Observational Study

Medicine

OPEN

Short-term prognosis for hepatocellular carcinoma patients with lung metastasis

A retrospective cohort study based on the SEER database

Shicheng Chen, MBBS[a,b], Xiaowen Li, MBBS[a,b], Yichao Liang, MM[c], Xinyu Lu, MBBS[a,b], Yingyi Huang, MBBS[d], Jiajia Zhu, MBBS[e], Jun Li, MM[a,b,*]

Abstract

Our study aimed to develop a prediction model to predict the short-term mortality of hepatocellular carcinoma (HCC) patients with lung metastasis. The retrospective data of HCC patients with lung metastasis was from the Surveillance, Epidemiology, and End Results registration database between 2010 and 2015. 1905 patients were randomly divided into training set (n = 1333) and validation set (n = 572). There were 1092 patients extracted from the Surveillance, Epidemiology, and End Results database 2015 to 2019 as the validation set. The variable importance was calculated to screen predictors. The constructed prediction models of logistic regression, random forest, broad learning system, deep neural network, support vector machine, and naïve Bayes were compared through the predictive performance. The mortality of HCC patients with lung metastasis was 51.65% within 1 month. The screened prognostic factors (age, N stage, T stage, tumor size, surgery, grade, radiation, and chemotherapy) and gender were used to construct prediction models. The area under curve (0.853 vs. 0.771) of random forest model was more optimized than that of logistic regression model in the training set. But, there were no significant differences in testing and validation sets between random forest and logistic regression models. The value of area under curve in the logistic regression model was significantly higher than that of the broad learning system model (0.763 vs. 0.745), support vector machine model (0.763 vs. 0.689) in the validation set, and higher than that of the naïve Bayes model (0.775 vs. 0.744) in the testing model. We further chose the logistic regression prediction model and built the prognostic nomogram. We have developed a prediction model for predicting short-term mortality with 9 easily acquired predictors of HCC patients with lung metastasis, which performed well in the internal and external validation. It could assist clinicians to adjust treatment strategies in time to improve the prognosis.

Abbreviations: AUC = the area under curve, BLS = broad learning system, DNN = deep neural network, HCC = hepatocellular carcinoma, PPV = positive predictive value, ROC = the receiver operating characteristic, SEER = the Surveillance, Epidemiology, and End Results, SVM = support vector machine.

Keywords: hepatocellular carcinoma, lung metastasis, nomogram, SEER

1. Introduction

Hepatocellular carcinoma (HCC) is the seventh most common cancer in the world and the second leading cause of cancer-related deaths.[1,2] According to the data from GLOBOCAN, in North America, the incidence risk of HCC in 2018 was estimated to be 6.6 per 100,000 men and 3.4 per 100,000 women.[1] The rate of global mortality associated with HCC in 2018 was 8.5 per 100,000 person-years, and the prognosis of HCC worldwide was very poor.[1] HCC had a great tendency to invade the portal vein and hepatic vein, leading to intrahepatic and extrahepatic metastasis, a well-known significant adverse prognostic factor.[3,4] Studies showed that the prognosis of patients with extrahepatic metastasis was worse than that of patients without extrahepatic metastasis.[5,6] The metastasis sites included lung metastasis, brain metastasis, bone metastasis, lymph node metastasis, etc.[7]

The lung was the most common site of extrahepatic metastasis, and the prognosis of patients was extremely poor.[5] More than

The study was funded by Science and Technology Planning Project of Guangdong Province of China (2017A020213018).

The authors have no conflicts of interest to disclose.

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

Supplemental Digital Content is available for this article.

[a] Department of Traditional Chinese Medicine, Nanfang Hospital, Guangzhou, P. R. China, [b] School of Chinese Medicine, Southern Medical University, Guangzhou, P. R. China, [c] Department of Hepatology, TCM-Integrated Hospital of Southern Medical University, Guangzhou, P. R. China, [d] Department of Neurology, Guangzhou First People's Hospital, Guangzhou, P. R. China, [e] Department of Neurology, Nanfang Hospital, Guangzhou, P. R. China.

**Correspondence: Jun Li, Department of Traditional Chinese Medicine, Nanfang Hospital of Southern Medical University, No. 1838 North Guangzhou Avenue, Guangzhou, Guangdong 510515, P. R. China (e-mail: junlicm@163.com).*

How to cite this article: Chen S, Li X, Liang Y, Lu X, Huang Y, Zhu J, Li J. Short-term prognosis for hepatocellular carcinoma patients with lung metastasis: A retrospective cohort study based on the SEER database. Medicine 2022;101:45(e31399).

Received: 4 January 2022 / Received in final form: 23 September 2022 / Accepted: 28 September 2022

http://dx.doi.org/10.1097/MD.0000000000031399

70% of deaths in HCC patients were caused by lung metastasis.[8] Therefore, it is of great significance to establish models for predicting the prognosis of HCC patients with lung metastasis.[9] Previous studies have established models for the risk of bone metastases, brain metastases, and extrahepatic metastases in HCC patients.[10–12] Meanwhile, there were several prognostic models for HCC patients with bone and lung metastases.[10,13] However, the prognostic model of HCC patients with lung metastasis was not satisfied, and the model has not been optimized, which needs further clarification in population-based studies.

Therefore, this study aimed to develop 6 prediction models for predicting short-term mortality in HCC patients with lung metastasis using easy-to-collect demographic and clinicopathological variables and obtain a more optimal model based on the predictive performance of the 6 models. Clinicians could utilize the predictive model to predict the course of HCC patients more realistically, accurately and timely and to optimize the choice of treatment options.

2. Methods

2.1. Study population

The study was a retrospective cohort study. A total of 1934 HCC patients with lung metastasis were extracted from the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) database[14] between 2010 and 2015 as training and testing sets, and a total of 1386 HCC patients with lung metastasis were extracted from the SEER database between 2015 and 2020 as the validation set. The data on cancer diagnosis, treatment, and survival for about 30% of the U.S. population was included in the SEER database of the National Cancer Institute. The eligible patients were randomly divided into training and testing sets according to the ratio of 7:3. All the data collected by SEER was de-identified, and the study was exempted from the Institutional Ethics Committee.

2.2. Data collection

Variable definitional information was encoded at diagnosis and made available in the SEER database. For each eligible patient, the following data were collected: age, gender, race, marital status, T stage, N stage, grade [I (well differentiated), II (moderately differentiated), III (poorly differentiated), and IV (undifferentiated/anaplastic)], tumor size (≤5 cm, >5 cm), bone metastasis (yes/no), brain metastasis (yes/no), intrahepatic metastasis (yes/no), and treatment methods (including surgery, chemotherapy, and radiation). The follow-up outcome we focus on was the 1-month all-cause mortality obtained from the SEER project. The average follow-up time was 3.855 months.

