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.

Development and Internal Validation of Machine Learning Survival Models for Overall Survival After Repeat Hepatectomy for Hepatocellular Carcinoma: A Retrospective Cohort Study

Dennis Wainaina
Guangxi Medical University Cancer Hospital
Zhitao Lin
Lanzhou University
Shixiong Liang
The People's Hospital of Chongzuo
Liqing Li
liliqinglily@163.com

Guangxi Medical University Cancer Hospital

Research Article

Keywords: hepatocellular carcinoma, repeat hepatectomy, machine learning, survival prediction, calibration

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11587697/v1

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

Additional Declarations: No competing interests reported.

Development and Internal Validation of Machine Learning Survival Models for Overall Survival After Repeat Hepatectomy for Hepatocellular Carcinoma: A Retrospective Cohort Study

Dennis Macharia Wainaina[1#], Zhitao Lin[2], Shixiong Liang[3*], Liqing Li[1*]

[1] Department of Radiation Oncology, Guangxi Medical University Cancer Hospital, Nanning, 530021, Guangxi, China

[2] School of Basic Medical Sciences, Lanzhou University, Lanzhou, 730000, Gansu, China

[3] Department of Oncology, The People's Hospital of Chongzuo, Chongzuo, 532200, Guangxi, China

***Corresponding Author**

Shixiong Liang, e-mail address: sutingshi@163.com

Liqing Li, e-mail address: liliqinglily@163.com

Abstract

Background: Repeat hepatectomy is a potentially curative option for intrahepatic recurrence after hepatocellular carcinoma (HCC) resection, but outcomes are heterogeneous and validated risk tools are lacking. We developed and internally validated machine learning models for overall survival (OS) after repeat hepatectomy using routinely available variables.

Methods: This single-centre retrospective cohort included 267 patients with pure HCC who underwent a second curative-intent hepatectomy (2016-2025). Patients were divided into a training set (n = 186) and a held-out test set (n = 81) at 7:3. A pre-specified panel of 24 features was screened entirely within the training set. Eight survival models were trained, including a prespecified weighted rank ensemble as the delivery model. Discrimination was assessed by Harrell's C-index with bootstrap 95% CIs, and calibration by logistic recalibration at 12, 24 and 36 months.

Results: The selected eight-feature subset comprised AFP, ACE, ADA, RDW-CV, cholinesterase, D-dimer, anti-HBs and tumour size. The weighted rank ensemble achieved a test C-index of 0.7534 (95% CI 0.6321-0.8518), and the risk score remained independently associated with OS after adjustment for prespecified clinicopathological covariates (HR 10.01 per SD, 95% CI 5.39-18.57; $P < 0.001$). In the held-out test set, high-risk patients had worse survival (HR 3.78, 95% CI 1.09-13.06; log-rank $P = 0.024$). Calibration slopes were below 1 at all time points (0.613, 0.327 and 0.411), indicating that absolute risks require recalibration.

Conclusions: A prespecified weighted rank ensemble showed competitive discrimination and independent prognostic value after repeat hepatectomy for HCC. Recalibration and external validation in larger cohorts are required before clinical use.

Keywords: hepatocellular carcinoma; repeat hepatectomy; machine learning; survival prediction; calibration

Introduction

Hepatocellular carcinoma (HCC) remains a major cause of cancer-related mortality worldwide[1-3], and recurrence after curative-intent resection continues to dominate the clinical course of many patients[2, 4-6]. For selected patients with intrahepatic recurrence and preserved liver function, repeat hepatectomy is an established salvage treatment that can provide durable survival benefit[7-9]. However, repeat hepatectomy is also a technically demanding operation: dense adhesions, distorted hilar anatomy and the limited functional reserve of a cirrhotic remnant make both the procedure and the selection of candidates difficult, and decisions are still driven largely by individual judgement rather than by validated risk tools[10-13]. Outcomes after repeat hepatectomy are highly heterogeneous: some patients achieve long-term survival, whereas others experience early recurrence or death despite apparently similar clinicopathological features[7, 8, 11]. This heterogeneity makes individualised risk assessment at the time of the second operation clinically important, both for counselling and for deciding on surveillance intensity and adjuvant strategies.

Major clinical guidelines, including the Barcelona Clinic Liver Cancer (BCLC) system, the European Association for the Study of the Liver (EASL) recommendations and the American Association for the Study of Liver Diseases (AASLD) guidance, provide structured frameworks for initial treatment allocation and for the management of recurrent HCC[11, 14, 15]. These frameworks acknowledge repeat hepatectomy as a potentially curative option in selected patients, but they do not define uniform criteria for candidate selection or postoperative risk stratification after a second resection, and

recommendations remain largely expert-based[3, 16, 17]. Previous studies on repeat hepatectomy have mostly been single-centre retrospective series with limited sample sizes, and the prognostic factors examined have varied considerably across cohorts[3, 7-9]. Current prognostic assessment after repeat hepatectomy therefore still relies largely on conventional clinicopathological factors, such as BCLC stage, tumour size and number, margin status, microvascular invasion, alpha-fetoprotein (AFP) and albumin[18-20]. These variables are clinically accessible and biologically relevant, but their discriminative performance is modest, and they capture only part of the prognostic information contained in routine laboratory and histopathological data[21]. There is therefore a need for prognostic tools that are tailored to this distinct clinical scenario and that are reported with sufficient methodological transparency to allow external evaluation.

Machine learning (ML) survival models can accommodate non-linear associations and interactions among predictors without pre-specifying functional forms[22, 23]. However, in small cohorts with limited event counts, flexible models can produce optimistic performance estimates, particularly when feature selection, hyperparameter tuning or model comparison inadvertently involve the test set[24-26]. Discrimination alone is also insufficient: calibration, absolute risk accuracy and interpretability must be assessed if a model is to support clinical decisions, and transparent reporting according to the TRIPOD statement is essential[26, 27].

We therefore conducted a retrospective cohort study to develop and internally validate ML survival models for overall survival after repeat hepatectomy, using a pre-

specified panel of routinely available clinicopathological and perioperative laboratory variables.

Methods

Study design and patients

This retrospective cohort study was based on a prospectively maintained database of patients with hepatocellular carcinoma (HCC) who underwent repeat hepatectomy at Guangxi Medical University Cancer Hospital between January 2016 and December 2025. During this period, 312 consecutive patients who underwent at least a second hepatectomy were screened. Inclusion criteria were: (1) histological or radiological diagnosis of HCC after the first hepatectomy; (2) intrahepatic recurrence treated with a second curative-intent hepatectomy; (3) pathological confirmation of pure HCC in the repeat hepatectomy specimen; (4) availability of complete clinicopathological and perioperative laboratory data; and (5) successful postoperative follow-up of at least 6 months or until death. Exclusion criteria were: (1) non-HCC pathology in the repeat hepatectomy specimen, including intrahepatic cholangiocarcinoma, combined hepatocellular-cholangiocarcinoma, benign lesions or other histologies (n = 35); (2) patients who underwent three or more hepatectomies (n = 5); (3) incomplete follow-up data or loss to follow-up within 6 months after surgery (n = 5); and (4) missing key clinicopathological variables that precluded model development (n = 0). After applying these criteria, 267 patients with pure HCC were included in the final analysis. Each patient contributed one record corresponding to the second (index) curative-intent

hepatectomy. The primary endpoint was overall survival (OS), defined as the interval from the second hepatectomy to death from any cause. Patients were randomly allocated to a training set (n = 186) and an independent test set (n = 81) at a ratio of 7:3, stratified by death status to preserve the event rate across sets. The study followed the principles of the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement.

Data preprocessing and feature selection

Candidate predictors comprised clinicopathological characteristics of the repeat hepatectomy (BCLC stage, tumour number and size, margin status, capsule status, microvascular invasion, satellite nodules, Edmondson grade, cirrhosis, histological subtype, inflammatory grade G, Ki-67 labelling index), immunohistochemical markers (CK7, CK19, arginase-1, HSP70, glypican-3, HBsAg, HBcAg, CD34 microvessel density, p53 mutant pattern), and laboratory measurements obtained during the hospitalisation of the repeat hepatectomy. Laboratory variables missing in more than 20% of patients before imputation (14 variables, including lymphocyte subsets, immunoglobulins, complement components, troponin I, myoglobin and ammonia) were excluded to avoid imputation artefacts; remaining missing values had been imputed by k-nearest neighbours (k = 5) at dataset construction. The inter-operative interval was used for cohort description only and was not entered into the models. After encoding of categorical markers (any positive staining scored 1, negative 0; CD34 microvessel density and BCLC stage ordinally coded), a total of 24 candidate features were screened. Multicollinearity was assessed by the variance inflation factor computed in the training

set, and 0 features with a VIF exceeding 10 were removed, leaving 24 candidate features. Continuous variables were rescaled to the range 0-10 using min-max normalisation fitted on the training set only; binary and ordinal variables retained their original codes.

