Am J Cancer Res 2021;11(2):576-589
www.ajcr.us /ISSN:2156-6976/ajcr0119185

Original Article

Deep learning radiomics model accurately predicts hepatocellular carcinoma occurrence in chronic hepatitis B patients: a five-year follow-up

Jieyang Jin[1,2*], Zhao Yao[3*], Ting Zhang[1,2*], Jie Zeng[1,2], Lili Wu[1,2], Manli Wu[1,2], Jinfen Wang[1,2], Yuanyuan Wang[3,4], Jinhua Yu[3,4#], Rongqin Zheng[1,2#]

[1]Department of Ultrasound, The Third Affiliated Hospital of Sun Yat-Sen University, 600 Tianhe Road, Guangzhou, China; [2]Guangdong Key Laboratory of Liver Disease Research, The Third Affiliated Hospital of Sun Yat-Sen University, 600 Tianhe Road, Guangzhou, China; [3]Department of Electronic Engineering, Fudan University, Shanghai, China; [4]The Key Laboratory of Medical Imaging Computing and Computer Assisted Intervention of Shanghai, China. [*]Equal contributors and co-first authors. [#]Equal contributors.

Received July 29, 2020; Accepted November 18, 2020; Epub February 1, 2021; Published February 15, 2021

Abstract: An early and accurate prediction of hepatocellular carcinoma (HCC) is beneficial for individualized treatment and follow-up of chronic hepatitis B (CHB) patients. We aimed to establish a prediction model for HCC by radiomics analysis in CHB patients and compare performance with liver stiffness measurement (LSM) and other clinical prognostic scores. Initially, 1215 patients were included and finally 434 CHB patients with 5-year follow-up were enrolled, 96.3% of them underwent liver biopsy. Deep learning radiomics analysis was performed on 2170 two-dimensional shear wave elastography (2D-SWE) and corresponding B-mode ultrasound (US) images. These high-throughput imaging features were also combined with low-dimensional serological clinical data by deep learning radiomics to establish different HCC prediction models and to overcome challenges of an unbalanced sample. The best model which is simple with high accuracy was selected. Prediction performance of the selected model was compared with LSM and other clinical prognostic scores. During 5-year follow-up, 32 (7.4%) of 434 patients developed HCC. The best prediction model was HCC-R, which included 2D-SWE and B-mode US images, sex and age. This model showed a high predictive value with areas under the receiver operating characteristic curve (AUCs) of 0.981, 0.942 and 0.900 in training, validation and testing cohorts for predicting 5-year prognosis of HCC. These predictive values were significantly higher than that of LSM (AUC: 0.676~0.784, $p < 0.05$) and better than that of other clinical prognostic scores (AUC: 0.544~0.869). HCC-R radiomics model based on 2D-SWE and B-mode US images, sex and age comprehensively reflected biomechanical and morphological information of patients and can accurately predict HCC occurrence; thus, this model has great value for treatment and follow-up of CHB patients.

Keywords: Radiomics, prediction, hepatocellular carcinoma, elastography, ultrasound

Introduction

Hepatocellular carcinoma (HCC) is ranked as the sixth most common neoplasm and the third leading cause of cancer-related death [1]. Most HCC cases (80%) occur in sub-Saharan Africa and eastern Asia, especially China, where the main risk factors are chronic hepatitis B (CHB) infection. China is a large country with more than 90 million people infected with hepatitis B [2]. It is of great clinical significance to detect HCC in the early stage to promote prognosis and the quality of disease prevention and therapeutic effect of patients. Therefore, an accurate evaluation of the risk of HCC can enable clinicians to start treatment as early as possible. Furthermore, it is necessary to conduct individualized follow-up for high-risk patients with effective antiviral therapy, and on the other hand, avoid radical treatment for low-risk patients to reduce the waste of medical resources.

Current clinical methods used serological indicators such as “Guide with Age, Gender, hepatitis B virus deoxyribonucleic acid (HBV-DNA), Core promoter mutations and Cirrhosis” (GAG-HCC) score [3], “Cirrhosis, Age, Male sex, and

Diabetes mellitus" (CAMD) score [4], "Hepatocellular Carcinoma after Hepatitis B e-antigen Seroclearance" (HCC-ESC) score [5] and "Age, Albumin, Bilirubin, HBV-DNA, and Clinical Cirrhosis" (CU-HCC) score [6] have been applied to predict HCC development but with unsatisfactory effective for HCC development at 5-year with areas under receiver operating characteristic curve (AUC) of 0.699~0.757. At present, there is a lack of highly accurate, truly effective, recognizable, operable and convenient methods to predict HCC development.

Recently, some studies have shown that the accuracy of HCC prediction can be improved by using transient elastography (TE) to measure the liver stiffness combined with clinical biochemical indicators, with the AUC for HCC prediction at 5-year of 0.759~0.806 [7]. However, the application of TE has several limitations, such as the inability to measure in patients with ascites [8]. Two-dimensional shear wave elastography (2D-SWE) has the ability to overcome the limitations of TE, and its accuracy in assessing the degree of fibrosis appeared similar to that of TE. Furthermore, 2D-SWE showed better performances than TE in the assessment of liver fibrosis especially in cirrhotic (F4) and advanced fibrotic (≥ F3) patients with higher sensitivity and specificity [9]. Moreover, 2D-SWE provided both elastography and B-mode ultrasound (US) images in real time, making it possible to perform radiomics analysis for clinical diagnosis and prognosis.

Equipped with machine learning technique, radiomics can extract, quantify and select high-throughput image features from medical images, and thus has the potential to uncover disease characteristics that fail to be caught by naked eyes [10]. Our previous study [11] showed that deep learning radiomics of 2D-SWE significantly improved performance of noninvasive liver fibrosis detection, which is similar to the histopathological diagnosis with AUC for F3 and F4 at 0.97 to 1.00. The development of CHB patients into HCC is a long process. In the long-term US follow-up of CHB patients, there must be important image features indicating quantitative change to qualitative change, that is, from the progress of hepatitis to HCC.

Therefore, the purpose of this study was to establish a prediction model based on deep learning radiomics analysis on 2D-SWE images with corresponding B-mode US images, clinical information and serological data for predicting the occurrence of HCC in CHB patients, and to compare the model with existing clinical prognostic scores.

Patients and methods

Study design

This was a retrospective study. Enrolled patients were divided into training, validation and testing cohort. Each patient has different sets of features: 2D-SWE images, B-mode US images, serological data, and demographic features. In training cohort, different models with one or more sets of features are trained to predict the occurrence of HCC. We compared all the models with each other and selected the best one as the new radiomics prediction model, HCC-R. The performance of HCC-R was validated and compared with that of liver stiffness measurement (LSM) and many other clinical prognostic scores in training, validation and testing cohorts.

Patients

We included CHB patients who underwent 2D-SWE examination in the hospital between 1 April 2011 and 1 September 2017. Inclusion criteria were as follows: (1) older than 18 years old; (2) hepatitis B surface antigen positive for more than 6 months; (3) with available B-mode US and 2D-SWE examination and images; (4) CHB patients underwent liver biopsy or cirrhotic patients diagnosed by clinical diagnostic criteria [3]. Exclusion criteria were as follows: (1) follow-up period less than 5 years; (2) pre-existing HCC or other hepatic metastases cancer; (3) coinfection with human immunodeficiency virus or other viral hepatitis; (4) combined with alcoholic liver disease, autoimmune liver disease or another liver diseases; (5) missing important serological data; (6) unsuccessful B-mode US or 2D-SWE measurements or unavailable images. All the patients were given standard medical treatments [12]. This study was approved by the local medical ethics committee of the hospital. The study protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki as reflected in a priori approval by the institution's human research committee. Written informed consent was obtained from each patient included in the study.

Clinical and laboratory data collection

At baseline, clinical characteristics [sex, age, height, weight, body mass index (BMI) and antiviral therapy history] and serological data [aspartate aminotransferase (AST), alanine aminotransferase (ALT), albumin (ALB), total bilirubin (TB), gamma-glutamyl transpeptidase (GGT), alkaline phosphatase (ALP), platelets (PLT), prothrombin time (PT), serum α-fetoprotein (AFP) and HBV-DNA] of each patient were collected.

Clinical prognostic scores of hepatocellular carcinoma

Clinical prognostic scores included GAG-HCC score [3], CAMD score (Supplementary Table 1) [4], HCC-ESC score [5] and CU-HCC score (Supplementary Table 2) [6] (for more details see Supplementary Materials).

Ultrasonography and two-dimensional shear wave elastography

US and 2D-SWE examinations were performed by two operators (both with more than 2 years of US experience and 500 2D-SWE examinations), using the Aixplorer US system (SuperSonic Imagine, Aix-en-Provence, France) with a convex broadband probe (SC6-1, 1-6 MHz). The 2D-SWE protocol has been described in our previous studies [13, 14], which is also recommended by European Federation of Societies for Ultrasound in Medicine and Biology guideline [8] (for more details see Supplementary Materials). Five independent 2D-SWE measurements were taken, and the median value of five 2D-SWE measurements was recorded as LSM for statistical analysis. Measurements were considered failed if little or no signal was obtained in the region of interest (ROI). Five 2D-SWE images with corresponding liver parenchyma images from each patient were obtained and stored.

Liver biopsy

Liver biopsy with US guidance was performed in the right lobe of liver within 3 days of the US and 2D-SWE examinations by using a 16 or 18 G needle (Bard Magnum, GA, USA). All the biopsy specimens were analyzed by two liver pathologists (with more than 10 years of experience). Unqualified samples, including portal tracts less than 6 or length less than 15 mm, were excluded. Liver fibrosis was staged according to METAVIR scoring system [15] as follows: F0, no fibrosis; F1, portal fibrosis without septa; F2, portal fibrosis and few septa; F3, numerous septa without cirrhosis; F4, cirrhosis. Necro-inflammatory activity of liver was graded according to METAVIR scoring system as follows: A0 for none; A1 for mild; A2 for moderate; A3 for severe. Liver steatosis was graded based on Brunt scoring system as follows: S0, minimal steatosis (< 5%); S1, mild steatosis (5%-33%); S2, moderate steatosis (> 33%-66%); S3, severe steatosis (> 66%).

