


MRI-Based Deep-Learning Model for Distant Metastasis-Free Survival in Locoregionally Advanced Nasopharyngeal Carcinoma

Lu Zhang, MD,¹ Xiangjun Wu, MD,^{2,3} Jing Liu, MD,¹ Bin Zhang, MD,¹ Xiaokai Mo, MD,¹ Qiuying Chen, MD,¹ Jin Fang, MD,¹ Fei Wang, MD,¹ Minmin Li, MD,¹ Zhuozhi Chen, MD,¹ Shuyi Liu, MD,¹ Luyan Chen, MD,¹ Jingjing You, MD,¹ Zhe Jin, MD,¹ Binghang Tang, PhD,^{4*} Di Dong, PhD,^{2,3*} and Shuixing Zhang, PhD^{1*} 

Background: Distant metastasis is the primary cause of treatment failure in locoregionally advanced nasopharyngeal carcinoma (LANPC).

Purpose: To develop a model to evaluate distant metastasis-free survival (DMFS) in LANPC and to explore the value of additional chemotherapy to concurrent chemoradiotherapy (CCRT) for different risk groups.

Study Type: Retrospective.

Population: In all, 233 patients with biopsy-confirmed nasopharyngeal carcinoma (NPC) from two hospitals.

Field Strength: 1.5T and 3T.

Sequence: Axial T₂-weighted (T₂-w) and contrast-enhanced T₁-weighted (CET₁-w) images.

Assessment: Deep learning was used to build a model based on MRI images (including axial T₂-w and CET₁-w images) and clinical variables. Hospital 1 patients were randomly divided into training (*n* = 169) and validation (*n* = 19) cohorts; Hospital 2 patients were assigned to a testing cohort (*n* = 45). LANPC patients were divided into low- and high-risk groups according to their DMFS (*P* < 0.05). Kaplan–Meier survival analysis was performed to compare the DMFS of different risk groups and subgroup analysis was performed to compare patients treated with CCRT alone and treated with additional chemotherapy to CCRT in different risk groups, respectively.

Statistical Tests: Univariate analysis was performed to identify significant clinical variables. The area under the receiver operating characteristic (ROC) curve (AUC) was used to assess the model performance.

Results: Our deep-learning model integrating the deep-learning signature, node (N) stage (from TNM staging), plasma Epstein–Barr virus (EBV)-DNA, and treatment regimens yielded an AUC of 0.796 (95% confidence interval [CI]: 0.729–0.863), 0.795 (95% CI: 0.540–1.000), and 0.808 (95% CI: 0.654–0.962) in the training, internal validation, and external testing cohorts, respectively. Low-risk patients treated with CCRT alone had longer DMFS than patients treated with additional chemotherapy to CCRT (*P* < 0.05).

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.27308

Received Feb 18, 2020, Accepted for publication Jul 18, 2020.

*Address reprint requests to: S.Z., D.D., and B.T., Department of Radiology, The First Affiliated Hospital of Jinan University, No. 613, Huangpu West Road, Tianhe District, Guangzhou, Guangdong 510627, China. E-mail: shui7515@126.com (S.Z.); di.dong@ia.ac.cn (D.D.); jmftbh@sina.com (B.T.)

The first three authors contributed equally to this work.

Contract grant sponsor: National Natural Science Foundation of China; Contract grant number: 81571664, 81871323, 81801665, 91959130, 81971776, 81771924;

Contract grant sponsor: National Natural Science Foundation of Guangdong Province; Contract grant number: 2018B030311024; Contract grant sponsor:

Scientific Research General Project of Guangzhou Science Technology and Innovation Commission; Contract grant number: 201707010328; Contract grant sponsor: China Postdoctoral Science Foundation; Contract grant number: 2016M600145

From the ¹Department of Radiology, The First Affiliated Hospital of Jinan University, Guangzhou, China; ²School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China; ³CAS Key Lab of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China; and

⁴Department of Radiology, Zhongshan Hospital of Sun Yat-sen University, Zhongshan, China

Additional supporting information may be found in the online version of this article

Data Conclusion: The proposed deep-learning model, based on MRI features and clinical variates, facilitated the prediction of DMFS in LANPC patients.

Level of Evidence: 3.

Technical Efficacy Stage: 4.

J. MAGN. RESON. IMAGING 2020.

NASOPHARYNGEAL CARCINOMA (NPC) is an uncommon malignancy worldwide, and is highly associated with a specific endemic distribution as Epstein–Barr virus (EBV) infection.^{1,2} Approximately 129,000 new NPC patients were diagnosed worldwide in 2018,³ of which 70% were classified as locoregionally advanced stage (stage III–IVB), which is associated with poor treatment response and unfavorable prognosis.^{4,5} Although the local recurrence rate of NPC has been significantly reduced by radiotherapy and concurrent chemoradiotherapy (CCRT) treatment strategies, distant metastasis (DM), which occurs in more than 30% of NPC patients, is a leading cause of treatment failure.⁶ Moreover, the treatment response of NPC patients with DM is unsatisfactory, with a low 5-year survival rate of less than 5%.⁷

To improve distant metastasis-free survival (DMFS) in locoregionally advanced NPC (LANPC), the addition of induction (IC) or adjuvant (AC) chemotherapy to CCRT has been proposed.^{8,9} However, the effect of adding IC or AC to CCRT remains controversial due to the inconsistent results of several randomized controlled trials. A phase III study of 172 patients with advanced NPC found no significant differences between IC (combined gemcitabine, carboplatin, and paclitaxel [GCP]) plus CCRT and CCRT alone with regard to survival or locoregional control.¹⁰ In contrast, multicenter phase III trials reported that IC plus CCRT could marginally improve distant failure-free survival in LANPC.¹¹ Thus, identifying risk factors to predict DMFS in LANPC patients is critical to aid clinicians in developing effective treatment strategies. Node (N) stage, plasma EBV-DNA, and micro-RNAs (miRNAs) are common risk factors used as predictors for NPC therapeutic efficacy in clinical practice.^{12,13} However, these factors may not adequately reflect intratumoral differences in NPC. Previous studies have noted that the expression of certain genes could reflect intratumoral heterogeneity,¹⁴ and that gene expression was closely associated with NPC prognosis and treatment.¹⁵ A gene expression signature⁴ consisting of 13 genes could be used to predict the risk of DM in LANPC patients and provide a reference for LANPC patients who could benefit from CCRT. However, an investigation of new factors that can reflect intratumoral heterogeneity is still urgently needed to predict DMFS in LANPC patients.

