Journal of Hepatocellular Carcinoma

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Machine-Learning Based Prognostic Model for Predicting Early Recurrence in HCC Patients After Hepatectomies: An Explainable AI Approach

Heng-Yuan Hsu[1], Jiunn-Chang Lin[2–4], Chun-Wei Huang[1,5], Song-Fong Huang[1,6], Chun-Yi Wu[1], Tun-Sung Huang[2–4], Hung-Fei Lai[2–4], Ming-Chin Yu[1,7]

[1]Division of General Surgery, Department of Surgery, New Taipei Municipal Tucheng Hospital, Built and Operated by Chang Gung Medical Foundation, New Taipei, 23652, Taiwan; [2]Department of Surgery, MacKay Memorial Hospital, Taipei, 104217, Taiwan; [3]MacKay Junior College of Medicine, Nursing, and Management, New Taipei, 112021, Taiwan; [4]Department of Medicine, MacKay Medical University, New Taipei, 252005, Taiwan; [5]Graduate Institute of Business Administration, College of Management, Fu Jen Catholic University, New Taipei, 242062, Taiwan; [6]Department of Biological Science and Technology, National Yang Ming Chiao Tung University, Hsinchu, 30068, Taiwan; [7]College of Medicine, Chang Gung University, Taoyuan, 33305, Taiwan

Correspondence: Ming-Chin Yu, Division of General Surgery, Department of Surgery, New Taipei Municipal Tucheng Hospital, built and operated by Chang Gung Medical Foundation, No. 6, Sec. 2, Jincheng Road, Tucheng Dist, New Taipei, 23652, Taiwan, Fax +886-2-8273-1845, Email mingchin2000@gmail.com

Purpose: Early recurrence within 24 months post-resection remains a primary driver of poor prognosis in hepatocellular carcinoma (HCC). In the absence of standardized adjuvant guidelines, robust postoperative risk stratification is critical. We evaluated explainable machine learning (ML) architectures to optimize risk modeling using readily accessible parameters.

Patients and Methods: This retrospective, multicenter study analyzed 1,681 HCC patients undergoing curative-intent hepatectomy at Chang Gung institutions (2007–2020) as the training cohort. External validation was conducted using an independent cohort (n = 251) from Mackay Memorial Hospital. Four algorithms—random survival forest, Cox-nnet, LASSO, and extreme gradient boosting (XGBoost)—were trained using 5-fold cross-validation. Missing data were handled via k-nearest neighbors imputation. Discriminative capacity was assessed using the concordance index (C-index), and feature significance was decoded through SHAP values.

Results: The XGBoost framework yielded optimal discrimination, achieving a high training C-index of 0.98. During independent external validation, the C-index attenuated to a robust 0.72, reflecting expected adjustments for baseline institutional heterogeneities. Multivariable Cox and SHAP analyses consistently identified five pivotal predictors: sex, preoperative treatment, tumor size, satellite lesions, and vascular invasion. The derived nomogram enabled effective patient risk-tiering ($p < 0.0001$), although absolute recurrence probabilities were systematically overestimated in the external validation cohort.

Conclusion: While the XGBoost model exhibits expected calibration shifts across disparate cohorts, it provides robust, cross-center discriminative generalizability for categorical risk stratification. Rather than serving as an absolute probability estimator, this explainable model functions as a reliable clinical tool to selectively identify high-risk candidates for intensive imaging surveillance. Geographically and ethnically diverse prospective validation remains required prior to broader clinical deployment.

Keywords: hepatocellular carcinoma, liver resection, nomogram, SHAP, multicenter validation, machine learning

Introduction

Hepatocellular carcinoma (HCC) remains a global health challenge, ranking among the leading causes of cancer-related mortality worldwide.[1] While curative hepatectomy offers the best chance for long-term survival, postoperative recurrence remains the primary obstacle to achieving satisfactory oncological outcomes, with five-year recurrence rates lingering over 50%.[2] This high incidence of recurrence places a significant burden on both patients and clinicians. Currently, the lack of standardized adjuvant therapies and a consensus on recurrence management underscores the critical need for early recurrence prediction. Early recurrence is an independent adverse prognostic factor for HCC after curative treatment.[3,4]

https://doi.org/10.2147/JHC.S618264
Received: 19 April 2026
Accepted: 18 June 2026
Published: 24 June 2026

Identifying high-risk patients for intensive surveillance and timely intervention remains the most effective strategy to improve prognosis.[2]

Extensive research has identified various clinicopathological predictors of HCC recurrence, a field to which our group has previously contributed.[2,4,5] Established predictors such as tumor size, vascular invasion, and satellite lesions have been associated with recurrence, yet their integration into accurate predictive frameworks remains suboptimal.[2] Conventional statistical methods often struggle to integrate the increasingly complex and multidimensional nature of clinical data. Furthermore, manual data processing is not only labor-intensive but also susceptible to human error. In contrast, machine learning (ML)—a subset of artificial intelligence (AI)—offers a robust framework for identifying intricate patterns within vast datasets.[6] By iteratively optimizing predictive algorithms, ML can enhance the accuracy of risk modeling and adapt to new data inputs.[7] For instance, Guo et al recently developed a metabolomic state-integrated nomogram with multi-dimensional data, including 168 metabolomic profiles alongside clinicopathological factors. Their framework demonstrated outstanding performance in predicting long-term liver-related events and complications, yielding time-dependent area under the curve (AUC) of 0.930 at 3 years, 0.889 at 5 years, and 0.861 at 10 years in the validation cohort, respectively.[8]

Due to the absence of overt clinical symptoms in patients with recurrent HCC, early detection remains heavily predicated upon serial surveillance imaging. In this context, AI-enhanced analysis substantially mitigates interpretation variability while augmenting diagnostic precision.[9] Timely detection of recurrent tumors followed by prompt intervention is well-documented to improve long-term oncological outcomes. To date, several investigators have leveraged ML to refine HCC prognostication and treatment strategies. For instance, Wu et al constructed an ML framework using multi-institutional data from three medical centers (n = 636 total) to predict survival in HCC patients with concurrent type 2 diabetes mellitus (T2DM). Their Random Survival Forest (RSF) model achieved external validation AUCs of 0.862, 0.815, and 0.798 for 1-, 2-, and 3-year survival, respectively.[10] Similarly, Zhu et al developed an ML prognostic model tailored for hepatitis B-associated HCC following radical resection. This model yielded concordance indices (C-indices) of 0.736 and 0.629 for the training and validation cohorts, respectively, demonstrating robust capacity in predicting 1-, 3-, and 5-year survival rates.[11]

