Neuro-Oncology

24(2), 289–299, 2022 | https://doi.org/10.1093/neuonc/noab151 | Advance Access date 26 June 2021

Deep learning-based automatic tumor burden assessment of pediatric high-grade gliomas, medulloblastomas, and other leptomeningeal seeding tumors

Jian Peng,[†] Daniel D. Kim,[†] Jay B. Patel, Xiaowei Zeng, Jiaer Huang, Ken Chang, Xinping Xun, Chen Zhang, John Sollee, Jing Wu, Deepa J. Dalal, Xue Feng, Hao Zhou, Chengzhang Zhu, Beiji Zou, Ke Jin, Patrick Y. Wen, Jerrold L. Boxerman, Katherine E. Warren, Tina Y. Poussaint, Lisa J. States, Jayashree Kalpathy-Cramer, Li Yang, Raymond Y. Huang, and Harrison X. Bai

Department of Neurology, The Second Xiangya Hospital of Central South University, Changsha, Hunan, China (J.P., X.Z., X.X., C.Z., L.Y.); Department of Diagnostic Imaging, Rhode Island Hospital and Alpert Medical School of Brown University, Providence, Rhode Island, USA (D.D.K., J.S., J.L.B., H.X.B.); Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA (R.Y.H.); Department of Radiology, Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA (J.B.P., K.C., J.K.-C.); Department of Radiology, The Second Xiangya Hospital of Central South University, Changsha, Hunan, China (J.W.); Center for Neuro-Oncology, Dana Farber Cancer Institute, Boston, Massachusetts, USA (P.Y.W.); Department of Pediatrics, Dana Farber Cancer Institute, Boston, Massachusetts, USA (K.E.W.); Department of Radiology, Boston Children's Hospital, Boston, Massachusetts, USA (T.Y.P.); Department of Radiology, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA (D.J.D., L.J.S.); Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia, USA (X.F.); School of Computer Science and Engineering, Central South University, Changsha, Hunan, China (J.H., C.Z., B.Z.); Department of Radiology, Hunan Children's Hospital, Changsha, Hunan, China (K.J.); Department of Neurology, Xiangya Hospital of Central South University, Changsha, Hunan, China (H.Z.)

[†]These authors contributed equally to this work.

Corresponding Author: Li Yang, MD, Department of Neurology, The Second Xiangya Hospital of Central South University, No. 139 Middle Renmin Road, Changsha, Hunan 410011, China (yangli762@csu.edu.cn).

Abstract

Background. Longitudinal measurement of tumor burden with magnetic resonance imaging (MRI) is an essential component of response assessment in pediatric brain tumors. We developed a fully automated pipeline for the segmentation of tumors in pediatric high-grade gliomas, medulloblastomas, and leptomeningeal seeding tumors. We further developed an algorithm for automatic 2D and volumetric size measurement of tumors.

Methods. The preoperative and postoperative cohorts were randomly split into training and testing sets in a 4:1 ratio. A 3D U-Net neural network was trained to automatically segment the tumor on T1 contrast-enhanced and T2/FLAIR images. The product of the maximum bidimensional diameters according to the RAPNO (Response Assessment in Pediatric Neuro-Oncology) criteria (AutoRAPNO) was determined. Performance was compared to that of 2 expert human raters who performed assessments independently. Volumetric measurements of predicted and expert segmentations were computationally derived and compared.

Results. A total of 794 preoperative MRIs from 794 patients and 1003 postoperative MRIs from 122 patients were included. There was excellent agreement of volumes between preoperative and postoperative predicted and manual segmentations, with intraclass correlation coefficients (ICCs) of 0.912 and 0.960 for the 2 preoperative and 0.947 and 0.896 for the 2 postoperative models. There was high agreement between AutoRAPNO scores on predicted segmentations and manually calculated scores based on manual segmentations (Rater 2 ICC = 0.909; Rater

3 ICC = 0.851). Lastly, the performance of AutoRAPNO was superior in repeatability to that of human raters for MRIs with multiple lesions.

Conclusions. Our automated deep learning pipeline demonstrates potential utility for response assessment in pediatric brain tumors. The tool should be further validated in prospective studies.

Key Points

1. A deep learning pipeline for automatic pediatric brain tumor segmentation was built.
2. Excellent agreement between predicted and human segmentation volumes (ICC ≥ 0.9).
3. Excellent agreement between manual and automatic 2D measurements (ICC > 0.85).

Importance of the Study

Longitudinal measurement of tumor burden with MRI is essential for response assessment in pediatric tumors of the brain. Current practice requires human experts to manually estimate the size based on bidimensional diameters. This process is subject to intra- and interrater variability, especially when done on patients with multiple complex lesions. In this study, we developed an algorithm based on deep learning that automatically segments tumors on pre- and postoperative MRIs and estimates tumor volume and the product of maximum bidimensional diameters according to the RAPNO criteria (AutoRAPNO). The models demonstrated excellent performance, with high repeatability and agreement with human raters. This tool has been released as open-source so that it may be utilized in clinical trials and routine clinical practice to accurately and efficiently assess treatment response after further validation.

Tumors of the central nervous system (CNS) are the second most common pediatric malignancy, and tumors of the brain are the most common cause of cancer-related deaths in children.[1] Medulloblastomas (MBL), primitive neuroectodermal tumors of cerebellar origin, represent up to 20% of all malignant pediatric brain tumors.[1] The standard therapeutic strategy for MBL typically follows a mixed approach of surgical resection of the primary mass followed by radiation and chemotherapy; complete or near-total primary tumor resection is associated with improved outcomes.[2,3] High-grade gliomas (HGG) are difficult to treat given their biological and histological heterogeneity.[4] HGG comprise approximately 8%-12% of pediatric CNS tumors and are the leading cause of cancer-related death in those younger than 19 years.[4] While the standard of care remains surgical resection with or without adjuvant chemotherapy or radiotherapy, 5-year progression-free survival remains low.[1,5,6] For those with recurrent HGG, overall survival with treatment is 5.6 months.[7] Calculating the size of the postresection residual tumor allows for the most accurate prediction of prognosis.[1,5]

