Multiparametric-MRI-Based Radiomics Model for Differentiating Primary Central Nervous System Lymphoma From Glioblastoma: Development and Cross-Vendor Validation

Wei Xia, PhD,[1,2,3†] Bin Hu, MD,[3†] Haiqing Li, MD,[3] Chen Geng, PhD,[1,2,3] Qiuwen Wu, MS,[1,3] Liqin Yang, PhD,[1,3] Bo Yin, MD,[3] Xin Gao, PhD,[2] Yuxin Li, MD,[1,3*] and Daoying Geng, MD[1,3*]

Background: Preoperative differentiation of primary central nervous system lymphoma (PCNSL) from glioblastoma (GBM) is important to guide neurosurgical decision-making.
Purpose: To validate the generalization ability of radiomics models based on multiparametric-MRI (MP-MRI) for differentiating PCNSL from GBM.
Study Type: Retrospective.
Population: In all, 240 patients with GBM (n = 129) or PCNSL (n = 111).
Field Strength/Sequence: 3.0T scanners (two vendors). Sequences: fluid-attenuation inversion recovery, diffusion-weighted imaging (DWI), and contrast-enhanced T_1-weighted imaging (CE-T_1WI). Apparent diffusion coefficients (ADCs) were derived from DWI.
Assessment: Cross-vendor and mixed-vendor validation were conducted. In cross-vendor validation, the training set was 149 patients' data from vendor 1, and test set was 91 patients' data from vendor 2. In mixed-vendor validation, a training set was 80% of data from both vendors, and the test set remained at 20% of data. Single and multisequence radiomics models were built. The diagnoses by radiologists with 5, 10, and 20 years' experience were obtained. The integrated models were built combining the diagnoses by the best-performing radiomics model and each radiologist. Model performance was validated in the test set using area under the ROC curve (AUC). Histological results were used as the reference standard.
Statistical Tests: DeLong test: differences between AUCs. *U*-test: differences of numerical variables. Fisher's exact test: differences of categorical variables.
Results: In cross-vendor and mixed-vendor validation, the combination of CE-T_1WI and ADC produced the best-performing radiomics model, with AUC of 0.943 vs. 0.935, P = 0.854. The integrated models had higher AUCs than radiologists, with 5 (0.975 vs. 0.891, P = 0.002 and 0.995 vs. 0.885, P = 0.007), 10 (0.975 vs. 0.913, P = 0.029 and 0.995 vs. 0.900, P = 0.030), and 20 (0.975 vs. 0.945, P = 0.179 and 0.995 vs. 0.923, P = 0.046) years' experiences.
Data Conclusion: Radiomics for differentiating PCNSL from GBM was generalizable. The model combining MP-MRI and radiologists' diagnoses had superior performance compared to the radiologists alone.
Level of Evidence: 4
Technical Efficacy Stage: 2

J. MAGN. RESON. IMAGING 2020.

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.27344

Received Jun 4, 2020, Accepted for publication Aug 11, 2020.

*Address reprint requests to: Y.L. or D.G., Department of Radiology, Huashan Hospital, Fudan University, 12 Wulumuqi Middle Road, Shanghai 200040, China. E-mail: liyuxin@fudan.edu.cn (Li); gengdy@163.com (Geng)
†Wei Xia and Bin Hu are primary authors and contributed equally.
Contract grant sponsor: National Natural Science Foundation of China; Contract grant numbers: 61801474, 61672236; Contract grant sponsor: Suzhou Science and Technology Plan Project; Contract grant number: SYG201908; Contract grant sponsor: Science and Technology Commission of Shanghai Municipality; Contract grant number: 19411951200.

From the [1]Academy for Engineering and Technology, Fudan University, Shanghai, China; [2]Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, China; and [3]Department of Radiology, Huashan Hospital, Fudan University, Shanghai, China

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

PRIMARY CENTRAL NERVOUS SYSTEM LYMPHOMA (PCNSL) and glioblastoma (GBM) are two commonly diagnosed malignant primary brain tumors.[1] The accurate differentiation of PCNSL from GBM is clinically crucial due to the different treatment strategies between them.[2–4] The current treatment guidelines adopt aggressive resection for GBM, while for PCNSL noninvasive therapy including chemotherapy, targeted therapies, or whole brain radiotherapy are mostly recommended.[5] Although the reference standard for tumor diagnosis is histological examination after stereotactic biopsy, this has a 6.6% probability of severe complications such as cerebral hemorrhage or death.[6] In addition, the stereotactic biopsy is more challenging to be carried out for the deep part of the mid-brain than the superficial part.[7] Therefore, accurate noninvasive diagnosis of PCNSL and GBM is important to guide neurosurgical decision-making.

Magnetic resonance imaging (MRI) generally enables the noninvasive differentiation of typical PCNSL from GBM and advanced MRI techniques, such as dynamic contrast-enhanced imaging, arterial spin labeling, magnetic resonance spectroscopy, and dynamic susceptibility-weighted contrast-enhanced imaging, have been helpful in the differentiation of more complex cases.[8] However, extra expense and time are needed to perform advanced MRI techniques and these are generally not performed in routine clinical practice.[9] Conventional multiparametric-MRI (MP-MRI) including contrast-enhanced T_1-weighted imaging (CE-T_1WI), fluid-attenuation inversion recovery (FLAIR), and diffusion-weighted imaging (DWI) are almost always performed. However, using these conventional techniques differentiation of PCNSL from GBM is challenging. Atypical PCNSL with necrosis and hemorrhage, for example, may mimic GBM, while atypical GBM without visible necrosis is similar to PCNSL.[10–12] Constructing a diagnostic model through machine-learning techniques, which fully utilizes the conventional MRI data and could be very useful.