Several studies have applied machine learning to HCC prognosis; however, the clinical translation of ML prognostic models in hepatology remains substantially constrained. First, current multi-omics-based frameworks frequently rely on high-dimensional, specialized profiles that incur prohibitive financial and logistical costs, restricting their scalability in routine clinical practice. Second, existing algorithms are often developed using localized cohorts with limited sample sizes, which undermines their statistical power and prevents robust algorithmic saturation. Third, many models lack genuine external validation, relying instead on split-samples from the same homogeneous registry—a strategy that inherently inflates performance metrics and obscures real-world generalizability. Furthermore, despite numerous predictive models, their clinical adoption remains limited due to a lack of interpretability and user-friendly tools for bedside decision-making. To address these limitations, we developed and externally validated an explainable machine learning model integrating clinicopathological data, with emphasis on interpretability and clinical applicability through a nomogram-based risk stratification system.

Materials and Methods

Study Population

Between January 2007 and December 2020, we retrospectively reviewed 1,940 consecutive patients with HCC who underwent curative-intent hepatectomy performed by the same surgical team across Chang Gung Memorial Hospital (CGMH) and New Taipei municipal Tucheng Hospital (built and operated by Chang Gung Medical Foundation). The inclusion criteria focused on patients receiving their primary curative resection. Exclusion criteria were defined as follows: (1) histologically confirmed combined hepatocellular-cholangiocarcinoma; (2) non-primary or palliative intent hepatectomies; (3) age <18 or >85 years; (4) a concurrent or prior history of other malignancies; (5) Barcelona Clinic Liver Cancer (BCLC) stage D;[12] and (6) perioperative mortality, subsequent liver transplantation, or loss to follow-up within 24 months postoperatively (Figure 1).

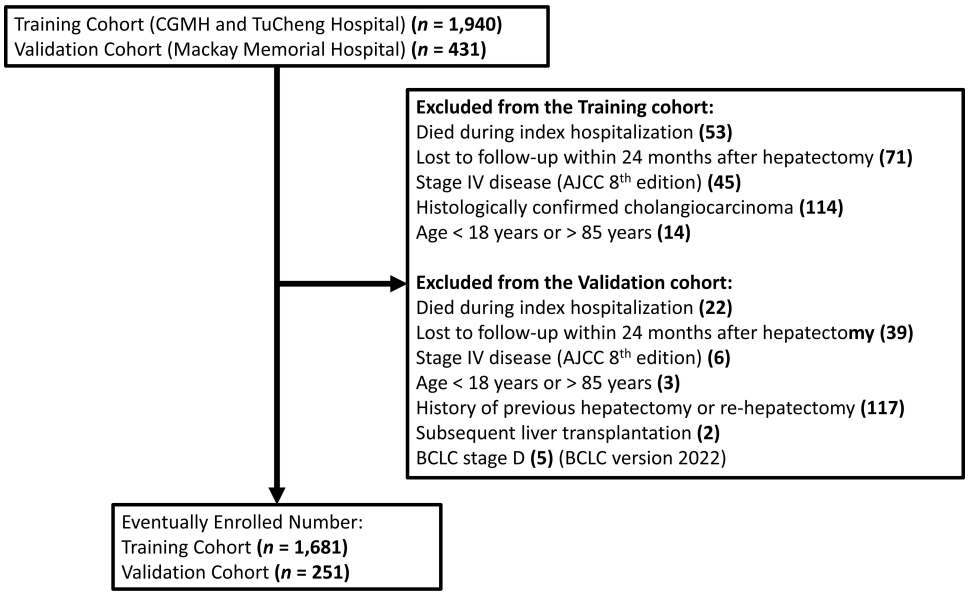

Figure 1 Flowchart of patient selection. After strict application of the inclusion and exclusion criteria, a total of 1,681 patients from Chang Gung Memorial Hospital (CGMH) and Tucheng Municipal Hospital were enrolled into the training cohort to develop the ML-based predictive model. An independent cohort of 251 patients from Mackay Memorial Hospital (MMH) served as the validation cohort for external validation. Exclusion criteria varied slightly across centers due to data availability.

Surgical approaches, including open laparotomy, minimally invasive surgery, or hybrid techniques, were selected based on the IWATE criteria and the lead surgeon's discretion.[13] All resected specimens were evaluated by board-certified pathologists to confirm the diagnosis and staging, the latter of which followed the American Joint Committee on Cancer (AJCC) 8th edition Cancer Staging Manual.[14] For model development and robust performance assessment, the CGMH and Tucheng Hospital cohort served as the internal training dataset. An independent cohort from Mackay Memorial Hospital (MMH) was utilized as the external validation dataset to evaluate the model's predictive accuracy and generalizability. This study was conducted in accordance with the ethical guidelines of the Declaration of Helsinki and was approved by the Institutional Review Board (IRB) of CGMH (IRB No: 202401522B0). The requirement for informed consent was waived due to the retrospective nature of the study, and strict measures were implemented to ensure the de-identification and confidentiality of all patient data.

Definition of Early Recurrence and Postoperative Follow-Up

Postoperative surveillance was standardized for all patients in the study cohort. A baseline triphasic computed tomography (CT) scan of the liver was performed within one month after surgery to confirm curative resection. Subsequently, patients underwent regular follow-up every three months during the first two years, which included physical examinations, liver function tests, serum alpha-fetoprotein (AFP) levels, and liver imaging (CT or magnetic resonance imaging [MRI]). If recurrence was clinically suspected due to rising AFP levels or suspicious nodules, additional diagnostic workups, including gadoxetic acid-enhanced MRI or ultrasound-guided biopsy, were conducted. Early recurrence was defined as the documented return of HCC within 24 months of the primary curative hepatectomy.[15] Diagnosis of recurrence was established based on typical dynamic imaging hallmarks (eg., arterial phase hyperenhancement with portal venous or delayed phase washout), significantly elevated AFP levels, or histological confirmation. Management of recurrent HCC was determined by a multidisciplinary tumor board based on the patient's liver functional reserve and tumor burden.