Treatment response and tumor progression are assessed in MBL and other leptomeningeal seeding tumors by taking the sum of the products of perpendicular diameters of up to 4 measurable contrast-enhanced lesions on T1-weighted imaging sustained for at least 4 weeks.[3] The most recent RAPNO (Response Assessment in Pediatric Neuro-Oncology) criteria for pediatric HGG are similar, although a maximum of 3 target lesions is recommended.[4] Magnetic resonance imaging (MRI) for response assessment during clinical trials is performed at a minimum interval of 3 months for MBL and recently diagnosed (<1 year) HGG.[3,4] For relapsing HGG, MRIs may be performed every 2 months.[4] Although the manual method of 2-dimensional tumor size determination is the current standard, several adult, and pediatric studies have shown that this approach demonstrates considerable inter-observer variability.[8–12] Moreover, manual delineation is time-consuming and labor-intensive.[13] While volumetric segmentation has recently been recognized as superior to traditional 2D linear measurements, existing methods are poorly validated.[14]

Building accurate, reproducible, and efficient automated tools to segment tumor volume for MRI-based assessment of treatment response is an essential step in facilitating the use of 3D tumor volume as an endpoint in clinical trials. Given recent advances in computing power, deep learning has emerged as the most promising approach for automating the segmentation of medical images.[15,16] Deep learning is superior to radiomics and other machine learning approaches, which rely on hand-crafted features.[17] The convolutional neural network (CNN), which is the basis of deep learning, can be trained with raw image data to predict defined outputs. Recent work by our group

and colleagues using this approach has succeeded in developing accurate and reproducible volumetric segmentations of pre- and postoperative gliomas in adults.[18–20] Whether a similar approach can be successfully used in pediatric gliomas, MBL, and other pediatric leptomeningeal seeding tumors has not yet been determined.

Materials and Methods

Patient Cohort

We retrospectively collected imaging and clinical data from pediatric patients with intracranial brain tumors for our preoperative cohort from January 2011 to December 2018 and admitted them to 4 large academic hospitals in Hunan Province, China and from January 2005 to December 2019 and admitted them to the Children's Hospital of Philadelphia (CHOP). For our postoperative cohort, similar data were collected for pediatric patients specifically with HGG, MBL, or other leptomeningeal seeding tumors from the 4 hospitals in Hunan. Leptomeningeal seeding tumors were defined as tumor subtypes that frequently metastasize throughout the CNS and included MBL, glioblastomas, anaplastic astrocytomas, embryonal tumors, germ cell tumors, and choroid plexus papillomas in our study. Inclusion criteria for HGG, MBL, and leptomeningeal seeding tumors were only applied to the postoperative cohort as this is the target population for response assessment. Patients were excluded if they were over 18 years of age, had missing pathological reports, incomplete imaging data or sequences, or insufficient clinical follow-up information. The preoperative cohort was defined as the baseline brain MRI scans performed before receiving surgical resection, radiotherapy, chemotherapy, or chemoradiotherapy. The postoperative cohort was defined as the follow-up brain MRI scans performed every 2-3 months after treatment until progression or last follow-up date. The institutional review boards of all involved institutions approved this study, and the requirement for informed consent was waived.

In the preoperative cohort, axial T2-weighted turbo spin echo (T2) and contrast-enhanced axial T1-weighted turbo spin echo (T1ce) sequences were used. For the postoperative cohort, axial FLAIR (fluid-attenuated inversion recovery), T1ce, and nonenhanced axial T1-weighted turbo spin echo (T1) sequences were used. MRI acquisition parameters are shown in Supplementary Figures 1–3.

Manual Tumor Segmentation and RAPNO Measurements

For the preoperative cohort, manual segmentation of T2 hyperintensity was performed by a neuro-oncologist (H.Z., Rater 1) with 9 years of experience, and manual segmentation of contrast-enhancing tumor was performed by another neuro-oncologist (J.P., Rater 2) with 7 years of experience, using the Level Tracing and Threshold tools in 3D Slicer (v.4.10). For the postoperative cohort, segmentation of FLAIR hyperintensity, segmentation of contrast-enhancing tumor, and RAPNO measurements were performed by Rater 2 and repeated independently by a separate neuro-oncologist (X.Z, Rater 3) with 6 years of experience to assess interrater variability. RAPNO measurements were conducted as delineated in the RAPNO criteria.[3]

Deep Learning-Based T2/FLAIR Hyperintensity and Contrast-Enhancing Tumor Segmentation

The preoperative and postoperative cohorts were randomly split into training and testing sets using a 4:1 ratio. About 20% of the training set was used as the validation set. Images were split such that scans from the same patient were in the same cohort. The training set was only used for training the model, while all model metrics were assessed on the testing set.

To train the model, segmentations by Rater 3 were randomly selected over segmentations by Rater 2 to use as the ground truth for loss function minimization. Performance of the model, however, was determined by comparing the model predictions to both rater's segmentations on an independent test set.

Preprocessing, Model Training, and Postprocessing

Preprocessing, model training, and postprocessing were performed using DeepNeuro (v2) with Tensorflow 2.0 backend.[21] Images were preprocessed before training and predicting. Images were first reoriented to right, anterior, inferior (RAI) orientation, resampled to isotropic resolution, and co-registered to the same anatomical template using 3D Slicer (v4.10). BSpline interpolation was used for image resampling, while nearest neighbor was used for ground truth segmentation resampling. Skull stripping was then performed with Robust Brain Extraction (ROBEX)[22] (Supplementary Figure 4) and N4 bias correction was applied with Advanced Normalization Tools (ANTs).[23] Finally, images were normalized.

A 3D U-Net neural network architecture with 5 levels was used (Supplementary Figure 5).[24] Four neural networks were created: (1) T2 sequences as input for preoperative, T2 hyperintensity segmentation, (2) T1ce and T2 sequences as input for preoperative, contrast-enhancing tumor segmentation, (3) FLAIR and T1ce sequences as input for postoperative, FLAIR hyperintensity segmentation, and (4) FLAIR, T1ce, and T1 sequences as input for postoperative, contrast-enhancing tumor segmentation. AdamW optimizer[25] was used for training with an initial learning rate of 0.0001 and initial weight decay of 0.00002, and Cosine Annealing[26] was used as a learning rate scheduler. Models optimized a soft Dice loss function on the validation set:

$$D(p, g) = \frac{2\Sigma_i p_i g_i}{\Sigma_i (p_i + g_i) + \alpha}$$

where D is the dice, ρ is the probability output of the model, g is the ground truth, and α is a constant. Max pooling and rectified linear unit (ReLU) activation function

was applied at each layer. Upsampling was performed with Trilinear interpolation,[27] and normalization was performed using Group Normalization.[28] At every iteration of training, 2 patches of size 128 × 128 × 128 were extracted with a 25% bias to tumor lesions from each input imaging modality and subjected to scaling, rotation, shear, translation, and left/right patch flipping data augmentations to artificially increase the training set and reduce overfitting.[29] During validation and testing, the full image with each respective imaging modality was inputted into the model for prediction. Models returned a probability map connecting each voxel to the probability of region of interest (ROI). A probability cutoff of 0.5 was used for binary ROI labeling. All models were trained using a 16 GB NVIDIA V100 Tensor Core graphical processing unit (GPU).

