Journal Healthcare Treatment Development
Vol: 06, No. 02, July-Dec 2026
ISSN: 2799-1148, DOI: 10.55529/jhtd.62.1.14

Research Paper

An interpretable integer score for angiographic coronary artery disease in the Cleveland cohort: derivation and cross-validated comparison with machine-learning models

Inamzhanova Gulnara[1*], Matsakov Kanybek[2], Begaly uulu Samidin[3]
[1*,3]Department of Clinical Disciplines, Jalal-Abad International University, 720906, 24 Kurmanbek Str., Manas, Kyrgyzstan.
[2]Department of Morphological Disciplines, Jalal-Abad International University, 720906, 24 Kurmanbek Str., Manas, Kyrgyzstan.

Article Info

Article History:
Received: 10 March 2026
Revised: 21 May 2026
Accepted: 28 May 2026
Published: 18 November 2026

Keywords:
Coronary Artery Disease
Diagnostic Prediction
Clinical Risk Score
Decision Curve Analysis
Cleveland Heart Disease Data

ABSTRACT

Background: Machine-learning classifiers are often proposed for the non-invasive diagnosis of coronary artery disease (CAD), but they are rarely compared with simple scores that clinicians can compute by hand.
Objective: To compare a transparent integer points score with six alternative models for the diagnosis of angiographic CAD.
Methods: The public Cleveland Clinic angiography cohort was used (303 patients, 138 with at least 50% stenosis in a major coronary artery). Nine clinically dichotomised items were defined (male sex, age ≥55 years, asymptomatic chest pain, exercise-induced angina, ST depression ≥1 mm, flat or downsloping ST segment, maximum heart rate <140 bpm, at least one vessel visible on fluoroscopy, and a thallium perfusion defect), weighted by rounding twice their logistic coefficients, and recalibrated to absolute risk. The score was compared with a unit-weight item count, logistic regression with all 13 predictors, L1-penalised logistic regression, random forest, gradient boosting and a support vector machine. Evaluation used 20 repeats of stratified five-fold cross-validation, with all derivation steps confined to training folds.
Results: The points score (range 0–19) achieved a mean area under the receiver operating characteristic curve (AUC) of 0.915, a Brier score of 0.113 and a calibration slope of 0.88. Full logistic regression performed similarly (AUC 0.912), whereas random forest (0.902), support vector machine (0.894) and gradient boosting (0.887) had lower AUC than the score in every one of the 20 repeats. Observed CAD prevalence rose from 0% (0/30) at 0–2 points to 96.8% (60/62) at ≥15 points. Decision curve analysis showed net benefit for the score comparable to or greater than that of the ensemble models across threshold probabilities of 0.10–0.70. The score was well calibrated in both women (observed-to-expected ratio 1.00) and men (1.00), whereas the random forest over-predicted risk in women.

Check for updates

Journal Healthcare Treatment Development (JHTD)

Conclusions: In this dataset, a transparent 19-point score performed at least as well as machine-learning ensembles while offering advantages in interpretability and bedside use. External validation in contemporary cohorts is required before clinical application.

Corresponding Author:
Inamzhanova Gulnara
Department of Clinical Disciplines, Jalal-Abad International University, 720906, 24 Kurmanbek Str., Manas, Kyrgyzstan.
Email: 0309asem@gmail.com

Copyright © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Cardiovascular diseases remain the leading cause of death worldwide, and ischaemic heart disease accounts for the largest share of that burden [1]. For a patient with chest pain or suspected myocardial ischaemia, the first clinical question is the probability that obstructive coronary artery disease (CAD) is present. Current guidelines use this pre-test probability to decide whether a patient needs no further testing, non-invasive functional or anatomical imaging, or invasive coronary angiography [2], [3]. Accurate and well-calibrated estimates of CAD probability therefore have direct consequences for patient safety, radiation exposure and health-care costs.

Probability estimation in this setting has a long history. Diamond and Forrester showed in 1979 that age, sex and the character of chest pain could be combined with Bayes' theorem to estimate the likelihood of CAD [4]. Detrano and colleagues extended this approach with exercise test and fluoroscopic findings and evaluated a new probability algorithm in patients referred for angiography in Cleveland, Budapest, Long Beach and Switzerland [5]. Later models updated and extended the classical predictors for contemporary populations [6]. In parallel, integer points scores such as the Framingham risk categories and the HEART score for emergency chest pain have shown that multivariable models can be translated into simple sums that clinicians can compute without a computer [7], [8], [9].

The Cleveland subset of the Detrano data has since become one of the most popular benchmark datasets in medical machine learning. A large number of studies have applied random forests, gradient boosting, support vector machines and neural networks to it, often reporting high classification accuracy. Several features of this literature, however, limit its clinical value. Models are frequently compared only with each other rather than with a simple clinical score; accuracy at a single threshold is reported instead of calibration and clinical utility; and the internal validation is often a single train-test split, which is unstable in a sample of 303 patients [10], [11]. Moreover, the most widely circulated copy of the data has categorical codes and the outcome label permuted relative to the original documentation, which can silently invert the meaning of predictors if not checked.

There are good reasons to expect that complexity offers little here. A systematic review found no average performance benefit of machine learning over logistic regression in clinical prediction studies at low risk of bias [12], and the classical observation that "improper" unit-weighted linear models often perform almost as well as optimally weighted ones suggests that, when predictors are strong and the sample is small, the weighting matters less than the choice of predictors [13]. For high-stakes decisions, there is also a principled argument for preferring inherently interpretable models whenever they perform comparably [14], [15].

This study therefore asked whether an interpretable integer points score, built from nine clinically familiar items, can match machine-learning ensembles for the diagnosis of angiographic CAD in the Cleveland cohort. The specific aims were: (i) to derive a points score with transparent weights; (ii) to compare its discrimination, calibration and net benefit with six alternative models under leakage-free repeated cross-validation; and (iii) to examine performance separately in women and men, a subgroup in which CAD prediction is known to be more difficult [16].

2. METHODOLOGY

2.1 Data Source and Participants

The Cleveland Clinic Foundation cohort of the heart disease data of Detrano. was analysed [5]. It comprises 303 consecutive patients referred for coronary angiography, of whom 138 (45.5%) had angiographic CAD in the copy analysed here, defined as at least 50% diameter narrowing in at least one major coronary artery. The data are de-identified and publicly available, and no new data were collected.