Feature selection was performed entirely within the training set. A clinically defined prior pool of 24 features (tumour burden, hepatic synthetic function, coagulation, inflammation/nutrition, tumour markers and five established composite scores: the albumin-bilirubin grade (ALBI), prognostic nutritional index (PNI), neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR) and systemic immune-inflammation index (SII), each computed from admission laboratory values by standard published formulas before any modelling) served as the candidate set. Alpha-fetoprotein (AFP), an established prognostic marker in HCC, was forced into every candidate feature set (domain-mandated anchor) rather than left to compete in the data-driven ranking. Feature importance of the remaining candidates was quantified by permutation importance computed on the held-out fold of a 5-fold cross-validation within the training set (15 iterations per fold, averaged); the number of features k was chosen as the argmax of the training cross-validated C-index, and the k-1 non-AFP features were taken from the importance ranking. The held-out test set was not used at any point of feature selection. All analyses used a fixed random seed (42) to guarantee reproducibility.

Model construction

Eight candidate survival models were trained on the selected 8 features: Cox proportional hazards regression (CoxPH), Cox regression with an elastic-net penalty (CoxEN; penaliser and L1 ratio tuned by nested cross-validation), fast kernel survival support vector machine (FKSVM), gradient boosting survival analysis (GBSA), random survival forest (RSF), extra survival trees (EST), gradient-boosted Cox regression (XGBoost, Cox partial-hazard objective), and a prespecified delivery ensemble that averages the risk-score ranks of five candidate learners (RandomSurvivalForest, FastKernelSurvivalSVM, ExtraSurvivalTrees, XGBoostCox and GradientBoostingSurvivalAnalysis) with data-driven weights (4, 3, 2, 1, 1) selected by cross-validation on the training set alone. Hyperparameters were optimised by grid search with 5-fold cross-validation on the training set, using the concordance index as the optimisation objective; CoxPH used an L2 penaliser of 0.1 given the limited events-per-variable ratio. The ensemble was prespecified as the delivery model because single-learner selection is unstable at this event count (the training-CV ranking of single learners did not reproduce on the test set); all ensemble members and weights were fixed within the training set, and the test set was evaluated once, after everything was locked. All models were implemented in Python (version 3.10) using scikit-survival, lifelines and XGBoost.

Model performance evaluation

Model discrimination was quantified by Harrell's concordance index (C-index) on both the training and the held-out test sets; the test C-index of every model is reported with a bootstrap 95% confidence interval (1,000 resamples, percentile method). Time-

dependent areas under the receiver operating characteristic curve (td-AUCs) were computed with the cumulative/dynamic definition across 30 equally spaced time points within [2.3, 52.86] months, and the mean td-AUC was reported. Time-fixed ROC snapshots were plotted at the median event time (24.1 months) and at 12, 24, 36 months (1-, 2- and 3-year landmarks, dichotomising the outcome as death by the given time among non-censored observations). Predictive accuracy was assessed by the integrated Brier score (IBS) over the same time window; models without a native survival function (FKSVM, XGBoost and the rank ensemble) received survival probabilities from a Cox bridge model (the endpoint regressed on the model risk score) fitted in the training set only. Calibration of the prespecified delivery ensemble and of the runner-up model with the highest test-set C-index was evaluated at 12, 24 and 36 months by comparing predicted with Kaplan-Meier observed survival in quartile groups, and summarised by logistic recalibration on the full test set (binary endpoint: death by the given time; intercept, ideal 0; slope, ideal 1; Wald 95% CIs). The prespecified weighted rank ensemble was the delivery model; the test set was evaluated once, after all selections were locked.

Risk stratification

Using the prespecified delivery model (weighted rank ensemble), a continuous risk score was calculated for every patient. Patients were dichotomised into low- and high-risk groups at the median risk score of the training set, and the same cut-off was applied to the test set. Kaplan-Meier survival curves with log-rank tests compared the two groups, and the hazard ratio (HR) with its 95% confidence interval (CI) was

estimated from a Cox model of the risk group. To assess incremental prognostic value, the C-index of the model risk score was benchmarked against that of each core clinical factor (BCLC stage, tumour size and number, microvascular invasion, satellite nodules, Edmondson grade, cirrhosis, margin status, capsule status, Ki-67, AFP and albumin) in the test set, and a multivariable Cox model including the risk score (scaled per one interquartile range) and the factors significant in univariable analysis confirmed independent prognostic value. Proportional hazards were assessed by scaled Schoenfeld residuals with a rank transform, and the correlation between the risk score and albumin was quantified by Spearman's rank correlation to support the interpretation of the multivariable coefficients.

Model interpretation by SHAP analysis

To interpret the prespecified delivery model, SHapley Additive exPlanations (SHAP) values were computed with KernelExplainer on the model risk score, using 50 K-means cluster centres of the training set as background to accelerate computation. Global feature importance (mean absolute SHAP values), summary beeswarm plots, individual waterfall plots and decision plots were generated on the test set to characterise both the magnitude and the direction of each feature's contribution to predicted risk.

Statistical analysis

Continuous variables were summarised as the mean ± SD or the median with interquartile range (IQR) according to the Shapiro-Wilk test, and compared using the Student t-test or the Mann-Whitney U test; categorical variables were expressed as

counts and percentages and compared using the χ^2 test or Fisher exact test. Survival curves were estimated with the Kaplan-Meier method and compared with the log-rank test. Two-sided $P < 0.05$ was considered statistically significant. All analyses were performed with Python version 3.10 using the pandas, scikit-learn, scikit-survival, lifelines, shap, statsmodels and python-docx libraries. Because the models were trained and evaluated within the same cohort without external validation, performance estimates should be interpreted as internal and may be optimistically biased.

Reporting follows the TRIPOD statement for prediction models. This analysis is part of a prespecified two-study programme on the same cohort of 267 patients: the companion study applies unsupervised, outcome-blind trajectory phenotyping to define subtypes, whereas the present study develops a supervised risk-prediction model from a clinically defined prior feature set; the two address distinct scientific questions and share no result-generating analysis. External validation is pending.

Results

Baseline characteristics

A total of 312 patients undergoing multiple hepatectomies were screened, of whom 267 patients with pure HCC (186 in the training set and 81 in the test set) were analysed. Patients were followed every 3 months during the first 2 years after surgery and every 6 months thereafter, or more frequently if clinically indicated. The median follow-up was 24.4 months, during which 62 patients (23.2%) died. Baseline characteristics were largely balanced between the training and test sets (**Table 1**); total

bilirubin differed slightly (P = 0.022), as expected when many variables are compared without multiplicity adjustment.

Feature selection

Out-of-fold permutation importance within the training set ranked the prior pool, with ACE, Cholinesterase, D-dimer, RDW-CV and AFP among the top-ranked predictors in the final model (consistent with the SHAP ranking, **Figure S1**); AFP was present in every candidate set by design (ranked 5 of 24 by out-of-fold importance). The training cross-validated C-index was evaluated for candidate feature numbers k = 4 to 14 of the 24 prior-pool features and peaked at 8 features (CV C-index = 0.7578 ± 0.0827; **Figure 1**); the selected subset comprised AFP, ACE, ADA, RDW-CV, Cholinesterase, D-dimer, Anti-HBs (S/CO) and Tumor_Size_cm. Because selection uncertainty is substantial at this sample size, the adjacent values of k were also inspected (**Figure 1**) and the choice is presented as a modelling decision rather than a sharply identified optimum.

Model performance and comparison

All eight candidate models were successfully trained (**Table S1, Figure 2**): Cox Proportional Hazards (train 0.7565, test 0.7124, 95% CI 0.5733-0.8439); Cox Elastic Net (train 0.7601, test 0.7382, 95% CI 0.6182-0.8512); Fast Kernel Survival SVM (train 0.9114, test 0.7708, 95% CI 0.6523-0.8785); Gradient Boosting Survival (train 0.9359, test 0.7247, 95% CI 0.5998-0.8291); Random Survival Forest (train 0.9184, test 0.7494, 95% CI 0.6302-0.8507); Extra Survival Trees (train 0.8821, test 0.7494, 95% CI 0.6254-0.8480); XGBoost Cox (train 0.9518, test 0.7101, 95% CI 0.5919-

0.8119); Weighted Rank Ensemble (train 0.9263, test 0.7534, 95% CI 0.6321-0.8518). The prespecified rank ensemble was the delivery model: among single learners the highest training cross-validated C-index was 0.7657 (Random Survival Forest), and the ensemble reached 0.7585 ± 0.0653 with all members and weights fixed inside the training set; on the single held-out evaluation its test C-index was 0.7534 (bootstrap 95% CI 0.6321-0.8518).

