Early prognostication of overall survival for pediatric diffuse midline gliomas using MRI radiomics and machine learning

Xinyang Liu, Zhifan Jiang, Holger R. Roth, Syed Muhammad Anwar, Erin R. Bonner, Aria Mahtabfar, Roger J. Packer, Anahita Fathi Kazerooni, Miriam Bornhorst, Marius George Linguraru

Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital (XL, ZJ, SMA, MGL)

Brain Tumor Institute, Children's National Hospital (ERB, RJP, MB)

School of Medicine and Health Sciences, George Washington University (SMA, ERB, MGL)

NVIDIA (HRR)

Center for Data-Driven Discovery in Biomedicine (D3b), Children's Hospital of Philadelphia (AFK, AM)

Department of Neurosurgery, University of Pennsylvania (AFK)

Center for AI & Data Science for Integrated Diagnostics (AI2D) and Cneter for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania (AFK)

Running title: Survival prognostication for pediatric DMG

ABSTRACT

Background: Diffuse midline gliomas (DMG) are aggressive pediatric brain tumors. MRI is the standard non-invasive tool for DMG diagnosis and monitoring. We developed an automatic pipeline to segment subregions of DMG and select radiomic features to predict patient overall survival (OS).

Methods: We acquired diagnostic and post-radiation therapy (RT) multisequence MRI (T1, T1ce, T2, and T2 FLAIR) and manual segmentations of 53 (internal cohort) and 16 (external cohort) DMG patients. We pretrained a deep learning model on an adult brain tumor dataset, and finetuned the model on our internal cohort to segment tumor core (TC) and whole tumor (WT). PyRadiomics and sequential feature selection were used for feature extraction and selection based on the segmented volumes. Two machine learning models were trained on our internal cohort to predict patient 1-year survival from diagnosis. One model used only diagnostic features (baseline study) and the other used both diagnostic and post-RT features (post-RT study).

Results: For segmentation, Dice score (mean [median]±SD) was 0.91 (0.94)±0.12/0.74 (0.83)±0.32 for TC and 0.88 (0.91)±0.07/0.86 (0.89)±0.06 for WT of internal/external cohorts. For OS prediction, accuracy was 77%/81% for the baseline study and 85%/78% for the post-RT study of internal/external cohorts. Our results suggest post-RT features are more discriminative and reliable compared with diagnostic features. Smaller post-RT TC/WT volume ratio indicates longer OS. Our model predicts with high accuracy which patients have short OS.

Conclusions: We demonstrated how a fully automatic approach to compute imaging biomarkers of DMG from multisequence MRI can accurately and non-invasively predict overall survival for impacted pediatric patients.

KEYWORDS

diffuse midline glioma (DMG); magnetic resonance imaging; tumor segmentation; radiomics; machine learning

KEYPOINTS

This is the first fully automatic deep learning/machine learning MRI study to predict DMG survival.

Post-radiation therapy features are more discriminative and reliable than diagnostic features.

Smaller post-radiation therapy tumor core/whole tumor volume ratio indicates better prognosis.

IMPORTANCE OF STUDY

Previous studies on pediatric DMG prognostication relied on manual tumor segmentation, which is time-consuming and has high inter-operator variability. There is a great need for non-invasive prognostic imaging tools that can be universally used. Such tools should be automatic, objective, and easy to use in multi-institutional clinical trials. We developed a fully automatic imaging tool to segment subregions of DMG and select radiomic features to predict patient overall survival (OS). Our acquired 4 sequences of MRI for each patient, at both diagnostic and post-radiation therapy from 2 institutions, were more comprehensive than previous studies. The proposed method achieved high accuracy in DMG segmentation and survival prediction, especially for patients having short OS. The proposed method will be the foundation of increasing the utility of MRI as a tool for predicting clinical outcome, stratifying patients into risk-groups for improved therapeutic management and monitoring therapeutic response with greater sensitivity and an opportunity to adapt treatment.

Introduction

Diffuse midline gliomas (DMG), including diffuse intrinsic pontine gliomas (DIPG), are aggressive central nervous system (CNS) pediatric tumors located in the brainstem, thalamus, spinal cord and cerebellum.[1] As one of the most devastating pediatric cancers, DMG represents about 10–15% of all pediatric CNS tumors, with an estimated 300 new cases diagnosed annually in the USA.[2] Most DMGs occur between the ages of 5 and 10 years, with a peak at 7 years.[3] There is currently no curative therapy for DMG and radiation therapy (RT) remains the standard treatment with only transitory benefits.[4] Despite numerous clinical trials of new agents and novel therapeutic approaches over the last few decades,[5] disease outcomes remain dismal with a median overall survival (OS) of less than 1 year, a 2-year OS rate of less than 10%,[6] and a 5-year OS rate of less than 1%.[7]

Magnetic resonance imaging (MRI) is the standard noninvasive tool for DMG diagnosis and monitoring of tumor response to therapy. For DIPG, typical MRI findings include a T1-hypointense and T2-hyperintense lesion involving greater than 50% of the pons.[8] MRI features have been used to predict H3K27M mutation status[9] and correlate with patient prognosis.[10-15] However, the features utilized in these studies were either simple without high-dimensional image features[10,11,13-15] or only based on texture analysis.[12] The statistical analysis that most of these studies relied on tend to identify inconsistent and inconclusive biomarkers among different studies and datasets. For example, a study of 357 pediatric DIPG demonstrated that although many MRI features, such as tumor size, enhancement and necrosis etc., were strongly associated with survival on univariable analysis, very few were significantly associated with survival on multivariable analysis.[11] These findings suggest only relying on statistical analysis of conventional MRI findings may not be sufficient to predict OS in DMGs.

