Table S1. Acquisitions parameters in the training and test cohorts.

| Sequence          | Training cohort and internal test cohort | External test cohort |
|-------------------|------------------------------------------|----------------------|
|                   | TR (ms) | TE (ms) | Pixel spacing (mm) | FOV (mm) | Slice thickness (mm) | TR (ms) | TE (ms) | Pixel spacing (mm) | FOV (mm) | Slice thickness (mm) |
|                   | Median (IQR) | Range   | Median (IQR) | Range   | Median (IQR) | Range   |
| Axial T1WI        | 466.0 (466.0-534.0) | 10.0 (9.4-10.0) | 0.72 (0.72-0.75) | 200×200-260×260 | 4.0-5.0 | 572.0 (572.0-572.0) | 9.5 (9.4-9.5) | 0.69 (0.69-0.69) | 160×220-200×220 | 4.0-5.0 |
| Axial T2WI FS     | 5650.0 (3007.5-7400.0) | 78.0 (72.0-78.0) | 0.73 (0.72-0.75) | 200×200-260×260 | 4.0-5.0 | 4980.0 (4710.0-4980.0) | 71.0 (71.0-71.0) | 0.52 (0.52-0.52) | 150×220-230×230 | 4.0-5.0 |
| Axial CE T1WI FS  | 466.0 (466.0-501.3) | 10.0 (10.0-10.0) | 0.72 (0.72-0.75) | 200×200-260×260 | 5.0 | 676.0 (676.0-826.0) | 9.5 (9.5-9.5) | 0.69 (0.69-0.69) | 150×220-180×230 | 3.0-5.0 |

**Abbreviations:** T1WI = T1-weighted image, T2WI FS = T2-weighted fat-suppressed image, CE-T1WI FS = Contrast-enhanced T1-weighted fat-suppressed image; TR = repetition time; TE = echo time; FOV = field of view; IQR, interquartile range.
Table S2. Network structure of ResNet18.

| stage | output   | ResNet18                      |
|-------|----------|-------------------------------|
| Conv1 | 256 × 256| conv 7 * 7, 64, stride 2      |
| Conv2 | 128 × 128| max pool, 3 * 3, stride 2     |
|       |          | [conv, 3 * 3, 64] × 2         |
| Conv3 | 64 × 64  | [conv, 3 * 3, 128] × 2        |
| Conv4 | 32 × 32  | conv, 3 * 3, 256, 2           |
| Conv5 | 16 × 16  | conv, 3 * 3, 512, 2           |
|       | 1 × 1    | average pool, 4-d fc, softmax |

Note: Each convolutional operation is followed by batch normalization operation and the activation function of all intermediate layers is ‘ReLU’ function. The default stride in all operation is 1. Conv, convolution; fc, fully connection.
| Holdout test cohort | First fold | Second fold | Third fold |
|---------------------|------------|-------------|------------|
| Selected handcrafted features | T1W_original_shape_LeastAxisLength | T2W_log.sigma.2.0.mm.3D_glcmm_ClusterShade | T1W_wavelet.HL_firstorder_Skewness |
| | T2W_wavelet.HH_glcmm_ClusterShade | T2W_log.sigma.2.0.mm.3D_glcmm_ClusterShade | T2W_log.sigma.3.0.mm.3D_glrmm_LowGrayLevelRunEmphasis |
| | T2W_wavelet.HL_glszm_SizeZoneNonUniformityNormalized | T2W_original_glszm_SmallAreaLowGrayLevelEmphasis | CE-T1W_wavelet.HH_firstorder_Skewness |
| | CE-T1W_wavelet.HH_glrmm_LongRunLowGrayLevelEmphasis | T2W_original_glszm_SmallAreaLowGrayLevelEmphasis | CE-T1W_wavelet.HH_firstorder_Skewness |
| | CE-T1W_wavelet.HH_glrmm_ShortRunLowGrayLevelEmphasis | T2W_log.sigma.3.0.mm.3D_glcmm_ClusterShade | CE-T1W_wavelet.HH_firstorder_Skewness |
| C-index (95% CI) | Training cohort | 0.775 (0.697-0.854) | p < 0.001 | 0.706 (0.618-0.795) | p < 0.001 | 0.648 (0.537-0.758) | p = 0.009 |
| | Internal test cohort | 0.795 (0.684-0.907) | p < 0.001 | 0.699 (0.537-0.861) | p = 0.016 | 0.713 (0.592-0.833) | p < 0.001 |
| | External test cohort | 0.626 (0.426-0.825) | p = 0.22 | 0.651 (0.484-0.819) | p = 0.076 | 0.546 (0.339-0.752) | p = 0.66 |

