Supplementary material

Image Quantification

Radiomic Feature Extraction

Conventional radiomics features include shape, histogram-based, gray level co-occurrence matrix based, gray level run length matrix based, gray level size zone matrix based, gray level dependency matrix based, and neighboring gray-tone difference matrix based features, as standardized by the Imaging Biomarker Standardization Initiative were extracted using NovoUltrasound Kit (NUK, version 1.5.0, GE Healthcare Shanghai). Laplacian of Gaussian filter (σ = 0.1, 0.2, 0.3, 0.4, 0.5), wavelet transform, and two-dimensional local binary pattern transform (radius = 1) were applied to the original ultrasound images for extraction of higher-order features. A bin width of 2 was used for grayscale discretization.

Deep Learning-based Radiomic Feature Extraction

Deep Learning-Based Radiomics features were extracted from ResNet50 and Swin transformer pre-trained on the ImageNet dataset. The ResNET50 contains a convolutional layer, four ResNet blocks including 3, 4, 6, and 3 residual layers, respectively, an average pooling layer, and a fully connected layer. The four ResNet blocks stack in a hierarchical way, which extracts image features from the texture-related features to semantic features. The last average pooling and fully connected layers were removed for deep learning feature extraction. Four average pooling layers were used to convert feature maps to feature vectors in the four ResNet blocks. Then the ROI of the tumor area was resized and fed to the adjusted ResNet50. Finally, the four feature vectors were concatenated to form deep learning-based radiomics features. Workflow of DL-based features extraction was shown in Figure 2.

Prognostic Model Construction and Validation

Development of Deep Learning-based Radiomics Score

Firstly, the univariate Cox proportional hazard regression analysis was constructed to select the significative radiomics and deep-learning features related to survival time and status, and the features with Harrell's concordance index (C-index) , were used for further survival analysis. Then, L1-penalized logistic regression Cox regression was applied to identify a subset of features with the best prognostic value and least inter-feature correlations. The penalty term α for L1-penalized logistic regression Cox regression was tuned by five-fold internal cross-validation. The remained features were then fitted to a multivariate Cox model with stepwise selection. The survival hazard of each case was calculated based on the final model, denoted as the deep learning radiomics score.

Prognostic Model Using Clinical variables and DL Radiomic Score

To refine prognostic outcome predictions, prognostic values of clinical variables and DL radiomic score were evaluated. Candidate clinical variables are age, gender, HBsAg-positive, alpha-fetoprotein, carcinoembryonic antigen, carbohydrate antigen 125, carbohydrate antigen 199, tumor long and shot axis length, echotexture, enhancement patterns of arterial phase, portal phase and late phase, presence of enhancing capsules, presence of multiple lesions, presence of satellite nodules, unsmooth tumor margins and early reoccurrence defined by tumor progression within one year after surgery. The clinical and DL radiomic score combined model was constructed through multivariate Cox regression with backward stepwise selection.

Validation of Prognostic Models

Patients were stratified into high-risk and low-risk groups based on their corresponding survival hazards. The stratification threshold was determined by X-tile software in the training cohort. Kaplan-Meier curves were plotted for each risk group to observe patient survival behaviors. Harrell's C-index was used to measure the prognostic performance of the DL radiomic score/prognostic model. A C-index closer to 1 indicates excellent performance. To quantify the relative improvement in prediction accuracy, Net Reclassification Improvement was calculated. The overall performance of these models is evaluated by prediction error curves and composite Brier scores.

Construction and Validation of Model for Early Reoccurrence Prediction

For prediction of early reoccurrences, conventional and deep learning-based radiomic features were examined with univariate logistic regression. Maximum relevance minimum redundancy algorithm was applied to the feature set consisting of significant features from univariate analysis to produce a smaller subset of features critical to early reoccurrence prediction. L1-penalized logistic regression was then utilized to this feature subset for further feature selection and shrinkage. The final features were incorporated in a multivariate logistic regression with stepwise selection to produce the predictive model for early reoccurrence. The model's output signifies probability of having tumor progression within one year after initial treatment, denoted as the DL radiomics reoccurrence score. Similar to the construction of the prognostic model, clinical variables were combined with DL reoccurrence score using a multivariate logistic regression with a backward approach.

The model's prediction performance was evaluated using the area under the receiver (AUC) operating characteristic curve (ROC), which indicates the sensitivity and false-positive rates with different probability thresholds. A decision curve analysis was constructed to assess the

clinical benefit of the models. The decision probability threshold above which the patient would be considered to have an early reoccurrence was determined by maximizing Youden's index on the ROC curve. Accuracy, sensitivity, and specificity calculated at the decision probability threshold were used for diagnostic evaluation.

We also assessed the ability of the DL radiomics model to improve the ability of 3 clinicians (11 years experience; 5 years experience; 2 years experience respectively) to predict ER, with or without the assistance of the DL radiomics model. To demonstrate the impact of the DL radiomics model on clinician-individualized assessment performance, 3 clinicians independently reassessed each patient's ER status on the same day after accounting for the DL radiomics model predictions.

