Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Interpretable Integer Risk Score for Osteosarcoma Survival: Development and External Validation in a Multicenter Cohort

Zhihao Ma
Chinese PLA General Hospital
Jiayan Wu
Chinese PLA General Hospital
Dajiang Sun
Hebei North University
Yifei Zheng
Hebei North University
Liting Li
Chinese PLA General Hospital
Junxiao Su
Tsinghua University
Wenhao Hu
Luckyhao555@163.com

Chinese PLA General Hospital
Xuesong Zhang
Chinese PLA General Hospital

Article

Keywords: machine learning, risk score, survival, AutoScore, osteosarcoma

Posted Date: October 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11281109/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Osteosarcoma prognosis is heterogeneous, and existing risk stratification tools are limited. We aimed to develop and externally validate an interpretable integer-based scoring system for predicting overall survival in patients with osteosarcoma.

Methods

This retrospective prognostic study used data from the Surveillance, Epidemiology, and End Results (SEER) database for model development (2010 to 2022) and an independent multicenter Chinese cohort for external validation (2010 to 2024). The SEER cohort was randomly split into training (70%) and internal validation (30%) sets. Candidate predictors included demographic, clinical, and treatment-related variables available at diagnosis. The AutoScore-Survival framework was applied to derive an integer score. Model performance was assessed using C-index, time-dependent AUC, calibration metrics, and decision curve analysis.

Results

The development cohort included 2,755 patients (906 deaths, 32.9%) and the external validation cohort 816 patients (314 deaths, 38.5%). The final score (0 to 19 points) included seven variables: lung metastasis, tumor grade, age, tumor size, primary site, lymph node metastasis, and chemotherapy. In external validation, the score achieved a C-index of 0.705 (95% CI, 0.675–0.735) and time-dependent AUCs of 0.729, and 0.742 at 3, and 5 years. Calibration was good at 3 years (O/E ratio = 1.156), with some underestimation at 5 years (O/E = 1.368). Decision curve analysis showed positive net benefit at 3 and 5 years.

Conclusions

This integer risk score provides a practical and interpretable tool for osteosarcoma risk stratification with acceptable discrimination and calibration in external validation. Future prospective multi-center studies are needed to confirm its clinical utility across broader populations and to explore integration of molecular markers for improved model performance.

Introduction

Osteosarcoma is the most common primary malignant bone tumor, accounting for a substantial proportion of cancer-related morbidity and mortality in children and adolescents [1, 2]. Despite the multimodal therapy combining neoadjuvant chemotherapy, extensive resection, and postoperative

chemotherapy having made progress, the prognosis of osteosarcoma remains highly heterogeneous. Patients with localized disease have a 5-year overall survival (OS) rate of 60% to 70%, whereas those with metastatic disease have only approximately 20% [3, 4]. This disparity underscores the need for accurate and personalized tools to guide risk stratification and treatment planning.

Machine learning (ML) offers advantages over traditional Cox regression in capturing non-linear relationships and handling high-dimensional data. Several studies have applied ML models for osteosarcoma and reported good discrimination. However, most of them lack independent external validation, leaving their generalizability uncertain [5, 6]. Furthermore, the “black-box” nature of many ML models limits their interpretability and clinical adoption [7]. Consequently, there is a conspicuous lack of a widely applicable and highly interpretable risk assessment tool for osteosarcoma patients in the bedside setting. The AutoScore-Survival framework provides a practical tool for long-term risk stratification [8]. The scoring system derived from this framework is simpler and more interpretable [9, 10]. It balances discriminative ability with calibration and uses a point-based format, making it easy to facilitate clinical implementation. Although AutoScore-Survival has been successfully applied in other clinical settings, it has not yet been used in osteosarcoma [11–13].

To fill this gap, we used the AutoScore-Survival to develop an interpretable survival prediction model for osteosarcoma and performed independent external validation in a multicenter Chinese cohort. We aimed to construct a simple and practical risk score to predict prognosis of osteosarcoma, thereby facilitating individualized postoperative management and resource allocation.

Patients/Methods

Data Source and Patient Population

This retrospective prognostic study followed the TRIPOD + AI guideline [14]. The development cohort was derived from the Surveillance, Epidemiology, and End Results (SEER) database, which is widely used for cancer survival research owing to its high data quality and long-term follow-up. We extracted data on all patients with a first diagnosis of osteosarcoma between 2010 and 2022 [15]. The external validation cohort comprised patients from three Chinese hospitals. Consecutive patients diagnosed with osteosarcoma between January 2010 and December 2024 were retrospectively enrolled. The end of follow-up was March 2026. Inclusion criteria were (1) histologically confirmed primary osteosarcoma, (2) underwent surgical treatment, and (3) complete survival records. Exclusion criteria were (1) secondary osteosarcoma, (2) prior or concurrent other malignancies, and (3) loss to follow-up or unknown vital status. Finally, the development cohort included 2,755 patients and the external validation cohort 816 patients (Fig. 1). The SEER cohort was randomly split into a training (70%, n = 1,928) and an internal validation (30%, n = 827) sets. The study was approved by the ethics committee of our Hospital. Written informed consent was waived owing to the retrospective and de-identified nature of the data.

Outcomes Definition and Candidate Variables

The primary outcome was OS, defined as the time from diagnosis to death from any cause. Survival were evaluated at 1, 3, and 5 years after diagnosis. In the SEER cohort, vital status and survival time were derived from the “Cause of Death and Follow-up” category. In the external cohort, vital status was obtained through active telephone follow-up. We conducted a systematic review of the variables available in the SEER database and identified thirteen candidate variables. Demographic features included age, sex, and race and ethnicity. Clinical parameters included primary site, histologic subtype, grade, tumor size, distant metastases (lung, lymph node, and other sites), time from diagnosis to treatment, vital status, treatment-related variables, and OS. Twenty-five SEER cases with a survival time of 0 months were set to 0.5 months to satisfy survival model requirements. To ensure diagnostic consistency and contemporaneity, all primary cases were reclassified in accordance with the histologic subtypes defined by the WHO Classification of Tumors of Soft Tissue and Bone (5th Edition, 2020) [16]. The treatment groups were categorized as surgery alone, neoadjuvant chemotherapy followed by surgery (Preop CT), surgery followed by adjuvant chemotherapy (Postop CT), or perioperative chemotherapy (Periop CT). The complete variable definitions were presented in **Supplement** Table 1.