Abbreviations: C-index, Harrell's concordance indices; CI, confidence intervals; T1W, T1-weighted; T2W, T2-weighted; CE-T1W, contrast enhanced T1-weighted; val-cohort, validation cohort.
Table S4. The formula of the radiomics signature.

| Denotations of features | T1W          | T2W          | CE-T1W       |
|-------------------------|--------------|--------------|--------------|
|                         | $x_1$        | $y_1$        | $z_1$        |
|                         | original_shape_LeastAxisLength | log.sigma.2.0.mm.3D_glcM_ClusterShade | wavelet.HH_glrLm_LongRunLowGrayLevelEmphasis |
|                         | $x_2$        | $y_2$        | $z_2$        |
|                         | wavelet.HL_firstorder_Skewness | wavelet.HL_glszm_SizeZoneNonUniformityNormalized | wavelet.HH_glrLm_ShortRunLowGrayLevelEmphasis |
|                         | $y_3$        | $y_3$        | $z_3$        |
|                         | wavelet.HH_glcM_ClusterShade | wavelet.HH_glcM_ClusterShade | log.sigma.5.0.mm.3D_glcM_ClusterShade |
|                         | $y_4$        | $y_4$        | $z_4$        |
|                         | original_glszm_SmallAreaLowGrayLevelEmphasis | original_glszm_SmallAreaLowGrayLevelEmphasis | wavelet.LH_glcM_Autocorrelation |
|                         | $y_5$        | $y_5$        | $z_5$        |
|                         | log.sigma.3.0.mm.3D_glrLm_LowGrayLevelRunEmphasis | log.sigma.3.0.mm.3D_glrLm_LowGrayLevelRunEmphasis | wavelet.HH_firstorder_Skewness |

In cross-validation, when fold0 is holdout validation cohort, the first model’s signature is:

$$s_1 = 0.4429 \times x_1 - 1.0095 \times y_1 - 0.4349 \times y_2 - 0.5750 \times y_3 - 3.9568 \times z_1 + 2.6610 \times z_2$$

When fold1 is holdout validation cohort, the second model’s signature is:

$$s_2 = 0.4128 \times x_1 - 0.7420 \times y_4 - 1.3693 \times y_1$$

When fold2 is holdout validation cohort, the third model’s signature is:

$$s_3 = 0.4031 \times x_2 - 0.7458 \times y_5 - 0.4153 \times z_3 + 0.5570 \times z_4 - 0.5574 \times z_5$$

The final radiomic signature is the max value of these three cross-validation signatures:

$$Signature = \max(s_1, s_2, s_3)$$
Table S5. Univariate analysis of clinical parameters in the training cohort (n=132).

| Characteristics                                      | Hazard ratio (95% CI) | p-value‡ |
|------------------------------------------------------|-----------------------|----------|
| Age, years                                           | 1.56 (0.78-3.13)      | 0.2      |
| Sex, female vs. male                                 | 0.34 (0.12-0.98)      | 0.037*   |
| T-stage, I-II vs. III-IV                             | 2.26 (0.98-5.26)      | 0.051    |
| N-stage, 0-I vs. II-III                              | 2.93 (1.20-7.12)      | 0.013*   |
| Overall stage, I-III vs. IV                          | 1.89 (0.93-3.83)      | 0.073    |
| Smoking index, < 15 pack-years vs. >= 15 pack-years  | 1.22 (0.50-2.97)      | 0.66     |
| Family history of cancer, yes vs. no                 | 0.44 (0.10-1.84)      | 0.25     |
| Anemia, yes vs. no†                                   | 0.69 (0.32-1.49)      | 0.34     |
| TILs grading, high vs. low                           | 0.49 (0.24-1.01)      | 0.049*   |
| Presence of sarcomatoid tumor cells                  | 0.74 (0.31-1.81)      | 0.51     |
| Intra-tumor necrosis, yes vs. no                     | 1.59 (0.73-3.44)      | 0.23     |
| Lymphonode necrosis, yes vs. no                      | 1.84 (0.91-3.73)      | 0.086    |
| Lower neck involvement                               | 1.59 (0.68-3.71)      | 0.28     |
| LDH, abnormal vs. normal*                            | 1.00 (0.99-1.00)      | 0.25     |
| pEBV DNA, ≥ 4000 copy/ml vs. < 4000 copy/ml *        | 1.00 (1.00-1.00)      | 0.012*   |
| RT technique, 2DCRT or 3DCRT vs. IMRT                | 0.64 (0.28-1.46)      | 0.28     |
| Chemotherapy combination, ICT+CCRT vs the remaining  | 1.19 (0.46-3.10)      | 0.72     |