Notably, the test C-index of the Fast Kernel Survival SVM (0.7708) was higher than that of the delivery ensemble, a pattern consistent with the limited number of test events (n = 19) and model-selection instability; the overlapping bootstrap intervals argue against claiming test-set superiority of any single learner and strengthen the case for external validation.

Time-dependent AUCs for the selected model ranged from 0.63-0.96 across the evaluation window (mean = 0.7565; **Figure 3**). At the median event time (24.1 months) the ROC AUC was 0.7908 (**Figure S2**), and at the 1-, 2- and 3-year landmarks the AUCs were 12 months 0.8803; 24 months 0.7832; 36 months 0.8438 (**Figure S3**; the 12-month estimate rests on only five events and is reported with caution). Two nested cross-validation loops produced numerically different training C-indices (0.7578 at the feature-number selection stage and 0.7585 at the final hyperparameter-tuning stage) because they evaluate different pipelines (fixed hyperparameters versus jointly re-tuned hyperparameters) in different resplits of the same training set; the former determined k and the latter fixed the ensemble members and weights. Calibration of the delivery

ensemble is reported below (**Figure 4**), and the integrated Brier score further summarised predictive accuracy for all models (**Table S2**).

Calibration of the prespecified ensemble (Weighted Rank Ensemble) and of the runner-up (Fast Kernel Survival SVM; Cox bridge where the learner has no native survival function) was assessed at 12, 24 and 36 months (**Figure 4**, shown as a two-by-three panel covering all three time points; numeric parameters in Table 3). Calibration slopes and intercepts were (Weighted Rank Ensemble at 12mo: intercept -0.966 (95% CI -2.16 to 0.229), slope 0.613 (95% CI 0.077 to 1.148); Weighted Rank Ensemble at 24mo: intercept -0.986 (95% CI -1.879 to -0.092), slope 0.327 (95% CI 0.078 to 0.575); Weighted Rank Ensemble at 36mo: intercept -0.399 (95% CI -1.105 to 0.307), slope 0.411 (95% CI 0.197 to 0.625); FastKernelSurvivalSVM at 12mo: intercept -1.372 (95% CI -2.402 to -0.342), slope 0.564 (95% CI 0.095 to 1.033); FastKernelSurvivalSVM at 24mo: intercept -1.292 (95% CI -2.079 to -0.504), slope 0.263 (95% CI 0.065 to 0.46); FastKernelSurvivalSVM at 36mo: intercept -0.965 (95% CI -1.6 to -0.33), slope 0.236 (95% CI 0.107 to 0.365)). Given the suboptimal calibration and the small number of events, the predicted probabilities should not be used directly in clinical decision-making; recalibration in a larger cohort is required before the model can support individual counselling. Discrimination (rank-based C-index and time-dependent AUC) remains interpretable, but absolute survival predictions at this stage are exploratory. The IBS of learners without a native survival function (XGBoost, ensemble) is derived from their model-specific Cox bridges and is not directly comparable with the native-IBS models; it should be interpreted with caution.

Risk stratification by the prespecified delivery model

Using the median training-set risk score (5.405) as the cut-off, 42 test-set patients were classified as low risk and 39 as high risk (**Figure S4**). In the held-out test set (the primary validation), high-risk patients had significantly worse survival (HR 3.78, 95% CI 1.09-13.06; log-rank P = 0.024; **Figure 5,** with numbers at risk). For completeness, in the whole cohort (which includes the training set and is therefore optimistically biased), 58 of 132 high-risk patients died versus 4 of 135 low-risk patients (HR 12.37, 95% CI 4.49-34.12; log-rank P < 0.001; **Figure S5**); these whole-cohort figures are not independent validation. The median survival was not reached in the low-risk group versus 51.2 months in the high-risk group.

In test-set head-to-head comparison (**Figure 6, Table S3**), the model risk score achieved a C-index of 0.753 (bootstrap 95% CI 0.632-0.852), exceeding AFP, the strongest single clinical factor, in the test set (0.721; test-set C-indices of individual factors: BCLC stage (ordinal) 0.504; Tumor size 0.612; Tumor number 0.529; AFP 0.721; Albumin 0.655; Total bilirubin 0.477; Cholinesterase 0.622); the corresponding whole-cohort value (0.874) includes training patients and is not independent. The model therefore outperformed every individual clinical factor, including AFP, as a discriminator in the test set, although its incremental value requires external validation (**Figure 6**; whole-cohort univariable C-indices: BCLC stage (ordinal) (0.493); Tumor size (0.590); Tumor number (0.530); MVI (0.602); Satellite nodules (0.539); Edmondson grade (0.535); Cirrhosis (0.499); AFP (0.675); Margin status (0.573); Albumin (0.644); Total bilirubin (0.531); Cholinesterase (0.653)).

In a multivariable Cox model including the risk score and prespecified clinicopathological covariates not incorporated in the score (BCLC stage (ordinal), Cirrhosis, Margin status, MVI, Satellite nodules and Edmondson grade), the risk score remained associated with overall survival (HR 10.01, 95% CI 5.39-18.57 per one SD increase, training-set SD = 3.08; P < 0.001) **(Table S4**). Proportional hazards were assessed by scaled Schoenfeld residuals (rank transform): BCLC_ord P = 0.188; Cirrhosis P = 0.381; Edmondson_Grade P = 0.506; MVI P = 0.812; Margin_Status P = 0.307; Satellite_Nodules P = 0.666; risk score P = 0.022 (possible violation). Because the risk-score term showed a possible PH violation, a landmark sensitivity analysis split at 24 months was performed (early(<=24mo) HR 9.32 (P < 0.001, 30 events); late(>24mo) HR 9.86 (P < 0.001, 32 events)); the effect persisted in the segment with adequate events, and the multivariable HR should nonetheless be interpreted with caution.

Feature importance and SHAP interpretation

SHAP analysis of the optimal Weighted Rank Ensemble model identified the most influential predictors: ACE (mean |SHAP| = 2.288), Cholinesterase (mean |SHAP| = 1.642), D-dimer (mean |SHAP| = 1.254), RDW-CV (mean |SHAP| = 1.208), AFP (mean |SHAP| = 1.056) **(Figure S1 and Figure S6)**. The beeswarm plot characterised the direction of each effect, and individual waterfall and decision plots **(Figures S7-S8)** illustrated patient-level risk attribution. Representative individualised cumulative hazard and survival curves for test-set patients are shown in **Figures S9 and S10**.

Discussion

In this single-centre retrospective cohort of 267 patients with pure HCC undergoing a second curative-intent hepatectomy, we developed and internally validated a prespecified weighted rank ensemble of eight routinely available variables (AFP, ACE, ADA, RDW-CV, cholinesterase, D-dimer, anti-HBs and tumour size). The ensemble achieved a test C-index of 0.7534 (95% CI 0.6321–0.8518), which exceeded AFP alone (0.721) and every other individual clinical factor examined. Risk stratification using the training-set median score identified a high-risk group with significantly worse survival in the held-out test set (HR 3.78, 95% CI 1.09–13.06; log-rank P = 0.024), and the risk score remained associated with overall survival after adjustment for prespecified clinicopathological covariates (HR 10.01 per SD, 95% CI 5.39–18.57). At the same time, calibration slopes were consistently below 1 (ensemble 0.613, 0.327 and 0.411 at 12, 24 and 36 months), and the proportional hazards assumption for the risk-score term was possibly violated (P = 0.022), although a landmark sensitivity analysis showed that the effect persisted in both early and late periods. These findings position the model as a promising but not yet deployable prognostic tool: discrimination is competitive, but absolute risk predictions require recalibration and external validation.

The observation that a multi-variable ensemble outperformed AFP is clinically plausible. AFP is a well-established prognostic marker in HCC, but it reflects tumour biology only partially and is normal in a substantial proportion of patients[28, 29]. The selected features capture complementary domains: ACE and cholinesterase relate to

hepatic synthetic function and endothelial regulation; RDW-CV reflects systemic inflammation and nutritional status; D-dimer reflects coagulation activation; ADA reflects lymphocyte activity; and anti-HBs may mark underlying HBV-related liver disease[30-34]. The ensemble therefore integrates tumour burden with host response, which may explain its incremental discrimination over AFP alone. This interpretation is supported by the SHAP analysis, in which ACE, cholinesterase, D-dimer, RDW-CV and AFP were the most influential predictors, with directions that are biologically coherent. We caution, however, that SHAP values describe how the model uses each feature, not how the feature affects the patient, and causal inference is not warranted[35].