2.3. Development of prediction models for predicting 1-month all-cause mortality

First, we conducted a descriptive analysis and compared the variables between the training set and the testing set. Second, using random forest to filter variables in the training set based on variable importance as prognostic factors for 1-month all-cause mortality.[15]

Third, the identified prognostic variables and sex (a common important influencing factor)[16] were utilized to construct prediction models of logistic regression, random forest,[17] broad learning system (BLS),[18] deep neural network (DNN),[19] support vector machine (SVM),[20] and naïve Bayes,[21] respectively.

Random forest models contained a collection of decision trees. In the process of building each decision tree, different random subsets of the variables from the training dataset were selected to establish how best to partition the dataset at each node.[17] This study used the parameters of n_estimators was 300, max_depth was 7, random_state = 0, bootstrap = TRUE, criterion="gini," min_samples_split = 2, min_samples_leaf = 1, and min_weight_fraction_leaf = 0.0.

BLS[18] is a deep structure neural network model based on Random Vector Functional-link Neural Network. Deep learning transformed and extracted features based on the data and became the feature layer as the input layer of the entire neural network. When the feature layer was merged with the enhancement layer, the pseudo-inverse was used to calculate the weight 1. This study used N1 (the number of feature nodules per window) = 10, N2 (the number of windows per node) = 9, and N3 (the number of enhanced nodes) = 30 to learn features, and it was 120 features.

DNN[19] includes 3 layers, input layer, hidden layer, and output layer. Each layer is fully connected. The training set was used as the input layer, and the sample features were progressively obtained through the hidden layer, and then the features in the output layer were predicted. This study used the parameters of hidden_layer_sizes = 310, activation = relu, the hidden layer activation function is the rectified linear unit function, solver="adam," "adam" refers to a stochastic gradient-based optimizer, and the random seed is 2021. 30 hidden layers were used in this study.

SVM is a classifier that classifies data by finding the decision boundary with the largest margin. This study used the nonlinear SVM method, C-Support Vector Classification.[20] The nonlinear classifier introduces a kernel function to map the features in the original space to a higher dimension for classification. Regularization parameter C was 1.00. The strength of the regularization is inversely proportional to C. The penalty is a squared l2 penalty. The parameter of kernel = rbf specifies the kernel type to be used in the algorithm and the random seed is 2021.

Naïve Bayes[21] uses the Bayesian formula to calculate the posterior probability (conditional probability of object feature

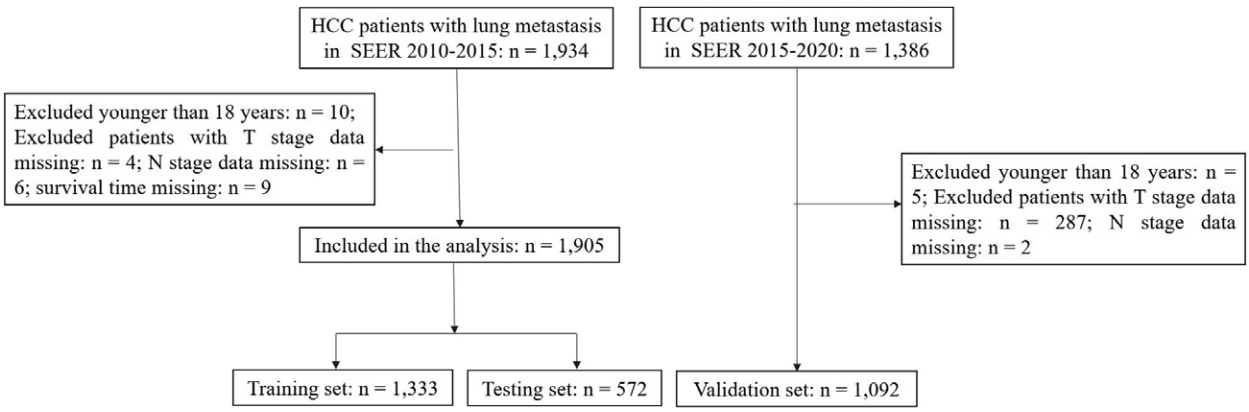

Figure 1. Flowchart of the systematic selection process.

Table 1

Demographic and clinicopathological characteristics of selected patients.

Variables	Total (n = 1905)	Groups		Statistic	P
		Training set (n = 1333)	Validation set (n = 572)		
Age, mean ± SD	62.56 ± 11.48	62.57 ± 11.67	62.52 ± 11.05	$t = 0.10$	.923
Gender, n (%)					
Female	398 (20.89)	282 (21.16)	116 (20.28)	$\chi^2 = 0.186$	.667
Male	1507 (79.11)	1051 (78.84)	456 (79.72)		
Race, n (%)					
White	1187 (62.31)	822 (61.67)	365 (63.81)	–	.793
Black	329 (17.27)	236 (17.70)	93 (16.26)		
Other	383 (20.10)	271 (20.33)	112 (19.58)		
Unknown	6 (0.31)	4 (0.30)	2 (0.35)		
Marital status, n (%)					
Married	851 (44.67)	596 (44.71)	255 (44.58)	$\chi^2 = 0.849$	.838
Single	520 (27.30)	357 (26.78)	163 (28.50)		
Divorced/widowed/separated	444 (23.31)	315 (23.63)	129 (22.55)		
Unknown	90 (4.72)	65 (4.88)	25 (4.37)		
T stage, n (%)					
T1	384 (20.16)	258 (19.35)	126 (22.03)	$\chi^2 = 7.377$	.117
T2	170 (8.92)	112 (8.40)	58 (10.14)		
T3	681 (35.75)	478 (35.86)	203 (35.49)		
T4	233 (12.23)	178 (13.35)	55 (9.62)		
TX	437 (22.94)	307 (23.03)	130 (22.73)		
N stage, n (%)					
N0	1171 (61.47)	824 (61.82)	347 (60.66)	$\chi^2 = 1.272$	.530
N1	374 (19.63)	253 (18.98)	121 (21.15)		
NX	360 (18.90)	256 (19.20)	104 (18.18)		
Grade, n (%)					
I	87 (4.57)	60 (4.50)	27 (4.72)	$\chi^2 = 0.071$	.995
II	212 (11.13)	149 (11.18)	63 (11.01)		
III/IV	263 (13.81)	185 (13.88)	78 (13.64)		
Unknown	1343 (70.50)	939 (70.44)	404 (70.63)		
Tumor size, n (%)					
≤5 cm	284 (14.91)	186 (13.95)	98 (17.13)	$\chi^2 = 3.594$	.166
>5 cm	1018 (53.44)	726 (54.46)	292 (51.05)		
Unknown	603 (31.65)	421 (31.58)	182 (31.82)		
Bone metastasis, n (%)					
No	1533 (80.47)	1062 (79.67)	471 (82.34)	$\chi^2 = 3.578$	.167
Yes	314 (16.48)	233 (17.48)	81 (14.16)		
Unknown	58 (3.04)	38 (2.85)	20 (3.50)		
Brain metastasis, n (%)					
No	1795 (94.23)	1255 (94.15)	540 (94.41)	$\chi^2 = 0.549$	.760
Yes	40 (2.10)	30 (2.25)	10 (1.75)		
Unknown	70 (3.67)	48 (3.60)	22 (3.85)		
Intrahepatic metastasis, n (%)					
No	1685 (88.45)	1171 (87.85)	514 (89.86)	$\chi^2 = 2.732$	.255
Yes	188 (9.87)	141 (10.58)	47 (8.22)		
Unknown	32 (1.68)	21 (1.58)	11 (1.92)		
Surgery, n (%)					
No	1868 (98.06)	1307 (98.05)	561 (98.08)	$\chi^2 = 0.002$	.968
Yes	37 (1.94)	26 (1.95)	11 (1.92)		
Chemotherapy, n (%)					
No	1272 (66.77)	886 (66.47)	386 (67.48)	$\chi^2 = 0.186$	.666
Yes	633 (33.23)	447 (33.53)	186 (32.52)		
Radiation, n (%)					
No	1741 (91.39)	1217 (91.30)	524 (91.61)	$\chi^2 = 0.049$	.825
Yes	164 (8.61)	116 (8.70)	48 (8.39)		