†For male, hemoglobin < 130 g/L is defined Anemia; for female, hemoglobin < 120 g/L is defined Anemia.
‡p-values are computed using the likelihood ratio test.
* A p-value with an asterisk reaches the statistically significant level.
*#p-values were calculated in subsets of 85 patients with available pretreatment LDH and 73 with pEBV DNA, respectively.

Abbreviations: CI, confidential interval; TILs, tumoral infiltrating lymphocytes; LDH, lactate dehydrogenase; pEBV DNA, plasma Epstein–Barr Virus DNA; 2DCRT, 2-dimensional conformal radiotherapy; 3DCRT, 3-dimensional conformal radiotherapy; IMRT, intensity-modulated radiotherapy; ICT, induction chemotherapy; CCRT, concomitant chemoradiotherapy.
Figure S1. The model selection by average performance of 100 random 3-fold cross-validations in the SYSU5 cohort (n=176).
Figure S2. The classification loss and the DeepSurv loss of deep convolutional neural network during histopathologic signature building in the SYSU5 cohort (n=176)  
Abbreviations: Int-test cohort, internal test cohort.
Figure S3. Kaplan-Meier survival curves with secondary endpoints for the low-risk and high-risk groups in the training, internal and external test cohorts (n=220). Abbreviations: Int-test cohort, internal test cohort; Ext-test cohort, external test cohort.
Figure S4. Stratification analysis by the clinical risk factors in the training cohort (n=132).
Figure S5. Time-independent ROC curves comparing the predictive power of different models for FFS in the training, internal and external test cohorts. ROC, receiver operator characteristic; AUC, area under the curve;
Figure S6. The pathway landscape of genetic in the biological test cohort (n=16). Abbreviations: TILs, tumoral infiltrating lymphocytes.
Figure S7. Typical images from two patients with different levels of the radiomic signatures. The tumoral ROIs were shown in red on the axial T2WI FS MRI sequences. The two patients shared similar stages and pathological signatures, while yielded distinct treatment outcomes. 

a) clinical stage T4N1M0 IVa; pretreatment pEBV DNA level 510 copy/ml; pathological signature 0.11; radiomic signature 2.66; metastasis after 6.5 months and death after 9.2 months.

b) clinical stage T4N1M0 IVa; pretreatment pEBV DNA level 5300 copy/ml; pathological signature 0.10; radiomic signature 0.84; 26.5 months failure-free survival up to the last follow-up.
Text S1. Treatment details for the training, internal and external test cohorts (n=220).

89.1% (196/220) of the patients in the training, internal and external test cohorts were treated with intensity-modulated radiotherapy (IMRT) and the remaining 10.9% (24/220) were treated with 2-dimensional or 3D conformal radiotherapy (2DCRT or 3DCRT). All patients received cumulative radiation doses as ≥66 Gy to the primary tumor and ≥50 Gy to the bilateral cervical lymph nodes and potential sites of local infiltration. All patients were treated with 30-35 fractions with five daily fractions per week for 6-7 weeks. Platinum-based chemotherapy was administered to 95.9% (211 of 220) patients, including concomitant chemoradiotherapy (CCRT), sequential chemotherapy (induction chemotherapy and/or adjuvant chemotherapy). Of these patients, 70.9% (156/220) patients received CCRT+ sequential chemotherapy and 14.5% (32/220) received CCRT only. In total, 7/220 (3.2%) patients did not received chemotherapy, with 2 (29%) having stage I-II disease, 2 (29%) patients having age more than 65 years, and the other 3 (42%) patients for refusal and/or economic reasons.
Text S2. Detail of the panel sequencing protocol.