Radiomic features, which reflect the translation of medical images into quantitative image features via machine-learning algorithms, are able to capture intratumoral heterogeneity in a noninvasive manner.¹⁶ For example, the result of our previous study on extracting features from magnetic

resonance imaging (MRI) showed that the radiomic features reflecting intratumoral heterogeneity were associated with advanced NPC.¹⁷ Deep learning is a subset of machine learning that can rapidly extract information from different types of images using deep convolution neural networks, which comprise a powerful complement to clinical variables for more precise diagnosis and treatment.^{18,19} Compared to conventional radiomic methods that extract quantitative features manually, deep learning can mine available information by transforming the feature engineering step into a learning step.²⁰ Recently, deep-learning models have been successfully applied to both solid tumors and nontumor disease, such as epidermal growth factor receptor mutation status prediction from lung cancer computed tomography (CT) images,²¹ classification and mutation prediction from non-small cell lung cancer histopathology images,²² and classification of pulmonary tuberculosis from chest radiography.²³ In NPC, recent studies have used deep learning to automatically segment the primary tumor volume based on MR or CT images.^{24,25} However, it is unclear whether deep-learning models can predict DMFS in LANPC patients who have received CCRT with or without additional chemotherapy.

The purpose of this study was to build an MRI-based combined model as a novel tool for pretreatment prediction of DMFS in LANPC using data from LANPC patients who have received CCRT with or without additional chemotherapy. We also investigated the value of additional chemotherapy to CCRT for different risk groups.

Materials and Methods

This retrospective study was approved by the Institutional Review Board of each center. Written informed consent was waived by the Institutional Review Board. In our trial, we retrospectively included patients with incident, primary, and biopsy-confirmed NPC from two independent centers. A total of 188 patients admitted from March 2009 to December 2018 at Hospital 1 and 45 patients admitted from August 2008 to December 2016 at Hospital 2 were included. Patients were enrolled in the study if they met the following criteria: 1) previously untreated, biopsy-proven NPC; 2) stage III–IVb disease according to the 7th American Joint Committee on Cancer (AJCC) TNM staging manual²⁶; 3) underwent a pretreatment and prebiopsy nasopharyngeal and neck MRI scan (including axial T₂-weighted [T₂-w] images and contrast-enhanced T₁-weighted [CET₁-w]); 4) underwent a complete pretreatment physical examination, including chest radiography, abdominal ultrasound, skeletal scintigraphy, or whole-body fluorodeoxyglucose positron emission tomography (PET) / computed tomography (CT),

blood and biochemical profile, and plasma EBV DNA level; and 5) regular follow-up: every 1–3 months during the first years, every 6 months in years 2–5, and annually thereafter. The exclusion criteria were as follows: i) evidence of distant metastasis at initial treatment; ii) history of MRI contraindications; iii) history of previous or synchronous malignant tumors; or iv) previous treatment for NPC.

Eligible patients at Hospital 1 were randomly divided into a training cohort ($n = 169$) and an internal validation cohort ($n = 19$) at a ratio of 9:1. Eligible patients at Hospital 2 were assigned to an external testing cohort ($n = 45$). The baseline clinical characteristics that were collected included age, sex, tumor (T) stage, N stage, histological subtype, plasma EBV-DNA, and treatment regimen(s). The T and N staging was carried out according to the TNM staging. The histological tumor subtypes were categorized according to the World Health Organization (WHO) standards as follows: type I (differentiated keratinizing carcinoma), type II (differentiated non-keratinizing carcinoma), and type III (undifferentiated non-keratinizing carcinoma).²⁷

The primary endpoint was DMFS, which was calculated from the date of diagnosis to the date of distant metastasis detection or the date of censoring (the date on which the patient was last known to be distant metastasis-free). Patients were censored if they were still alive on December 31, 2018, which was the date of the last follow-up. If DM was suggested by any medical reports (chest radiographs, abdominal ultrasound, PET/CT, and/or skeletal scintigraphy), then the potentially involved sites underwent additional examinations, including CT, MRI, PET/CT, and/or biopsy. If the presence of DM was immediately confirmed by these additional examinations, the diagnosis was accepted. If additional examinations were not feasible or yielded negative results, follow-up examinations were performed every 3 months for at least 12 months. Patients were considered to have locoregional disease if the lesion remained unchanged during the follow-up period. If lesion enlargement was observed, DM was considered present. The details of the treatment protocols are presented in Supplementary Methods 1.

MR Image Acquisition and Quality Assessment

Overall, 189 patients underwent nasopharyngeal and neck MRI with a 1.5T MR scanner (Optima, TwinSpeed, GE Healthcare, Milwaukee, WI, $n = 105$; Achieva, Philips Healthcare, Best, Netherlands, $n = 84$), and 44 patients underwent nasopharyngeal and neck MRI with a 3.0T MR scanner (Discovery, TwinSpeed, GE Healthcare, $n = 32$; Ingenia, Philips Healthcare, $n = 12$). Gd-DTPA (Magnevist, Bayer Schering, Berlin, Germany) was administered as a bolus through a peripheral vein (dosage: 0.1 mmol/kg; rate: 2.0 mL/s followed by a 10-mL flush of saline) via a power injector at the eighth dynamic scan. The MRI acquisition parameters are presented in Table 1. Axial T₂-w and CET₁-w digital imaging and communications in medicine (DICOM) images were retrieved from the picture archiving and communication systems (PACS).

Since the images were retrieved from different hospitals, the quality of the images occasionally differed. Thus, to assess image quality objectively and quantitatively, the structural similarity index (SSIM) was used. The SSIM combined three features: luminance comparison, contrast comparison, and structural comparison. SSIM analysis was performed in Python (<https://www.python.org/>), and

the detailed calculation methods were referenced from the study by Wang et al.²⁸ In the current study, the SSIM of each corresponding patch under the sliding window of the two pictures was calculated. The pictures of a patient in Hospital 1 were successively matched with the pictures of all patients in Hospital 2. Then the mean value of the overall SSIM of each patient was calculated and termed MSSIM. For more than 85% of the patients MSSIM ≥ 0.9 was considered to indicate good consistency of image quality.

MR Image Segmentation and Preprocessing

ITK-SNAP software (open source software; <https://itk.org/>) was used to manually segment regions of interest (ROIs) in both axial T₂-w and CET₁-w images. All slices of the entire tumor were delineated by a radiologist with 12 years of experience in head and neck MR image interpretation (reader 1, F.J.). If the tumor could invade the skull base, the radiologist would reference precontrast T₁-w and CET₁-w images to segment ROIs. After 1 month, 30 patients were randomly selected and the images were segmented again by reader 1 and by another radiologist with 11 years of clinical experience (reader 2, W.F.) to evaluate the intra- and interobserver reproducibility of radiomic features across segmentations. An example of the ROIs segmentation is shown in Fig. 1.

For natural images, there are tens of millions of annotated public datasets. During the training model, the whole image was directly used as the input of deep-learning networks, which leads to very good performance. For medical images, the sample size was typically smaller than that for natural images. Therefore, we removed the background noise and used only the tumor area segmented by radiologists so that the network learning could learn the essential features more rapidly. We removed the background by setting the image intensity beyond the tumor area to 0 using the “NumPy” packages in Python. In addition, because the MR images in this study were scanned by different MRI scanners using different scan protocols, the intensity ranges of the images were normalized from 0–255, also using the “NumPy” packages in Python. Finally, to eliminate different distributions of grayscale values, we also performed histogram equalization in Python using the “OpenCV” packages (<https://pypi.org/project/opencv-python/>).