Machine learning has been widely used to predict survival or discriminate between certain groups in studies of other brain tumors such as glioblastoma multiforme (GBM) and pediatric low-grade gliomas.[16-19] For DMG, machine learning-based regression models were proposed to correlate with patient prognosis based on extracted MRI radiomic features.[20,21] However, these studies only focused on imaging data from diagnosis, and the tumors were segmented manually which is generally believed to be time-consuming and has high inter-operator variability. Studies have demonstrated that semiautomated DMG volume measurements are more accurate, prognostically-relevant, and consistent than manual measurements.[14,15] In addition to diagnostic scans, it is also important to consider longitudinal data at post-treatment timepoints.[10]

With new therapeutic strategies currently under investigation for DMG, including epigenetic therapy and immunotherapy,[22] there is a great need for non-invasive prognostic imaging tools that can be universally used to accurately identify which patients are at risk for the most rapid deterioration, and thereby assist clinical trial eligibility and therapy planning. Such tools should be automatic, objective, and easy to use in multi-institutional clinical trials. With the vast advancements in deep learning techniques, there has been tremendous success in automatic segmentation of brain tumors from MRI, including adult and[23,24] pediatric brain tumors,[25,26] including our previous work focused on the segmentation of pediatric DMG[27,28]. These advancements have the potential to create a fully automatic, image-based radiomic analysis and DMG prognostic tool.

In this work, we developed a novel imaging tool to process and analyze DMG patient's MRI data with the goal of predicting their 1-year OS. One year is the median OS of our internal cohort and it is also close to the median OS reported on larger DIPG studies (11 months).[11] Therefore, accurate prediction of patient's 1-year OS could have profound impact on the clinical management of DMG. The proposed tool is fully automatic, including multisequence MRI preprocessing, deep learning-based segmentation of subregions of DMG, radiomic feature extraction and selection, and machine learning-based OS prediction. The proposed method was

trained and validated on an internal cohort from Children's National Hospital (CNH) to investigate the accuracy of OS prediction in 1) a baseline study using MR images obtained only at diagnosis, and 2) a post-RT study using MR images obtained at both diagnosis and post-RT. The method was further tested on an external DMG dataset collected from Children's Brain Tumor Network (CBTN).

Materials and Methods

Study Cohort

For this 2-center retrospective study, institutional review board approval was obtained at both participating institutions. Our internal cohort includes 53 pediatric and adolescent patients diagnosed with DMG between 2005-2022 (F=29, M=24) at CNH. The median patient age at diagnosis is 6.5 years with a range of 3.2–25.9 years. The median OS is 12 months with a range of 3.3–132 months from diagnosis (1 patient is still alive).

The external cohort includes 16 pediatric patients diagnosed with DMG between 2005-2022 (F=9, M=7), collected through CBTN from Children's Hospital of Philadelphia (CHOP). The median age at diagnosis is 9.4 years with a range of 3.8–18.2 years. The median OS is 9.6 months with a range of 1.3–27.1 months from diagnosis.

MRI Data

Both institutions used similar scanners and protocols which varied among patients and timepoints because of retrospective data collection. For each patient, 4 MRI sequences at diagnosis and/or post-RT were collected including T1-weighted (T1), contrast-enhanced T1 (T1ce), T2-weighted (T2), and T2-weighted-Fluid-Attenuated Inversion Recovery (T2 FLAIR). The collected MRIs were acquired at 1.5 or 3 T magnet, with 2D or 3D acquisition, using scanners from GE Healthcare, Siemens AG, and Toshiba. T1 and T1ce MRIs included T1 SE, T1 FSE, T1 MPRAGE, and T1 SPGR. T2 MRI included T2 SE, T2 FSE, T2 FRFSE and T2 propeller. T2 FLAIR MRI included

those with and without post-gadolinium. Slice thickness ranges of 0.5–6 mm and matrix ranges of (256–512)×(256–512). All images were collected in the DICOM image format.

Manual segmentation of DMG was used as the ground truth for training our deep learning segmentation model. It was performed under the supervision of 2 expert neurooncologists using ITK-SNAP[29]. Inter-expert variability was resolved through consensus. Because necrosis/cyst is not consistently identifiable for DMG, 2 labels were created: tumor core (TC) and whole tumor (WT). TC includes the Gd-enhancing tumor appeared as enhancement on T1ce MRI, and the necrotic/cystic core appeared as hypointense on T1ce MRI. WT includes TC and the peritumoral edematous/infiltrated tissue which is defined as the abnormal hyperintense signal on the T2 FLAIR MRI.

Automatic DMG Segmentation

Despite the tremendous success of deep learning-based automatic segmentation for adult GBMs, directly using these methods on rare pediatric brain tumors remains challenging[30]. While GBMs and DMGs share several clinical properties, they have distinctive characteristics as well, especially in their location in the brain and radiologic presentation. Our approach was to transfer knowledge learnt from GBM segmentation to DMG segmentation.

The Brain Tumor Segmentation (BraTS) challenge is an ongoing annual event that has been held since 2012. We acquired imaging data of 1,251 GBM patients that was publicly available from the BraTS 2021 challenge.[31] For each patient, 4 MRI sequences (T1, T1ce, T2, and T2 FLAIR) and manual segmentations of subregions of GBM were provided.

The winning method of the BraTS 2020 challenge was nnU-Net[24],a popular and robust semantic deep-learning segmentation method. It analyzes the training data and automatically configures a matching U-Net[32]-based segmentation pipeline. Figure 1 shows the model architecture of our transfer learning-based approach using nnU-Net. It includes a pretraining phase, which trained nnU-Net using the BraTS 2021 challenge dataset. Because nnU-Net

automatically determines the segmentation pipeline based on the specific dataset, we first had the segmentation pipeline planned based on the DMG dataset, and then used the planned pipeline to prepare the BraTS dataset and perform pretraining. The pretrained network weights were then used as initialization to finetune the model using the DMG dataset, which was preprocessed to be compatible with the BraTS format. Preprocessing was performed in an automatic fashion and included N4 bias correction[33], rigid registration to the SRI-24 Atlas[34], and skull stripping[35]. The output of the model was the predicted TC and WT volumes, which were used as input to the radiomic feature extraction step.