After prediction, predicted segmentations were resampled back to their original resolutions for comparison with the manual segmentation.

AutoRAPNO Algorithm

To further streamline the pipeline, an algorithm was developed to automatically calculate the product of the 2D diameters given a contrast-enhancing tumor segmentation. First, the algorithm looked for connected lesions and calculated bidimensional measurements for each connected lesion. To do this, the algorithm did an exhaustive search for the largest line segment consisting entirely of tumor and its respective perpendicular (tolerance = ±5 degrees) at each axial slice. Following criteria for measurable lesions, lengths were zeroed if either bidimensional measurements were ≤8 mm (2× slice thickness).[3] Bidimensional measurements were then multiplied, and the largest product was selected as the cross-sectional area for this connected lesion. If more than one connected lesion existed, up to the top 4 products were summed and returned as the RAPNO score.[3]

Statistical Analysis

Model segmentation outputs were evaluated using the Sørensen-Dice coefficient. As the main objective of automatic segmentation is for response assessment, volumes and RAPNO measurements were also evaluated by the Spearman rank correlation coefficient ρ and intraclass correlation coefficient (ICC). Spearman rank correlation coefficient was calculated using the Scipy Stats package from Python 3.7.9.[30,31] A 2-way model assessing agreement and returning single score ICCs from the Interrater Reliability and Agreement (IRR) package in R 4.0.2 was used to generate ICC values.[32–35] With response assessment being correlated to the amount of change in tumor size, the ICC values of the delta volumes and RAPNO scores, defined as the change in amount since the prior visit, were also calculated for the postoperative cohort.

Code Availability

The code used for preprocessing/postprocessing, model training/predicting, and AutoRAPNO are available on https://github.com/naddan27/AutoPNeuro. The 4 pretrained models are also available. Accessed April 5, 2021.

Results

Patient Cohort

T2 hyperintensity segmentation of the preoperative cohort was available on 794 T2 brain MRIs from 794 patients. Eighty-five of these patients were from the 4 hospitals in Hunan, while 709 were from CHOP. T1ce segmentation of the preoperative cohort was available on 683 T1ce brain MRIs from 683 patients (85 Hunan, 598 CHOP) among the above 794 patients. There were 111 patients from CHOP without T1ce MRI sequences. In the cohort for preoperative T1ce segmentation, 39 patients were excluded due to skull stripping failure that removed the image at the ROI, and 6 were excluded due to co-registration failure, for a final T1ce preoperative patient cohort size of 638.

Nineteen FLAIR hyperintensity segmentations of the postoperative cohort were incomplete and therefore excluded. FLAIR hyperintensity segmentation was available on 492 FLAIR brain MRIs from 122 patients. T1ce segmentation of the postoperative cohort was available on 511 T1ce brain MRIs from 122 patients. All data comprising the postoperative cohort were from the 4 hospitals in Hunan. Characteristics of patients from the postoperative cohort are shown in Table 1. The characteristics of patients from the preoperative cohort are shown in Supplementary Tables 1–3.

Deep Learning-Based T2/FLAIR Hyperintensity and Contrast-Enhancing Tumor Segmentation

For T2 hyperintensity segmentation of the preoperative cohort, the mean Dice score of the model was 0.724 (95% confidence interval [CI]: 0.684-0.764), the median Dice score was 0.819 (interquartile range [IQR]: 0.711-0.879), and the volume ICC value was 0.912 ($P < .001$). For contrast-enhancing tumor segmentation of the preoperative cohort, the mean Dice score was 0.724 (95% CI: 0.672-0.775), the median Dice score was 0.843 (IQR: 0.677-0.909), and the volume ICC value was 0.960 ($P < .001$).

For FLAIR hyperintensity segmentation of the postoperative cohort, the volume ICC was 0.947 ($P < .001$). For contrast-enhancing tumor segmentation of the postoperative cohort, the volume ICC was 0.896 ($P < .001$). Comparisons between automatically and manually derived volumes are shown across all 4 models in Figure 1.

Examples of model output are shown in Figure 2A and B. Examples from the postoperative cohort were specifically chosen over those in the preoperative to display as these lesions tend to be more complex in shape and have necrosis in the middle, therefore depicting uncompromised model performance despite such challenges.

Table 1 Study Population Characteristics: Postoperative Cohort

Characteristics	Number (Percentage)
Median age at diagnosis in years: mean (range)	10.6 (0.2-17.9)
Sex	
Male	72 (59%)
Female	50 (41%)
High-grade glioma (HGG) group	43
Glioblastoma	25 (20.5%)
Anaplastic astrocytoma	18 (14.8%)
Non-HGG group	79
Medulloblastoma	30 (24.6%)
Embryonal tumor group	
Atypical teratoid rhabdoid tumor	12 (9.8%)
Pineoblastoma	3 (2.5%)
Primitive neuroectodermal tumor	4 (3.3%)
Germ cell tumor group	
Germinoma	26 (21.3%)
Germ cell tumor	3 (2.5%)
Choroid plexus papilloma	1 (0.7%)
Treatment modalities	
Chemotherapy only	23 (18.9%)
Radiotherapy only	23 (18.9%)
Chemoradiotherapy	76 (62.2%)
Surgical extent	
Biopsy only	41 (33.6%)
Partial resection	52 (42.6%)
Gross total resection	29 (23.8%)
Leptomeningeal seeding or not	
With leptomeningeal seeding	38 (31.2%)
Without leptomeningeal seeding	84 (68.8%)

Interrater Agreement for RAPNO Scores

There was high agreement between manually and automatically calculated RAPNO scores on the postoperative cohort. The ICC value between AutoRAPNO scores on predicted segmentations and AutoRAPNO scores on manual segmentations was greatest at 0.933 ($P < .001$) (Figure 3A). The ICC value between RAPNO scores from Rater 2 and from Rater 3 was 0.909 ($P < .001$) (Figure 3B). The ICC value between AutoRAPNO on the predicted segmentation and Rater 2 was 0.909 ($P < .001$) (Figure 3C), and the ICC value between AutoRAPNO on the predicted segmentation and Rater 3 was 0.851 (Figure 3D). Examples of AutoRAPNO output are shown in Figure 2C.

