Supplemental material

Supplemental Methods

1. MRI imaging acquisition

All images were acquired by 1.5-T Siemens Healthineers (Munich, Germany) MRI scanners. The median value of in-plane resolutions was 0.81 mm with a range from 0.31 to 1.88 mm. The MRI slice thickness was 5mm with reconstructed image matrix size of 512 * 512. The repetition time (TR) and echo time (TE) was 550 ms (median, range: 298-770) and 18 ms (median, range: 4-28) respectively for T1WI sequence, and 4095 (median, range: 1086-10098) and 85 ms (median, range: 30-120) respectively for T2FS sequence.

2. Neoadjuvant chemotherapy regimens

The main chemotherapeutic agents for osteosarcoma in this study include adriamycin (ADM), cisplatin (DDP), high dose methotrexate (HDMTX), ifosfamide (IFO), vincristine (VCR) and dacarbazine (DTIC). For almost 20 years, high-dose and multi-drug neoadjuvant chemotherapy has been the gold standard for osteosarcoma treatment. The NAC regimens included Coss protocol (combined use of MTX, DDP and ADM, with or without IFO), MAID protocol (consisting of MTX, ADM, IFO, and DTIC) or T12 protocol (combined use of the VCR and MTX) and HDMAP protocol (combined use of the high-dose MTX, ADM and DDP). In this study, the most common chemotherapy regimen of each hospital was the Coss regimen，consisting of ADM, DDP, HDMTX and IFO. The dose ranges and the duration of action of drugs are expressed as follows: ADM 90 mg/m2 (3w), DDP 120 ~ 140 mg/m2 (2w), MTX 8 ~ 10 g/m2 (2w), IFO 15 g/m2 (3w), and the drug duration is 4-6 cycles (2-3 months).

3. Automatic segmentation strategy

Before segmentation, pre-processing was performed on MRI image stack to focus only on the target anatomical part to perform the segmentation algorithm efficiently and faster. The bounding-box covering target anatomical part along image stack were created by thresholding operation and cropped throughout all slices programmatically. Next, images were normalized to total 256 grey levels (0-255) to facilitate further operation.

Automatic tumor segmentation was performed to delineate the whole tumor at baseline and follow-up. Unsupervised clustering-based algorithms, including Simple-lineariterative-clustering Superpixels (SLIC-S) and Fuzzy c-means clustering (FCM) 2 were used for automated tumor segmentation, as recommended by previous studies 3. SLIC-S generates supervoxels by clustering voxels based on intensity similarity and proximity in plane. FCM employs fuzzy partitioning of voxel intensities to classify the voxels into specified number of clusters with a measurement of cluster centers such that the dissimilarity measure among clusters is minimized. Tumor regions was observed as hyperintense region and within the tumor depending upon tumor heterogeneity, two to three levels of different intensities were observed. Healthy muscle tissues, bone, and background were observed to have lower signal intensity levels. SLIC-S and FCM both the algorithms were applied in MRI images to generate clusters of voxels depending upon their intensity similarity and hyperintense clusters were selected programmatically to segment the tumor volume.

4. Deep learning (DL) features

4.1 CNNs architecture

Six representative pretrained CNN architectures were applied on T1WI and T2FS sequences for the extraction of DL-based features, including Xception, VGG16, VGG19, ResNet50, InceptionV3, and InceptionResNetV2. These six CNNs were commonly used and pre-trained by the large-scale and well-annotated ImageNet database. This published research released dataset containing enormous object categories and manually annotated training images, the optimization hyperparameters of which was not tuned permitting a broader generalization on other datasets. After tumor segmentation, prepared slices of the whole tumor volume would be ready as the input of the pre-trained CNNs to generate DL-based features. The models are publicly accessed by Keras and TensorFlow open-source code (https://github.com/fchollet/deep-learning-models/).

4.2 Elimination of the last fully-connected layer

The convolutional base is connected by a fully-connected layer for the pre-trained models. We removed the last fully-connected layer, and then different CNNs reached various numbers of feature maps (2048 for ResNet50, InceptionV3 and Xception, 512 for VGG16 and VGG19, and 1536 for InceptionResNetV2) from the new output of these models.

