ACCSCIENCE PUBLISHING

Eurasian Journal of Medicine and Oncology

ORIGINAL RESEARCH ARTICLE

An integrated clinical risk scoring and machine learning framework for risk stratification of hepatocellular carcinoma: Development and real-world clinical validation

Yingying Luo[1†], Zengbei Yuan[2†], Yingkang Lin[3†], XiaoLei Huang[1], Feiyang Pu[4], Rui Li[1], Lihong Liu[1*], and Zhihua Ni[1*]

[1]Department of Radiology, Shanghai Baoshan Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai, China

[2]School of Medical Technology and Engineering, Fujian Medical University, Fuzhou, Fujian, China

[3]Department of Rehabilitation, Guangdong Work Injury Rehabilitation Hospital, Guangzhou, Guangdong, China

[4]Department of Nuclear Medicine, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China

[†]*These authors contributed equally to this work.*

***Corresponding authors:**
Lihong Liu
(llh656@qq.com);
Zhihua Ni
(19713115@qq.com)

Citation: Luo Y, Yuan Z, Lin Y, *et al*. An integrated clinical risk scoring and machine learning framework for risk stratification of hepatocellular carcinoma: Development and real-world clinical validation. *Eurasian J Med Oncol.*
doi: 10.36922/EJMO026310393

Received: July 28, 2026

Revised: September 10, 2026

Accepted: September 17, 2026

Published online: November 16, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract

Introduction: Hepatocellular carcinoma (HCC) remains a major cause of cancer-related mortality, highlighting the need for more accurate risk stratification approaches.

Objective: To develop and validate an integrated clinical risk scoring (CRS) and machine learning (ML) system to accurately identify HCC, overcoming the limitations of traditional alpha-fetoprotein (AFP) screening, especially in AFP-negative individuals.

Methods: A retrospective cross-sectional diagnostic study utilized data from two institutions (development cohort n = 5,000; external validation cohort n = 443) with 12 routine clinical variables. Multivariable logistic regression was used to construct the CRS, and ensemble ML models were developed. Performance was assessed via area under the receiver operating characteristic curve (AUC), quantitative calibration measures, decision curve analysis (DCA), net reclassification improvement (NRI), and integrated discrimination improvement (IDI). We applied development-trained models directly to the independent cohort for strict external validation.

Results: Hepatitis C, diabetes and hepatitis B were identified as the strongest independent risk factors for HCC. In the development cohort, the CatBoost model demonstrated the best overall discriminatory ability (AUC = 0.9773), significantly outperforming the single AFP indicator (0.6377) and CRS (0.9385); notably, in the independent validation cohort, CRS demonstrated superior stability in terms of AUC (0.9849) and risk reclassification ability (NRI/IDI), performing slightly better than CatBoost (AUC = 0.9816). DCA, NRI, and IDI confirmed significant incremental clinical net benefit and risk reclassification over AFP-only screening (p < 0.001).

Conclusion: This integrated HCC risk stratification system effectively overcomes traditional AFP screening limitations, particularly in AFP-negative populations. The observed bias-variance trade-off supports a tiered deployment strategy: CRS for robust screening in resource-limited settings, and ML for precision assessment in well-resourced centers.

Keywords: Hepatocellular carcinoma; Machine learning; Clinical risk scoring; Alpha-fetoprotein

1. Introduction

Primary liver cancer primarily comprises three major histological subtypes: hepatocellular carcinoma (HCC), intrahepatic cholangiocarcinoma, and combined hepatocellular-cholangiocarcinoma.[1] HCC is the most prevalent of these subtypes, accounting for around 75%–85% of all primary liver cancer cases.[2] According to the latest statistics from the National Cancer Center, there were approximately 316,000 liver cancer-related deaths in China, ranking it as the second leading cause of cancer-related mortality and imposing a substantial disease burden.[3] Due to the lack of specific clinical manifestations in the early stages, most HCC patients are diagnosed at an advanced stage, missing the window for curative therapies and severely compromising their overall survival and prognosis.[4,5] Therefore, establishing robust and precise risk assessment and early screening strategies for HCC is of paramount clinical significance to improve early diagnosis rates, optimize clinical decision-making, and ultimately enhance patient outcomes.

Currently, screening populations at high risk of HCC primarily relies on alpha-fetoprotein (AFP) combined with hepatic imaging.[6] However, the usefulness of AFP as a standalone serological biomarker is limited due to its suboptimal sensitivity and specificity.[7] AFP levels often remain normal in the early stages of HCC and may even stay within the normal range in some patients with advanced disease.[8] Furthermore, conditions such as active chronic hepatitis or liver cirrhosis can cause spurious AFP elevations, further compromising its diagnostic specificity.[9] Ultrasonography, the most widely utilized imaging modality, has limited sensitivity for detecting minute early-stage lesions, resulting in an overall early diagnosis rate of less than 30% in China.[10,11] Consequently, relying on AFP combined with conventional imaging does not meet the demands of precise HCC screening. There is an urgent need to develop robust risk stratification models that integrate multidimensional clinical information to enhance the identification of high-risk individuals and improve the efficiency of early diagnosis.[12]

In recent years, advancements in machine learning (ML) have introduced a novel approach to predicting disease risk.[13] Compared with traditional statistical methods, ML is better at identifying complex, non-linear relationships and high-order interactions between clinical variables.[14] This makes it particularly well-suited to modeling heterogeneous clinical data. Previous studies have shown that ensemble learning algorithms such as extreme gradient boosting (XGBoost) and CatBoost exhibit superior predictive performance in oncological risk prediction, consequently garnering widespread attention.[15,16] Furthermore, for successful clinical translation, it is critical that a model can maintain robust performance on independent clinical data.[17] However, the majority of current studies develop and validate their models using single-center data, lacking independent external validation.[18] This yields insufficient evidence regarding model generalizability and clinical applicability. Therefore, validation utilizing multicenter or real-world clinical cohorts is essential to bolster clinician confidence and ensure the robust generalizability of these predictive models.[19]

Based on the aforementioned background, the target population for this study comprises adults with underlying chronic liver conditions or metabolic risk factors who are candidates for routine HCC surveillance. This study aimed to develop and validate a comprehensive HCC risk stratification system integrating a clinical risk score (CRS) with ML using bicentric cohorts. We systematically compared the predictive performance of the CRS and ML models against traditional AFP screening, with a specific focus on their utility in AFP-negative populations, and evaluated their clinical net benefit and incremental predictive value through rigorous external validation. While advanced molecular and metabolic markers are increasingly used in specialized settings, they remain inaccessible in many real-world scenarios. Therefore, our framework is specifically designed to augment existing AFP-based surveillance by integrating multidimensional, routinely available clinical data to refine risk stratification, particularly in the challenging AFP-negative subgroup. Ultimately, this study seeks to establish an evidence-based, tiered decision-support framework that balances predictive performance with clinical interpretability, providing a reliable foundation for precision screening and individualized risk management in high-risk HCC populations.

2. Materials and methods

2.1. Enrolled participants

This study employed a bicentric retrospective cross-sectional diagnostic study design, comprising a large-

scale public database for model development and an independent real-world clinical dataset for strict external validation. The index date for each patient was defined as the date of their initial clinical presentation or admission. The outcome (HCC+ or HCC−) was determined based on definitive diagnostic criteria established at or shortly after this index date. Crucially, all 12 routine clinical predictor variables were extracted from baseline medical records at the time of this initial presentation, prior to or concurrent with the definitive HCC diagnosis, thereby mitigating any potential reverse causation bias.

Specifically, the development cohort (Center 1) was derived from a publicly available Kaggle database (https://www.kaggle.com/datasets/miadul/predict-liver-cancer-from-and-clinical-features), including 5,000 patients (1,089 HCC cases [HCC+]), which was used for feature evaluation, model training, and hyperparameter optimization. The independent validation cohort (Center 2) consisted of a real-world clinical dataset collected from our institution, including 443 patients (84 HCC+), serving strictly to evaluate the generalizability of the optimized technical pipeline. No model refitting or new hyperparameter tuning was performed on the Center 2 data.

The overall methodological workflow encompassed the following core modules: (i) data quality control and preprocessing; (ii) screening of clinical risk factors and construction of a CRS; (iii) development and hyperparameter tuning of MLs in Center 1; (iv) strict external validation and comparison of multi-strategy predictive performance; (v) analysis of the independent predictive value of AFP and evaluation of clinical net benefit (decision curve analysis [DCA]), net reclassification improvement (NRI), and integrated discrimination improvement (IDI); and (vi) incremental model gain and performance analysis.