In the preoperative cohort, the ICC value between AutoRAPNO scores on predicted segmentations and AutoRAPNO scores on manual segmentations was 0.940 ($P < .001$).

Of the 99 samples in the postoperative contrast-enhancing tumor segmentation test set, there were 81, 11, 1, and 6 scans with 1, 2, 3, and 4 or more connected lesions, respectively, in both segmentations performed by Rater 2 and 3. The performance between AutoRAPNO and human raters was compared among scans with multiple connected lesions. For Rater 2, the ICC value was highest comparing manual RAPNO with AutoRAPNO on ground truth summing the top 4 lesions (0.972, $P < .001$) and lowest with AutoRAPNO on ground truth on just the largest lesion (0.795, $P < .001$). For Rater 3, the ICC value was highest comparing manual RAPNO with AutoRAPNO on ground truth on just the largest lesion (0.881, $P < .001$) and lowest with AutoRAPNO on ground truth summing the top 4 lesions (0.633, $P = .021$).

Agreement in Longitudinal Changes in Size

The ICC values between automatically and manually calculated delta FLAIR volumes, contrast-enhancing tumor volumes, and RAPNO scores was 0.870 ($P < .001$), 0.799 ($P < .001$), and 0.795 ($P < .001$), respectively, in the postoperative cohort (Figure 4).

Correlation Between RAPNO Measures and Volume

The Spearman correlation value between manual volumes and manual RAPNO scores and between manual volumes and AutoRAPNO on predicted segmentations was 0.957 ($P < .001$) and 0.853 ($P < .001$), respectively (Figure 5).

Volume Segmentation and AutoRAPNO Speed Performance

For the preoperative cohort, the median times to get volumes and AutoRAPNO scores were 0.151 seconds (IQR: 0.092-0.285) and 27.67 seconds (IQR: 4.74-107.716), respectively, for each scan. For the postoperative cohort, the median times to get volumes and AutoRAPNO scores were 0.054 seconds (IQR: 0.021-0.071) and 29.104 (IQR: 4.873-223.884), respectively. AutoRAPNO score times were positively skewed, with few samples taking over 30 minutes for the preoperative cohort and over 150 minutes for the postoperative cohort. In comparison, median manual RAPNO score calculation time was 56.34 seconds (IQR: 23.04-505.64).

Discussion

In this study, we demonstrate that a fully automated, deep learning-based pipeline can be used to calculate tumor volumes and RAPNO measurements in the brain with high accuracy in pediatric HGG, MBL, and other leptomeningeal seeding tumors. After images were preprocessed, brain-extracted, and normalized, they were used to train a model with a 3D U-Net neural network architecture to automatically segment tumor lesions in pre- and postoperative MRIs. To further increase the clinical utility of the pipeline, an algorithm (AutoRAPNO) was developed to

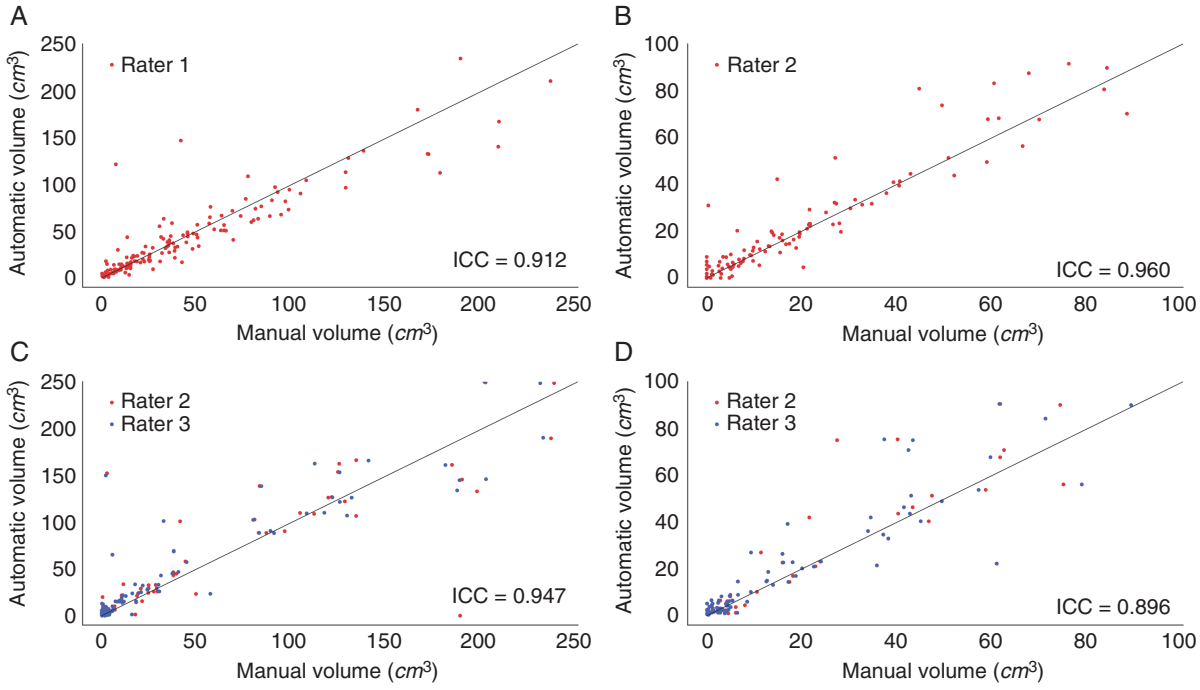

Fig. 1 There is a high correlation between volumes of automatic and manual segmentations. Correlations are shown for the (A) T2 hyperintensity in the preoperative cohort, (B) contrast-enhancing tumor in the preoperative cohort, (C) FLAIR hyperintensity in the postoperative cohort, and (D) contrast-enhancing tumor in the postoperative cohort. Data and ICC values for the test set are shown. Line of identity (y = x) is displayed in all plots. Abbreviations: FLAIR, fluid-attenuated inversion recovery; ICC, intraclass correlation coefficient.

automatically calculate RAPNO scores given a contrast-enhancing tumor segmentation. There was excellent agreement between the automatic pipeline and human raters in both lesion volume and RAPNO scores.