The copy of the data used here is a widely distributed version in which several categorical variables and the outcome were recoded. Before analysis, every variable was mapped back to the coding of the original repository documentation and checked against its published frequencies. The outcome label was inverted relative to the original diagnosis variable; chest pain, resting electrocardiogram, ST-segment slope and thallium categories were relabelled; and five out-of-range values of the number of vessels coloured by fluoroscopy and two out-of-range thallium values were treated as missing, corresponding to the missing values documented in the original file. All category frequencies after recoding, including the outcome (165 without and 138 with disease, against 164 and 139 in the original file), matched the documented distribution to within one patient. The workflow of the study is shown in Figure 1.

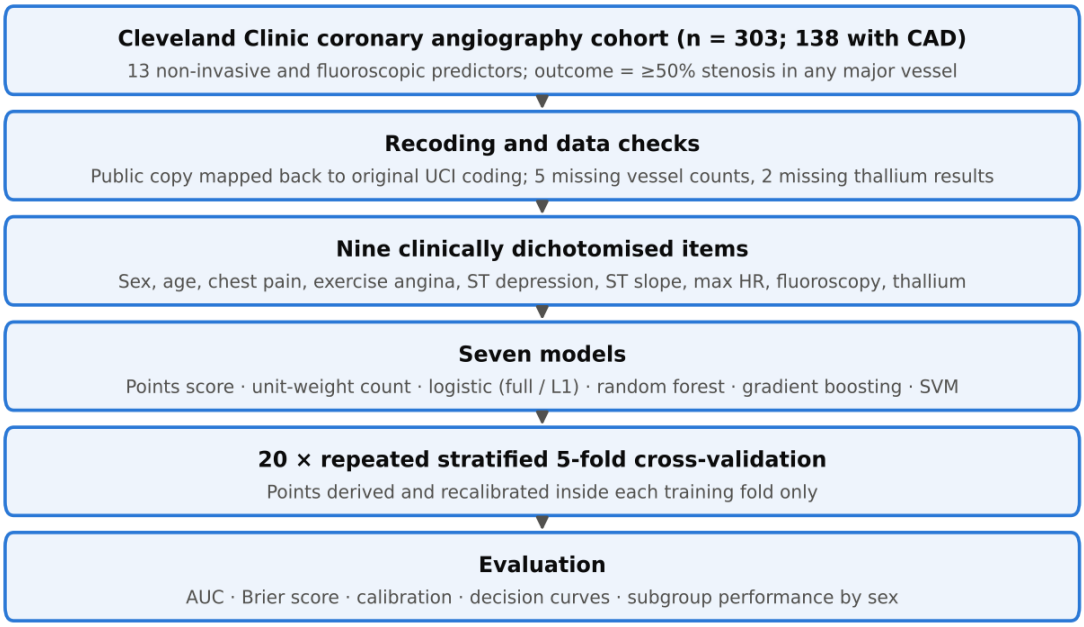


Figure 1. Study Workflow. All Derivation Steps, Including the Estimation of Item Weights, the Assignment of Points and the Recalibration of the Score, Were Repeated Inside Every Training Fold

2.2 Candidate Predictors

Thirteen predictors were available: age, sex, chest pain type (typical angina, atypical angina, non-anginal pain or asymptomatic), resting blood pressure, serum cholesterol, fasting blood glucose above 120 mg/dL, resting electrocardiogram (normal, ST-T wave abnormality or left ventricular hypertrophy), maximum heart rate achieved during exercise, exercise-induced angina, exercise-induced ST depression

relative to rest, slope of the peak exercise ST segment, number of major vessels (0–3) coloured by fluoroscopy and thallium scintigraphy result (normal, fixed defect or reversible defect).

2.3 Derivation of the Points Score

Nine binary items were defined a priori using conventional clinical cut-points rather than data-driven thresholds: male sex; age 55 years or older; asymptomatic chest pain; exercise-induced angina; ST depression of 1 mm or more; flat or downsloping peak exercise ST segment; maximum heart rate below 140 beats per minute; at least one vessel coloured by fluoroscopy; and any thallium perfusion defect. Dichotomisation loses information [17], but it is how clinicians usually record and communicate these findings and is a prerequisite for a hand-calculable score. Resting blood pressure, cholesterol, fasting glucose and resting electrocardiogram were not included because of their weak univariable association with angiographic disease in this cohort (Figure 2) and in the original report [5].

The nine items were entered in an unpenalised logistic regression. Following the approach used for Framingham-type scores [8], each item was assigned points equal to twice its coefficient rounded to the nearest integer, so that one point corresponds to approximately 0.5 units on the log-odds scale; items with a negative coefficient would have received zero points. The total score was then converted to absolute risk by a logistic recalibration of the outcome on the score. A simpler variant was also evaluated: a unit-weight count of the nine items, which ignores the estimated weights entirely [13] and was converted to risk by the same logistic recalibration.

2.4 Comparator Models

Five comparator models used all 13 predictors, with median imputation of continuous and most-frequent imputation of categorical missing values, and dummy coding of categorical predictors: (a) logistic regression with a weak ridge penalty ($C = 1$) and standardised continuous predictors; (b) L1-penalised (lasso) logistic regression ($C = 0.3$) [18]; (c) a random forest of 500 trees with at least three observations per leaf [19]; (d) histogram-based gradient boosting with learning rate 0.05, 150 iterations and at most six leaves per tree [20]; and (e) a support vector machine with a radial basis function kernel and Platt-scaled probability outputs [21]. Hyperparameters were fixed at conventional, moderately regularised values to avoid optimism from tuning on a small sample. All analyses used Python with scikit-learn [22] and statsmodels.

2.5 Validation and Performance Measures

Performance was estimated by stratified five-fold cross-validation repeated 20 times, giving 100 fits per model. Every derivation step, including imputation, scaling, estimation of item coefficients, assignment of points and recalibration of the score, was performed within the training fold only. The 138 events provide about 15 events per parameter for the nine-item score but fewer than 8 per parameter for the dummy-coded 13-predictor models, so the sample is adequate for a compact model but leaves limited margin for more flexible learners [23].

