scientific reports

Check for updates

OPEN

A nomogram and random forest model for predicting liver metastasis in patients with early-onset colorectal cancer

Xingzhi Han[1], Xueying Bai[3], Qun Zhang[2]✉ & Xiaoping Qian[1]✉

The incidence of colorectal cancer (CRC) in individuals under the age of 50 has increased. Liver metastasis (LM) is the most common metastasis in CRC patients and is associated with a poor prognosis. This study aimed to use public databases to identify the risk factors for LM in early-onset colorectal cancer (EOCRC) patients and develop a nomogram to quantify the risk of LM. We retrospectively collected data of EOCRC patients diagnosed from 2010 to 2015 in the Surveillance, Epidemiology, and End Results (SEER) database. Univariate and multivariate logistic analysis were used to screen and validate the risk factors for LM in EOCRC patients, and a nomogram was established based on these factors. Calibration curve, area under the receiver operating curve (AUC), and decision curve analysis (DCA) were developed to evaluate the accuracy of the model. A total of 2567 EOCRC patients were included and randomly divided into a training set ($n=1797$) and a validation set ($n=770$) at a ratio of 7:3. Univariate and multivariate analyses showed that N stage, pretreatment CEA, bone metastasis, and lung metastasis were independent risk factors. The AUCs of the training set and validation set were 0.7958 and 0.7653, respectively, and the calibration curve also demonstrated good accuracy and predictive ability. DCA indicated that it was more clinically relevant than the traditional TN staging. We constructed a Random Forest model, and calculated the SHapley Additive exPlanations (SHAP) values to determine variables importance and visualize the results. We developed a nomogram to predict the risk of LM in EOCRC patients, and the model was internally validated with good accuracy and reliability. It can assist doctors in risk assessment and clinical decision-making.

Keywords Colorectal cancer, Liver metastasis, Nomogram, SEER

Colorectal cancer (CRC) is the third most common tumor and ranks second in cancer-related deaths worldwide[1]. Due to the implementation of CRC prevention and screening programs for the 50–75 age group, the overall incidence of CRC has decreased by approximately 2–3% annually[2,3]. In sharp contrast, the incidence of CRC in individuals under 50 has been increasing[4,5]. Early-onset colorectal cancer (EOCRC) is defined as CRC diagnosed in individuals under 50. From 2000 to 2013,the incidence of EOCRC increased by 22%[2,6], accounting for 10% of all newly diagnosed CRC cases[7]. It is projected to become the leading cause of cancer death among Americans aged 20 to 49 by 2030[4]. Due to delayed diagnosis, potential differences in tumor biology, or other as-yet-undetermined factors, EOCRC patients are more likely to have metastasis at the time of diagnosis compared to those with late-onset CRC[8,9]. EOCRC represents a significant cancer burden among young people.

Among all CRC patients, approximately 20% have metastasis at the initial diagnosis, and 50% will develop metastasis later on[10]. Distant metastasis is the main cause of death for CRC patients, and liver metastasis is the most common route of late-stage distant spread. 20–25% of CRC patients exhibit liver metastasis at the time of diagnosis, and up to 50% may develop liver metastasis after resection of the primary tumor. Radical resection and chemotherapy are the main treatment options for colorectal cancer liver metastasis (CRLM) patients. Early detection of CRC liver metastasis followed by radical resection can lead to a 5-year survival

[1]Department of Comprehensive Cancer Centre, Affiliated Hospital of Medical School, Nanjing Drum Tower Hospital, Nanjing University, 321 Zhong Shan Rd, Nanjing 210008, People's Republic of China. [2]Comprehensive Cancer Centre of Nanjing Drum Tower Hospital, Medical School of Nanjing University, Clinical Cancer Institute of Nanjing University, 321 Zhong Shan Rd, Nanjing 210008, People's Republic of China. [3]Nanjing Drum Tower Hospital Clinical College of Traditional Chinese and Western Medicine, Nanjing University of Chinese Medicine, Nanjing, China. ✉email: qunzhangnora@outlook.com; xiaopingqian@nju.edu.cn

rate of 30–57%, in contrast, patients who are not eligible for resection have a 5-year survival rate of less than 5%[11,12]. However, only 10%–20% of patients are eligible for surgical treatment, and the 5-year survival rate is less than 50%. Traditional chemotherapy affects all rapidly dividing cells, leading to significant toxicity and frequent drug resistance. Paradoxically, chemotherapy may even promote CRC metastasis[13]. Therefore, early detection of CRC-related liver metastasis and targeted intervention are crucial for improving patient prognosis. However, there is currently a lack of effective methods in clinical practice to predict these liver metastases early. Routine examinations for CRC patients include contrast-enhanced CT to monitor distant metastasis[14]. EOCRC patients are relatively young, and the follow-up and monitoring period is long. However, multiple exposures to ionizing radiation carry the risk of secondary malignancies. MRI is more suitable for patients with high suspicion of liver metastasis on CT but cannot be confirmed, but it has the problems of long examination time and high economic cost.

Nomograms have been accepted as visual predictive tools for statistical regression models, which can help clinicians make accurate decisions and promote the development of precision medicine[15]. Machine learning algorithms have emerged as powerful tools that are increasingly utilized in cancer research[16,17]. Random Forest (RF) inherently possesses the capability to capture nonlinear relationships among variables and model feature interactions by generating multiple decision trees and aggregating their individual predictions, thereby establishing itself as a prominent ensemble learning approach. SHapley Additive Explanations (SHAP) is a model-agnostic interpretability method designed to elucidate the outcomes generated by machine learning models[17,18]. By integrating these two methodologies, it is possible not only to assess the reliability of model predictions but also to obtain valuable insights into the influence of individual features on the prediction process.

In recent years, based on bioinformatics methods and open-access cancer patient data, we have been able to explore independent risk factors for tumor metastasis. The publicly accessible SEER database includes data on cancer patients from 18 registration sites, covering approximately 28% of the US cancer patient population[19]. This work aims to construct a reliable nomogram for predicting liver metastasis in early-onset colorectal cancer patients, thereby assisting clinicians in more precisely tailoring treatment plans to reduce the metastasis rate and improve survival rates.

Materials and methods

Study population

The data used in this study were obtained from SEER Stat (version 8.4.3). Information on CRC patients under the age of 50 from 2010 to 2015 was extracted from the SEER database. Since all SEER database information is publicly available and patient information is anonymized, ethical approval and informed consent from patients are not required. Inclusion criteria: (1) Patients under the age of 50; (2) Pathologically diagnosed with CRC; (3) CRC as the first primary tumor; (4) Sufficient variable information including demographics and clinical pathology. Exclusion criteria: (1) Survival time less than one month; (2) Unknown marital status; (3) Unknown bone/brain/liver/lung metastasis status; (4) Pretreatment CEA unknown; (5) Grade unknown; (6) Tx/Nx; (7) Not the first primary tumor; (8) Tumor size unknown. After screening, 2567 eligible patients were included in the study. The flowchart of patient inclusion and exclusion is shown in Fig. 1.