Deep-Learning Model Construction

We constructed a deep convolution model based on the CET₁-w and T₂-w images to predict the DMFS of LANPC patients. Deep convolutional neural networks are mostly used in image processing networks—such as visual geometry group (VGG) networks, residual networks (ResNet), and dense networks (DenseNet)—which are widely used for both natural and medical images. In this study, our residual network was derived from a 50-layer residual network architecture, including 49 convolution layers and a fully connected layer.²⁹ The network structure is shown in Supplementary Methods 3. As with the original residual network architecture, batch normalization was used after each convolutional layer. This forced network activations to follow a unit Gaussian distribution after each update, preventing internal covariate shifts and overfitting.

For training images, we selected the slice with the largest tumor area annotated derived from the corresponding ROIs using an automatic algorithm. This algorithm was used to calculate the dot product of the tumor region matrix segmented by radiologists, and

TABLE 1. MR Image Acquisition Parameters

Scanner	N	Sequence	TR (msec)	TE (msec)	NEX	FOV (mm)	Acquisition matrix size	Reconstructed matrix size	Slice thickness (mm)	Spacing between slices (mm)
1.5T GE	105	Axial T ₂ WI	2720–5866	81–113	1	200 × 200–250 × 250	288 × 224	512 × 512	4	1
		Axial CE-T ₁ WI	440–600	8–11	1	200 × 200–250 × 250	256 × 224	512 × 512	4	1
3T GE	32	Axial T ₂ WI	4682–8002	85–124	1	200 × 200–250 × 250	288 × 256	512 × 512	4	1
		Axial CE-T ₁ WI	400–700	6.7–10	1	200 × 200–250 × 250	288 × 256	512 × 512	4	1
1.5T Philips	84	Axial T ₂ WI	2949	120	1	200 × 200–230 × 230	288 × 198	512 × 512	5	1
		axial CE-T ₁ WI	630	15	1	200 × 200–230 × 230	260 × 162	512 × 512	5	1
3.0T Philips	12	Axial T ₂ WI	2851	90	1	200 × 200–230 × 230	288 × 223	512 × 512	5	1
		Axial CE-T ₁ WI	595	7.1	1	200 × 200–230 × 230	256 × 207	512 × 512	5	1

Axial T₂WI = axial T₂-weighted spin-echo images; axial CE-T₁WI = axial contrast-enhanced T₁-weighted spin-echo images; TR = repetition time; TE = echo time; FOV = field of view; NEX = number of excitations.

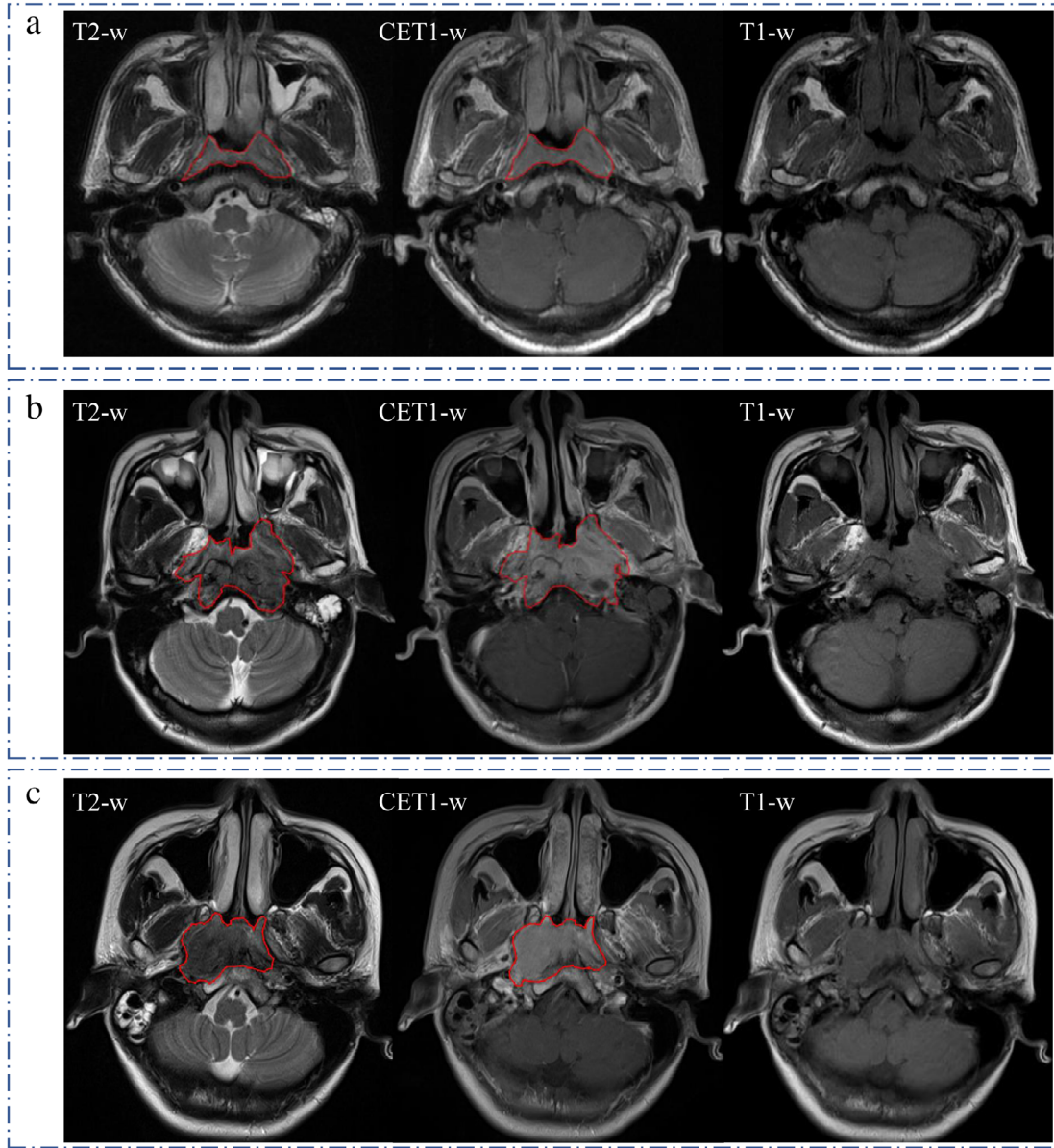


FIGURE 1: Example of ROI segmentation. (a) The ROI of a patient with noninvaded skull-base bone segmented in T_2 -weighted (T_2 -w) and contrast-enhanced T_1 -weighted (CET_1 -w) images. **(b,c)** The ROI of patients with invaded the skull-base bone segmented in T_2 -w and CET_1 -w, and referenced precontrast T_1 -w images (T_1 -w).

the slice with the highest dot product was selected. We used CET_1 -w and T_2 -w images successively to train the max slice deep residual networks and to obtain independent CET_1 -w and T_2 -w prediction models. We also used combined CET_1 -w and T_2 -w images to train the max slice ensemble deep residual networks and obtain a combined prediction model. We compared the performances of the CET_1 -w, T_2 -w, and combined models. Finally, the max slice deep-learning signature was built based on the optimal model. We also created multiple-slice deep residual networks based on three slices (the max slices of the tumor area and the two contiguous slices) using a similar approach. For the max slice networks and the multiple slices networks, the inputs were $512 \times 512 \times 1$ and $512 \times 512 \times 3$, respectively, and all the input images were resized to the same size. The output of the network was the predicted probability value of the DMFS, and the optimal

threshold was determined using the training set by the Delong test.