to be classified) according to the prior probability (probability obtained from the training set) of the feature, and selects the class with the largest probability value as the class to which the feature belongs. The prior probability $P(Y = C_k) = m_k/m$, m is the total number of training set samples, m_k is the number of training set samples whose output is the kth category, P (Y = 0) = 48.24%, and P (Y = 1) = 51.76.

Fourth, the receiver operating characteristic (ROC) curve for each prediction model was generated. The prognostic nomogram was generated for the logistic regression model.[22] Delong's test for comparing the area under the curve (AUC) values of different prediction models. Patients in the testing and validation set were used to verify the prediction effect of the prediction model, using specificity, positive predictive value (PPV), and negative predictive value to evaluate the model. Five-fold cross validation was used to increase the performance.

2.4. Statistical analysis

Shapiro–Wilk was used to test the normality of measurement data. Normally distributed data were described by

mean ± standard deviation, and the *t* test was used. Non-normally distributed data were described by median (interquartile range), and the Mann–Whitney *U* rank sum test was used. The count data was described by the number of cases and the composition ratio [n (%)], using Chi-square or Fisher's exact test. $P < .05$ (2 sides) was statistical significance. All statistical analyses were performed by SAS v. 9.4 (SAS Institute, Cary, NC), Python software v. 3.7.4 (Python Software Foundation, DE), and R v. 4.20 (R Foundation for Statistical Computing, Vienna, Austria) software.

3. Results

3.1. Characteristics of the study population

A total of 1934 HCC patients with lung metastasis were extracted from the SEER database 2010 to 2015 as training and testing sets. We excluded patients younger than 18 (n = 10); clinicopathologic variables of interest were missing (n = 10); the information on survival time was missing (n = 9). According to the selection process, a total of 1905 patients were included in training and testing sets. There were 1386 cases of HCC complicated with lung metastasis extracted from the SEER database 2015 to 2019, 5 patients younger than 18, 287 patients with missing T stage, and 2 patients with missing N stage were excluded as the validation set (n = 1092). The flow chart of the systematic selection process was shown in Figure 1.

The mortality of HCC patients with lung metastasis was 51.65% (984/1905) within 1 month. The average follow-up time was 3.855 months. The average age was 62.56 ± 11.48 years old. There were 398 (20.89%) females and 1507 (79.11%) males. For race, 1187 (62.31%) were white, 329 (17.27%) were black, 383 (20.10%) were other races. 984 (51.65%) patients died within 1 month. More detailed information was shown in Table 1. Meanwhile, 1333 patients were incorporated into the training set, and the remaining 572 patients were incorporated into the testing set. And no significant differences were found in demographic and clinical variables between the training and testing sets (Table 1).

3.2. Prognostic factors for HCC patients with lung metastasis

The random forest model (Fig. 2) shows that the prognostic variable for HCC patients with lung metastasis was ranked according to variable importance. The top 10 characteristics were selected, which were chemotherapy, age, radiation, tumor size, grade, N stage, surgery, and T stage among the 14 variables. These 8 variables and gender (a common prognostic factor) were selected to build 6 prediction models, logistic regression, random forest, BLS, DNN, SVM, and naïve Bayes, respectively. The distribution of survival and death characteristics was shown in Table 2.

3.3. Development of the 6 prediction models

Based on the screened 8 variables and gender, the 6 prediction models (logistic regression model, random forest, BLS, DNN, SVM, and naïve Bayes) were developed. The multivariate analysis of the logistic regression model was in Table S1, Supplemental Digital Content, http://links.lww.com/MD/H770. The formula for prediction was: Logit $(P) = \ln (P/1-P) = 0.002$ age + 0.249 male + 0.045 T2 + 0.175 T3 + 0.395 T4 + 0.155 TX + 0.204 N1 – 0.128 NX + –0.284 (Grade II) + 0.489 (Grade III/IV) + 0.399 (Grade unknown) + 0.157 (tumor size > 5 cm) + 0.484 (tumor size was unknown) – 1.274 surgery – 2.075 chemotherapy – 1.236 radiation. And a prognostic nomogram was established for the 1-month survival in HCC patients with lung metastasis (Fig. 3). A dynamic nomogram is available at https://jingting.shinyapps.io/DynNomapp/.