Follow-up and diagnosis of hepatocellular carcinoma

Patients were followed up for five year until HCC occurrences, death or the end of the study. Liver biopsy or contrast-enhanced imaging (contrast-enhanced ultrasound, multiphasic computed tomography or dynamic contrast-enhanced magnetic resonance imaging) were performed if suspected during the US or clinical examination. HCC was diagnosed according to the latest European Association for the Study of the Liver (EASL) guideline [16] based on pathological confirmation or contrast-enhanced imaging criteria for nodule(s) ≥ 1 cm in diameter.

Radiomics analysis and establishment of prediction model of hepatocellular carcinoma

The HCC and non-HCC patients were divided at a ratio of 3:1:1 respectively, and 3/5 patients, 1/5 patients and 1/5 patients were used as training, validation and testing cohorts. For radiomics analysis, two square patches were first manually extracted containing 2D-SWE and B-mode US ROIs (**Figure 1A**), and then the ROIs were resized to the same size and normalized. By back propagation, the ROIs were fed into network and the parameters of the model were updated automatically [17]. After the training process was complete, the network was considered as an encoder that could convert high-throughput image features into low-dimensional features. We selected the penultimate fully connected (FC) layer outputs as deep features. Sparse representation and support vector machine (SVM) were finally used for feature reduction and classification, respectively [18] (**Figure 1B**) (for more details see Supplementary Table 3).

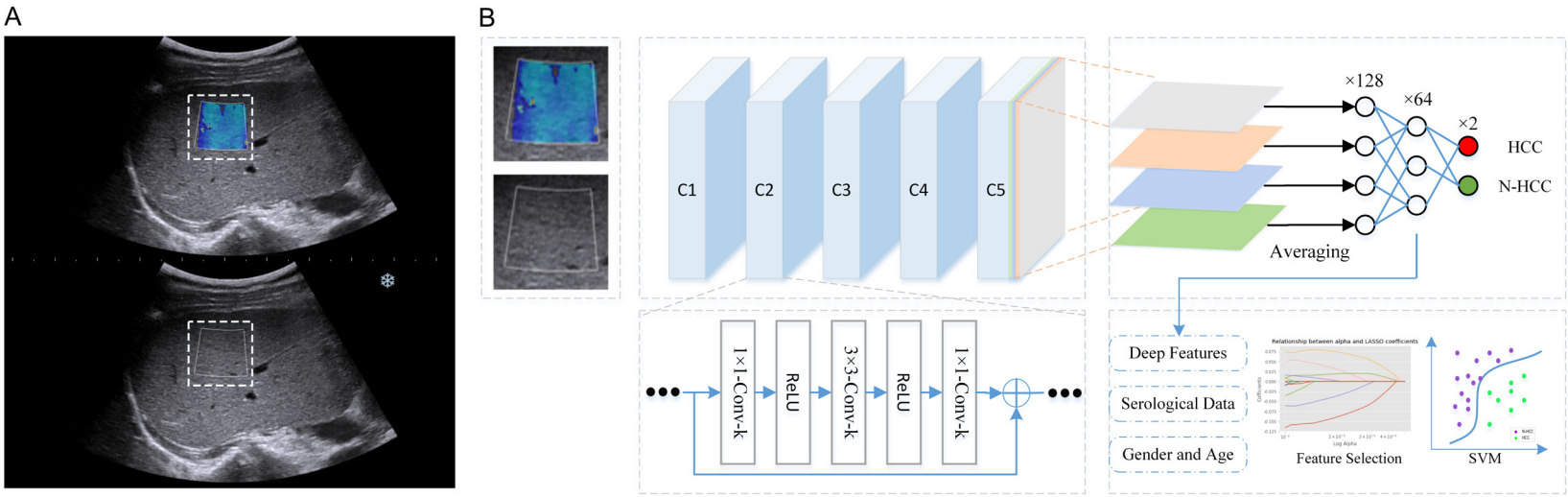

Figure 1. Illustration of selection of ROIs and radiomics analysis flow chart. A. 2D-SWE ROI and corresponding B-mode US ROI (white dotted area). B. Radiomics analysis flow chart. C1-C5 represents the residual block. 2D-SWE ROIs and B-mode US ROIs extracted from training cohort optimize model parameters automatically. The network features combined with clinical features then used to train SVM model.

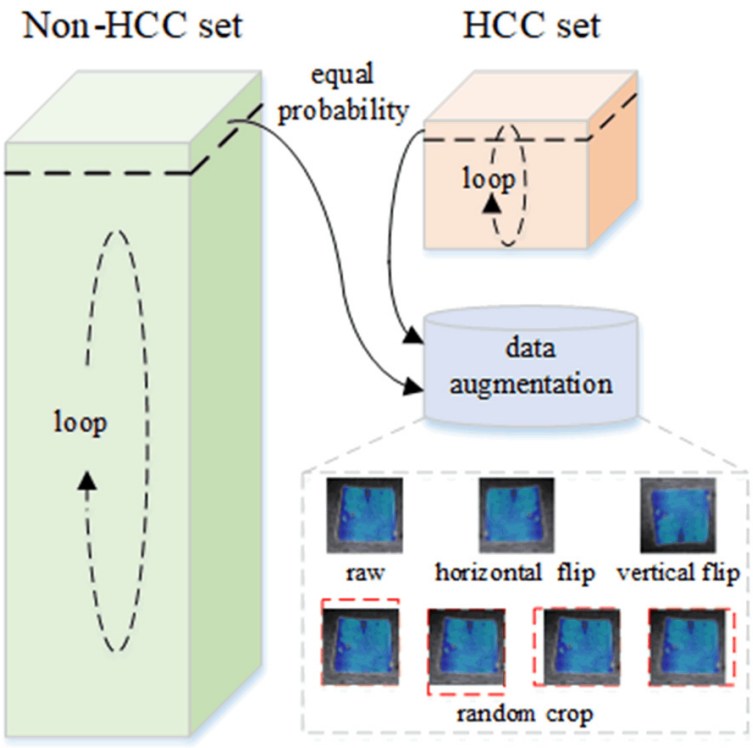

Figure 2. The flow chart of data balancing and data augmentation. The loop-based data generator acquired data from Non-HCC set and HCC set with equal probability for each training step, and then the data was fed into the data augmentation module.

Integrated modeling of image features and low-dimensional clinical information

The trained network was an encoder and the outputs of the penultimate FC layer were used as image features. To integrate the image features and clinical information, we first encoded nonnumeric clinical information (for example, gender was encoded as 0 or 1) and then normalized them. There were three types of features including image features (2D-SWE image features and B-mode US image features), serological features, and demographic features. By combining different sources of features, ten sets of features were obtained and modeled. Then, the prediction performance of these models was evaluated.

Data augmentation of very asymmetric sample data

The challenges in training cohort was the insufficient and unbalanced data. Due to the insufficiency of sample, a transfer learning strategy was adopted to alleviate the problem of overfitting [19]. In terms of data, online data augmentation was used, which added random vertical and horizontal flip modules and random crop modules to the data stream. In terms of model, the L^2 regularization loss was added to control the complexity of models. In terms of the unbalanced data, a loop-based training procedure was used when optimizing model parameters. Instead of randomly using samples from training cohort, we first divided training cohort into HCC set and non-HCC set, and used samples to optimize model parameters from the two sets with equal probability for each training step. The loop speed of two sets was inversely proportional to the number of samples in the set. Combined with the above online data augmentation, the sample diversity can be guaranteed to the greatest extent (for more details see Supplementary Materials). The flow chart of data balancing and data augmentation is shown in **Figure 2**.

Comparison of HCC-R and other methods and statistical analysis

Statistical analysis was performed using Statistical Package for Social Science (version 13.0, SPSS Inc.) or MedCalc software (version 11.2, MedCalc Software). Continuous data were expressed as means ± standard deviation (SD) or medians with interquartile range (IQR) and were compared by Student's t-test or Mann-Whitney U test. Categorical data were expressed as numbers with percentages and were compared by chi-square test or Fisher's exact test. Univariate and multivariate logistic regression analyses were used to assess associations between individual variables and clinical outcomes.

The performance of HCC-R was compared with that of other methods, including LSM and different clinical prognostic scores, in all the train-

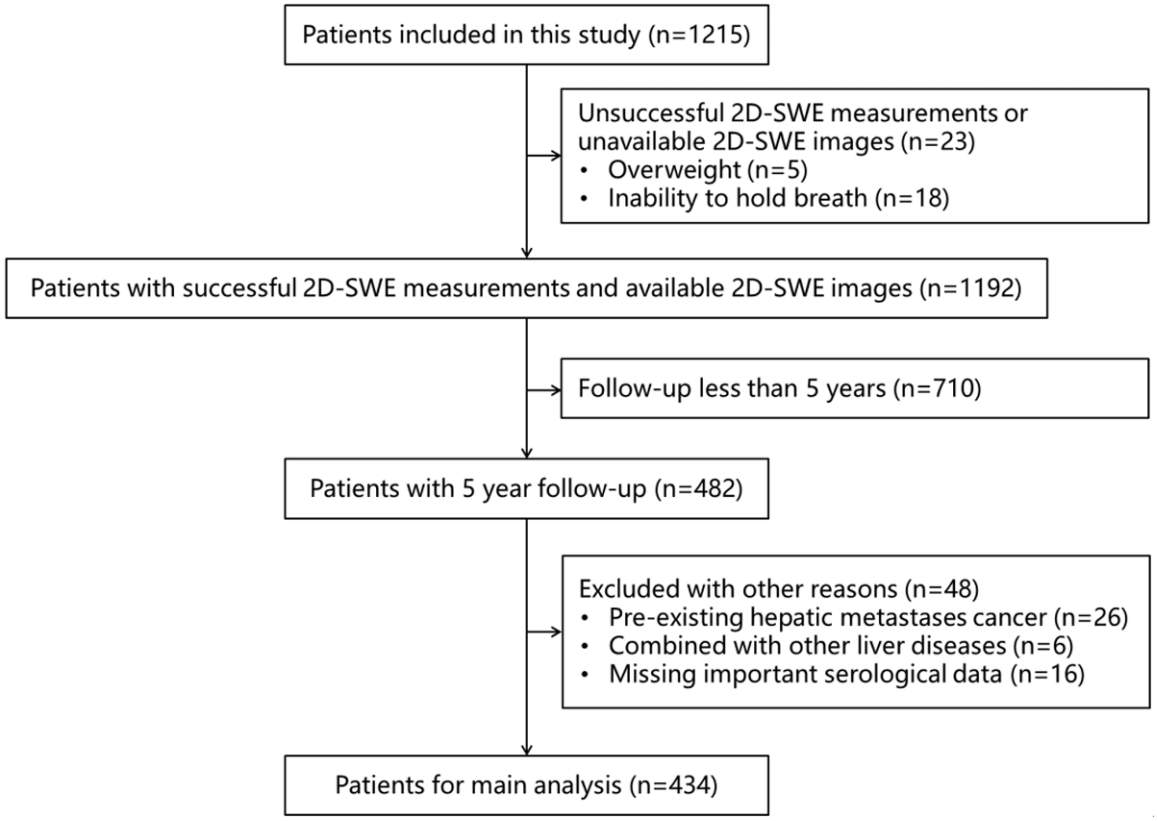

Figure 3. The results of patient enrolments in this study.