The root mean square prop optimizer was used in the training process, which could optimize the parameters of the deep-learning network.³⁰ The range of the batch size was [1, 2, 4, 8, 16], the range of the learning rate was [0.1, 0.01, 0.001, $10e-4$, $10e-5$], and the corresponding range of the learning rate decay was [0.01, 0.001, $10e-4$, $10e-5$, $10e-6$]. The optimal hyperparameter combinations and model weights were selected based on the results of the internal validation set. The deep-learning model was constructed in Python with Keras (v. 2.1.3), based on TensorFlow (v. 1.4.1).

Then a DMFS predictive model, called the MRI-based combined model, was built by integrating the deep-learning signature and clinical variables that had prognostic significance for DMFS with coefficients weighted by logistic regression analysis of the

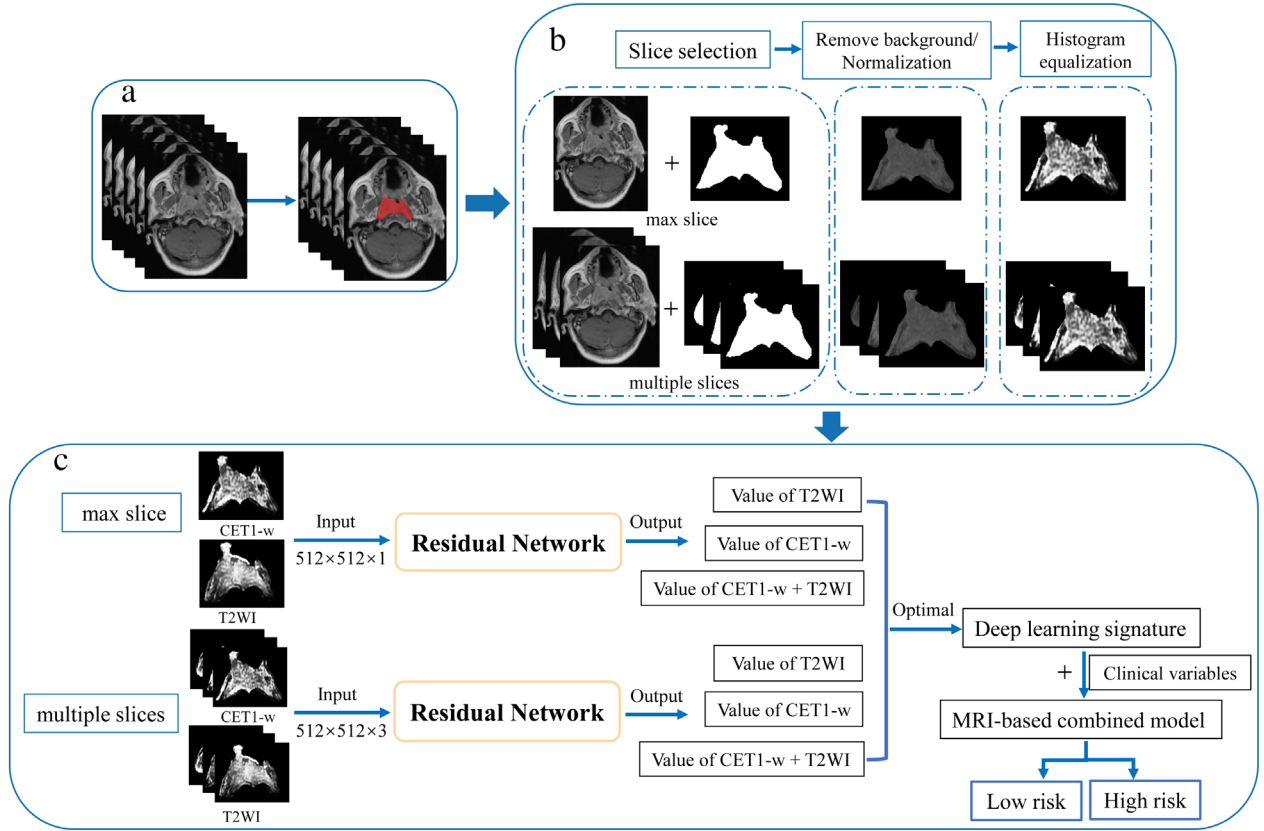


FIGURE 2: Flow of the MRI-based combined model. (a) MR image acquisition and segmentation. (b) Imaging preprocessing: Slice selection, Remove background/Normalization, Histogram equalization. (c) Residual network and MRI-based combined model construction.

training cohort. The MRI-based combined model workflow is presented in Fig. 2. Based on the MRI-based combined model for DMFS, the median of the predictive risk scores was defined as the optimal cutoff to classify patients into low- and high-risk groups. We built the models on the training cohort, optimized the performance of the models on the internal validation cohort, and tested the stability of the models on the external testing cohort.

Radiomic Signature and Clinical Model Building Design

We built a radiomic signature based on radiomic feature extraction and selection. We performed standardized feature extraction on the same max slice of the tumor used in the deep-learning model from both the axial T₂-w and CET₁-w images using the PyRadiomics platform.³¹ Then univariate and multivariate analyses were performed for feature selection; the multivariate analysis included the minimum redundancy-maximum relevance (mRMR), the least absolute shrinkage and selection operator (LASSO), and the Akaike information criterion (AIC) algorithms.^{32–34} Similarly, we also built a multiple slices radiomic signature based on the same three slices as was used for the deep-learning signature—namely, the max slices of the tumor area and the contiguous slices. We used the remaining optimal features to construct radiomic signatures by logistic regression analysis. Furthermore, to explore the repeatability of feature selection, we sequentially performed univariate analysis, LASSO, and AIC algorithms to select features. A clinical model was also built

based on the clinical variables that had prognostic significance for DMFS by logistic regression analysis. The median of the predictive risk scores was also defined as the optimal cutoff to classify patients into low- and high-risk groups in both the radiomic signatures and clinical model for DMFS.

