**Supplementary Method**

**PET/CT Imaging Protocol**

18F-FDG PET/CT scans were performed using a Discovery ST-16 (GE Medical Systems, Milwaukee, WI, USA) at our center. Imaging was performed using a combination PET/CT scanner according to PET/CT tumor imaging guidelines.First, patients fasted for 6 h prior to the PET/CT scan. Iohexol (Omnipaque; GE Healthcare) was used as the oral contrast agent with 500 ml water. The blood glucose levels obtained immediately before 18F-FDG injection were maintained at <11.1mmol/L to all patients. Then, the imaging was performed 45-60 mins after injection of 3.7 Mbq/kg of body weight (0.1 mCi/kg) of 18F-FDG. Next, a low-dose multislice CT scan was acquired using a 16-slice multidetector scanner (parameters: 180-250 mA, 140 kV, pitch 1.375 mm, and slice thickness 3.75 mm) of areas from the skull base to the upper thigh base with the patient in shallow breathing. All patients’ arms were up because arms downing may result in artifact of the liver FDG signal due to beam-hardening effects. The standard uptake value (SUV) was calculated using the body weight. A standard whole-body PET scan was performed in 2D mode with an acquisition time of 3 mins per bed position (six-eight bed positions) covering the same field as the CT scan. The acquired data were reconstructed using an iterative algorithm, and CT images without contrast enhancement were acquired for attenuation correction. Finally, the acquisition data were transferred to a workstation (AW Server 2.0; GE Healthcare) for processing and interpretation. The CT, PET and coregistered PET/CT images were evaluated by two experienced nuclear medicine physicians in transaxial, coronal and sagittal views. Difference was solved by consensus.

**Deep learning feature extraction**

Since medical images of different anatomical structure exhibit large variation, which could affect the feature representation seriously, we constructed four different deep convolutional neural networks (DCNNs) for deep learning feature extraction on the four groups of ROIs respectively. As shown in **Figure S3**, we trained the models to recognize whether recurrence, metastasis or death happened to make the features captured by DCNNs relevant to the prognostic information of patients based on large amount of 2D patches extracted from the slices. Then, for feature extraction, the PET/CT slices were input into the hierarchical convolutional structure, and several feature maps were generated after the final convolutional layer of the DCNNs. Subsequently, the statistical characteristics of feature maps were quantified as the deep learning features. The implementation details are introduced below.

The DCNNs for CT image had three convolutional layer groups, followed by a dropout layer and a 2-way fully-connected layer with softmax classifier. Each convolutional layer group was composed of four convolutional layers and a max pooling layer. The 2×2 max pooling was used in the first and second convolutional layer group, while the global max pooling was used in the final convolutional layer group. The convolutional layers all had filters with kernel size 2×2, stride size 1 and pad size 1. Batch normalization (BN) was used after convolution for faster convergence and controlling overfitting (1). The LeakyReLU with a negative slop coefficient of 0.01 was used as nonlinearity activation function to address the problem of gradient vanishing (2). The number of kernels of the convolutional and pooling layers was set to 16 in the first and second convolutional layer group, and was set to 8 in the final convolutional layer group. Considering the low pixel resolution and the small image size of PET images, the DCNNs for PET image only had two convolutional layer groups.

In the training step, the training dataset was composed of the patches extracted from the CT images or the PET images, while ensuring the segmented ROI covered one-third at least. The data augmentation strategies of random rotation, translation and transposition were used. The imbalance between the classes was eliminated during data augmentation. We randomly selected 20% cases from the training set to tune the hyper-parameters of the model. Finally, the labeled training dataset contained around 47,000 64×64 patches for CT tumor ROI, around 73,000 64×64 patches for CT lymph node ROI, around 32,000 16×16 patches for PET tumor ROI and around 59,000 16×16 patches for PET lymph node ROI. The network parameters were initialized by Xavier (3) and were optimized using stochastic gradient descent optimizer. Each mini-batch contained 256 patches. L2 regularization was used to control overfitting. The values of the L2 regularization coefficients, the dropout rate and the learning rate were selected based on the 20% training cases. We stopped training when the validation performance did not increase after 5 epochs.

For feature extraction, we input the informative slice (i.e. the slice that had the discriminative information for prediction) into the hierarchical convolutional structure and generated 8 feature maps after the final convolutional layer. Then, the deep learning features were extracted by quantifying the characteristics of ROI in these feature maps. To discover the informative slice, we extracted patches from five slices which had the largest segmented regions, ensuring the patches covered the ROIs as much as possible, and inputted them into the DCNN successively. As the DCNN has peaky responses on the discriminative patches, the patch which yielded the highest prediction value was selected. We extracted deep learning features from not a single patch but the whole segmented ROI in the slice to obtain more robust and more significant features. Noted that we calculated not only the maximum value but also other statistical values, which were the same as the “Histogram features” described in the “Hand-crafted feature extraction” section below. Therefore, there were a total of 136 (i.e. 8×17) deep learning features extracted from each ROI.