Radiomics is a form of machine learning that is used to extract high-throughput quantitative image features and train a predictive model.[13] Based on the large number of image features extracted from MRI that describe tumor heterogeneity, radiomics has shown great potential in building models that are capable of differentiating PCNSL from GBM.[9,14–17] However, as different MRI sequences have been introduced into clinical practice, it is important to determine the optimal combination of MRI sequences for the development of the radiomics model. In addition, previous studies have been conducted with data acquired from a single MR scanner.[9,14–17] This has resulted in concerns about the risk of overfitting the models that may be biased by subtle differences between MRI hardware or imaging parameters. The generalizability of radiomics models to correctly interpret data acquired by different MR scanners with different protocol parameters is important[18] and would allow the application of radiomics in clinical practice. Furthermore, while in previous studies the performance of radiomics models were evaluated against radiologists,[9,14,16,17] models were initially developed as support tools to assist radiologists rather than to replace them.[19] It would therefore be helpful to investigate the benefit of integrating the diagnosis by the radiomics model with that of the radiologist.

The aim of this study was to develop and validate the generalizability of multiparametric-MRI (MP-MRI)-based radiomics models for differentiating PCNSL from GBM and to assess the additional benefit of integrating the radiomics model with the radiologists' diagnoses.

Materials and Methods

Patients

The Institutional Review Board of our center approved this retrospective study, and the requirement for evidence of informed consent was waived. The inclusion criteria were as follows: 1) histologically proven PCNSL or GBM from March 2011 to March 2019; 2) preoperative MRI. The exclusion criteria were as follows: i) lacking any one of the following conventional MRI sequences: CE-T_1WI, FLAIR, DWI; ii) received prior treatment before MRI scanning; iii) image data with miscellaneous artifacts. In total, 240 patients were enrolled in this study, including 129 patients with GBM and 111 patients with PCNSL.

Image Acquisition

MRI was performed on 3T scanners with 8-channel head coils. MR images of 149 patients were acquired with two Verio 3T scanners (Vendor 1: Siemens, Erlangen, Germany), and MR images of 91 patients were obtained by one Signa 3T scanner (Vendor 2: GE Healthcare, Milwaukee, WI). The MRI protocols are shown in Table 1. FLAIR and DWI images were obtained. Based on DWI images, the ADC maps were generated automatically by the MRI workstation. After intravenous injection of gadopentetate dimeglumine (0.1 mmol/kg), axial CE-T_1WI images were obtained.

Radiomics Feature Extraction

For each patient the tumor volume of interest (VOI) was delineated manually using Medical Imaging Interaction Toolkit (MITK) software (v. 2013.12.0; http://www.mitk.org/). The tumor VOI was drawn on the FLAIR images, covering the tumor tissue and peripheral edema, and was adjusted following viewing of DWI and CE-T_1WI images. To enable the VOI to be used with images from all MRI sequences, the CE-T_1WI, DWI, and ADC images were resampled and aligned to the same resolution, spacing, and position as the FLAIR images using the open-source Insight Segmentation and Registration Toolkit (ITK, v. 4.7.2; https://itk.org/).[20] To standardize the MR images from all sequences, the mean value and the standard deviation of intensity in the images from each MRI volume were calculated, and each was normalized by the z-score method, which consisted of subtracting the mean intensity and dividing by the standard deviation of intensity.[21–23]

TABLE 1. MRI Scanning Protocols

Sequences	Verio 3T (1)	Verio 3T (2)	Signa 3T
T_1WI	TR/TI/TE = 2000/860/9 msec, matrix size = 320*199, FOV = 213*240, slice thickness = 8 mm, slice spacing = 0.94 mm.	TR/TI/TE = 2000/857/17 msec, matrix size = 256*168, FOV = 201*230, slice thickness = 8 mm, slice spacing = 0.9 mm.	TR/TI/TE = 1935/750/21 msec, matrix size = 288*192, FOV = 240*240, slice thickness = 8 mm, slice spacing = 0.47 mm.
FLAIR	TR/TI/TE = 9000/2501/94 msec, matrix size = 256*160, FOV = 213*240, slice thickness = 8 mm, slice spacing = 0.45 mm.	TR/TI/TE = 9000/2500/102 msec, matrix size = 256*190, FOV = 201*230, slice thickness = 8 mm, slice spacing = 0.45 mm.	TR/TI/TE = 8602/2100/123 msec, matrix size = 288*192, FOV = 240*240, slice thickness = 8 mm, slice spacing = 0.47 mm.
DWI (b = 0, 1000 s/mm^2)	TR/TE = 6600/100 msec, matrix size = 192*192, FOV = 240*240, slice thickness = 8 mm, slice spacing = 1.77 mm.	TR/TE = 5000/104 msec, matrix size = 192*192, FOV = 229*229, slice thickness = 8 mm, slice spacing = 1.2 mm.	TR/TE = 4800/74 msec, matrix size = 128*130, FOV = 240*240, slice thickness = 8 mm, slice spacing = 0.94 mm.

TR = repetition time; TI = inversion time; TE = echo time; FOV = field of view; T_1WI = T_1-weighted imaging; FLAIR = fluid-attenuation inversion recovery; DWI = diffusion-weighted imaging.

MRI feature extraction was conducted using an open-source Python package Pyradiomics (v. 2.1.2; http://www.radiomics.io/pyradiomics.html).[24] A total of 851 image features were calculated for each MRI volume, including 14 shape-based features, 18 first-order statistics features, 75 texture features, and 744 wavelet features. The shape-based features were extracted in 3D by using shape descriptors to quantify the shape of the tumor VOI. First-order statistic features described the distribution of voxel intensities within the tumor VOI. Texture features employed gray-level matrixes to represent the spatial heterogeneity of intensities within the tumor VOI, with the bin width of intensity being set to 32. To extract more image features quantifying the tumor heterogeneity, the 3D wavelet filtering was applied to each MRI volume. The 3D wavelet filtering decomposed the original volume V into eight decompositions. Let L and H be the low-pass and the high-pass filtering, the wavelet decompositions of V can be labeled as V_LLL, V_LLH, V_LHL, V_LHH, V_HLL, V_HLH, V_HHL, and V_HHH. For instance, V_HHL is obtained from x-directional high-pass filtering, y-directional high-pass filtering, and z-directional low-pass filtering of V. The obtained decompositions have the same size of the original image. For each of the eight decompositions, the 18 first-order statistical features and the 75 texture features were calculated, thus the corresponding 744 wavelet features were obtained. The wavelet filtering was implemented by PyWavelets package (v. 1.0.1). The details of all features are described online (https://pyradiomics.readthedocs.io/en/2.1.2/features.html).[24]