Statistical Analysis

The distribution of clinical variables between the training, internal validation, and external testing cohorts was compared using the analysis of variance (ANOVA) (or independent-samples *t*-test, when appropriate) for continuous variables and the Kruskal–Wallis H test (or χ^2 test, when appropriate) for categorical variables. We performed univariate analysis using the independent-samples *t*-test (or the Mann–Whitney *U*-test, when appropriate) and χ^2 tests (or Fisher's exact test, when appropriate) to identify clinical variables with prognostic significance for DMFS for developing the clinical model and for inclusion in the MRI-based combined model. Mann–Whitney *U*-tests and χ^2 tests were also used for feature selection prior to the radiomics model generation. All tests were two-sided, and a value of $P < 0.05$ was considered statistically significant. The intra- and interobserver reproducibility of features from different tumor ROIs (radiologists F.J. and W.F.) were assessed by the intraclass correlation coefficient (ICC) with an ICC > 0.75 being regarded as good. Kaplan–Meier survival curves and the log-rank test were used to compare the DMFS between the low- and high-risk groups in different models and treatment protocols. We also

performed subgroup survival analysis between patients treated with CCRT alone and additional chemotherapy plus CCRT in different risk groups. Survival analysis and Kaplan–Meier survival curves were processed in R using the “survival” and “survcomp” packages. All statistical analyses were performed using R (v. 3.2.1, Vienna, Austria), Python (v. 3.6.5, <https://www.python.org/>), and the Statistical Package for Social Sciences (SPSS) software (v. 23.0, IBM, Armonk, NY).

Results

The baseline clinical characteristics of the training, internal validation, and external testing cohorts are presented in Table 2 and show statistically significant differences in N stage, EBV-DNA, and treatment regimen (all $P < 0.05$). The mean DMFS was 43.07, 41.68, and 50.98 months in training, internal validation, and external testing cohorts, respectively. The mean overall survival (OS) was 59.8, 61.84, and 70.38 months in the training, internal validation, and external testing cohorts, respectively.

Validation of the Deep-Learning Model

After the quality assessment of the MR images, 86.67% of the images satisfied SSIM ≥ 0.9 in CET₁-w imaging, and 88.87% of the images satisfied SSIM ≥ 0.9 in T₂-w imaging. This indicated a good consistency in image quality between the two hospitals. The intra-/interobserver reproducibility of the features extracted from different tumor ROIs resulted in an ICC ≥ 0.75 , which indicates good agreement of the image segmentation.

By using an end-to-end deep-learning approach, we determined the predictive information from the T₂-w and CET₁-w imaging series. The AUC of the deep learning and radiomic signatures based on the max tumor slice and multiple slices are presented in Table 3. We found that the max tumor slice from CET₁-w imaging had greater predictive value than the equivalent slice from T₂-w imaging alone or CET₁-w and T₂-w imaging combined. Furthermore, the models based on the max tumor slice demonstrated better performance than the models based on multiple slices (the max slice and two contiguous slices). Accordingly, a deep-learning signature was constructed based on the CET₁-w images. This yielded an AUC of 0.769 (95% confidence interval [CI]: 0.694–0.843) in the training cohort, 0.705 (95% CI: 0.456–0.954) in the internal validation cohort, and 0.783 (95% CI: 0.604–0.962) in the external testing cohort. After univariate analysis for the significant clinical variables, only N stage, EBV-DNA, and treatment regimen were significantly associated with DMFS in LANPC patients (all $P < 0.05$). Thus, the MRI-based combined model was built by integrating the deep-learning signature with these three clinically significant variables. The deep-learning model generated an accurate prediction of DMFS in LANPC patients, yielding an AUC of 0.796 (95% CI: 0.729–0.863) in the

training cohort, 0.795 (95% CI: 0.540–1.000) in the internal validation cohort, and 0.808 (95% CI: 0.654–0.962) in the external testing cohort.

Examination of the Radiomic Signature and Clinical Model

We extracted features from T₂-w and CET₁-w images based on the max tumor slice and multiple slices. After feature selection, the radiomic signature based on multiple slices was found to perform better than that based on the max slice alone (Table 3). Five image features were retained: T₁_original_shape_Maximum 3D Diameter, T₁_wavelet.LLL_firstorder_Minimum, T₁_squareroot_firstorder_90 Percentile, T₁_wavelet.HHH_GLSZM_Large Area High Gray Level Emphasis, and T₁_wavelet.HHL_GLCM_Imc2. The radiomic signature yielded an AUC of 0.789 (95% CI: 0.714–0.864) in the training cohort, 0.761 (95% CI: 0.537–0.986) in the internal validation cohort, and 0.765 (95% CI: 0.569–0.961) in the external testing cohort. Besides, the features apart from those selected by the two different algorithm combinations were different. The details are presented in Table S1 in the Supplemental Material.

We also built a clinical model based on the clinical variables. After univariate analysis for the significant clinical variables, only the N stage, EBV-DNA, and treatment regimen were used to build the clinical model. The clinical model yielded an AUC of 0.698 (95% CI: 0.614–0.781) in the training cohort, 0.727 (95% CI: 0.474–0.981) in the internal validation cohort, and 0.649 (95% CI: 0.453–0.845) in the external testing cohort. Thus, the MRI-based combined model had better predictive performance than the deep-learning signature, radiomic signature, and clinical model. The receiver operating characteristic curves are presented in Fig. 3.

Survival Analysis for Different Treatment Regimens Based on the Models

According to the MRI-based combined model, deep-learning signature, radiomic signature, and clinical model the DMFS of the high-risk group patients was significantly shorter than that of the low-risk group patients ($P < 0.05$; Fig. 4a–d). There was a significant difference between patients treated with CCRT alone and patients treated with additional chemotherapy plus CCRT in the low-risk group ($P < 0.05$; Fig. 4e). The DMFS of patients treated with CCRT alone was longer than that of patients treated with additional chemotherapy plus CCRT in the low-risk group. However, there was no significant difference between patients treated with CCRT alone and patients treated with additional chemotherapy plus CCRT in the high-risk group ($P = 0.62$; Fig. 4f).

TABLE 2. Baseline Characteristics of Patients With LANPC for the Training, Internal Validation, and External Testing Cohorts

	Training cohort (<i>n</i> = 169)	Internal validation cohort (<i>n</i> = 19)	External testing cohort (<i>n</i> = 45)	<i>P</i> value
Age (mean, years)	41.94 (18–71)	44.16 (27–68)	43.16 (20–64)	0.486
Sex				0.459
Men	125 (73.96%)	16 (84.21%)	36 (80.00%)	
Women	44 (26.04%)	3 (15.79%)	9 (20.00%)	
T stage				0.178
T1	6 (3.55%)	0	6 (13.33%)	
T2	30 (17.75%)	3 (15.79%)	9 (20.00%)	
T3	88 (52.07%)	12 (63.16%)	22 (48.89%)	
T4	45 (26.63%)	4 (21.05%)	8 (17.78%)	
N stage				0.020
N0	8 (4.73%)	4 (21.05%)	1 (2.22%)	
N1	46 (27.22%)	3 (15.79%)	7 (15.56%)	
N2	84 (49.70%)	10 (52.63%)	33 (73.33%)	
N3	31 (18.34%)	2 (10.53%)	4 (8.89%)	
Staging				0.223
III	99 (58.58%)	14 (73.68%)	33 (73.33%)	
IVa	41 (24.26%)	4 (21.05%)	8 (17.78%)	
IVb	29 (17.16%)	1 (5.26%)	4 (8.89%)	
Histology				0.880
Differentiated keratinizing	0	0	0	
Differentiated non-keratinizing	8 (4.73%)	1 (5.26%)	3 (6.67%)	
Undifferentiated non-keratinizing	161 (95.27%)	18 (94.74%)	42 (93.33%)	
EBV DNA (copies/mL)				<0.001
<4000	98 (57.99%)	11 (57.89%)	40 (88.89%)	
≥4000	71 (42.01%)	8 (42.11%)	5 (11.11%)	
Chemotherapeutic regimens				0.005
CCRT	54 (31.95%)	8 (42.11%)	5 (11.11%)	
IC+CCRT	100 (59.17%)	8 (42.11%)	38 (84.44%)	
CCRT+AC	15 (8.88%)	3 (15.78%)	2 (4.44%)	