3.4. Comparison of the AUC of the 6 prediction models

In the training set, the AUC of the random forest model was significantly higher than the logistic regression model [0.853 (95% CI: 0.832–0.873) vs. 0.771 (95% CI: 0.746–0.796), $P < .001$], there were no significant differences in testing and validation sets between the random forest model and logistic regression model (Table 3). In the validation set, the value of AUC in the logistic regression model was significantly higher than that of the BLS model [0.763 (95% CI: 0.734–0.791) vs. 0.745 (95% CI: 0.716–0.774), $P = .006$], in Table 3. Compared with the SVM model, the logistic regression model had higher values of AUC in the training set [0.771 (95% CI: 0.746–0.796) vs. 0.690 (95% CI: 0.661–0.719), $P < .001$], testing set [0.775 (95% CI: 0.736–0.813) vs. 0.704 (95% CI: 0.661–0.746), $P < .001$], and validation set [0.763 (95% CI: 0.734–0.791) vs. 0.689 (95% CI: 0.657–0.720), $P < .001$]. In Table 3, the AUC of the naïve Bayes model was

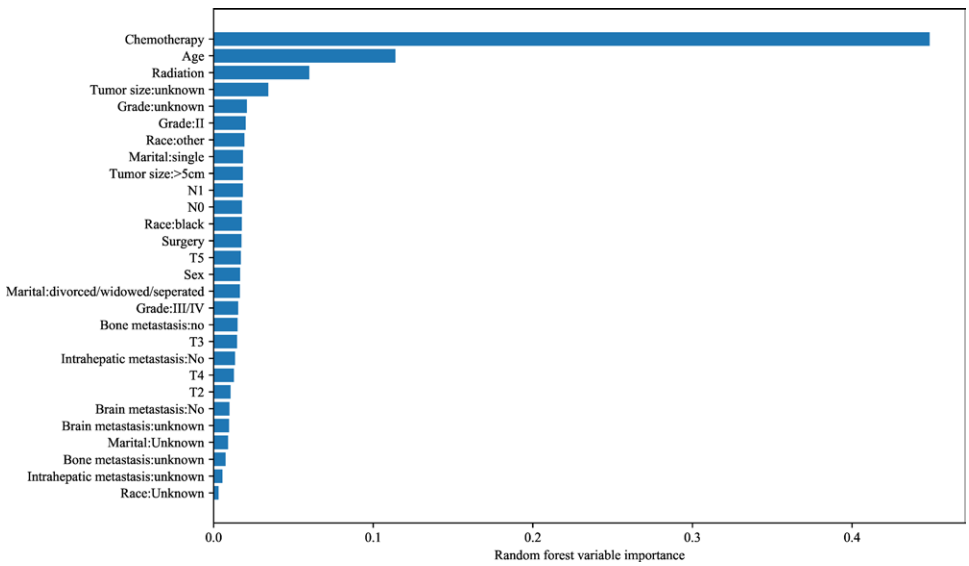

Figure 2. Variable importance from the random forest model.

Table 2

Distribution of survival and death in training set.

Variables	Total (n = 1333)	Groups	
		Survival (n = 643)	Death (n = 690)
Age, mean ± SD	62.57 ± 11.67	62.35 ± 12.03	62.78 ± 11.33
Gender, n (%)			
Female	282 (21.16)	144 (22.40)	138 (20.00)
Male	1051 (78.84)	499 (77.60)	552 (80.00)
Race, n (%)			
White	822 (61.67)	376 (58.48)	446 (64.64)
Black	236 (17.70)	121 (18.82)	115 (16.67)
Other	271 (20.33)	143 (22.24)	128 (18.55)
Unknown	4 (0.30)	3 (0.47)	1 (0.14)
Marital status, n (%)			
Married	596 (44.71)	304 (47.28)	292 (42.32)
Single	357 (26.78)	159 (24.73)	198 (28.70)
Divorced/widowed/separated	315 (23.63)	146 (22.71)	169 (24.49)
Unknown	65 (4.88)	34 (5.29)	31 (4.49)
T stage, n (%)			
T1	258 (19.35)	133 (20.68)	125 (18.12)
T2	112 (8.40)	61 (9.49)	51 (7.39)
T3	478 (35.86)	244 (37.95)	234 (33.91)
T4	178 (13.35)	81 (12.60)	97 (14.06)
TX	307 (23.03)	124 (19.28)	183 (26.52)
N stage, n (%)			
N0	824 (61.82)	418 (65.01)	406 (58.84)
N1	253 (18.98)	112 (17.42)	141 (20.43)
NX	256 (19.20)	113 (17.57)	143 (20.72)
Grade, n (%)			
I	60 (4.50)	38 (5.91)	22 (3.19)
II	149 (11.18)	92 (14.31)	57 (8.26)
III/IV	185 (13.88)	86 (13.37)	99 (14.35)
Unknown	939 (70.44)	427 (66.41)	512 (74.20)
Tumor size, n (%)			
≤5 cm	186 (13.95)	99 (15.40)	87 (12.61)
>5 cm	726 (54.46)	381 (59.25)	345 (50.00)
Unknown	421 (31.58)	163 (25.35)	258 (37.39)
Bone metastasis, n (%)			
No	1062 (79.67)	505 (78.54)	557 (80.72)
Yes	233 (17.48)	121 (18.82)	112 (16.23)
Unknown	38 (2.85)	17 (2.64)	21 (3.04)
Brain metastasis, n (%)			
No	1255 (94.15)	610 (94.87)	645 (93.48)
Yes	30 (2.25)	14 (2.18)	16 (2.32)
Unknown	48 (3.60)	19 (2.95)	29 (4.20)
Intrahepatic metastasis, n (%)			
No	1171 (87.85)	567 (88.18)	604 (87.54)
Yes	141 (10.58)	70 (10.89)	71 (10.29)
Unknown	21 (1.58)	6 (0.93)	15 (2.17)
Surgery, n (%)			
No	1307 (98.05)	621 (96.58)	686 (99.42)
Yes	26 (1.95)	22 (3.42)	4 (0.58)
Chemotherapy, n (%)			
No	886 (66.47)	288 (44.79)	598 (86.67)
Yes	447 (33.53)	355 (55.21)	92 (13.33)
Radiation, n (%)			
No	1217 (91.30)	557 (86.63)	660 (95.65)
Yes	116 (8.70)	86 (13.37)	30 (4.35)

significantly higher than the logistic regression model in the training set [0.771 (95% CI: 0.746–0.796) vs. 0.746 (95% CI: 0.719–0.772), $P < .001$] and testing set [0.775 (95% CI: 0.736–0.813) vs. 0.744 (95% CI: 0.703–0.785), $P = .004$]. The ROC curves of training, testing, and validation sets in the logistic regression model (Fig. 4A), random forest model (Fig. 4B), BLS model (Fig. 4C), DNN model (Fig. 4D), SVM model (Fig. 4E), and naïve Bayes (Fig. 4F) were established. We chose the logistic regression prediction model through a comprehensive comparison of the AUC and interpretability of the 6 models.