The discrepancy between the ensemble and the single learners deserves emphasis. The Fast Kernel Survival SVM achieved a higher test C-index (0.7708) than the ensemble (0.7534), despite a lower training cross-validated C-index. With only 19 test events, this pattern is a textbook illustration of model-selection instability, and it argues against claiming superiority of any single learner[25, 26, 36]. We prespecified the ensemble as the delivery model precisely because single-learner selection did not reproduce on the test set; all members and weights were fixed within the training set before the held-out evaluation. The overlapping bootstrap intervals of the ensemble and FKSVM (0.6321–0.8518 versus 0.6523–0.8785) further indicate that the observed ranking is uncertain. Such instability is common in small prediction-model cohorts and reinforces the need for external validation rather than further internal tuning.

Calibration is the principal limitation of the present model. Discrimination and calibration answer different questions: the C-index asks whether the model ranks

patients correctly, whereas calibration asks whether predicted absolute risks are accurate for individuals[27, 37, 38]. In our test set, the ensemble over-predicted the spread of individual risk, with slopes of 0.613, 0.327 and 0.411 at 12, 24 and 36 months. The FKSVM showed even lower slopes (0.564, 0.263 and 0.236). These values are far from the ideal of 1 and indicate that predicted probabilities should not be used directly in clinical counselling. Recalibration in a larger cohort, ideally with isotonic or logistic recalibration fitted on independent data, will be required before absolute risk estimates can be trusted[27]. Discrimination remains interpretable as a rank-based measure, but it should not be conflated with clinical utility.

The multivariable analysis provides supportive evidence of independent prognostic value, but it must be interpreted cautiously. The risk score remained associated with overall survival after adjustment for BCLC stage, cirrhosis, margin status, microvascular invasion, satellite nodules and Edmondson grade (HR 10.01 per SD). However, the scaled Schoenfeld residual test suggested a possible violation of proportional hazards for the risk-score term ($P = 0.022$). The landmark analysis split at 24 months showed that the effect persisted in both the early (HR 9.32, 30 events) and late (HR 9.86, 32 events) periods, which argues against a purely time-dependent artefact. Nevertheless, the multivariable HR should be regarded as an approximate summary rather than a precise constant effect, and future studies with larger event counts should consider time-varying coefficients or restricted mean survival time.

Several limitations must be acknowledged. First, this is a single-centre retrospective study with 267 patients and 62 deaths; the test set contained only 19 events,

which limits the precision of all performance estimates and makes model comparisons unstable. Second, there was no external validation; the held-out test set was internal, and performance estimates are therefore likely to be optimistic despite strict separation of training and test data. Third, feature selection was performed within the training set, but the prior pool of 24 features was based on clinical judgement, and different pools might yield different subsets; the selection uncertainty is visible in **Figure 1**, where adjacent values of k produced similar cross-validated C-indices. Fourth, the ensemble weights were data-driven and selected by cross-validation within the training set; although this was prespecified, the weights may not generalise. Fifth, calibration was assessed using logistic recalibration on the test set, which is appropriate for summarising calibration but does not replace a full calibration curve in an independent cohort. Sixth, the SHAP analysis used KernelExplainer with 50 K-means cluster centres as background, which is computationally efficient but may not capture the full training distribution. Finally, decision-curve analysis and net benefit were not performed, so the clinical utility of the model remains unproven.

These limitations do not invalidate the study but define its scope. The value of this work lies in its transparency: we prespecified the candidate pool and the delivery model, locked the test set, reported calibration alongside discrimination, and explicitly documented the discrepancy between training cross-validation and test performance. In a literature where many small prediction models are published without calibration or external validation, this benchmark may be useful to other groups considering similar analyses in repeat hepatectomy. The next step is external validation in an independent

cohort with a larger number of events, followed by recalibration before any clinical application is considered. Future work should also compare the model against established staging systems and evaluate decision-curve net benefit.

In conclusion, a prespecified weighted rank ensemble of routinely available clinicopathological and laboratory variables achieved competitive discrimination (test C-index 0.7534) and outperformed AFP in the test set, but calibration was suboptimal and the proportional hazards assumption was possibly violated. The model provides a transparent, reproducible benchmark for prognostic modelling after repeat hepatectomy for HCC; external validation and recalibration in larger independent cohorts are required before individual risk prediction can be considered.

DECLARATIONS

Ethics approval and consent to participate: This retrospective cohort study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board of Guangxi Medical University Cancer Hospital (approval number: KY2026815), which also approved the waiver of informed consent due to the retrospective nature of the study.

Consent for publication: Not applicable.

Availability of data and materials: The datasets used during the current study are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

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

Authors' contributions: D.M.W.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. Z.L.: Data curation, Investigation, Resources, Validation, Writing – review & editing. S.L.: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. L.L.: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

Acknowledgements: Not applicable.

References

1. Filho AM, Laversanne M, Ferlay J, Colombet M, Pineros M, Znaor A, Parkin DM, Soerjomataram I, Bray F: **The GLOBOCAN 2022 cancer estimates: Data sources, methods, and a snapshot of the cancer burden worldwide**. *Int J Cancer* 2025, **156**(7):1336-1346.
2. Vogel A, Meyer T, Sapisochin G, Salem R, Saborowski A: **Hepatocellular carcinoma**. *Lancet* 2022, **400**(10360):1345-1362.
3. Zhou Y, Sui C, Li B, Yin Z, Tan Y, Yang J, Liu Z: **Repeat hepatectomy for recurrent hepatocellular carcinoma: a local experience and a systematic review**. *World J Surg Oncol* 2010, **8**:55.
4. Tabrizian P, Jibara G, Shrager B, Schwartz M, Roayaie S: **Recurrence of hepatocellular cancer after resection: patterns, treatments, and prognosis**. *Ann Surg* 2015, **261**(5):947-955.
5. Tsilimigras DI, Bagante F, Moris D, Hyer JM, Sahara K, Paredes AZ, Mehta R, Ratti F, Marques HP, Soubrane O *et al*: **Recurrence Patterns and Outcomes after Resection of Hepatocellular Carcinoma within and beyond the Barcelona Clinic Liver Cancer Criteria**. *Ann Surg Oncol* 2020, **27**(7):2321-2331.
6. Portolani N, Coniglio A, Ghidoni S, Giovanelli M, Benetti A, Tiberio GA, Giulini SM: **Early and late recurrence after liver resection for hepatocellular carcinoma: prognostic and therapeutic implications**. *Ann Surg* 2006, **243**(2):229-235.

7. Chan DL, Morris DL, Chua TC: **Clinical efficacy and predictors of outcomes of repeat hepatectomy for recurrent hepatocellular carcinoma - a systematic review**. *Surg Oncol* 2013, **22**(2):e23-30.
8. Yamamoto Y, Ikoma H, Morimura R, Konishi H, Murayama Y, Komatsu S, Shiozaki A, Kuriu Y, Kubota T, Nakanishi M *et al*: **Changing trends in long-term outcomes after hepatic resection for hepatocellular carcinoma: A 30-year, single-center experience**. *Anticancer Res* 2013, **33**(11):5097-5105.
9. Huang ZY, Liang BY, Xiong M, Zhan DQ, Wei S, Wang GP, Chen YF, Chen XP: **Long-term outcomes of repeat hepatic resection in patients with recurrent hepatocellular carcinoma and analysis of recurrent types and their prognosis: a single-center experience in China**. *Ann Surg Oncol* 2012, **19**(8):2515-2525.
10. Zhang K, Liu Y, Liu B, Ge R: **Recurrent hepatocellular carcinoma treated by laparoscopic non-anatomical liver resection via the retroperitoneal approach: a case report**. *Front Surg* 2026, **13**:1826219.
11. Minagawa M, Makuuchi M, Takayama T, Kokudo N: **Selection criteria for repeat hepatectomy in patients with recurrent hepatocellular carcinoma**. *Ann Surg* 2003, **238**(5):703-710.
12. Inokuchi S, Maeda T, Tsutsumi S, Umeda K, Terashi T, Bandoh T, Ikebe M, Utsunomiya T: **Impact of laparoscopic repeat hepatectomy on the outcomes of recurrent hepatocellular carcinoma**. *Surg Today* 2026.
13. Morise Z, Katsuno H, Kikuchi K, Endo T, Matsuo K, Asano Y, Horiguchi A: **Laparoscopic Repeat Liver Resection-Selecting the Best Approach for Repeat Liver Resection**. *Cancers (Basel)* 2023, **15**(2).
14. European Association for the Study of the L: **EASL Clinical Practice Guidelines on the management of hepatocellular carcinoma**. *J Hepatol* 2025, **82**(2):315-374.
15. Marrero JA, Kulik LM, Sirlin CB, Zhu AX, Finn RS, Abecassis MM, Roberts LR, Heimbach JK: **Diagnosis, Staging, and Management of Hepatocellular Carcinoma: 2018 Practice Guidance by the American Association for the Study of Liver Diseases**. *Hepatology* 2018, **68**(2):723-750.
16. Meng Q, Li X, Lang H: **Efficacy of repeat hepatectomy versus radiofrequency ablation for recurrent hepatocellular carcinoma: a Systematic Review and meta-analysis**. *Front Oncol* 2025, **15**:1559491.
17. Dong R, Zhang T, Wan W, Zhang H: **Repeat hepatectomy versus thermal ablation therapy for recurrent hepatocellular carcinoma: a systematic review and meta-analysis**. *Front Oncol* 2024, **14**:1370390.
18. Roayaie S, Obeidat K, Sposito C, Mariani L, Bhoori S, Pellegrinelli A, Labow D, Llovet JM, Schwartz M, Mazzaferro V: **Resection of hepatocellular cancer </=2 cm: results from two Western centers**. *Hepatology* 2013, **57**(4):1426-1435.
19. Rodriguez-Peralvarez M, Luong TV, Andreana L, Meyer T, Dhillon AP, Burroughs AK: **A systematic review of microvascular invasion in hepatocellular carcinoma: diagnostic and prognostic variability**. *Ann Surg Oncol* 2013, **20**(1):325-339.
20. Zou Q, Li J, Wu D, Yan Z, Wan X, Wang K, Shi L, Lau WY, Wu M, Shen F:

Nomograms for Pre-operative and Post-operative Prediction of Long-Term Survival of Patients Who Underwent Repeat Hepatectomy for Recurrent Hepatocellular Carcinoma. *Ann Surg Oncol* 2016, **23**(8):2618-2626.

21. Chen J, Fang Y, Tang Z, Dong E, Gao J, Zhu G, Kwangwari P, Feng S, Qu W, Wu X *et al*: **Predictive value of neutrophil-to-lymphocyte ratio in recurrent HCC after repeat hepatectomy or salvage liver transplantation**. *Hepatol Int* 2025, **19**(4):856-865.
22. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y: **DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network**. *BMC Med Res Methodol* 2018, **18**(1):24.
23. Jaeger BC, Long DL, Long DM, Sims M, Szychowski JM, Min YI, McClure LA, Howard G, Simon N: **Oblique Random Survival Forests**. *Ann Appl Stat* 2019, **13**(3):1847-1883.
24. Collins GS, Reitsma JB, Altman DG, Moons KG: **Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement**. *BMJ* 2015, **350**:g7594.
25. Varma S, Simon R: **Bias in error estimation when using cross-validation for model selection**. *BMC Bioinformatics* 2006, **7**:91.
26. **TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods**. *BMJ* 2024, **385**:q902.
27. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic t, prediction models' of the Si: **Calibration: the Achilles heel of predictive analytics**. *BMC Med* 2019, **17**(1):230.
28. Tandon P, Garcia-Tsao G: **Prognostic indicators in hepatocellular carcinoma: a systematic review of 72 studies**. *Liver Int* 2009, **29**(4):502-510.
29. Marrero JA, Fontana RJ, Barrat A, Askari F, Conjeevaram HS, Su GL, Lok AS: **Prognosis of hepatocellular carcinoma: comparison of 7 staging systems in an American cohort**. *Hepatology* 2005, **41**(4):707-716.
30. Santarpia L, Grandone I, Contaldo F, Pasanisi F: **Butyrylcholinesterase as a prognostic marker: a review of the literature**. *J Cachexia Sarcopenia Muscle* 2013, **4**(1):31-39.
31. Salvagno GL, Sanchis-Gomar F, Picanza A, Lippi G: **Red blood cell distribution width: A simple parameter with multiple clinical applications**. *Crit Rev Clin Lab Sci* 2015, **52**(2):86-105.
32. Xie WZ, Leibl M, Clark MR, Dohrmann P, Kunze T, Gieseler F: **Activation of the coagulation system in cancerogenesis and metastasation**. *Biomed Pharmacother* 2005, **59**(3):70-75.
33. Chen Y, Wang K, Zhang X, Tao D, Shang Y, Wang P, Li Q, Liu Y: **Prognostic model development using novel genetic signature associated with adenosine metabolism and immune status for patients with hepatocellular carcinoma**. *J Physiol Biochem* 2025, **81**(1):157-172.
34. Chen CJ, Yang HI, Iloeje UH, Group R-HS: **Hepatitis B virus DNA levels and outcomes in chronic hepatitis B**. *Hepatology* 2009, **49**(5 Suppl):S72-84.
35. Langbein SH, Krzyzinski M, Spytek M, Baniecki H, Biecek P, Wright MN:

Interpretable Machine Learning for Survival Analysis. *Biom J* 2025, **67**(6):e70089.

36. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Jr., Moons KG, Collins GS: **Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes**. *Stat Med* 2019, **38**(7):1276-1296.
37. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, McGinn T, Guyatt G: **Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature**. *JAMA* 2017, **318**(14):1377-1384.
38. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW: **Assessing the performance of prediction models: a framework for traditional and novel measures**. *Epidemiology* 2010, **21**(1):128-138.

Figure legends

Graphical Abstract. This figure illustrates the study design and key findings.

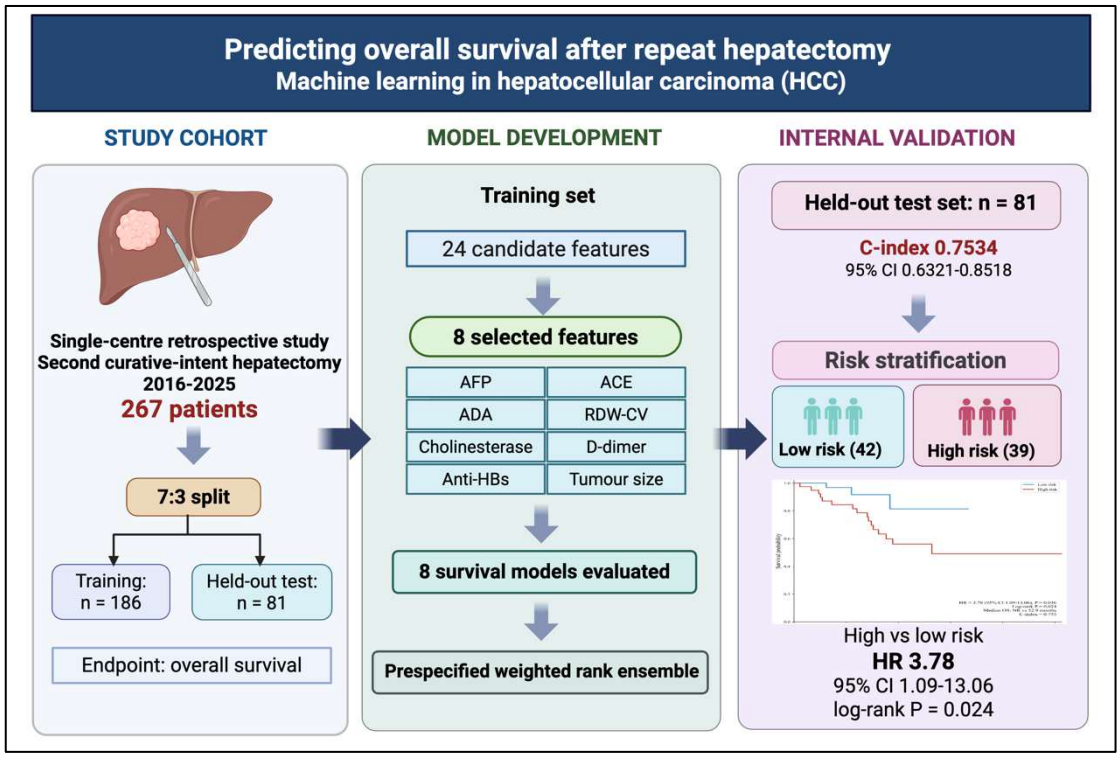


Figure 1. Feature-number selection within the training set. Training 5-fold cross-validated C-index (mean ± SD) for candidate feature numbers k = 4 to 14 drawn from the 24-feature clinical prior pool. AFP was forced into every candidate set by design. The test set was not involved at any point of feature selection.