**Hand-crafted feature extraction**

A filtering process was performed to implement image smoothing and image difference. Separable filtering was used to avoid the multi-dimensional convolution. The convolution was performed with a low-/high-pass “Coiflet 1” wavelet filter along x-/y-direction separately. Consider *L* and *H* to be a low-pass and high-pass functions respectively, *X* to be the original PET/CT image and the filtered results of *X* to be labelled as and . Two new images were obtained by filtering the original image.

For each ROI, a total of 133 3D quantitative imaging features were extracted based on the original image and its corresponding filtered image, most of which were referenced from (4-6). The formulas for computing these features were described in detail below. For feature standardization, all the extracted features of each patient were standardized by z-score method based on the parameters calculated in the training set.

Part 1: Shape features

Shape features describe the morphological property of the ROI and were generated from only the image without filtration.

1. Maximum radius:

The maximum 3D tumor radius is measured as the half of the largest pairwise Euclidean distance, between voxels on the boundary of the tumor region.

1. Volume:

The area of the region is determined by counting the number of pixels in the region.

1. Surface area:

The Surface area of the region is determined by counting the number of pixels on the boundary.

1. Surface area to volume ratio:

Part 2: Histogram features

Histogram features describe the gray level distribution of the image and were generated from the image without/after filtration.

1. Energy:
2. Entropy:
3. Standard entropy:
4. Kurtosis:
5. Maximum:

The maximum intensity value of .

1. Mean:
2. Mean absolute deviation:

The mean of the absolute deviations of all voxel intensities around the mean intensity value.

1. Median:

The median intensity value of .

1. Minimum:

The minimum intensity value of .

1. Mass:

The sum intensity value of .

1. Range:

The range of intensity values of .

1. Root mean square (RMS):
2. Skewness:
3. Standard deviation:
4. Uniformity:
5. Standard uniformity:
6. Variance:

where indicates the image with voxels and is the mean of . denotes the histogram divided by discrete intensity levels.

Part 3: Gray-Level Co-Occurrence Matrix (GLCM) features

Gray-Level Co-Occurrence Matrix (GLCM) features were generated from the image without/after filtration.

A GLCM is defined as , a matrix with size describing the second-order joint probability function of an image, where the th element represents the number of times the combination of intensity levels and occur in two pixels in the image, that are separated by a distance of pixels in direction , and is the number of discrete gray level intensities (4,7).

In this study, distance was set to 1 and direction to each of the 4 directions in slices, yielding a total of 4 gray level co-occurrence matrices for each image. From these gray-level co-occurrence matrices, several textural features were derived. Each GLCM based feature was then calculated as the mean of the feature calculations for each of the 4 directions.

Let:

be the co-occurrence matrix for an arbitrary and ,

be the number of discrete intensity levels in the image,

be the mean of ,

be the mean of ,

be the mean of ,

be the standard deviation of ,

be the standard deviation of ,

1. Autocorrelation:
2. Cluster prominence:
3. Cluster shade:
4. Cluster tendency:
5. Contrast:
6. Correlation:
7. Dissimilarity:
8. Energy:
9. Entropy:
10. Homogeneity 1:
11. Homogeneity 2:
12. Variance:

Part 4: Gray-Level Run-Length Matrix (GLRLM) features

Gray-Level Run-Length Matrix (GLRLM) features were generated from the image without/after filtration.

A gray level run is defined as the length in number of pixels, of consecutive pixels that have the same gray level value. In a GLRLM , the th element describes the number of times a gray level appears consecutively in the direction specified by , and is the number of discrete gray level intensities(4,8).

In this study, a GLRLM was computed for every slice, from which the below textural features were derived. Each GLRLM feature was then calculated as the mean of the feature values for each of the x-/y-directions.

Let:

be the th entry in the given run-length matrix for a direction ,

the number of discrete intensity values in the image,

the number of different run lengths,

the number of voxels in the image.

1. Short run emphasis (SRE):
2. Long run emphasis (LRE):
3. Gray level non-uniformity:
4. Run length non-uniformity:
5. Run percentage:
6. Low gray level run emphasis (LGLRE):
7. High gray level run emphasis (HGLRE):
8. Short run low gray level emphasis (SRLGLE):
9. Short run high gray level emphasis (SRHGLE):
10. Long run low gray level emphasis (LRLGLE):
11. Long run high gray level emphasis (LRHGLE):
12. Mean:
13. Energy:
14. Entropy:

**Feature selection and radiomics signature building**

We built two radiomics signatures reflecting the phenotypic characteristics of the primary tumor and the lymph nodes in CT and PET images respectively as independent predictors of disease-free survival (DFS), i.e. the CT-based signature and the PET-based signature. In order to avoid model over-fitting and improve the performance, we implemented feature selection and signature building described below to match the sample size.