ing, validation and testing cohorts. To compare the predictive value of HCC-R with that of other clinical prognostic methods, we calculated AUCs and compared them using Delong test. The sensitivity, specificity, positive predictive values, negative predictive values, positive likelihood ratio and negative likelihood ratio of maximizing the Youden index on the estimated receiver operating characteristic (ROC) were calculated. Statistical significance level was defined as $P < 0.05$.

Results

Baseline characteristics

Between 1 April 2011 and 1 September 2017, a total of 1215 patients met the inclusion criteria. After excluding 781 patients (**Figure 3** and Supplementary Materials), 434 CHB patients with 2170 2D-SWE images were eventually enrolled in the final analysis. Among all the patients, 418 (96.3%) went through liver biopsy with qualified samples. The remaining 16 patients were diagnosed with cirrhosis according to the clinical criteria.

During 5-year follow-up, 32 patients (7.4%) of all the 434 patients developed HCC, 12 of which were diagnosed by liver biopsy and 20 which were diagnosed by contrast-enhanced imaging. Among all the patients, 262 patients with 1310 images, 86 patients with 430 images and 86 patients with 430 images were assigned to training, validation and testing cohorts, respectively. Baseline characteristics are shown in **Table 1** and details of antiviral therapy are shown in Supplementary Table 4.

Development and validation of radiomics prediction model of hepatocellular carcinoma

Factors associated with hepatocellular carcinoma: By comparing with HCC and non-HCC groups in training cohort, patients who developed HCC were older, had higher levels of LSM and lower levels of GGT and PLT than those without HCC (**Table 2**). Among all the characteristics, sex, age, ALB, and LSM were identified in the univariate analysis. Based on multivariate regression analysis, age and LSM were identified as independent risk factors for HCC development. Patients who developed HCC had a

Table 1. Baseline characteristics of patients

Variables	Training cohort	Validation cohort	Testing cohort
Number of patients (%)	262	86	86
Male, n (%)	209 (79.8)	58 (67.4)	68 (79.1)
Age (y)	37.7 ± 10.5	38.3 ± 10.8	36.8 ± 11.4
BMI (kg/m^2)	22.4 ± 3.3	22.3 ± 3.6	22.0 ± 3.2
AST (U/L)	33.0 (26.0-51.8)	35.5 (25.0-53.8)	30.0 (24.0-44.3)
ALT (U/L)	44.5 (29.0-76.0)	42.5 (26.0-66.0)	37.0 (26.3-60.5)
ALB (g/L)	43.8 (40.1-46.0)	42.6 (38.9-45.3)	43.5 (41.1-46.1)
TB (μmol/L)	14.4 (10.7-20.6)	15.0 (11.3-20.3)	12.8 (10.3-16.7)
GGT (U/L)	35.0 (23.0-80.0)	44.0 (18.3-94.5)	27.5 (20.0-48.8)
ALP (U/L)	69.0 (56.0-87.0)	68.0 (57.0-82.5)	68.0 (55.0-86.0)
PLT (* 10^9/L)	178.0 (144.8-220.3)	182.0 (145.0-217.0)	197.0 (157.5-223.5)
PT (s)	13.2 (12.7-13.8)	13.2 (12.7-13.9)	13.4 (13.0-14.0)
AFP (ng/mL)	4.0 (2.7-9.4)	3.7 (2.7-8.6)	2.9 (2.1-5.1)
HBV-DNA (IU/ml)	1.3 * 10^6 (7.9 * 10^4-1.9 * 10^7)	5.9 * 10^5 (5.5 * 10^4-2.9 * 10^7)	2.7 * 10^5 (1.5 * 10^4-2.2 * 10^6)
LSM (kPa)	10.7 ± 6.7	10.2 ± 6.4	9.5 ± 6.2
Antiviral therapy before LSM (%)	51 (19.5)	16 (18.6)	9 (10.5)
Antiviral therapy after LSM (%)	197 (75.2)	63 (73.3)	55 (64.0)
HCC cases (%)	20 (7.6)	6 (7.0)	6 (7.0)
Patients with liver biopsy (%)	252	83	83
Fibrosis stages			
F0	41 (16.3)	12 (14.5)	20 (24.1)
F1	77 (30.6)	23 (27.7)	29 (34.9)
F2	51 (20.2)	24 (28.9)	11 (13.3)
F3	37 (14.7)	11 (13.3)	12 (14.5)
F4	46 (18.3)	13 (15.7)	11 (13.3)
Necro-inflammation grades			
A0	9 (3.6)	6 (7.2)	4 (4.8)
A1	96 (38.1)	25 (30.1)	44 (53.0)
A2	76 (30.2)	28 (33.7)	19 (22.9)
A3	71 (28.2)	24 (28.9)	16 (19.3)
Steatosis grades			
S0	187 (74.2)	64 (77.1)	60 (72.3)
S1	52 (20.6)	15 (18.1)	19 (22.9)
S2	7 (2.8)	4 (4.8)	2 (2.4)
S3	6 (2.4)	0 (0)	2 (2.4)

Data are presented as mean ± standard deviation, median (IQR) or n (%). *P* values were calculated between training, validation and testing cohorts. BMI, body mass index; AST, aspartate aminotransferase; ALT, alanine aminotransferase; ALB, albumin; TB, total bilirubin; GGT, gamma-glutamyl transpeptidase; ALP, alkaline phosphatase; PLT, platelet count; PT, prothrombin time; AFP, serum α-fetoprotein; HBV-DNA, Hepatitis B virus deoxyribonucleic acid; LSM, liver stiffness measurement; HCC, hepatocellular carcinoma.

higher stage of fibrosis (62.5% were histological F4 or clinically diagnosed with cirrhosis) and a higher grade of inflammation (57.9% were A3), but had a lower grade of steatosis (63.2% were S0).

Visualization of feature distribution: Due to the high dimensional characteristics of deep learning features in diagnosis model, it is difficult to understand the diagnosis efficiency of the model explicitly. To prove the validity of the network feature, we used the t-Stochastic Neighbour Embedding (t-SNE) method to reduce the 64-dimensional features to three [20], as displayed them in **Figure 4**. The scatter diagram describes the spatial distribution of network features after dimension reduction by t-SNE method. The network features of HCC and non-HCC cases were clearly distinguishable, which proved that the 2D-SWE and B-mode US images do contain the important features for HCC prediction and the model could encode the images into distinguishing features.

Establishment and selection of radiomics prediction models: We compared relevant parameters including 2D-SWE images, B-mode US images, sex, age, serological data (AST, ALT, TB,

Table 2. Baseline characteristics of the HCC and Non-HCC CHB patients in training cohort

Variables	HCC (n = 20)	Non-HCC (n = 242)	*P* values
Male, n (%)	19 (95.0)	190 (78.5)	0.077
Age (y)	47.7 ± 10.6	36.9 ± 10.1	< 0.001
BMI (kg/m^2)	23.6 ± 4.1	22.3 ± 3.2	0.162
AST (U/L)	35.5 (27.5-55.8)	33.0 (26.0-51.0)	0.510
ALT (U/L)	40.5 (30.3-90.0)	45.5 (29.0-75.8)	0.901
ALB (g/L)	39.0 (36.9-45.3)	43.9 (40.8-46.0)	0.060
TB (μmol/L)	17.3 (12.9-23.6)	14.3 (10.6-19.9)	0.093
GGT (U/L)	43.0 (28.0-114.0)	35.0 (23.0-79.0)	0.043
ALP (U/L)	80.0 (67.0-108.0)	69.0 (56.0-87.0)	0.093
PLT (* 10^9/L)	122.0 (80.3-168.5)	180.0 (148.8-222.5)	< 0.001
PT (s)	14.0 (13.5-15.3)	13.2 (12.6-13.8)	0.102
AFP (ng/mL)	4.5 (2.9-14.9)	3.9 (2.7-8.0)	0.257
HBV-DNA (IU/ml)	1.6 * 10^6 (2.1 * 10^5-5.1 * 10^6)	1.3 * 10^6 (7.2 * 10^4-1.9 * 10^7)	0.593
2D-SWE (kPa)	14.2 ± 8.1	10.4 ± 6.6	0.015

Data are presented as mean ± standard deviation, median (IQR) or n (%). *P* values were calculated between patients developed HCC and others. BMI, body mass index; AST, aspartate aminotransferase; ALT, alanine aminotransferase; ALB, albumin; TB, total bilirubin; GGT, gamma-glutamyl transpeptidase; ALP, alkaline phosphatase; PLT, platelet count; PT, prothrombin time; AFP, serum α-fetoprotein; HBV-DNA, Hepatitis B virus deoxyribonucleic acid; 2D-SWE, two-dimensional shear wave elastography; HCC, hepatocellular carcinoma.