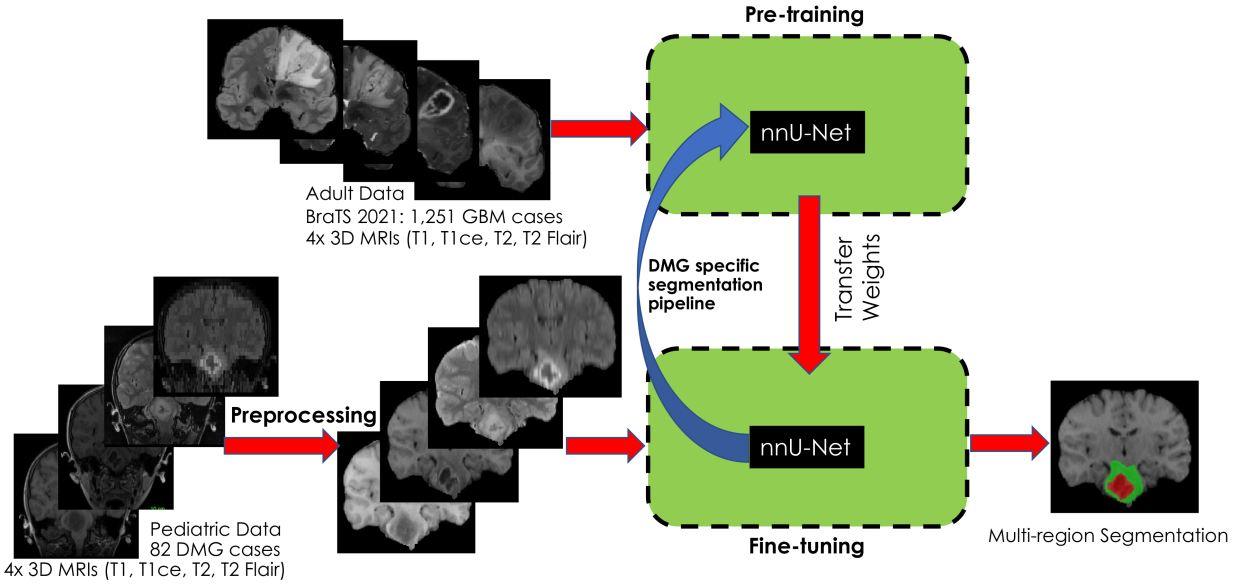

Figure 1. Model architecture of our DMG segmentation method.

Many DMG cases do not have or have very small TC volumes. While small TC volumes are unlikely to affect OS prediction, they may cause misleading segmentation evaluation. We therefore imposed a postprocessing step on the predicted TC volumes before calculating the evaluation metrics. TC/WT denoted the ratio of TC volume over WT volume. The predicted volumes were first cleaned by removing small (i.e., <130 voxels) disconnected regions, followed

by setting the TC volume to be 0 if TC/WT<4%. This postprocessing was determined by optimization on the BraTS-PEDs 2023 dataset.[36,37]

Experiments and Evaluation for Tumor Segmentation

45/53 CNH patients with manual segmentations were used for training and validation of the segmentation model. Scans at diagnosis and post-RT of the same patient were counted for the purpose of segmentation. This yielded a total of 82 cases from the 45 patients. Specifically, 41/82 scans were acquired at diagnosis, 34/82 scans were acquired within 1-month post-RT, and the rest of 7 scans were acquired 2–4 months post-RT.

The 82 DMG cases were randomly divided into 5 splits, and 5-fold cross-validation was performed to obtain the predicted TC and WT volumes. Dice coefficient and volume similarity (VS) were used as evaluation metrics to compare the predicted and ground truth segmentations. VS is defined as VS=1-VD, where VD (volume distance) is calculated as the absolute volume difference divided by the sum of the compared volumes[38]. Comparison between predicted and ground truth in small or absent TC volumes tend to produce extreme metrics (e.g., Dice score=0 or 1). To void bias to small volumes, we did not evaluate segmentation performance if 0<TC/WT<4% for both predicted and ground truth segmentations. If TC/WT=0 for both, the metrics were set to be 1.

After 5-fold cross-validation, we trained a final model with all 82 cases and used it to predict TC and WT volumes for the rest 8/53 internal patients without manual segmentations and 16 external patients, of which 14 with manual segmentations were used in the external test set.

Radiomic Feature Extraction

Based on automatically segmented DMG volumes, we used the open-source PyRadiomics software[39] to extract radiomic features including 13 shape features and 91 gray level features. Please refer to Supplemental Appendix S1 for a complete list of features. The gray level features

included: 18 first order features, 22 gray level co-occurrence matrix (GLCM) features, 16 gray level size zone matrix (GLSZM) features, 16 gray level run length matrix (GLRLM) features, 5 neighboring gray tone difference matrix (NGTDM) features, and 14 gray level dependence matrix (GLDM) features. In addition, we added 2 clinical features (i.e., sex and age), and 2 shape features of interest: brain volume and relative tumor volume (DMG volume divided by the brain volume). Because gray level features are susceptible to inter-scanner variation due to different acquisition protocol[40], image gray levels were normalized by removing the mean and scaling to unit variance before the features were calculated.

The baseline study employed 401 features including 37 shape features and 4 sets of 91 gray level features (1 set for each MRI sequence). The shape features included sex, age, brain volume, 14 shape features (i.e., 13 from PyRadiomics and relative DMG volume) for WT, 10 shape features for TC, and 10 shape features for the ratio between TC and WT (TC/WT). Because many DMG cases did not have TC volume, 4 features (elongation, flatness, surface area to volume ratio, and sphericity) having measurements of TC in the denominator of their calculation were excluded, because their definitions were not valid with 0 volume. The gray level features were calculated based on WT volumes.