EBV-DNA = plasma Epstein–Barr virus DNA; CCRT = concurrent chemoradiotherapy; IC = induction chemotherapy; AC = adjuvant chemotherapy.

Discussion

In this study, we built an MRI-based combined model integrating a deep-learning signature and clinical variables—

namely, the N stage, EBV-DNA, and treatment regimen—which could predict DMFS in LANPC patients. Compared with clinical-based and deep-learning signature-based, and

TABLE 3. AUC of the Deep Learning and the Radiomic Signatures Based on Max Slice and Multiple Slices

Methods	Slices	Image series	Training cohort	Internal validation cohort	External testing cohort
Deep learning	Max slice	CET ₁ -w	0.769 (0.694–0.843)	0.705 (0.456–0.954)	0.783 (0.604–0.962)
		T ₂ -w	0.553 (0.461–0.645)	0.636 (0.326–0.947)	0.573 (0.369–0.777)
		CET ₁ -w + T ₂ -w	0.763 (0.691–0.836)	0.773 (0.523–1.000)	0.732 (0.573–0.892)
	Multiple slices	CET ₁ -w	0.595 (0.504–0.685)	0.364 (0.090–0.637)	0.788 (0.640–0.936)
		T ₂ -w	0.617 (0.528–0.707)	0.557 (0.249–0.864)	0.609 (0.423–0.794)
		CET ₁ -w + T ₂ -w	0.616 (0.527–0.704)	0.578 (0.271–0.888)	0.677 (0.493–0.860)
Radiomic	Max slice	CET ₁ -w	0.674 (0.589–0.758)	0.511 (0.208–0.769)	0.730 (0.580–0.880)
		T ₂ -w	0.663 (0.574–0.753)	0.608 (0.321–0.895)	0.643 (0.461–0.824)
		CET ₁ -w + T ₂ -w	0.681 (0.594–0.767)	0.585 (0.284–0.886)	0.696 (0.538–0.853)
	Multiple slices	CET ₁ -w	0.789 (0.714–0.864)	0.761 (0.537–0.986)	0.765 (0.569–0.961)
		T ₂ -w	0.784 (0.709–0.859)	0.761 (0.537–0.986)	0.639 (0.447–0.831)
		CET ₁ -w + T ₂ -w	0.801 (0.729–0.872)	0.739 (0.486–0.992)	0.684 (0.487–0.882)

T₂-w = T₂-weighted imaging; CET₁-w = contrast-enhanced T₁-weighted imaging.

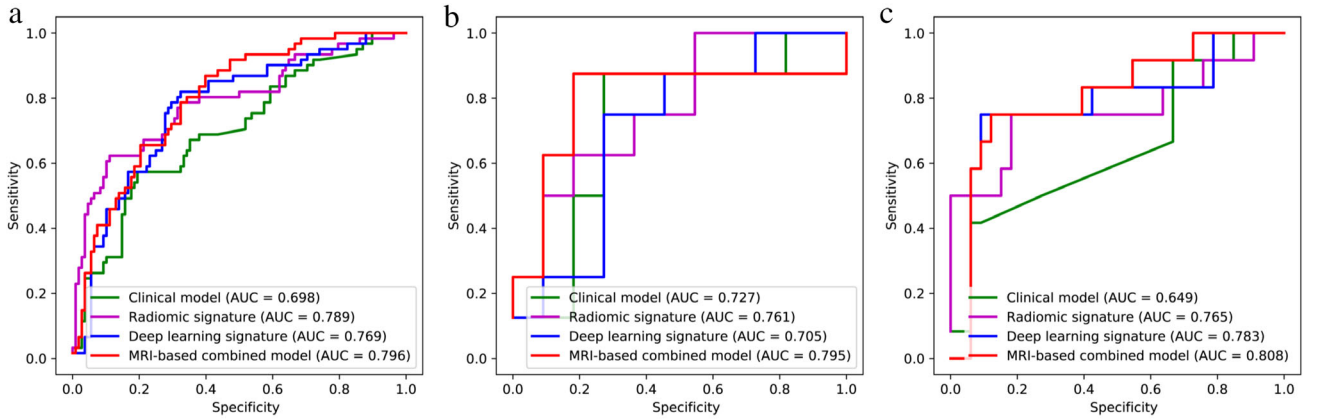


FIGURE 3: ROC curves for a newly constructed deep-learning model, deep-learning signature, radiomic signature, and clinical model for the (a) training, (b) internal validation, and (c) external testing cohorts of LANPC patients.

radiomic signature-based models, this combined model had a better predictive performance. Furthermore, the deep-learning signature based on the largest tumor slice and radiomic signature based on multiple slices in CET₁-w images had the best performance among signature models. Additional subgroup analysis showed that, based on the MRI-based deep-learning model, the DMFS of low-risk patients treated with CCRT alone was longer than it was for patients treated with additional chemotherapy plus CCRT.

Our deep-learning model was predominantly based on MR image processing because MRI is a standard examination technique for primary tumors and offers outstanding image resolution. Compared with hand-crafted radiomics methods, the deep-learning model was easy to operate because it only required inputting the MR images to end-to-end output a predictive value. Additionally, the MRI-based model used a range of features from visual characteristics to abstract mappings that are associated with DMFS. In our study, there was