2.2. Data preprocessing and quality control

As real-world clinical data was used, no automatic imputation or algorithmic correction was applied. Instead, a rigorous combined manual and statistical quality control protocol was implemented. This protocol specifically comprised:

(i) Data integrity verification: Screening for missing values and duplicate records to ensure completeness and non-redundancy across both centers.
(ii) Outlier screening and clinical verification. The interquartile range method was used to identify potential outliers in continuous variables. Any extreme values identified were only retained after clinical assessment and medical record review confirmed their validity in real-world clinical scenarios.
(iii) Multicollinearity assessment: A correlation coefficient matrix was used to evaluate multicollinearity among the variables. For highly correlated variables, retention was determined based on a clinical value assessment conducted by physicians. Concurrently, MLs with inherent robustness to multicollinearity were selected to mitigate potential interference.

This study incorporated 12 routine clinical variables (Table 1), covering dimensions such as demographics, lifestyle factors, infection history, medical history, and laboratory investigations. All variables were pre-specified for inclusion based on clinical relevance, with no variables eliminated solely on the basis of statistical significance. Detailed population information is provided in Table S1.

2.3. Clinical risk factor analysis and construction

Initially, univariate logistic regression analysis was performed to evaluate the association between each clinical variable and the risk of HCC development, yielding odds ratios (ORs) and their corresponding 95% confidence intervals (CIs). Subsequently, all 12 pre-specified clinical variables were incorporated into a multivariable logistic regression model to determine the regression coefficient (β), OR, and 95% CI for each predictor.

Based on the multivariable logistic regression results, a CRS was constructed utilizing the following formula (**Equation 1**):

$$CRS = \Sigma(\beta i \times Xi) \quad (1)$$

where βi denotes the multivariable regression coefficient of the i-th variable, and Xi represents the corresponding value of that variable.

To enable straightforward clinical interpretation, the raw CRS was transformed using a linear scale and standardized to a range of 0–100. Patients were then categorized as low-, intermediate-, or high-risk based on the CRS tertiles, and the actual incidence rates of HCC across these risk categories were compared to validate the clinical usefulness of the categorization.

2.4. Development of machine learning models

Based on the same 12 clinical features, this study developed and compared three ML models:

(i) Logistic regression (LR): This served as the conventional baseline model, with continuous variables standardized via *Z*-score normalization prior to model fitting.[20]
(ii) XGBoost: An efficient ensemble learning algorithm operating within a gradient boosting framework. Hyperparameter optimization was conducted using

Table 1. The demographic information of enrolled subjects (mean ± standard deviation)

Characteristic	Center 1				Center 2			
	Total (*n*)	HCC+	HCC−	*p*-value	Total (*n*)	HCC+	HCC−	*p*-value
Number	5,000	1,089 (21.8%)	3,911 (78.2%)		443	84 (19.0%)	359 (81.0%)	
Age (years), mean ± SD (range)	57 ± 27	66 ± 22 [30–84]	54 ± 26 [30–84]	<0.001	64 ± 12	70 ± 10 [50–97]	63 ± 12 [18–88]	0.024
Gender				0.956				0.004
Male	2,494 (49.9%)	544 (50%)	1,950 (49.9%)		260 (58.7%)	61 (72.6%)	199 (55.4%)	
Female	2,506 (50.1%)	545 (50%)	1,961 (50.1%)		183 (41.3%)	23 (27.4%)	160 (44.6%)	
AFP				<0.001				<0.001
<25 ng/mL	4,037 (80.7%)	709 (65.1%)	3,328 (85.1%)		399 (90.1%)	40 (47.6%)	359 (100%)	
>25 ng/mL	963 (19.3%)	380 (34.9%)	583 (14.9%)		44 (9.9%)	44 (52.4%)	0	
BMI (kg/m^2), mean ± SD		27.9 ± 8.3	25.5 ± 4.9	<0.001		23.2 ± 3.6	24.5 ± 3.3	0.68

Note: Data presented as *n* (%), unless stated otherwise.
Abbreviations: AFP: Alpha-fetoprotein; BMI: Body mass index; HCC: Hepatocellular carcinoma; SD: Standard deviation.

5-fold cross-validation coupled with grid search.[21]

(iii) CatBoost: A gradient boosting algorithm utilizing symmetric (oblivious) trees which offers native support for categorical features and effectively mitigates overfitting. Hyperparameter optimization was performed in a similar manner via 5-fold cross-validation and grid search.[22]

During the training and validation phases, model performance was rigorously assessed using 5-fold cross-validation. The primary evaluation metrics comprised the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, and the F1-score. All models were implemented in a Python environment (version 3.9) using Scikit-learn and associated open-source ML libraries. The ML models were developed and optimized using the Center 1 cohort, where hyperparameter tuning was conducted via 5-fold cross-validation coupled with grid search. Subsequently, the real-world clinical dataset (Center 2) served strictly as an independent validation cohort. To robustly evaluate the generalizability of this optimized technical pipeline on a smaller, independent dataset, we applied 5-fold cross-validation in Center 2. Crucially, no new hyperparameter search or model refitting was performed during this phase; each fold was trained using the exact, pre-determined hyperparameters established in Center 1. This approach assesses the stability of the fixed technical pipeline in a new population without introducing data leakage.

To rigorously prevent data leakage and ensure the independence of the evaluation metrics, a strict leakage-free pipeline was implemented during the 5-fold cross-validation. Specifically, all data preprocessing steps, including *Z*-score normalization for continuous variables, were calculated exclusively using the parameters (mean and standard deviation) of the training folds and subsequently applied to the untouched test fold. No global scaling or data-driven feature selection was performed prior to data splitting; all 12 variables were pre-specified based on clinical relevance. Furthermore, hyperparameter optimization via grid search was conducted strictly within the training folds, and no automatic imputation was applied to the real-world data. This protocol ensured that no information from the test folds leaked into the model training or tuning phases, guaranteeing that the reported high AUC values reflect genuine predictive performance.

2.5. Alpha-fetoprotein-independent sensitivity analysis

To evaluate the reliance of the ML models on the conventional tumor biomarker AFP, an AFP-independent sensitivity analysis was conducted. First, each ML model was retrained after excluding the AFP variable, and the corresponding variations in the AUC were compared. Second, within the subgroup with normal AFP levels (AFP < 25 ng/mL), the discriminative performance of both the full model and the AFP-excluded model was evaluated to

verify the clinical screening value of the models in AFP-negative high-risk populations.[23]

2.6. Evaluation of clinical utility

We employed DCA to evaluate the clinical net benefit of different predictive strategies. The net benefit of each model was calculated across a range of threshold probabilities (from 0 to 1). Additionally, reference lines representing the "treat all" and "treat none" strategies were plotted to assess the practical utility of each model visually across various clinical decision thresholds.

2.7. Evaluation of model incremental value

Net reclassification improvement and IDI were used to quantify the incremental predictive value of the novel models compared to the baseline models.[24] To mitigate the sensitivity of categorical NRI to arbitrary risk boundaries, we additionally calculated the category-free NRI (cfNRI), which assesses the overall upward or downward shift in predicted probabilities across the entire cohort without predefined categories. For the categorical NRI, risk stratification thresholds of 10% and 30% were set to simulate clinical decision zones: low risk (<10%, routine surveillance), intermediate risk (10%–30%, intensive 6-month surveillance), and high risk (>30%, immediate diagnostic workup). To quantify statistical uncertainty, 95% CIs for all NRI and IDI estimates were calculated using a 1,000-iteration bootstrap resampling procedure. Statistical significance was evaluated using a *Z*-test.

To comprehensively assess the incremental predictive value, six pairwise model comparisons were made with the baseline models: CRS vs. AFP, XGBoost vs. AFP, CatBoost vs. AFP, XGBoost vs. CRS, CatBoost vs. CRS, and CatBoost vs. XGBoost.

All statistical analyses and ML models were conducted in a Python environment using associated open-source libraries, including scikit-learn. Continuous variables were initially assessed for normality. Those exhibiting a normal distribution were expressed as mean ± standard deviation, and between-group comparisons were conducted using the independent-samples *t*-test. Those exhibiting a non-normal distribution were presented as median (interquartile range), and between-group comparisons were performed using the Mann–Whitney *U* test. Categorical variables were reported as frequencies and percentages, and between-group comparisons were performed using the chi-squared test or Fisher's exact test as appropriate. Metrics for evaluating model performance, including DCA, NRI, and IDI, were computed using specialized statistical and algorithmic packages. All statistical tests were two-tailed, and a *p*-value of less than 0.05 was considered statistically significant.