Machine Learning Based Prognostic Model

Clinical and pathological variables were comprehensively collected for all patients. Missing data were imputed using the k-nearest neighbors (KNN) algorithm to maintain dataset integrity. The parameter K was configured based on

cohort scale, utilizing K = 60 for the large-scale training cohort and K = 16 for the smaller external validation cohort. To ensure model robustness and minimize bias, a 5-fold cross-validation was employed during the training phase. We developed prognostic models using four distinct machine learning algorithms: Random Survival Forests (RSF), Cox Proportional Hazards Neural Network (Cox-nnet), Least Absolute Shrinkage and Selection Operator (LASSO) regression, and eXtreme Gradient Boosting (XGBoost). The predictive performance of these candidate models was evaluated and compared using the concordance index (C-index) to identify the optimal architecture for predicting HCC recurrence.

To enhance the interpretability of the final model, feature importance was quantified using gain and weight values, further supplemented by SHapley Additive exPlanations (SHAP) beeswarm plots to visualize the contribution of individual parameters.[16] Based on the optimized model, a novel nomogram was constructed to estimate the 2-year recurrence-free survival (RFS) for patients following hepatectomy. The study cohort was subsequently stratified into high- and low-risk groups based on the nomogram-derived scores. RFS curves were generated via the Kaplan–Meier method and compared using the Log rank test. The clinical utility and accuracy of the nomogram were rigorously validated through C-index calculation, calibration curves (using 1000 bootstrap resamples), decision curve analysis (DCA), and receiver operating characteristic (ROC) curve analysis. To move beyond single-metric reliance on the C-index, model prediction error was supplementarily evaluated using time-dependent Brier scores at 24 months.

Statistical Analysis

For baseline characteristics, categorical variables were expressed as frequencies and percentages, compared using Pearson's chi-squared test. Continuous variables were presented as medians with interquartile ranges (IQR) and compared using the Mann–Whitney *U*-test or Student's *t*-test, as appropriate. Risk factors for recurrence were initially screened using univariate Cox regression analysis; variables with $p < 0.05$ were subsequently included in a multivariate Cox proportional hazards model. To justify sample size adequacy under the retrospective design, an Events Per Variable (EPV) analysis was executed. Statistical significance was defined as a two-tailed *p*-value < 0.05. All computational analyses were performed using R software (version 4.4.2) and Python (version 3.11.11).

Results

Baseline Clinicopathological Characteristics

Following a rigorous screening process based on the inclusion and exclusion criteria, 1,681 patients from the CGMH and Tucheng Hospital cohorts were enrolled in the training dataset, while 251 patients from Mackay Memorial Hospital (MMH) formed the independent external validation dataset (Figure 1). The demographic and clinicopathological variables are summarized in Table 1. Significant baseline clinical heterogeneity was observed between the two cohorts. Specifically, the training cohort exhibited a lower prevalence of comorbidities (54.2% vs. 61.4%, $p = 0.04$) and fewer preoperative treatments (11.1% vs. 39.8%, $p < 0.001$) compared to the validation cohort. Regarding liver function and operative factors, the training cohort had lower indocyanine green retention at 15 minutes (ICG R15) (7.5% vs. 9.9%, $p < 0.001$) and total bilirubin levels (0.6 vs. 0.9 mg/dL, $p < 0.001$), but underwent longer operative durations (270 vs. 223 min, $p < 0.001$) with more frequent major hepatectomies (28.9% vs. 19.5%, $p = 0.003$) and inflow control (68.5% vs. 48.4%, $p < 0.001$). Pathologically, the training cohort showed higher rates of cirrhosis (47.4% vs. 36.0%, $p < 0.001$) and high-grade Edmondson-Steiner tumors (38.9% vs. 23.1%, $p < 0.001$), but a lower incidence of satellite lesions (5.5% vs. 17.5%, $p < 0.001$).

The median follow-up duration was 64.7 months (IQR, 38.5–119.9) for the training cohort and 58.2 months (IQR, 32.1–91.1) for the validation cohort. The overall survival (OS) rates in the training cohort at 1, 3, and 5 years were 92%, 78%, and 67%, respectively, compared to 91%, 75%, and 64% in the validation cohort ($p = 0.998$; Supplemental Figure 1). Notably, the 24-month early recurrence rate was significantly higher in the training cohort (38.3%) than in the validation cohort (17.9%, $p < 0.001$; Supplemental Figure 1). Despite these heterogeneities in baseline characteristics, the validation cohort provides a robust platform to test the adaptability of the proposed prognostic model.

Table 1 Baseline Demographic and Clinicopathological Characteristics of the Study Cohorts

Variables[a]	Training Cohort (CGMH and Tucheng Hospital) (n = 1,681)	Validation Cohort (Mackay Memorial Hospital) (n = 251)	*p*-value
Age (years)	61.6 (52.5–69.1)	63.0 (55.0–69.5)	0.30
Sex (Male)	1,301 (77.4)	181 (72.1)	0.08
Comorbidity (Yes)	911 (54.2)	154 (61.4)	0.04
DM (Yes)	401 (23.9)	85 (33.9)	< 0.001
HTN (Yes)	618 (36.8)	72 (28.7)	0.02
ESRD (Yes)	35 (2.1)	1 (0.4)	0.08
HBV (Positive)	1,007 (61.4)	138 (61.1)	0.99
HCV (Positive)	492 (30.8)	73 (32.3)	0.71
Pre-OP treatment[b] (Yes)	186 (11.1)	100 (39.8)	< 0.001
ICG R15 (%)	7.5 (4.2–12.4)	9.9 (6.0–14.9)	< 0.001
Platelet (1000/uL)	174.0 (133.0–221.0)	174.5 (130.3–227.8)	0.67
INR	1.1 (1.0–1.1)	1.1 (1.0–1.1)	0.23
Albumin (g/dL)	4.2 (3.9–4.5)	4.2 (3.8–4.6)	0.83
ALT (IU/L)	35.0 (24.0–58.0)	34.5 (25.0–56.0)	0.83
Bil-T (mg/dL)	0.6 (0.5–0.9)	0.9 (0.6–1.1)	< 0.001
AFP (ng/mL)	15.2 (4.6–197.3)	15.3 (5.8–218.2)	0.28
OP time (min)	270.0 (210.0–345.0)	223.0 (175.0–273.0)	< 0.001
Blood loss (mL)	250.0 (100.0–550.0)	500.0 (137.5–1000.0)	< 0.001
Major hepatectomy[c] (Yes)	485 (28.9)	49 (19.5)	0.003
Inflow control (Yes)	1,112 (68.5)	120 (48.4)	< 0.001
Tumor size (cm)	3.5 (2.3–5.8)	3.5 (2.5–6.0)	0.32
Cirrhosis[d] (Yes)	795 (47.4)	90 (36.0)	< 0.001
Encapsulation (Yes)	1,405 (83.9)	209 (83.9)	1.0
Satellite lesion (Yes)	92 (5.5)	44 (17.5)	< 0.001
Margin < 0.5 (cm)	756 (45.1)	134 (53.6)	0.02
Grade III,IV / II,I	650 (38.9) / 1020 (61.1)	58 (23.1) / 193 (76.9)	< 0.001
Vascular invasion			0.84
No	1,088 (64.9)	166 (66.7)	
Microvascular	460 (27.4)	64 (25.7)	
Gross	128 (7.7)	19 (7.6)	