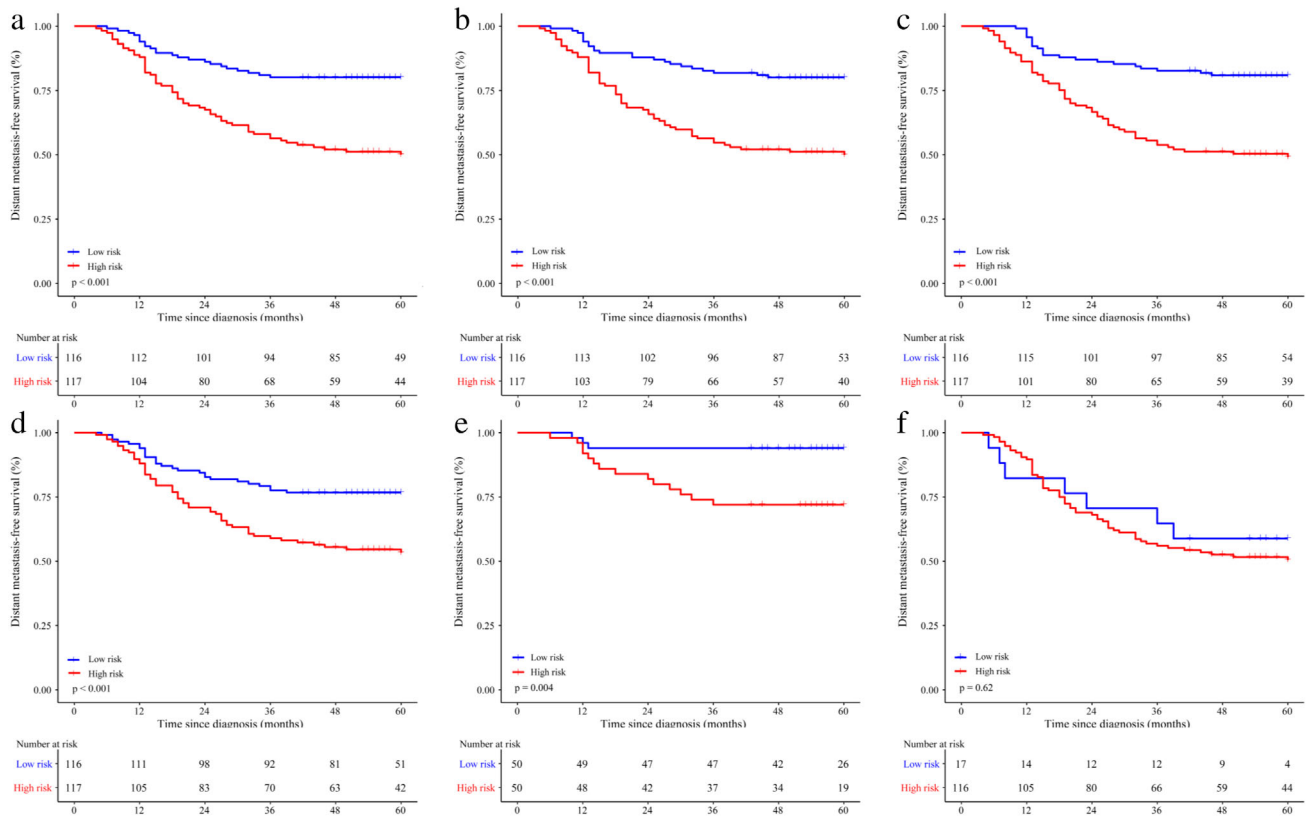


FIGURE 4: Kaplan–Meier curves for DMFS in LANPC patients. The patients were classified into a high DM risk group and a low DM risk group by the (a) deep learning model, (b) deep learning signature, (c) radiomic signature, and (d) clinical model. The DMFS rate of high-risk patients is significantly lower than that of low-risk patients ($P < 0.001$). Kaplan–Meier curves comparing the DMFS between patients treated with additional chemotherapy plus concurrent chemoradiotherapy (IC/AC+CCRT) to those treated with concurrent chemoradiotherapy alone (CCRT) and CCRT alone in low-risk group (e) and (f) high-risk group.

no significant difference between deep learning and radiomic signatures when they were used independently. After integrating clinical variables, however, the MRI-based combined model could predict DMFS more accurately than the radiomic signature model among the training, validation, and testing cohorts. Wang et al²¹ also suggested that a deep-learning model has better performance than conventional radiomic and clinical models. Indeed, the MRI-based combined model could not only automatically extract relevant features to evaluate the DMFS but could also classify LANPC patients into high and low DMFS risk groups using its deep-learning algorithm.

The radiomic signature constructed based on multiple slices exhibited better performance than a radiomic signature based on the largest tumor slice alone. This result may arise from the fact that radiomics is well suited to solving small-sample problems and a radiomic signature based on multiple slices may have more extracted image features available for calculation. However, while deep learning makes better and more complicated decisions that generally involve massive data when the sample is not sufficiently large, multiple slices could lead to overfitting or increased noise. Our results showed that the performance of the deep-learning signature

based on the largest tumor slice is better than that based on multiple slices. Ge et al³⁵ reported that using a single 2D slice image to build a deep-learning model could mitigate overfitting. Furthermore, Han et al³⁶ showed that a 2D slice could provide sufficient contextual information for the model to reliably separate locally similar voxels and produce very accurate results. Interestingly, both the deep learning and radiomic signatures based on CET₁-w images performed better than those based on T₂-w images or combined CET₁-w and T₂-w images. This may be because T₂-w images mainly provide anatomical information, whereas CET₁-w images evaluate blood supply, which is a key determinant for establishing and determining prognosis. Jiang et al³⁷ also suggested that using CET₁-w images to build the model produces better results than T₂-w images.

The efficacy of adding IC or AC to the CCRT regimen in the treatment of LANPC patients is still controversial. Interestingly, our study showed that low-risk patients who received IC or AC plus CCRT had a lower DMFS rate than patients who received CCRT alone, which was inconsistent with the results of previous studies that suggested IC plus CCRT could improve the DMFS rate¹¹ or long-term locoregional control.³⁸ It is worth noting that Grade 3 to

4 toxicities of additional chemotherapy plus CCRT were previously reported in LANPC patients.^{10,39} We speculated that additional chemotherapy increasing acute hematological toxicities and reducing patient tolerance to CCRT might contribute to poor treatment efficacy. In addition, the retrospective nature of the study could cause patient distributional unbalance of the chemotherapeutic regimens, which might affect our results; further, prospective studies are needed to validate the outcomes.

Pretreatment plasma EBV-DNA and N stage were reported as important risk factors for NPC patients and have implications for DMFS.¹³ Our previous study suggested that N stage is a component of an MRI-based radiomic model for DM prediction. In the present study, N stage and plasma EBV-DNA were also strongly associated with DM. By integrating the N stage, EBV-DNA, and deep-learning signature, our MRI-based combined model had better predictive efficacy than clinical-based models for the training, validation, and testing cohorts.

Limitations

First, the retrospective nature of the study substantially limits the value of the deep-learning model and the collected information. The distribution of the N stage, plasma EBV-DNA, and treatment regime had a significant difference between the training/validation and testing cohorts. However, despite the difference in data distribution, the performance of the models was good, and indicated good model generalization. In order to generalize our results to other populations, further external and prospective validation in other endemic and nonendemic areas is still required. Second, we compared the LASSO algorithm with univariate and multivariate analyses that were performed sequentially during feature selection. The repeatability of radiomic signatures was poor in this study, and the remaining features did not overlap. Third, because the different scan protocols between hospitals had significant impacts on image quality, the generalizability of the models is indirectly decreased. In our study, after image collection, we performed an image quality assessment to filter out unsatisfactory images.