3.5. Predictive performance of the 6 prediction models

Specificity, PPV, and negative predictive value of prediction models in training, testing, and validation sets were shown in Table 4. The specificity of the Logistic regression model was 0.912 (95% CI: 0.888–0.936) in the validation set, which was higher than those of BLS [0.662 (95% CI: 0.621–0.702)], DNN [0.633 (95% CI: 0.593–0.674)], SVM [0.630 (95% CI: 0.589–0.671)], and naïve Bayes [0.645 (95% CI: 0.604–0.685)] models. The Logistic regression model obtained a higher value of PPV than BLS, DNN, SVM, and naïve Bayes prediction models. The mean score of 5-fold cross validation

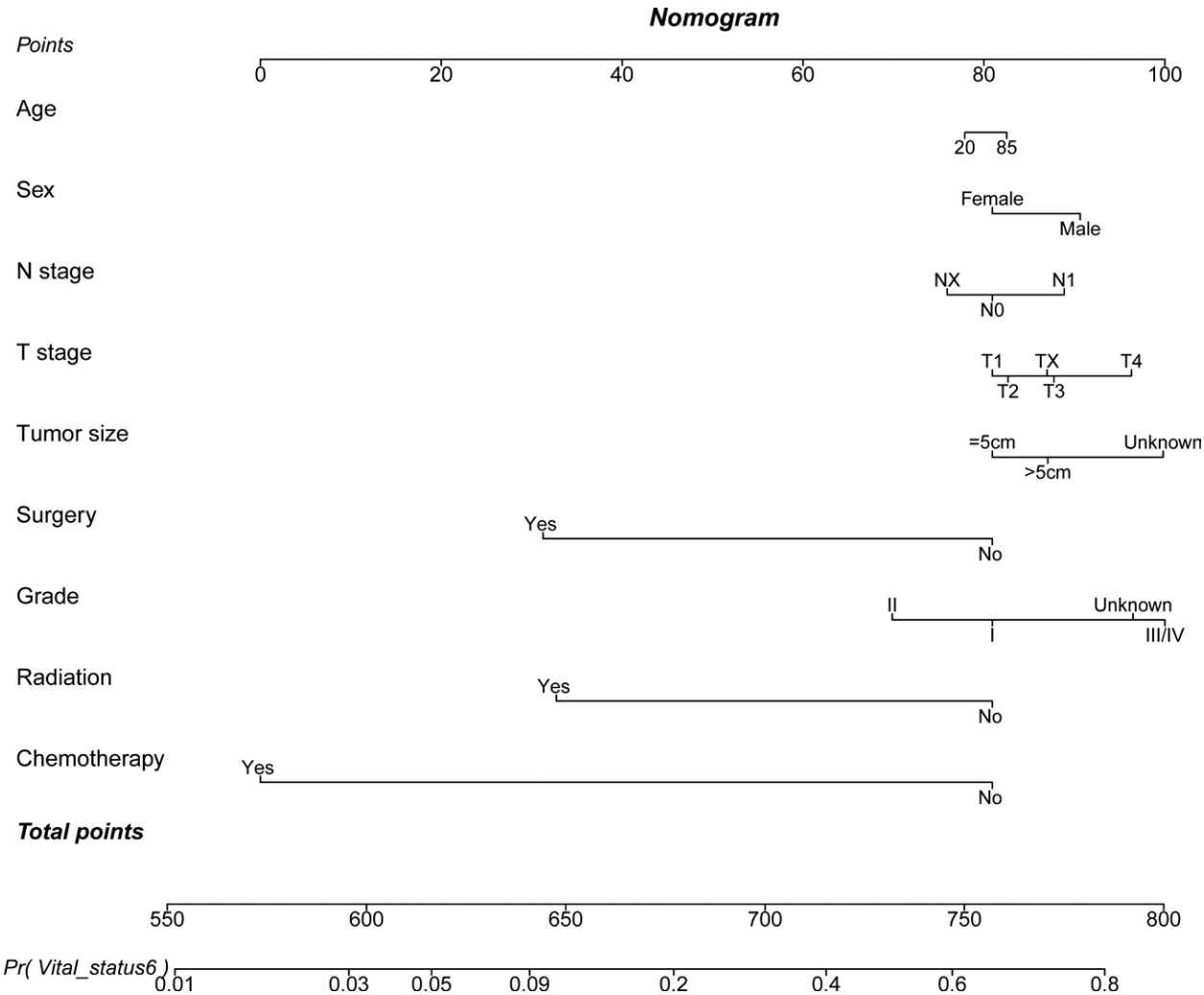

Figure 3. Nomogram for predicting 1-month survival of HCC patients with lung metastasis. HCC = hepatocellular carcinoma.

Table 3

Comparison of AUC in training, testing, and validation sets in the 6 prediction models.

	AUC	Training set	Testing set	Validation set
Model (95% CI)	Logistic	0.771 (0.746–0.796)	0.775 (0.736–0.813)	0.763 (0.734–0.791)
	Random forest	0.853 (0.832–0.873)	0.764 (0.724–0.803)	0.762 (0.734–0.791)
	BLS	0.775 (0.750–0.800)	0.772 (0.733–0.811)	0.745 (0.716–0.774)
	DNN	0.772 (0.746–0.797)	0.762 (0.722–0.801)	0.757 (0.728–0.786)
	SVM	0.690 (0.661–0.719)	0.704 (0.661–0.746)	0.689 (0.657–0.720)
	Naïve Bayes	0.746 (0.719–0.772)	0.744 (0.703–0.785)	0.758 (0.729–0.786)
DeLong test *P*	Random forest vs. Logistic	<.001*	.156	.970
	BLS vs. Logistic	.502	.725	.006*
	DNN vs. Logistic	.854	.066	.271
	SVM vs. Logistic	<.001*	<.001*	<.001*
	Naïve Bayes vs. Logistic	<.001*	.004*	.527

AUC = area under the curve, BLS = broad learning system, CI = confidence interval, DNN = deep neural network, SVM = support vector machine.
*$P < .05$.

in the Logistic regression model (0.72) was better than the others (Table 5).

The results of the Hosmer-Lemeshow test in the training set was $\chi^2 = 6.051$ ($P = .642$) and in the validation set was $\chi^2 = 8.975$ ($P = .344$), suggesting that the overall fitting degree of the model was great. This study chose the logistic regression prediction model because of its good prediction.