Table 1
Baseline characteristics of patients in the study cohorts

Predictor	ALL (n = 3571)	Training Set (n = 1928)	Internal validation set (n = 827)	External validation set (n = 816)	P Value
Age, y	17.0 (25.0)	17.0 (30.0)	17.0 (25.0)	17.0 (15.0)	< 0.001
Sex					
Male	2041 (57.2)	1070 (55.5)	454 (54.9)	517 (63.4)	< 0.001
Female	1530 (42.8)	858 (44.5)	373 (45.1)	299 (36.6)	
Race					
American Indian/Alaska Native	26 (0.7)	17 (0.9)	9 (1.1)	0 (0.0)	< 0.001
Asian or Pacific Islander	1098 (30.7)	202 (10.5)	80 (9.7)	816 (100.0)	
Black	382 (10.7)	263 (13.6)	119 (14.4)	0 (0.0)	
White	2045 (57.3)	1431 (74.2)	614 (74.2)	0 (0.0)	
Missing/Unknown	20 (0.6)	15 (0.8)	5 (0.6)	0 (0.0)	
Primary site					
Appendicular skeleton	2865 (80.2)	1475 (76.5)	645 (78.0)	745 (91.3)	< 0.001
Axial skeleton	706 (19.8)	453 (23.5)	182 (22.0)	71 (8.7)	
Histologic subtype					
9180/3	3172 (88.8)	1710 (88.7)	734 (88.8)	728 (89.2)	< 0.001
9184/3	60 (1.7)	37 (1.9)	16 (1.9)	7 (0.9)	
9187/3	34 (1.0)	7 (0.4)	5 (0.6)	22 (2.7)	
9192/3	98 (2.7)	56 (2.9)	25 (3.0)	17 (2.1)	

The SEER cohort had a follow-up range of 0.5–155.0 months, and the external validation cohort had a follow-up range of 4.4–248.4 months. Data are presented as Median (Interquartile Range) or n (%). P values were calculated using Kruskal-Wallis test (continuous variables) or chi-square test (categorical variables) comparing the three cohorts. Preop, preoperative; Postop, postoperative; Periop, perioperative.

Predictor	ALL (n = 3571)	Training Set (n = 1928)	Internal validation set (n = 827)	External validation set (n = 816)	P Value
9193/3	42 (1.2)	19 (1.0)	3 (0.4)	20 (2.5)	
9194/3	165 (4.6)	99 (5.1)	44 (5.3)	22 (2.7)	
Tumour grade					
Low	462 (12.9)	229 (11.9)	77 (9.3)	156 (19.1)	< 0.001
High	2541 (71.2)	1338 (69.4)	586 (70.9)	617 (75.6)	
Missing/Unknown	568 (15.9)	361 (18.7)	164 (19.8)	43 (5.3)	
Tumour size, mm	82.0 (57.0)	84.0 (60.0)	85.0 (65.0)	75.0 (45.0)	< 0.001
Lung metastases					
No	3077 (86.2)	1679 (87.1)	697 (84.3)	701 (85.9)	0.008
Yes	464 (13.0)	228 (11.8)	121 (14.6)	115 (14.1)	
Missing/Unknown	30 (0.8)	21 (1.1)	9 (1.1)	0 (0.0)	
Lymph node metastases					
No	3312 (92.7)	1776 (92.1)	761 (92.0)	775 (95.0)	< 0.001
Yes	87 (2.4)	45 (2.3)	12 (1.5)	30 (3.7)	
Missing/Unknown	172 (4.8)	107 (5.5)	54 (6.5)	11 (1.3)	
Other site metastases					
No	3444 (96.4)	1893 (98.2)	814 (98.4)	737 (90.3)	< 0.001
Yes	101 (2.8)	15 (0.8)	7 (0.8)	79 (9.7)	

The SEER cohort had a follow-up range of 0.5–155.0 months, and the external validation cohort had a follow-up range of 4.4–248.4 months. Data are presented as Median (Interquartile Range) or n (%). P values were calculated using Kruskal-Wallis test (continuous variables) or chi-square test (categorical variables) comparing the three cohorts. Preop, preoperative; Postop, postoperative; Periop, perioperative.

Predictor	ALL (n = 3571)	Training Set (n = 1928)	Internal validation set (n = 827)	External validation set (n = 816)	P Value
Missing/Unknown	26 (0.7)	20 (1.0)	6 (0.7)	0 (0.0)	
Chemotherapy					
No	532 (14.9)	349 (18.1)	127 (15.4)	56 (6.9)	< 0.001
Preop chemotherapy	658 (18.4)	439 (22.8)	202 (24.4)	17 (2.1)	
Postop chemotherapy	488 (13.7)	299 (15.5)	122 (14.8)	67 (8.2)	
Periop chemotherapy	1890 (52.9)	839 (43.5)	376 (45.5)	675 (82.7)	
Missing/Unknown	3 (0.1)	2 (0.1)	0 (0.0)	1 (0.1)	
Radiotherapy					
No	3266 (91.5)	1757 (91.1)	760 (91.9)	749 (91.8)	< 0.001
Preop radiotherapy	39 (1.1)	14 (0.7)	9 (1.1)	16 (2.0)	
Postop radiotherapy	260 (7.3)	155 (8.0)	58 (7.0)	47 (5.8)	
Periop radiotherapy	2 (0.1)	2 (0.1)	0 (0.0)	0 (0.0)	
Missing/Unknown	4 (0.1)	0 (0.0)	0 (0.0)	4 (0.5)	
Diagnosis to treatment interval, d	14.0 (25.0)	16.0 (24.0)	16.0 (26.8)	8.0 (27.0)	< 0.001
Vital status					
Alive	2351 (65.8)	1294 (67.1)	555 (67.1)	502 (61.5)	0.013
Death	1220 (34.2)	634 (32.9)	272 (32.9)	314 (38.5)	
Median survival, m	42.0 (72.0)	39.0 (73.0)	39.0 (68.8)	60.0 (89.0)	< 0.001

The SEER cohort had a follow-up range of 0.5–155.0 months, and the external validation cohort had a follow-up range of 4.4–248.4 months. Data are presented as Median (Interquartile Range) or n (%). P values were calculated using Kruskal-Wallis test (continuous variables) or chi-square test (categorical variables) comparing the three cohorts. Preop, preoperative; Postop, postoperative; Periop, perioperative.