Supplementary Table 1 Relevant factors for construction model of early recurrence

Reoccurrence Model	Cox Regression Model			
	OR	[0.025	0.975]	P
Clinical Model				
Satellite Nodules	216.134	9.180	5089.831	0.001
Constant	0.068	0.014	0.342	0.001
Clinical + Radiomics				
DL Radiomics Reoccurrence Score	132.847	33.818	521.652	<0.001
Satellite Nodules	340.495	10.247	11304.996	0.001
Constant	0.004	0.001	0.033	<0.001

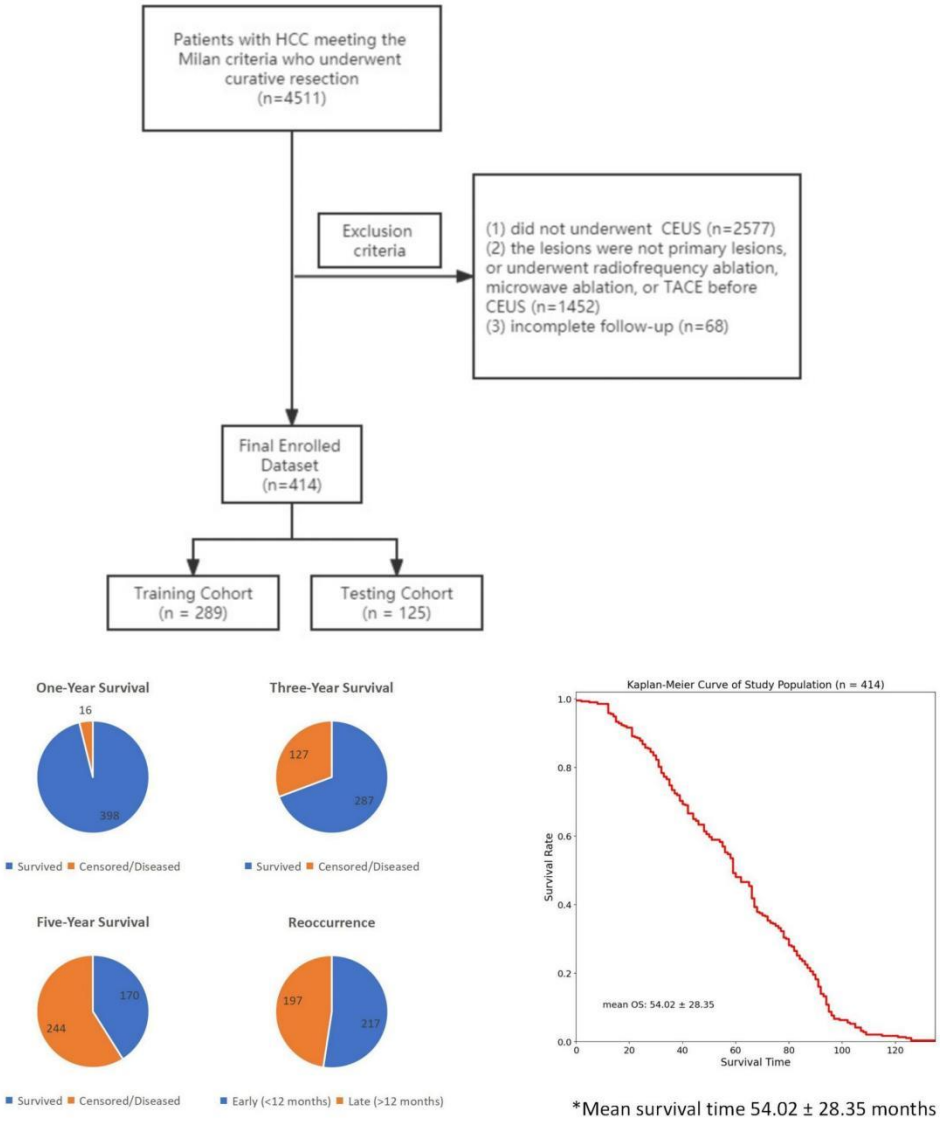
DL Radiomics, Deep Learning based Radiomics.

Supplementary Table 2 Relevant factors for construction model of overall survival

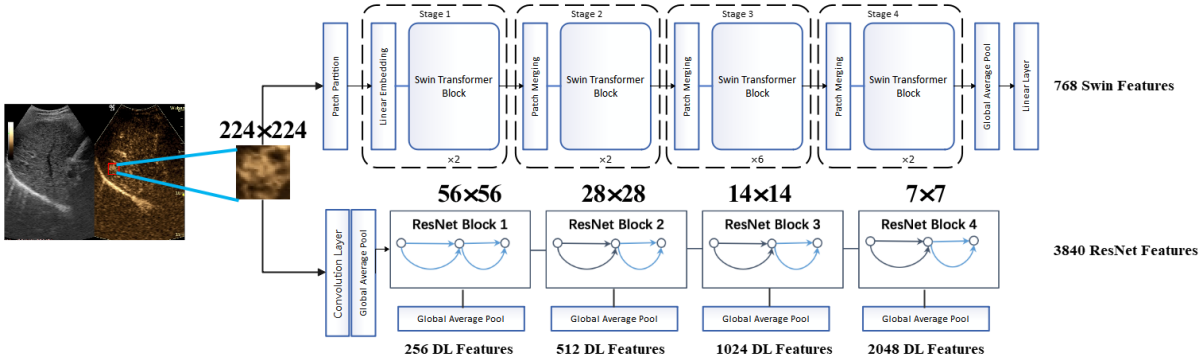
Survival Model	Cox Regression Model			
	OR	[0.025	0.975]	P
Clinical Model				
Age	1.02	1.01	1.03	<0.005
CA119	0.64	0.42	0.98	0.04
Tumor Size y	1.13	1.04	1.22	<0.005
Echogenicity	0.74	0.59	0.94	0.01
Clinical + Radiomics				
Age	1.01	1	1.03	0.02
CA119	0.6	0.38	0.92	0.02
Tumor Size y	1.11	1.03	1.19	0.01
Echogenicity	0.82	0.65	1.04	0.1
DL Radiomics	4.33	3.45	5.45	<0.005

CA199, Carbohydrate antigen199; DL Radiomics, Deep Learning based Radiomics.

Figure Legends



SupplementaryFigure 1. Flow chart for patient selection.



SupplementaryFigure 2. Development flow chart of the deep learning-based radiomics model.