The post-RT study employed 1,576 features including 120 shape features and 1,456 gray level features. The shape features included sex, age, skull-stripped brain volumes at diagnosis and post-RT, 28 WT shape features (14 at diagnosis and 14 post-RT), changes of 14 WT shape features (post-RT values minus values at diagnosis), relative changes of 14 WT shape features (changes divided by values at diagnosis), 20 TC shape features (10 at diagnosis and 10 post-RT), changes of 10 TC shape features, 20 TC/WT shape features (10 at diagnosis and 10 post-RT), and changes of 10 TC/WT shape features. We did not consider relative changes of TC and TC/WT features because measurements related to TC at diagnosis could be 0, which would make the definition of relative change invalid. The gray level features included 4 sets of 91 gray level

features at diagnosis, 4 sets of 91 gray level features post-RT, changes of 4 sets of 91 gray level features, and relative changes of 4 sets of 91 gray level features.

Feature Selection

On the training data, feature selection was performed prior to prediction to avoid overfitting. In the first step, feature filtering was performed using the Mann-Whitney U test comparing feature values between short OS (<365 days) and long OS (≥365 days). 69 features with $p<0.05$ were selected for the post-RT study. For the baseline study, because there was only 1 feature with $p<0.05$, we selected 40 (i.e., 10% of 401) features with the smallest p-values.

As a common requirement for many machine learning models, the selected feature values in the previous step were standardized by removing the mean and scaling to unit variance. Sequential feature selection (SFS) was then performed to select the optimal number of discriminative features for each study. Let n be the desired number of features. The algorithm added 1 feature at an iteration to form a feature subset in a greedy fashion until n was reached. At each iteration, the algorithm went through each feature not currently in the feature subset and chose the feature to add such that the new feature subset achieved the best accuracy in the leave-one-out cross-validation. For leave-one-out cross-validation, we trained a linear support vector machine (SVM) to classify between short OS and long OS using all subjects in our internal cohort except for 1, which was used for testing. This process was repeated iteratively until all patients were tested. We employed the linear kernel for the SVM model because it is less prone to overfitting than non-linear kernels for a small dataset. The number of selected features was limited to less than 10% of the number of patients to avoid overfitting the model to the training data.

Experiments and Evaluation for OS Prediction

Images at diagnosis of 52/53 CNH patients were used for training and validation in the baseline study. 26/52 patients had long OS, i.e., survival greater than 1 year from diagnosis. One patient did not have images of all 4 MRI sequences at diagnosis, but the post-RT images were used for training the segmentation model. Images at diagnosis and within 3 months post-RT of 41/52 patients were used for training and validation in the post-RT study. 19/41 patients had long OS.

After feature selection using leave-one-out cross-validation, the final SVM model was trained with all internal patients with the selected features. Validation of the final model on the internal dataset was reported. The final model was used to predict OS based on the same selected features on the external dataset. For the baseline study, 16 external patients (7 had long OS) were tested. 9/16 external patients who had post-RT imaging (<3 months) were tested in the post-RT study. 5/9 patients had long OS.

Results

Segmentation Results

Table 1 shows performance of the proposed automatic DMG segmentation method evaluated on the internal and external datasets. Machine learning-based brain tumor segmentation algorithms need to be evaluated on out-of-distribution data to access generalizability, reflective of tumor heterogeneity[41]. Metrics of WT segmentation for the external cohort (0.86 mean Dice score and 0.91 mean volume similarity) were similarly well as those obtained for the internal cohort (0.88 mean Dice score [Student's t-test p=0.44] and 0.93 mean volume similarity [p=0.28]). This suggests our method can be successfully generalized for segmenting WT volume of images from outside sources. On the other hand, metrics of TC segmentation for the external cohort (0.74 mean Dice score and 0.81 mean volume similarity) were less accurate than those obtained for the internal cohort (0.91 mean Dice score [p=0.002] and 0.93 mean volume similarity [p=0.006]). It is worth mentioning the median Dice score (0.83) and volume similarity (0.99) of TC segmentation for the external cohort were improved from the mean. This indicates the model

performed well on TC segmentation for most external cases (12/14). We noticed that 2 cases generating poor metrics (Dice<0.2) showed significant under-segmentation of TC volumes. Figure 2 shows qualitative segmentation results on the diagnosis and post-RT images of a DMG patient. The Dice scores for this case were 0.92 (diagnostic TC), 0.92 (diagnostic WT), 0.97 (post-RT TC), and 0.93 (post-RT WT).

Table 1. Mean (median) and standard deviation of Dice coefficient and volume similarity calculated by comparing predicted tumor core (TC) and whole tumor (WT) volumes and those segmented manually. Results shown include validation on the internal cohort (82 cases) and testing on the external cohort (14 cases).

Evaluation dataset	TC Dice	WT Dice	TC vol. similarity	WT vol. similarity
Internal cohort	0.91 (0.94) ± 0.12	0.88 (0.91) ± 0.07	0.94 (0.99) ± 0.10	0.93 (0.96) ± 0.07
External cohort	0.74 (0.83) ± 0.32	0.86 (0.89) ± 0.06	0.81 (0.99) ± 0.34	0.91 (0.93) ± 0.07