Sample Size and Missing Data

The minimum sample size was estimated using the events-per-variable criterion. With 13 candidate predictors (expanding to approximately 30 after one-hot encoding) and an events per variable threshold of 20, a minimum of 600 death events was required. Missing values were handled using the missForest algorithm, which predicts missing values by leveraging information from all other available variables and has shown favorable performance compared with other common imputation techniques [17, 18]. To prevent data leakage, the imputer was fitted exclusively on the training set and then applied to the internal and external validation sets.

Statistical Analysis

First, six supervised ML models were compared to identify the optimal backbone, including Ridge-penalized Cox regression (Ridge-Cox), Cox regression with a strong L2 penalty (Strong Ridge-Cox), Cox regression with an elastic net penalty (Elastic Net-Cox), Gradient Boosting Survival Analysis (GBSA), Componentwise Gradient Boosting for Cox regression (CompGB), and Fast Survival Support Vector Machine (FastSurvivalSVM). Continuous variables were standardized, and categorical variables were one-hot encoded. All preprocessing steps were fitted exclusively on the training set and applied to the validation sets. The optimal model was selected based on the highest Harrell's C-index in the internal validation set [19]. After the optimal model was determined, we applied the AutoScore-Survival framework to generate an interpretable scoring system. Model discrimination was assessed using Harrell’s C-index with 95% confidence intervals (CI) and time-dependent area under the receiver operating characteristic curve (AUC). Calibration was evaluated using calibration plots, calibration slope, observed-to-expected (O/E) ratio, and Brier score. Clinical utility was assessed using decision curve analysis. Patients were classified into risk groups based on score tertiles in the external validation cohort. Subgroup analyses were performed in the external validation cohort, with C-indexes calculated for each group. Feature importance was assessed using permutation importance on the external validation set. Hazard ratios and their ranking from the final model were reported to illustrate the direction and relative strength of each predictor. All analyses were performed using Python 3.14 with scikit-survival, pandas, numpy, matplotlib, and scikit-learn.

Results

Patient Characteristics and Cohort Description

Baseline characteristics of the development cohort (median survival 39.0 months, range 0.5–155.0 months) and the external cohort (median survival 60.0 months, range 4.4–248.4 months) are described in Table 1. In the development cohort, 906 patients (32.9%) died, compared with 314 patients (38.5%) in the external validation cohort. Patients in the external validation cohort had a higher proportion of appendicular primary sites (91.3% vs. 77.1%, $P < 0.001$), high-grade osteosarcoma (75.6% vs. 69.9%, $P < 0.001$), and perioperative chemotherapy (82.7% vs. 44.9%, $P < 0.001$). The overall Kaplan-Meier curves

showed similar trends between the two cohorts (P = 0.167, **Supplementary Fig. 1A**). Missing rates are shown in **Supplementary Fig. 1B**.

Model Development and External Validation

Detailed prediction performance of the 6 models was shown in Table 2. In the internal validation set, Ridge-Cox achieved the highest C-index of 0.707 (95% CI, 0.673–0.739) and was chosen as the primary model. Time-dependent AUCs for Ridge-Cox were 0.781 (95% CI, 0.720–0.840), 0.740 (95% CI, 0.697–0.780), and 0.729 (95% CI, 0.684–0.774) at 1, 3, and 5 years (Fig. 2A). The corresponding Brier scores were 0.075, 0.170, and 0.195 (Fig. 2B). In the external validation, Ridge-Cox achieved a C-index of 0.700 (95% CI, 0.670–0.727). Time-dependent AUCs were 0.781 (95% CI, 0.665–0.885) at 1 year, 0.721 (95% CI, 0.676–0.760) at 3 years, and 0.734 (95% CI, 0.695–0.773) at 5 years (Fig. 2A). Brier scores were 0.030, 0.190, and 0.250, respectively (Fig. 2B).

Table 2
Detailed prediction performance of machine learning models.

Internal Validation	Ridge-Cox	Strong Ridge-Cox	Elastic Net-Cox	GBSA	CompGB	FastSurvivalSVM
C-index (95% CI)	0.707 (0.673–0.739)	0.705 (0.673–0.738)	0.698 (0.664–0.732)	0.701 (0.671–0.736)	0.683 (0.648–0.718)	0.707 (0.673–0.736)
Time-dependent AUC at 1 year (95% CI)	0.781 (0.720–0.840)	0.778 (0.716–0.845)	0.763 (0.703–0.820)	0.778 (0.716–0.834)	0.743 (0.675–0.803)	0.781 (0.721–0.839)
Time-dependent AUC at 3 years (95% CI)	0.740 (0.697–0.780)	0.737 (0.695–0.782)	0.729 (0.687–0.771)	0.730 (0.685–0.773)	0.715 (0.667–0.758)	0.742 (0.704–0.783)
Time-dependent AUC at 5 years (95% CI)	0.729 (0.684–0.774)	0.728 (0.690–0.767)	0.722 (0.679–0.761)	0.720 (0.678–0.763)	0.707 (0.663–0.749)	0.727 (0.684–0.770)
External Validation						
C-index (95% CI)	0.700 (0.670–0.727)	0.702 (0.674–0.730)	0.698 (0.664–0.727)	0.711 (0.680–0.737)	0.664 (0.632-0.700)	0.709 (0.676–0.737)
Time-dependent AUC at 1 year (95% CI)	0.781 (0.665–0.885)	0.786 (0.689–0.877)	0.779 (0.680–0.863)	0.790 (0.696–0.875)	0.745 (0.649–0.838)	0.792 (0.698–0.885)
Time-dependent AUC at 3 years (95% CI)	0.721 (0.676–0.760)	0.725 (0.685–0.768)	0.720 (0.675–0.761)	0.732 (0.689–0.775)	0.677 (0.629–0.723)	0.731 (0.686–0.773)
Time-dependent AUC at 5 years (95% CI)	0.734 (0.695–0.773)	0.737 (0.695–0.778)	0.731 (0.692–0.768)	0.755 (0.718–0.789)	0.688 (0.642–0.730)	0.743 (0.699–0.782)

Data are point estimates with 95% CIs from 500 bootstrap resamples. The primary model was selected as the algorithm with the highest internal C-index among models supporting survival probability predictions. CI, confidence interval; AUC, area under the curve; Ridge-Cox, Ridge-penalized Cox regression; Strong Ridge-Cox, Cox regression with a strong L2 penalty; Elastic Net-Cox, Cox regression with an elastic net penalty; GBSA, Gradient Boosting Survival Analysis; CompGB,

Internal Validation	Ridge-Cox	Strong Ridge-Cox	Elastic Net-Cox	GBSA	CompGB	FastSurvivalSVM

Componentwise Gradient Boosting for Cox regression; FastSurvivalSVM, Fast Survival Support Vector Machine.