4.3 Addition of max pooling layer and feature extraction

With the utility of a global pooling window, local data would be concentrated into a decreased dimensionality. After Step 4.2, for models with more than one-dimensional features, we got feature maps with height and width dimensions corresponding to location invariance in the input layer. After global max pooling, each feature map vector was transformed to a maximal raw value among them. During this step, the feature maps were transformed to numeric values, as representational of DL-based features.

5. Handcrafted radiomics feature

Handcrafted radiomics features were computed from the radiologist-drawn ROIs using an open-source python package PyRadiomics (version 2.1.2) on T1WI and T2FS sequences respectively. The online documentation of PyRadiomics package depicts the detailed formation of radiomics features (https://pyradiomics.readthedocs.io/en/latest/features.html). We set resampled voxel sizes as 1× 1 × 1 mm^3 voxels for the slice thickness standardization, the bin width of image intensities as 25 HU, and voxel array shift as 1000. To allow the involvement of the whole tumor and avoid interference from the air and bone tissues, segmented voxels were resampled with the range of 50 to 400 HU. Defined radiomic image features without/with wavelet filtration were extracted to interpret tumor characteristics comprehensively. Wavelet filtration filtered original image with two pass filters, high pass filter (H) and low pass filter (L) for three directions, x, y and z respectively, which represented a total of eight different combinations of decompositions. The extracted radiomics features could be classified into three groups: (a) First-order statistics; (b) Shape features; (c) Second-order statistics. Most radiomics features mentioned above showed consistency with feature definitions in accordance with the IBSI guidelines.

There are differences in gray value discretization for the fixed bin size type and resampling, both of which cannot be corrected by customization settings alone and require replacement by custom functions (shown in the Pyradiomics documents). There are two features available in PyRadiomics without definitions in the IBSI, Total Energy and Standard Deviation. Entropy in Pyradiomics is defined by IBSI named Intensity Histogram Entropy. Uniformity in Pyradiomics is defined by IBSI named Intensity Histogram Uniformity. Mesh Volume in Pyradiomics is defined named Volume. Voxel Volume in Pyradiomics is defined in IBSI named Approximate Volume. Joint Energy in Pyradiomics is defined by IBSI named Angular Second Moment. Maximum Probability in Pyradiomics

is defined by IBSI named Joint maximum. Sum of Squares in Pyradiomics is defined by IBSI named Joint Variance. The PyRadiomics kurtosis is not corrected, whereas IBSI kurtosis is corrected by -3, yielding 0 for normal distributions. Despite these features above, the remaining features are consistent with the IBSI definitions.

6. Prediction model construction for response to NAC

For model building and unbiased performance evaluation on the training set, 10 iterations of five-fold nested cross validation were performed using the code published by Deist et al. built upon the "caret" package. The total training set was split into five subgroups (outer folds). Each Subgroup was then split once more for five times (inner folds). Hyperparameters were optimized as part of the inner folds. The selected hyperparameters were then used for testing on the five outer folds. The total mean area under the receiver operator characteristic (ROC) curve (AUC) over all outer folds was calculated for model comparison. The hyperparameter combination with the best mean performance was used to retrain a final model on the whole training set. Model performance with the optimal hyperparameter combination was assess using cross validation.

7. Statistical analysis

The MRI-derived features were harmonized to reduce the multi-scanner disturbance induced by different protocols. Feature robustness for inter-observer and intra-obsercer reproducibility was assessed by intra-class correlation coefficients (ICCs) using "irr" R package. The Logistic regression for response of NAC and the Cox proportional hazard (CPH) analyses for overall survival (OS) was performed using the "survival" R package. The KM analyses with log-rank tests were used to compare survival differences on the DL-based signature and response to NAC by "survminer" R package. Discrimination abilities were tested by Harrell's concordance indices (C-index), Brier score, and calibration curves using "rms", "survival" and "Hmisc" R packages. The integrated nomogram for response to NAC was depicted by "Hmisc" package. The decision curve analysis was performed by "rmda" R package.