4. Discussion

HCC is a highly aggressive tumor, it is prone to extrahepatic metastasis, occurring in 14.0% to 36.7% of patients.[23] The lung is the most common distant metastasis and the prognosis of HCC patients with lung metastasis is extremely poor.[24] In our study, we established and validated 6 prediction models for predicting 1-month all-cause mortality for HCC patients

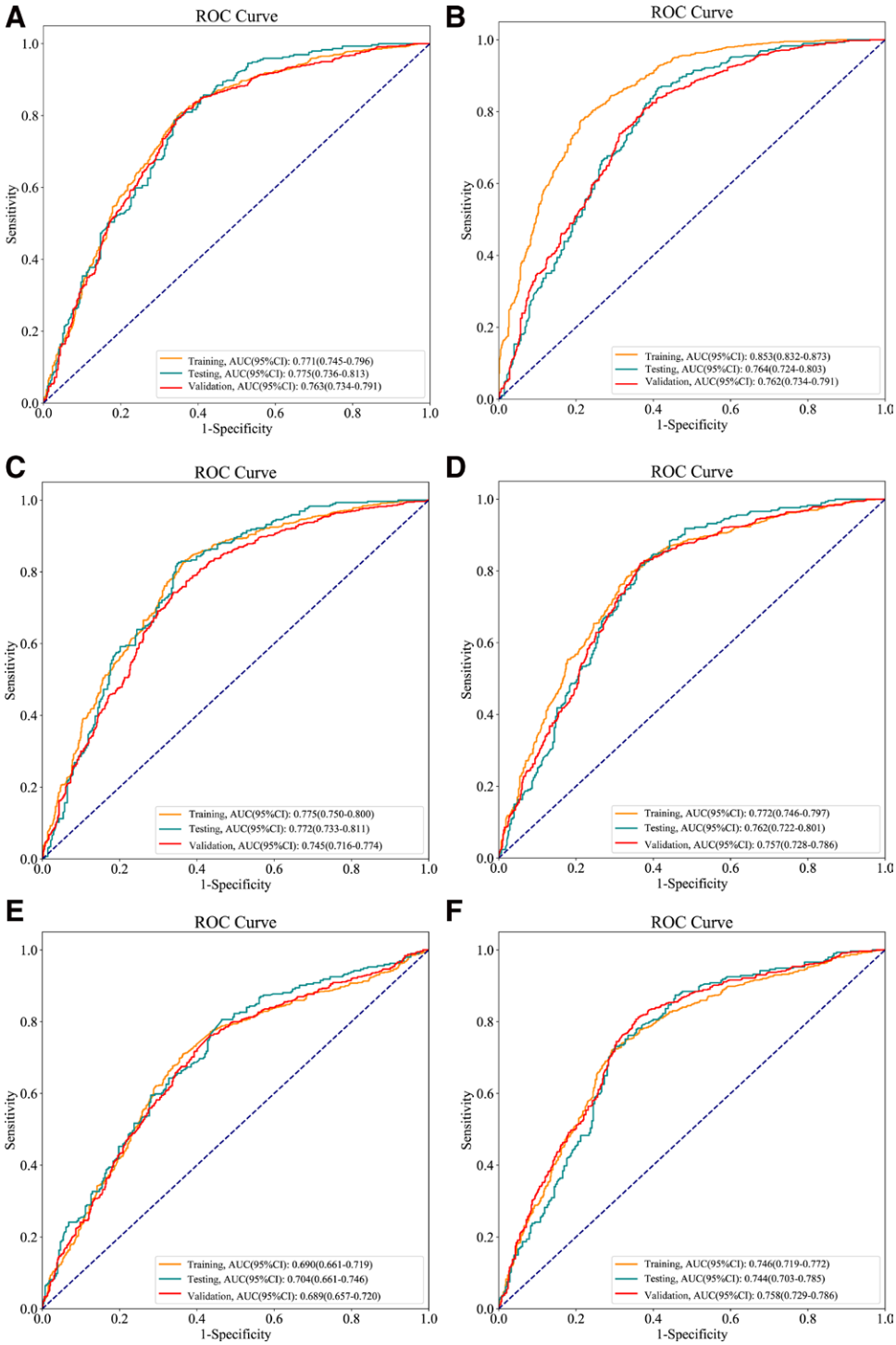

Figure 4. ROC curve of the training set, testing set, and validation set in logistic regression model (A); random forest model (B); BLS model (C); DNN model (D); SVM model (E); and naïve Bayes (F). BLS = broad learning system, DNN = deep neural network, ROC = the receiver operating characteristic, SVM = support vector machine.

with lung metastasis. We finally chose the more optimized logistic regression model due to prediction performance and model visualization. By obtaining the data of 9 easily accessible variables (age, N stage, T stage, tumor size, surgery, grade, radiation, chemotherapy, and gender) on the dynamic nomogram of each HCC patient with lung metastasis, the total score could be calculated (https://jingting.shinyapps.io/DynNomapp/). Then, the all-cause mortality of HCC patients with lung metastasis could be predicted, which could provide guidance for further clinical management. Moreover, the logistic regression model demonstrated excellent performance in the prediction of HCC patients with lung metastasis, which may make the individualized clinical decision and surveillance more accurate.

For HCC patients with extrahepatic metastasis, the disease progresses rapidly, and the severe patients may die quickly. Clinically, an appropriate prognostic judgment index or tool is needed to guide the individualized treatment of HCC patients

Table 4

Performance of prediction models in training, testing, and validation sets.

		Model (95% CI)					
		Logistic	Random forest	BLS	DNN	SVM	Naïve Bayes
Training set	Specificity	0.653 (0.616–0.690)	0.788 (0.757–0.820)	0.628 (0.591–0.666)	0.656 (0.620–0.693)	0.636 (0.599–0.673)	0.708 (0.672–0.743)
	PPV	0.712 (0.680–0.744)	0.797 (0.767–0.827)	0.707 (0.675–0.738)	0.714 (0.682–0.746)	0.677 (0.643–0.711)	0.726 (0.692–0.759)
	NPV	0.753 (0.717–0.788)	0.765 (0.732–0.797)	0.780 (0.744–0.816)	0.752 (0.717–0.788)	0.672 (0.634–0.709)	0.702 (0.667–0.737)
Testing set	Specificity	0.608 (0.551–0.665)	0.694 (0.640–0.748)	0.647 (0.591–0.704)	0.622 (0.565–0.679)	0.579 (0.521–0.637)	0.683 (0.629–0.738)
	PPV	0.688 (0.639–0.736)	0.704 (0.651–0.757)	0.712 (0.664–0.760)	0.697 (0.649–0.746)	0.639 (0.587–0.691)	0.711 (0.660–0.762)
	NPV	0.758 (0.702–0.814)	0.677 (0.623–0.731)	0.776 (0.722–0.830)	0.769 (0.714–0.824)	0.649 (0.590–0.709)	0.709 (0.655–0.763)
Validation set	Specificity	0.912 (0.888–0.936)	0.949 (0.931–0.968)	0.662 (0.621–0.702)	0.633 (0.593–0.674)	0.630 (0.589–0.671)	0.645 (0.604–0.685)
	PPV	0.773 (0.716–0.830)	0.767 (0.690–0.844)	0.698 (0.661–0.735)	0.702 (0.667–0.737)	0.658 (0.619–0.697)	0.705 (0.670–0.740)
	NPV	0.548 (0.515–0.581)	0.517 (0.486–0.549)	0.710 (0.670–0.750)	0.771 (0.732–0.811)	0.649 (0.608–0.690)	0.761 (0.721–0.800)