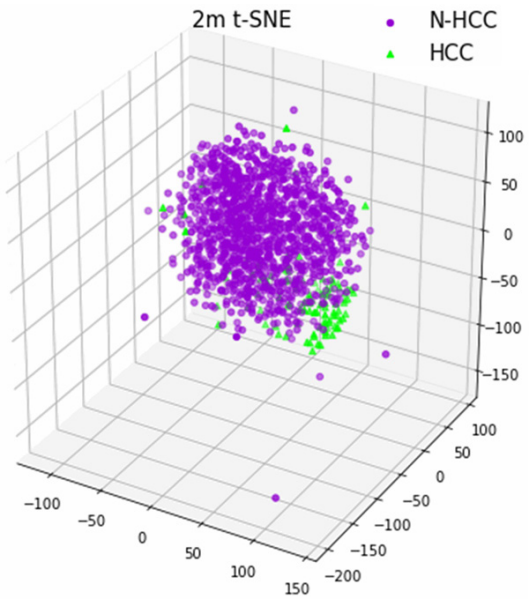

Figure 4. The spatial distribution of network features after dimension reduction by t-SNE. Purple dots and green triangles correspond to the features of non-HCC cases and HCC cases, respectively.

ALB) and built ten different models, named R1 to R10 (Supplementary Table 5). In the training cohort, all the models demonstrated a very high prognostic accuracy with the AUCs more than 0.970 in predicting HCC. Since R8 (B-mode US images + 2D-SWE images + sex + age) did not include serological data, and reached 0.942 and 0.900 in validation and testing cohorts, respectively, with no significant difference in R10 (B-mode US images + 2D-SWE images + sex + age + serological data). Therefore, we chose R8 as the best radiomics prediction model and named it HCC-R.

Comparison of HCC-R and other methods

Performance of HCC-R in predicting HCC in training cohort was significantly higher than that of LSM, GAG-HCC, CAMD, HCC-ESC and CU-HCC (**Table 3** and **Figure 5**). Comparable results were observed in validation and testing cohorts: performance of HCC-R was the best in both validation and testing cohorts for predicting HCC, with the AUCs of 0.942 in validation cohort and 0.900 in testing cohort. Predictive ability of HCC-R was significantly higher than LSM, GAG-HCC, HCC-ESC and CU-HCC in validation cohort, and significantly higher than LSM and CU-HCC in testing cohort (**Table 3**).

Prediction performance of HCC-R and liver stiffness measurement in different transaminase levels and under different antiviral therapy conditions

Performance of HCC-R and LSM for predicting HCC in different transaminase levels and un-

Table 3. Comparison of diagnostic performance of HCC-R, LSM and other clinical prognostic scores in predicting HCC in training, validation and testing cohorts

	Scores	N	AUC	SEN (%)	SPE (%)	PPV (%)	NPV (%)	PLR	NLR	P
Training cohort	HCC-R	262	0.981 (0.965-0.996)	100.0 (82.4-100.0)	96.7 (93.6-98.6)	70.4 (49.4-86.5)	100.0 (98.4-100.0)	30.4 (29.7-31.1)	0 (-)	
	LSM	262	0.676 (0.615-0.732)	65.0 (40.8-84.6)	66.5 (60.2-72.4)	13.8 (7.6-22.5)	95.8 (91.6-98.3)	1.9 (1.4-2.7)	0.5 (0.3-1.0)	< 0.001
	GAG-HCC	239	0.869 (0.820-0.909)	88.9 (65.3-98.6)	80.1 (74.2-85.1)	26.7 (16.0-39.8)	98.9 (96.0-99.9)	4.5 (3.7-5.3)	0.1 (0.04-0.5)	0.011
	CAMD	262	0.797 (0.743-0.844)	60.0 (36.1-80.9)	93.8 (90.0-96.5)	44.4 (25.5-64.7	96.6 (93.4-98.5)	9.7 (6.8-13.9)	0.4 (0.2-0.9)	0.006
	HCC-ESC	239	0.763 (0.703-0.815)	83.3 (58.6-96.4)	62.0 (55.2-68.4)	15.2 (8.7-23.8)	97.9 (93.8-99.6)	2.2 (1.7-2.8)	0.3 (0.1-0.8)	< 0.001
	CU-HCC	239	0.801 (0.744-0.849)	88.9 (65.3-98.6)	70.1 (63.6-76.1)	19.5 (11.6-29.7)	98.7 (95.5-99.8)	3.0 (2.5-3.6)	0.2 (0.04-0.6)	0.002
Validation cohort	HCC-R	86	0.942 (0.874-1.000)	83.3 (35.9-99.6)	95.0 (87.7-98.6)	55.6 (21.2-86.3)	98.7 (92.9-100.0)	16.7 (11.6-23.9)	0.2 (0.02-1.3)	
	LSM	86	0.722 (0.615-0.813)	66.7 (22.3-95.7)	78.8 (68.2-87.1)	11.6 (3.9-25.1)	100.0 (88.8-100.0)	1.8 (1.4-2.4)	0 (-)	0.014
	GAG-HCC	75	0.703 (0.586-0.803)	100.0 (47.8-100.0)	45.7 (33.7-58.1)	38.5 (13.9-68.4)	98.2 (90.5-100.0)	6.7 (4.6-9.6)	0.2 (0.03-1.3)	0.040
	CAMD	86	0.743 (0.637-0.831)	66.7 (22.3-95.7)	73.8 (62.7-83.0)	16.0 (4.4-36.6)	96.7 (88.7-99.6)	2.5 (1.4-4.5)	0.5 (0.1-1.5)	0.063
	HCC-ESC	75	0.606 (0.486-0.717)	60.0 (14.7-94.7)	78.6 (67.1-87.5)	16.7 (3.6-41.4)	96.5 (87.8-99.6)	2.8 (1.4-5.8)	0.5 (0.2-1.6)	< 0.001
	CU-HCC	75	0.544 (0.425-0.660)	60.0 (14.7-94.7)	72.9 (60.9-82.8)	13.6 (2.9-34.9)	96.2 (86.9-99.6)	2.2 (1.1-4.6)	0.6 (0.2-1.7)	0.006
Testing cohort	HCC-R	86	0.900 (0.717-1.000)	83.3 (35.9-99.6)	96.3 (89.4-99.2)	62.5 (22.1-92.7)	98.7 (93.0-100.0)	22.2 (15.5-31.9)	0.2 (0.02-1.4)	
	LSM	86	0.784 (0.683-0.866)	83.3 (35.9-99.6)	78.8 (68.2-87.1)	22.7 (7.8-45.4)	98.4 (91.5-100.0)	3.9 (2.7-5.7)	0.2 (0.03-1.3)	0.032
	GAG-HCC	71	0.815 (0.705-0.897)	80.0 (28.4-99.5)	93.9 (85.2-98.3)	50.0 (13.9-86.1)	98.4 (91.5-100.0)	13.2 (8.5-20.5)	0.2 (0.03-1.6)	0.170
	CAMD	86	0.863 (0.771-0.927)	83.3 (35.9-99.6)	78.8 (68.2-87.1)	22.7 (7.8-45.4)	98.4 (91.5-100.0)	3.9 (2.7-5.7)	0.2 (0.03-1.3)	0.514
	HCC-ESC	71	0.779 (0.665-0.869)	80.0 (28.4-99.5)	90.9 (81.3-96.6)	40.0 (12.2-73.8)	98.4 (91.1-100.0)	8.8 (5.6-13.7)	0.2 (0.03-1.5)	0.132
	CU-HCC	71	0.773 (0.658-0.864)	80.0 (28.4-99.5)	90.9 (81.3-96.6)	40.0 (12.2-73.8)	98.4 (91.1-100.0)	8.8 (5.6-13.7)	0.2 (0.03-1.5)	0.047

HCC-R: B-mode US images + 2D-SWE images + sex + age. Data in parentheses are 95% CIs. *P* values were calculated between HCC-R and other methods in training, validation and testing cohorts. LSM, liver stiffness measurement; ROC, receiver operating characteristic; N, number of patients; AUC, area under the receiver operating characteristic curve; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value; PLR, positive likelihood ratio; NLR, negative likelihood ratio.

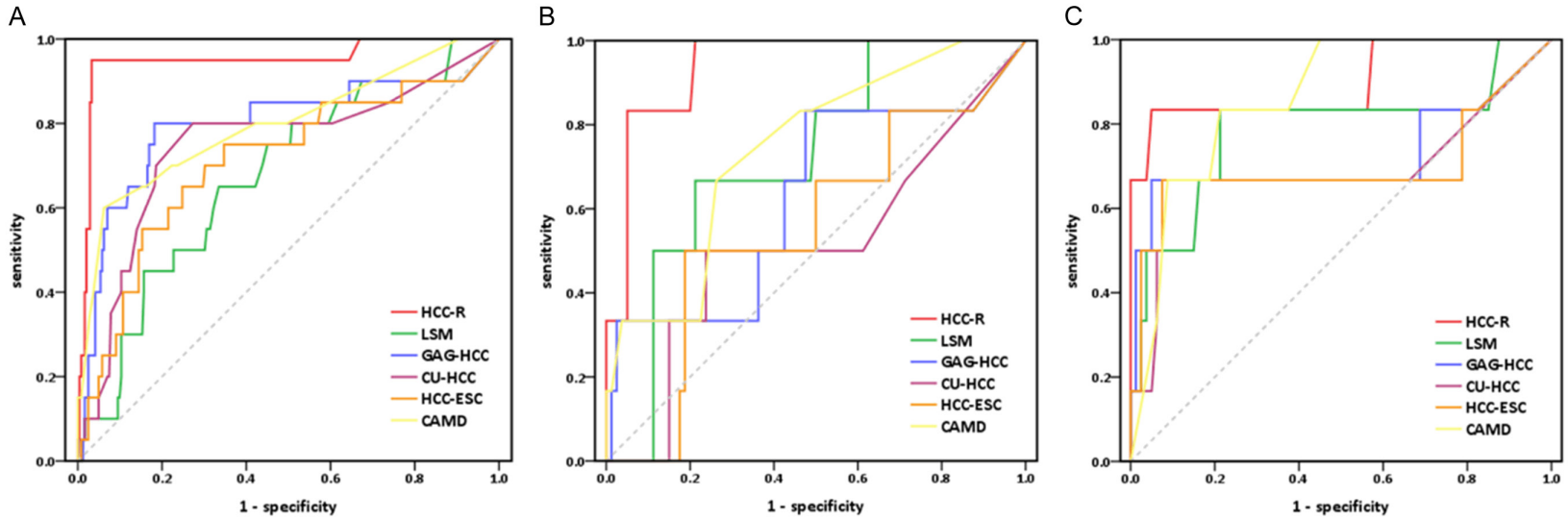

Figure 5. Comparison of ROC curves between HCC-R, LSM and other clinical prognostic scores in predicting HCC in training (A), validation (B) and testing (C) cohorts.