While a recent study developed a model with a 2D ResNeXt deep learning architecture to predict the presence, relative location, and subtype of pediatric posterior fossa tumors, no study has described fully automated segmentation and volume estimation of pediatric brain tumors with deep learning.[17] Given that age-related changes in brain structure are nonlinear and differ substantially between tissue types, it was unknown whether automated deep learning-based segmentation techniques developed and validated in adult subjects would perform well in children. Additionally, there are technical challenges with automated segmentation that are unique to children. These may include increased movement-related artifacts or noise due to lack of child-appropriate equipment, presence of dental braces, heterogeneity in brain size and maturational stage, and lack of age-specific brain atlases for normalization.[36] Despite these challenges, this study has demonstrated that deep learning can be used to accurately segment brain tumors in the pediatric population. Even without removing patients due to artifact, excellent performance was maintained by stripping potential artifacts from the image inputs like the presence of dental braces with the skull stripping brain extraction algorithm (Supplementary Figure 4). Furthermore, models are robust to even difficult segmentations, as shown by the high agreement with human raters of overall metrics despite inclusion into the dataset of tumor subtypes that are typically infiltrative where margins are difficult to define. An example of model prediction on infiltrative tumors is shown in Supplementary Figure 6.

The model also demonstrates high performance in postoperative MRIs, which often have brain distortion and surgical cavities that can complicate segmentation. As depicted in Figure 2A and B, postoperative lesions may be complex in shape with associated necrosis. In fact, the ICC for automatically vs manually derived volumes was slightly higher for the postoperative model for FLAIR hyperintensity segmentation (0.947) than the preoperative model for T2 hyperintensity segmentation (0.912).

There was high agreement between both human vs human RAPNO measures and human vs AutoRAPNO measures. Importantly, the ICC value between AutoRAPNO scores on automatically predicted segmentations and AutoRAPNO scores on manually segmented scans was higher than the ICC value between AutoRAPNO scores on the automatically predicted segmentations and manually calculated scores on manually segmented scans. This goes to show that when response assessment, specifically 2D measurements, is the primary criteria for model metrics, the performance of the model is ostensibly compromised by using the manual RAPNO score calculation, which would include human variability, as the ground truth.

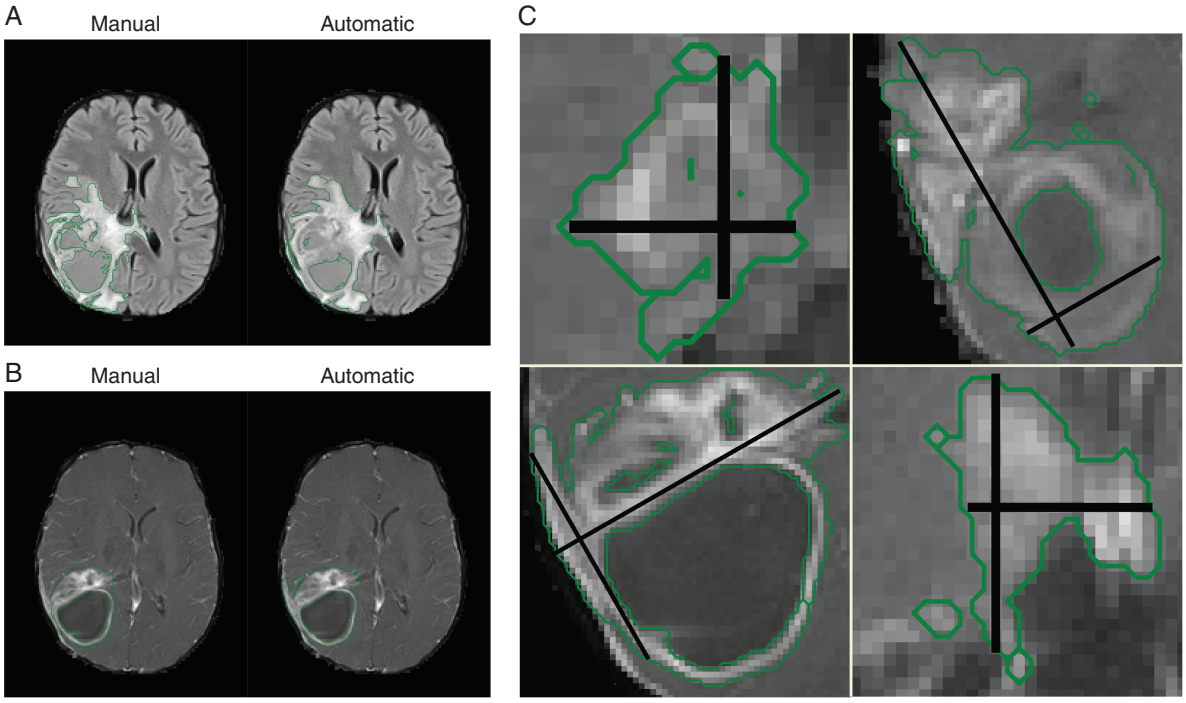

Fig. 2 Examples of model segmentation for (A) T2 FLAIR hyperintensity and (B) contrast-enhancing tumor regions in the postoperative cohort. (C) Examples of AutoRAPNO applied on predicted contrast-enhancing tumor regions in the postoperative cohort. Abbreviation: FLAIR, fluid-attenuated inversion recovery.

When manual segmentation is performed and subsequent RAPNO scores are derived, there exist multiple points at which human inter and intra-rater variability can be introduced. Thus, by utilizing deep learning-based automatic segmentation and AutoRAPNO, measurements are more standardized and highly repeatable.

To test agreement between human raters and AutoRAPNO specifically on MRIs with complex lesions, we identified a subset of scans in the postoperative contrast-enhancing tumor segmentation test set with multiple lesions. When performing manual RAPNO measurements, distinct lesions can lead to errors if they are not appropriately treated as separate. Failing to recognize separate lesions and incorporating them into the RAPNO score can underestimate tumor size and consequently affect response assessment. An important finding of our study is that AutoRAPNO performed better than human raters on MRIs with multiple contrast-enhancing connected lesions. This is evidenced by the fact that the ICC value between scores by Rater 3 and AutoRAPNO on ground truth was greater when only the largest lesion was accounted for (0.881) than when the top 4 lesions were summed (0.633). The lower ICC between scores when the top 4 lesions were considered is attributable to Rater 3 interpreting separate lesions as single lesions. Interestingly, for Rater 2, the highest ICC with AutoRAPNO scores was when 4 lesions were taken into account (0.972); when only the largest lesion was considered, the ICC was 0.795. This therefore highlights the superior repeatability of AutoRAPNO as compared to manually derived scores, which can differ between even highly experienced raters due to varied interpretation of complex lesions on MRI.