Data collection

Variables obtained in the selected cohort included: clinical characteristics (age at diagnosis, sex, ethnicity, marital status, tumor location, grade, T/N stage, histology, distant metastasis, tumor size and pretreatment CEA), and treatment-related information (surgery, radiotherapy, and chemotherapy). Patients were classified as black, white or other (Alaskan native/American Indian, Pacific/Asian Islander), tumor sites were classified as left colon, right colon and rectum, tumor size was classified as <5 cm and ≥ 5 cm. Grade is divided into Well/moderately and Poorly/undifferentiated. Histology is divided into Adenocarcinoma and Non-adenocarcinoma. CEA was divided into negative/norma and positive/elevated.

Statistical analysis

Statistical analysis in this study were performed using SPSS26.0 and R software (version 4.5.1). $^{*}P < 0.05$ indicated statistical significance, and the nomogram was established by "rms" package in R software. The specific process of constructing the predictive model and nomogram is as follows: First, all patients were randomly divided into a training set and a validation set at a ratio of 7:3, and baseline information was compared using Chi-square test and Mann-Whitney' U test. Training sets are used for nomogram development and validation sets are used for validation.

We then performed a univariate logistic analysis to identify factors associated with liver metastasis. Variables with *P*-values less than 0.05 in univariate analysis were included in multivariate logistic regression to determine independent risk factors for LM in patients with EOCRC. It is imperative to note that due to temporal circularity between surgical/radiation/chemtherapy and LM diagnosis, treatment variables were excluded from the predictive model.

In addition, We also plotted the ROC curve and calculated the area under the curve (AUC) to estimate the discrimination of the model. The calibration and decision curve analysis (DCA) were developed to further validate the model. In addition, ROC curves for all the independent variables are generated to compare the AUC of the nomogram with all the independent variables. The RF model was constructed using the "randomForest" package, and the SHAP values were calculated using the "kernelshap" package. SHAP was employed to explain the impact of each feature on the prediction output of the model, and the relative importance of the features was visualized.

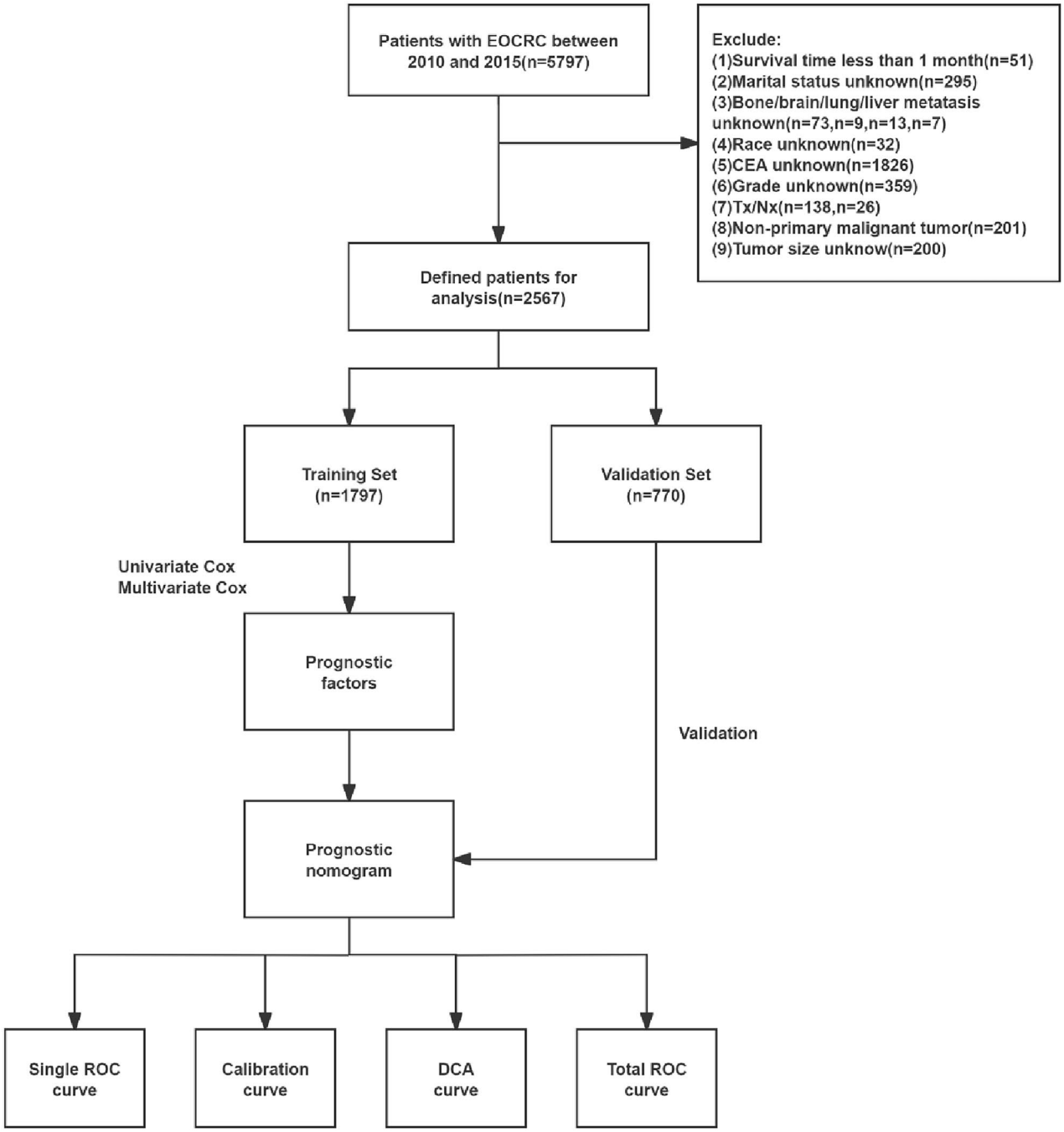

Fig. 1. Flowchart for inclusion and exclusion of early-onset colorectal cancer.

Results

Clinical features

A total of 2567 EOCRC patients between 2010 and 2015 were obtained from the SEER database. In the entire cohort, all patients were under 50 years of age, 15.7% developed liver metastases, 1368 (53%) were under 40 years of age, 72% were white, and 61% were married. The most common tumor location was the left colon (1130,44.0%), most of the patients received surgical treatment for the primary tumor (2447,95.3%), 708 (27.6%) patients received radiotherapy, and 41.6% of patients had elevated CEA before treatment. Patients were randomly assigned to a training set ($N=1797$) and a validation set ($N=770$). There was no significant difference in clinical features between the two groups ($P>0.05$). The results are shown in Table 1.