Generalizability and Interpretability of the Primary Model

The external validation cohort was divided into three equal-sized risk groups (n = 272 each) based on the predicted risks from the Ridge-Cox model. Kaplan-Meier curves showed early and sustained separation (P<0.001, **Supplementary Fig. 2**). For the death events at 1 year in the external cohort was limited (n = 19), we focused the risk stratification assessment at 3 and 5 years. External calibration curves are shown in Figs. 2C and 2D. Observed event rates increased stepwise with higher predicted risk: 3-year risks were 0.115, 0.168, and 0.430 for the low-, intermediate-, and high-risk groups, respectively; corresponding 5-year risks were 0.184, 0.276, and 0.607. At 5 years, observed risks exceeded predicted risks, suggesting systematic underestimation.

Permutation importance identified lung metastasis (mean decrease in C-index: 0.058) and tumor grade (0.047) as top contributors to model discrimination (**Supplement** Table 2). Hazard ratio ranking confirmed that lung metastasis and high tumor grade were the strongest predictors of mortality (**Supplementary Table 3**).

Development and Performance of the AutoScore-Survival Framework

The AutoScore-Survival framework was then applied to develop an integer scoring system for survival prediction. Model performance plateaued with seven variables (Fig. 3A), including lung metastasis, tumor grade, age at diagnosis, tumor size, primary site, lymph node metastasis, and chemotherapy. The resulting score ranged from 0 to 19, as shown in Table 3. Lung metastasis (4 points), high-grade tumor (5 points), age ≥ 60 years (4 points), and tumor size > 100 mm (3 points) received higher weights. Based on tertiles of the total score in the external validation cohort, patients were classified into low-risk (0–7 points; n = 417), intermediate-risk (8 points; n = 169), and high-risk (9–19 points; n = 230) groups. Kaplan-Meier curves showed clear gradient separation among the three groups (P<0.001, Fig. 3B). In external validation, the AutoScore achieved a C-index of 0.705 (95% CI, 0.675–0.735) and time-dependent AUCs of 0.778 (0.672–0.867), 0.729 (95% CI, 0.687–0.771), and 0.742 (95% CI, 0.706–0.780) at 1, 3, and 5 years (Table 4). These values were comparable to those of the full-variable Ridge-Cox model (external C-index 0.700), indicating that the simplified score maintained robust performance with fewer variables.

Table 3
AutoScore Integer Scoring System and Predicted Event Risk

AutoScore Scoring Table			
Variable	Category		Score points
Lung metastases	No/ Unknown		0
	Yes		4
Tumour grade	Low grade/ Unknown		0
	High grade		5
Age at diagnosis, y	< 19		0
	20–39		1
	40–59		3
	≥ 60		4
Tumour size, mm	≤ 50/ Unknown		0
	51–100		2
	> 100		3
Primary site	Appendicular skeleton		0
	Axial skeleton		2
Lymph node metastases	No/ Unknown		0
	Yes		2
Chemotherapy	Periop chemotherapy		0
	None/ Preop/ Postop/ Unknown		1
Risk stratification and probability			
Total points	Risk group	Predicted 3-year event risk	Predicted 5-year event risk
0–7	Low	10.7%	14.8%
8	Intermediate	19.1%	26.1%

Instructions: Sum the points for each patient. Locate the corresponding risk group and read the predicted probability of death (event risk) at 3, and 5 years after diagnosis. The predicted survival probability can be calculated as 1 minus the event risk. Risk group stratification was derived from tertiles of the total score distribution in the external validation cohort. Scores and predicted event risks were derived from the AutoScore-based Cox model. Total scores range from 0 to 19. Missing values are scored as 0 for lung metastasis, grade, tumour size, and lymph node; and as 1 for chemotherapy. Preop, preoperative; Postop, postoperative; Periop, perioperative.

AutoScore Scoring Table			
9–19	High	42.5%	53.0%

Instructions: Sum the points for each patient. Locate the corresponding risk group and read the predicted probability of death (event risk) at 3, and 5 years after diagnosis. The predicted survival probability can be calculated as 1 minus the event risk. Risk group stratification was derived from tertiles of the total score distribution in the external validation cohort. Scores and predicted event risks were derived from the AutoScore-based Cox model. Total scores range from 0 to 19. Missing values are scored as 0 for lung metastasis, grade, tumour size, and lymph node; and as 1 for chemotherapy. Preop, preoperative; Postop, postoperative; Periop, perioperative.

Table 4
AutoScore performance in the internal and external cohort

Discrimination ability	Internal Validation	External Validation
C-index (95% CI)	0.706 (0.672–0.737)	0.705 (0.675–0.735)
Time-dependent AUC at 1 year (95% CI)	0.778 (0.715–0.838)	0.778 (0.672–0.867)
Time-dependent AUC at 3 years (95% CI)	0.737 (0.693–0.780)	0.729 (0.687–0.771)
Time-dependent AUC at 5 years (95% CI)	0.721 (0.675–0.762)	0.742 (0.706–0.780)
Calibration ability		
Calibration slope at 3 years	0.951	0.809
Calibration slope at 5 years	0.861	0.840
O/E at 3 years	1.073	1.156
O/E at 5 years	1.163	1.368
Brier score at 3 years	0.169	0.188
Brier score at 5 years	0.194	0.248

Data are point estimates with 95% CIs from 1,000 bootstrap resamples. CI, confidence interval; AUC, area under the curve; O/E, observed-to-expected ratio

Calibration and Clinical Utility of the AutoScore-Survival Framework

Calibration plots are shown in Figs. 4A and 4B. The AutoScore was well calibrated at 3 years (O/E ratio = 1.156, calibration slope = 0.809), underestimated at 5 years (O/E = 1.368, calibration slope = 0.840). Decision curve analysis demonstrated a positive net benefit across a range of clinically meaningful threshold probabilities at 3 and 5 years (Figs. 4C and 4D).

Subgroup Analysis of the AutoScore-Survival Framework

Subgroup analysis (**Supplementary Table 4**) showed that the AutoScore performed better in women (C-index: 0.716), patients younger than 20 years (0.700), and those with low-grade tumors (0.732). However, discrimination was lower in patients with lung metastasis (0.551), high-grade tumors (0.662), and those aged 40 years or older (0.632). This suggested that the scoring system had limited ability to stratify high-risk patients with greater internal heterogeneity.