Radiomics Model Development

To improve the generalizability of the radiomics model, the features with low intra- or interobserver reproducibility were excluded.[20,25] Two radiologists (B.H. and H.L.) with 5 years of experience performed the same delineation of the tumor VOI for all patients: radiologist 1 delineated the tumor VOI twice at different times and radiologist 2 carried out the delineation once. The radiomics features were calculated after each delineation and intra- and interobserver reproducibility determined for each feature. Features with low reproducibility (intra- or interobserver intraclass coefficient [ICC] below 0.75) were excluded.[20] In addition, pairwise feature Spearman correlation coefficients (SCCs) were calculated to build a correlation matrix, and the feature pairs with SCC higher than 0.9 were identified as highly correlated.[13,20] In each highly correlated feature pair, the SCCs between a feature and all the other features was calculated, and the feature with the larger mean SCC was considered to be redundant and excluded. Subsequently, a minimum redundancy maximum relevance (mRMR) feature selection method[26] was employed to control the number of features remaining to within 1/10 of the number of cases to reduce the risk of model overfitting.[27]

The selected radiomics features were input to the least absolute shrinkage and selection operator (LASSO) for radiomics model building. LASSO is a generalized linear model (GLM) that performs both feature selection and regularization to enhance the classification accuracy and interpretability of the model,[28] and has shown advantages over other classifiers in radiomics studies.[13,18] For each MRI sequence, a single-sequence radiomics model was trained using 10-fold cross-validation. Multisequence radiomics models were generated by integrating single-sequence radiomics models using multivariable logistic regression with all possible combinations of sequences. In total, four single-sequence radiomics models and 11 multisequence radiomics models were built. The single-sequence radiomics models were the linear weighted sum of radiomics features, and the multisequence radiomics models were the linear weighted sum of the outputs of single-sequence radiomics models. The outputs of models were transformed to probabilities by sigmoid function. The radiomics models allowed assigning patients with a radiomics score that was the diagnostic probability of PCNSL.

Comparison and Integration of Diagnosis by Model and Radiologists

Three radiologists (Y.Y., Y.T., and Y.L. with over 5, 10, and 20 years' experience in neuroradiology) blinded to the histological results were assigned to review MRI images and classify the cases as PCNSL or GBM independently. The radiologists with over 5, 10, and 20 years' experience were regarded as junior radiologist, intermediate-level radiologist, and senior radiologist, respectively. For each radiologist, an integrated model was built using multivariable logistic regression to make a final decision by combining the diagnoses by radiologist with the best-performing radiomics model.

Model Validation

A cross-vendor validation was conducted to test the cross-vendor generalizability of the models developed; the images of the 149 patients acquired by the Vendor 1's scanners were grouped as the training set for model development; and the images of the 91 patients acquired by the Vendor 2's scanner were used as the test set. All models were built using the training set and independently validated on the test set.

In addition, a mixed-vendor validation was conducted to test the robustness of the model development procedure. The models were trained on the dataset and consisted of 80% of the patients' images from Vendor 1 and 80% of the patients' images from Vendor 2. Then the models were tested on a dataset of the remaining 20% patients' images from both vendors. The results of mixed-vendor validation were compared to the cross-vendor validation.

Statistical Analysis

The diagnostic performances of the models were assessed using receiver operating characteristic (ROC) curve analysis and measured by the area under the ROC curve (AUC). Differences between ROC curves were assessed by the DeLong test using MedCalc software (v. 11.4.2.0, http://www.medcalc.be/).[29] By designating the patients with PCNSL as positive cases, the sensitivity, specificity, and accuracy of the models were calculated. In statistical tests of clinical characteristics, the Mann–Whitney *U*-test was used for numerical variables, and Fisher's exact test was used for categorical variables. The intra- and interobserver reproducibility of features was determined by the ICC. The feature selection, the development and validation of models, and statistical tests were conducted using R software (v. 3.6.2, https://www.r-project.org/). $P < 0.05$ was considered statistically significant. The flowchart of model development and validation is depicted in Fig. 1.

Results

Patient characteristics are provided in Table 2. There were no significant differences in the age, gender, or pathology between the training set and the test set in both cross-vendor and mixed-vendor validations.

A total of 851 radiomics features were extracted for each MRI sequence. After the features with low reproducibility were excluded, this was reduced to 669, 501, 711, and 631 for CE-T_1WI, FLAIR, DWI, and ADC, respectively. Further exclusion of redundant features following identification of highly correlated feature pairs resulted in 141 (179) CE-T_1WI features, 85 (85) FLAIR features, 151 (140) DWI features, and 133 (141) ADC features remained in cross-vendor validation (mixed-vendor validation). After LASSO feature selection, there were 6 (6), 3 (1), 2 (4), and 5 (6) features in the CE-T_1WI, FLAIR, DWI, and ADC-based