Discrimination was assessed by the area under the receiver operating characteristic curve (AUC); overall accuracy by the Brier score [24]; and calibration by the calibration slope and calibration intercept, together with calibration plots based on octiles of predicted risk [25]. Sensitivity and specificity were reported at a probability threshold of 0.5. Metrics were computed for each repeat on the pooled out-of-fold predictions and summarised as the mean and the 2.5th–97.5th percentile range over the 20 repeats; paired differences in AUC relative to the points score were computed within repeats, and the number of repeats in which the comparator exceeded the score was recorded. Clinical utility was assessed with decision curve analysis [26]. Observed risk within score bands was summarised with Wilson 95% confidence intervals [27]. Finally, discrimination and the observed-to-expected (O/E) ratio were calculated separately for women and men. Reporting follows the TRIPOD+AI recommendations where applicable [11], [28].

3. RESULTS AND DISCUSSION

3.1 Patient Characteristics

Table 1 describes the cohort. Patients with CAD were older (56.6 vs 52.5 years), more often male (82.6% vs 56.4%), more often had asymptomatic chest pain (75.4% vs 23.6%) and exercise-induced angina (55.1% vs 13.9%), had a lower maximum heart rate (139 vs 159 bpm), greater ST depression (1.6 vs 0.6 mm), more vessels coloured by fluoroscopy (1.1 vs 0.3) and a much higher frequency of reversible thallium defects (64.5% vs 17.0%). Resting blood pressure and cholesterol differed only marginally ($p = 0.035$ and $p = 0.036$), and fasting glucose did not differ ($p = 0.74$).

Table 1. Baseline Characteristics by Angiographic Coronary Artery Disease Status

Characteristic	All (n = 303)	No CAD (n = 165)	CAD (n = 138)	p
Age, years	54.4 ± 9.1	52.5 ± 9.6	56.6 ± 8.0	<0.001
Resting BP, mmHg	131.6 ± 17.5	129.3 ± 16.2	134.4 ± 18.7	0.035
Serum cholesterol, mg/dL	246.3 ± 51.8	242.2 ± 53.6	251.1 ± 49.5	0.036
Maximum heart rate, bpm	149.6 ± 22.9	158.5 ± 19.2	139.1 ± 22.6	<0.001
ST depression, mm	1.0 ± 1.2	0.6 ± 0.8	1.6 ± 1.3	<0.001
Vessels on fluoroscopy, n	0.7 ± 0.9	0.3 ± 0.6	1.1 ± 1.0	<0.001
Male sex	207 (68.3)	93 (56.4)	114 (82.6)	<0.001
Fasting glucose >120 mg/dL	45 (14.9)	23 (13.9)	22 (15.9)	0.744
Exercise-induced angina	99 (32.7)	23 (13.9)	76 (55.1)	<0.001
Chest pain				<0.001
Typical angina	23 (7.6)	16 (9.7)	7 (5.1)	
Non-anginal pain	87 (28.7)	69 (41.8)	18 (13.0)	
Atypical angina	50 (16.5)	41 (24.8)	9 (6.5)	
Asymptomatic	143 (47.2)	39 (23.6)	104 (75.4)	
Resting ECG				0.007
LV hypertrophy	147 (48.5)	68 (41.2)	79 (57.2)	
Normal	152 (50.2)	96 (58.2)	56 (40.6)	
ST-T abnormality	4 (1.3)	1 (0.6)	3 (2.2)	
Peak exercise ST slope				<0.001
Downsloping	21 (6.9)	9 (5.5)	12 (8.7)	
Upsloping	142 (46.9)	107 (64.8)	35 (25.4)	
Flat	140 (46.2)	49 (29.7)	91 (65.9)	
Thallium scan				<0.001
Fixed defect	18 (5.9)	6 (3.6)	12 (8.7)	
Normal	166 (54.8)	130 (78.8)	36 (26.1)	
Reversible defect	117 (38.6)	28 (17.0)	89 (64.5)	

The pattern of association is summarized in Figure 2. The strongest rank correlations with CAD were for a thallium defect ($\rho = 0.53$), asymptomatic chest pain (0.52), number of vessels on fluoroscopy (0.49), exercise-induced angina (0.44), maximum heart rate (−0.43) and ST depression (0.42), whereas resting blood pressure, cholesterol and fasting glucose were weakly correlated (0.12, 0.12 and 0.03). Several exercise variables were moderately correlated with each other, notably ST depression and flat or downsloping ST slope ($\rho = 0.58$) and maximum heart rate with ST slope (−0.47).

Figure 3 shows how strongly CAD prevalence varied across the categorical findings. Prevalence was 73% in patients with asymptomatic chest pain but only 18–30% in those with typical, atypical or non-anginal pain. This inversion of the usual clinical gradient probably reflects the referral setting, in which patients without symptoms were referred for angiography because of abnormal test results. CAD was

present in 22% of patients with a normal thallium scan compared with 76% of those with a reversible defect, and rose from 26% with no fluoroscopy to 85% with three.