Discussion

In this study, we developed a simple and practical survival prediction score for osteosarcoma using the AutoScore-Survival. In external validation, the score achieved a C-index of 0.705, good calibration at 3 years, and positive net benefit on decision curve analysis. It uses seven routinely available variables and can be readily implemented in clinical decision-making.

ML offers advantages over traditional regression, but complex “black-box” model often hinder clinical adoption [20]. Most previous studies lacked independent external validation or had inadequate sample sizes, limiting their generalizability to patients from different regions [21]. Consequently, there is no widely accepted and interpretable risk score for osteosarcoma that could be rigorously validated. Our model aims to address this gap in current clinical practice. The AutoScore-Survival framework first ranks all candidate variables by importance using random survival forests and then converts the selected core variables into integer scores based on Cox regression, generating a simple clinical scoring system [8, 22]. Rather than pursuing maximal accuracy, we prioritized real-world applicability. Our score facilitates real-time risk stratification and shared decision-making. Our model finally achieved a C-index of 0.705, with AUCs of 0.729, and 0.742 at 3, and 5 years. Although lower than some published results, these values still reached a clinically acceptable level [6, 7]. One reason may be that our model used only routine clinical variables and did not include molecular markers or radiomic features. The SEER database lacks known strong prognostic factors such as tumor necrosis rate after chemotherapy, serum lactate dehydrogenase, and alkaline phosphatase. Without these key variables, the performance advantage of ML models is limited. Besides, osteosarcoma is characterized by substantial biological heterogeneity with considerable genomic and molecular diversity across patients [23]. These limitations also define a natural ceiling for predictive performance based on conventional variables.

Calibration showed good performance at 3 years (O/E = 1.156, slope = 0.809), with underestimation at 5 years (O/E = 1.368). We speculate that the 1-year overestimation is largely due to sparse events in the external cohort (n = 19). The 5-year underestimation may reflect underlying cohort differences. The external validation cohort had longer median survival (60.0 vs. 39.0 months) but higher mortality rate (38.5% vs. 32.9%), which may largely attributable to longer follow-up (up to 248.4 vs. 155.0 months). Longer follow-up captured more long-term events, leading to observed deaths exceeding those predicted. In addition, the external cohort had a higher proportion of patients receiving perioperative chemotherapy (82.7% vs. 44.9%), which may have also influenced long-term survival patterns. Decision curve analysis showed a positive net benefit at 3 and 5 years, confirming that the tool can be expected to guide clinical decisions for osteosarcoma.

Subgroup analysis showed limited discrimination in lung metastasis, older age, and high-grade tumor subgroups. One possible explanation is that these patients already have poor prognosis with limited survival variance, making further stratification difficult. In addition, factors not included in the score, such as metastatic burden of lung, comorbidities, chemotherapy responses, and the complexity of the osteosarcoma genome, may influence the prognosis [24–26]. These factors may dilute discriminative ability in high-risk subgroups. Therefore, although the AutoScore provides reliable risk stratification for the general population, its application in extremely high-risk subgroups should be complemented by clinical judgment.

Several limitations should be considered. First, the SEER database does not include several known prognostic factors (lactate dehydrogenase, alkaline phosphatase, Huvos grade, and immunohistochemical or molecular markers) [27–29]. Incorporating these variables might improve discrimination, particularly in high-risk subgroups. How to integrate these multi-dimensional prognostic factors to optimize the model is an important direction for future research. Second, our sample included only surgically treated patients. The outcomes of patients managed non-operatively remain unclear. For high-risk patients who may not derive substantial benefit from surgery due to comorbidities or extensive metastases, alternative treatment strategies should be carefully evaluated in the future. Future research should focus on prospective multi-center cohorts and external validation across different demographic subgroups, which are crucial for verifying the practicality of this tool in various patient populations and clinical settings.

Conclusions

We developed and externally validated an interpretable scoring system for predicting survival in osteosarcoma. The model achieves acceptable discrimination and calibration and provides a positive net benefit on decision curve analysis. This tool may facilitate personalized risk stratification and shared decision-making in clinical practice. Future prospective studies are warranted to confirm its utility across broader populations.

Abbreviations

ML, Machine learning; OS, Overall survival; AUC, Area under the receiver operating characteristic curve; CI, Confidence interval; HR, Hazard ratio; O/E, observed-to-expected.

Declarations

Ethics approval and consent to participate:

This research was approved by the ethics committee of PLA general hospital. Written informed consent was waived owing to the retrospective and deidentified nature of the data. All aspects of the study adhered to the Declaration of Helsinki.

Consent for publication:

Not applicable.

Funding:

This study was supported by the National Natural Science Foundation of China (Grant No. 12404501).

Author Contribution

ZM participated in the design of the study and drafted the manuscript. JW, DS, YZ, and LL prepared the data. ZM, JW, DS, and JS performed the statistical analysis. JS and WH conceived of the study and helped to draft the manuscript. XZ finally approved the submitted version. All authors read and approved the final manuscript.

Acknowledgements:

Not applicable.

Authors' information:

Data Availability

The datasets generated and analyzed during the current study are available in the Github repository, https://github.com/mazhihao0802/Mechain-Learning-Method.git

References

1. Biermann, J. S. et al. Bone Cancer, Version 2.2025, NCCN Clinical Practice Guidelines In Oncology. *J. Natl. Compr. Canc Netw.* ;**23**(4). (2025).
2. Kim, C. et al. Osteosarcoma in Pediatric and Adult Populations: Are Adults Just Big Kids? Cancers. (*Basel)* ;**15**(20). (2023).
3. Smeland, S. et al. Survival and prognosis with osteosarcoma: outcomes in more than 2000 patients in the EURAMOS-1 (European and American Osteosarcoma Study) cohort. *Eur. J. Cancer*. **109**, 36–50 (2019).
4. Yan, L., Dong, X. & Chen, S. Advances in osteosarcoma research: Pathogenesis and emerging therapeutic strategies. *Crit. Rev. Oncol. Hematol.* **222**, 105252 (2026).
5. Alharran, A. M. et al. Machine learning algorithms and web-based prognostic tool for different histological subtypes of osteosarcoma: a retrospective cohort. *Ann. Med. Surg. (Lond)*. **88** (1), 125–