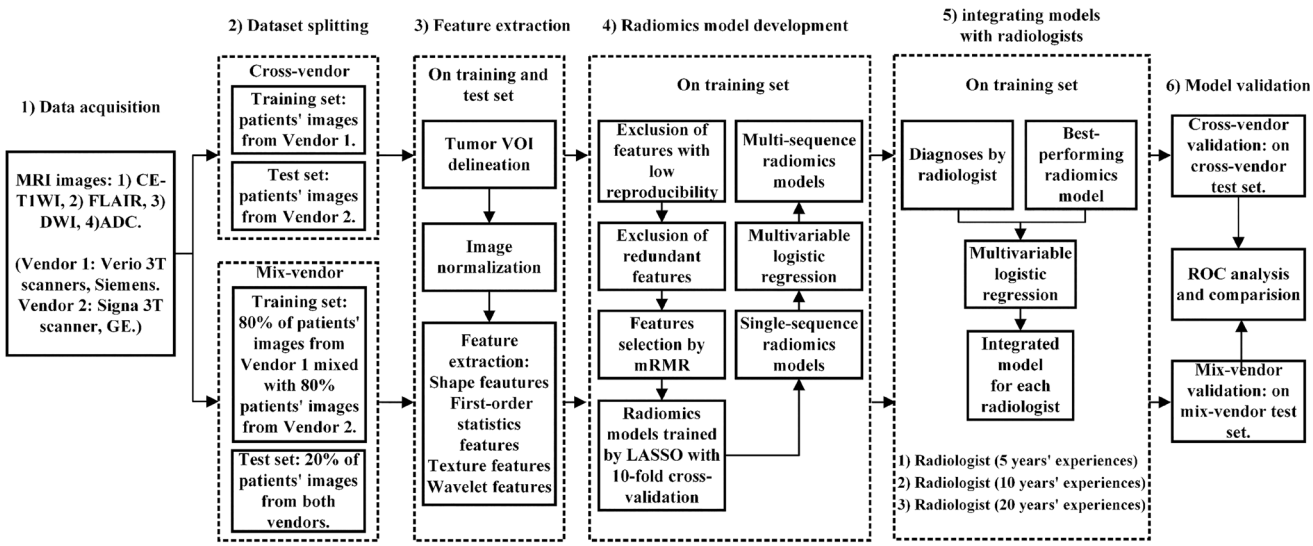

FIGURE 1: Flowchart of model development and validation. 1) Data acquisition: the MRI images were acquired from two vendors' scanners; 2) Dataset splitting: The cross-vendor approach split the data as training and test sets according to vendor type, and the mixed-vendor approach split the data from both vendors as training and test sets by a ratio of 8:2; 3) Feature extraction: the radiomics features were extracted from the delineated tumor VOI; 4) Radiomics model development: on each training set, the feature reduction and selection were conducted, and the single-sequence radiomics models were built by LASSO with 10-fold cross validation, then the multisequence models were built by integrating single-sequence models; 5) integrating models with radiologists: on each training set, the integrated model was built by combining the diagnoses by each radiologist with the best-performing radiomics model; 6) model validation: the cross-vendor and mixed-vendor-based models were tested and compared.

TABLE 2. Patient Profiles of Training and Test Set

		Cross-vendor validation			Mixed-vendor validation		
Characteristic	**Overall**	**Training set**	**Test set**	***P*-value**	**Training set**	**Test set**	***P*-value**
Age (years)	53.8 ± 13.0	53.3 ± 12.7	54.6 ± 13.5	0.332	54.1 ± 13.0	52.7 ± 13.2	0.486
Gender				0.892			1.000
Male	142	89	53		114	28	
Female	98	60	38		78	20	
Pathology				0.350			0.258
GBM	129	84	45		107	22	
PCNSL	111	65	46		85	26	
Total number	240	149	91		192	48	

Ages are shown as mean ± standard deviation, and others are number of patients. The *P* values are derived from the comparison between training set and test set.

radiomics models in cross-vendor validation (mixed-vendor validation), respectively. The details of the single-sequence radiomics model are listed in Table S1–S4 of Supplement A, and the details of multisequence radiomics models are listed in Table S5–S15 of Supplement A.

The test performances of the models are listed in Table 3. In both cross-vendor and mixed-vendor validation, among the single-sequence radiomics models the model based on CE-T_1WI achieved the best performance with the highest AUC. For the multisequence radiomics models, the best-performing radiomics model was derived from the combination of CE-T_1WI and the ADC map (CE-T_1WI + ADC model). The ROC curves of the CE-T_1WI + ADC model is given in Fig. 2. In addition, there were no statistically significant differences in the AUC of models between cross-vendor and mixed-vendor validation.

The diagnostic performances of the three independent radiologists are shown in Table 3 and the corresponding ROC curves are presented in Fig. 2. In both cross-vendor and mixed-vendor validation, the AUC of CE-T_1WI + ADC model was comparable to that of the senior radiologist (0.943 vs. 0.945, P = 0.948 and 0.935 vs. 0.923, P = 0.824). Integrating the diagnoses of the junior radiologist and the CE-T_1WI + ADC model resulted in significantly higher AUC than the junior radiologist (0.975 vs. 0.891, P = 0.002 and 0.995 vs. 0.885, P = 0.007) and intermediate-level radiologist (0.975 vs. 0.913, P = 0.029 and 0.995 vs. 0.900, P = 0.030), and higher than the senior radiologist (0.975 vs. 0.945, P = 0.179 and 0.995 vs. 0.923, P = 0.046). Integrating the diagnoses of the intermediate-level radiologist and the CE-T_1WI + ADC model resulted in a significant improvement in AUC compared with the intermediate-level radiologist (0.981 vs. 0.913, P = 0.009 and 0.997 vs. 0.900, P = 0.019), and higher than the senior radiologist (0.981 vs. 0.945, P = 0.088 and 0.997 vs. 0.923, P = 0.045). Integrating the diagnoses of the senior radiologist and the CE-T_1WI + ADC model also resulted in improvement in AUC compared with the senior radiologist alone (0.980 vs. 0.945, P = 0.074 and 0.995 vs. 0.923, P = 0.046). The integrated models' parameters are given in Table S16 of Supplement A. The ROC curves of the integrated models are given in Fig. 2.