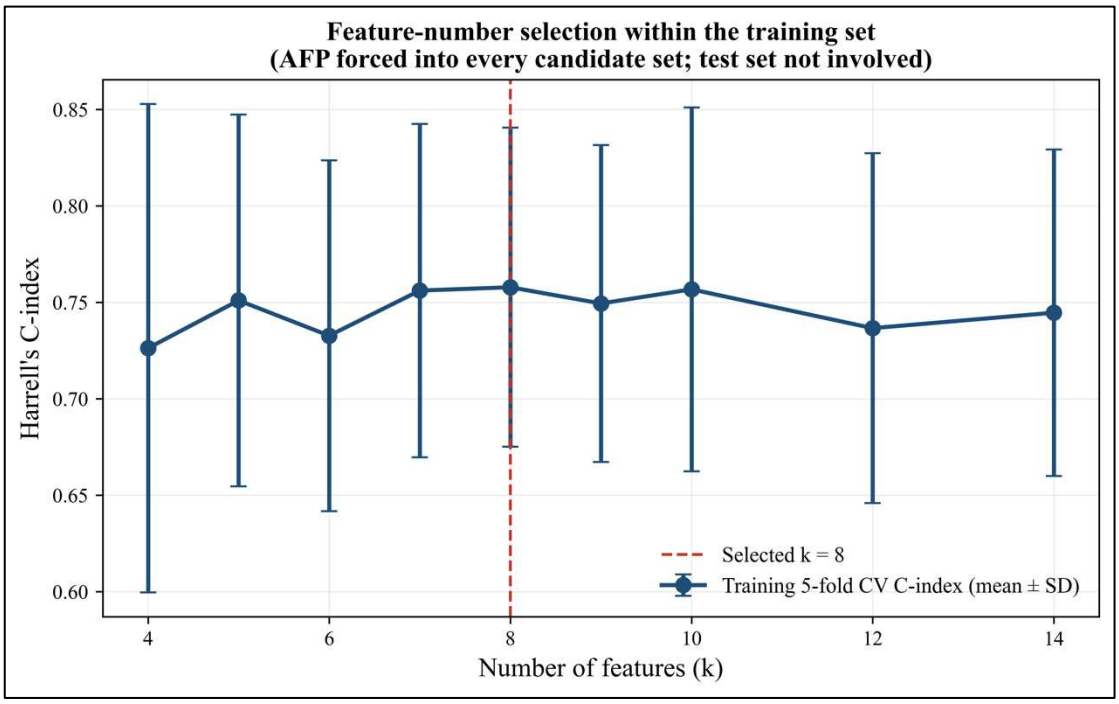


Figure 2. Discrimination of the eight candidate survival models. Training and held-out test-set Harrell C-index for CoxPH, CoxEN, FKSVM, GBSA, RSF, EST, XGBoost Cox and the prespecified weighted rank ensemble.

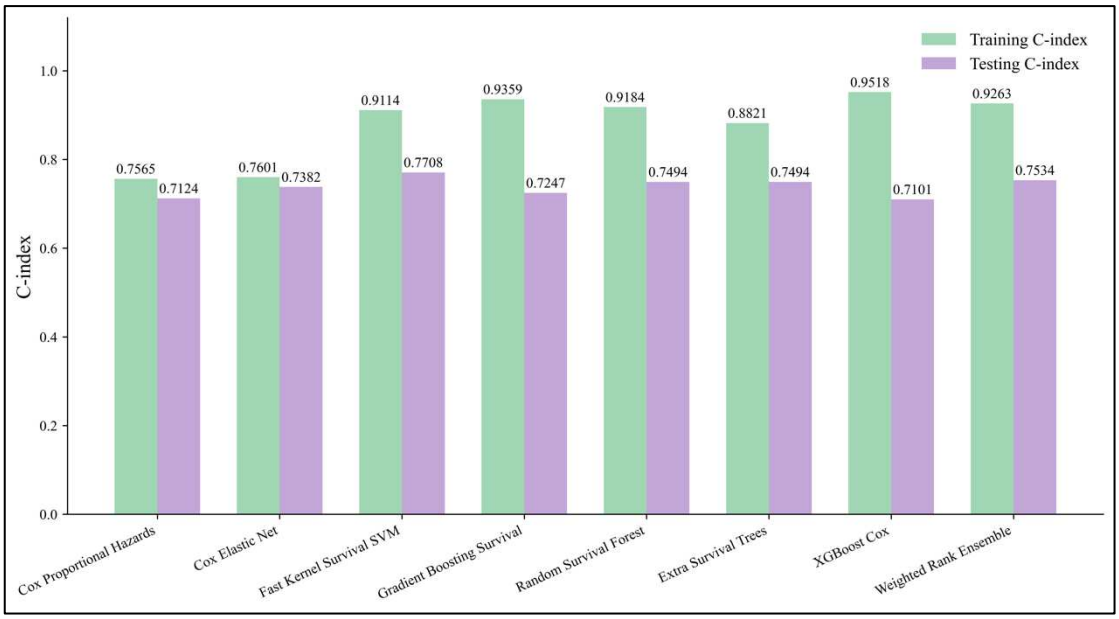


Figure 3. Time-dependent AUC curves (cumulative/dynamic definition) of the eight candidate models on the held-out test set across the evaluation window.

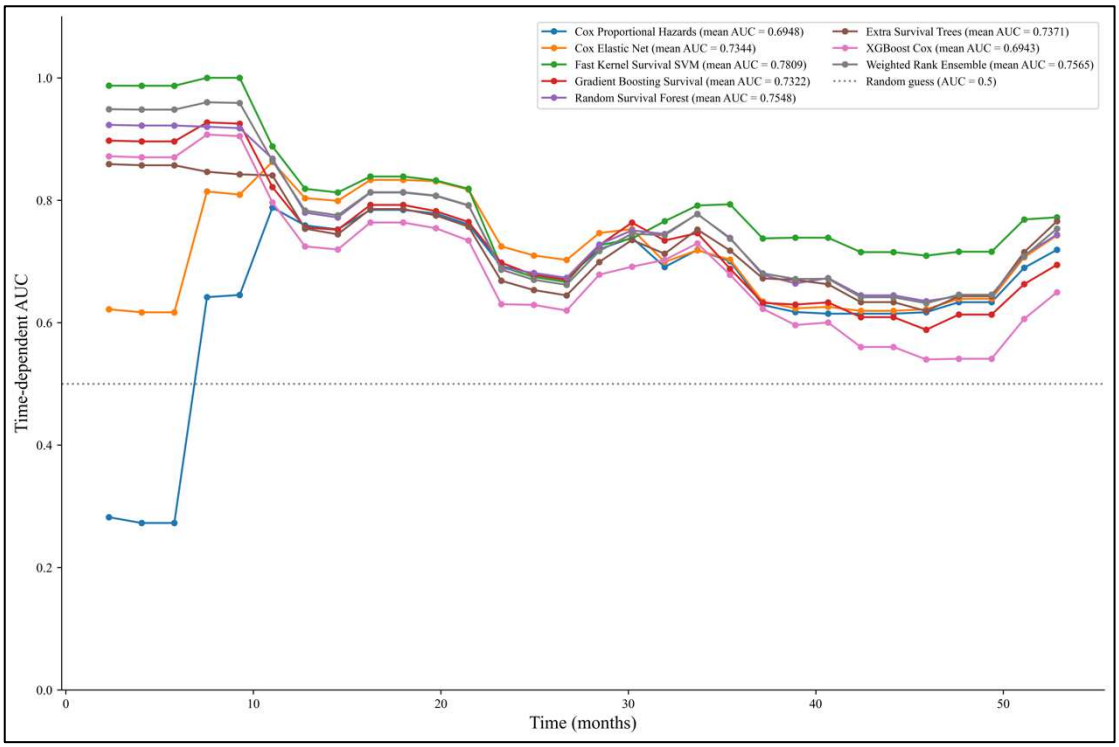


Figure 4. Calibration panels at 12, 24 and 36 months on the test set. Top row: the prespecified weighted rank ensemble (delivery model). Bottom row: the highest test-performing single learner (Fast Kernel Survival SVM). Dots represent quartile groups of predicted survival; shaded bands represent bootstrap 95% CI of Kaplan-Meier estimates; the dashed line indicates perfect calibration. Insets give the logistic recalibration slope and intercept with 95% CIs.