Supplemental Table 1 Univariate and multivariate analysis on clinical factors associated with response to neoadjuvant chemotherapy in adolescents and young adults with osteosarcoma in the training cohort

Characteristics	Univariate analysis		Multivariate analysis	
	HR (95% CI)	*P* value	HR (95% CI)	*P* value
Gender (Female: Male)	0.917 (0.406-2.073)	0.835		
Age (≤14: >14 years old)	0.121 (0.040-0.365)	<0.001	0.123 (0.040-0.381)	<0.001
Tumor size (≤10:>10 cm)	0.798 (0.311-2.052)	0.640		
Location (Extremity: Trunk)	0.373 (0.160-0.873)	0.023	0.385 (0.152-0.977)	0.045
HGB (Normal: Abnormal)	1.429 (0.576-3.541)	0.441		
Platelet (Normal: Abnormal)	0.527 (0.181-1.535)	0.241		
WBC (Normal: Abnormal)	1.127 (0.348-3.650)	0.843		
ALB (Normal: Abnormal)	1.772 (0.712-4.407)	0.219		
AKP (Normal: Abnormal)	0.713 (0.291-1.744)	0.459		
LDH (Normal: Abnormal)	0.513 (0.196-1.346)	0.175		
Duration of hospitalization (≤14: >14 days)	0.586 (0.257-1.336)	0.204		

Abbreviations: HR, Hazard ratio; CI, Confidence interval; HGB, Hemoglobin; WBC, White blood cell; ALB, Serum albumin; ALP, Alkaline phosphatase; LDH, Lactate dehydrogenase.

Supplemental Table 2 Handcrafted radiomics features selected for prediction of response to neoadjuvant chemotherapy in adolescents and young adults with osteosarcoma

MRI sequence	Index	Filter [a]	Feature class	Feature
T1	1	T1_Wavelet [c](LHH)	First order [d]	Minimum
	2	T1_Wavelet (HLL)	First order	Maximum
	3	T1_Wavelet (LLH)	First order	90 Percentile
	4	T1_Wavelet (HLH)	GLCM	Maximal Correlation Coefficient
	5	T1_Wavelet (LLL)	GLCM	Maximal Correlation Coefficient
	6	T1_Wavelet (HHL)	GLSZM	Gray Level Variance
T2	1	T2_Original [b]	GLSZM	Small Area Emphasis
	2	T2_Wavelet (HHL)	GLCM	Maximal Correlation Coefficient
	3	T2_Wavelet (LLL)	GLCM	Maximal Correlation Coefficient
	4	T2_Wavelet (HLH)	First order	Entropy
	5	T2_Wavelet (LLL)	GLDM	Dependence Non Uniformity Normalized
T1+T2	1	T1_Wavelet (LHH)	First order	Minimum
	2	T2_Wavelet (LHH)	First order	Root Mean Squared
	3	T1_Wavelet (LHL)	GLCM	SumAverage
	4	T1_Wavelet (LLL)	GLCM	Maximal Correlation Coefficient
	5	T1_Wavelet (LLL)	GLRLM	Run Entropy

Abbreviations: MRI, Magnetic resonance imaging; GLSZM, Gray Level Size Zone Matrix Features; GLCM, Gray Level Co-occurrence Matrix Features; GLDM, .Gray Level Dependence Matrix Features; GLRLM, . Gray Level Run Length Matrix.

Note:

a: HLH, HLL, HHL, HHH and LLL, representative of high pass or low pass filter on the X, Y, Z three dimensions (H, high pass filter; L, low pass filter);

b: Original, original images without any filter used;

c: Wavelet, wavelet filtrated image;

d: First order, first order statistics.

Supplemental Table 3 Number of features selected and used in construction of DL-based models for prediction of response to neoadjuvant chemotherapy in adolescents and young adults with osteosarcoma

Feature extractor	MRI sequence	Number of selected features
Xception	T1	5
	T2	8
	T1+T2	6
VGG16	T1	4
	T2	7
	T1+T2	6
VGG19	T1	3
	T2	4
	T1+T2	3
ResNet50	T1	4
	T2	7
	T1+T2	6
InceptionV3	T1	6
	T2	6
	T1+T2	6
InceptionResNetV2	T1	5
	T2	5
	T1+T2	4

Abbreviations: DL, Deep learning; SVM, Support vector machine, MRI, Magnetic resonance imaging.