130 (2025).
6. Cao, P. et al. Machine learning-based individualized survival prediction model for prognosis in osteosarcoma: Data from the SEER database. *Med. (Baltim).* **103** (39), e39582 (2024).
7. Hao, Y. et al. Machine learning for predicting the survival in osteosarcoma patients: Analysis based on American and Hebei Province cohort. *Biomol. Biomed.* **23** (5), 883–893 (2023).
8. Xie, F. et al. A Machine Learning-Based Automatic Clinical Score Generator and Its Application to Mortality Prediction Using Electronic Health Records. *JMIR Med. Inf.* **8** (10), e21798 (2020).
9. Yuan, H. et al. AutoScore-Imbalance: An interpretable machine learning tool for development of clinical scores with rare events data. *J. Biomed. Inf.* **129**, 104072 (2022).
10. Xie, F. et al. AutoScore-Survival: Developing interpretable machine learning-based time-to-event scores with right-censored survival data. *J. Biomed. Inf.* **125**, 103959 (2021).
11. Sharma, G. et al. Development and Validation of a Hybrid Machine Learning Model to Predict Lung Transplant Outcomes. *JAMA Netw. Open.* **8** (11), e2545369 (2025).
12. Chu, C. et al. Development and validation of an interpretable clinical scoring model to monitor the progression of preclinical Alzheimer's disease. *Alzheimers Res. Ther.* **17** (1), 268 (2025).
13. Ji, S. et al. Acute Care for Elders Risk Score: A Practical Machine Learning-Based Tool for Screening High-Risk Older Inpatients. *J. Am. Med. Dir. Assoc.* **26** (10), 105818 (2025).
14. Collins, G. S. et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
15. Surveillance, E. *and End Results (SEER) Program. SEER*Stat Database: Incidence - SEER Research Data, 17 Registries, Nov 2024 Sub (2000–2022)* (National Cancer Institute, DCCPS, Surveillance Research Program, 2025).
16. WHO Classification of Tumours Editorial Board. *WHO Classification of Tumours of Soft Tissue and Bone* 5th edn (International Agency for Research on Cancer, 2020).
17. Stekhoven, D. J. & Bühlmann, P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28** (1), 112–118 (2011).
18. Waljee, A. K. et al. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open.* ;**3**(8). (2013).
19. Verbakel, J. Y. et al. ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *J. Clin. Epidemiol.* **126**, 207–216 (2020).
20. Jiang, J. et al. Predictive model for the 5-year survival status of osteosarcoma patients based on the SEER database and XGBoost algorithm. *Sci. Rep.* **11** (1), 5542 (2021).
21. Spreafico, M., Hazewinkel, A. D., van de Sande, M. A. J., Gelderblom, H. & Fiocco, M. Machine Learning versus Cox Models for Predicting Overall Survival in Patients with Osteosarcoma: A Retrospective Analysis of the EURAMOS-1 Clinical Trial Data. *Cancers (Basel)* ;**16**(16). (2024).

22. Xie, F. et al. Development and Assessment of an Interpretable Machine Learning Triage Tool for Estimating Mortality After Emergency Admissions. *JAMA Netw. Open.* **4** (8), e2118467 (2021).
23. Zhou, Y. et al. Single-cell RNA landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. *Nat. Commun.* **11** (1), 6322 (2020).
24. Bull, E. C. et al. Targeting metastasis in paediatric bone sarcomas. *Mol. Cancer*. **24** (1), 153 (2025).
25. Mustokoweni, S. et al. Correlation of High-Grade Osteosarcoma Response to Chemotherapy with Enhanced Tissue Immunological Response: Analysis of CD95R, IFN-γ, Catalase, Hsp70, and VEGF. *Virchows Arch.* **484** (6), 925–937 (2024).
26. Espejo Valle-Inclan, J. et al. Ongoing chromothripsis underpins osteosarcoma genome complexity and clonal evolution. *Cell* ;**188**(2). (2025).
27. Bernthal, N. M. et al. Long-term results (> 25 years) of a randomized, prospective clinical trial evaluating chemotherapy in patients with high-grade, operable osteosarcoma. *Cancer* **118** (23), 5888–5893 (2012).
28. Rubio-San-Simón, A. et al. Prognostic factors in patients with relapsed high-grade osteosarcoma: a systematic review. *Br. J. Cancer*. **133** (7), 1020–1028 (2025).
29. Bacci, G. et al. Grade of chemotherapy-induced necrosis as a predictor of local and systemic control in 881 patients with non-metastatic osteosarcoma of the extremities treated with neoadjuvant chemotherapy in a single institution. *Eur. J. Cancer*. **41** (14), 2079–2085 (2005).