The chord diagrams[30] are illustrated in Fig. 3 and show the corrective effect of the integrated models. The chord diagrams of the cross-vendor validation are plotted in Fig. 3a–c. In Fig. 3a, it can be seen that six out of 10 patients who were misdiagnosed by the junior radiologist were correctly diagnosed by the integrated model. In Fig. 3b, there are four out of eight patients who were misdiagnosed by the intermediate-level radiologist and were correctly diagnosed by the integrated model. In Fig. 3c, it can be seen that three out of five patients who were misdiagnosed by the senior radiologist were correctly diagnosed by the integrated model, but there was one patient correctly diagnosed by the senior radiologist who was misdiagnosed by the integrated model. Similar results can be found in the mixed-vendor validation and shown in Fig. 3d–f. The MR images of representative cases are shown in Fig. 4.

Discussion

In this study we developed radiomics models using the image features extracted from MP-MRI for differentiating PCNSL from GBM. The MR images used in routine radiology workflow—CE-T_1WI, FLAIR, DWI, and ADC—were used

TABLE 3. Test Performance of Models and Radiologists

	Cross-vendor validation				Mixed-vendor validation				
Models and radiologists	**AUC**	**ACC**	**SEN**	**SPE**	**AUC**	**ACC**	**SEN**	**SPE**	***P*-value**
CE-T_1WI	0.937	0.890	0.870	0.911	0.933	0.896	0.846	0.954	0.937
FLAIR	0.897	0.868	0.804	0.933	0.884	0.854	0.730	1.000	0.846
DWI	0.905	0.857	0.783	0.933	0.886	0.833	0.692	1.000	0.757
ADC	0.925	0.868	0.782	0.956	0.900	0.854	0.846	0.864	0.648
CE-T_1WI + DWI	0.927	0.890	0.870	0.911	0.925	0.917	0.923	0.909	0.973
CE-T_1WI + FLAIR	0.937	0.890	0.870	0.911	0.918	0.895	0.923	0.863	0.716
CE-T_1WI + ADC	0.943	0.912	0.891	0.933	0.935	0.917	0.923	0.909	0.854
FLAIR+DWI	0.898	0.868	0.804	0.933	0.886	0.833	0.846	0.818	0.852
FLAIR+ADC	0.926	0.879	0.804	0.956	0.913	0.854	0.769	0.954	0.788
DWI + ADC	0.928	0.890	0.826	0.956	0.897	0.854	0.807	0.909	0.656
CE-T_1WI + FLAIR+DWI	0.939	0.890	0.848	0.933	0.916	0.895	0.923	0.864	0.665
CE-T_1WI + FLAIR+ADC	0.940	0.901	0.870	0.933	0.923	0.896	0.923	0.864	0.728
CE-T_1WI + DWI + ADC	0.935	0.879	0.870	0.889	0.927	0.917	0.923	0.909	0.859
FLAIR+DWI+ ADC	0.917	0.824	0.740	0.911	0.900	0.812	0.730	0.909	0.744
CE-T_1WI + FLAIR+DWI + ADC	0.930	0.868	0.847	0.889	0.923	0.875	0.884	0.863	0.879
Junior radiologist	0.891	0.890	0.804	0.978	0.885	0.875	0.769	1.000	0.903
Integrated model_junior	0.975	0.956	0.935	0.978	0.995	0.958	0.923	1.000	0.265
Intermediate-level radiologist	0.913	0.912	0.870	0.956	0.900	0.896	0.846	0.954	0.814
Integrated model_intermediate	0.981	0.945	0.935	0.956	0.997	0.979	1.000	0.954	0.349
Senior radiologist	0.945	0.945	0.913	0.978	0.923	0.917	0.846	1.000	0.607
Integrated model_senior	0.980	0.956	0.935	0.978	0.995	0.979	1.000	0.954	0.383

CE-T_1WI = contrast-enhanced T_1-weighted imaging; FLAIR = fluid-attenuation inversion recovery; DWI = diffusion-weighted imaging; ADC = apparent diffusion coefficient; Integrated model_x = integrating the diagnoses by the x (junior, intermediate, or senior) radiologist and the CE-T_1WI + ADC model; AUC = area under the receiver operating characteristic curve; ACC = accuracy; SEN = sensitivity; SPE = specificity. The sensitivity and specificity were calculated by designating PCNSL as a positive case. The *P* values are derived from the comparisons of ROC curves between cross-vendor and mixed-vendor validation.

individually and in all combinations for model building. The diagnostic performance of models were cross-vendor and mixed-vendor validated, and good results were acquired in both validations. The CE-T_1WI model achieved the best performance with the highest AUC among all single-sequence models. For the multisequence radiomics models, the CE-T_1WI + ADC model achieved the best performance. The models were compared with three radiologists, and the CE-T_1WI + ADC model was comparable to the senior radiologist. Furthermore, the integrated models were built by combining the diagnoses by the CE-T_1WI + ADC model and the radiologists to provide a final decision. The integrated models achieved better performance than radiologists or radiomics models alone.

Both cross-vendor and mixed-vendor validations were performed, and there were no statistically significant differences in the AUC of models between the two validations. In both validations, the best-performing radiomics model achieved good performance that was comparable to the senior radiologist. The encouraging performance suggested that the radiomics approach was robust and generalizable, regardless of specific MR vendor and protocol.