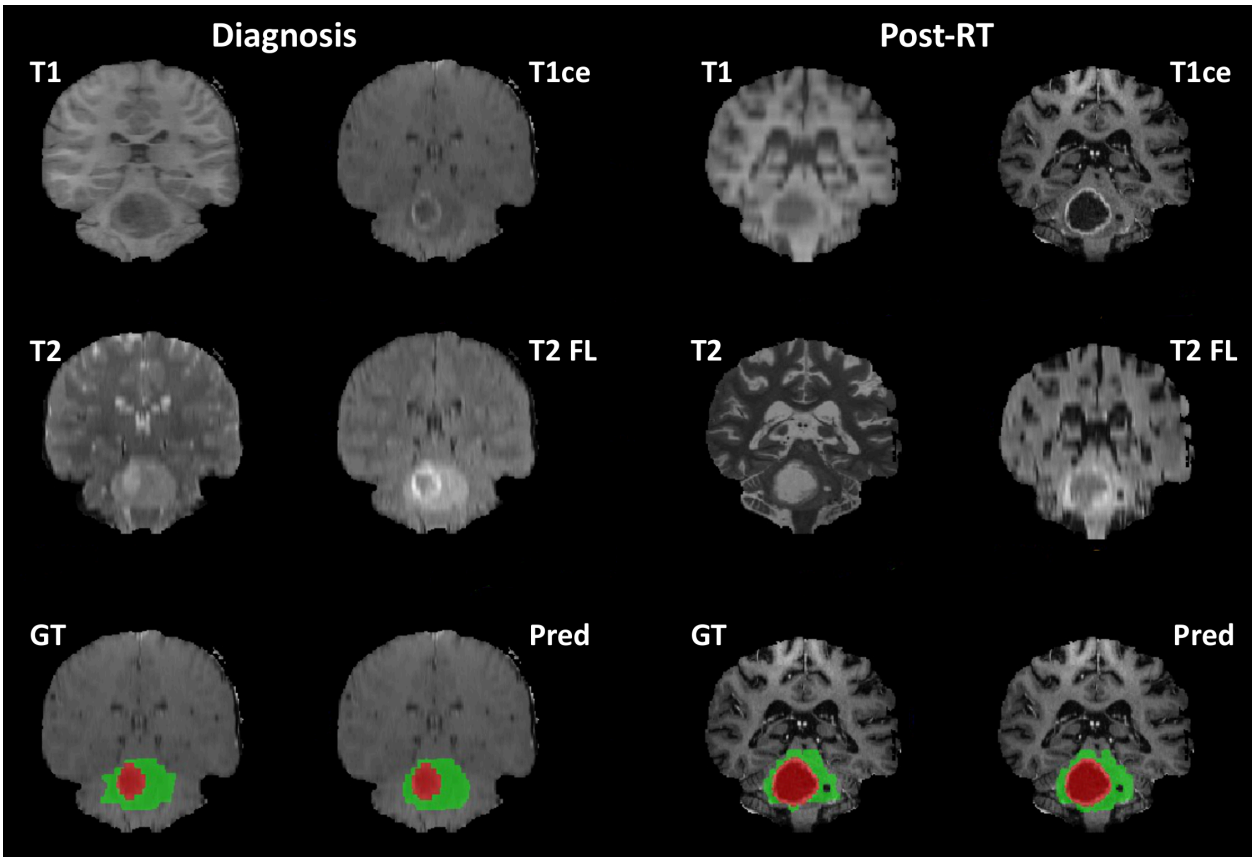

Figure 2. Qualitative segmentation results on the diagnosis and post-RT images of a DMG patient from the internal cohort. The figure shows 4 MRI sequences after preprocessing, the ground truth (GT) segmentation, and the predicted (Pred) segmentation generated by our method (red: TC, red + green: WT).

OS Prediction Results

Table 2 shows results of the proposed OS prediction method. Because identifying patients with higher risk (i.e., OS<365 days) is more important, we adjusted model parameters to maximize accuracy or sensitivity. In general, the results suggest that adding post-RT data improved prediction accuracy and sensitivity over the baseline. The evaluation metrics on our external cohort were comparable to those obtained on the internal cohort, demonstrating good generalization of our machine learning predictive model.

Table 2. Results of the proposed OS prediction method. OS<365 days is considered positive.

	Internal cohort (52 subjects)			External cohort (16 subjects)		
Study	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Baseline max accuracy	77%	81%	73%	81%	89%	71%
Baseline max sensitivity	62%	92%	31%	75%	100%	43%
	Internal cohort (41 subjects)			External cohort (9 subjects)		
Post-RT max accuracy (also max sensitivity)	85%	100%	68%	78%	100%	60%

The number of selected features for the baseline and post-RT studies is 5 and 4, respectively. We list below the selected features for each study, along with their p-values of Mann-Whitney *U* test between short and long OS computed on our internal cohort. The features are listed in the order of their relevance to OS prediction.

Selected 5 features for the baseline study:

- Information measure of correlation (Imc1) on T2 FLAIR (p=0.118): quantifies the complexity of the texture and is related to GLCM.
- High gray level zone emphasis on T1 (p=0.231): measures the distribution of the higher gray level values and is related to GLSZM.
- The median gray level value on T2 FLAIR (p=0.173)
- Skewness on T2 (p=0.061): measures the asymmetry of the distribution of gray level values about the mean value.
- The 10th percentile of gray level value on T2 FLAIR (p=0.217)

Selected 4 features for the post-RT study:

- The ratio of maximum 2D diameter (coronal plane) between post-RT TC and post-RT WT (p=0.017). The maximum 2D diameter is the largest pairwise Euclidean distance between tumor surface mesh vertices on a 2D plane.

- The 10th percentile of gray level value on post-RT T1ce (p=0.027).
- The ratio of minor axis length between post-RT TC and post-RT WT (p=0.002). The minor axis length is the second-largest axis length of principal component analysis performed on the volume.
- Root mean squared on post-RT T1ce (p=0.006): is the square-root of the mean of all the squared gray level values.

Figure 3 shows the comparison between short OS and long OS for the selected features of the 2 studies. A visual example of radiomics is shown in Fig. 4.