BLS = broad learning system, CI = confidence interval, DNN = deep neural network, NPV = negative predictive value, PPV = positive predictive value, SVM = support vector machine.

Table 5

Five-fold cross validation of the 6 prediction models.

	Logistic	Random forest	BLS	DNN	SVM	Naïve Bayes
Mean	0.72	0.73	0.72	0.72	0.53	0.70
Standard deviation	0.01	0.01	0.006	0.01	0.005	0.017

BLS = broad learning system, DNN = deep neural network, SVM = support vector machine.

with lung metastasis. Recent studies have constructed several prognostic assessment tools for predicting bone metastases,[10] brain metastases,[25] and lymph node metastases,[26] these models possessed good predictive capabilities. Hu et al[7] established the nomograms to predict the prognosis of HCC patients with bone metastasis with AUCs of 6- and 9-months survival prediction were 0.698 and 0.770 in the validation set, respectively. The model discrimination was at an acceptable level, it was still not ideal enough. In practice, its accuracy may reduce in different patients to some extent, so the use efficiency in practice may be lower than at present.[9] The prognosis model we established for HCC patients with lung metastasis had a higher AUC value of 0.775 in the testing set and 0.763 in the validation set, which may be more referenced in practical applications. In addition, a prognostic nomogram including age, T stage, surgical approach, and chemotherapy for HCC patients with lung metastasis was conducted by Ye et al[13] using SEER data. The calibration curve showed that the predicted survival rate was close to the actual survival rate, which was similar to the consistent results of our model. The consistency between the predicted and the actual data was better, which also verified good repeatability of the nomogram prediction model.[9] The study of Ye et al[13] only included the 4 variables mentioned above into the nomogram, and our nomogram contained 9 variables (age, gender, N stage, T stage, tumor size, surgery, grade, radiation, and chemotherapy). Although the number of variables we included in the prediction model was not the same, we established 6 models and selected the more optimized model based on the interpretability and predictive performance, which was the logistic regression model. Our study provided reference clinically, prognostic judgment tools based on big data were still needed for lung metastasis in HCC patients.

It has been reported that several demographic characteristics were independent factors affecting the prognosis of HCC patients with distant metastases.[13,27] Studies have shown that men over 45 and women over 55 have become an important factor in the death of HCC,[13] and young patients had relatively better survival compared with elderly patients.[27] Although the variable screening based on the variable importance did not filter gender, we still added the gender characteristic to the nomogram. The American Joint Committee on Cancer staging system is widely accepted and used. Higher T and N staging and higher tumor grades were associated with poor prognosis, and it was no exception in HCC patients with lung metastases.[28] T stage not only contained information about tumor size, but also included the number of vascular invasion and primary tumor lesions, which were related to the prognosis of the patient.[29] Patients with a lower degree of differentiation had a worse prognosis.[27] HCC patients with extrahepatic metastasis such as lung metastasis were considered to advanced. It was recommended that systemic chemotherapy had a positive effect on the survival of patients with advanced HCC.[30,31] Some studies also believed that the survival of HCC patients may be related to hepatectomy and radiotherapy of lung metastasis.[32,33] A multicenter study in Korea showed that liver resection provides a survival benefit compared with nonsurgical treatment for patients.[30] Surgical resection should be considered for select individuals with intermediate- or advanced-stage disease.[33]

In this study, 6 prediction models were established based on the selected prognostic factors, including logistic regression, random forest, BLS, DNN, SVM, and naïve Bayes models. Finally, the logistic regression model was selected from the 6 prediction models. A dynamic nomogram with a good discriminating ability of lung metastasis in HCC patients was generated. The prognostic factors used in the nomogram were simple, easy to obtain, and operable. But there were still certain limitations. On the 1 hand, although we have internally and externally verified the prediction models, it still needs external data support and verification due to the lack of sufficient external data in our study, so the general applicability of this research needs to be verified. On the other hand, this study was retrospective. There were too many missing serum alpha-fetoprotein data in the database to be analyzed even though serum alpha-fetoprotein expression is quite important for HCC patients. The data such as biomarkers,[34] drug use,[30] and genes[35] that may be related to the prognosis of HCC patients with lung metastasis were not available in the SEER database. Further studies with large samples and multi-center prospective data were required.

5. Conclusion

A logistic regression prediction model predicting 1-month all-cause mortality we created may be individual and convenient

using 9 easily acquired variables, including age, gender, N stage, T stage, tumor size, surgery, grade, radiation, and chemotherapy for HCC patients with lung metastasis. Based on the dynamic nomogram, clinicians could estimate the risk of a patient's death based on individualized prediction scores, clinicians could more accurately balance the benefits and risks, and formulate a reasonable long-term treatment strategy.

Author contributions

SC and JL designed the study. SC wrote the manuscript. XL, YL, XL, YH and JZ collected, analyzed and interpreted the data. JL critically reviewed, edited and approved the manuscript. All authors read and approved the final manuscript.