Supplemental Table 4 Predictive performances of DL-based models constructed by features extracted from different layers of ResNet50 algorithm for prediction of response to neoadjuvant chemotherapy in adolescents and young adults with osteosarcoma

Layer	MRI sequence	Training cohort						Testing cohort					
		AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
Res2b	T1	0.846	77.66	66.04	92.68	92.11	67.86	0.611	60.00	36.36	88.89	80.00	53.33
	T2	0.810	77.66	73.58	82.93	84.78	70.83	0.504	50.00	27.27	77.78	60.00	46.67
	T1+T2	0.861	81.91	86.79	75.61	82.14	81.58	0.689	70.00	68.18	72.22	75.00	65.00
Res3d	T1	0.871	77.66	67.92	90.24	90.00	68.52	0.641	60.00	50.00	72.22	68.75	54.17
	T2	0.838	73.40	62.26	87.80	86.84	64.29	0.553	55.00	36.36	77.78	66.67	50.00
	T1+T2	0.889	78.72	73.58	85.37	86.67	71.43	0.583	47.50	31.82	66.67	53.85	44.44
Res4f	T1	0.945	86.17	83.02	90.24	91.67	80.43	0.753	67.50	59.09	77.78	76.47	60.87
	T2	0.914	86.17	86.79	85.37	88.46	83.33	0.503	50.00	45.45	55.56	55.56	55.56
	T1+T2	0.926	84.04	79.25	90.24	91.30	77.08	0.727	67.50	59.09	77.78	76.47	60.87
Res5c	**T1**	**0.909**	**80.85**	**83.02**	**78.05**	**83.02**	**78.05**	**0.649**	**60.00**	**59.09**	**61.11**	**65.00**	**55.00**
	T2	**0.912**	**81.91**	**75.47**	**90.24**	**90.91**	**74.00**	**0.639**	**60.00**	**63.64**	**55.56**	**63.64**	**55.56**
	T1+T2	**0.948**	**86.17**	**83.02**	**90.24**	**91.67**	**80.43**	**0.770**	**65.00**	**54.55**	**77.78**	**75.00**	**58.33**
FC1000	T1	0.828	76.60	81.13	70.73	78.18	74.36	0.679	62.50	50.00	77.78	73.33	56.00
	T2	0.818	75.53	79.25	70.73	77.78	72.50	0.513	52.50	59.09	44.44	56.52	47.06
	T1+T2	0.852	78.72	81.13	75.61	81.13	75.61	0.515	47.50	54.55	38.89	51.27	41.18

Abbreviations: DL, Deep learning; MRI, Magnetic resonance imaging; AUC, Area under the receiver operating characteristic curve; PPV, Positive predictive value; NPV, Negative predictive value.

Supplemental Table 5 Univariate and multivariate analysis on clinical factors for predicting overall survival of in adolescents and young adults with osteosarcoma in the training cohort

Characteristic	Univariate analysis		Multivariate analysis	
	HR (95% CI)	*P* value	HR (95% CI)	*P* value
Response to neoadjuvant chemotherapy	14.875 (1.999-110.694)	0.008	8.887 (1.123-70.325)	0.038
Gender	1.712 (0.730-4.018)	0.217		
Age, years	0.639 (0.187-2.185)	0.475		
Tumor size, cm	1.124 (0.410-3.084)	0.821		
Location	1.049 (0.446-2.464)	0.913		
HGB, g/L	1.964 (0.580-6.655)	0.278		
Platelet, 10^9/L	0.558 (0.184-1.698)	0.304		
WBC, 10^9/L	2.968 (0.398-22.147)	0.289		
ALB, g/L	0.955 (0.372-2.449)	0.923		
AKP, U/L	0.533 (0.220-1.291)	0.163		
LDH, U/L	0.913 (0.355-2.348)	0.851		
Duration of hospitalization, day	0.802 (0.344-1.872)	0.610		
DL-based signature	6.148 (2.564-14.741)	<0.001	3.376 (1.368-8.331)	0.008

Abbreviations: HR, Hazard ratio; CI, Confidence interval; HGB, Hemoglobin; WBC, White blood cell; ALB, Serum albumin; ALP, Alkaline phosphatase; LDH, Lactate dehydrogenase; DL, Deep learning.