3. Results

This study used a retrospective design based on a bicentric cohort to develop and validate a risk stratification model for HCC (Figure 1). The overall methodological workflow was grounded in rigorous data quality control and analysis of baseline characteristics. Firstly, independent risk factors were identified using multivariable logistic regression to construct a CRS, which enabled the quantitative assessment of risk and clinical stratification of patients. Next, ML models, notably CatBoost, were developed using the full set of clinical features. These models were then systematically compared with an AFP-only model and the CRS model. Furthermore, an AFP-independent sensitivity analysis was conducted to verify the predictive value of comprehensive clinical information in isolation. Finally, DCA, NRI, and IDI were used to evaluate the clinical net benefit and incremental predictive value of each model comprehensively.

3.1. Results of clinical risk score analysis

Multivariable logistic regression analysis identified 10 statistically significant independent risk factors for HCC ($p < 0.05$) in the Center 1 cohort. The most robust risk factors for HCC were hepatitis C (OR = 16.1135, 95% CI: 11.7527–22.0926, $p < 0.001$), diabetes (OR = 15.0162, 95% CI: 11.5792–19.4733, $p < 0.001$), and hepatitis B (OR = 13.0912, 95% CI: 9.9651–17.1979, $p < 0.001$), in descending order of ORs. Although gender (OR = 0.9767, $p = 0.8172$) and physical activity level (OR = 0.9857, $p = 0.8336$) did not reach statistical significance, they were retained in the model based on their established clinical relevance and to prevent omitted variable bias (Table 2).

In the Center 2 cohort, the multivariable logistic regression analysis exhibited a distinct pattern compared to that observed in Center 1 due to the relatively small sample size ($n = 443$) and the highly imbalanced distribution of certain variables. The cirrhosis history (OR = 4.69×10^9, $p < 0.001$) and physical activity level (OR = 416.6553, $p < 0.001$) were identified as the most significant risk factors, while hepatitis B (OR = 0.0138, $p < 0.001$) and alcohol consumption (coefficient = −9.4179 [OR = 0.0001], $p < 0.001$) emerged as protective factors. Hepatitis C was an invariant variable with a value of zero in Center 2 (as none of the patients had a history of hepatitis C infection), thereby providing no discriminative information. The remaining variables, including age, BMI, AFP, sex, family history of malignancy, diabetes mellitus, and smoking status, did not reach statistical significance (Table 3).

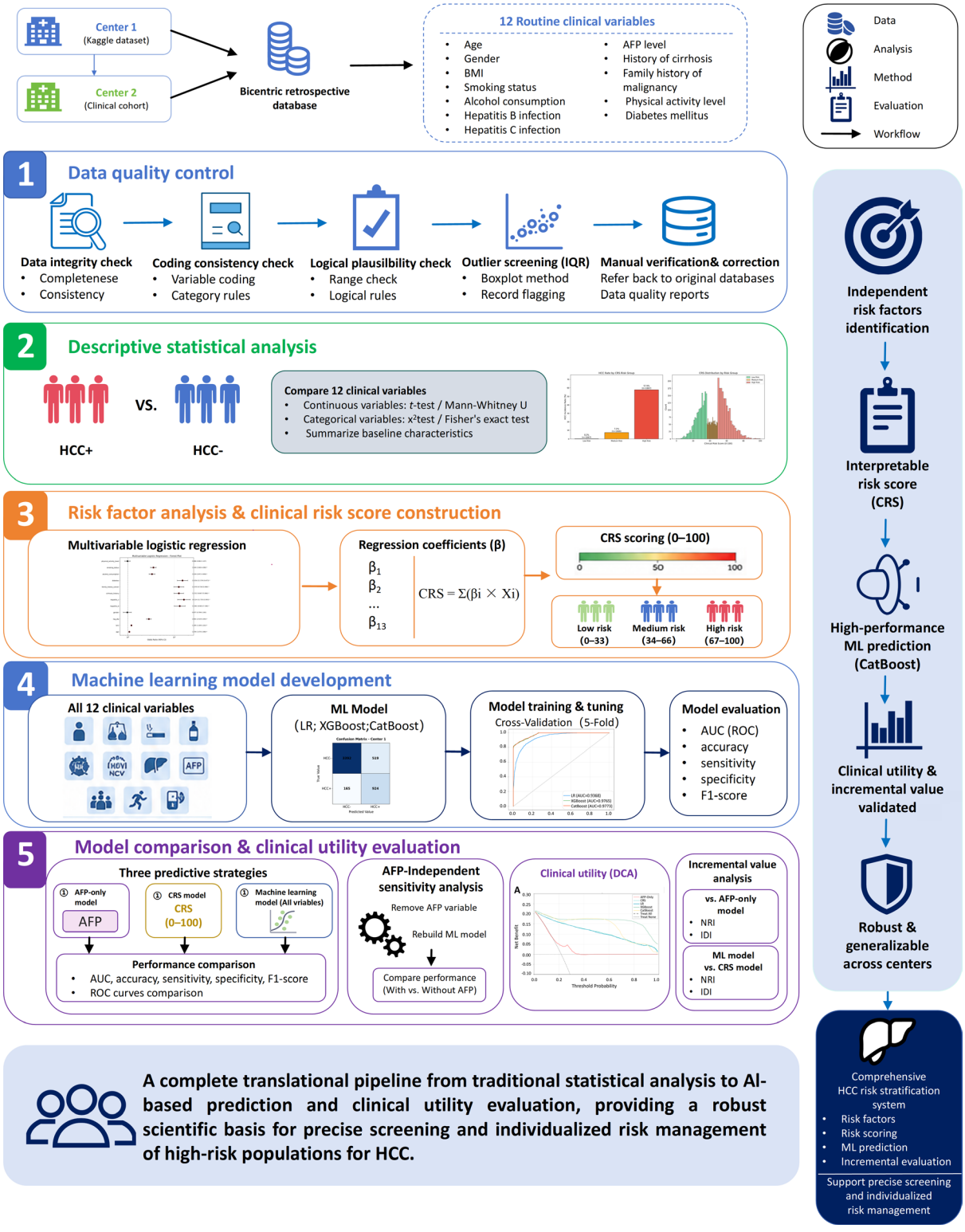


Figure 1. Overview of the framework for the risk stratification model for hepatocellular carcinoma (HCC). Created using Microsoft PowerPoint. Abbreviations: AFP: Alpha-fetoprotein; AUC: Area under the receiver operating characteristic curve; IDI: Integrated discrimination improvement; ML: Machine learning; NRI: Net reclassification improvement.

Table 2. Results of multivariate logistic regression (Center 1)

Variable	Coefficient	Odds ratio (95% confidence interval)	*p*-value
Age (years)	0.0749	1.0778 (1.0697–1.0860)	<0.001
Body mass index (kg/m^2)	0.1783	1.1952 (1.1694–1.2216)	<0.001
Log alpha-fetoprotein	1.0442	2.8411 (2.5188–3.2046)	<0.001
Gender	−0.0236	0.9767 (0.7993–1.1933)	0.8172
Hepatitis B	2.5719	13.0912 (9.9651–17.1979)	<0.001
Hepatitis C	2.7797	16.1135 (11.7527–22.0926)	<0.001
Cirrhosis history	2.5267	12.5120 (9.8667–15.8664)	<0.001
Family history of cancer	2.5233	12.4695 (9.7375–15.9680)	<0.001
Diabetes	2.7091	15.0162 (11.5792–19.4733)	<0.001
Alcohol consumption	1.1997	3.3191 (2.8705–3.8378)	<0.001
Smoking status	1.2666	3.5487 (3.0804–4.0881)	<0.001
Physical activity level	−0.0144	0.9857 (0.8622–1.1270)	0.8336

Table 3. Results of multivariate logistic regression (Center 2)

Variable	Coefficient	Odds ratio (95% confidence interval)	*p*-value
Age (years)	0.3752	1.4554 (0.2050–10.3320)	0.7075
Body mass index (kg/m^2)	−0.0191	0.9810 (0.1382–6.9647)	0.9847
Log alpha-fetoprotein	1.6115	5.0102 (0.7057–35.5692)	0.1071
Gender	−0.2383	0.7879 (0.1110–5.5938)	0.8116
Hepatitis B	−4.2860	0.0138 (0.0019–0.0977)	<0.001
Hepatitis C	0.0000	1.0000 (0.1409–7.0993)	1.0000
Cirrhosis history	22.2693	4.69×10^9 (0.66×10^9–33×10^9)	<0.001
Family history of cancer	0.0871	1.0910 (0.1537–7.7456)	0.9306
Diabetes	0.0784	1.0815 (0.1523–7.6782)	0.9375
Alcohol consumption	−9.4179	0.0001 (0.0000–0.0006)	<0.001
Smoking status	0.8062	2.2395 (0.3155–15.8989)	0.4201
Physical activity level	6.0323	416.6553 (58.6894–2957.9720)	<0.001

3.2. Clinical risk score

The CRS, which was constructed using multivariable logistic regression coefficients, demonstrated excellent discriminative performance across both centers. Risk stratification based on CRS tertiles revealed a distinct risk gradient in both cohorts (Table 4). In Center 1, the incidence rates of HCC were 0.18%, 7.20%, and 57.95% in the low-, intermediate-, and high-risk groups, respectively. In Center 2, the corresponding rates were 0%, 1.36%, and 55.41%.