(*Continued*)

Table 1 (Continued).

Variables[a]	Training Cohort (CGMH and Tucheng Hospital) (n = 1,681)	Validation Cohort (Mackay Memorial Hospital) (n = 251)	*p*-value
AJCC 8th staging			1.0
IA	256 (15.2)	38 (15.1)	
IB	647 (38.5)	98 (39.0)	
II	456 (27.2)	67 (26.7)	
III	321 (19.1)	48 (19.1)	

Notes: [a] Categorical variables are expressed as numbers (%), and continuous variables are expressed as medians (interquartile ranges). Only patients with available data were analyzed. [b] Preoperative treatments include ablation, targeted therapy, immunotherapy, transarterial chemoembolization, chemotherapy, radiotherapy, proton therapy, thalidomide, and others.[c] Major hepatectomy includes tri-segmentectomy, right/left lobectomy, and extended right/left lobectomy.[d] Liver cirrhosis is pathologically defined by an Ishak fibrosis score of F5–F6.

Abbreviations: AFP, alpha-fetoprotein; AJCC, American Joint Committee on Cancer; ALT, alanine aminotransferase; Bil-T, total bilirubin; CGMH, Chang Gung Memorial Hospital; DM, diabetes mellitus; ESRD, end-stage renal disease; HBV, hepatitis B virus; HCV, hepatitis C virus; HTN, hypertension; ICG R15, indocyanine green retention at 15 minutes; INR, international normalized ratio.

Development and Performance Comparison of Machine Learning Models

To develop a robust prognostic framework, we integrated comprehensive clinicopathological variables into four machine learning algorithms: RSF, Cox-nnet, LASSO regression, and XGBoost. Dataset integrity was maintained by employing the KNN algorithm for missing data imputation, while 5-fold cross-validation was implemented to mitigate overfitting. To maximize accuracy without discarding high-order nonlinear interactions, the optimal XGBoost model was configured under an Accelerated Failure Time (AFT) architecture using all available features. Mathematical overfitting was actively suppressed through a combination of strict tree-depth constraints, L1/L2 regularization penalties (α=0.01, λ=2), and an automated early-stopping mechanism, with the comprehensive algorithmic hyperparameter configuration registry detailed in Supplemental Figure 2. Model performance was rigorously evaluated using the C-index across both training and validation datasets. As summarized in Table 2, the XGBoost model achieved the highest predictive performance, with C-indices of 0.98 (95% CI: 0.97–0.98), 0.72 (95% CI: 0.65–0.80), and 0.95 (95% CI: 0.94–0.96) for the training, validation, and combined datasets, respectively. While LASSO regression demonstrated the most consistent performance across cohorts (validation C-index: 0.77; 95% CI: 0.70–0.83), XGBoost was selected as the optimal model due to its superior global discriminative capacity.

Feature importance was further elucidated through SHAP beeswarm plots, alongside weight and gain analyses of the XGBoost architecture (Figure 2).[16,17] This interpretable AI approach identified five predominant predictors for early HCC recurrence: sex, preoperative treatment, tumor size, satellite lesions, and vascular invasion. Notably, these findings were highly congruent with the results from traditional univariate and multivariate Cox regression analyses (Supplemental Table 1), where the same variables emerged as independent risk factors. This cross-methodological consistency underscores the biological relevance and predictive reliability of the identified features in the context of post-hepatectomy prognosis. When integrating these five terminal variables into the clinical nomogram, alpha-fetoprotein

Table 2 Predictive Performance of the Machine Learning Models Across the Training and External Validation Cohorts

	Training Dataset	Validation Dataset	All
Random Survival Forests (RSF)	0.81 (0.79–0.82)	**0.77 (0.70–0.83)**	0.79 (0.77–0.82)
Cox Proportional Hazards Neural Network (Cox-nnet)	0.97 (0.97–0.98)	0.65 (0.58–0.74)	0.93 (0.92–0.94)
Least Absolute Shrinkage and Selection Operator (Lasso regression)	0.75 (0.73–0.77)	**0.77 (0.70–0.83)**	0.72 (0.70–0.74)
Extreme Gradient Boosting (XGBoost)	**0.98 (0.97–0.98)**	0.72 (0.65–0.80)	**0.95 (0.94–0.96)**

Notes: Bold text indicates the highest predictive performance (C-index) within each dataset category. The "All" column represents a pooled population (n = 1,932) combining both the training and validation cohorts to compute global 95% confidence intervals via bootstrapping.