Conceptualization: Shicheng Chen, Jun Li.
Data curation: Xiaowen Li, Yichao Liang, Xinyu Lu, Yingyi Huang, Jiajia Zhu.
Formal analysis: Xiaowen Li, Yichao Liang, Xinyu Lu, Yingyi Huang, Jiajia Zhu.
Funding acquisition: Jun Li.
Investigation: Xiaowen Li, Yichao Liang, Xinyu Lu, Yingyi Huang, Jiajia Zhu.
Methodology: Xiaowen Li, Yichao Liang, Xinyu Lu, Yingyi Huang, Jiajia Zhu.
Project administration: Jun Li.
Writing – original draft: Shicheng Chen.
Writing – review & editing: Shicheng Chen, Jun Li.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel R, Torre L, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68:394–424.
[2] Hartke J, Johnson M, Ghabril M. The diagnosis and treatment of hepatocellular carcinoma. Semin Diagn Pathol. 2017;34:153–9.
[3] Kokudo T, Hasegawa K, Matsuyama Y, et al. Survival benefit of liver resection for hepatocellular carcinoma associated with portal vein invasion. J Hepatol. 2016;938–43.
[4] Kokudo T, Hasegawa K, Yamamoto S, et al. Surgical treatment of hepatocellular carcinoma associated with hepatic vein tumor thrombosis. J Hepatol. 2014.
[5] Wang YK, Bi XY, Li ZY, Zhao H, Cai JQ. A new prognostic score system of hepatocellular carcinoma following hepatectomy. Zhonghua zhong liu za zhi. 2017;39:903–9.
[6] Elmoghazy W, Ahmed K, Vijay A, et al. Hepatocellular carcinoma in a rapidly growing community: epidemiology, clinico-pathology and predictors of extrahepatic metastasis. Arab J Gastroenterol. 2019;20:38–43.
[7] Komatsu S, Kido M, Tanaka M, et al. Clinical significance of hepatectomy for hepatocellular carcinoma associated with extrahepatic metastases. Dig Surg. 2020;37:411–9.
[8] Fang T, Lv H, Lv G, et al. Tumor-derived exosomal miR-1247-3p induces cancer-associated fibroblast activation to foster lung metastasis of liver cancer. Nat Commun. 2018;9:191.
[9] Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. Lancet Oncol. 2015;16:e173–80.
[10] Hu C, Yang J, Huang Z, et al. Diagnostic and prognostic nomograms for bone metastasis in hepatocellular carcinoma. BMC Cancer. 2020;20:1–11.
[11] Lee CH, Chang CJ, Lin YJ, et al. Nomogram predicting extrahepatic metastasis of hepatocellular carcinoma based on commonly available clinical data. JGH Open. 2019;3:38–45.
[12] Chen QF, Huang T, Shen L, Li W. Predictive value of a nomogram for hepatocellular carcinoma with brain metastasis at initial diagnosis: a population-based study. PLoS One. 2019;14:e0209293.
[13] Ye G, Wang L, Hu Z, et al. Risk and prognostic nomograms for hepatocellular carcinoma with newly-diagnosed pulmonary metastasis using SEER data. Peer J. 2019;7:e7496.
[14] Gloeckler Ries LA, Reichman ME, Lewis DR, Hankey BF, Edwards BK. Cancer survival and incidence from the Surveillance, Epidemiology, and End Results (SEER) program. Oncologist. 2003;8:541–52.
[15] Ishwaran H. Variable importance in binary regression trees and forests. Electr J Stat. 2007.
[16] Frager S, Schwartz J. Hepatocellular carcinoma: epidemiology, screening, and assessment of hepatic reserve. Curr Oncol. 2020;27:S138–43.
[17] Breiman. Random forests. Mach Learn. 2001; 45:5–32.
[18] Chen C, Liu Z. Broad learning system: an effective and efficient incremental learning system without the need for deep architecture. IEEE Trans Neural Netw Learn Syst. 2018;29:10–24.
[19] Crabbé A, Cahy T, Somers B, Verbeke L, Van Coillie F. Neural network MLP classifier. 2020.
[20] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2007, 2(3, article 27).
[21] Gondy LA, Thomas C, Bayes N. Programs for machine learning. Adv Neural Inf Process Syst. 1993;79:937–44.
[22] Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000;56:337–44.
[23] Natsuizaka M, Omura T, Akaike T, et al. Clinical features of hepatocellular carcinoma with extrahepatic metastases. J Gastroenterol Hepatol. 2005;20:1781–7.
[24] Hu Z, Huang P, Zhou Z, et al. Aggressive intrahepatic therapies for synchronous hepatocellular carcinoma with pulmonary metastasis. Clin Transl Oncol. 2018;20:729–39.
[25] Chen Q-F, Huang T, Shen L, Li W. Predictive value of a nomogram for hepatocellular carcinoma with brain metastasis at initial diagnosis: a population-based study. PLoS One. 2019;14:e02093.
[26] Wee CW, Kim K, Chie EK, Yu SJ, Kim YJ, Yoon JH. Prognostic stratification and nomogram for survival prediction in hepatocellular carcinoma patients treated with radiotherapy for lymph node metastasis. Br J Radiol. 2016;89:20160383.
[27] Zhan H, Zhao X, Lu Z, Yao Y, Zhang X. Correlation and survival analysis of distant metastasis site and prognosis in patients with hepatocellular carcinoma. Front Oncol. 2021;11:652768.
[28] Beumer B, Buettner S, Galjart B, et al. Systematic review and meta-analysis of validated prognostic models for resected hepatocellular carcinoma patients. Eur J Surg Oncol. 2021.
[29] Ca'Granda ON, Alta AdRN , Civico-Di Cristina-Benfratelli S. Selection of treatment modalities for hepatocellular carcinoma at stages T1 and T2: a preliminary analysis based on the Surveillance, Epidemiology, and End Results registry database. J BUON. 2018;23:611–21.
[30] Lin S, Hoffmann K, Schemmer P. Treatment of hepatocellular carcinoma: a systematic review. Liver Cancer. 2012;1:144–58.
[31] Wada Y, Takami Y, Matsushima H, et al. The safety and efficacy of combination therapy of sorafenib and radiotherapy for advanced hepatocellular carcinoma: a retrospective study. Intern Med. 2018;57:1345–53.
[32] Kim H, Ahn S, Hong S, et al. Survival benefit of liver resection for Barcelona Clinic Liver Cancer stage B hepatocellular carcinoma. Br J Surg. 2017;104:1045–52.
[33] Yang JD, Hainaut P, Gores GJ, Amadou A, Plymoth A, Roberts LR. A global view of hepatocellular carcinoma: trends, risk, prevention and management. Nat Rev Gastroenterol Hepatol. 2019;16:589–604.
[34] Ge Y, Mu W, Ba Q, et al. Hepatocellular carcinoma-derived exosomes in organotropic metastasis, recurrence and early diagnosis application. Cancer Lett. 2020;477:41–8.
[35] Shin D, Jo J, Kim S, et al. Midkine is a potential therapeutic target of tumorigenesis, angiogenesis, and metastasis in non-small cell lung cancer. Cancers. 2020;12.