Risk factors of LM in EOCRC patients

Univariate logistic regression was used to analyze and screen risk factors related to liver metastasis in training set. The results showed grade, T stage, N stage, CEA, bone metastasis and lung metastasis are related to liver

Variable	Total($n=2567$)	Training($n=1797$)	Validation($n=770$)	*P*-value
Race				0.465
White	1842(71.8%)	1300(72.4%)	542(70.4%)	
Black	282(11.0%)	189(10.5%)	93(12.1%)	
Other	443(17.2%)	308(17.1%)	135(17.5%)	
Age				0.139
<45	1368(53.5%)	940(52.3%)	428(55.6%)	
≥45	1199(46.7%)	857(47.7%)	342(44.4%)	
Sex				0.071
Male	1409(54.9%)	965(53.7%)	444(57.7%)	
Female	1159(45.1%)	832(46.3%)	326(42.3%)	
Marital status				0.762
Married	1560(60.8%)	1097(61.0%)	464(60.3%)	
Unmarried	1007(39.2%)	700(39.0%)	306(39.7%)	
Tumor location				0.233
Right colon	692(27.0%)	502(27.9%)	190(24.7%)	
Left colon	1130(44.0%)	780(43.4%)	350(45.4%)	
Rectum	745(29.0%)	515(28.7%)	230(29.8%)	
Grade				0.543
Well/moderately	2070(80.6%)	1443(80.3%)	627(81.4%)	
Poorly/undifferentiated	497(19.4%)	354(19.7%)	143(18.6%)	
Histology				0.245
Adenocarcinoma	2363(92.1%)	1662(92.5%)	701(91.0%)	
Non-adenocarcinoma	204(7.9%)	135(7.5%)	69(9.0%)	
T stage				0.859
T1-T2	457(17.8%)	322(17.9%)	135(17.5%)	
T3-T4	2110(82.2%)	1475(82.1%)	635(82.5%)	
N stage				0.725
N0	975(38.0%)	687(38.2%)	288(37.4%)	
N1-N2	1592(62.0%)	1110(61.8%)	482(62.6%)	
Radiotherapy				0.555
Yes	708(27.6%)	489(27.2%)	219(28.4%)	
No/unknown	1859(72.4%)	1308(72.8%)	551(71.6%)	
Chemotherapy				0.261
Yes	1914(74.6%)	1328(73.9%)	586(76.1%)	
No/unknown	653(25.4%)	469(26.1%)	184(23.9%)	
CEA				0.892
negative/normal	1500(58.4%)	1048(58.3%)	452(58.7%)	
positive/elevated	1067(41.6%)	749(41.7%)	318(41.3%)	
Bone metastasis				0.14
No	2553(99.5%)	1790(99.6%)	763(99.1%)	
Yes	14(0.5%)	7(0.4%)	7(0.9%)	
Brain metastasis				0.588
No	2563(99.8%)	1795(99.9%)	768(99.7%)	
Yes	4(0.2%)	2(0.1%)	2(0.3%)	
Liver metastasis				0.669
No	2164(84.3%)	1519(84.5%)	645(83.8%)	
Yes	403(15.7%)	278(15.5%)	125(16.2%)	
Lung metastasis				0.559
No	2477(96.5%)	1731(96.3%)	746(96.9%)	
Yes	90(3.5%)	66(3.7%)	24(3.1%)	
Tumor Size				0.701
<5 cm	1310(51.0%)	922(51.3%)	388(50.4%)	
≥5 cm	1257(49.0%)	875(48.7%)	382(49.6%)	
Surgery of primary site				0.476
No	120(4.7%)	88(4.9%)	32(4.2%)	
Yes	2447(95.3%)	1709(95.1%)	738(95.8%)	

Table 1. Clinicopathological characteristic of 2567 EOCRC patients.

metastasis in EOCRC patients. Subsequently, multivariate logistic regression model was used to analyze and screen independent risk factors related to LM, and the results showed that N stage, CEA, bone metastasis and lung metastasis are independent risk factors for LM. However, grade and T stage were not independent risk factors for LM, and the results were shown in Table 2.

Development and validation of the predictive nomogram

We constructed a nomogram based on a multivariate logistic regression analysis to predict liver metastases in patients with EOCRC (Fig. 2), showing that each variable in the nomogram was assigned a score from 0 to 100 reflecting their contribution to the predictive model (Table 3). The nomogram showed that CEA, bone metastasis and lung metastasis are the most critical factors affecting LM. N stage also has a certain impact on LM. The ROC curve showed that the model was accurate and effective, and the AUC in the training cohort was 0.7958 (95%CI,

Variable	OR	95%CI	*P*-value	OR	95%CI	*P*-value
Race						
White	Ref					
Black	1.272	0.859–1.883	0.231			
Other	0.783	0.542–1.132	0.193			
Age						
<45	Ref					
≥45	1.279	0.990–1.652	0.06			
Sex						
Male	Ref					
Female	0.891	0.689–1.153	0.38			
Marital status						
Married	Ref					
Unmarried	0.812	0.622–1.061	0.126			
Tumor location						
Right colon	Ref					
Left colon	1.308	0.962–1.779	0.087			
Rectum	0.819	0.570_1.176	0.28			
Grade						
Well/moderately	Ref			Ref		
Poorly/undifferentiated	1.498	1.111–2.020	0.008	1.305	0.932–1.827	0.121
Histology						
Adenocarcinoma	Ref					
Non-adenocarcinoma	0.562	0.312–1.010	0.054			
T stage						
T1-T2	Ref			Ref		
T3-T4	1.504	1.040–2.176	0.03	0.758	0.488–1.179	0.219
N stage						
N0	Ref			Ref		
N1-N2	3.041	2.217–4.172	<0.001	2.565	1.792–3.670	<0.001
Pretreatment CEA						
negative/normal	Ref			Ref		
positive/elevated	6.6	4.875–8.935	<0.001	5.735	4.155–7.917	<0.001
Bone metastasis						
No	Ref			Ref		
Yes	33.485	4.016-279.231	0.001	27.395	2.717-276.247	0.005
Brain metastasis						
No	Ref					
Yes	<0.001	<0.001	0.999			
Lung metastasis						
No	Ref			Ref		
Yes	14.861	8.635–25.574	<0.001	11.850	6.482–21.664	<0.001
Tumor Size						
<5 cm	Ref					
≥5 cm	1.129	0.944–1.575	0.129			

Table 2. Univariate and multivariate logistic analysis of LM in EOCRC patients.