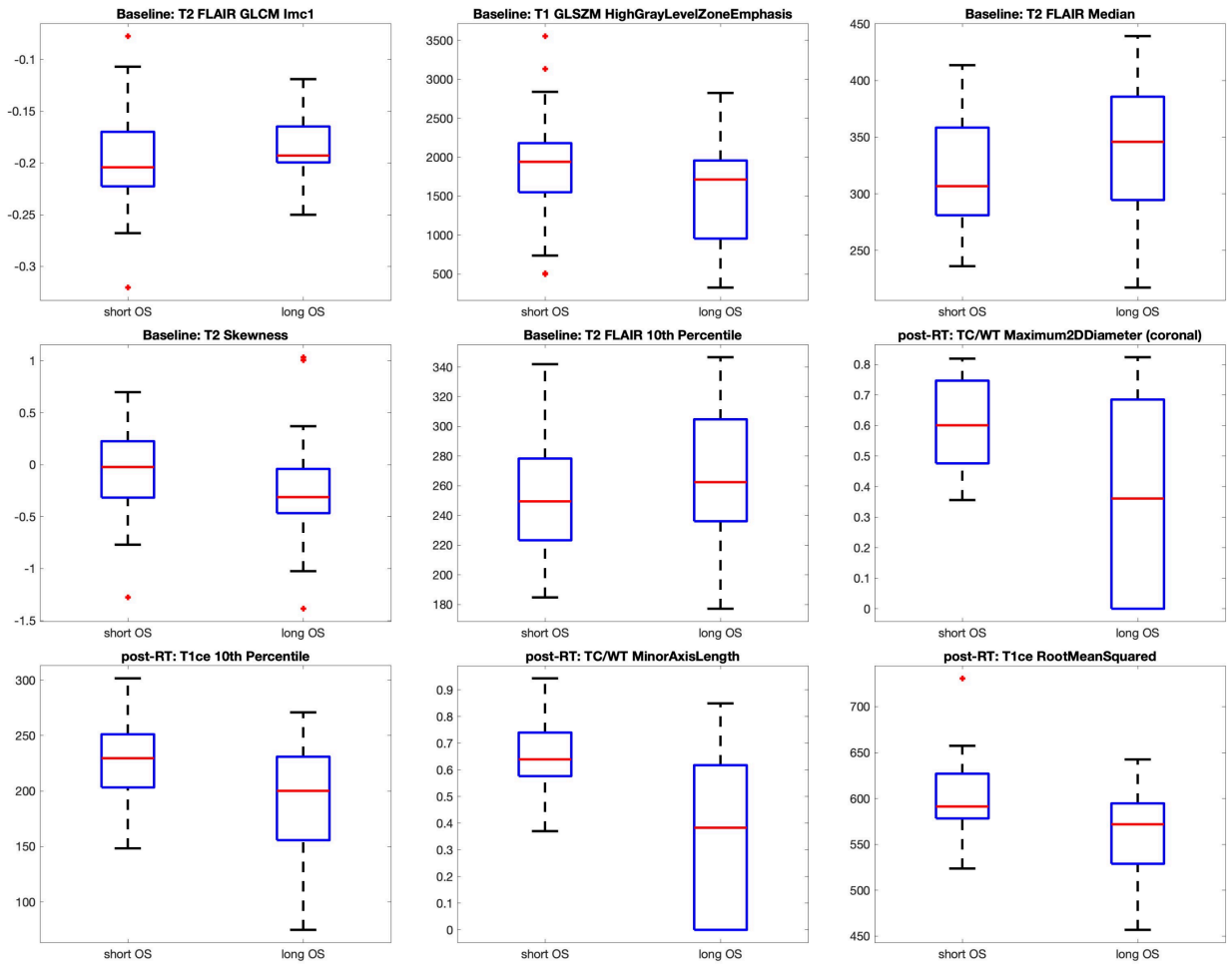

Figure 3. Comparison between short OS and long OS for the selected features of the baseline and the post-RT studies. Data of both internal and external cohorts were considered.

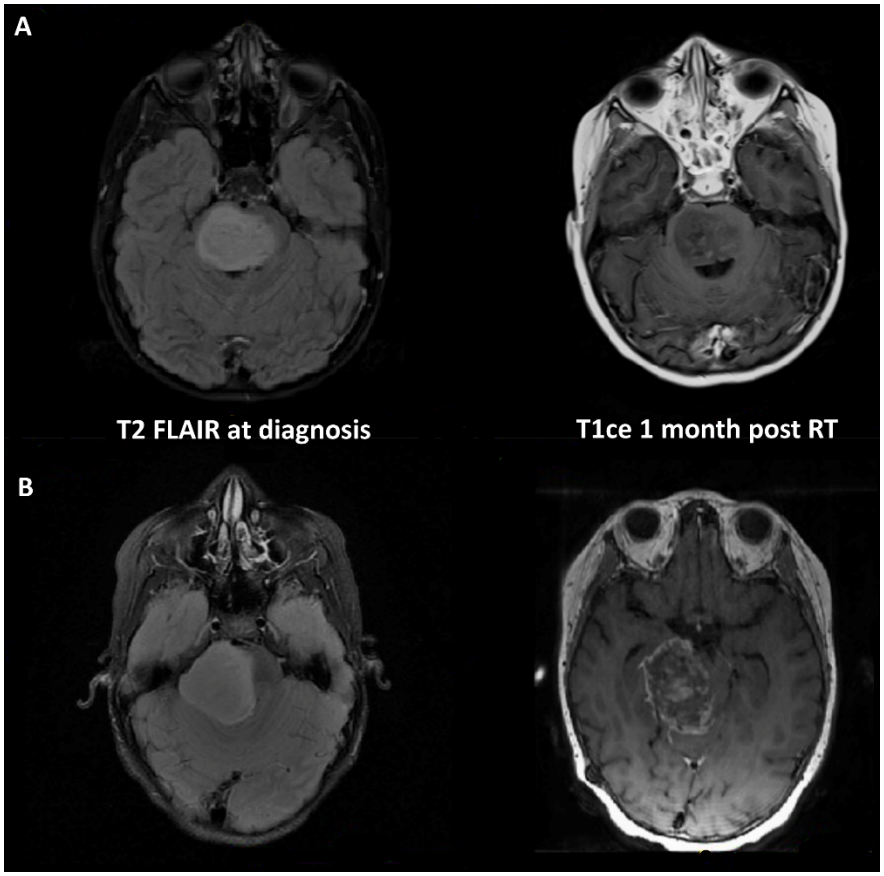

Figure 4. MRI of patients who survived 21 months (A) and 3 months (B) from our internal cohort. Diagnostic T2 FLAIR shows there is more heterogeneous intensity distribution in WT of patient A than B. Post-RT T1ce shows the TC/WT ratio of patient A is smaller than that of patient B.