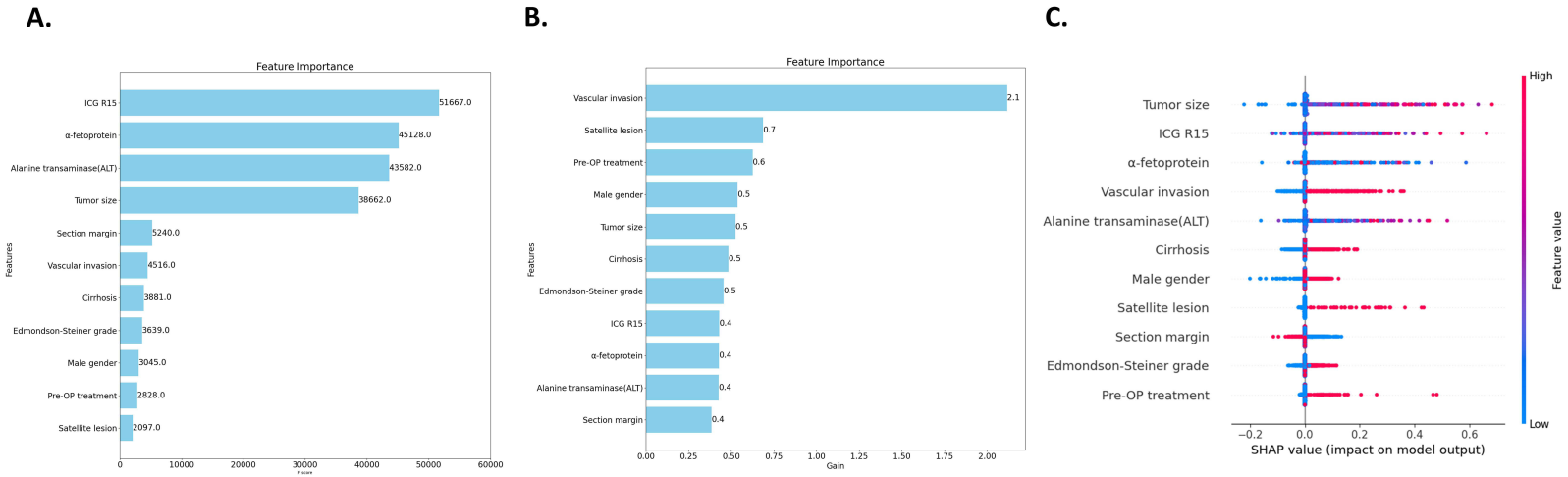

Figure 2 Feature importance analysis of the XGBoost model. (**A**) Weight values (*F*-score) of the features. Weight represents the total number of times a feature is used to split the data across all trees. A high *F*-score indicates that the model frequently uses a specific feature for granular modifications during the prediction process. ICG R15, AFP, ALT, and tumor size are the four parameters with the highest weight values. (**B**) Gain values of the features. Gain represents the average improvement in model accuracy brought by splits using a specific feature. Vascular invasion, satellite lesions, preoperative treatment, sex, and tumor size are the five parameters with the highest gain values. (**C**) SHAP beeswarm plot. This plot illustrates the directional impact of each feature on the model's predictions. The majority of features tend to increase the predicted recurrence risk, whereas wider resection margins demonstrate a protective (negative) effect on the model output.

(AFP)—despite demonstrating high model weight—was intentionally excluded due to its severe, non-normal right-skewed distribution, which rendered it unsuited for standard linear bedside point-scaling.

Development and Validation of the ML-Based Nomogram

In the training cohort (n = 1,681) with 643 early recurrence events, evaluated against the five terminal nomogram predictors, the resulting EPV exceeded 120. This far surpasses the standard methodological threshold (EPV > 10–20), ensuring highly stable and non-overparameterized prognostic estimates. Utilizing the key features identified by the XGBoost model, a clinical nomogram was constructed to predict the 2-year early recurrence risk for HCC patients following curative resection (Figure 3). The nomogram demonstrated robust predictive performance, yielding C-indices of 0.709 and 0.733 for the training and validation datasets, respectively. To evaluate its clinical utility, we calculated total nomogram scores for all patients and stratified them into high-risk and low-risk groups based on the median score. Kaplan–Meier analysis revealed that the low-risk group experienced significantly longer RFS compared to the high-risk group in both the training and validation cohorts (both $p < 0.0001$; Figure 3), confirming the discriminative power of the XGBoost-derived features.

The accuracy of the prognostic model was further rigorously assessed using calibration curves, which illustrated a high degree of concordance between the predicted probabilities and actual 2-year recurrence outcomes in the training cohort but not the validation cohort (Figure 4). The model achieved a robust Brier score of 0.1183 in the training cohort (null model: 0.1451). In the external validation cohort, the Brier score was 0.0364 (null model: 0.1190), demonstrating excellent overall mean squared error profiles despite the calibration overestimation. Decision curve analysis (DCA) substantiated the clinical net benefit of the nomogram across a wide range of threshold probabilities, consistently outperforming the "treat all" and "treat none" strategies (Figure 4).[18] Furthermore, ROC curve analysis was employed to compare the nomogram's performance against individual clinical variables. In the training cohort, the nomogram achieved an AUC of 0.727, surpassing individual predictors such as sex (0.541), preoperative treatment (0.533), tumor size (0.682), satellite lesions (0.544), and vascular invasion (0.660) (Figure 4 and Supplemental Figure 3). These findings collectively validate the nomogram as a reliable and superior tool for early recurrence risk assessment in clinical practice.

Discussion

Clinical Implications of Early Recurrence in HCC

Early recurrence of HCC, typically occurring within 24 months post-resection, remains the most formidable obstacle to long-term survival.[15] In our training cohort, 38.3% of patients experienced early recurrence, a figure consistent with major international reports.[2] Early recurrence is often driven by intrahepatic metastasis and reflects aggressive tumor biology, making it a critical target for predictive modeling.[19] Given the current global absence of standardized adjuvant therapies, early detection remains heavily predicated upon intensive surveillance imaging. In this context, AI-enhanced radiological workflows have been shown to substantially mitigate interpretation variability while augmenting early diagnostic precision, a translational paradigm successfully replicated across various cross-disciplinary medical frameworks.[9] Timely detection of sub-clinical recurrent lesions followed by prompt intervention significantly improves long-term oncological outcomes. Consequently, our explainable XGBoost-AFT model acts as an essential tactical gateway: by pre-emptively stratifying patients into distinct risk tiers, it empowers clinicians to efficiently allocate advanced, high-precision imaging protocols and aggressive surveillance resources to individuals at the highest risk for early recurrence.