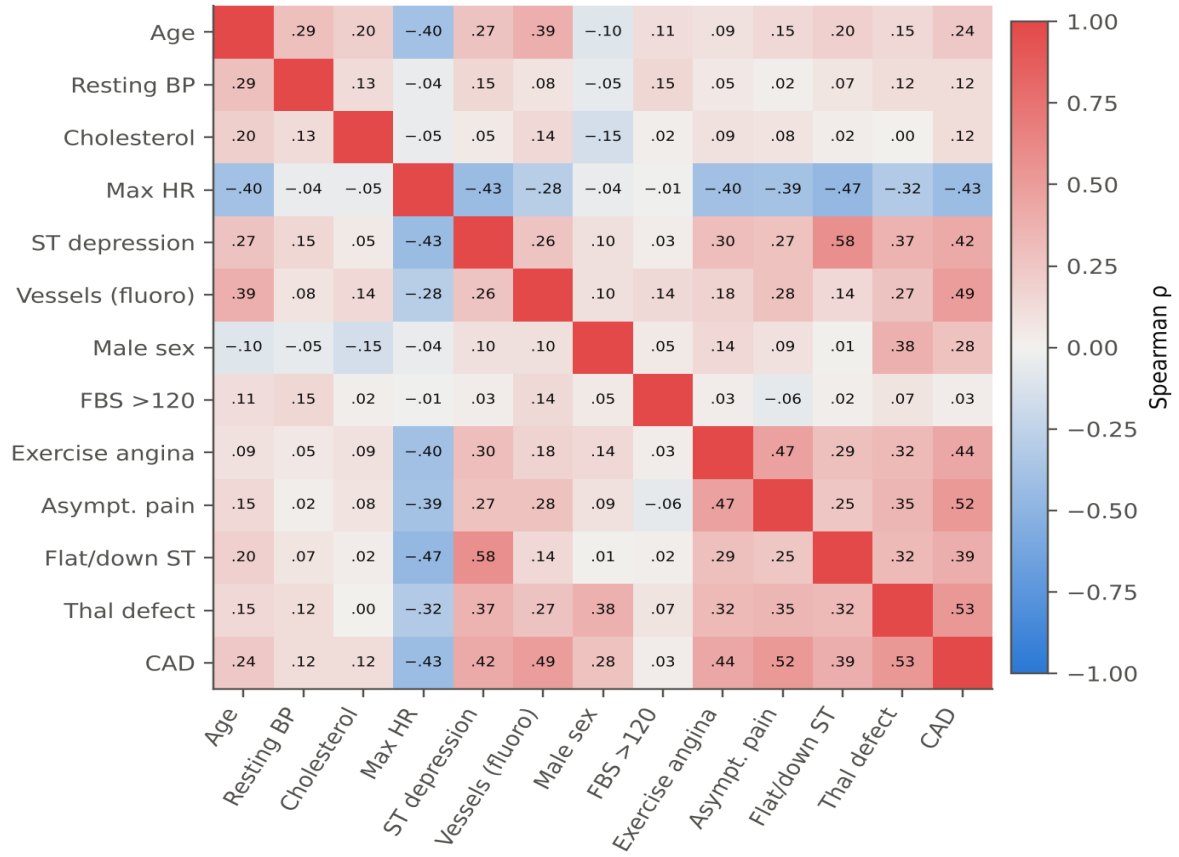


Figure 2. Spearman Rank Correlations Among Continuous Predictors, Binary Predictors, Three Dichotomised Score Items and the Outcome (CAD). Red Indicates Positive and Blue Negative Correlation

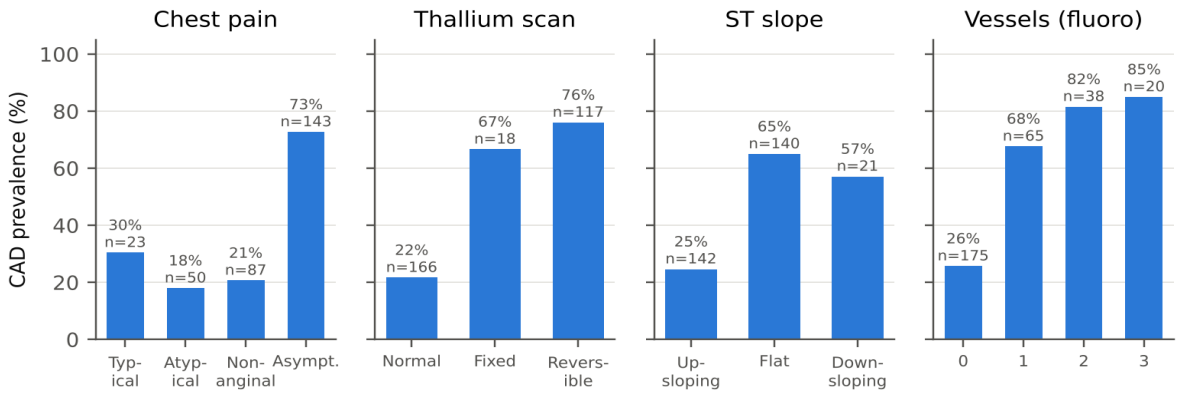


Figure 3. Prevalence of Angiographic CAD across Categories of Chest Pain, Thallium Scan, Peak Exercise ST Slope and Number of Vessels Coloured by Fluoroscopy. Labels Give Prevalence and the Number of Patients in Each Category

3.2 The Points Score

The multivariable coefficients and assigned points are shown in Table 2 and Figure 4. At least one vessel on fluoroscopy (adjusted OR 8.91, 95% CI 4.02–19.75) and asymptomatic chest pain (5.79, 2.70–12.43) received four points each; male sex (3.76, 1.51–9.38) received three points; thallium defect, flat or downsloping ST segment and exercise-induced angina received two points each; and age 55 years or older

and ST depression of at least 1 mm received one point each. Maximum heart rate below 140 bpm added little information once the other exercise findings were known (OR 1.22, 0.50–2.96) and received zero points, so the score in practice uses eight of the nine items and ranges from 0 to 19.

Table 2. Items of the Points Score: Prevalence, Adjusted Log-odds and Assigned Points (Full-sample Fit)

Item	Prevalence (%)	β	Adjusted OR (95% CI)	p	Points
Male sex	68.3	1.32	3.76 (1.51–9.38)	0.005	3
Age ≥55 years	52.5	0.54	1.72 (0.81–3.67)	0.160	1
Asymptomatic chest pain	47.2	1.76	5.79 (2.70–12.43)	<0.001	4
Exercise-induced angina	32.7	0.85	2.34 (1.03–5.32)	0.043	2
ST depression ≥1 mm	45.2	0.56	1.76 (0.80–3.88)	0.163	1
Flat/downsloping ST	53.1	1.16	3.18 (1.36–7.40)	0.007	2
Max HR <140 bpm	28.1	0.20	1.22 (0.50–2.96)	0.665	0
≥1 vessel on fluoroscopy	40.6	2.19	8.91 (4.02–19.75)	<0.001	4
Thallium defect	44.6	1.20	3.32 (1.53–7.21)	0.002	2
Maximum possible score					19