der different antiviral therapy conditions was shown in Supplementary Tables 6 and 7. Performances of HCC-R was not significantly different with regard to different AST or ALT levels. On the other hand, AUCs of LSM in AST ≤ 2×ULN and ALT ≤ 2×ULN group (AUC: 0.732~0.734) were higher than that in AST > 2×ULN and ALT > 2×ULN group (AUC: 0.582~0.689), although there was no significant difference between them. Also, there was no significant difference in the AUCs between HCC-R and LSM under different antiviral therapy conditions before LSM. Accuracy of HCC-R in predicting HCC was significantly higher than that of LSM in each stratification ($P < 0.001$).

Discussion

To our knowledge, this is the first study that applied radiomics method to predict HCC in patients with CHB. Our study enrolled 2170 images obtained from 434 CHB patients with 5-year follow-up duration in order to establish a new model to predict HCC. According to radiomics analysis and modeling, we built the HCC-R, which was based on 2D-SWE and B-mode US images as well as clinical information and serological data, with a better prediction performance than LSM and clinical prognostic scores.

In this study, we used single or multiple parameters to build different kinds of prediction models, and aimed to determine which model was the most suitable one for clinical use. This series of work is too difficult to finish merely by manpower, but can be easily achieved by radiomics. After radiomics analysis and optimization of parameters, we have built many models during our study. Compared with other models, R8 was not only accuracy in HCC prediction, but also simple as it was only composed of two images (2D-SWE and B-mode US images) from US examination and two basic information (sex and age) that could be easily acquired from each patient. In this way, operators only need to provide sex and age of the patient and perform a routine US with 2D-SWE examination, which is particularly convenient and valuable for clinical applications. According to prediction accuracy and clinical practicability, we finally chose R8 as our new HCC prediction model and named it HCC-R. The radiomics signatures of model R1 to R10 are summarized in Supplementary Materials. In addition to the two clinical information (sex and age), there are 12 deep features that jointly trained the HCC-R model.

A recent cohort study [7] of CHB patients demonstrated that LSM by TE was available in HCC prediction and produced a new prognostic score based on TE values (including age, albumin, HBV-DNA and LSM) with AUC of 0.83 for 5-year prognosis. However, in our study HCC-R showed higher accuracy with AUC of 0.981, 0.942 and 0.900 in training, validation and testing cohorts for 5-year prognosis. The reasons for such excellent performance are as follows: First, 2D-SWE provided both elastography and B-mode US images instead of just LSM value on TE, making it possible to perform radiomics analysis. By using high-throughput computer image analysis, radiomics analyzed every pixel within 2D-SWE and B-mode US images, which contain more accurate information than that of LSM value alone. In this way, the AUC of HCC-R, i.e. radiomics prediction model, was not only higher than AUC obtained LSM by using 2D-SWE (AUC: 0.676~0.784) in our study but also by that using TE (AUC: 0.73) in literature [7]. Second, 2D-SWE is relatively new shear-wave technique which can be used in patients with ascites. The diagnostic performance of 2D-SWE is even higher than that of TE, especially for advanced fibrosis (≥ F3) and cirrhosis (F4) [21]. Furthermore, HCC-R integrates both biomechanical and morphological information, which is more comprehensive and objective than serological data and LSM value.

The other study similar with us proposed a radiomics model to calculate HCC risk score in cirrhotic patients with indeterminate liver nodules [22]. Based on triphasic CT scans data, the author extracted massive quantitative imaging features and trained a radiomics signature with AUC of 0.70. In model building process, a CNN model is applied to extract high level vision features. During classifier training, deep features combined with clinical features jointly construct model and thus the model achieve promising performance. In addition, our model aims to predict prognostic score in patients with CHB for 5-years prognosis of HCC, focusing on the future.

The latest Elastography Guidelines and Recommendations [8, 23, 24] note that LSM might be affected by AST or ALT levels, which was also demonstrated in our previous studies [25]. To investigate whether AST or ALT level affects performances of HCC-R and LSM or not, stratification analysis was performed. The results showed that the accuracy of HCC-R was higher than that of LSM in each stratification, and neither AST level nor ALT level impacted the performance of HCC-R in HCC prediction. These findings revealed that HCC-R is also suitable for patients with high AST and/or ALT levels, which indicates that radiomics analysis may avoid the influence of elevated transaminase level on the detection of liver stiffness.

Currently, many well-established clinical prognostic scores, such as GAG-HCC [3], CAMD [4], HCC-ESC [5] and CU-HCC [6], have been used to predict risk of HCC. Parameters like sex, age and cirrhosis considered the essential features in these scores (CU-HCC contains the latter two while other scores contain all of them). Surprisingly, our results demonstrated that sex and age were extremely important features which were also included in HCC-R. Furthermore, assessing cirrhosis with radiomics is better than assessing cirrhosis with clinical methods, since radiomics could analyze many more features and details that may provide valuable information in clinical practice, as confirmed in our previous study. On the other hand, there were still some common limitations to these clinical prognostic scores, such as the components involved were insufficient. In contrast, our research applied radiomics analysis to contain many more features with advanced computing methods and revealed HCC-R as a comprehensive and accurate approach in HCC prediction. By comparing the prognostic accuracy of HCC-R with that of other methods, we found that the AUC of CAMD were higher in testing cohort than in training cohort, which might be due to the small sample size of testing cohort. However, HCC-R achieved a much better accuracy in all the training, validation and testing cohorts for HCC prediction.

The strengths of radiomics analysis applied in this study were as follows. First of all, we were the first to effectively combine the high-throughput US image features with low-dimensional clinical information to build a new radiomics model, which was more comprehensive. Second, to overcome the challenge of unbalanced sample size, we proposed an effective data amplification strategy in the training of deep learning network, which greatly improved the prognostic efficiency of our model. In addition, we have already developed the software according to the results of this study and uploaded it to the website (https://drive.google.com/drive/folders/140h9OUbH1JIN7UVs7foFk7q70B8LJCV6?usp=sharing), making it valuable for clinical guidance and application.

We must admit that the population size of our study, especially that of HCC patients, was limited, and unbalanced data distribution might compromise the efficacy of HCC-R. However, previous studies focusing on these areas showed that the 5-year cumulative incidence of HCC was 4.3-8.7% [26, 27], which is consistent with the data in our current study. Additionally, according to EASL guideline [16], approximately 2% of hepatitis B virus (HBV) related infected cirrhotic patients (F4) develop HCC per year, and even fewer patients with other stages of fibrosis (F0-F3). Since the patients enrolled in our study were suffered from different stages of liver fibrosis, which might cause the number of events to be relatively small. In fact, our study not merely focused on decompensated cirrhotic patients, but also enrolled patients with compensated cirrhosis and early fibrosis stages. Currently, many CHB patients, especially those with early stages of fibrosis, do not pay any attention to regular follow-up and treatment until complications (such as HCC) occur. Since HCC-R can accurately predict HCC in patients with all stages of fibrosis, it may arouse attention of many more patients and their physicians and play a key role in clinical practice and personalized medicine.

The major limitation of our study is that this was a single-disease investigation that considered only HBV-infected patients. This was because that HBV infection is the main cause of HCC worldwide, mostly in Asia and sub-Saharan Africa [28], which led us to pay specific attention to CHB patients. In the future, we need further prospective multicenter studies that involve more patients, not only with HBV infection, but also with hepatitis C virus infec-

tion, nonalcoholic fatty liver disease, alcoholic liver disease and other etiologies, to optimize and improve HCC-R. Also, we would like to find out whether HCC-R is valuable for predicting tumor recurrence or died in patients with HCC.

In conclusion, radiomics model HCC-R based on 2D-SWE and B-mode US images, sex and age of patients can accurately predict HCC occurrence in CHB patients. It can overcome the influence of elevated transaminase level in LSM and is beneficial to HCC surveillance, individualized treatment and clinical outcomes.

Acknowledgements

The study was supported by grants from the National Natural Science Foundation of China (No. 81827802), Shanghai Science and Technology Innovation Plan (19441903100).

Disclosure of conflict of interest

None.

Address correspondence to: Rongqin Zheng, Department of Ultrasound, The Third Affiliated Hospital of Sun Yat-Sen University, Guangzhou 510630, China; Guangdong Key Laboratory of Liver Disease Research, The Third Affiliated Hospital of Sun Yat-Sen University, Guangzhou 510630, China. Tel: +86-20-85252010; E-mail: zhengrq@mail.sysu.edu.cn; Jinhua Yu, Department of Electronic Engineering, Fudan University, Shanghai 200433, China. Tel: +86-21-65643202; E-mail: jhyu@fudan.edu.cn

References

[1] Forner A, Reig M and Bruix J. Hepatocellular carcinoma. Lancet 2018; 391: 1301-1314.

[2] Chen S, Li J, Wang D, Fung H, Wong LY and Zhao L. The hepatitis B epidemic in China should receive more attention. Lancet 2018; 391: 1572.

[3] Yuen MF, Tanaka Y, Fong DY, Fung J, Wong DK, Yuen JC, But DY, Chan AO, Wong BC, Mizokami M and Lai CL. Independent risk factors and predictive score for the development of hepatocellular carcinoma in chronic hepatitis B. J Hepatol 2009; 50: 80-88.