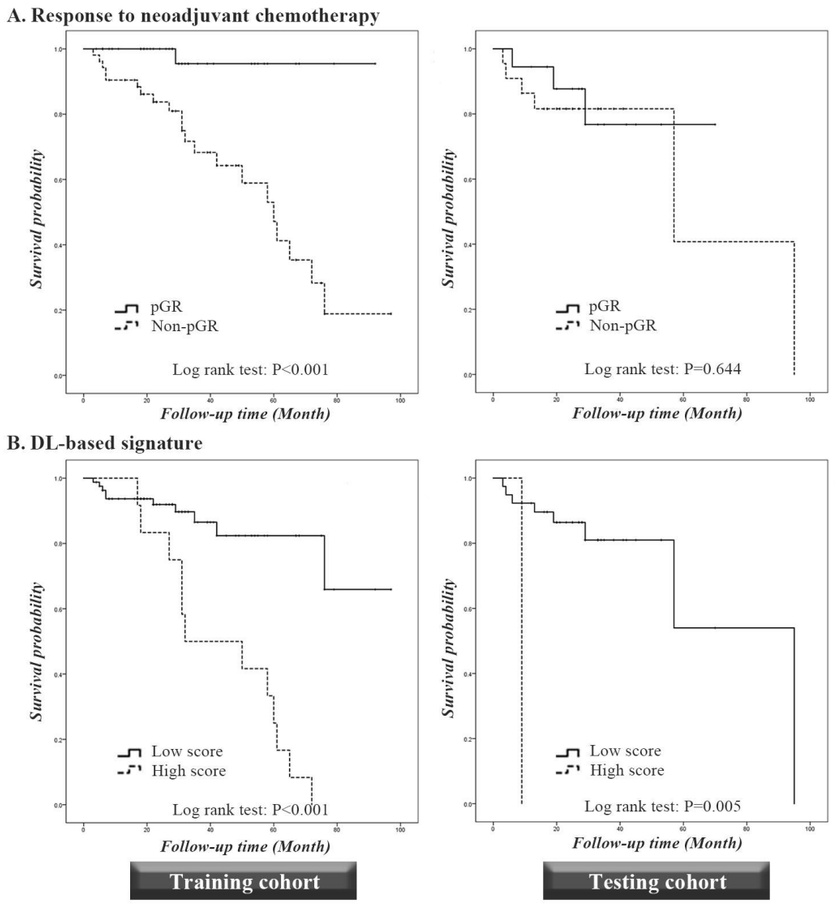
Supplemental Table 6 Kaplan-Meier analysis of response to neoadjuvant chemotherapy and the DL-based signature in association with overall survival in adolescents and young adults with osteosarcoma

Variable	Group	Training database		Testing database	
		Mean survival time (95%CI)	Log rank test	Mean survival time (95%CI)	Log rank test
Response to neoadjuvant chemotherapy	pGR	89.136 (83.653-94.620)	<0.001	58.509 (46.818-70.201)	0.644
	Non-pGR	55.945 (45.707-66.183)		63.341 (36.791-89.891)	
DL-based signature [a]	Low score	80.698 (70.986-90.411)	<0.001	69.545 (49.284-89.806)	0.005
	High score	43.500 (32.477-54.523)		9.000 (9.000-9.000)	

Abbreviations: DL, Deep learning; CI, Confidence interval; pGR, Pathological good response.

Note：

a: The DL-based signature was derived from the best DL-based model considering AUC of the testing cohort from the prediction of response to neoadjuvant chemotherapy.



Supplemental Figure 1 Kaplan-Meier analysis of response to neoadjuvant chemotherapy and the DL-based signature in association with overall survival in adolescents and young adults with osteosarcoma. Kaplan Meier survival curves for patients' OS displaying patient stratification by response to neoadjuvant chemotherapy (A) and the DL-based signature (B) in the training (Left) and testing (Right) patient cohorts. Abbreviations: OS, Overall survival; DL, Deep learning.