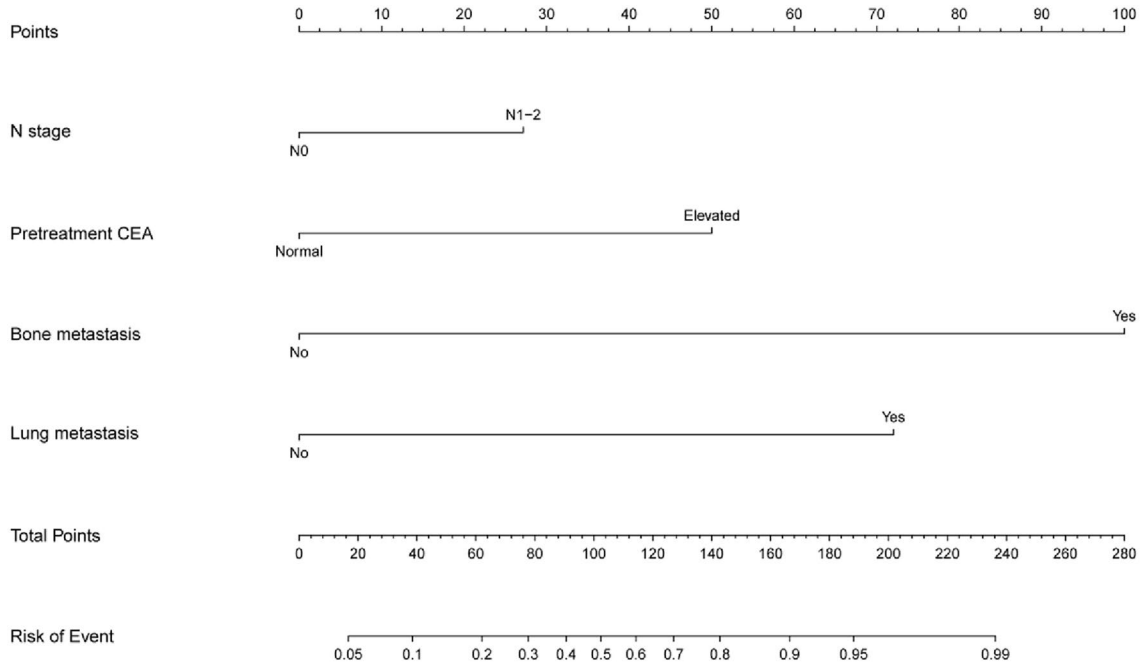

Fig. 2. A nomogram for predicting liver metastasis in EOCRC patients.

Variables	Points	Variables	Points
N stage		LM probability	Total-points
N0	0	10%	38
N1-2	27.16442	20%	62
Pretreatment CEA		30%	78
Normal	0	40%	91
Elevated	50.03282	50%	102
Bone metastasis		60%	114
No	0	70%	127
Yes	100	80%	143
Lung metastasis		90%	166
No	0	95%	188
Yes	72.02402		

Table 3. Nomogram scoring system.

0.771–0.8206) (Fig. 3A), the AUC in the validation cohort was 0.7653 (95%CI, 0.7218–0.8088) (Fig. 3B). Both calibration plots showed high consistency between the observed and predicted results (Fig. 4A-B).DCA has been shown to have clinical value in training and validation sets and has higher predictive accuracy than TN staging (Fig. 5A-B). Finally, the AUC of the combined nomogram is greater than that of any independent predictor (Fig. 6A-B).

RF machine learning model

The AUC of the RF model in predicting the risk of liver metastasis in training set was 0.885 (95%CI, 0.870–0.901), and the AUC in the validation set was 0.750 (95%CI, 0.704–0.796) (Fig. 7A). We examined the variable importance of RF model, Mean Decrease Accuracy (MDA) analysis identified CEA, lung metastasis, and N stage as the most influential predictors (Fig. 7B), these three variables were also included in the logistic regression nomogram. Next, we used the SHAP algorithm to enhance model interpretability. The SHAP summary plot (Fig. 8A) and beeswarm plot (Fig. 8B) ranked the features in the model and the impact of each feature on the model's predictive performance. The dependence plot illustrated the relationship between the variables in the model and their corresponding SHAP values (8 C). Among the 14 variables included in the model, CEA exhibited the largest mean absolute SHAP value.

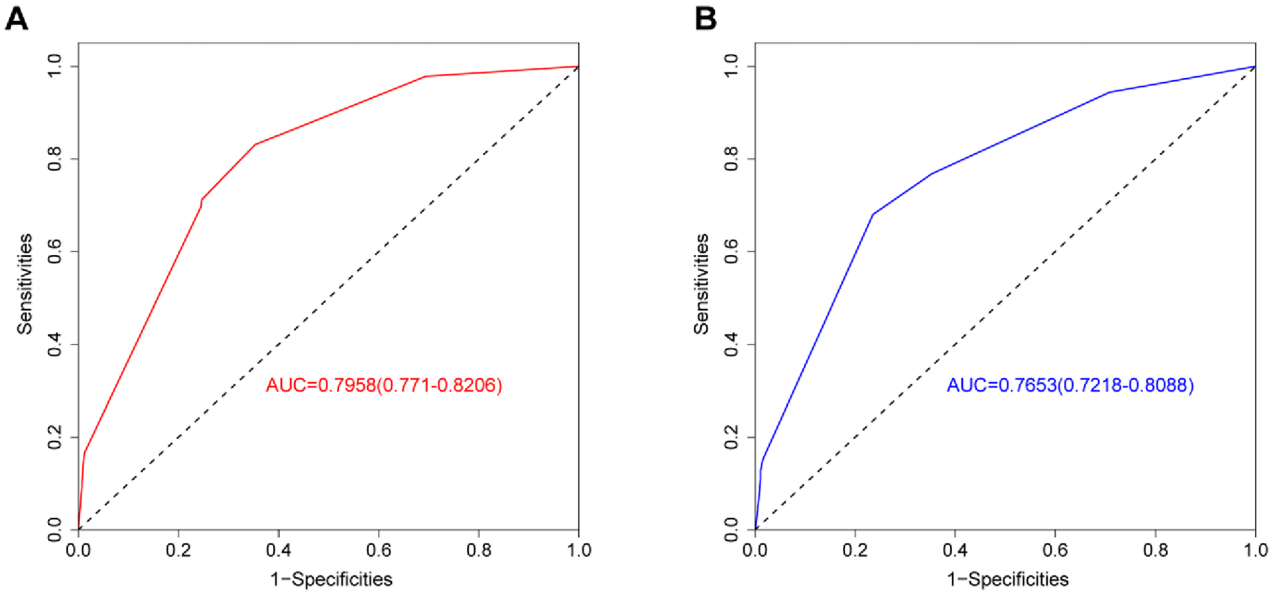

Fig. 3. The receiver operating characteristic curve of nomogram in training set (**A**); The receiver operating characteristic curve of nomogram in validation set (**B**).

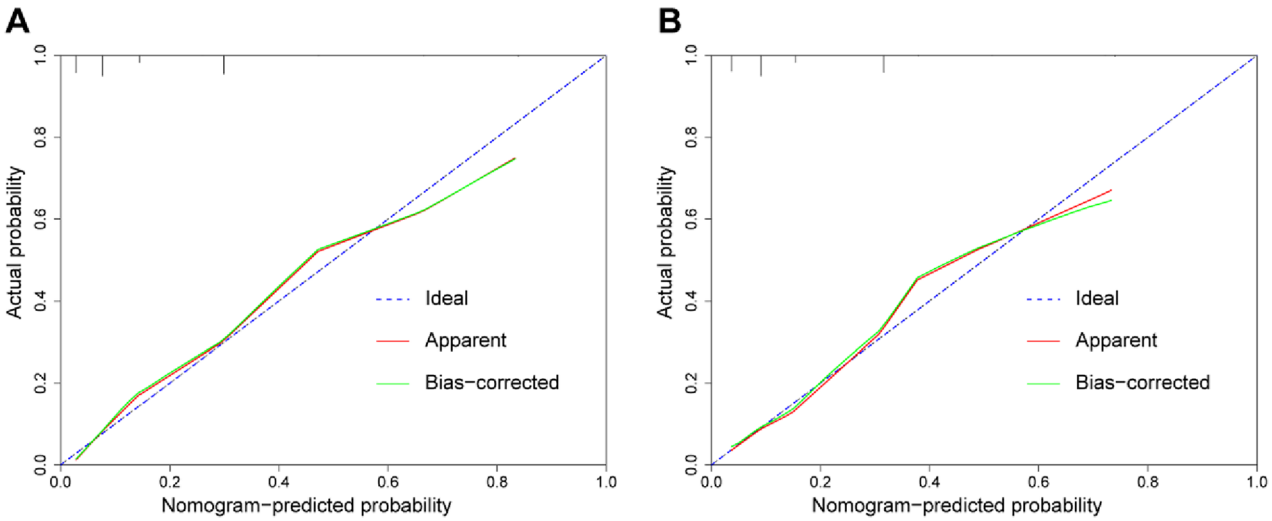

Fig. 4. The calibration curve of nomogram in the training set (**A**) and validation set (**B**).