3.3. Performance of machine learning models

The results of the 5-fold cross-validation for the three ML models with full features across both centers are summarized in Table 5 and Figures 2 and 3. In the Center 1 development cohort, CatBoost achieved optimal performance (AUC = 0.9773, ACC = 0.9590), outperforming LR significantly (AUC = 0.9368, ACC = 0.8632). When evaluating the independent validation cohort (Center 2) using the fixed technical pipeline optimized in Center 1, CatBoost maintained high predictive efficacy (AUC = 0.9816, ACC = 0.9684), as summarized by the 5-fold cross-validation results in Table 5. Therefore, all structures are based on the mean value obtained from 5-fold cross-validation.

3.4. Comparison of different predictive strategies

Table 6 and Figure 4 summarize the performance comparison among the five predictive strategies. Across both centers, the CRS and ML models significantly outperformed the AFP-only model. In the Center 1 cohort specifically, the AFP-only model yielded a suboptimal AUC of 0.6377. In contrast, the CRS achieved an AUC of 0.9385 (an absolute AUC increase of 30.08%), and CatBoost reached an AUC of 0.9773 (an absolute AUC increase of 33.96%). Similarly, in the Center 2 cohort, the AFP-only model demonstrated an AUC of 0.8031, compared to 0.9849 for the CRS (an improvement of 18.18%) and 0.9816 for CatBoost (an improvement of 17.85%). Interestingly, in the Center 2 validation cohort, the CRS exhibited a slightly higher AUC (0.9849) than CatBoost (0.9816). This suggests that, under certain data structures, a simple linear scoring model can perform similarly to complex ML models.

3.5. Alpha-fetoprotein-independent sensitivity analysis

Table 7 summarizes the variations in model performance following the exclusion of AFP. In Center 1, excluding AFP resulted in a 3.55% and 3.56% decrease in the AUC for both XGBoost and CatBoost, respectively. Conversely, in Center 2, the change in AUC was minimal (≤0.75%), suggesting that the models are not overly dependent on AFP. Furthermore, a subgroup analysis of patients with normal AFP levels (AFP < 25 ng/mL) showed that the discriminative performance of the models excluding AFP was comparable to, or even slightly better than, that of the full models. For example, in the Center 1 cohort, the full CatBoost model yielded an AUC of 0.9773, whereas the AFP-excluded model achieved an AUC of 0.9417, demonstrating virtually identical performance. These results highlight that comprehensive clinical information can sustain robust risk identification capabilities in AFP-negative patients.

3.6. Quantitative calibration assessment

Beyond visual calibration curves, quantitative calibration measures were explicitly computed to assess model reliability. In the Center 1 internal cohort, CatBoost demonstrated excellent discrimination but exhibited an extreme calibration slope (3.017), indicating overconfident probability outputs typical of tree-based ensembles. The CRS and LR maintained near-ideal calibration slopes (0.998 and 1.030, respectively) and intercepts close to zero.

In the strict external validation (Center 2), the calibration slopes for all models dropped significantly (<1), reflecting

Table 4. Results of clinical risk score stratification

Risk group	Center 1			Center 2		
	n	HCC cases	HCC rate	*n*	HCC cases	HCC rate
Low risk	1,667	3	0.18%	148	0	0%
Intermediate risk	1,666	120	7.20%	147	2	1.36%
High risk	1,667	966	57.95%	148	82	55.41%

Abbreviation: HCC: Hepatocellular carcinoma.

Table 5. Five-fold cross-validation performance of the machine learning models

Center	Model	AUC	ACC	SEN	SPE	F1
Center 1	LR	0.9368	0.8632	0.8485	0.8673	0.7299
	XGBoost	0.9765	0.9516	0.8347	0.9841	0.8825
	CatBoost	0.9773 [0.9733–0.9810]	0.9590 [0.9540–0.9648]	0.8182 [0.7968–0.8423]	0.9982 [0.9967–0.9992]	0.8968 [0.8834–0.9113]
Center 2	LR	0.9754	0.9549	0.9286	0.9610	0.8864
	XGBoost	0.9797	0.9436	0.9762	0.9359	0.8677
	CatBoost	0.9816 [0.9568–0.9958]	0.9684 [0.9503–0.9819]	0.9405 [0.8902–0.9875]	0.9749 [0.9580–0.9888]	0.9186 [0.8743–0.9545]

Abbreviations: ACC: Accuracy; AUC: Area under the receiver operating characteristic curve; LR: Logistic regression; SEN: Sensitivity; SPE: Specificity; XGBoost: Extreme gradient boosting.

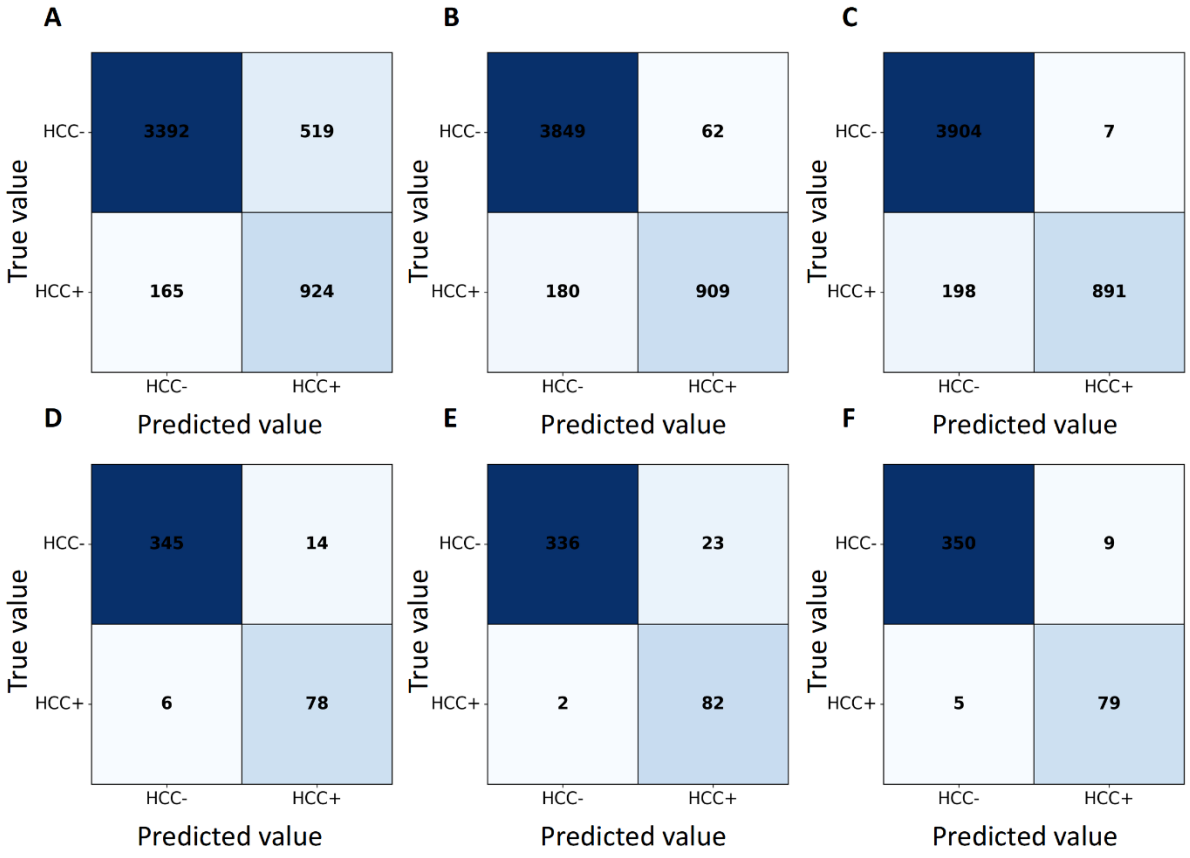


Figure 2. Confusion matrices using baseline machine learning models on different datasets for (A) LR, (B) XGBoost, and (C) CatBoost in Center 1 and (D) LR, (E) XGBoost, and (F) CatBoost in Center 2
Abbreviations: HCC: Hepatocellular carcinoma; LR: Logistic regression; XGBoost: Extreme gradient boosting.