[4] Hsu YC, Yip TC, Ho HJ, Wong VW, Huang YT, El-Serag HB, Lee TY, Wu MS, Lin JT, Wong GL and Wu CY. Development of a scoring system to predict hepatocellular carcinoma in Asians on antivirals for chronic hepatitis B. J Hepatol 2018; 69: 278-285.

[5] Fung J, Cheung KS, Wong DK, Mak LY, To WP, Seto WK, Lai CL and Yuen MF. Long-term outcomes and predictive scores for hepatocellular carcinoma and hepatitis B surface antigen seroclearance after hepatitis B e-antigen seroclearance. Hepatology 2018; 68: 462-472.

[6] Wong VW, Chan SL, Mo F, Chan TC, Loong HH, Wong GL, Lui YY, Chan AT, Sung JJ, Yeo W, Chan HL and Mok TS. Clinical scoring system to predict hepatocellular carcinoma in chronic hepatitis B carriers. J Clin Oncol 2010; 28: 1660-1665.

[7] Wong GL, Chan HL, Wong CK, Leung C, Chan CY, Ho PP, Chung VC, Chan ZC, Tse YK, Chim AM, Lau TK and Wong VW. Liver stiffness-based optimization of hepatocellular carcinoma risk score in patients with chronic hepatitis B. J Hepatol 2014; 60: 339-345.

[8] Dietrich C, Bamber J, Berzigotti A, Bota S, Cantisani V, Castera L, Cosgrove D, Ferraioli G, Friedrich-Rust M, Gilja O, Goertz R, Karlas T, de Knegt R, de Ledinghen V, Piscaglia F, Procopet B, Saftoiu A, Sidhu P, Sporea I and Thiele M. EFSUMB guidelines and recommendations on the clinical use of liver ultrasound elastography, update 2017 (Long Version). Ultraschall Med 2017; 38: e16-e47.

[9] Zeng J, Liu GJ, Huang ZP, Zheng J, Wu T, Zheng RQ and Lu MD. Diagnostic accuracy of two-dimensional shear wave elastography for the non-invasive staging of hepatic fibrosis in chronic hepatitis B: a cohort study with internal validation. Eur Radiol 2014; 24: 2572-2581.

[10] Gillies RJ, Kinahan PE and Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016; 278: 563-577.

[11] Wang K, Lu X, Zhou H, Gao Y, Zheng J, Tong M, Wu C, Liu C, Huang L, Jiang Ta, Meng F, Lu Y, Ai H, Xie XY, Yin LP, Liang P, Tian J and Zheng R. Deep learning radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study. Gut 2019; 68: 729-741.

[12] Terrault NA, Bzowej NH, Chang KM, Hwang JP, Jonas MM and Murad MH. AASLD guidelines for treatment of chronic hepatitis B. Hepatology 2016; 63: 261-283.

[13] Zheng J, Guo H, Zeng J, Huang Z, Zheng B, Ren J, Xu E, Li K and Zheng R. Two-dimensional shear-wave elastography and conventional US: the optimal evaluation of liver fibrosis and cirrhosis. Radiology 2015; 275: 290-300.

[14] Jin JY, Zheng YB, Zheng J, Liu J, Mao YJ, Chen SG, Gao ZL and Zheng RQ. 2D shear wave elastography combined with MELD improved prognostic accuracy in patients with acute-on-chronic hepatitis B liver failure. Eur Radiol 2018; 28: 4465-4474.

[15] Bedossa P and Poynard T. An algorithm for the grading of activity in chronic hepatitis C. The

METAVIR Cooperative Study Group. Hepatology 1996; 24: 289-293.
[16] Galle PR, Forner A, Llovet JM, Mazzaferro V, Piscaglia F, Raoul JL, Schirmacher P and Vilgrain V. EASL clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol 2018; 69: 182-236.
[17] Rumelhart DE, Hinton GE and Williams RJ. Learning representations by back-propagating errors. Nature 1986; 323: 533-536.
[18] Wu G, Chen Y, Wang Y, Yu J, Lv X, Ju X, Shi Z, Chen L and Chen Z. Sparse representation-based radiomics for the diagnosis of brain tumors. IEEE Trans Med Imaging 2018; 37: 893-905.
[19] Yosinski J, Clune J, Bengio Y and Lipson H. How transferable are features in deep neural networks? Adv Neural Inf Process Syst 27 (Nips 2014) 2014; 27.
[20] van der Maaten L and Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008; 9: 2579-2605.
[21] Cassinotto C, Lapuyade B, Mouries A, Hiriart JB, Vergniol J, Gaye D, Castain C, Le Bail B, Chermak F, Foucher J, Laurent F, Montaudon M and De Ledinghen V. Non-invasive assessment of liver fibrosis with impulse elastography: comparison of Supersonic Shear Imaging with ARFI and FibroScan(R). J Hepatol 2014; 61: 550-557.
[22] Mokrane FZ, Lu L, Vavasseur A, Otal P, Peron JM, Luk L, Yang H, Ammari S, Saenger Y, Rousseau H, Zhao B, Schwartz LH and Dercle L. Radiomics machine-learning signature for diagnosis of hepatocellular carcinoma in cirrhotic patients with indeterminate liver nodules. Eur Radiol 2020; 30: 558-570.
[23] Ferraioli G, Filice C, Castera L, Choi BI, Sporea I, Wilson SR, Cosgrove D, Dietrich CF, Amy D, Bamber JC, Barr R, Chou YH, Ding H, Farrokh A, Friedrich-Rust M, Hall TJ, Nakashima K, Nightingale KR, Palmeri ML, Schafer F, Shiina T, Suzuki S and Kudo M. WFUMB guidelines and recommendations for clinical use of ultrasound elastography: part 3: liver. Ultrasound Med Biol 2015; 41: 1161-1179.
[24] Barr RG, Ferraioli G, Palmeri ML, Goodman ZD, Garcia-Tsao G, Rubin J, Garra B, Myers RP, Wilson SR, Rubens D and Levine D. Elastography assessment of liver fibrosis: society of radiologists in ultrasound consensus conference statement. Radiology 2015; 276: 845-861.
[25] Zeng J, Zheng J, Jin JY, Mao YJ, Guo HY, Lu MD, Zheng HR and Zheng RQ. Shear wave elastography for liver fibrosis in chronic hepatitis B: adapting the cut-offs to alanine aminotransferase levels improves accuracy. Eur Radiol 2019; 29: 857-865.
[26] Wong GL, Chan HL, Chan HY, Tse PC, Tse YK, Mak CW, Lee SK, Ip ZM, Lam AT, Iu HW, Leung JM and Wong VW. Accuracy of risk scores for patients with chronic hepatitis B receiving entecavir treatment. Gastroenterology 2013; 144: 933-944.
[27] Jung KS, Kim SU, Song K, Park JY, Kim DY, Ahn SH, Kim BK and Han KH. Validation of hepatitis B virus-related hepatocellular carcinoma prediction models in the era of antiviral therapy. Hepatology 2015; 62: 1757-1766.
[28] Villanueva A. Hepatocellular carcinoma. N Engl J Med 2019; 380: 1450-1462.

Supplementary Materials

Patients and methods

GAG-HCC score [1]

14 * sex (male = 1; female = 0) + age (in years) + 3 * HBV-DNA levels (copies/mL in log) + 33 * cirrhosis (presence = 1; absence = 0).

CAMD score [2]

Supplementary Table 1. CAMD score

Variables	CAMD Score
Cirrhosis	
No cirrhosis	0
Cirrhosis with age < 40 years	10
Cirrhosis with age ≥ 40 years	6
Age	
Age: < 40 years	0
Age: 40-49 years	5
Age: 50-59 years	8
Age: 60 years or older	10
Male Sex	
Female sex	0
Male sex	2
Diabetes Mellitus	
Not diabetic	0
Diabetic	1

HCC-ESC score [3]

Age (years) + 20 * sex (male = 1; female = 0) + 29 * cirrhosis (presence = 1; absence = 0) + 5 * DNA (log IU/mL) + 31 * ALT group (flares or persistently abnormal ALT = 1; otherwise = 0) + 23 * hypoalbuminemia (< 39 g/L, presence = 1; absence = 0).

CU-HCC score [4]

Supplementary Table 2. CU-HCC score

Factor	Score
Age, years	
> 50	3
≤ 50	0
Albumin, g/L	
≤ 35	20
> 35	0
Bilirubin, μmol/L	
> 18	1.5
≤ 18	0

HBV-DNA, log copies/mL	
≤ 4	0
4-6	1
> 6	4
Cirrhosis	
Yes	15
No	0

Ultrasonography and two-dimensional shear wave elastography

All of the patients were fasted for at least 2 hours and rested for at least 10 minutes before two-dimensional shear wave elastography (2D-SWE) measurements. 2D-SWE was performed on the right lobe of liver for each patient during a brief breath hold, in supine position through a right intercostal space with right arm elevated above head. The size of region of interest (ROI) for 2D-SWE was 4 cm×3 cm, and it was located in liver parenchyma approximately 1 cm beneath liver capsule free of large vessels or bile ducts. A 2 cm diameter circular Q-Box was placed in the ROI area, then the elastography values were automatically calculated and displayed.

HCC-R model and training process

The HCC-R model is mainly based on ResNet50 and inherits all convolution layers of it [5]. Globally, one path of the HCC-R model is stacked from a 224 * 224 * 3 input layer, a 7 * 7 convolutional layer, a max-pooling layer and a series of residual blocks, following a global average pooling (GAP) layer and three fully connected (FC) layers with 128, 64 and 2 neurons, respectively. For the 2D-SWE and US ROIs inputs, a parallel path is added and both share parameters. After encoded by GAP layer, the two modal features were concatenated and used as the input of the following FC layers. We disentangle context between two modal image information by feature fusion. The detailed structure of the model is shown in **Table 1**.