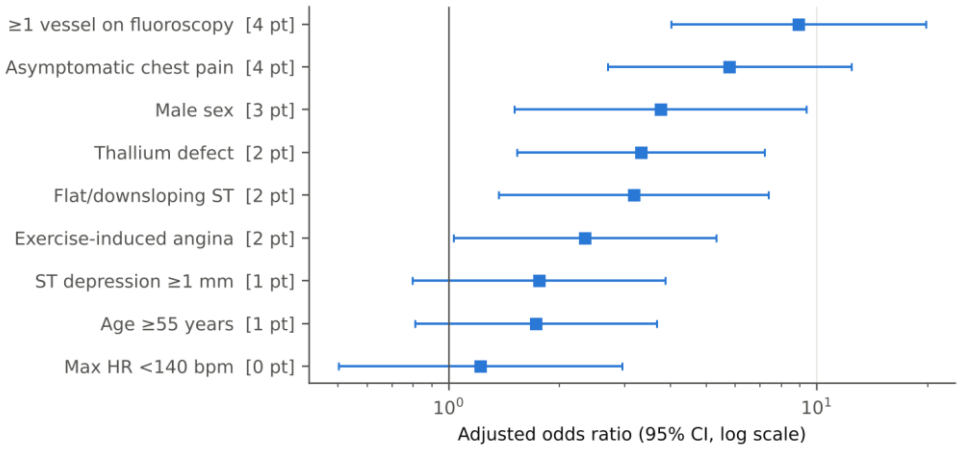


Figure 4. Adjusted Odds Ratios (95% Confidence Intervals, Log Scale) of the Nine Score Items from the Full-sample Logistic Model. The Number of Points Assigned to Each Item Is Shown in Brackets

The relationship between the score and risk was steep and monotonic (Table 3 and Figure 5). None of the 30 patients with 0–2 points had CAD (95% CI 0–11.4%), and only 6.7% of the 45 patients with 3–4 points did. Risk crossed 50% between 9–10 points (37.0%) and 11–12 points (73.5%) and reached 92.9% at 13–14 points and 96.8% at 15 points or more. The recalibrated risk function, logit(p) = −4.91 + 0.507 × score, followed the observed proportions closely, with every predicted range overlapping the corresponding confidence interval; the largest deviation occurred at 9–10 points, where the model predicted 42–54% against an observed 37% (95% CI 22–56%).

Table 3. Observed and Model-predicted Risk of Coronary Artery Disease across Points-score Bands

Score Band	n	CAD Events	Observed % (95% Wilson CI)	Predicted % (Range)
0–2	30	0	0.0 (0.0–11.4)	0.7–2.0
3–4	45	3	6.7 (2.3–17.9)	3.3–5.3
5–6	36	4	11.1 (4.4–25.3)	8.5–13.4
7–8	41	10	24.4 (13.8–39.3)	20.5–29.9
9–10	27	10	37.0 (21.5–55.8)	41.5–54.1
11–12	34	25	73.5 (56.9–85.4)	66.2–76.5
13–14	28	26	92.9 (77.4–98.0)	84.4–90.0

Score Band	n	CAD Events	Observed % (95% Wilson CI)	Predicted % (Range)
≥15	62	60	96.8 (89.0–99.1)	93.7–99.1

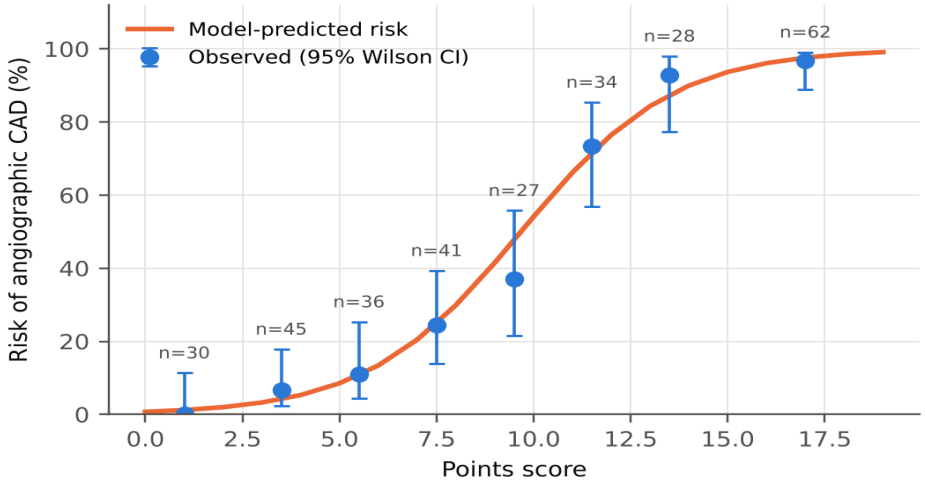


Figure 5. Observed Prevalence of CAD (Points with 95% Wilson Confidence Intervals) by Points-score Band and the Risk Predicted by the Recalibrated Score (Line). Numbers Above the Points Give the Number of Patients per Band.

3.3 Cross-Validated Comparison with Other Models

Table 4 summarises the cross-validated performance of all seven models. The points score achieved the highest mean AUC, 0.915 (range across repeats 0.905–0.922), and the lowest Brier score, 0.113. Full logistic regression with all 13 predictors performed almost identically (AUC 0.912, Brier 0.115), and the unit-weight count of the nine items was only slightly lower (0.908, Brier 0.117), demonstrating that most of the predictive information lies in the choice of items rather than in their exact weights.