Discussion

The incidence of EOCRC has increased globally among both men and women[20,21]. The most common symptoms of EOCRC are abdominal pain and rectal bleeding, and it often occurs in the left colon and rectum[4], it is often diagnosed at an advanced stage, which may be due to the lack of screening for early lesions. Moreover, although early-onset metastatic CRC has more favorable baseline characteristics such as fewer comorbidities, better physical condition, higher frequency of surgery and radiotherapy, higher chemotherapy doses, and fewer treatment-related adverse events compared with late-onset metastatic CRC, the survival rate of early-onset metastatic CRC has not improved[22]. This may be due to delayed diagnosis, potential differences in tumor biology, or other as-yet-undetermined factors that increase the aggressiveness of the disease in these patients. Colorectal cancer metastasis often affects various organs, with colorectal cancer liver metastasis (CRLM) being the most common, and its severity is associated with a poor prognosis[23]. In this study, we developed a predictive model to predict the risk of liver metastasis in patients with early-onset colorectal cancer.

Nomograms predict specific outcomes based on the combination of crucial predictive factors and have become widely used practical clinical tools in cancer research. Each variable in a nomogram is assigned a score of 100, and patients obtain a total score by summing the scores of each variable. The higher the score, the higher the risk of liver metastasis. In this work, we used four variables to construct a nomogram, including N stage,

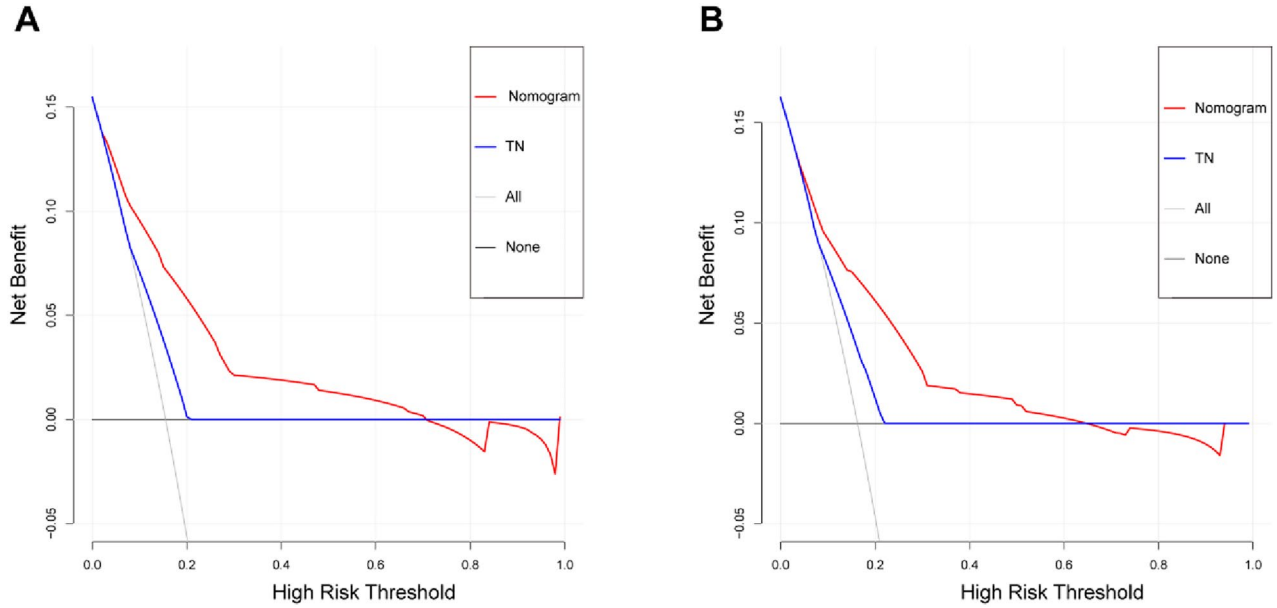

Fig. 5. Comparison of decision curve analysis between the predictive nomogram and TN staging in the training set (**A**) and validation set (**B**).

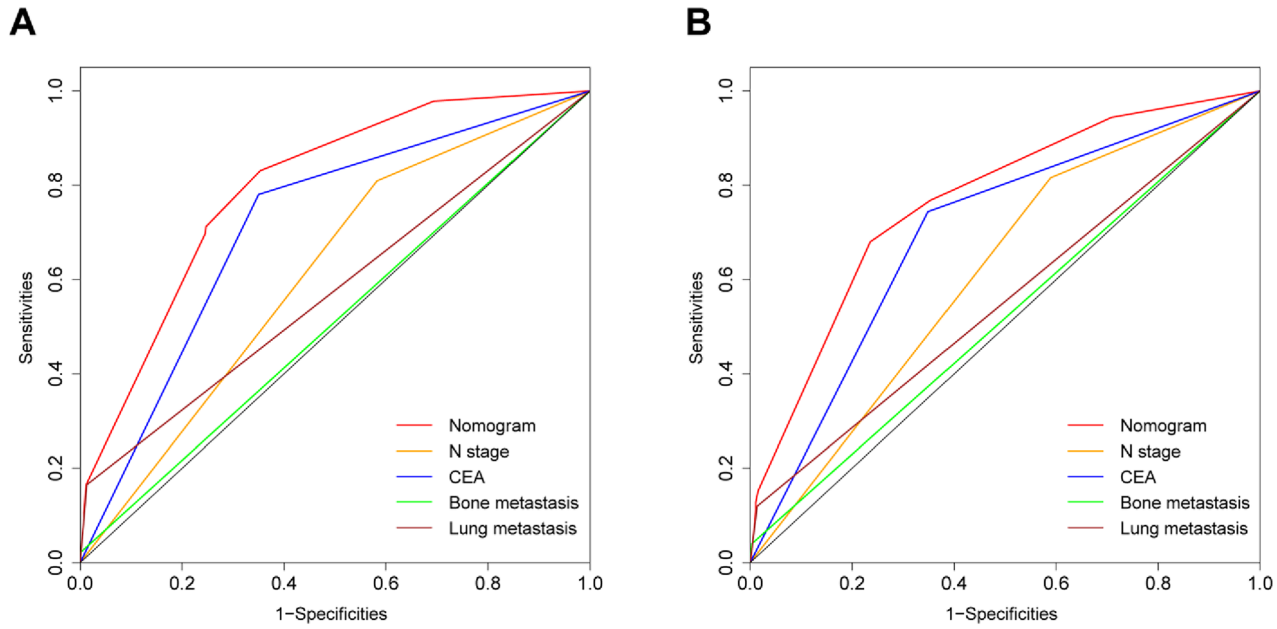

Fig. 6. Comparison of the values of area under the curve between nomogram and single independent risk factor in the training set (**A**) and validation set (**B**).