Supplementary Table 3. Detailed model structure of HCC-R

Layer name	Output size	HCC-R
Conv1	112×112	7×7,64, stride 2
Conv2_x	56×56	[1×1,64 3×3,64 1×1,256] ×3
Conv3_x	28×28	[1×1,128 3×3,128 1×1,512] ×4
Conv4_x	14×14	[1×1,256 3×3,256 1×1,1024] ×6
Conv5_x	7×7	[1×1,512 3×3,512 1×1,2048] ×3
Fc1	1×1	Average pool, 256-d fc
Fc2*	1×1	64-d fc
Fc3	1×1	2-d fc, softmax

*Note that the gray shading (Fc2) refers to deep feature output layer. Conv, convolution; Fc, fully connected.

In order to accelerate the convergence of the model and reduce the training difficulty, we use the parameters pre-trained on the ImageNet dataset to initialize the convolution layers [6]. The three fully connected layers were initialized by Xavier weights. The cross-entropy of the output and the label was cal-

culated as the loss function. We set learning rate to 1e-4 and the Adam optimizer is used to optimize the objective function and update the model parameters with batch size 48. The maximum training step is set to 5000 and the learning rate decayed by 1/2 at 1000 and 3000 steps. After the training is finished, we select the last but one fully connected layer outputs as deep features with dimension of 64.

For the feature selection, a wrapper based method was applied, which considered the effect of a subset of features on each iteration. According to the idea of sparse representation, in each iteration step, the subset features are used to represent the classification labels [7]. And the above process can be formulated as:

$$\hat{\alpha}^{(k)} = \underset{\alpha}{\operatorname{argmin}} \| Y^{(k)} - D^{(k)}\alpha \|_2^2 + \lambda \| \alpha \|_0$$

Where $Y^{(k)}$ is the sample label of kth iteration. $D^{(k)}$ is the feature set of kth iteration. α is the sparse coefficients and a lager α_i means higher importance. By using OMP algorithm, the k-iteration results can be obtained:

$$\alpha^{(k)} = \frac{1}{k}\sum_{i=1}^{k}\hat{\alpha}^{(i)}$$

The loop ended when the residual of regression reaches a specific ε.

For the feature classification, the SVM model is used and we adjust the penalty coefficients of different sample categories to reduce the influence of imbalance data [8]. The mathematical formulates can be written as:

$$\min_{\omega,b,\xi} \frac{1}{2}\omega^T\omega + C_+ \sum_{y_i=1}\xi_i + C_- \sum_{y_i=-1}\xi_i$$

subject to $y_i(\omega^T\phi(x_i) + b) \geq 1-\xi_i$

$\xi_i \geq 0$, $i = 1,\ldots, l$

Where $\phi(\cdot)$ is the kernel function. ω and b determine the classification hyperplane. ξ_i is a small constant and $C_{+(-)}$ is the penalty coefficients.

Data processing and l^2 regularization

In the training cohort, there are 226 non-HCC samples and 21 HCC samples. For each samples, 5 US images are enrolled for analysis and thus non-HCC set and HCC set have 1130 and 105 images, respectively. Before fed into the model, the images were randomly selected from the two sets in equal probability, then perform data augmentation and normalization. To statistical model performance, the model predicts 5 images of each sample and votes to determine the final prediction result.

In the training process, we adopted a transfer learning strategy to save training cost. To reduce the risk of overfitting, we used l^2 regularization to constrain the complexity of the model. The mathematical process can be formulated as:

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \sum_i L(y_i, f(x_i;\omega)) + \lambda \| \omega \|_2$$

Where ω is model parameters and $L(\cdot)$ is loss function. x_i and y_i correspond to sample and label, respectively. By adjusting the constant λ, model complexity can be controlled.

Radiomics signature

To build the radiomics, we extract 64 deep features and 6 clinical features (sex, age, serological data (AST, ALT, TB, ALB)). A wrapper based feature selection algorithm is used to reduce feature dimension and the radiomics signatures of model R1 to R10 are summarized below:

R1: 14 deep features.

R2: age + 9 deep features.

R3: age + AST + ALT + 27 deep features.

R4: 25 deep features.

R5: age + 12 deep features.

R6: age + AST + ALT + 17 deep features.

R7: 18 deep features.

R8: sex + age + 12 deep features.

R9: sex + age + AST + ALT + 9 deep features.

R10: ALT + TB + 13 deep features.

Results

Baseline characters

Patients were excluded because of unsuccessful 2D-SWE measurements (5 patients for overweight, 18 patients for inability to hold breath), follow-up less than five years (710 patients), combined with other liver diseases (3 patients combined with hepatitis C virus infection, 2 patients combined with alcoholic liver disease, 1 patient combined with primary biliary cholangitis, 25 patients combined with HCC, 1 patient combined with hepatic metastases cancer), and missing important serological data (16 patients).

Supplementary Table 4. Antiviral therapy before and after LSM

Antiviral therapy	Number of patients
Before LSM	76
Interferon-based therapy	1
Lamivudine	5
Telbivudine	22
Adefovir	6
Entecavir	38
Tenofovir	5
After LSM	319
Interferon-based therapy	66
Lamivudine	12
Telbivudine	81
Adefovir	51
Entecavir	181
Tenofovir	57

LSM, liver stiffness measurement.

Supplementary Table 5. Diagnostic performance of different radiomics prediction models in predicting HCC in training, validation and testing cohorts

		AUC	SEN (%)	SPE (%)	PPV (%)	NPV (%)	PLR	NLR	P
Training cohort n = 262	R1	0.979 (0.964-0.995)	100.0 (82.4-100.0)	96.3 (93.1-98.3)	67.9 (47.6-84.1)	100.0 (98.4-100.0)	27.0 (26.3-27.7)	0 (-)	0.2738
	R2	0.980 (0.965-0.995)	100.0 (82.4-100.0)	95.1 (91.5-97.4)	61.3 (42.2-78.2)	100.0 (98.4-100.0)	20.3 (19.7-20.8)	0 (-)	0.3284
	R3	0.985 (0.971-0.998)	100.0 (82.4-100.0)	97.1 (94.2-98.8)	73.1 (51.7-88.7)	100.0 (98.4-100.0)	34.7 (34.0-35.5)	0 (-)	0.8424
	R4	0.983 (0.970-0.997)	100.0 (82.4-100.0)	96.7 (93.6-98.6)	70.4 (49.4-86.5)	100.0 (98.4-100.0)	30.4 (29.7-31.1)	0 (-)	0.6349
	R5	0.984 (0.970-0.997)	100.0 (82.4-100.0)	96.7 (93.6-98.6)	70.4 (49.4-86.5)	100.0 (98.4-100.0)	30.4 (29.7-31.1)	0 (-)	0.5983
	R6	0.977 (0.959-0.994)	100.0 (82.4-100.0)	96.3 (93.1-98.3)	67.9 (47.6-84.1)	100.0 (98.4-100.0)	27.0 (26.3-27.7)	0 (-)	0.1466
	R7	0.985 (0.973-0.998)	100.0 (82.4-100.0)	96.7 (93.6-98.6)	70.4 (49.4-86.5)	100.0 (98.4-100.0)	30.4 (29.7-31.1)	0 (-)	0.8820
	R8	0.981 (0.965-0.996)	100.0 (82.4-100.0)	96.7 (93.6-98.6)	70.4 (49.4-86.5)	100.0 (98.4-100.0)	30.4 (29.7-31.1)	0 (-)	0.2080
	R9	0.990 (0.980-1.000)	100.0 (82.4-100.0)	97.9 (95.3-99.3)	79.2 (57.8-92.9)	100.0 (98.5-100.0)	48.6 (47.7-49.5)	0 (-)	0.3060
	R10	0.986 (0.974-0.998)	100.0 (82.4-100.0)	97.1 (94.2-98.8)	73.1 (51.7-88.7)	100.0 (98.4-100.0)	34.7 (34.0-35.5)	0 (-)	-
Validation cohort n = 86	R1	0.896 (0.759-1.000)	83.3 (35.9-99.6)	93.8 (86.0-97.9)	50.0 (18.7-81.3)	98.7 (92.8-100.0)	13.3 (9.3-19.2)	0.2 (0.02-1.3)	0.4155
	R2	0.902 (0.759-1.000)	83.3 (35.9-99.6)	96.3 (89.4-99.2)	62.5 (22.1-92.7)	98.7 (93.0-100.0)	22.2 (15.5-31.9)	0.2 (0.02-1.4)	0.5731
	R3	0.875 (0.694-1.000)	83.3 (35.9-99.6)	95.0 (87.7-98.6)	55.6 (21.2-86.3)	98.7 (92.9-100.0)	16.7 (11.6-23.9)	0.2 (0.02-1.3)	0.3759
	R4	0.860 (0.652-1.000)	83.3 (35.9-99.6)	92.5 (84.4-97.2)	45.5 (15.6-78.0)	98.7 (92.8-100.0)	11.1 (7.7-16.0)	0.2 (0.03-1.3)	0.3557
	R5	0.883 (0.736-1.000)	83.3 (35.9-99.6)	92.5 (84.4-97.2)	45.5 (15.6-78.0)	98.7 (92.8-100.0)	11.1 (7.7-16.0)	0.2 (0.03-1.3)	0.2793
	R6	0.892 (0.747-1.000)	83.3 (35.9-99.6)	92.5 (84.4-97.2)	45.5 (15.6-78.0)	98.7 (92.8-100.0)	11.1 (7.7-16.0)	0.2 (0.03-1.3)	0.3998
	R7	0.925 (0.837-1.000)	83.3 (35.9-99.6)	95.0 (87.7-98.6)	55.6 (21.2-86.3)	98.7 (92.9-100.0)	16.7 (11.6-23.9)	0.2 (0.02-1.3)	0.8119
	R8	0.942 (0.874-1.000)	83.3 (35.9-99.6)	95.0 (87.7-98.6)	55.6 (21.2-86.3)	98.7 (92.9-100.0)	16.7 (11.6-23.9)	0.2 (0.02-1.3)	0.2032
	R9	0.913 (0.740-1.000)	83.3 (35.9-99.6)	100.0 (95.5-100.0)	100.0 (47.8-100.0)	98.8 (93.3-100.0)	0 (-)	0.2 (0.02-1.3)	0.8690
	R10	0.921 (0.829-1.000)	83.3 (35.9-99.6)	93.8 (86.0-97.9)	50.0 (18.7-81.3)	98.7 (92.8-100.0)	13.3 (9.3-19.2)	0.2 (0.02-1.3)	-
Testing cohort n = 86	R1	0.831 (0.510-1.000)	83.3 (35.9-99.6)	97.5 (91.3-99.7)	71.4 (29.0-96.3)	98.7 (93.1-100.0)	33.3 (23.3-47.8)	0.2 (0.02-1.6)	0.3385
	R2	0.852 (0.581-1.000)	83.3 (35.9-99.6)	95.0 (87.7-98.6)	55.6 (21.2-86.3)	98.7 (92.9-100.0)	16.7 (11.6-23.9)	0.2 (0.02-1.3)	0.3150
	R3	0.842 (0.588-1.000)	83.3 (35.9-99.6)	96.3 (89.4-99.2)	62.5 (22.1-92.7)	98.7 (93.0-100.0)	22.2 (15.5-31.9)	0.2 (0.02-1.4)	0.1651
	R4	0.856 (0.710-1.000)	66.7 (22.3-95.7)	93.8 (86.0-97.9)	44.4 (13.7-78.8)	97.4 (90.9-99.7)	10.7 (6.0-18.8)	0.4 (0.09-1.5)	0.2605
	R5	0.866 (0.779-1.000)	83.3 (35.9-99.6)	78.8 (68.2-87.1)	22.7 (7.8-45.4)	98.4 (91.5-100.0)	3.9 (2.7-5.7)	0.2 (0.03-1.3)	0.0941
	R6	0.867 (0.661-1.000)	83.3 (35.9-99.6)	95.0 (87.7-98.6)	55.6 (21.2-86.3)	98.7 (92.9-100.0)	16.7 (11.6-23.9)	0.2 (0.02-1.3)	0.1181
	R7	0.865 (0.608-1.000)	83.3 (35.9-99.6)	98.8 (93.2-100.0)	83.3 (31.1-99.8)	98.7 (93.2-100.0)	66.7 (46.6-95.4)	0.2 (0.01-2.4)	0.3624
	R8	0.900 (0.717-1.000)	83.3 (35.9-99.6)	96.3 (89.4-99.2)	62.5 (22.1-92.7)	98.7 (93.0-100.0)	22.2 (15.5-31.9)	0.2 (0.02-1.4)	0.5429
	R9	0.881 (0.667-1.000)	83.3 (35.9-99.6)	97.5 (91.3-99.7)	71.4 (29.0-96.3)	98.7 (93.1-100.0)	33.3 (23.3-47.8)	0.2 (0.02-1.6)	0.3201
	R10	0.910 (0.748-1.000)	83.3 (35.9-99.6)	96.3 (89.4-99.2)	62.5 (22.1-92.7)	98.7 (93.0-1.0)	22.2 (15.5-31.9)	0.2 (0.02-1.4)	-