Among single-sequence radiomics models, the CE-T_1WI-based model achieved the highest AUC. CE-T_1WI is the preferred MRI sequence in brain tumor diagnosis and can clearly show the tumor entity and necrosis. Heterogeneous enhancement was commonly observed in GBM, while

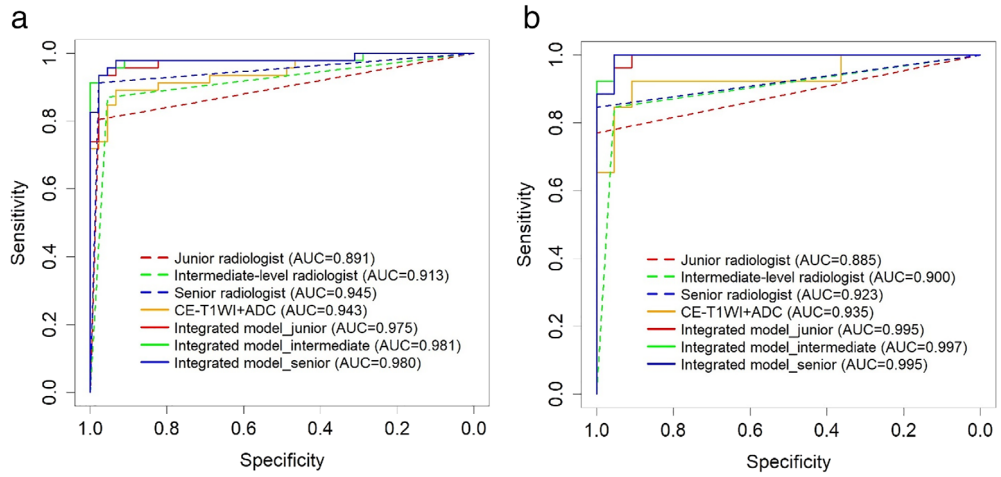

FIGURE 2: The ROC curves of radiologists and models. (a) Cross-vendor validation, (b) Mixed-vendor validation.

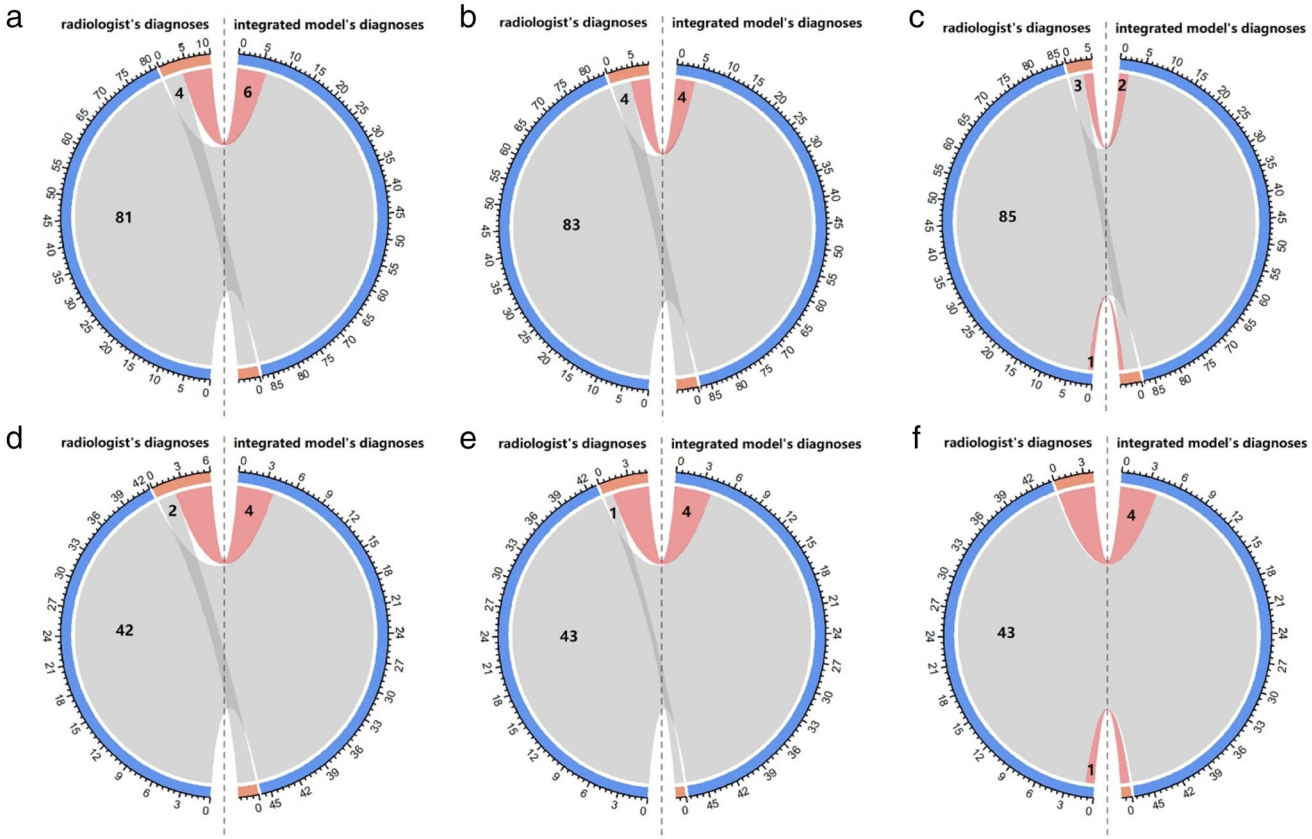

FIGURE 3: The chord diagrams for showing the corrective effect of the integrated models. (a–c) The chord diagrams for comparison of the diagnoses by junior, intermediate-level, and senior radiologist and the corresponding integrated model in cross-vendor validation. (d–f) The chord diagrams for comparison of the diagnoses by junior, intermediate-level, and senior radiologist and the corresponding integrated model in mixed-vendor validation.

homogenous enhancement was always seen in PCNSL. Previous radiomics studies based on CE-T_1WI showed that the CE-T_1WI-based model could achieve promising diagnostic performance, with results similar to ours.[14–16] The ADC map-based model showed the second highest diagnostic performance. ADC values measure the water diffusional restriction and can indicate the cellularity, necrosis, and cystic degeneration in tumors.[31] Previous studies have shown that cellularity was higher in PCNSL compared to GBM.[32] In the features that make up the ADC model, there is a feature named original_firstorder_10Percentile, which is the 10th percentile of ADC values within the tumor VOI, which suggests that this feature is important for differentiating PCNSL from GBM, and this is consistent with a previous study.[33] In the CE-T_1WI and ADC-based model, the wavelet features had the largest number, and similar findings were reported in