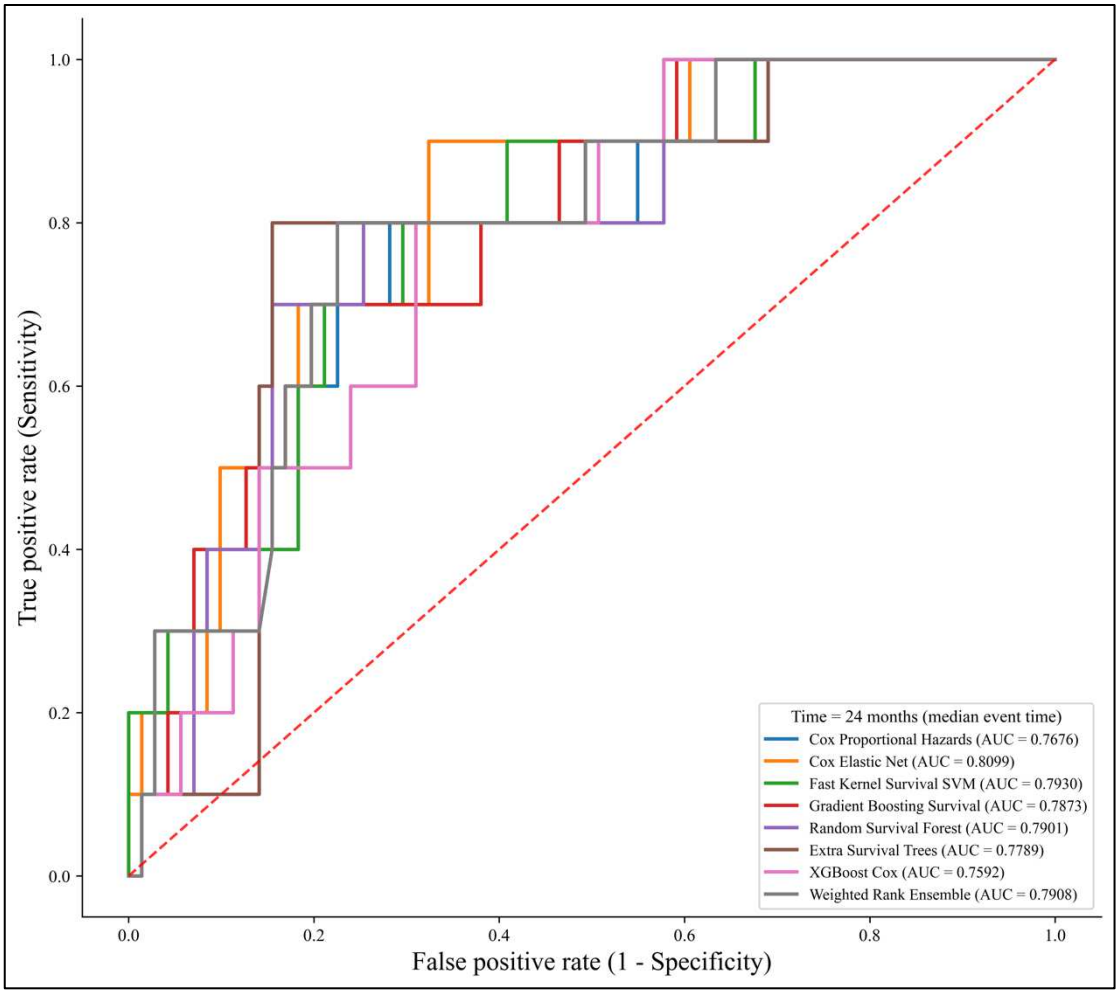


Figure 5. Kaplan-Meier survival by risk group in the held-out test set (primary validation) with numbers at risk. Patients were dichotomised at the training-set median risk score (5.405).

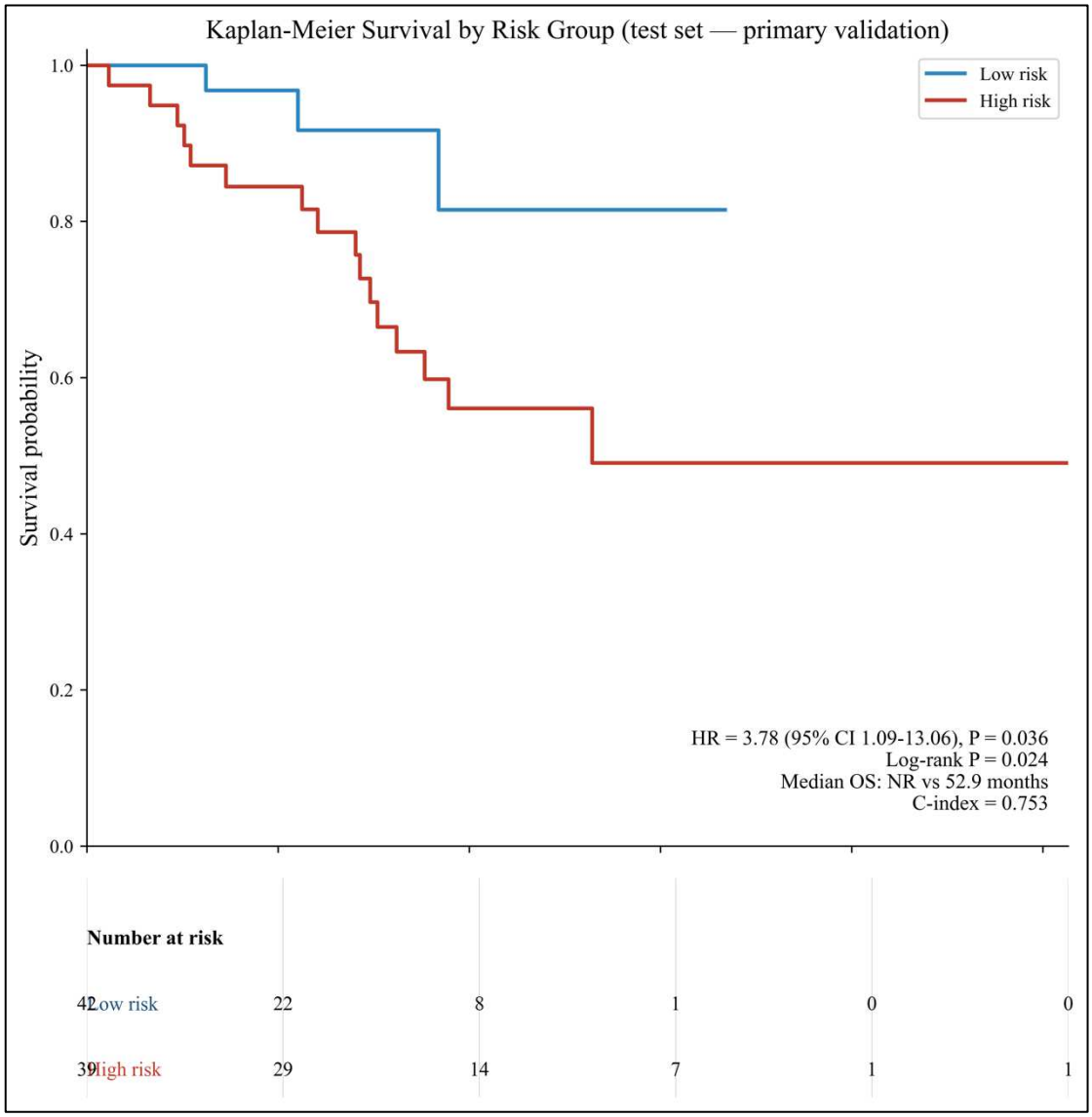


Figure 6. Test-set C-index of the model risk score versus individual clinical factors. The model outperformed AFP, the strongest single clinical factor, in the test set (0.753 vs 0.721).