Table 4. Cross-validated Performance of Seven Models (Mean over 20 Repeats of Stratified 5-fold Cross-validation)

Model	AUC (2.5th–97.5th Pct.)	Brier	Cal. Slope	Cal. Intercept	Sens. %	Spec. %
Points score (9 items)	0.915 (0.905–0.922)	0.113	0.88	+0.001	81.6	87.1
Unit-weight count	0.908 (0.905–0.911)	0.117	0.98	+0.001	80.4	87.3
Logistic regression (13 vars)	0.912 (0.904–0.919)	0.115	0.92	-0.007	79.4	87.8
L1 logistic regression	0.900 (0.893–0.907)	0.125	1.13	-0.027	78.2	84.9
Random forest	0.902 (0.893–0.909)	0.129	1.38	+0.008	78.1	84.7
Gradient boosting	0.887 (0.872–0.897)	0.135	0.80	-0.002	78.4	84.6
SVM (RBF)	0.894 (0.880–0.903)	0.128	1.01	+0.021	77.8	85.8

The machine-learning models were consistently inferior. Relative to the points score, the mean paired difference in AUC was −0.014 for the random forest (range −0.023 to −0.008), −0.022 for the support vector machine (−0.030 to −0.013) and −0.028 for gradient boosting (−0.041 to −0.020), and none of these models exceeded the score in any of the 20 repeats. The L1-penalised logistic regression also performed slightly worse (−0.015), probably because the penalty strength shrank informative coefficients in this small sample. Full logistic regression exceeded the score in only 4 of 20 repeats (mean difference −0.004, range −0.011 to +0.004). The ROC curves of the first repeat (Figure 6) illustrate how closely the curves of the score and logistic regression track each other, and how the ensembles fall slightly below them over parts of the curve.

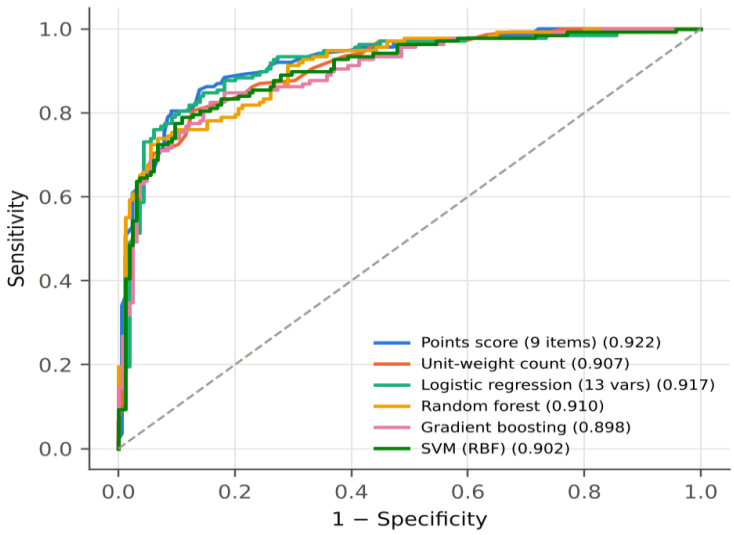


Figure 6. Receiver Operating Characteristic Curves of Six Models from the First Cross-validation Repeat; the AUC of Each Model in That Repeat Is Given in Brackets. Table 4 Reports Means Over All 20 Repeats

At the 0.5 probability threshold, the points score had a sensitivity of 81.6% and a specificity of 87.1%, both higher than those of the random forest (78.1% and 84.7%) and gradient boosting (78.4% and 84.6%).

3.4 Calibration and Clinical Utility

Calibration differed more between models than discrimination did (Figure 7a). The points score and the unit-weight count were close to ideal on average (calibration intercepts 0.001 and slopes 0.88 and 0.98). The random forest had a calibration slope of 1.38, indicating predictions compressed towards the middle of the risk range: it under-estimated risk in high-risk patients and over-estimated it in low-risk patients. Gradient boosting showed the opposite pattern, with a slope of 0.80 indicating over-confident, overly extreme predictions. The support vector machine, whose outputs were explicitly calibrated by Platt scaling, had a slope close to one (1.01) but lower discrimination.

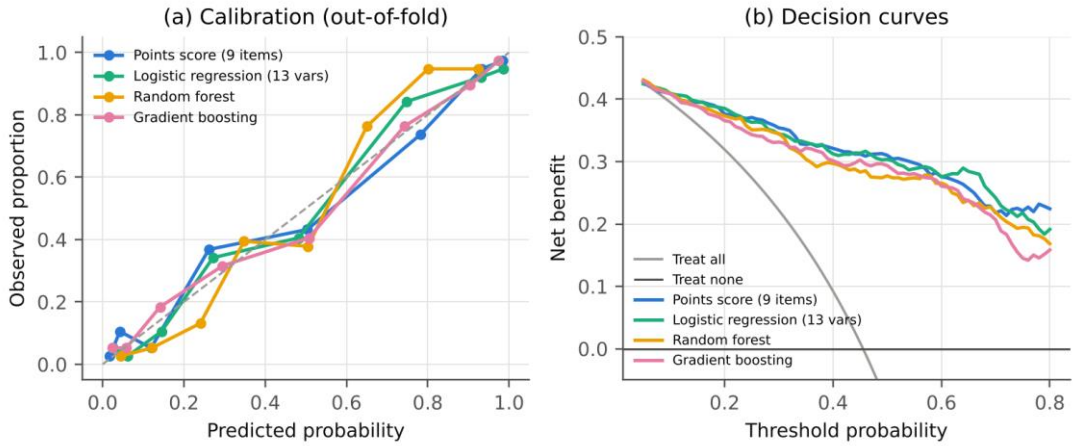


Figure 7. (a) Calibration of the points score, logistic regression, random forest and gradient boosting, based on out-of-fold predictions averaged over 20 repeats and grouped into octiles. (b) Decision curves showing net benefit across threshold probabilities, compared with treating (testing) all patients and none

Decision curves Figure 7b) showed that all models provided greater net benefit than referring to every patient for angiography at thresholds of 0.10 and above. The advantage grew rapidly with the threshold: at 0.30, the net benefit of the points score was 0.354 against 0.222 for referring everyone,

equivalent to avoiding about 31 unnecessary angiograms per 100 patients without missing any additional disease (net reduction in interventions = (0.354 − 0.222) × (1 − 0.3)/0.3 × 100). Across thresholds of 0.10–0.70, the net benefit of the score was equal to or higher than that of the random forest and gradient boosting at almost every threshold (for example, 0.310 vs 0.277 and 0.294 at a threshold of 0.50) and within about 0.025 of full logistic regression.