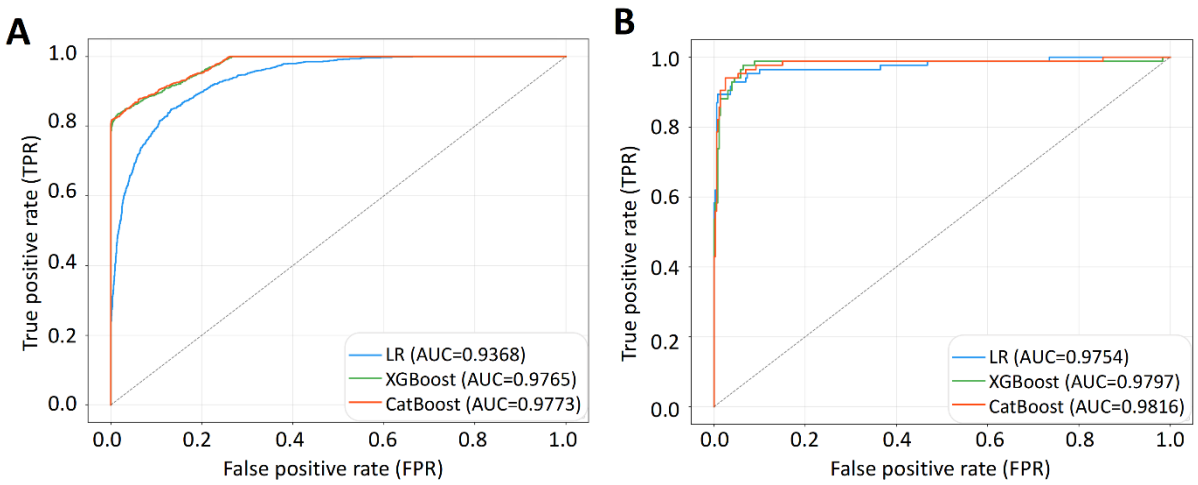


Figure 3. The receiver operating characteristic curves of different datasets for Center 1 (A) and Center 2 (B)
Abbreviations: AUC: Area under the receiver operating characteristic curve; LR: Logistic regression; XGBoost: Extreme gradient boosting.

Table 6. Performance comparison of the five predictive strategies

Center	Strategy	AUC	ACC	SEN	SPE	F1
Center 1	AFP-only	0.6377	0.7286	0.4848	0.7965	0.4376
	CRS	0.9385	0.8622	0.8577	0.8635	0.7305
	LR (Full)	0.9368	0.8632	0.8485	0.8673	0.7299
	XGBoost (Full)	0.9765	0.9516	0.8347	0.9841	0.8825
	CatBoost (Full)	0.9773	0.9590	0.8182	0.9982	0.8968
Center 2	AFP-only	0.8031	0.9187	0.6786	0.9749	0.7600
	CRS	0.9849	0.9865	0.9405	0.9972	0.9634
	LR (Full)	0.9754	0.9549	0.9286	0.9610	0.8864
	XGBoost (Full)	0.9797	0.9436	0.9762	0.9359	0.8677
	CatBoost (Full)	0.9816	0.9684	0.9405	0.9749	0.9186

Abbreviations: ACC: Accuracy; AFP: Alpha-fetoprotein; AUC: Area under the receiver operating characteristic curve; CRS: Clinical risk score; LR: Logistic regression; SEN: Sensitivity; SPE: Specificity; XGBoost: Extreme gradient boosting.

the expected compression of predicted probabilities when applied to a new distribution. Notably, the CRS achieved a Brier score of 0.0990 and a calibration intercept of 3.314, compared to CatBoost (Brier score = 0.1085; intercept = 1.617), further confirming its superior calibration stability and robustness in small-sample external cohorts.

3.7. Evaluation of clinical utility

Decision curve analysis showed that, within a risk threshold range of 5% to 50%, both ML models and the CRS had a consistently higher net benefit than the AFP-only model and the "treat all" strategy (Table 8 and Figure 5). Specifically, at a 10% risk threshold in Center 1, the net benefit for CatBoost was 0.1950, followed by the CRS (0.1821). In contrast, the AFP-only model and the "treat all" strategy yielded lower net benefits of 0.1295 and 0.1318, respectively. At a 30% threshold, CatBoost's benefit (0.1721) remained substantially higher than the

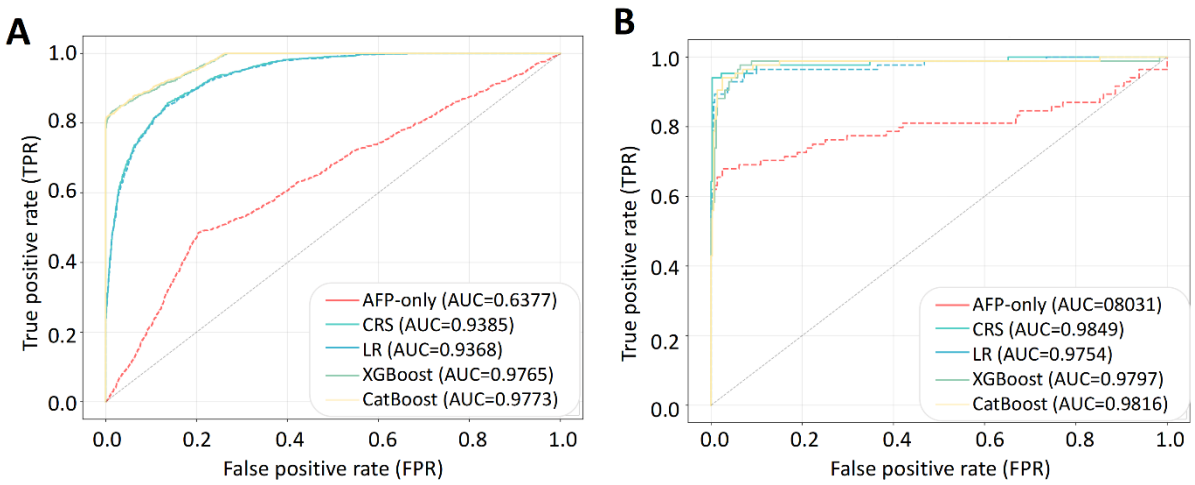


Figure 4. The receiver operating characteristic curves of different predictive strategies for (A) Center 1 and (B) Center 2
Abbreviations: AFP: Alpha-fetoprotein; AUC: Area under the receiver operating characteristic curve; CRS: Clinical risk score; LR: Logistic regression; XGBoost: Extreme gradient boosting.

Table 7. Variations in model performance following the exclusion of alpha-fetoprotein

Center	Model	AUC	AUC without AFP	AUC delta	AUC relative change
Center 1	LR	0.9368	0.9143	0.0225	2.25%
	XGBoost	0.9765	0.9410	0.0355	3.55%
	CatBoost	0.9773	0.9417	0.0356	3.56%
Center 2	LR	0.9754	0.9772	−0.0018	−0.18%
	XGBoost	0.9797	0.9722	0.0075	0.75%
	CatBoost	0.9816	0.9826	−0.0010	−0.10%

Abbreviations: AFP: Alpha-fetoprotein; AUC: Area under the receiver operating characteristic curve; LR: Logistic regression; XGBoost: Extreme gradient boosting.

CRS (0.1416) and the AFP-only model's (0.0234), while the treat all strategy's net benefit had declined to a negative value (−0.1128). In Center 2, at a 10% threshold, the CRS, CatBoost, and XGBoost models achieved identical net benefits, all averaging approximately 0.1750, which was significantly higher than that of the AFP-only model (0.1147). Furthermore, at a 50% threshold, the CRS's net benefit remained at 0.1715, whereas the treat-all strategy exhibited a substantial negative net benefit of −0.6370. Collectively, these findings confirm that predictive models based on comprehensive clinical information provide a consistent positive net benefit for clinical decision-making across a broad range of risk thresholds, demonstrating clear superiority over conventional AFP-based screening strategies.