Both the volume calculations and AutoRAPNO measurements were achieved efficiently for the pre- and postoperative MRIs. For the preoperative cohort, the median time to volume output was 0.151 seconds, and for the postoperative cohort, the median time to volume output was 0.054 seconds. The median time required for AutoRAPNO to calculate scores was 27.67 seconds for preoperative MRIs and 29.104 seconds for postoperative MRIs. The AutoRAPNO time requirement was related to the complexity of the lesion shape, with a few MRIs taking over 30 minutes for the preoperative cohort and over 150 minutes for the postoperative cohort. This finding further supports the use of volumetric measurements rather than 2D measurements for response assessment. Given that volume measurements can be automatically calculated in less than a second, our algorithm demonstrates potential utility for extremely rapid assessment of tumor burden in clinical practice.

The strong correlation between RAPNO measures and volumes indicates that RAPNO guidelines can be easily adapted for 3D measurements in the future. Current 2D guidelines show excellent consistency in treatment response evaluation for MBL and other leptomeningeal seeding tumors.[37] As the RAPNO criteria uses percentage rather than absolute changes in tumor size,[3] the strong

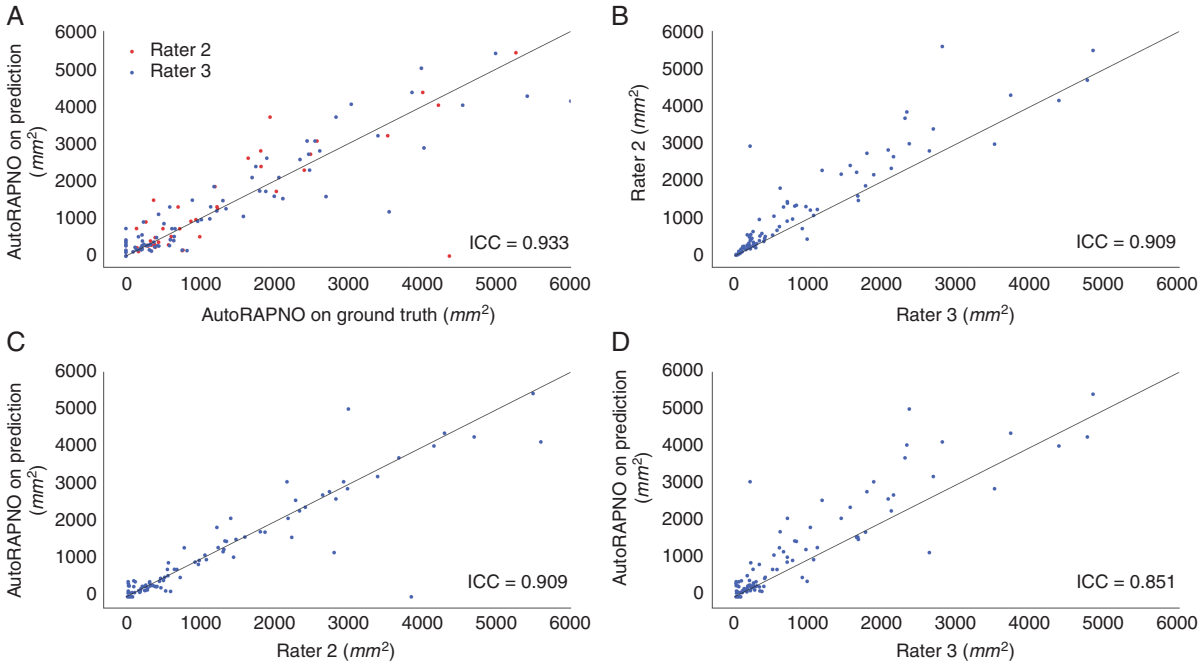

Fig. 3 There is a high correlation between automatically and manually calculated RAPNO scores. The correlation between AutoRAPNO applied on prediction and ground truth labels is greater than that between the automatic RAPNO scores on prediction labels and manually calculated scores. Correlation is shown for the (A) AutoRAPNO on predicted label vs AutoRAPNO on ground-truth label, (B) Rater 2 vs Rater 3, (C) AutoRAPNO on predicted label vs Rater 2, and (D) AutoRAPNO on the predicted label vs Rater 3. Data and ICC values for the test set from the postoperative cohort are shown. Line of identity (y = x) is displayed in all plots. Abbreviations: ICC, intraclass correlation coefficient; RAPNO, Response Assessment in Pediatric Neuro-Oncology.

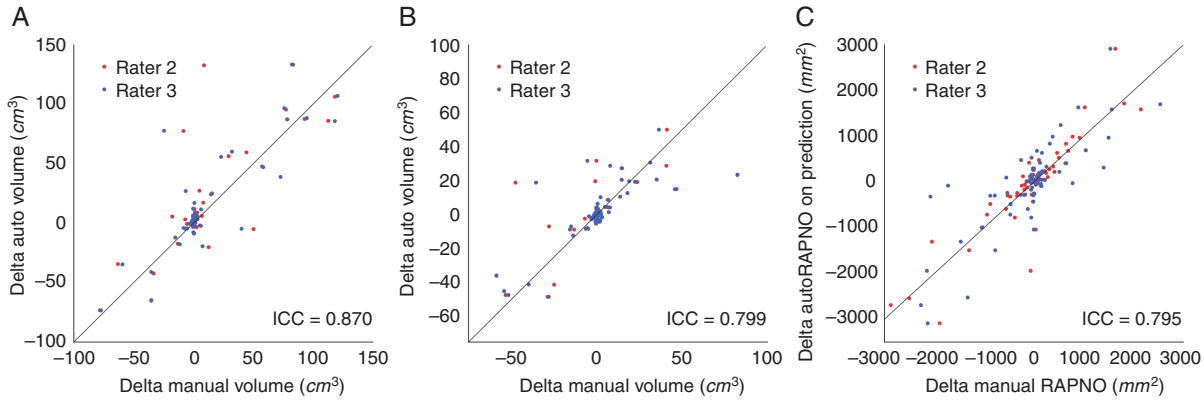

Fig. 4 There was high agreement in longitudinal changes between volumes of automatic and manual FLAIR hyperintensity segmentations. There was moderately high agreement in longitudinal changes between volumes of automatic and manual contrast-enhancing tumor segmentations and between AutoRAPNO on predicted segmentations and manually calculated RAPNO scores. Correlations between delta measurements are shown for (A) FLAIR hyperintensity volumes, (B) contrast-enhancing tumor volumes, and (C) RAPNO scores for the postoperative cohort. Data and ICC values for the test set from the postoperative cohort are shown. Line of identity (y = x) is displayed in all plots. Abbreviations: FLAIR, fluid-attenuated inversion recovery; ICC, intraclass correlation coefficient; RAPNO, Response Assessment in Pediatric Neuro-Oncology.