3.5 Performance in Women and Men

Results by sex are shown in Table 5. CAD prevalence was 25.0% in women and 55.1% in men. All three models evaluated discriminated better in women than in men, which probably reflects case mix rather than prevalence itself: a larger proportion of women were clearly low risk, so cases and non-cases were easier to separate. The points score was well calibrated in both sexes (O/E ratio 1.00 in each), and full logistic regression was nearly so (0.94 in women and 1.01 in men). The random forest, however, over-predicted risk in women by about a quarter (mean predicted risk 31.5% vs observed 25.0%; O/E 0.79) and under-predicted it slightly in men (O/E 1.06). Such sex-specific miscalibration would systematically lead to over-testing of women if the forest were used to guide referral.

Table 5. Performance by Sex (Out-of-fold Predictions Averaged over 20 Repeats)

Sex	Model	n	CAD (%)	Mean Predicted	O/E	AUC	Brier
Women	Points score	96	24 (25.0%)	25.0%	1.00	0.942	0.070
Women	Logistic regression	96	24 (25.0%)	26.7%	0.94	0.932	0.073
Women	Random forest	96	24 (25.0%)	31.5%	0.79	0.928	0.090
Men	Points score	207	114 (55.1%)	55.0%	1.00	0.893	0.130
Men	Logistic regression	207	114 (55.1%)	54.4%	1.01	0.889	0.132
Men	Random forest	207	114 (55.1%)	51.9%	1.06	0.877	0.145

3.6 Discussion

In the Cleveland angiography cohort, a nine-item integer score, in effect eight items summing to at most 19 points, discriminated angiographic CAD as well as logistic regression with all 13 predictors and better than random forest, gradient boosting and a support vector machine, while being well calibrated overall and within each sex. The advantage of the score over the ensembles was small in absolute terms (0.014–0.028 in AUC) but highly consistent, appearing in all 20 cross-validation repeats, and it extended to calibration and net benefit, which are the properties that determine whether model-guided decisions are better than default strategies [25], [26].

These findings agree with the broader evidence that flexible learners rarely outperform well-specified logistic models on small clinical datasets with strong, largely additive predictors [12]. In this cohort, most of the diagnostic signal comes from a handful of findings, chiefly fluoroscopic calcification, thallium defects, the symptom pattern, exercise ECG changes and male sex, whose effects appear to combine roughly additively on the log-odds scale. Tree ensembles spend their flexibility searching for interactions and non-linearities that the data cannot support, which increases variance; the random forest compensates by averaging, which produced the compressed probabilities seen in its calibration slope of 1.38, while gradient boosting over-fitted, which produced over-confident predictions. The near-equivalence of the unit-weight count and the weighted score echoes the classic observation that, once the right predictors are chosen, the precise weights matter surprisingly little [13].

The interpretability of the score has practical value beyond its accuracy. Every point can be traced to a familiar clinical finding, the arithmetic can be done at the bedside or in a referral form, and implausible results can be checked by a clinician. The score also makes the structure of the decision explicit: a patient with no fluoroscopic calcification, a normal thallium scan and non-anginal pain cannot exceed 9 points, whereas any combination of calcification and asymptomatic presentation already reaches 8. Such transparency supports shared decision-making and audit and is particularly desirable for high-stakes diagnostic decisions [14]. Optimisation-based methods for integer scores could improve sparsity further

[15], but the simple rounding approach used here, standard in the Framingham tradition [8], already performed as well as the full logistic model.

The sex-specific results deserve attention. CAD in women is often under-recognised, presents with atypical symptoms and has a different relationship to non-invasive test results [16]. In this cohort, the points score retained calibration in women, whereas the random forest over-estimated their risk by about a quarter. The higher AUC in women than in men contrasts with the expectation that diagnosis is harder in women; it should be interpreted cautiously, because it rests on only 24 events in women and reflects a referral population in which most women had a clearly low-risk profile. A model that is well calibrated for the whole cohort can thus be poorly calibrated for an important subgroup, a pattern that is invisible in studies reporting only overall accuracy, and subgroup calibration should be part of every evaluation of a diagnostic model [11].

Several features of the score limit its direct clinical use. It includes the number of vessels coloured by fluoroscopy and a thallium scan, which are investigations in their own right rather than pre-test information, and these two items carry six of the 19 points. The score is therefore best understood as a post-test integration of exercise testing, myocardial perfusion and fluoroscopy findings, not as a first-line pre-test probability tool of the type recommended by current guidelines [2], [3], [6]. Second, the high prevalence of CAD among asymptomatic patients reflects referral for angiography after abnormal tests, a form of verification bias that would not apply in an unselected chest-pain population; the weight given to asymptomatic presentation should not be transferred to primary care. Third, the cohort is decades old, small and drawn from a single referral centre, so both the absolute risks and the relative weights would need to be re-estimated in a contemporary population before use [10], [29].

3.6.1 Strengths and Weaknesses of the Study

The strengths of the study include a careful audit and recoding of the widely circulated version of the dataset, an a priori definition of score items with conventional cut-points, a comparison with six alternative models under identical leakage-free repeated cross-validation, and a focus on calibration, net benefit and subgroup performance rather than accuracy alone.

The limitations follow largely from the data. With 303 patients and 138 events, all performance estimates are internal and carry non-trivial uncertainty, and the variability observed across repeats reflects partition rather than full sampling variability. Hyperparameters of the machine-learning models were fixed rather than tuned; careful nested tuning might narrow the gap, but on a dataset of this size it would also add optimism and instability [23]. Dichotomising continuous predictors sacrifices information [17], although the score lost nothing measurable relative to the full continuous model. The few missing values were imputed simply, which is unlikely to matter given their small number. Finally, external validation, for example in the Hungarian, Swiss and Long Beach series collected alongside the Cleveland data [5] or in modern registries, is needed to establish transportability.

4. CONCLUSION

A transparent integer score built from nine conventional clinical, exercise-test and imaging items, eight of which carry points summing to a maximum of 19, matched logistic regression and outperformed random forest, gradient boosting and support vector machine models for the diagnosis of angiographic CAD in the Cleveland cohort. It was well calibrated overall and in both sexes and provided at least as much net benefit as the ensembles across clinically relevant thresholds. These results argue that, for small clinical datasets with strong additive predictors, interpretable scores should be the default benchmark against which machine-learning methods are judged, and that calibration and subgroup performance should be reported alongside discrimination. Prospective external validation in contemporary chest-pain populations is the necessary next step.