Firstly, the inter/intra-class correlation coefficients (ICCs) were used to evaluate the intra- and inter-observer agreement of features extraction. An ICC greater than 0.8 presents good agreement. Only the stable and reproducible features were remained.

Secondly, univariate analysis was performed for each feature. Features with *P* < 0.1 were considered to be associated with DFS potentially and were selected into the following process (9).

Thirdly, the Pearson correlation coefficients (hereafter denoted r) between each pair of features were computed to analyze the linear correlation. The features were divided into different groups to ensure all pairs of features in a group had a |r| greater than 0.8. To remove the redundant features, only the most important prognostic feature of each group (i.e. the feature that yielded the lowest *P* value in univariate analysis) was remained.

Finally, the least absolute shrinkage and selection operator (LASSO) Cox regression method, which is suitable for the regression of high-dimensional data, was used to obtain the most useful prognostic combination of features. The radiomics score (Rad-score) was computed for each patient through a linear combination of selected features weighted by their respective coefficients and was used as the radiomics signature in this study.

**Statistical Analysis**

The differences of clinical characteristics between the training and test sets were assessed by Mann-Whitney U test for continuous variables and Fisher exact test or chi-square test for categorical variables. Univariate analysis was employed to assess the possible association of each predictor with the DFS. The relative hazard ratio (HR) and its 95% confidence interval (CI), and *p* values for chi-square according to likelihood-ratio test were calculated. To quantify the discrimination performance, Harrell’s concordance indices (C-index) for each predictor were also computed. The C-index is the fraction of all pairs of subjects whose predicted survival times are correctly ordered (i.e., concordant with actual survival times). C-index=0.5 indicates that the model is not better than random chance, and C-index=1 indicates that the model has perfect predictive accuracy. A statistical test was used to determine if the C-index estimate was significantly different from 0.5.

Multivariable Cox proportional hazard analysis, beginning with the statistically significant clinical characteristics and radiomics signatures, was conducted to develop an individually prognostic model for DFS, named as radiomics nomogram. Backward step-wise selection was applied by using the likelihood ratio test with Akaike’s information criterion (AIC) as the stopping rule. Moreover, to compare the prognostic value of radiomics signatures with pre-DNA, we also developed two clinical nomograms, one using clinical factors without pre-DNA (nomogram A) and another using clinical factors with pre-DNA (nomogram B). The radiomics nomogram was defined as nomogram C.

To compare the agreement between the actual outcomes and the predicted outcomes of the radiomics nomogram, the calibration curve were plotted, accompanied with the Hosmer-Lemeshow test. The prognostic or predictive accuracy of the radiomics nomogram was investigated by using time-dependent receiver operating characteristic (time-dependent ROC) analysis. We identified the threshold of the nomogram to split the patients into high-risk and low-risk groups using an algorithm of maximization of hazard ratio (10). To explore the association of the radiomics nomogram with DFS, we used Kaplan–Meier survival analysis. The curves of the high-risk and low-risk groups were compared using the log-rank test. The differences of OS, DMFS and LRRFS between the two groups were also assessed. With the concern that there may be confounding within the derived results, a stratified analysis was presented by the gender, age, overall stage and pre-DNA on the test set.

Statistical analysis was conducted with R software (version 3.4.4; <http://www.Rproject.org>) and MATLAB. A two-sided p value < 0.05 was used as the criterion to indicate a statistically significant difference.

**References**

1. Ioffe S, Szegedy C. Batch Normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of International Conference on Machine Learning. 2015;448-456.
2. Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. In Proceedings of the 30th International Conference on Machine Learning, Workshop on Deep Learning for Audio, Speech, and Language Processing. 2013;192.
3. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In Proceedings of 14th International Conference on Artificial Intelligence and Statistics. 2011;315-323.
4. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications. 2014;5:4006.
5. Huang Y, Liang C, He L, et al. Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. Journal of Clinical Oncology. 2016;34(18):2157-2164.
6. Kickingereder P, Burth S, Wick A, et al. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. Radiology. 2016;280(3):880-889.
7. Haralick RM, Shanmugam K. Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics. 1973;6:610-621.
8. Galloway MM. Texture analysis using gray level run lengths. Computer Graphics and Image Processing. 1975;4:172-179.
9. Boonstra K, Weersma RK, van Erpecum KJ, et al. Population‐based epidemiology, malignancy risk, and outcome of primary sclerosing cholangitis. Hepatology. 2013;58(6):2045-2055.
10. Contal C, O'Quigley J. An application of changepoint methods in studying the effect of age on survival in breast cancer. Computational Statistics & Data Analysis. 1999;30(3):253-270.