Advantages of Machine Learning Over Traditional Statistical Methods

A key finding of this study is the superior predictive performance of ML compared to traditional Cox proportional hazards regression. While traditional Cox regression aims to identify specific statistical relationships (*p*-values), ML algorithms are designed to maximize predictive accuracy by identifying complex patterns within the data. Although our univariate Cox analysis identified several significant factors—including male gender, preoperative treatment, ICG R15, and tumor characteristics (Supplemental Table 1)—traditional statistics often require the dichotomization of continuous

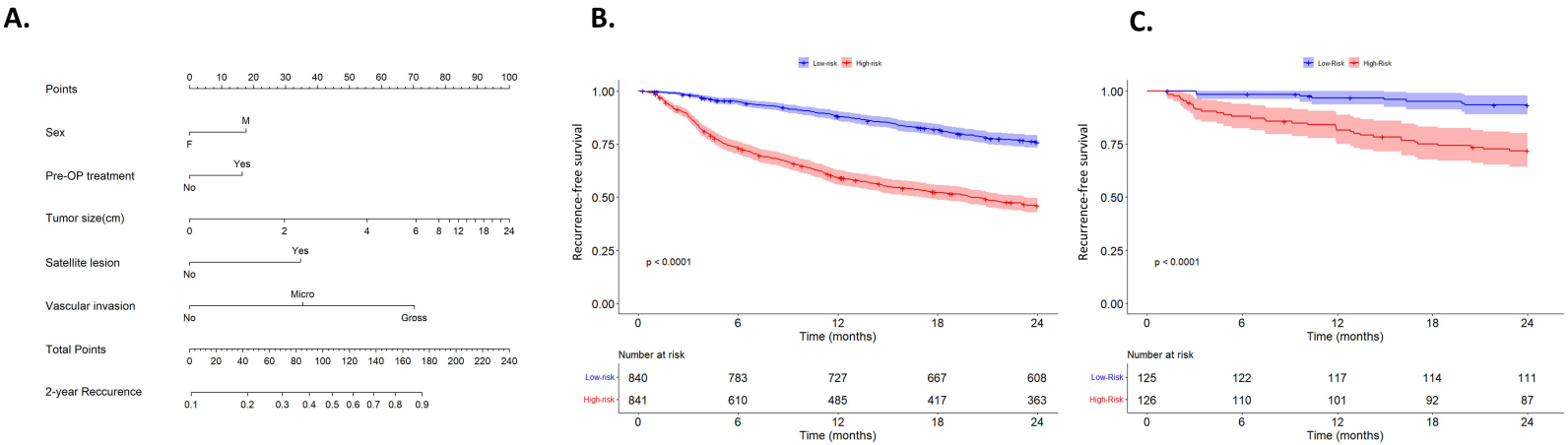

Figure 3 Development and performance of the clinical nomogram. (**A**) A clinical nomogram for predicting early recurrence in HCC patients post-hepatectomy, developed using the top five predictive features identified by the XGBoost model. The total score, calculated by summing the points for each variable, projects the absolute probability of 2-year early recurrence on the bottom scale. (**B**) Kaplan-Meier recurrence-free survival (RFS) curves for the training cohort and (**C**) the validation cohort, stratified by nomogram-derived risk scores, demonstrating robust discriminatory performance (both $p < 0.0001$).

https://doi.org/10.2147/JHC.S618264

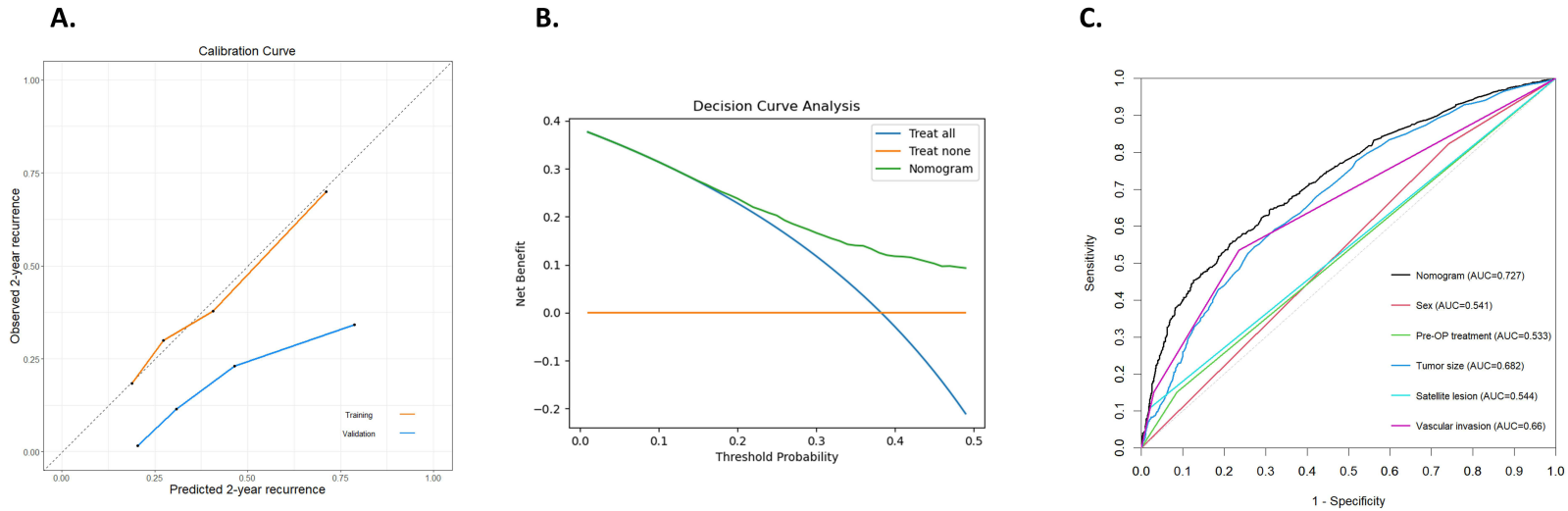

Figure 4 Clinical validation of the nomogram. (**A**) Calibration curves. These curves depict the relationship between the nomogram-predicted probability of early recurrence and the actual observed outcomes. The nomogram performs well in the training cohort; however, in the external validation cohort, it systematically overestimates the recurrence rate, particularly among patients at high predicted risk, due to baseline cohort heterogeneity. (**B**) Decision curve analysis (DCA). The DCA demonstrates the clinical net benefit of the nomogram across a wide range of threshold probabilities in the training cohort. (**C**) Receiver operating characteristic (ROC) curves. A comparison of the Area Under the Curve (AUC) values between the integrated nomogram and individual clinicopathological features for predicting early recurrence in the training cohort.