Discussion

To our best knowledge, this study is first in reporting a fully automatic, machine learning-based classification model to prognosticate DMG survival using MRI features. Our automatic DMG segmentation method generated accurate TC and WT segmentations. The mean Dice scores of 0.91 for TC and 0.88 for WT of cross-validation on the internal cohort were comparable to those reported for adult GBM segmentation using state-of-the-art deep learning models.[42,43] Although worse than the internal cohort, TC segmentation for the external cohort (mean Dice=0.74) is still

comparable to the 0.62–0.74 Dice scores reported in a recent study of automatic segmentation of subregions of pediatric brain tumors[26].

Based on manual segmentation, a recent study presented a machine learning-based regression model to correlate MRI radiomic features with DIPG prognosis.[20] The study employed T1ce and T2 MRI acquired at diagnosis, and found heterogeneous tumor pixel intensity or texture, such as the GLCM features, conferred a better prognosis. A similar pattern was found in our baseline study, where *GLCM Imc1 of WT on T2 FLAIR* was larger for the long OS group compared with the short OS group, although the difference was not significant (Fig. 3).

While diagnostic features were considered in the post-RT study, all the selected features in the post-RT study were related to post-RT measurements, and they were more discriminative in terms of statistical test ($p<0.05$) compared with those for the baseline ($p>0.05$). Shape features which are independent of scanner variation, were selected for the post-RT study whereas no shape feature was selected for the baseline. These results suggest post-RT features may be more discriminative and reliable compared with diagnostic features. This is verified by the improved prediction accuracy for our internal cohort, although results for the external cohort are comparable. Radiomic analysis allowed us to calculate complex shape features, such as the 2 selected ones in the post-RT study, which correlated to post-RT TC volume but is more discriminative as it was identified by our feature selection method. Based on our results, smaller or non-existent post-RT TC/WT ratio indicates longer OS. For both baseline and post-RT studies, our method produced high sensitivity and low specificity for both internal and external cohorts. It indicates the model predicts with high accuracy which patients have short OS.

Our study is not without limitations. Both of our internal and external cohorts represent small datasets, especially for the post-RT studies. The findings of this study need to be further verified with a larger DMG dataset. Better DMG segmentation and OS prediction models may be achieved by training on larger data and the fully automatic nature of the proposed method is well suited for multi-institutional collaboration. Radiomics are susceptible to bias and variation due to

numerous inter-scanner factors such as different acquisition protocols. Additional feature harmonization methods besides what was performed in our study could be used to remove scanner effects in brain MRI radiomic features.[40,44]

Conclusions

In this multi-institutional study, we demonstrated that a fully automatic approach to compute imaging biomarkers of diffuse midline gliomas from multisequence MRI can accurately and non-invasively predict overall survival for impacted pediatric patients. The proposed method can be used as the foundation of increasing the utility of MRI as a tool for predicting clinical outcome, stratifying patients into risk-groups for improved therapeutic management and monitoring therapeutic response with greater sensitivity and an opportunity to adapt treatment.

AKNOWLEDGMENTS

This work was supported by the National Institutes of Health under grant No. 5UH3CA236536-04.

References:

1. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-Oncology.* 2021;23:1231-1251.

2. Warren KE. Diffuse intrinsic pontine glioma: poised for progress. *Front Oncol*. 2012;2:205.

3. Di Ruscio V, Del Baldo G, Fabozzi F, et al. Pediatric Diffuse Midline Gliomas: an unfinished puzzle. *Diagnostics (Basel).* 2022;12(9):2064.

4. Hoffman LM, DeWire M, Ryall S, et al. Spatial genomic heterogeneity in diffuse intrinsic pontine and midline high-grade glioma: implications for diagnostic biopsy and targeted therapeutics. *Acta Neuropathol Commun.* 2016;4:1.

5 Espirito Santo V, Passos J, Nzwalo H, et al. Remission of pediatric diffuse intrinsic pontine glioma: case report and review of literature. *J Pediatric Neurosci*. 2021;16:1-4.

6. Rashed WM, Maher E, Adel M, et al. Pediatric diffuse intrinsic pontine glioma: where do we stand? *Cancer Matastasis Rev.* 2019;38(4):759-770.

7. Hayden E, Holliday H, Lehmann R, et al. Therapeutic targets in diffuse midline gliomas – an emerging landscape. *Cancers (Basel).* 2021;13(24):6251.

8. Barkovich AJ, Krischer J, Kun LE, et al. Brain stem gliomas: a classification system based on magnetic resonance imaging. *Pediatr Neurosurg.* 1990;16(2):73-83.

9. Chauhan RS, Kulanthaivelu K, Kathrani N, et al. Prediction of H3K27M mutaiton status of diffuse midline gliomas using MRI features. *J Neuroimaging.* 2021;31:1201-1210.