3.8. Evaluation of model incremental value

In Center 1, the CRS showed a greater ability to predict outcomes than the AFP-only model (Table 9, Figure 6, and Table S2). This was demonstrated by a categorical NRI of 1.1271 (95% CI: 0.9842–1.2650; event NRI = 0.5096, non-event NRI = 0.6175) and a cfNRI of 1.6543 (95% CI: 1.5120–1.7915), alongside an IDI of 0.5099 (95% CI: 0.4512–0.5685; $Z = 51.66$, $p < 0.001$). Furthermore, the ML models produced significant improvements compared to the CRS. XGBoost versus CRS produced a categorical NRI of 0.1470 (95% CI: 0.0580–0.2350; cfNRI = 0.2105, 95% CI: 0.1134–0.3052) and an IDI of 0.1843 (95% CI: 0.1250–0.2420; $Z = 14.14$, $p < 0.001$), while CatBoost versus CRS produced a categorical NRI of 0.1354 (95% CI: 0.0412–

Table 8. Decision curve analysis results of the different predictive strategies

Center	Threshold	AFP-only	CRS	LR	XGBoost	CatBoost	Treat all
Center 1	0.05	0.1770	0.1974	0.1975	0.2059	0.1974	0.1770
	0.1	0.1295	0.1821	0.1813	0.1923	0.1950	0.1318
	0.15	0.0825	0.1691	0.1690	0.1803	0.1821	0.0813
	0.2	0.0515	0.1591	0.1579	0.1741	0.1674	0.0246
	0.25	0.0389	0.1510	0.1501	0.1719	0.1655	−0.0396
	0.3	0.0234	0.1416	0.1407	0.1710	0.1721	−0.1128
	0.4	−0.0014	0.1263	0.1249	0.1731	0.1762	−0.2952
	0.5	0.0000	0.1120	0.1080	0.1721	0.1760	−0.5800
Center 2	0.05	0.1309	0.1805	0.1705	0.1817	0.1779	0.1473
	0.1	0.1147	0.1754	0.1724	0.1754	0.1746	0.1005
	0.15	0.1123	0.1759	0.1678	0.1720	0.1740	0.0482
	0.2	0.1186	0.1756	0.1632	0.1615	0.1671	−0.0106
	0.25	0.1182	0.1746	0.1663	0.1618	0.1633	−0.0770
	0.3	0.1171	0.1736	0.1606	0.1578	0.1565	−0.1529
	0.4	0.1122	0.1768	0.1551	0.1476	0.1543	−0.3418
	0.5	0.1106	0.1715	0.1489	0.1375	0.1465	−0.6370

Abbreviations: AFP: Alpha-fetoprotein; CRS: Clinical risk score; LR: Logistic regression; XGBoost: Extreme gradient boosting.

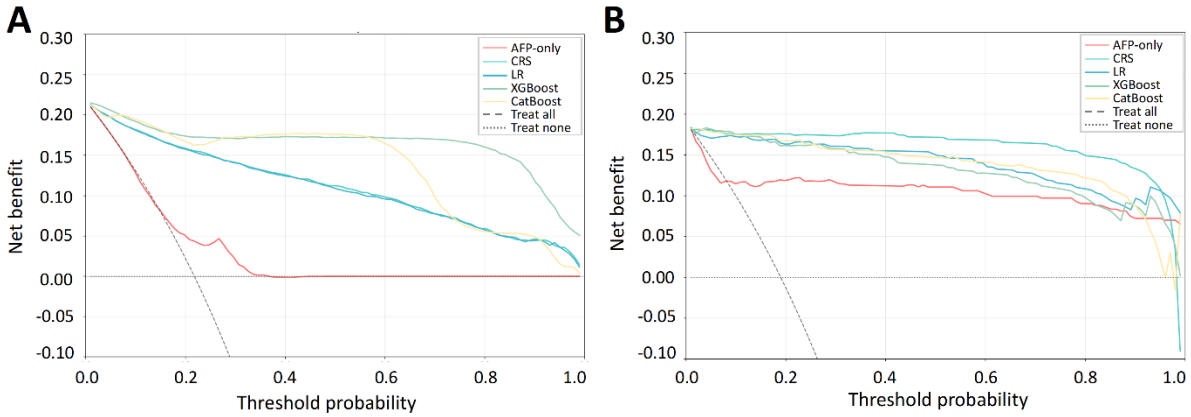


Figure 5. The DCA curves of different predictive strategies for (A) Center 1 and (B) Center 2
Abbreviations: AFP: Alpha-fetoprotein; CRS: Clinical risk score; LR: Logistic regression; XGBoost: Extreme gradient boosting.

0.2285; cfNRI = 0.1950, 95% CI: 0.0980–0.2910) and an IDI of 0.0279 (95% CI: 0.0150–0.0610; $Z = 2.36$, $p = 0.018$).

In Center 2, the CRS also showed statistically significant improvements over the AFP-only model (Table 9, Figure 6, and Table S2), with a categorical NRI of 0.5789 (95% CI: 0.3845–0.7620; event NRI = 0.2976, non-event NRI = 0.2813) and a cfNRI of 0.8932 (95% CI: 0.6815–1.0940), alongside an IDI of 0.3839 (95% CI: 0.2510–0.5150; $Z = 7.43$, $p < 0.001$). However, the incremental gains of the

MLs over the CRS were negative: XGBoost versus CRS resulted in a categorical NRI of −0.0651 (95% CI: −0.1540 to 0.0230; cfNRI = -0.0815, 95% CI: −0.1750 to 0.0120) and an IDI of −0.1656 (95% CI: −0.2510 to −0.0810; Z = −3.84, $p < 0.001$), and CatBoost versus CRS resulted in a categorical NRI of −0.0638 (95% CI: −0.1524 to 0.0245; cfNRI = −0.0912, 95% CI: −0.1856 to 0.0045) and an IDI of −0.1027 (95% CI: −0.1850 to −0.0210; $Z = -2.38$, $p = 0.017$). These findings, robustly supported by the CIs and the threshold-independent cfNRI, suggest that in the small-sample cohort of Center 2, the parsimonious CRS may be non-inferior to, or even superior to, complex ML models in terms of risk reclassification.

4. Discussion

In this study, we developed and validated a comprehensive risk stratification system for HCC by integrating a CRS with ML, using multicenter data. Through a systematic evaluation of predictive performance, clinical net benefit, and risk reclassification, we showed that this combined approach substantially outperforms traditional AFP-based screening. Crucially, our external real-world validation revealed the nuanced trade-offs between model complexity and stability in various clinical scenarios.

Regarding the 12 routine clinical variables incorporated into our framework, their inclusion was pre-specified based on established HCC risk factors, ensuring a holistic clinical profile. Beyond the prominent viral (hepatitis B/C) and metabolic (diabetes) drivers, other variables hold distinct clinical relevance. Demographic factors (age, sex) are foundational risk stratifiers, while lifestyle factors (smoking, alcohol consumption, physical activity) directly influence hepatic inflammation and metabolic homeostasis. Medical history variables (cirrhosis, family history of cancer) capture long-term cumulative risks. Although variables such as gender and physical activity level did not reach statistical significance in the large Center 1 cohort, and factors like body mass index or family history showed variable significance across cohorts, retaining them is clinically justified. Excluding variables solely based on statistical significance in a single cohort can introduce omitted variable bias and reduce the model's generalizability to diverse real-world populations where these factors may interact differently. Our comprehensive inclusion aligns with the multifactorial nature of HCC pathogenesis.

Our multivariable analysis revealed a changing landscape of HCC risk factors. In the large-scale Center 1 cohort, metabolic and viral factors (particularly diabetes mellitus, hepatitis C, and hepatitis B) emerged as the most

Table 9. Summary of net reclassification improvement and integrated discrimination improvement results

Center	Comparison	NRI total	IDI	NRI event	NRI non-event	*z*-score	*p*-value
Center 1	CRS vs. AFP	1.1271	0.5099	0.5096	0.6175	51.6601	<0.001
	XGBoost vs. AFP	1.2906	0.6942	0.5601	0.7305	74.4453	<0.001
	CatBoost vs. AFP	1.2726	0.5379	0.5822	0.6904	70.9622	<0.001
	XGBoost vs. CRS	0.1470	0.1843	0.0588	0.0882	14.1422	<0.001
	CatBoost vs. CRS	0.1354	0.0279	0.0781	0.0573	2.3595	0.0183
	CatBoost vs. XGBoost	−0.0160	−0.1563	0.0257	−0.0417	−13.7215	<0.001
Center 2	CRS vs. AFP	0.5789	0.3839	0.2976	0.2813	7.4286	<0.001
	XGBoost vs. AFP	0.5334	0.2182	0.2381	0.2953	3.9048	<0.001
	CatBoost vs. AFP	0.5382	0.2812	0.2262	0.3120	5.0329	<0.001
	XGBoost vs. CRS	−0.0651	−0.1656	−0.0595	−0.0056	−3.8416	<0.001
	CatBoost vs. CRS	−0.0638	−0.1027	−0.0833	0.0195	−2.3824	<0.05
	CatBoost vs. XGBoost	−0.0043	0.0630	−0.0238	0.0195	1.3100	0.1902

Abbreviations: AFP: Alpha-fetoprotein; CRS: Clinical risk score; IDI: Integrated discrimination improvement; NRI: Net reclassification improvement; XGBoost: Extreme gradient boosting.