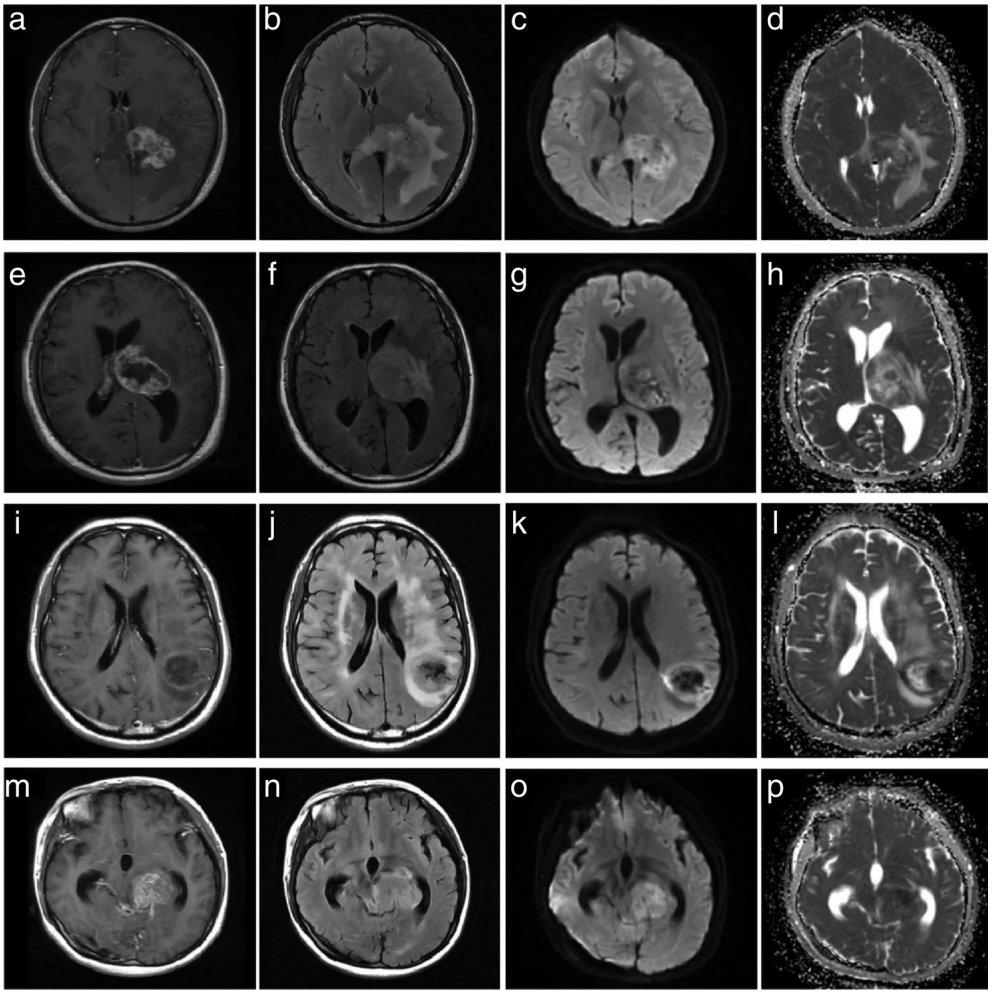

FIGURE 4: MR images of representative cases. (a,e,i,m) CE-T_1WI images. (b,f,j,n) FLAIR images. (c,g,k,o) DWI images. (d,h,l,p) ADC images. (a–d) A case with histologically confirmed PCNSL who was misdiagnosed as GBM by radiologists and were correctly diagnosed by the CE-T_1WI + ADC model. (e–h) A case with histologically confirmed GBM who was misdiagnosed as PCNSL by the junior radiologist and were correctly diagnosed by the CE-T_1WI + ADC model. (i–l) A case with histologically confirmed PCNSL who was misdiagnosed as GBM by the CE-T_1WI + ADC model and all radiologists. (m–p) A case with histologically confirmed GBM who was misdiagnosed as PCNSL by the CE-T_1WI + ADC model was correctly diagnosed by all radiologists.

previous studies that showed that wavelet features were the major components in radiomic models.[20,34,35] The FLAIR-based model had inferior performance compared with other sequences, as FLAIR mainly identifies edema that is not as significant as tumor tissue for differentiation in clinical practice. Among the multisequence radiomics models, the combination of CE-T_1WI and ADC achieved better performance than the combination of all sequences; this may due to the fact that DWI and FLAIR could not provide complementary information for CE-T_1WI and ADC, and introducing redundant information may decrease the model performance based on the principle of machine learning.

The diagnostic performance of the models were compared with the radiologists. The sensitivity of all radiologists was lower than the specificity. This is consistent with previous studies that showed that atypical PCNSL was a challenge in radiological diagnosis.[14,16,17] For all radiologists, the integrated model showed improvement in the AUC compared with that of the radiologists or radiomics model alone. In particular, for the junior and intermediate-level radiologists, the integrated model significantly improved the AUC compared to the radiologists themselves, and had higher AUC than the senior radiologist. This indicated that the machine-learning derived knowledge and human knowledge were mutually complementary and that the integration approach may facilitate effective cooperation between machines and human readers for more accurate diagnosis.

Limitations

First, the models developed need to be further validated in different centers before application in clinical work. The development tools for model building are publicly available and the features and coefficients used in the models have been provided, ensuring that they can be reproduced at other centers for validation. Second, while the models were developed to differentiate PCNSL from GBM, there are many

more types of brain lesions that should be included in future studies in order to build a nondichotomous diagnostic model that can realize fully automated diagnosis.

Conclusion

In this study we developed and validated MP-MRI-based radiomics models for differentiating PCNSL from GBM, which proved to be accurate and generalizable. The model has the potential to provide supplementary diagnoses to radiologists and to improve diagnostic performance.

Acknowledgment

The authors thank Dr. Yang Yu and Dr. Yin Tang for reviewing the MRI images.