Figures

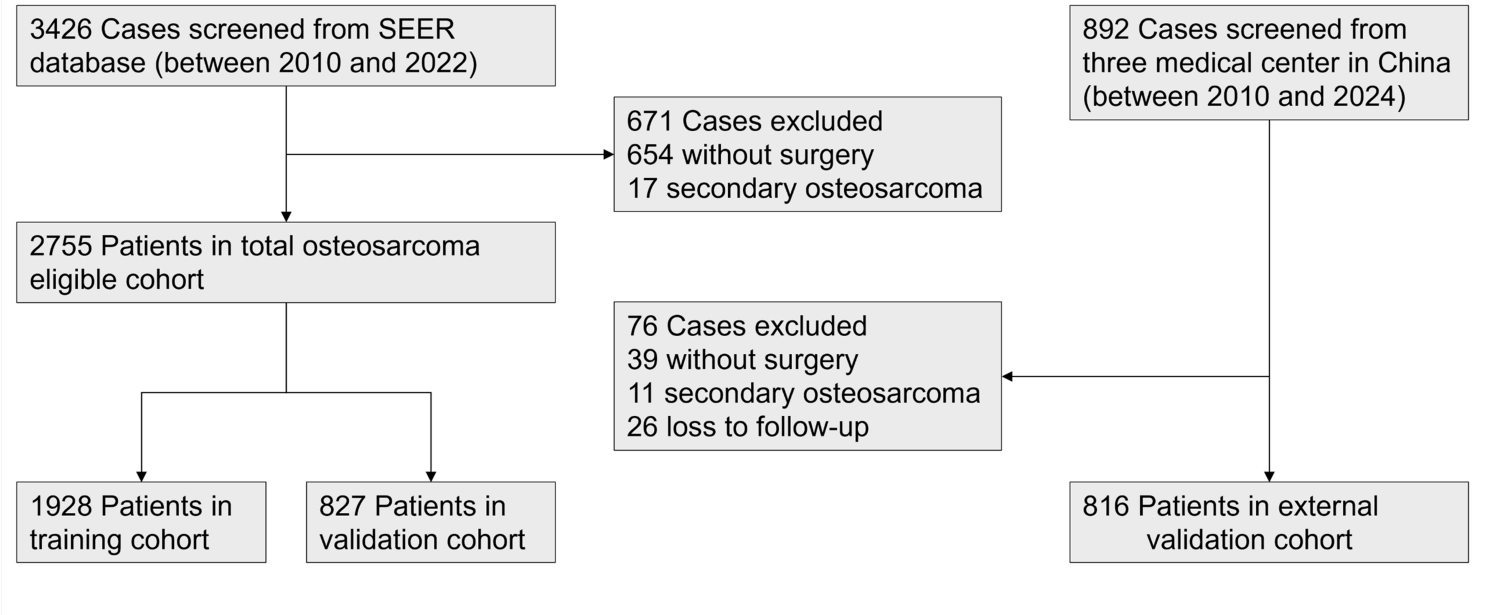


Figure 1

Study flow diagram for cohort selection and data splitting. A total of 2,755 patients from the SEER database (2010–2022) were included as the development cohort and randomly split into a training set (n = 1,928, 70%) and an internal validation set (n = 827, 30%) stratified by event status. An independent external validation cohort of 816 patients was derived from three Chinese tertiary referral centres (2010–2024). SEER indicates Surveillance, Epidemiology, and End Results.

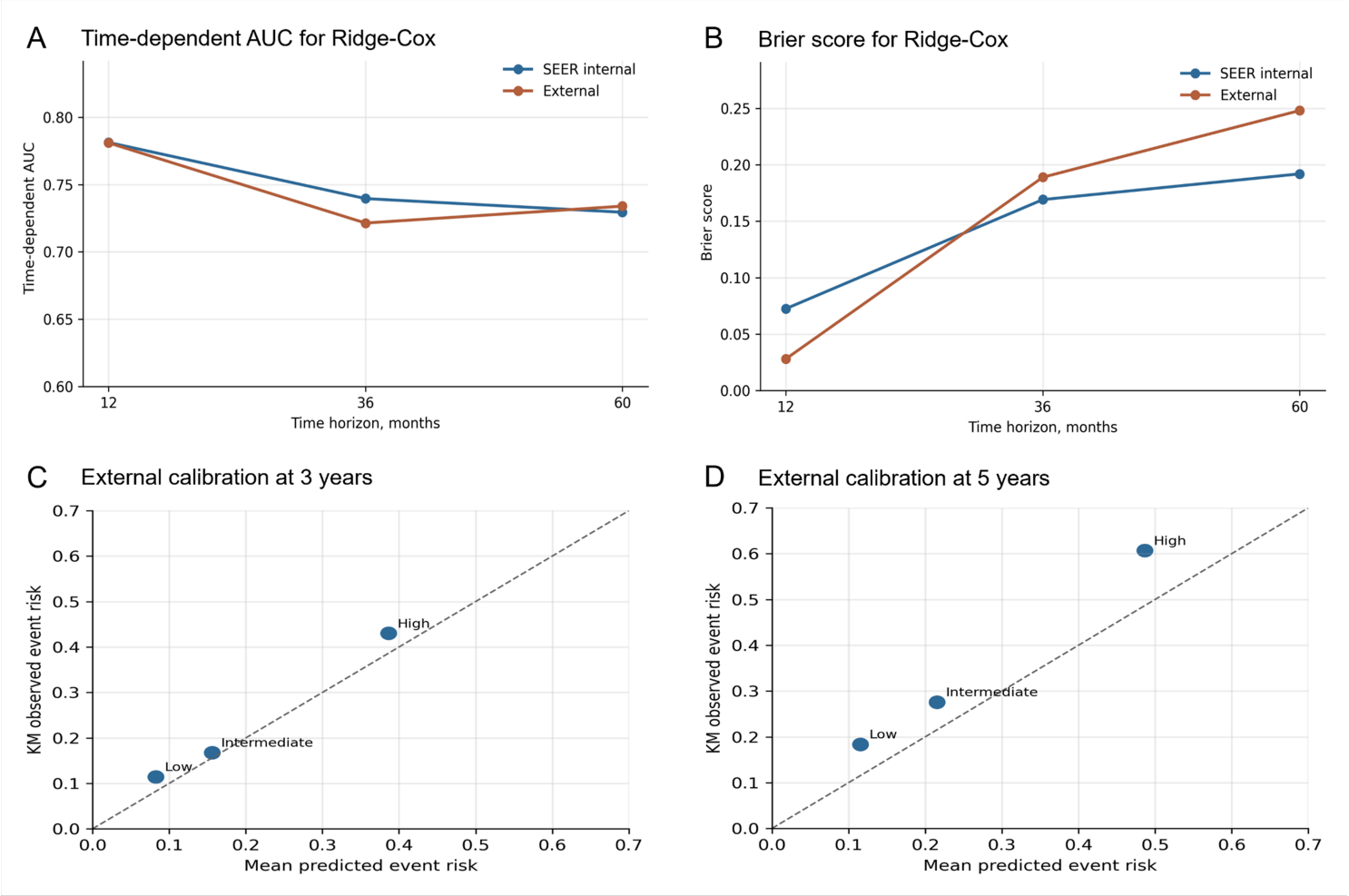


Figure 2

Performance of the primary Ridge-Cox mode. (A) Time-dependent AUC for the Ridge-Cox model in the internal and external validation cohorts at 1, 3, and 5 years. (B) Brier scores for the Ridge-Cox model at the same time points. Lower Brier scores indicate better predictive accuracy. (C-D) External calibration plots at 3 and 5 years. The external validation cohort was divided into three equal-sized risk groups (tertiles) based on predicted risks from the Ridge-Cox model. Mean predicted event risks are plotted against Kaplan-Meier observed event risks for each group. The diagonal dashed line represents perfect calibration. AUC, area under the curve.