https://doi.org/10.2147/JHC.S618264

variables (eg., ALT >2x ULN or AFP >200 ng/mL) to calculate hazard ratios. However, such data transformation can lead to a significant loss of granular information. In contrast, ML algorithms like XGBoost process continuous data in their original distribution, capturing non-linear relationships that traditional models might overlook. Our results demonstrate that directly inputting raw continuous data into the XGBoost model yielded higher gain values, significantly reducing the potential for human error in data preprocessing while maintaining high computational efficiency.

Deciphering the "Black Box": Model Interpretability via SHAP

The clinical adoption of ML has historically been hindered by its "black-box" nature, where the decision-making process remains opaque to clinicians. To address this, we employed an interpretable AI framework using XGBoost, which functions as an ensemble of gradient-boosted decision trees. Based on Gradient Boosting Decision Trees (GBDT), XGBoost iteratively adds new decision trees, and each new tree corrects the errors in the previous layer of the model, thereby gradually improving the predictive ability of the model.[17] By analyzing feature importance through weight and gain values in XGBoost, we quantified the purification effect and frequency of each variable within the model. The weight value (*F*-score) indicates the frequency with which a feature is utilized to split data across the ensemble of trees, reflecting its role as a pervasive modifier. Conversely, the gain value represents a feature's relative contribution to model accuracy, indicating its influence in optimizing the predictive splits. Furthermore, the SHAP beeswarm plot provided a nuanced visualization of how specific features influence the model's output directionally.[16] For instance, the SHAP analysis effectively illustrated the negative impact of narrow resection margins on recurrence-free survival (Figure 2). The high degree of congruence between the top five ML-identified predictors (sex, preoperative treatment, tumor size, satellite lesions, and vascular invasion) and those from multivariate Cox analysis reinforces the biological validity of our model and enhances its transparency for clinical use (Supplemental Table 1).

Model Validation and Universal Applicability

To translate our findings into a practical clinical tool, we developed a nomogram based on the five key features identified by XGBoost. This nomogram demonstrated robust risk stratification and predictive accuracy, with C-indices of 0.709 and 0.733 in the training and validation datasets, respectively (Figure 3). The calibration curve demonstrated strong predictive power in the training cohort. DCA and ROC curve comparisons further confirmed that our integrated nomogram offers superior clinical net benefit compared to any single clinicopathological factor, providing a reliable foundation for personalized clinical decision-making (Figure 4).[18]

Our nomogram provides distinct incremental value over established staging systems like BCLC or TNM. While traditional classifications focus on pre-therapeutic, anatomical tumor staging (eg., size, satellite lesions, vascular invasion) to guide initial treatment routing, our model is tailored specifically as a post-operative tool for patients undergoing their first curative hepatectomy. Crucially, our framework integrates key clinical features—biological sex and preoperative management—that are omitted in standard staging systems. This non-anatomical integration enables finer risk differentiation to actively guide intensive post-operative surveillance pathways.

Model Generalizability and Addressing Algorithmic Overfitting

We acknowledge that the pronounced discrepancy between our training and validation performance (C-index: 0.98 vs. 0.72) raises legitimate concerns regarding overfitting, an inherent challenge in gradient-boosted machine-learning architectures. To minimize this risk during development, we implemented a rigorous 5-fold cross-validation framework and leveraged a substantially large training cohort (n = 1,681) to ensure algorithmic saturation. Although we actively restricted mathematical overfitting through strict algorithmic hyperparameters (eg., depth constraints, L1/L2 regularization penalties and an automated early-stopping mechanism; Supplemental Figure 2), a significant discrepancy in recurrence-free survival was still observed between the two cohorts ($p < 0.001$, Supplemental Figure 1). This reflects profound baseline heterogeneities in patient selection, preoperative treatment pathways, and postoperative surveillance protocols between institutions (Supplemental Table 2). Consistent with recent high-impact literature where predictive metrics characteristically attenuate during external validation or extended follow-up, this cross-cohort divergence explains the calibration shift in the validation set.[8,10,11] Nevertheless, the fact that our nomogram sustained

a consistent, robust discriminative capability (C-index: 0.733) across such disparate populations underscores its real-world generalizability and mitigates the clinical impact of overfitting under rigorous stress-testing conditions.

Study Limitations

Despite the robust performance of our XGBoost-AFT prognostic model, several limitations warrant consideration. First, its retrospective nature inherently introduces selection bias and historical data gaps. Although we utilized KNN imputation to maintain dataset integrity, imputed values may not fully replicate individual biological variations. Second, marked clinical heterogeneity existed between cohorts, notably in preoperative treatment rates (11.1% vs. 39.8%, $p < 0.001$; Supplemental Table 2). While this divergence served as a rigorous stress test confirming the model's discriminative generalizability across distinct medical systems, it inevitably shifted probability calibration in the external validation set. Finally, the non-normal distribution of certain clinicopathological parameters poses an inherent challenge for algorithmic scaling. Future multi-center prospective trials with standardized protocols are essential to refine risk calibration and validate clinical utility across broader populations.

Conclusion

In conclusion, our study demonstrates that an explainable ML prognostic model, specifically driven by the XGBoost-AFT algorithm, delivers robust and clinically plausible feature integration for predicting early recurrence in HCC patients post-hepatectomy. However, the performance attenuation observed during independent cross-center testing underscores the inherent challenge of algorithmic generalizability under diverse institutional settings. While the model maintains consistent discriminative power, its precise risk-probability calibration is constrained by baseline clinical heterogeneity. Therefore, rather than serving as an absolute probability estimator, the derived nomogram should be responsibly implemented as a reliable tool for categorical postoperative risk stratification, guiding the intensification of surveillance pathways for patients identified as high-risk. Ultimately, multicenter prospective validation in geographically and ethnically diverse populations remains an essential prerequisite before broader clinical implementation can be recommended.