References

1. Lapointe S, Perry A, Butowski NA. Primary brain tumours in adults. Lancet 2018;392:432-446.
2. Hunt MA, Jahnke K, Murillo TP, Neuwelt EA. Distinguishing primary central nervous system lymphoma from other central nervous system diseases: A neurosurgical perspective on diagnostic dilemmas and approaches. Neurosurg Focus 2006;21:1-7.
3. DeAngelis LM. Primary CNS lymphoma: Treatment with combined chemotherapy and radiotherapy. J Neurooncol 1999;43:249-257.
4. Sanai N, Polley M-Y, McDermott MW, Parsa AT, Berger MS. An extent of resection threshold for newly diagnosed glioblastomas: Clinical article. J Neurosurg 2011;115:3-8.
5. Rubenstein JL, Gupta NK, Mannis GN, LaMarre AK, Treseler P. How I treat CNS lymphomas. Blood 2013;122:2318-2330.
6. Lu Y, Yeung C, Radmanesh A, Wiemann R, Black PM, Golby AJ. Comparative effectiveness of frame-based, frameless, and intraoperative magnetic resonance imaging–guided brain biopsy techniques. World Neurosurg 2015;83:261-268.
7. Cheng G, Yu X, Zhao H, et al. Complications of stereotactic biopsy of lesions in the sellar region, pineal gland, and brainstem: A retrospective, single-center study. Medicine 2020;99:e18572.
8. Suh CH, Kim HS, Jung SC, Park JE, Choi CG, Kim SJ. MRI as a diagnostic biomarker for differentiating primary central nervous system lymphoma from glioblastoma: A systematic review and meta-analysis: Differentiating PCNSL from glioblastoma. J Magn Reson Imaging 2019;50:560-572.
9. Suh HB, Choi YS, Bae S, et al. Primary central nervous system lymphoma and atypical glioblastoma: Differentiation using radiomics approach. Eur Radiol 2018;28:3832-3839.
10. Savage J, Quint D. Atypical imaging findings in an immunocompetent patient. Primary central nervous system lymphoma. JAMA Oncol 2015; 1:247-248.
11. Bühring U, Herrlinger U, Krings T, Thiex R, Weller M, Küker W. MRI features of primary central nervous system lymphomas at presentation. Neurology 2001;57:393-396.
12. Al-Okaili RN, Krejza J, Woo JH, et al. Intraaxial brain masses: MR imaging–based diagnostic strategyMixed-Initial experience. Radiology 2007;243:539-550.
13. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data. Radiology 2016;278:563-577.
14. Alcaide-Leon P, Dufort P, Geraldo AF, et al. Differentiation of enhancing glioma and primary central nervous system lymphoma by texture-based machine learning. Am J Neuroradiol 2017;38:1145-1150.
15. Chen Y, Li Z, Wu G, et al. Primary central nervous system lymphoma and glioblastoma differentiation based on conventional magnetic resonance imaging by high-throughput SIFT features. Int J Neurosci 2018;128:608-618.
16. Kang D, Park JE, Kim Y-H, et al. Diffusion radiomics as a diagnostic model for atypical manifestation of primary central nervous system lymphoma: Development and multicenter external validation. Neuro-Oncology 2018;20:1251-1261.
17. Yamashita K, Yoshiura T, Arimura H, et al. Performance evaluation of radiologists with artificial neural network for differential diagnosis of intra-axial cerebral tumors on MR images. Am J Neuroradiol 2008;29: 1153-1158.
18. Nguyen AV, Blears EE, Ross E, Lall RR, Ortega-Barnett J. Machine learning applications for the differentiation of primary central nervous system lymphoma from glioblastoma on imaging: A systematic review and meta-analysis. Neurosurg Focus 2018;45:E5.
19. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: Threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp 2018;2:35.
20. Meng X, Xia W, Xie P, et al. Preoperative radiomic signature based on multiparametric magnetic resonance imaging for noninvasive evaluation of biological characteristics in rectal cancer. Eur Radiol 2018;29(6): 3200-3209.
21. Reinhold JC, Dewey BE, Carass A, Prince JL: Evaluating the impact of intensity normalization on MR image synthesis. ArXiv181204652 Cs 2018.
22. Lu L, Lv W, Jiang J, et al. Robustness of radiomic features in [11C]choline and [18F]FDG PET/CT imaging of nasopharyngeal carcinoma: Impact of segmentation and discretization. Mol Imaging Biol 2016;18: 935-945.
23. Li Y, Jian J, Pickhardt PJ, et al. MRI-based machine learning for differentiating borderline from malignant epithelial ovarian tumors: A multicenter study. J Magn Reson Imaging 2020;52:897-904.
24. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017; 77:e104-e107.
25. 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. J Clin Oncol 2016;34:2157-2164.
26. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27:1226-1238.
27. Nie K, Shi L, Chen Q, et al. Rectal cancer: Assessment of neoadjuvant chemoradiation outcome based on radiomics of multiparametric MRI. Clin Cancer Res 2016;22:5256-5264.
28. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1-22.
29. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 1988;44:837-845.
30. Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in R. Bioinformatics 2014;30:2811-2812.
31. Chen L, Liu M, Bao J, et al. The correlation between apparent diffusion coefficient and tumor cellularity in patients: A meta-analysis. PLoS One 2013;8:e79008.
32. Chiavazza C, Pellerino A, Ferrio F, Cistaro A, Soffietti R, Rudà R. Primary CNS lymphomas: Challenges in diagnosis and monitoring. Biomed Res Int 2018;2018:1-16.
33. Choi YS, Lee H-J, Ahn SS, et al. Primary central nervous system lymphoma and atypical glioblastoma: Differentiation using the initial area under the curve derived from dynamic contrast-enhanced MR and the apparent diffusion coefficient. Eur Radiol 2017;27:1344-1351.
34. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5.
35. Wu W, Parmar C, Grossmann P, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. Front Oncol 2016;6.