pretreatment CEA, bone metastasis, and lung metastasis. As previously reported, the lung is the second most common target organ for CRC metastasis, accounting for 20–30% of distant metastases in CRC[24,25]. CRC bone metastasis is most often associated with liver and lung metastases, with 83% of patients with bone metastasis also having lung, liver, and brain metastases[26], and 57% having liver involvement[27], Isolated bone metastasis accounts for only 1–2%[28]. Therefore, whether other extrahepatic metastases have a synergistic effect on the development of liver metastasis is worthy of further study. We observed that CEA is also an important risk factor. According to previous reports, CEA is a prognostic indicator of the status of CRC patients and is involved in multiple steps of CRC-related liver metastasis[29,30]. Therefore, monitoring and follow-up of CEA in patients are also important. In addition, N stage is also one of the factors affecting tumor metastasis. The higher the N stage, the higher the possibility of tumor reaching the liver through the lymphatic system and blood circulation, and the greater the risk of liver metastasis[31,32]. As our results revealed, the AUC of the nomogram we constructed was 0.7958 (95%

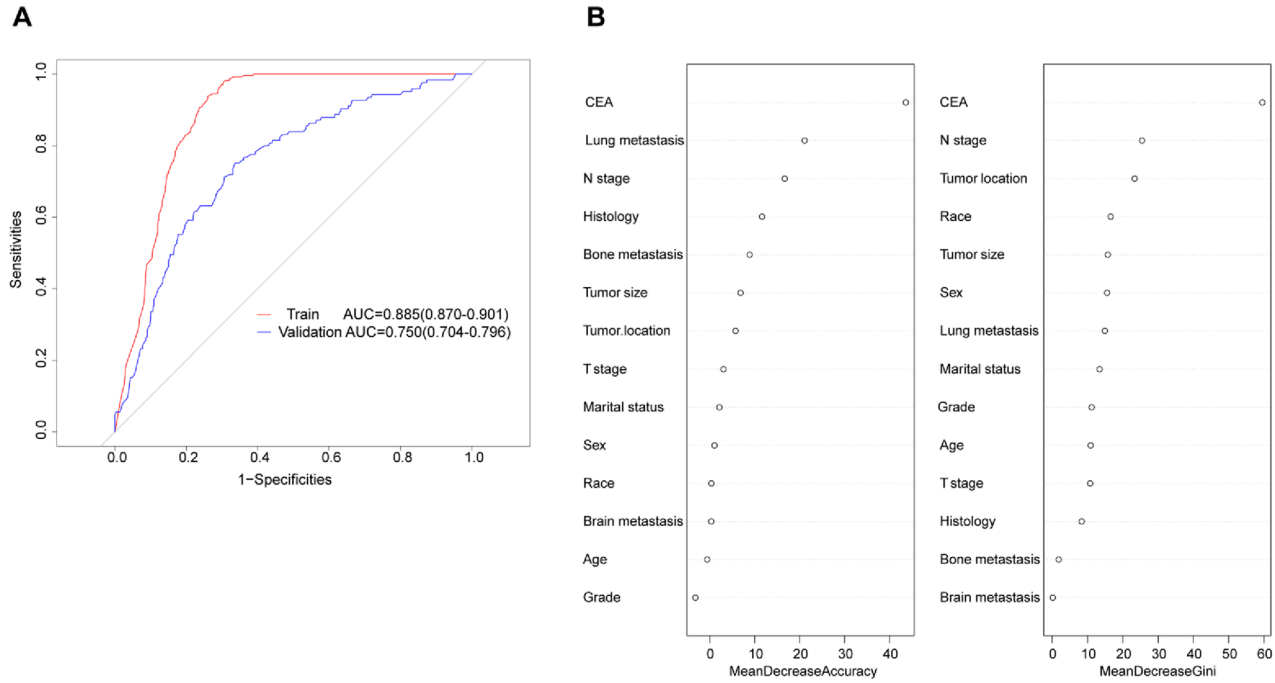

Fig. 7. Performance evaluation and feature analysis of the RF model. (**A**) Receiver operating characteristic curve of the RF model; (**B**) Variable importance plot showing the relative importance of features in the RF model.

CI, 0.771–0.8206) in the training set and 0.7653 (95% CI, 0.7218–0.8066) in the validation set, indicating its good predictive accuracy. The calibration curves and DCA results of the training set and validation set indicated that the model has good clinical applicability.

Although univariate analysis revealed that higher T stage and poor tumor differentiation were associated with liver metastasis, T stage and grade failed to achieve statistical significance ($P > 0.3$) in the multivariate analysis that incorporated N stage and other variables. This finding contrasts with the general consensus in tumor biology that the extent of local invasion and differentiation of the primary tumor are key drivers of tumor progression[33,34]. The non-significance of these variables may be attributed to potential collinearity among variables or the complexity of biological processes. For instance, T stage is clinically closely correlated with N stage, which reflects the degree of lymph node involvement[35,36]. In this context, the model tends to allocate predictive power to N stage, which is statistically more robust and more dominant in predicting liver metastasis. Although T stage and grade did not reach statistical significance in our analysis, their biological plausibility and potential collinearity with other variables underscore the need for further research to comprehensively elucidate the role of these factors within the context of our study.

SHAP-based RF analysis revealed that CEA and N stage as dominant predictors, and these variables were also included in the nomogram. Meanwhile, we observed discrepancies in the influential variables between the nomogram and machine learning models. The extremely low incidence of bone metastasis ($n = 14$, 0.5%), as a rare event, led to instability in the logistic regression predictive model and limited SHAP's capacity to capture its true association with liver metastasis. Although lung metastasis occurred more frequently than bone metastasis, it may exhibit a non-linear relationship with liver metastasis; in the presence of stronger signals from CEA and N stage, SHAP analysis failed to fully capture this relationship. On the other hand, despite race, tumor location, and gender showing no significant significance in univariate analysis, they ranked relatively high in SHAP, which may be attributed to their potential interactions with other variables—interactions that SHAP values can capture.

Logistic regression emphasizes linear associations and clinical interpretability, while the mean decrease in accuracy (MDA) of random forests focuses on the degree of model accuracy reduction upon variable permutation. In contrast, SHAP values quantify feature-level contributions to individual predictions by decomposing model outputs, thereby capturing both linear and non-linear relationships. The discrepancy in variable importance ranking between SHAP analysis and logistic regression underscores the importance of considering multiple modeling approaches. While logistic regression remains valuable for clinical translation, machine learning techniques provide critical insights into variable importance and complex relationships. Integrating traditional statistical methods with interpretable machine learning techniques enhances the robustness of predictive models. Future studies could explore advanced methodologies, such as deep learning, to capture complex interactions and improve clinical applicability.