R1: B-mode US images; R2: B-mode US images + sex + age; R3: B-mode US images + serological data; R4: 2D-SWE images; R5: 2D-SWE images + sex + age; R6: 2D-SWE images + serological data; R7: B-mode US images + 2D-SWE images; R8: B-mode US images + 2D-SWE images + sex + age; R9: B-mode US images + 2D-SWE images + serological data; R10: B-mode US images + 2D-SWE images + sex + age + serological data; Data in parentheses are 95% CIs. *P* values were calculated between R10 and other models in training, validation and testing cohorts. AUC, area under the receiver operating characteristic curve; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value; PLR, positive likelihood ratio; NLR, negative likelihood ratio.

Supplementary Table 6. Diagnostic performance of HCC-R and LSM for predicting HCC in different AST and ALT levels

		N	AUC	SEN (%)	SPE (%)	PPV (%)	NPV (%)	PLR	NLR	p
HCC-R	AST > 2×ULN	53	0.944 (0.844-0.988)	100.0 (39.8-100.0)	91.8 (80.4-97.7)	50.0 (13.9-86.1)	100.0 (92.1-100.0)	12.3 (11.3-13.3)	0 (-)	0.859
	AST ≤ 2×ULN	381	0.936 (0.907-958)	89.3 (71.8-97.7)	96.3 (93.8-98.0)	65.8 (48.4-80.5)	99.1 (97.5-99.8)	24.2 (21.3-27.6)	0.1 (0.03-0.4)	
	ALT > 2×ULN	110	0.974 (0.925-0.995)	94.4 (72.7-99.9)	96.7 (93.6-98.6)	68.0 (46.5-85.1)	99.6 (97.7-100.0)	28.8 (25.7-32.3)	0.1 (0.01-0.04)	0.208
	ALT ≤ 2×ULN	324	0.927 (0.893-0.953)	100.0 (63.1-100.0)	96.1 (90.3-98.9)	66.7 (33.4-90.8)	100.0 (96.3-100.0)	25.5 (24.5-26.5)	0 (-)	
LSM	AST > 2×ULN	53	0.582 (0.438-0.716)	75.0 (19.4-99.4)	63.3 (48.3-76.6)	14.3 (3.0-36.3)	96.9 (83.8-99.9)	2.0 (1.1-3.7)	0.4 (0.1-2.2)	0.433
	AST ≤ 2×ULN	381	0.734 (0.687-0.778)	67.9 (47.6-84.1)	71.7 (66.7-76.3)	16.0 (9.9-23.8)	96.6 (93.6-98.4)	2.4 (1.8-3.1)	0.5 (0.3-0.8)	
	ALT > 2×ULN	110	0.689 (0.593-0.774)	62.5 (24.5-91.5)	73.5 (63.9-81.8)	15.6 (5.2-33.1)	96.2 (89.2-99.2)	2.4 (1.4-4.1)	0.5 (0.2-1.3)	0.726
	ALT ≤ 2×ULN	324	0.732 (0.681-0.780)	54.2 (32.8-74.4)	86.0 (81.6-89.7)	23.6 (13.2-37.0)	95.9 (92.8-97.9)	3.9 (2.7-5.6)	0.5 (0.3-0.9)	

HCC-R: B-mode US images + 2D-SWE images + sex + age. Data in parentheses are 95% CIs. *P* values were calculated between different AST and ALT levels. LSM, liver stiffness measurement; AST, aspartate aminotransferase; ALT, alanine aminotransferase; ULN, upper limit of normal; N, number of patients; AUC, area under the receiver operating characteristic curve; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value; PLR, positive likelihood ratio; NLR, negative likelihood ratio.

Supplementary Table 7. Diagnostic performance of HCC-R for predicting HCC in different antiviral therapy situations before LSM

		N	AUROC	SEN (%)	SPE (%)	PPV (%)	NPV (%)	PLR	NLR	P
HCC-R	Antiviral therapy	76	0.987 (0.930-1.000)	100.0 (76.8-100.0)	95.2 (86.5-99.0)	82.4 (55.6-96.5)	100.0 (93.9-100.0)	20.7 (19.5-21.9)	0 (-)	0.081
	No antiviral therapy	358	0.906 (0.871-0.934)	83.3 (58.6-96.4)	95.9 (93.2-97.7)	51.7 (32.2-70.9)	99.1 (97.4-99.8)	20.2 (16.4-24.9)	0.2 (0.1-0.6)	
LSM	Antiviral therapy	76	0.713 (0.598-0.811)	57.1 (28.9-82.3)	80.7 (68.6-89.6)	40.0 (19.1-63.9)	89.3 (78.0-96.0)	3.0 (1.8-4.7)	0.5 (0.2-1.2)	0.960
	No antiviral therapy	358	0.718 (0.669-0.764)	72.2 (46.5-90.3)	66.8 (61.5-71.8)	10.3 (5.6-17.0)	97.8 (95.0-99.3)	2.2 (1.6-2.9)	0.4 (0.2-0.9)	

HCC-R: B-mode US images + 2D-SWE images + sex + age. Data in parentheses are 95% CIs. *P* values were calculated between different AST levels and ALT levels. LSM, liver stiffness measurement; N, number of patients; AUROC, area under the receiver operating characteristic curve; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value; PLR, positive likelihood ratio; NLR, negative likelihood ratio.

References

[1] Yuen MF, Tanaka Y, Fong DY, Fung J, Wong DK, Yuen JC, But DY, Chan AO, Wong BC, Mizokami M and Lai CL. Independent risk factors and predictive score for the development of hepatocellular carcinoma in chronic hepatitis B. J Hepatol 2009; 50: 80-88.

[2] Hsu YC, Yip TC, Ho HJ, Wong VW, Huang YT, El-Serag HB, Lee TY, Wu MS, Lin JT, Wong GL and Wu CY. Development of a scoring system to predict hepatocellular carcinoma in Asians on antivirals for chronic hepatitis B. J Hepatol 2018; 69: 278-285.

[3] Fung J, Cheung KS, Wong DK, Mak LY, To WP, Seto WK, Lai CL and Yuen MF. Long-term outcomes and predictive scores for hepatocellular carcinoma and hepatitis B surface antigen seroclearance after hepatitis B e-antigen seroclearance. Hepatology 2018; 68: 462-472.

[4] Wong VW, Chan SL, Mo F, Chan TC, Loong HH, Wong GL, Lui YY, Chan AT, Sung JJ, Yeo W, Chan HL and Mok TS. Clinical scoring system to predict hepatocellular carcinoma in chronic hepatitis B carriers. J Clin Oncol 2010; 28: 1660-1665.

[5] He K, Zhang X, Ren S and Sun J. Deep residual learning for image recognition. 2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr) 2016: 770-778.

[6] Yosinski J, Clune J, Bengio Y and Lipson H. How transferable are features in deep neural networks? Advances in Neural Information Processing Systems 27 (Nips 2014) 2014; 27.

[7] Wu G, Chen Y, Wang Y, Yu J, Lv X, Ju X, Shi Z, Chen L and Chen Z. Sparse representation-based radiomics for the diagnosis of brain tumors. IEEE Trans Med Imaging 2018; 37: 893-905.

[8] Chang CC and Lin CJ. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2011; 2: 1-27.