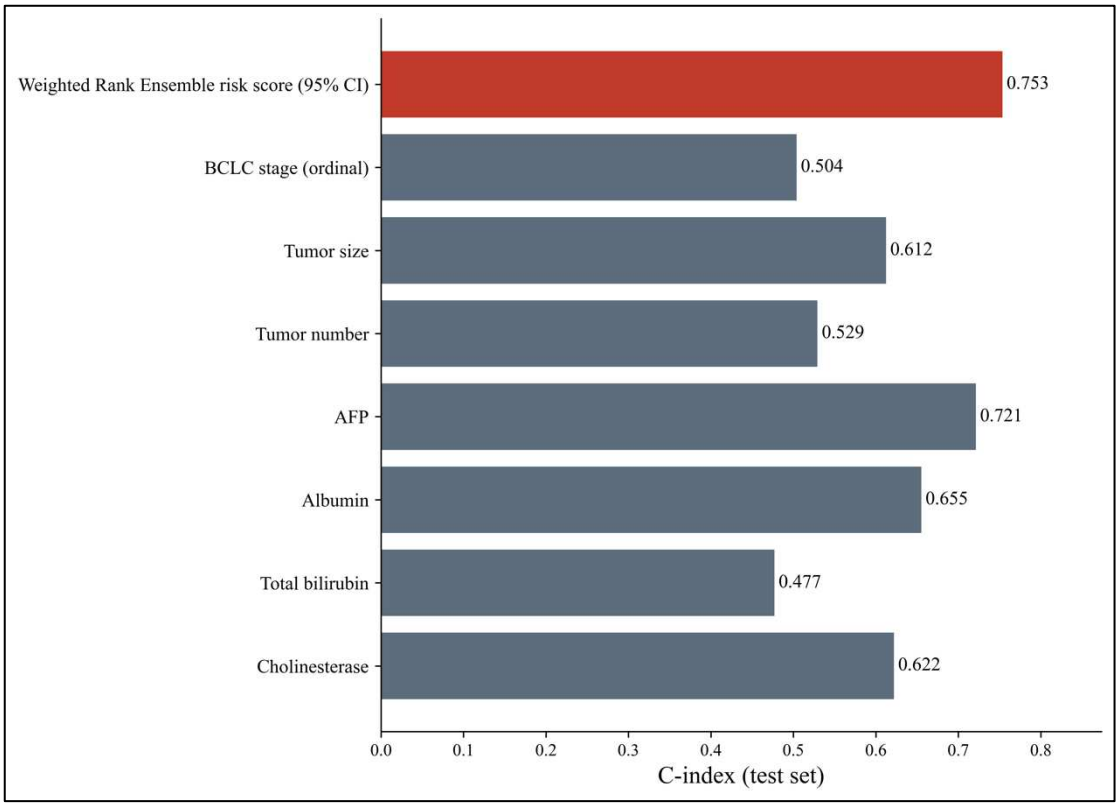


Table 1. Baseline characteristics of the study cohort.

Characteristic	Overall (N=267)	Training set (n=186)	Test set (n=81)	P value
Follow-up time, months	24.41 (12.75-42.79)	24.20 (12.38-42.80)	25.20 (14.03-42.02)	P = 0.453
Death, n (%)	62 (23.2)	43 (23.1)	19 (23.5)	P = 1.000
BCLC stage				P = 0.319
Stage 0	131 (49.1)	98 (52.7)	33 (40.7)	
Stage A	88 (33.0)	57 (30.6)	31 (38.3)	
Stage B	23 (8.6)	14 (7.5)	9 (11.1)	
Stage C	25 (9.4)	17 (9.1)	8 (9.9)	
Tumor number, n (267)				P = 0.364
Solitary	245 (91.8)	168 (90.3)	77 (95.1)	
2 tumours	20 (7.5)	16 (8.6)	4 (4.9)	
>=3 tumours	2 (0.7)	2 (1.1)	0 (0.0)	
Tumor size, cm	3.00 (2.10-4.40)	3.00 (2.00-4.28)	3.00 (2.30-4.50)	P = 0.690
Tumor size > 5 cm, n (%)	52 (19.5)	37 (19.9)	15 (18.5)	P = 0.926
Margin status				P = 0.570
Negative	207 (77.5)	141 (75.8)	66 (81.5)	
Narrow	34 (12.7)	25 (13.4)	9 (11.1)	
R1	26 (9.7)	20 (10.8)	6 (7.4)	
Capsule intact, n (267)				P = 0.236
Not intact	132 (49.4)	87 (46.8)	45 (55.6)	
Intact	135 (50.6)	99 (53.2)	36 (44.4)	
Microvascular invasion				P = 0.580
Absent	160 (59.9)	114 (61.3)	46 (56.8)	
Present	107 (40.1)	72 (38.7)	35 (43.2)	
Satellite nodules				P = 0.384
Absent	239 (89.5)	169 (90.9)	70 (86.4)	
Present	28 (10.5)	17 (9.1)	11 (13.6)	
Edmondson grade, n (267)				P = 0.412

Grade I	1 (0.4)	0 (0.0)	1 (1.2)	
Grade II	128 (47.9)	87 (46.8)	41 (50.6)	
Grade III	129 (48.3)	93 (50.0)	36 (44.4)	
Grade IV	9 (3.4)	6 (3.2)	3 (3.7)	
Cirrhosis				P = 0.681
No	112 (41.9)	76 (40.9)	36 (44.4)	
Yes	155 (58.1)	110 (59.1)	45 (55.6)	
Inflammation grade				P = 0.892
G1-2	188 (70.4)	130 (69.9)	58 (71.6)	
G3-4	79 (29.6)	56 (30.1)	23 (28.4)	
Ki-67, %	30.00 (20.00-50.00)	30.00 (20.00-50.00)	30.00 (20.00-45.00)	P = 0.883
P53 mutant pattern				P = 0.464
Negative pattern	60 (22.5)	39 (21.0)	21 (25.9)	
Mutant pattern	207 (77.5)	147 (79.0)	60 (74.1)	
AFP, ng/mL	10.63 (2.75-196.53)	16.27 (2.81-223.02)	5.59 (2.57-146.20)	P = 0.283
AFP > 400 ng/mL, n (%)	50 (18.7)	34 (18.3)	16 (19.8)	P = 0.910
ALBI grade score	-1.65 ± 0.37	-1.66 ± 0.36	-1.62 ± 0.39	P = 0.516
Prognostic nutritional index (PNI)	198.70 ± 17.91	198.07 ± 17.27	200.14 ± 19.33	P = 0.409
Neutrophil-to-lymphocyte ratio (NLR)	1.87 (1.38-2.54)	1.78 (1.36-2.51)	2.02 (1.42-2.56)	P = 0.283
Platelet-to-lymphocyte ratio (PLR)	101.37 (79.74-124.07)	101.59 (79.67-122.95)	101.14 (80.36-126.83)	P = 0.884
Systemic immune-inflammation index (SII)	309.37 (205.29-465.07)	298.63 (204.77-456.56)	350.55 (220.98-470.84)	P = 0.307
Albumin, g/L	39.67 ± 3.58	39.55 ± 3.45	39.96 ± 3.86	P = 0.410
Total bilirubin, umol/L	13.70 (10.70-17.15)	13.35 (10.43-16.30)	15.10 (11.80-19.20)	P = 0.022
ALT, U/L	23.00 (17.00-31.62)	23.00 (17.00-32.95)	24.00 (17.00-30.00)	P = 0.931
AST/ALT ratio	1.29 (1.00-1.62)	1.26 (1.00-1.63)	1.29 (1.00-1.59)	P = 0.755
GGT, U/L	32.00 (24.00-56.50)	32.00 (24.00-58.25)	34.00 (21.00-52.00)	P = 0.513

Prothrombin time, s	12.50 (11.70-13.20)	12.50 (11.70-13.20)	12.43 ± 1.32	P = 0.941
Platelets, 10^9/L	170.00 (135.00-207.10)	168.00 (131.25-207.60)	172.49 ± 53.48	P = 0.722
Hemoglobin, g/L	140.00 (129.75-150.00)	138.81 ± 14.65	141.19 ± 15.33	P = 0.239
WBC, 10^9/L	5.41 (4.42-6.45)	5.42 (4.40-6.30)	5.33 (4.45-6.52)	P = 0.375
Creatinine, umol/L	77.00 (69.00-89.00)	76.00 (69.00-88.00)	80.00 (71.00-90.00)	P = 0.193
Sodium, mmol/L	141.00 (140.00-142.00)	141.00 (140.00-142.00)	141.00 (140.00-142.00)	P = 0.922
CEA, ng/mL	2.04 (1.50-2.87)	2.09 (1.50-2.84)	1.94 (1.47-3.04)	P = 0.992
CA 19-9, U/mL	6.40 (3.35-12.15)	6.85 (3.23-13.78)	5.50 (3.60-9.40)	P = 0.099

Data are presented as mean ± SD for normally distributed continuous variables, median (IQR) for non-normally distributed continuous variables, and n (%) for categorical variables. Continuous variables were compared using the Student t-test or Mann-Whitney U test according to the Shapiro-Wilk test; categorical variables were compared using the χ^2 test or Fisher exact test. Laboratory variables reflect the hospitalisation of the second (index) hepatectomy. BCLC, Barcelona Clinic Liver Cancer; MVI, microvascular invasion; AFP, alpha-fetoprotein; ALT, alanine aminotransferase; AST, aspartate aminotransferase; GGT, gamma-glutamyl transferase; WBC, white blood cell count; ALBI, albumin-bilirubin grade; PNI, prognostic nutritional index; NLR, neutrophil-to-lymphocyte ratio; PLR, platelet-to-lymphocyte ratio; SII, systemic immune-inflammation index. The biphenotypic subtype denotes tumours with mixed hepatocellular-cholangiocarcinoma features on histology.

Supplementary Files

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

- SupplementaryMaterials.docx