Although the nomogram established based on the SEER database has good accuracy, it also has some potential limitations. Due to the limited factors included in the SEER database, some possible risk factors were not included in the study, such as alcohol consumption, dietary habits, chronic hepatitis and specific chemotherapy

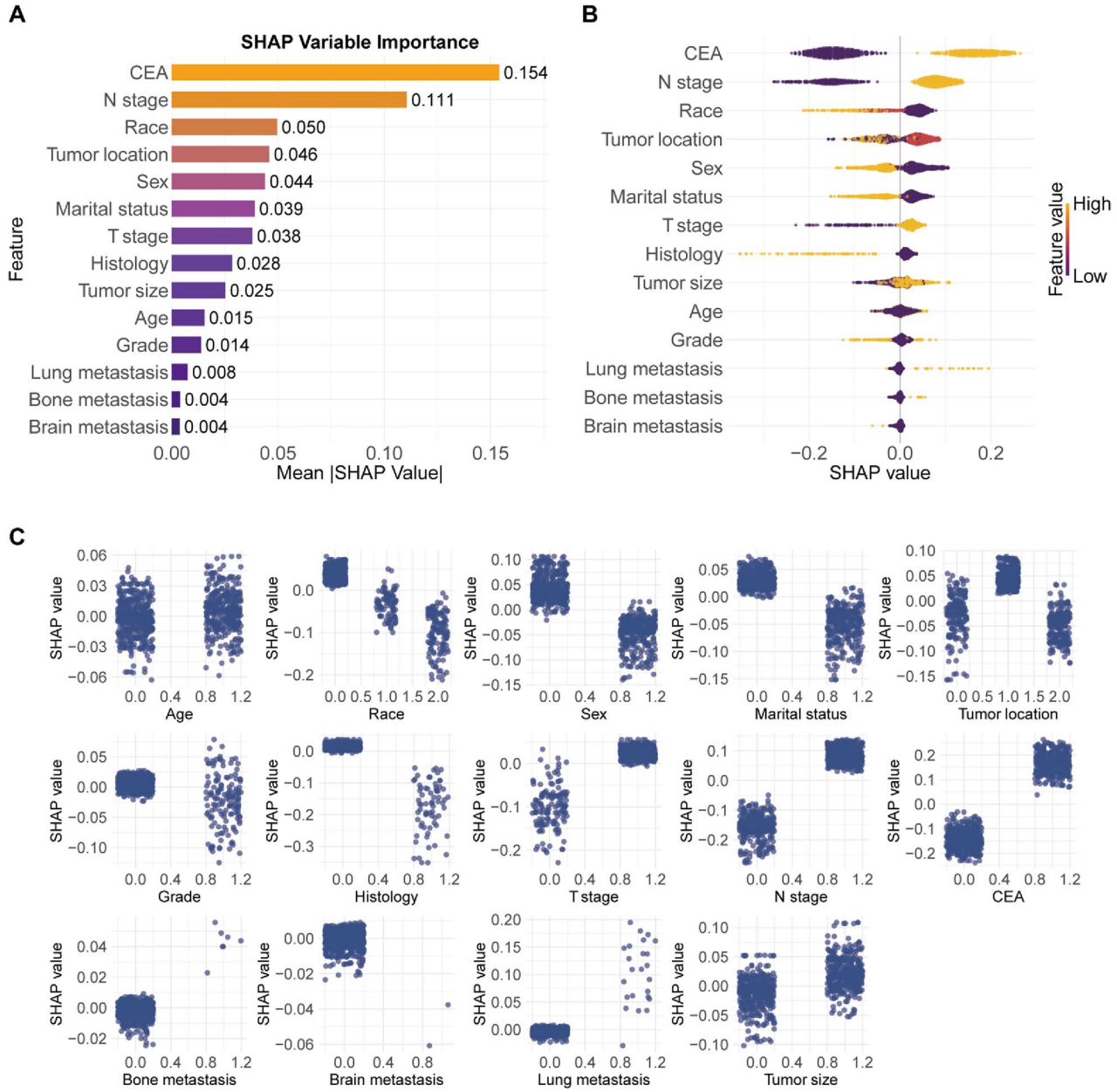

Fig. 8. SHAP value analysis of the RF model. (**A**) Summary plot showing the overall impact of features on model outputs; (**B**) Beeswarm plot displaying the distribution of SHAP values for each feature across all samples; (C) Dependence plot illustrating the relationship between variables and their SHAP values, with partial dependence trends.

regimens. Therefore, our model lacks the relationship between these factors and liver metastasis. In addition, studies based on SEER data are all retrospective, which may bring inevitable selection bias and information bias. Among the included variables, rare events like bone metastasis may inflate effect sizes and reduce model stability, necessitating cautious interpretation. Therefore, future prospective studies using large, multi-center samples are needed to further validate the nomogram.

Conclusion

Our research utilized data from the SEER database to develop a predictive model for CRC patients under the age of 50. We found that N stage, pretreatment CEA, bone metastasis, and lung metastasis are risk factors for liver metastasis in EOCRC patients. These factors are relatively easy to obtain in most hospitals. Additionally, we developed a new nomogram that can predict the risk of liver metastasis in EOCRC patients, and the nomogram demonstrated good accuracy, reliability, and clinical applicability. This convenient and visual tool can assist clinicians in risk assessment and prognosis prediction. Machine learning models based on RF have enhanced our understanding of liver metastasis in EOCRC patients and facilitated physicians in making informed decisions.

Data availability

The datasets generated and/or analyzed in this study are available in the SEER database(https://seer.cancer.gov/).Data are available upon reasonable request from the lead contact, Qun Zhang(qunzhangnora@outlook.com).

Received: 6 April 2025; Accepted: 29 August 2025
Published online: 25 September 2025