Conclusion

Our MRI-based combined model incorporating a deep-learning signature and clinical features, including the N stage, plasma EBV-DNA, and treatment regimen, could be used to evaluate the risk of DMFS in LANPC patients and could provide a complementary tool for making treatment decisions.

Conflict of interest

The authors declare that they have no competing interests.

References

1. Lin JC, Wang WY, Chen KY, et al. Quantification of plasma Epstein-Barr virus DNA in patients with advanced nasopharyngeal carcinoma. *N Engl J Med* 2004;350(24):2461-2470.
2. Kamran SC, Riaz N, Lee N. Nasopharyngeal carcinoma. *Surg Oncol Clin N Am* 2015;24(3):547-561.
3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018; 68(6):394-424.
4. Tang XR, Li YQ, Liang SB, et al. Development and validation of a gene expression-based signature to predict distant metastasis in locoregionally advanced nasopharyngeal carcinoma: A retrospective, multicentre, cohort study. *Lancet Oncol* 2018;19(3):382-393.
5. Jiang C, Chen J, Xie S, et al. Evaluation of circulating EBV microRNA BART2-5p in facilitating early detection and screening of nasopharyngeal carcinoma. *Int J Cancer* 2018;143(12):3209-3217.
6. Ren X, Yang X, Cheng B, et al. HOPX hypermethylation promotes metastasis via activating SNAIL transcription in nasopharyngeal carcinoma. *Nat Commun* 2017;8:14053.
7. Wang WY, Twu CW, Chen HH, et al. Plasma EBV DNA clearance rate as a novel prognostic marker for metastatic/recurrent nasopharyngeal carcinoma. *Clin Cancer Res* 2010;16(3):1016-1024.
8. Chan AT, Gregoire V, Lefebvre JL, et al. Nasopharyngeal cancer: EHNS-ESMO-ESTRO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2012;23(Suppl 7):vii83-vii85.
9. Liu L-T, Chen Q-Y, Tang L-Q, et al. Neoadjuvant or adjuvant chemotherapy plus concurrent CRT versus concurrent CRT alone in the treatment of nasopharyngeal carcinoma: A study based on EBV DNA. *J Natl Compr Canc Netw* 2019;17:703-710.
10. Tan T, Lim W-T, Fong K-W, et al. Concurrent chemo-radiation with or without induction gemcitabine, carboplatin, and paclitaxel: A randomized, phase 2/3 trial in locally advanced nasopharyngeal carcinoma. *Int J Radiat Oncol Biol Phys* 2015;91:952-960.
11. Cao S-M, Yang Q, Guo L, et al. Neoadjuvant chemotherapy followed by concurrent chemoradiotherapy versus concurrent chemoradiotherapy alone in locoregionally advanced nasopharyngeal carcinoma: A phase III multicentre randomised controlled trial. *Eur J Cancer* 2017;75:14-23.
12. Liu N, Chen NY, Cui RX, et al. Prognostic value of a microRNA signature in nasopharyngeal carcinoma: A microRNA expression analysis. *Lancet Oncol* 2012;13(6):633-641.
13. Leung SF, Zee B, Ma BB, et al. Plasma Epstein-Barr viral deoxyribonucleic acid quantitation complements tumor-node-metastasis staging prognostication in nasopharyngeal carcinoma. *J Clin Oncol* 2006;24(34):5414-5418.
14. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 2012;366(10):883-892.
15. Wang S, Li X, Li ZG, et al. Gene expression profile changes and possible molecular subtypes in differentiated-type nonkeratinizing nasopharyngeal carcinoma. *Int J Cancer* 2011;128(4):753-762.
16. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: The bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017;14:749-762.
17. Zhang B, Tian J, Dong D, et al. Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. *Clin Cancer Res* 2017;23(15):4259-4269.
18. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436-444.
19. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. *Annu Rev Biomed Eng* 2017;19:221-248.
20. Schmidhuber J. Deep learning in neural networks: An overview. *Neural Netw* 2015;61:85-117.

21. Wang S, Shi J, Ye Z, et al. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. *Eur Respir J* 2019;53:1800986.
22. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* 2018;24:1559-1567.
23. Lakhani P, Sundaram B. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017;284:574-582.
24. Lin L, Dou Q. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. *Radiology* 2019; 291(3):677-686.
25. Liang S, Tang F, Huang X, et al. Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. *Eur Radiol* 2019;29(4): 1961-1967.
26. Edge SB, Compton CC. The American Joint Committee on Cancer: The 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann Surg Oncol* 2010;17(6):1471-1474.
27. Ou SH, Zell JA, Ziogas A, et al. Epidemiology of nasopharyngeal carcinoma in the United States: Improved survival of Chinese patients within the keratinizing squamous cell carcinoma histology. *Ann Oncol* 2007; 18(1):29-35.
28. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. *IEEE Trans Image Process* 2004;13(4):600-612.
29. He KM, Zhang XY, Ren SQ, et al. *Deep residual learning for image recognition*. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE 2016; pp. 770-778.
30. Tieleman T, Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA Neural Networks Mach Learn 2012;4(2):26-31.
31. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 2017; 77(21):e104-e107.
32. De Jay N, Papillon-Cavanagh S, Olsen C, et al. mRMRe: An R package for parallelized mRMR ensemble feature selection. *Bioinformatics* 2013; 29(18):2365-2368.
33. Vasquez MM, Hu C, Roe DJ, Chen Z, Halonen M, Guerra S. Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: Simulation and application. *BMC Med Res Methodol* 2016;16(1):154.
34. Vrieze SI. Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol Methods* 2012;17(2): 228-243.
35. Ge C, Gu IY, Jakola AS, et al. Deep learning and multi-sensor fusion for glioma classification using multistream 2D convolutional networks. *Conf Proc IEEE Eng Med Biol Soc* 2018; 2018: 5894-5897.
36. Han X. MR-based synthetic CT generation using a deep convolutional neural network method. *Med Phys* 2017;44(4):1408-1419.
37. Jiang X, Li J, Kan Y, et al. MRI based radiomics approach with deep learning for prediction of vessel invasion in early-stage cervical cancer. *IEEE/ACM Trans Comput Biol Bioinform* 2020;1.
38. Zhao C, Miao J-J, Hua Y-J, et al. Locoregional control and mild late toxicity after reducing target volumes and radiation doses in patients with locoregionally advanced nasopharyngeal carcinoma treated with induction chemotherapy (IC) followed by concurrent chemoradiotherapy: 10-year results of a phase 2 study. *Int J Radiat Oncol Biol Phys* 2019;104(4):836-844.
39. Wang C, Tang X, Wang J, Song J, Xu Y. Induction chemotherapy plus concurrent chemoradiotherapy vs concurrent chemoradiotherapy in elderly patients with advanced nasopharyngeal carcinoma. *Otolaryngol Head Neck Surg* 2017;157(2):233-238.