Acknowledgment

The authors acknowledge Dr Robert Detrano and colleagues at the Cleveland Clinic Foundation, who collected the original data, and the curators who made the heart disease data publicly available.

Funding Information

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Inamzhanova Gulnara	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓
Matsakov Kanybek	✓	✓		✓		✓	✓	✓	✓	✓		✓	✓	✓
Begaly uulu Samidin	✓		✓	✓	✓		✓	✓		✓	✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this article.

Informed Consent

Not applicable. This study analysed publicly available, de-identified data; no participants were recruited.

Ethical Approval

This study is a secondary analysis of a fully de-identified, publicly available dataset. No new data were collected from human participants, and institutional ethical approval and informed consent were therefore not required.

Data Availability

The Cleveland heart disease data analysed in this study are publicly available from the UCI Machine Learning Repository and its open mirrors. The recoded dataset and the analysis code (Python, scikit-learn) that reproduce all tables and figures are available from the corresponding author on reasonable request.

Use of AI Tools

Generative AI tools were used to assist with code development, statistical analysis and drafting of the manuscript. The authors reviewed and verified all analyses, results and text and take full responsibility for the content of this article.

REFERENCES

[1] Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. J Am Coll Cardiol. 2020;76(25):2982–3021. doi.org/10.1016/j.jacc.2020.11.010

[2] Knuuti J, Wijns W, Saraste A, Capodanno D, Barbato E, Funck-Brentano C, et al. 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes. Eur Heart J. 2020;41(3):407–477. doi.org/10.1093/eurheartj/ehz425

[3] Gulati M, Levy PD, Mukherjee D, Amsterdam E, Bhatt DL, Birtcher KK, et al. 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR guideline for the evaluation and diagnosis of chest pain. Circulation. 2021;144(22):e368–e454. doi.org/10.1161/CIR.0000000000001029
[4] Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. N Engl J Med. 1979;300(24):1350–1358. doi.org/10.1056/NEJM197906143002402
[5] Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. Am J Cardiol. 1989;64(5):304–310. doi.org/10.1016/0002-9149(89)90524-9
[6] Genders TSS, Steyerberg EW, Alkadhi H, Leschka S, Desbiolles L, Nieman K, et al. A clinical prediction rule for the diagnosis of coronary artery disease: validation, updating, and extension. Eur Heart J. 2011;32(11):1316–1330. doi.org/10.1093/eurheartj/ehr014
[7] Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837–1847. doi.org/10.1161/01.CIR.97.18.1837
[8] Sullivan LM, Massaro JM, D'Agostino RB Sr. Presentation of multivariate data for clinical use: the Framingham Study risk score functions. Stat Med. 2004;23(10):1631–1660. doi.org/10.1002/sim.1742
[9] Six AJ, Backus BE, Kelder JC. Chest pain in the emergency room: value of the HEART score. Neth Heart J. 2008;16(6):191–196. doi.org/10.1007/BF03086144
[10] Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd ed. Cham: Springer; 2019. doi.org/10.1007/978-3-030-16399-0
[11] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi.org/10.1136/bmj-2023-078378
[12] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. doi.org/10.1016/j.jclinepi.2019.02.004
[13] Dawes RM. The robust beauty of improper linear models in decision making. Am Psychol. 1979;34(7):571–582. doi.org/10.1037/0003-066X.34.7.571
[14] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–215. doi.org/10.1038/s42256-019-0048-x
[15] Ustun B, Rudin C. Supersparse linear integer models for optimized medical scoring systems. Mach Learn. 2016;102(3):349–391. doi.org/10.1007/s10994-015-5528-6
[16] Shaw LJ, Bugiardini R, Merz CNB. Women and ischemic heart disease: evolving knowledge. J Am Coll Cardiol. 2009;54(17):1561–1575. doi.org/10.1016/j.jacc.2009.04.098
[17] Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006;332(7549):1080. doi.org/10.1136/bmj.332.7549.1080
[18] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 1996;58(1):267–288. doi.org/10.1111/j.2517-6161.1996.tb02080.x
[19] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. doi.org/10.1023/A:1010933404324
[20] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–1232. doi.org/10.1214/aos/1013203451
[21] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–297. doi.org/10.1007/BF00994018
[22] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–2830. doi.org/10.48550/arXiv.1201.0490
[23] 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;368:m441. doi.org/10.1136/bmj.m441

[24] Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1–3. doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
[25] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. doi.org/10.1186/s12916-019-1466-7
[26] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–574. doi.org/10.1177/0272989X06295361
[27] Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927;22(158):209–212. doi.org/10.1080/01621459.1927.10502953
[28] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594. doi.org/10.1136/bmj.g7594
[29] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691–698. doi.org/10.1136/heartjnl-2011-301247

How to Cite: Inamzhanova Gulnara, Matsakov Kanybek, Begaly uulu Samidin (2026). An interpretable integer score for angiographic coronary artery disease in the Cleveland cohort: derivation and cross-validated comparison with machine-learning models. Journal Healthcare Treatment Development, 6(2), 1–14. https://doi.org/10.55529/jhtd.62.1.14

BIOGRAPHIES OF AUTHOR

	**Inamzhanova Gulnara**, is a Lecturer in the Department of Clinical Disciplines at Jalal-Abad International University, Manas, Kyrgyzstan, where the author teaches and conducts research in clinical medicine. For this article, Inamzhanova Gulnara was responsible for conceptualization, data curation, development of the integer risk score, formal analysis, writing – original draft. Email: 0309asem@gmail.com
	**Matsakov Kanybek**, is an Associate Professor in the Department of Morphological Disciplines at Jalal-Abad International University, Manas, Kyrgyzstan, where the author teaches and conducts research in the morphological (basic medical) sciences. For this article, Matsakov Kanybek was responsible for methodology, clinical interpretation of angiographic and cardiovascular variables, supervision, writing – review and editing. Email: kanybekmatsakov@gmail.com
	**Begaly uulu Samidin**, is a Lecturer in the Department of Clinical Disciplines at Jalal-Abad International University, Manas, Kyrgyzstan, where the author teaches and conducts research in clinical medicine. For this article, Begaly uulu Samidin was responsible for software, training and comparison of machine-learning ensembles, validation, visualization, writing – review and editing. Email: begalyuulusamidin@gmail.com