a) Panel introduction
Oseq-508TM (commercial pan-cancer product of the Beijing Genomics Institute, BGI) is a hybridization-capture based pan-cancer panel encompassing all genes that druggable by approved therapies and crucial pathways about tumorigenesis and tumor progression. The Oseq-508TM panel we used in this project totally contains 508 genes. The capture region covering all exons of those 508 genes and partly introns. Oseq-508TM is capable of detecting single nucleotide variants (SNVs), insertions and deletions (INDEL), copy number alterations, and structural rearrangements.

b) Genomic DNA preparation
Formalin-fixed paraffin-embedded (FFPE) tumour tissue (obtained at primary diagnosis) was provided by the pathology department and ethylenediaminetetraacetic acid (EDTA) blood samples were collected as normal tissue. The genomic DNA from tissues was prepared using the QIAamp DNAFFPE Tissue (Qiagen) and blood using QIAamp DNA Mini Kit (Qiagen) following the manufacturer’s instructions. The quantity and quality of DNA were determined by agarose gel electrophoresis. Mass spectrometric fingerprint genotyping of 21 common SNPs was performed for all the samples before sequencing, to verify that paired tumour and normal tissues were derived from the same patient. 1ug of genomic DNA from each sample was fragmented using a Covarias sonication system to mean sizes of 400 bp. After fragmentation, libraries were constructed according to the Illumina Paired-End protocol.

c) Sequencing and analysis workflow
Captured DNA fragments were sequenced on MGISEQ-2000 as paired-end 100-base pair reads. The analysis pipeline start from raw data produced by sequencer. Firstly, raw data were filtered by fastq_tools-0.1.0 to eliminate low quality reads and adapter contamination. The remaining reads by this step were aligned to human genome (hg 19) using BWA-backtrack algorithms. Picard-1.98 was used to mark PCR duplications and the Genome Analysis Toolkit was utilized to revise complex mutation alignments. Only the unique reads were used to call mutations. SOMATK_snv (custom developed by BGI) were used separately to generate somatic SNVs based on paired cancer and normal samples and the results were combined together as the initial somatic SNVs. Similarly, INDELs were called by SOMATK_indel (custom developed by BGI) softwares. And CNV or fusion mutation were analyzed by som_sv (custom developed by BGI). As the germline mutations were eliminated through cancer-normal paired analysis, the mutations obtained from above analysis were somatic mutations. Mapping quality, base quality and mutation frequency were used to filter real somatic mutations with high confidence coefficient.
Text S3. Magnetic resonance images acquisition parameters.

MR images of each patient were acquired from a 3.0-Tesla system (MAGNETOM Verio, Siemens Healthcare, Germany) in the two institutions (n=214) and a 1.5-Tesla system (Avanto, Siemens Healthcare, Germany) in the Guilin Medical University Affiliated hospital (n=6). Non-contrast-enhanced [T1-weighted MR imaging (T1WI) in the axial, coronal and sagittal plane, T2-weighted fat-suppressed MR imaging (T2WI FS in the axial plane) and contrast-enhanced (contrast-enhanced T1-weighted fat-suppressed MR imaging (CE T1WI FS) in the axial, coronal and sagittal plane) images were obtained.

During the image acquisition, patients were scanned from the suprasellar cistern to the inferior margin of the sternal end of the clavicle examined, with a head-and-neck combined coil. T1WI in the axial, coronal, and sagittal planes; and T2WI FS in the axial plane were obtained before injection of contrast material. At approximately 40 seconds after intravenous Gd-DTPA injection at a dose of 0.1 mmol/kg body weight (Magnevist; Schering, Berlin, Germany), axial and sagittal CE T1WI FS were performed sequentially. The acquisition parameters are shown in Table S1.
Text S4. Details of the imaging processing.

a) **Interpolation**

The in-plane pixel spacings vary in MR images. It is showed that some radiomic features of MR images are affected by the in-plane voxel size [1]. To avoid the bias of different spacings, we applied bilinear interpolation to standardize the in-plane voxel size of MR images to 0.5 mm.