References

1. Sung, H. et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249. https://doi.org/10.3322/caac.21660 (2021).
2. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2017. *CA Cancer J. Clin.* **67** https://doi.org/10.3322/caac.21387 (2017).
3. Edwards, B. K. et al. Annual report to the Nation on the status of cancer, 1975–2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. *Cancer* **116**, 544–573. https://doi.org/10.1002/cncr.24760 (2010).
4. Akimoto, N. et al. Rising incidence of early-onset colorectal cancer - a call to action. *Nat. Rev. Clin. Oncol.* **18**, 230–243. https://doi.org/10.1038/s41571-020-00445-1 (2021).
5. Stoffel, E. M. & Murphy, C. C. Epidemiology and mechanisms of the increasing incidence of colon and rectal cancers in young adults. *Gastroenterology* **158**, 341–353. https://doi.org/10.1053/j.gastro.2019.07.055 (2020).
6. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA Cancer J. Clin.* **68** https://doi.org/10.3322/caac.21442 (2018).
7. Zaborowski, A. M. et al. Characteristics of early-onset vs late-onset colorectal cancer: A review. *JAMA Surg.* **156**, 865–874. https://doi.org/10.1001/jamasurg.2021.2380 (2021).
8. Chen, F. W., Sundaram, V., Chew, T. A. & Ladabaum, U. Advanced-stage colorectal cancer in persons younger than 50 years not associated with longer duration of symptoms or time to diagnosis. *Clin. Gastroenterol. Hepatol.* **15** https://doi.org/10.1016/j.cgh.2016.10.038 (2017).
9. Kneuertz, P. J. et al. Overtreatment of young adults with colon cancer: more intense treatments with unmatched survival gains. *JAMA Surg.* **150**, 402–409. https://doi.org/10.1001/jamasurg.2014.3572 (2015).
10. Yoshino, T. et al. Biomarker analysis beyond angiogenesis: RAS/RAF mutation status, tumour sidedness, and second-line ramucirumab efficacy in patients with metastatic colorectal carcinoma from RAISE-a global phase III study. *Ann. Oncol.* **30**, 124–131. https://doi.org/10.1093/annonc/mdy461 (2019).
11. Stewart, C. L. et al. Cytoreduction for colorectal metastases: liver, lung, peritoneum, lymph nodes, bone, brain. When does it palliate, prolong survival, and potentially cure? *Curr. Probl. Surg.* **55**, 330–379. https://doi.org/10.1067/j.cpsurg.2018.08.004 (2018).
12. Margonis, G. A. et al. Impact of surgical margin width on recurrence and overall survival following R0 hepatic resection of colorectal metastases: a systematic review and meta-analysis. *Ann. Surg.* **267**, 1047–1055. https://doi.org/10.1097/SLA.0000000000002552 (2018).
13. Ma, Y., Guo, C., Wang, X., Wei, X. & Ma, J. Impact of chemotherapeutic agents on liver microenvironment: oxaliplatin create a pro-metastatic landscape. *J. Exp. Clin. Cancer Res.* **42**, 237. https://doi.org/10.1186/s13046-023-02804-z (2023).
14. Kuipers, E. J. & Grobbee, E. J. Personalised screening for colorectal cancer, ready for take-off. *Gut* **69**, 403–404. https://doi.org/10.1136/gutjnl-2019-319677 (2020).
15. Balachandran, V. P., Gonen, M., Smith, J. J. & DeMatteo, R. P. Nomograms in oncology: more than Meets the eye. *Lancet Oncol.* **16**, e173–e180. https://doi.org/10.1016/S1470-2045(14)71116-7 (2015).
16. Van Calster, B. & Wynants, L. Machine learning in medicine. *N Engl. J. Med.* **380**, 2588. https://doi.org/10.1056/NEJMc1906060 (2019).
17. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit. Med.* **5** https://doi.org/10.1038/s41746-022-00592-y (2022).
18. Bifarin, O. O. Interpretable machine learning with tree-based Shapley additive explanations: application to metabolomics datasets for binary classification. *PLoS One.* **18**, e0284315. https://doi.org/10.1371/journal.pone.0284315 (2023).
19. Hankey, B. F., Ries, L. A. & Edwards, B. K. The surveillance, epidemiology, and end results program: a National resource. *Cancer Epidemiol. Biomarkers Prev.* **8**, 1117–1121 (1999).
20. Vuik, F. E. et al. Increasing incidence of colorectal cancer in young adults in Europe over the last 25 years. *Gut* **68**, 1820–1826. https://doi.org/10.1136/gutjnl-2018-317592 (2019).
21. Siegel, R. L. et al. Global patterns and trends in colorectal cancer incidence in young adults. *Gut* **68**, 2179–2185. https://doi.org/10.1136/gutjnl-2019-319511 (2019).
22. Lipsyc-Sharf, M. et al. Survival in Young-Onset metastatic colorectal cancer: findings from cancer and leukemia group B (Alliance)/SWOG 80405. *J. Natl. Cancer Inst.* **114**, 427–435. https://doi.org/10.1093/jnci/djab200 (2022).
23. Tauriello, D. V. F., Calon, A., Lonardo, E. & Batlle, E. Determinants of metastatic competency in colorectal cancer. *Mol. Oncol.* **11** https://doi.org/10.1002/1878-0261.12018 (2017).
24. Hugen, N., van de Velde, C. J. H., de Wilt, J. H. W. & Nagtegaal, I. D. Metastatic pattern in colorectal cancer is strongly influenced by histological subtype. *Ann. Oncol.* **25**, 651–657. https://doi.org/10.1093/annonc/mdt591 (2014).
25. Li, J. et al. Expert consensus on multidisciplinary therapy of colorectal cancer with lung metastases (2019 edition). *J. Hematol. Oncol.* **12**, 16. https://doi.org/10.1186/s13045-019-0702-0 (2019).
26. Kanthan, R., Loewy, J. & Kanthan, S. C. Skeletal metastases in colorectal carcinomas: a Saskatchewan profile. *Dis. Colon Rectum.* **42**, 1592–1597 (1999).
27. Roth, E. S. et al. Does colon cancer ever metastasize to bone first? A Temporal analysis of colorectal cancer progression. *BMC Cancer.* **9**, 274. https://doi.org/10.1186/1471-2407-9-274 (2009).
28. Katoh, M., Unakami, M., Hara, M. & Fukuchi, S. Bone metastasis from colorectal cancer in autopsy cases. *J. Gastroenterol.* **30**, 615–618 (1995).
29. Thomas, D. S. et al. Evaluation of serum CEA, CYFRA21-1 and CA125 for the early detection of colorectal cancer using longitudinal preclinical samples. *Br. J. Cancer.* **113**, 268–274. https://doi.org/10.1038/bjc.2015.202 (2015).
30. Hatate, K. et al. Liver metastasis of colorectal cancer by protein-tyrosine phosphatase type 4A, 3 (PRL-3) is mediated through lymph node metastasis and elevated serum tumor markers such as CEA and CA19-9. *Oncol. Rep.* **20**, 737–743 (2008).
31. Compton, C. C. & Greene, F. L. The staging of colorectal cancer: 2004 and beyond. *CA Cancer J. Clin.* **54**, 295–308 (2004).
32. Wang, J. et al. Metastatic patterns and survival outcomes in patients with stage IV colon cancer: A population-based analysis. *Cancer Med.* **9**, 361–373. https://doi.org/10.1002/cam4.2673 (2020).
33. Sudo, M. et al. Long-term outcomes after surgical resection in patients with stage IV colorectal cancer: a retrospective study of 129 patients at a single institution. *World J. Surg. Oncol.* **17** https://doi.org/10.1186/s12957-019-1599-3 (2019).
34. Li, J. et al. TNM staging of colorectal cancer should be reconsidered by T stage weighting. *World J. Gastroenterol.* **20**, 5104–5112. https://doi.org/10.3748/wjg.v20.i17.5104 (2014).
35. Egashira, Y. et al. Analysis of pathological risk factors for lymph node metastasis of submucosal invasive colon cancer. *Mod. Pathol.* **17**, 503–511 (2004).

36. Kitajima, K. et al. Correlations between lymph node metastasis and depth of submucosal invasion in submucosal invasive colorectal carcinoma: a Japanese collaborative study. *J. Gastroenterol.* **39**, 534–543 (2004).

Acknowledgements
This research was supported by the National Natural Science Foundation of China(82303970),the Nanjing Health Science and Technology Development Key Program(ZKX21028),the Jiangsu Scientific and Technological Development of Traditional Chinese Medicine Key projects(ZD202227) and the Provincial Natural Science Foundation of Jiangsu(BK20211007).

Author contributions
X.H wrote the main manuscript.X.H and X.B prepared all figures and tables. QZ and XQ were responsible for funding acquisition for the study.All authors reviewed and approved the final manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Q.Z. or X.Q.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.