correlation suggests that volumes can substitute such 2D measurements and still achieve excellent consistency in treatment response evaluation. In smaller clinical trials or institutions with smaller volume of MBL and other leptomeningeal seeding tumor cases, the traditional approach where trained specialists, such as neuro-radiologists or

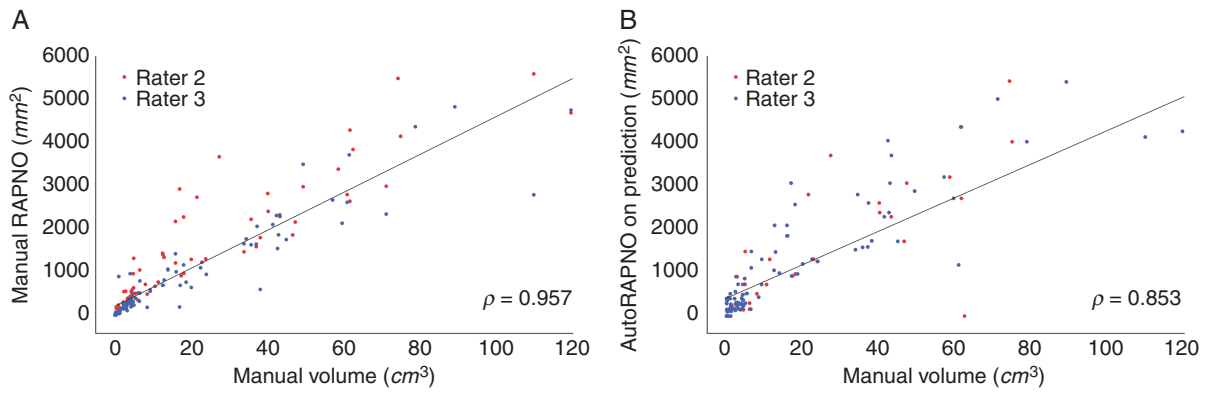

Fig. 5 Automatically and manually calculated RAPNO showed high correlation with volumes on manual segmentations. Correlations between (A) manual RAPNO scores and volumes of manual segmentations and (B) AutoRAPNO on predicted segmentations and volumes of manual segmentations of the postoperative cohort are shown. Data and Spearman ρ coefficients for the test set are shown. Line of identity (y = x) is displayed in all plots. Abbreviation: RAPNO, Response Assessment in Pediatric Neuro-Oncology.

oncologists, take 2D measurements for response assessment may be more ideal than 3D measurements, avoiding the need to segment the whole tumor, establish automatic segmentation pipelines, and allocate GPU/central processing unit (CPU) resources. However, in larger clinical trials and in institutions with limited human resources—either from high case volume or smaller number of trained specialists—the benefits of a deep learning segmentation pipeline outweigh the initial time and capital costs for setup. A deep learning segmentation pipeline is potentially more cost-effective than increasing volume of highly specialized human labor, scalable to demand, and excels in its remarkably fast volumetric calculations and robustness to inter and intra-rater variability. Our code is also contained within a Docker container, and therefore, a trained technician will be able to run the pipeline for inference with a few lines of code. If further training and model development are needed for fine-tuning of the model to an institution's patient demographic, more expertise would be required. As our high-performing deep learning models have already been trained on datasets segmented by specialists, institutions that employ deep learning segmentation pipelines would be able to dedicate fewer human resources for treatment response assessment.

There are some limitations to this study. First, the MBL RAPNO criteria were applied to non-MBL tumors, taking the sum of up to 4 target lesions.[3] For HGG, it is recommended that only up to 3 target lesions should be used.[4] We chose this approach in order to compare the overall performance of the model as a single collective cohort. However, there were only 6 MRIs for which 4 or more lesions were present, making the effect of this decision on the outcomes negligible. In addition, leptomeningeal disease was not assessed in the spine. Further deep learning classification models for leptomeningeal disease using spine MRIs should be investigated to supplement our brain tumor models for complete RAPNO response assessment. Finally, while these data were collected from multiple institutions, the institutions specifically in Hunan are in close proximity to each other and therefore patient demographics may be similar across these institutions. The tool should be validated in larger, prospective studies prior to widespread implementation.

In conclusion, we developed a fully automatic pipeline for the segmentation of tumors in pediatric MBL, HGG, and other leptomeningeal seeding tumors using deep learning. We further developed an algorithm that automatically calculates RAPNO scores for segmented tumors. The models demonstrate excellent performance in both pre- and post-operative MRIs, with high agreement in volume estimates and RAPNO scores with human experts. Additionally, we demonstrate that AutoRAPNO measurements are superior to those derived by human experts given the algorithm's ability to more accurately and consistently distinguish separate lesions. Finally, our study demonstrates that fully automated deep learning-based pipelines can successfully be applied to the pediatric population. This tool may aid clinicians and clinical trial investigators in response assessment for pediatric tumors.

Supplementary Material

Supplementary material is available at *Neuro-Oncology* online.

Keywords

brain | deep learning | response assessment | segmentation

Funding

This work was supported by the Natural Science Foundation of China (81971696 to L.Y.) and Sheng Hua Yu-Ying Project of Central South University to L.Y. Research reported in this publication was supported by a training grant from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under award number 5T32EB1680 to K.C. and by the NCI of the National Institutes of Health under Award Number F30CA239407 to K.C. This project was supported by the 111 project (No. B18059) to B.Z. This work is supported by the Scientific and Technological Innovation Leading Plan of High-tech Industry of Hunan Province (2020GK2021) to C.Z.

Conflict of interest statement. The authors declare no potential conflicts of interest.

Authorship statement. Conception and design: H.X.B., R.Y.H., L.Y., J.P., D.D.K., and K.C. Development of methodology: H.X.B., R.Y.H., J.L.B., P.Y.W., and K.E.W. Acquisition of data: J.P., X.Z., J.W., H.Z., C.Z., X.X., D.J.D., L.J.S., and K.J. Analysis and interpretation of data: J.P., D.D.K., X.Z., and J.H. Writing, review, and/or revision of the manuscript: J.B.P., J.P., D.D.K., H.X.B., R.Y.H., P.Y.W., K.C., K.E.W., J.K.-C., X.F., T.Y.P., and J.S. Administrative, technical, or material support (ie, reporting or organizing data, constructing databases): L.Y., D.J.D., L.J.S., X.F., C.Z., and B.Z. Study supervision: H.X.B., R.Y.H., and L.Y.