10. Löbel U, Hwang S, Edwards A, et al. Discrepant longitudinal volumetric and metabolic evolution of diffuse intrinsic pontine gliomas during treatment: implicaitons for current response assessment strategies. *Neuroradiology*. 2016;58(10):1027-1034.
11. Leach JL, Roebker J, Schafer A, et al. MR imaging features of diffuse intrinsic pontine glioma and relationship to overall survival: report from the International DIPG Registry. *Neuro Oncol*. 2020;22(11):1647-1657.
12. Szychot E, Youssef A, Ganeshan B, et al. Predicting outcome in childhood diffuse midline gliomas using magnetic resonance imaging based texture analysis. *Journal of Neuroradiology*. 2021;48(4):243-247.
13. Zhu X, Lazow MA, Schafer A, et al. A pilot radiogenomic study of DIPG reveals distinct subgroups with unique clinical trajectories and therapeutic targets. *Acta Neuropathol Commun*. 2021;9:14.
14. Gilligan LA, DeWire-Schottmiller MD, Fouladi M, et al. Tumor response assessment in diffuse intrinsic pontine glioma: comarison of semiautomated volumetric, semiautomated linear, and manual linear tumor measurement strategies. *Clinical Trial*. 2020;41(5):866-873.
15 Lazow MA, Nievelstein MT, Lane A, et al. Volumetric endpoints in diffuse intrinsic pontine glioma: comparison to cross-sectional measures and outcome correlations in the International DIPG/DMG Registry. *Neuro Nocol*. 2022;24(9):1598-1608.
16 Chang K, Zhang B, Guo X, et al. Multimodel imaging patterns predict survival in recurrent globlastoma patients treated with bevacizumab. *Neuro Oncol*. 2016;18(12):1680-1687.
17 Wagner MW, Hainc N, Khalvati F, et al. Radiomics of pediatric low-grade gliomas: toward a pretherapeutic differentiation of BRAF-mutated and BRAF-fused tumors. *AJNR Am J Neuroradiol*. 2021;42(4):759-765.
18 Li G, Li L, Li Y, et al. An MRI radiomics approach to predict survival and tumour-infiltrating macrophages in gliomas. *Brain*. 2022;145(3):1151-1161.
19 Moassefi M, Faghani S, Conte GM, et al. A deep learning model for discriminating true progression from pseudoprogression in glioblastoma patients. *J Neurooncol*. 2022;159(2):447-455.
20 Tam LT, Yeom KW, Wright JN, et al. MRI-based radiomics for prognosis of pediatric diffuse intrinsic pontine glioma: an international study. *Neurooncol Adv*. 2021;3(1):vdab042.
21 Wagner MW, Namdar K, Napoleone M, et al. Radiomic features based on MRI predict progression-free survival in pediatric diffuse midline glioma/diffuse intrinsic pontine glioma. *Canadian Association of Radiologists Journal*. 2023;74(1):119-126.
22. Long W, Yi Y, Chen S, et al. Potential new therapies for pediatric diffuse intrinsic pontine glioma. *Front Pharmacol*. 2017;8:495.
23. Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. *Proceedings of International MICCAI Brainlesion Workshop*. 2018;311-320.
24. Isensee F, Jaeger PF, Kohl SA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*. 2020;1-9.
25. Madhogarhia R, Kazerooni AF, Arif S, et al. Automated segmentation of pediatric brain tumors based on multi-parametric MRI and deep learning. *Proceedings of SPIE Medical Imaging*. 2022;120332R.
26. Kazerooni AF, Arif S, Madhogarhia R, et al. Automated tumor segmentation and brain tissue extraction from multiparametric MRI of pediatric brain tumors: A multi-institutional study. *Neurooncol Adv*. 2023;5(1):1-12.
27. Liu X, Bonner ER, Jiang Z, et al. From adult to pediatric: deep learning-based automatic segmentation of rare pediatric brain tumors. *Proceedings of SPIE Medical Imaging*. 2023; 1246505.
28. Liu X, Bonner ER, Jiang Z, et al. Automatic segmentation of rare pediatric brain tumors using knowledge transfer from adult data. *Proceedings of IEEE International Symposium on Biomedical Imaging*. 2023; In press.
29. Paul AY, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*. 2006;31(3):1116-1128.
30 Drai M, Testud B, Brun G, et al. Borrowing strength from adults: Transferability of AI algorithms for paediatric brain and tumour segmentation. *Eur J Radiol*. 2022;151:110291.

31 Baid U, Ghodasara S, Mohan S, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv:2107.02314*. 2021.
32 Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. *Proceedings of MICCAI*. 2015;234-241.
33 Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. 2010;29(6):1310-1320.
34 Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A. The SRI24 multichannel atlas of normal adult human brain structure. *Hum Brain Mapp*. 2010;31(5):798-819.
35 Thakur S, Doshi J, Pati S, et al. Brain extraction on MRI scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnositc training. *Neuroimage*. 2020;220:117081.
36 Kazerooni AF, Khalili N, Liu X, et al. The brain tumor segmentation (BRATS) challenge 2023: Focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). *arXiv:2305.17033*. 2023.
37 Capellan-Martin D, Jiang Z, Parida A, et al. Model ensemble for brain tumor segmentation in magnetic resonance imaging. *International MICCAI Brainlesion Workshop.* In press.
38 Cardenes R, Luis-Garcia R, Bach-Cuadra M. A multidimensional segmentation evaluation for medical image data. *Comput Methods Prog Biomed*. 2009;96(2):108-124.
39 van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104-e107.
40 Li Y, Ammari S, Balleyguier C, et al. Impact of preprocessing and harmonization methods on the removal of scanner effects in brain MRI radiomic features. *Cancers (Basel)*. 2021;13(12):3000.
41 Prabhudesai S, Wang NC, Ahluwalia V, et al. Strafification by tumor grade groups in a holistic evaluation of machine learning for brain tumor segmentation. *Front Neurosci*. 2021;15:740353.
42 Isensee F, Jaeger PF, Full PM, et al. nnU-Net for brain tumor segmentation. *International MICCAI Brainlesion Workshop*. 2020;118-132.
43 Aboian M, Bousabarah K, Kazarian E, et al. Clinical implementation of artificial intelligence in neuroradiology with development of a novel workflow-efficient picture archiving and communication system-based automated brain tumor segmentation and radiommic feature extraction. *Front Neurosci*. 2022;16:860208.
44 Stamoulou E, Spanakis C, Manikis GC, et al. Harmonization strategies in multicenter MRI-based radiomics. *J Imaging*. 2022;8(11):303.