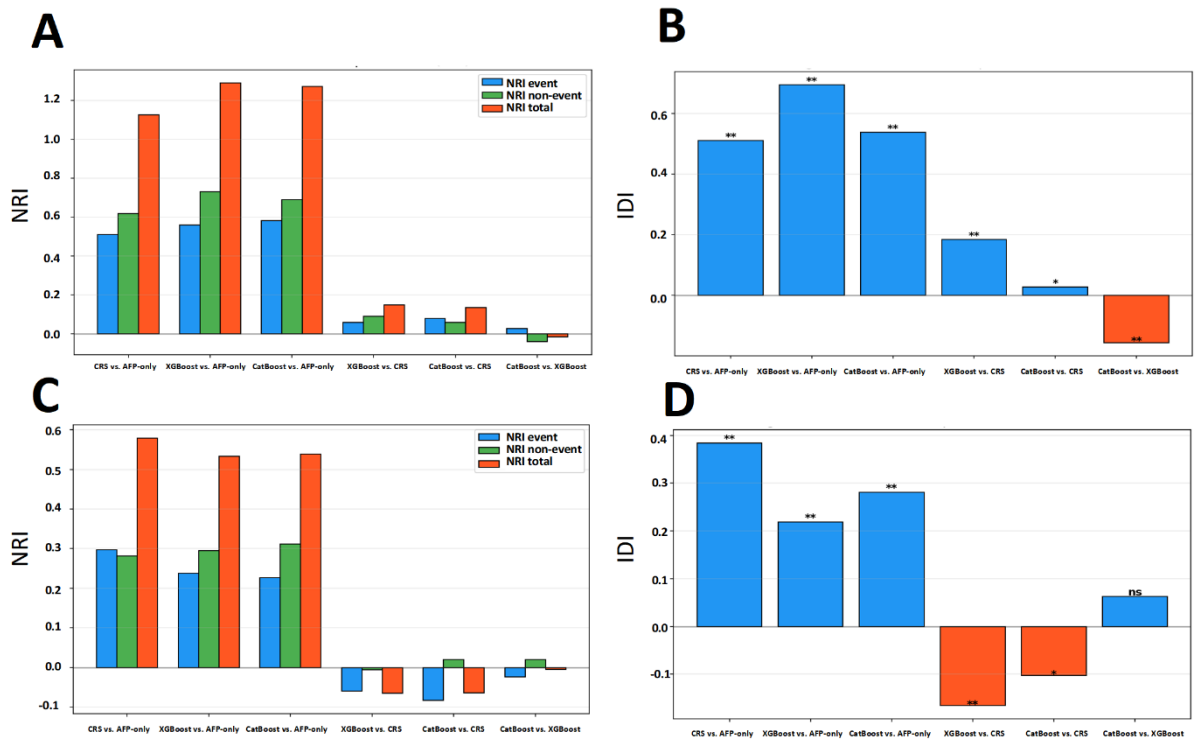


Figure 6. The bar plot of model incremental value with (A) NRI and (B) IDI in Center 1; (C) NRI and (D) IDI in Center 2. *$p < 0.05$, **$p < 0.001$ compared to the baseline model

Abbreviations: AFP: Alpha-fetoprotein; CRS: Clinical risk score; IDI: Integrated discrimination improvement; NRI: Net reclassification improvement; XGBoost: Extreme gradient boosting.

prominent independent predictors. The strong association with diabetes mellitus is consistent with the growing body of evidence linking metabolic dysfunction-associated steatotic liver disease to HCC pathogenesis.[25] In contrast, the risk factor profile in the smaller Center 2 cohort exhibited anomalous patterns, such as extremely high odds ratios for cirrhosis (OR = 4.69×10^9) and apparent protective effects for hepatitis B and alcohol consumption. Direct inspection of the baseline data (Table 1) confirms quasi-complete separation: for example, in the Center 2 HCC- group, the proportion of patients with AFP >25 ng/mL was 0% (0/359), and hepatitis C was an invariant with zero prevalence across the entire cohort. To quantitatively link these anomalies to sample size, we calculated the events per variable (EPV) ratio. While Center 1 had an EPV of ~90.7 (1,089 events/12 variables), ensuring stable maximum likelihood estimates, Center 2 had an EPV of only 7.0 (84 events/12 variables), falling below the critical threshold of 10. This severe sparse-data bias mathematically explains the coefficient explosion. We deliberately did not apply penalized logistic regression techniques (e.g., Firth correction or least absolute shrinkage and selection operator) to Center 2, as this cohort was strictly reserved for independent external validation of the pre-trained models; re-fitting the regression would violate the validation design. Instead, these anomalous Center 2 results serve to expose the vulnerabilities of traditional statistical modeling in small, skewed real-world datasets, thereby highlighting the superior robustness of our pre-trained CRS and ML models, which inherently mitigate such separation issues.

A key finding of our study is the empirical demonstration of the bias-variance trade-off in clinical predictive modeling. Although the ML models performed better in discriminating between outcomes in the large-scale cohort by capturing complex, nonlinear interactions among the clinical variables, this advantage was less pronounced in the smaller Center 2 cohort. Unlike many existing ML studies that solely pursue marginal AUC improvements using complex black-box algorithms, our study empirically demonstrates that a parsimonious linear scoring system (CRS) maintains robust generalizability in small real-world cohorts, whereas complex architectures are prone to overfitting noise. From a clinical perspective, this necessitates a tiered deployment strategy: the CRS is highly

recommended for rapid, interpretable initial screening in primary care or in settings with limited resources, while the ML models should be reserved for precision risk stratification in large-scale, well-resourced clinical centers.

Furthermore, our AFP-independent sensitivity analysis addresses a critical blind spot in current HCC surveillance. Conventional screening relies heavily on AFP, which has low sensitivity, particularly in the early stages of AFP-negative HCC. Our results showed that excluding AFP led to only a slight decrease in predictive performance, and the models retained excellent discriminative capacity, particularly within the AFP-normal subgroup. This suggests that integrating comprehensive clinical features, such as liver cirrhosis history, metabolic profiles and demographic factors, can effectively compensate for the absence of tumor markers. This finding provides a robust, evidence-based alternative for risk stratifying the substantial population of AFP-negative individuals at high risk, who are frequently missed by standard protocols.

Finally, the clinical utility of our integrated system was robustly confirmed through decision curve and reclassification analyses. The models consistently produced positive net benefits across a wide range of clinically relevant threshold probabilities. They significantly outperformed the "treat all" and "treat none" strategies, particularly at higher decision thresholds where precise risk stratification is crucial. However, the smaller cohort's negative incremental gains of ML over CRS further reinforce the principle of parsimony. This shows that the development of sophisticated algorithms must be balanced against the availability of clinical data, to ensure that the chosen predictive tool is optimally matched to the epidemiological and resource context of the target population.

Based on the empirical findings of this study and the nuanced trade-offs observed between model complexity and stability, we propose a tiered clinical deployment strategy for the integrated risk stratification system:

(i) For initial screening: In primary care or settings with limited resources, the CRS is highly recommended for rapidly screening high-risk populations. Its 0–100-point scale offers high interpretability and robust performance without the need for complex computational infrastructure.
(ii) For precision risk assessment: In well-resourced clinical centers where precise, personalized risk stratification is required, the ML module should be used to capture complex, nonlinear interactions among the 12 clinical variables.
(iii) For AFP-negative populations: The AFP-excluded model remains highly effective in identifying HCC risk. This strongly supports the routine incorporation of comprehensive clinical information into risk evaluations, overcoming the critical blind spots of conventional AFP-based surveillance.
(iv) For clinical decision-making: Optimal risk thresholds for clinical intervention should be established dynamically based on DCA results and carefully tailored to local healthcare resource availability, patient preferences and specific clinical contexts.

