliffe, Two Extra Components in the Brier Score Decomposition, Weather. Forecast. 23 (2008), 752-757. https://doi.org/10.1175/2007WAF2006116.1.

[34] T. Dimitriadis, T. Gneiting, A.I. Jordan, Stable Reliability Diagrams for Probabilistic Classifiers, Proc. Natl. Acad. Sci. USA 118 (2021), e2016191118. https://doi.org/10.1073/PNAS.2016191118.

[35] A. Swaminathan, U. Srivastava, L. Tu, I. Lopez, N.H. Shah, et al., Against Reflexive Recalibration: Towards a Causal Framework for Addressing Miscalibration, Diagn. Progn. Res. 9 (2025), 4. https://doi.org/10.1186/S41512-024-00184-2.

[36] L. Zhao, Y. Leng, Y. Hu, J. Xiao, Q. Li, et al., Understanding Decision Curve Analysis in Clinical Prediction Model Research, Postgrad. Med. J. 100 (2024), 512-515. https://doi.org/10.1093/POSTMJ/QGAE027.

[37] T. Honda, Y. Furuta, A. Maezono, S. Chen, Y. Ishida, et al., External Validation of a Risk Prediction Model for Atherosclerotic Cardiovascular Diseases in a Large National Health-Checkup and Claim Database, J. Am. Hear. Assoc. 14 (2025), e040386. https://doi.org/10.1161/JAHA.124.040386.

[38] I. Hozo, G. Guyatt, B. Djulbegovic, Decision Curve Analysis Based on Summary Data, J. Eval. Clin. Pract. 30 (2023), 281-289. https://doi.org/10.1111/JEP.13945.

[39] G. Netto Flores Cruz, K. Korthauer, Bayesian Decision Curve Analysis with Bayesdca, Stat. Med. 43 (2024), 6042-6058. https://doi.org/10.1002/SIM.10277.

[40] C. Chai, S.Z. Peng, R. Zhang, C.W. Li, Y. Zhao, Advancing Emergency Department Triage Prediction with Machine Learning to Optimize Triage for Abdominal Pain Surgery Patients, Surg. Innov. 31 (2024), 583-597. https://doi.org/10.1177/15533506241273449.

[41] Q. Chen, Y. Qin, Z. Jin, X. Zhao, J. He, et al., Enhancing Performance of the National Field Triage Guidelines Using Machine Learning: Development of a Prehospital Triage Model to Predict Severe Trauma, J. Med. Internet Res. 26 (2024), e58740. https://doi.org/10.2196/58740.

[42] A.J. Vickers, B. Van Claster, L. Wynants, E.W. Steyerberg, Decision Curve Analysis: Confidence Intervals and Hypothesis Testing for Net Benefit, Diagn. Progn. Res. 7 (2023), 11. https://doi.org/10.1186/S41512-023-00148-Y.

[43] L. Rountree, Y.T. Lin, C. Liu, M. Salvatore, A. Admon, et al., Reporting of Fairness Metrics in Clinical Risk Prediction Models Used for Precision Health: Scoping Review, Online J. Public Health Inform. 17 (2025), e66598. https://doi.org/10.2196/66598.

[44] H.T. Cronjé, A. Katsiferis, L.K. Elsenburg, T.O. Andersen, N.H. Rod, et al., Assessing Racial Bias in Type 2 Diabetes Risk Prediction Algorithms, PLOS Glob. Public Health 3 (2023), e0001556. https://doi.org/10.1371/JOURNAL.PGPH.0001556.

[45] T. Hastie, R. Tibshirani, Generalized Additive Models, Stat. Sci. 1 (1986), 297–310. https://doi.org/10.1214/SS/1177013604.

[46] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, et al., Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission, in: Proceedings of the 21th ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining, ACM, New York, 2015, pp. 1721-1730. https://doi.org/10.1145/2783258.2788613.

[47] M.T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 2016, pp. 1135-1144. https://doi.org/10.1145/2939672.2939778.

[48] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, et al., From Local Explanations to Global Understanding with Explainable AI for Trees, Nat. Mach. Intell. 2 (2020), 56-67. https://doi.org/10.1038/s42256-019-0138-9.

[49] S.M. Lundberg, S.I. Lee, A Unified Approach to Interpreting Model Predictions, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

[50] J.H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, Ann. Stat. 29 (2001), 1189-1232. https://doi.org/10.1214/AOS/1013203451.

[51] D.W. Apley, J. Zhu, Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models, J. R. Stat. Soc. Ser. B: Stat. Methodol. 82 (2020), 1059-1086. https://doi.org/10.1111/RSSB.12377.

[52] S. Kaufman, S. Rosset, C. Perlich, O. Stitelman, Leakage in Data Mining, ACM Trans. Knowl. Discov. Data 6 (2012), 1-21. https://doi.org/10.1145/2382577.2382579.

[53] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, et al., Model Cards for Model Reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA, 2019, pp. 220-229. https://doi.org/10.1145/3287560.3287596.

[54] C. Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, Nat. Mach. Intell. 1 (2019), 206-215. https://doi.org/10.1038/s42256-019-0048-x.

[55] J. Hippisley-Cox, C.A.C. Coupland, M. Bafadhel, R.E.K. Russell, A. Sheikh, et al., Development and Validation of a New Algorithm for Improved Cardiovascular Risk Prediction, Nat. Med. 30 (2024), 1440-1447. https://doi.org/10.1038/s41591-024-02905-y.

[56] M. Sud, A. Sivaswamy, A. Chu, P.C. Austin, T.J. Anderson, et al., Population-Based Recalibration of the Framingham Risk Score and Pooled Cohort Equations, J. Am. Coll. Cardiol. 80 (2022), 1330-1342. https://doi.org/10.1016/J.JACC.2022.07.026.

[57] K.R. van Daalen, D. Zhang, S. Kaptoge, E. Paige, E. Di Angelantonio, et al., Risk Estimation for the Primary Prevention of Cardiovascular Disease: Considerations for Appropriate Risk Prediction Model Selection, Lancet Glob. Health 12 (2024), e1343-e1358. https://doi.org/10.1016/S2214-109X(24)00210-9.