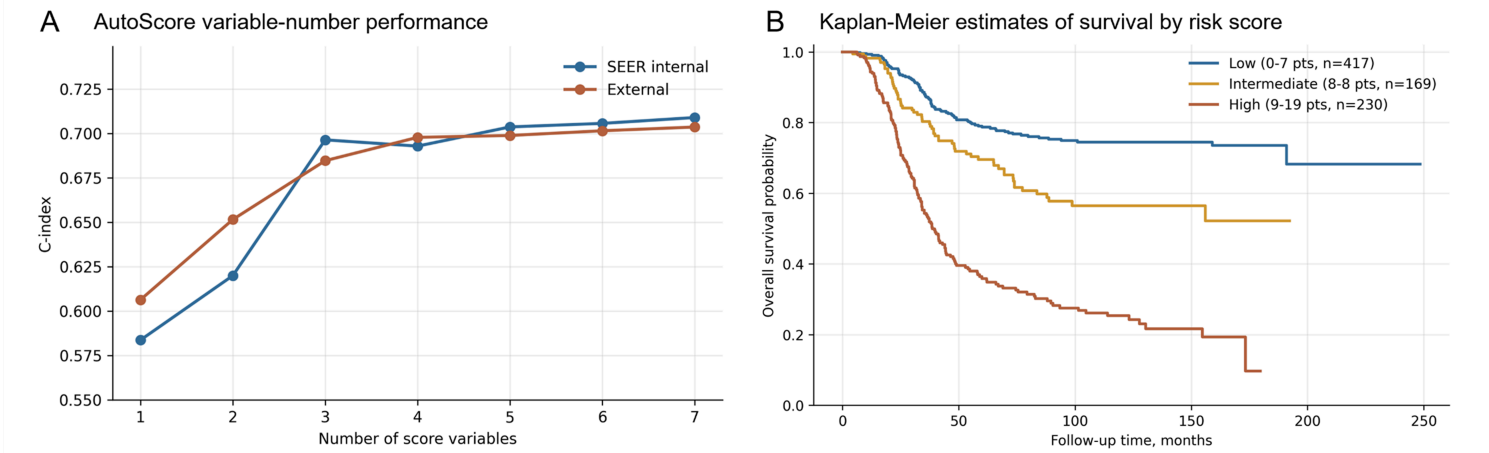


Figure 3

Development of the AutoScore integer scoring system. (A) C-index versus number of variables. Performance plateaued with seven variables. (B) Kaplan-Meier curves for low-, intermediate-, and high-risk groups in the external validation cohort.

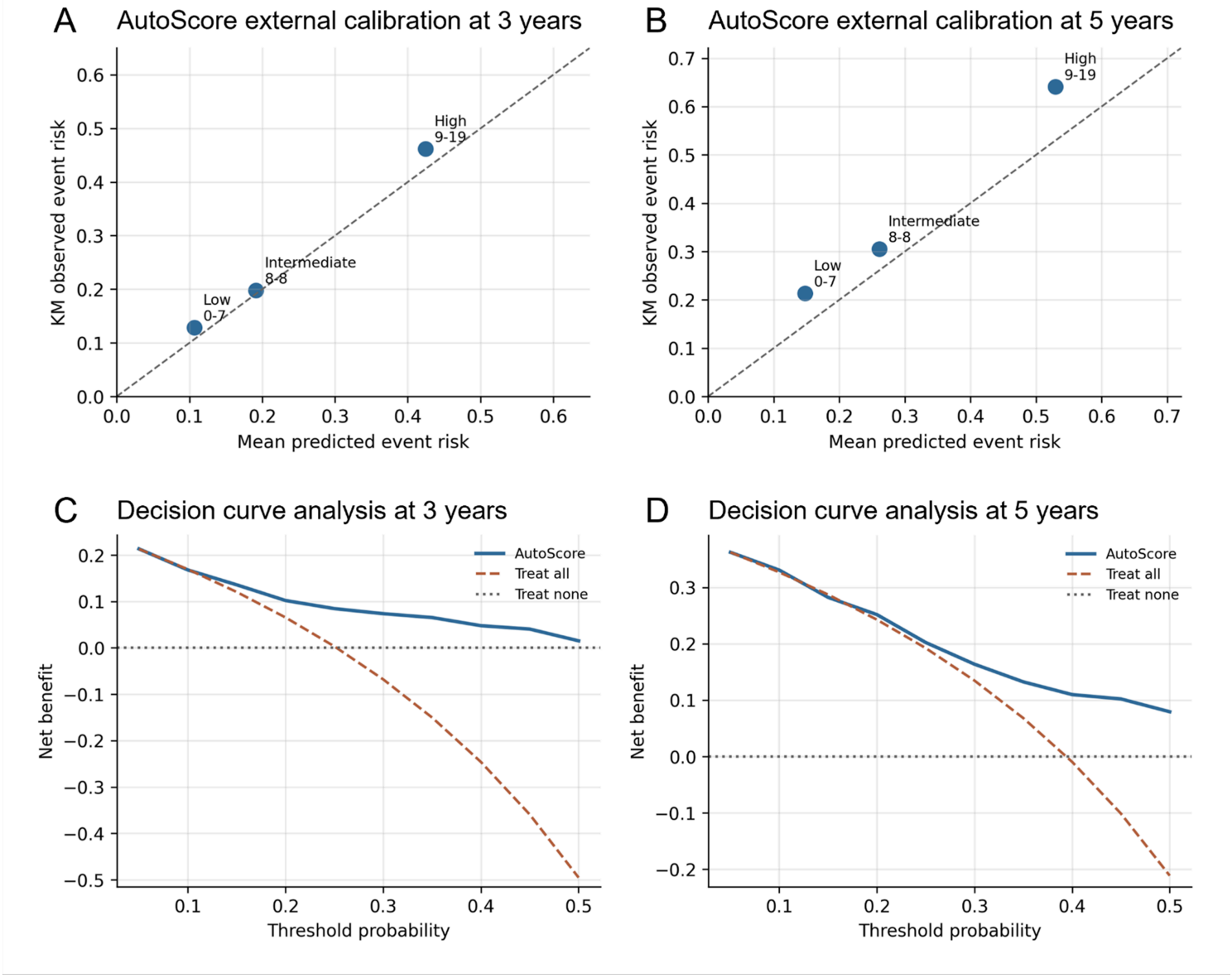


Figure 4

External validation and decision curve analysis of the AutoScore. (A-B) Calibration plots at 3, and 5 years. (C-D) Decision curve analysis at 3, and 5 years.

Supplementary Files

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

- Supplementaryfigure1.tif
- Supplementaryfigure2.tif
- SupplementaryTable1.docx
- SupplementaryTable2.docx

- SupplementaryTable3.docx
- SupplementaryTable4.docx