This study has several limitations that warrant consideration. Firstly, the retrospective study design is inherently subject to selection and information biases. Secondly, the sample size in Center 2 was relatively small (n = 443), and certain variables were distributed unevenly, which may affect the stability and generalizability of the models within this specific cohort. Thirdly, demographic and clinical differences between the patient populations of the two centers may affect the generalizability of the models across centers. Fourthly, while our model effectively integrates routine clinical and metabolic surrogates, it does not incorporate advanced molecular markers (e.g., circulating tumor DNA) or detailed virological metrics (e.g., quantitative hepatitis B virus DNA load, platelet counts) required for established scores like PAGE-B or REACH-B, as these were not uniformly available in our cohorts. This represents a limitation of the current study; however, it also highlights the practical value of our framework in providing an accessible screening layer for populations where such advanced testing is not feasible. Fifthly, this study did not incorporate radiomic features or genomic information. Future research should explore multimodal fusion models to enhance predictive accuracy further. Finally, temporal external validation was not performed, and the temporal stability of the models requires further evaluation in future prospective studies.

5. Conclusion

In conclusion, this study developed and validated an integrated HCC risk stratification system combining clinical risk scoring and ML based on 12 routine clinical variables. By overcoming the limitations of traditional AFP screening, particularly in AFP-negative populations and balancing model complexity with real-world stability, this framework provides a practical, tiered decision-support tool for precision screening and individualized risk management in high-risk populations.

Acknowledgments

None.

Funding

This work was supported by the Science and Technology Innovation Special Fund Project of Shanghai Baoshan District Science and Technology Committee (2023-E-17) and the Excellent Youth Program of Shanghai Baoshan District Health Committee (BSWSYC-2023-04).

Conflict of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization: Yingying Luo
Data curation: Yingying Luo, Zengbei Yuan, Yingkang Lin, Rui Li
Formal analysis: XiaoLei Huang, Feiyang Pu
Funding acquisition: Lihong Liu, Zhihua Ni
Investigation: Yingying Luo, Zengbei Yuan, Yingkang Lin
Methodology: Lihong Liu, Zhihua Ni
Supervision: Lihong Liu, Zhihua Ni
Validation: XiaoLei Huang, Feiyang Pu
Visualization: Rui Li
Writing–original draft: Yingying Luo, Zengbei Yuan, Yingkang Lin
Writing–review & editing: Lihong Liu, Zhihua Ni

Ethics approval and consent to participate

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Ethics Review Board of Shanghai Baoshan Hospital of Integrated Traditional Chinese and Western Medicine (#202607-01). All patients were exempted from signing the informed consent form.

Consent for publication

Not applicable.

Availability of data

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

Further disclosure

In accordance with the journal's policy, we confirm that no artificial intelligence (AI) writing tools or generative AI technologies were used in the drafting, editing, or preparation of this manuscript. All text and analysis are the original work of the authors.

References

1. Lee DD, Croome KP, Musto KR, *et al*. Liver transplantation for intrahepatic cholangiocarcinoma. *Liver Transpl*. 2018;24(5):634-644.

   doi: 10.1002/lt.25052

2. Sung H, Filho AM, Laversanne M, *et al*. Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 34 cancers in 186 countries. *CA Cancer J Clin*. 2026;76(4).

   doi: 10.3322/caac.70090

3. Han B, Zheng R, Zeng H, *et al*. Cancer incidence and mortality in China, 2022. *J Natl Cancer Cent*. 2024;4(1):47-53.

   doi: 10.1016/j.jncc.2024.01.006

4. Zhou J, Sun H, Wang Z, *et al*. Guidelines for the diagnosis and treatment of primary liver cancer (2022 edition). *Liver Cancer*. 2023;12(5):405-444.

   doi: 10.1159/000530495

5. de Lope CR, Tremosini S, Forner A, Reig M, Bruix J. Management of HCC. *J Hepatol*. 2012;56(suppl 1):S75-S87.

   doi: 10.1016/S0168-8278(12)60009-9

6. Tzartzeva K, Obi J, Rich NE, *et al*. Surveillance imaging and alpha fetoprotein for early detection of hepatocellular carcinoma in patients with cirrhosis: a meta-analysis. *Gastroenterology*. 2018;154(6):1706-1718.e1.

   doi: 10.1053/j.gastro.2018.01.064

7. Hanif H, Ali MJ, Susheela AT, *et al*. Update on the applications and limitations of alpha-fetoprotein for hepatocellular carcinoma. *World J Gastroenterol*. 2022;28(2):216-229.

   doi: 10.3748/wjg.v28.i2.216

8. Turshudzhyan A, Wu GY. Persistently rising alpha-fetoprotein in the diagnosis of hepatocellular carcinoma: a review. *J Clin Transl Hepatol*. 2022;10(1):159-163.

   doi: 10.14218/JCTH.2021.00176

9. Galle PR, Forner A, Llovet JM, *et al*. EASL Clinical Practice Guidelines: management of hepatocellular carcinoma. *J Hepatol*. 2018;69(1):182-236.

   doi: 10.1016/j.jhep.2018.03.019

10. Osho A, Rich NE, Singal AG. Role of imaging in management of hepatocellular carcinoma: surveillance, diagnosis, and treatment response. *Hepatoma Res*. 2020;6:55.

doi: 10.20517/2394-5079.2020.42

11. Shan S, Jia J. The clinical management of hepatocellular carcinoma in China: progress and challenges. *Clin Mol Hepatol*. 2023;29(2):339-341.

doi: 10.3350/cmh.2023.0077

12. Chang YP, Chen YC, Liu CH. Risk scores for stratifying hepatocellular carcinoma and optimizing surveillance strategies. *Cancers (Basel)*. 2026;18(1):158.

doi: 10.3390/cancers18010158

13. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56.

doi: 10.1038/s41591-018-0300-7

14. Zou ZM, Chang DH, Liu H, Xiao YD. Current updates in machine learning in the prediction of therapeutic outcome of hepatocellular carcinoma: what should we know? *Insights Imaging*. 2021;12(1).

doi: 10.1186/s13244-021-00977-9

15. Liu W, Wei R, Chen J, *et al*. Prognosis prediction and risk stratification of transarterial chemoembolization or intraarterial chemotherapy for unresectable hepatocellular carcinoma based on machine learning. *Eur Radiol*. 2024;34(8):5094-5107.

doi: 10.1007/s00330-024-10581-2

16. Zhang Y, Tong S, Yang J, *et al*. Explainable machine learning model for predicting the transarterial chemoembolization response and subtypes of hepatocellular carcinoma patients. *BMC Gastroenterol*. 2025;25(1).

doi: 10.1186/s12876-025-04105-5

17. Xu Y, Zhang B, Zhou F, *et al*. Development of machine learning-based personalized predictive models for risk evaluation of hepatocellular carcinoma in hepatitis B virus-related cirrhosis patients with low levels of serum alpha-fetoprotein. *Ann Hepatol*. 2024;29(6):101540.

doi: 10.1016/j.aohep.2024.101540

18. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63.

doi: 10.7326/M14-0697

19. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. *BMC Med*. 2023;21(1).

doi: 10.1186/s12916-023-02779-w

20. Imperiale TF, Monahan PO. Risk stratification strategies for colorectal cancer screening: from logistic regression to artificial intelligence. *Gastrointest Endosc Clin N Am*. 2020;30(3):423-440.

doi: 10.1016/j.giec.2020.02.004

21. Ghuriani V, Wassan JT, Tripathi P, Chauhan A. XGB-BIF: an XGBoost-driven biomarker identification framework for detecting cancer using human genomic data. *Int J Mol Sci*. 2025;26(12):5590.

doi: 10.3390/ijms26125590

22. Wu J, Tao G, Xie S, Yang H, Qi F, Bao N, *et al*. Prediction of three-year all-cause mortality in patients with heart failure and atrial fibrillation using the CatBoost model. *BMC Cardiovasc Disord*. 2025;25(1).

doi: 10.1186/s12872-025-04928-w

23. Huang C, Fang M, Feng H, *et al*. N-glycan fingerprint predicts alpha-fetoprotein negative hepatocellular carcinoma: a large-scale multicenter study. *Int J Cancer*. 2021;149(3):717-727.

doi: 10.1002/ijc.33564

24. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):157-172.

doi: 10.1002/sim.2929

25. Kalligeros M, Henry L, Younossi ZM. Metabolic dysfunction-associated steatotic liver disease and its link to cancer. *Metabolism*. 2024;160:156004.

doi: 10.1016/j.metabol.2024.156004