References

1. Udaka YT, Packer RJ. Pediatric brain tumors. *Neurol Clin*. 2018;36(3):533–556.
2. Albright AL, Sposto R, Holmes E, et al. Correlation of neurosurgical subspecialization with outcomes in children with malignant brain tumors. *Neurosurgery*. 2000;47(4):879–885; discussion 885–887.
3. Warren KE, Vezina G, Poussaint TY, et al. Response assessment in medulloblastoma and leptomeningeal seeding tumors: recommendations from the Response Assessment in Pediatric Neuro-Oncology committee. *Neuro Oncol*. 2018;20(1):13–23.
4. Erker C, Tamrazi B, Poussaint TY, et al. Response assessment in paediatric high-grade glioma: recommendations from the Response Assessment in Pediatric Neuro-Oncology (RAPNO) working group. *Lancet Oncol*. 2020;21(6):e317–e329.
5. Pollack IF. The role of surgery in pediatric gliomas. *J Neurooncol*. 1999;42(3):271–288.
6. Minturn JE, Fisher MJ. Gliomas in children. *Curr Treat Options Neurol*. 2013;15(3):316–327.
7. Kline C, Felton E, Allen IE, Tahir P, Mueller S. Survival outcomes in pediatric recurrent high-grade glioma: results of a 20-year systematic review and meta-analysis. *J Neurooncol*. 2018; 137(1):103–110.
8. Shah GD, Kesari S, Xu R, et al. Comparison of linear and volumetric criteria in assessing tumor response in adult high-grade gliomas. *Neuro Oncol*. 2006;8(1):38–46.
9. Vos MJ, Uitdehaag BMJ, Barkhof F, et al. Interobserver variability in the radiological assessment of response to chemotherapy in glioma. *Neurology*. 2003;60(5):826–830.
10. Hayward RM, Patronas N, Baker EH, et al. Inter-observer variability in the measurement of diffuse intrinsic pontine gliomas. *J Neurooncol*. 2008;90(1):57–61.
11. Deeley MA, Chen A, Datteri R, et al. Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. *Phys Med Biol*. 2011;56(14):4557–4577.
12. Boxerman JL, Zhang Z, Safriel Y, et al. Early post-bevacizumab progression on contrast-enhanced MRI as a prognostic marker for overall survival in recurrent glioblastoma: results from the ACRIN 6677/RTOG 0625 Central Reader Study. *Neuro Oncol*. 2013;15(7):945–954.
13. Ellingson BM, Bendszus M, Boxerman J, et al. Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials. *Neuro Oncol*. 2015;17(9):1188–1198.
14. Huang R. Response assessment in high-grade glioma: tumor volume as endpoint. *Neuro Oncol*. 2017;19(6):744–745.
15. Havaei M, Davy A, Warde-Farley D, et al. Brain tumor segmentation with deep neural networks. *Med Image Anal*. 2017;35:18–31. doi:10.1016/j.media.2016.05.004
16. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. 2017;36:61–78. doi:10.1016/j.media.2016.10.004
17. Quon JL, Bala W, Chen LC, et al. Deep learning for pediatric posterior fossa tumor detection and classification: a multi-institutional study. *Am J Neuroradiol*. 2020;41(9):1718–1725.
18. Chang K, Beers AL, Bai HX, et al. Automatic assessment of glioma burden: a deep learning algorithm for fully automated volumetric and bidimensional measurement. *Neuro Oncol*. 2019;21(11):1412–1422.
19. Feng X, Tustison NJ, Patel SH, Meyer CH. Brain tumor segmentation using an ensemble of 3D U-Nets and overall survival prediction using radiomic features. *Front Comput Neurosci*. 2020;14. doi:10.3389/fncom.2020.00025
20. Bakas S, Reyes M, Jakab A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. ArXiv181102629 Cs Stat. Published online April 23, 2019. http://arxiv.org/abs/1811.02629. Accessed January 16, 2021.
21. Beers A, Brown J, Chang K, et al. DeepNeuro: an open-source deep learning toolbox for neuroimaging. *Neuroinformatics*. 2021;19(1):127–140.
22. Iglesias JE, Liu C-Y, Thompson PM, Tu Z. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans Med Imaging*. 2011;30(9):1617–1634.
23. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. 2010;29(6):1310–1320.
24. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. ArXiv150504597 Cs. Published online May 18, 2015. http://arxiv.org/abs/1505.04597. Accessed January 16, 2021.
25. Loshchilov I, Hutter F. Decoupled weight decay regularization. ArXiv171105101 Cs Math. Published online January 4, 2019. http://arxiv.org/abs/1711.05101. Accessed January 16, 2021.
26. Loshchilov I, Hutter F. SGDR: stochastic gradient descent with warm restarts. ArXiv160803983 Cs Math. Published online May 3, 2017. http://arxiv.org/abs/1608.03983. Accessed January 16, 2021.
27. Bourke P. Trilinear interpolation. http://paulbourke.net/miscellaneous/interpolation/. Accessed January 16, 2021.

28. Wu Y, He K. Group normalization. ArXiv180308494 Cs. Published online June 11, 2018. http://arxiv.org/abs/1803.08494. Accessed January 16, 2021.
29. Dosovitskiy A, Springenberg JT, Riedmiller M, Brox T. Discriminative unsupervised feature learning with convolutional neural networks. *Adv Neural Inf Process Syst*. 2014;27:766–774.
30. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261–272.
31. Kokoska S, Zwillinger D. *CRC Standard Probability and Statistics Tables and Formulae, Student Edition*. 1st ed. Boca Raton, FL: CRC Press; 2000.
32. Gamer M, Lemon J, Fellows I, Singh P. Various coefficients of interrater reliability and agreement. https://cran.r-project.org/web/packages/irr/irr.pdf (Accessed January 2010).
33. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep*. 1966;19(1):3–11.
34. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996;1(1):30–46.
35. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–428.
36. Phan TV, Smeets D, Talcott JB, Vandermosten M. Processing of structural neuroimaging data in young children: bridging the gap between current practice and state-of-the-art methods. *Dev Cogn Neurosci*. 2018;33:206–223. doi:10.1016/j.dcn.2017.08.009
37. Peng J, Zhou H, Tang O, et al. Evaluation of RAPNO criteria in medulloblastoma and other leptomeningeal seeding tumors using MRI and clinical data. *Neuro Oncol*. 2020;22(10):1536–1544.