Abbreviations

AFT, accelerated failure time; AFP, alpha-fetoprotein; AI, artificial intelligence; AJCC, American Joint Committee on Cancer; AUC, area under the curve; BCLC, Barcelona Clinic Liver Cancer; CGMH, Chang Gung Memorial Hospital; C-index, concordance index; Cox-nnet, Cox proportional hazards neural network; CT, computed tomography; DCA, decision curve analysis; EPV, events per variable; GBDT, gradient boosting decision trees; HCC, hepatocellular carcinoma; ICG R15, indocyanine green retention at 15 minutes; IQR, interquartile range; IRB, Institutional Review Board; KNN, k-nearest neighbors; LASSO, least absolute shrinkage and selection operator; ML, machine learning; MMH, Mackay Memorial Hospital; MRI, magnetic resonance imaging; OS, overall survival; RFS, recurrence-free survival; ROC, receiver operating characteristic; RSF, random survival forest; SHAP, SHapley Additive exPlanations; T2DM, type 2 diabetes mellitus; XGBoost, eXtreme Gradient Boosting.

Data Sharing Statement

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Institutional Review Board Statement

This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (IRB No. 202401522B0) of Chang Gung Memorial Hospital (CGMH), Linkou Branch.

Informed Consent Statement

The requirement for informed consent was waived by the IRB due to the retrospective nature of the study. Strict measures were implemented to ensure the de-identification and confidentiality of all patient data.

Acknowledgments

We are grateful to all our colleagues at the Cancer Center, the Department of Pathology, and the Graduate Institute of Clinical Medical Sciences at Chang Gung University for their technical assistance. We also thank Yi-Ping Liu for assisting with data retrieval and processing.

Funding

This study was supported by the Chang Gung Medical Foundation, Taiwan (Grant No. CMRPVVN0142), and the National Science and Technology Council, Taiwan (Grant No. 114-2314-B-182A-139- for MCY).

Disclosure

Ming-Chin Yu reports Support for the manuscript from The project was from the Next-Generation Digital Healthcare Platform Plan of the Ministry of Health and welfare, during the conduct of the study; Grants or contracts from as item 1, outside the submitted work. The authors declare that they have no other competing interests in this work.

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74(3):229–263. doi:10.3322/caac.21834
2. Abdelhamed W, El-Kassas M. Hepatocellular carcinoma recurrence: predictors and management. *Liver Res*. 2023;7(4):321–332. doi:10.1016/j.livres.2023.11.004
3. Takeishi K, Maeda T, Tsujita E, et al. Predictors of intrahepatic multiple recurrences after curative hepatectomy for hepatocellular carcinoma. *Anticancer Res*. 2015;35(5):3061–3066.
4. Hsu HY, Tang JH, Huang SF, et al. Recurrence pattern is an independent surgical prognostic factor for long-term oncological outcomes in patients with hepatocellular carcinoma. *Biomedicines*. 2024;12(3):655. doi:10.3390/biomedicines12030655
5. Huang CW, Lin SE, Huang SF, et al. The Vessels That Encapsulate Tumor Clusters (VETC) Pattern Is a Poor Prognosis Factor in Patients with Hepatocellular Carcinoma: an Analysis of Microvessel Density. *Cancers*. 2022;14(21):5428. doi:10.3390/cancers14215428
6. Zeng J, Zeng J, Lin K, et al. Development of a machine learning model to predict early recurrence for hepatocellular carcinoma after curative resection. *Hepatobiliary Surg Nutr*. 2022;11(2):176–187. doi:10.21037/hbsn-20-466
7. Calderaro J, Seraphin TP, Luedde T, Simon TG. Artificial intelligence for the prevention and clinical management of hepatocellular carcinoma. *J Hepatol*. 2022;76(6):1348–1361. doi:10.1016/j.jhep.2022.01.014
8. Guo C, Liu Z, Fan H, et al. Machine-learning-based plasma metabolomic profiles for predicting long-term complications of cirrhosis. *Hepatology*. 2025;81(1):168–180. doi:10.1097/HEP.0000000000000879
9. Xu J, Gao C, Zhang J, et al. Advancements in imaging technologies and ai integration for neurodegenerative disease management: a narrative review. *Mol Imaging*. 2025;24:15353508251393056. doi:10.1177/15353508251393056
10. Wu L, Chen L, Zhang L, et al. A Machine Learning Model for Predicting Prognosis in HCC Patients With Diabetes After TACE. *J Hepatocell Carcinoma*. 2025;12:77–91. doi:10.2147/JHC.S496481
11. Zhu D, Tulahong A, Abuduhelili A, et al. Machine learning prognostic model for post-radical resection hepatocellular carcinoma in hepatitis b patients. *J Hepatocell Carcinoma*. 2025;12:353–365. doi:10.2147/JHC.S495059
12. Reig M, Sanduzzi-Zamparelli M, Forner A, et al. BCLC strategy for prognosis prediction and treatment recommendations: the 2026 update. *J Hepatol*. 2026;84(3):631–654. doi:10.1016/j.jhep.2025.10.020
13. Wakabayashi G. What has changed after the Morioka consensus conference 2014 on laparoscopic liver resection? *Hepatobiliary Surg Nutr*. 2016;5(4):281–289. doi:10.21037/hbsn.2016.03.03
14. Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin*. 2017;67(2):93–99. doi:10.3322/caac.21388
15. Imamura H, Matsuyama Y, Tanaka E, et al. Risk factors contributing to early and late phase intrahepatic recurrence of hepatocellular carcinoma after hepatectomy. *J Hepatol*. 2003;38(2):200–207. doi:10.1016/s0168-8278(02)00360-4
16. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017.
17. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016.;
18. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574. doi:10.1177/0272989X06295361
19. Poon RT, Fan ST, Ng IO, Lo CM, Liu CL, Wong J. Different risk factors and prognosis for early and late intrahepatic recurrence after resection of hepatocellular carcinoma. *Cancer*. 2000;89(3):500–507. doi:10.1002/1097-0142(20000801)89:3<500::AID-CNCR4>3.0.CO;2-O