b) **Image intensity normalization**

The intensity of MR images varies when images were scanned by different scanners or scanned with different acquisition parameter settings. In order to normalize the image intensity, we performed histogram matching after bilinear interpolation [2]:

$$I_1 = \begin{cases} 
1024 + (I - m_1) \times \frac{I_{50}-1024}{I_{m_1}-1024}, & I < I_{50} \\
1024 + (I - m_2) \times \frac{I_{95}-1024}{I_{95}-I_{50}}, & I \geq I_{50}
\end{cases} \quad \text{Eq. (S1)}$$

$$I_{\text{new}} = \begin{cases} 
0, & I_1 \leq 0 \\
I_1, & 0 < I_1 < 2048 \\
2048, & I_1 \geq 2048
\end{cases} \quad \text{Eq. (S2)}$$

where $m_1 = 2048 \times 5\%$, $m_2 = 2048 \times 95\%$, $I$ denotes MR image intensity and $I_p$ denotes value of the first $p\%$ of MR image intensity.

c) **Radiomic feature selection and signature building**

For each MR sequence, 788 features were selected, including five groups: histogram, shape, gray level dependency matrix (GLDM), gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM) and gray level cooccurrence matrix (GLCM).

In feature selection, first, we selected features with intra/inter-class correlation coefficients (ICCs) $> 0.75$ to select reproducible features. Then, for each split of the 3-fold cross-validation, we divided the SYSU5 cohort data into a training set and a holdout validation set. We used univariate analysis, minimal redundancy maximum relevance and random forest to select features in the training part, and we evaluated these selected features with Harrell's concordance index (C-index) to choose the best features selection methods. Therefore, three groups of features were selected from three splits, and were used to build three Cox proportional hazard models (CPHs). The maximum value of three CPHs was set as the radiomic signature. To obtain a robust radiomic signature, we repeated cross-validations for 100 times with randomly splits. For each repetition, we calculated the average training-C-index and the average validation-C-index of three CPHs in their respective training set and the holdout validation set. We finally compared the CPHs’ C-indices with the total average C-indices, and selected the CPHs which was nearest to the total average performance by Euclidian distance.

c) **Pathological image pre-processing**

The pathological images’ stain color was normalized by a structure-preserving method [3]. Then, the original images were randomly sliced into $512 \times 512$ sub-images. Using data augmentation technique, the images in the training cohort
were randomly rotated, added by Gaussian noise, or blurred, while the images in two test cohorts were not augmented.

d) Deep learning model training

First, histopathologic images were annotated by four labels: tumor area with predominant tumor nuclei (TN), intra-tumoral necrosis (ITN, if found), intra-tumoral lymphocytes infiltrating (I-TLI) and stromal tumoral lymphocytes infiltrating (S-TLI) area. We built a 4-class deep convolutional neural network classifier as a pre-trained feature extractor using pre-trained ResNet-18 on ImageNet dataset. Due to the unbalanced distribution of annotated labels (S-TLI: 1850 [21.7%], TN: 5798 [68.2%], I-TLI: 521 [6.0%], ITN: 337 [4.0%]), we fed each mini-batch with fixed distribution of labels (S-TLI vs. TN vs. I-TLI vs. ITN = 3:3:2:2) into the network in the training phase to avoid the model overfitting to a certain label.

Then, we kept the structure of this pre-trained model, except the last fully-connected layer, and replaced it with a two fully-connected layers, where the first one is a 238-node dropout layer with 0.36 dropout probability, and the second one is a one-node output layer. We fixed the weights of this model except the last two layers and trained it to fit FFS using the DeepSurv loss [4] and a multiple instance learning method [5]. Each sub-image of a patient is an instance, and only the sub-image with the maximal output value was used to back-propagation.

Finally, we fine-tuned all weights in this model with smaller learning rate to fit FFS with the multiple instance learning method. Model training was performed on the training cohort, and hyperparameter setting was performed on the internal test cohort. The output value of the fused model was set to be the pathological signature.

All deep learning models were trained on a workstation equipped with 64 Intel Xeon Gold 6130 CPU @ 2.10GHz (Intel; Santa Clara, CA) and one GPU of Nvidia Titan Xp (Nvidia; Santa Clara, CA).

e) statistical analysis

The “survival” package was used for Cox proportional hazards regression while the “rms” for nomograms and calibration curves. The “survcomp” package was used to calculate C-indices and hazard ratios and the “survivalROC” package was used to plot the time-independent ROC curves. The cut-off of continuous variables was determined by the “SurvMisc” package.

Reference

1. A Universal Intensity Standardization Method Based on a Many-to-One Weak-Paired Cycle Generative Adversarial Network for Magnetic Resonance Images
2. Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions
3. Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images
4. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network
5. Deep Multiple Instance Learning for Image Classification and Auto-annotation