Neuro-Oncology Advances

Automated pediatric brain tumor imaging assessment tool from CBTN: Enhancing suprasellar region inclusion and managing limited data with deep learning

Deep B. Gandhi[†,©], Nastaran Khalili[†], Ariana M. Familiar, Anurag Gottipati, Neda Khalili, Wenxin Tu, Shuvanjan Haldar, Hannah Anderson, Karthik Viswanathan, Phillip B. Storm, Jeffrey B. Ware, Adam Resnick, Arastoo Vossough, Ali Nabavizadeh[©] and , Anahita Fathi Kazerooni[©]

All author affiliations are listed at the end of the article

Corresponding Author: Anahita Fathi Kazerooni, PhD, Center for Data-Driven Discovery in Biomedicine (D³b), Children's Hospital of Philadelphia, Roberts Center for Pediatric Research, 2716 South St, Philadelphia, PA 19146, USA (anahitaf@upenn.edu).

[†]Equally contributing first authors.

Abstract

Background. Fully automatic skull-stripping and tumor segmentation are crucial for monitoring pediatric brain tumors (PBT). Current methods, however, often lack generalizability, particularly for rare tumors in the sellar/suprasellar regions and when applied to real-world clinical data in limited data scenarios. To address these challenges, we propose AI-driven techniques for skull-stripping and tumor segmentation.

Methods. Multi-institutional, multi-parametric MRI scans from 527 pediatric patients ($n = 336$ for skull-stripping, $n = 489$ for tumor segmentation) with various PBT histologies were processed to train separate nnU-Net-based deep learning models for skull-stripping, whole tumor (WT), and enhancing tumor (ET) segmentation. These models utilized single (T2/FLAIR) or multiple (T1-Gd and T2/FLAIR) input imaging sequences. Performance was evaluated using Dice scores, sensitivity, and 95% Hausdorff distances. Statistical comparisons included paired or unpaired 2-sample t-tests and Pearson's correlation coefficient based on Dice scores from different models and PBT histologies.

Results. Dice scores for the skull-stripping models for whole brain and sellar/suprasellar region segmentation were 0.98 ± 0.01 (median 0.98) for both multi- and single-parametric models, with significant Pearson's correlation coefficient between single- and multi-parametric Dice scores ($r > 0.80$; $P < .05$ for all). Whole tumor Dice scores for single-input tumor segmentation models were 0.84 ± 0.17 (median = 0.90) for T2 and 0.82 ± 0.19 (median = 0.89) for FLAIR inputs. Enhancing tumor Dice scores were 0.65 ± 0.35 (median = 0.79) for T1-Gd+FLAIR and 0.64 ± 0.36 (median = 0.79) for T1-Gd+T2 inputs.

Conclusion. Our skull-stripping models demonstrate excellent performance and include sellar/suprasellar regions, using single- or multi-parametric inputs. Additionally, our automated tumor segmentation models can reliably delineate whole lesions and ET regions, adapting to MRI sessions with missing sequences in limited data context.

Key Points

- Deep learning models for skull-stripping, including the sellar/suprasellar regions, demonstrate robustness across various pediatric brain tumor histologies.
- The automated brain tumor segmentation models perform reliably even in limited data scenarios.

Importance of the Study

We present robust skull-stripping models that work with single- and multi-parametric MR images and include the sellar/suprasellar regions in the extracted brain tissue. Since ~10% of the pediatric brain tumors originate in the sellar/suprasellar region, including the deep-seated regions within the extracted brain tissue makes these models generalizable for a wider range of tumor histologies. We also present 2 tumor segmentation models, one for segmenting whole tumor using T2/FLAIR images, and another for segmenting enhancing tumor region using T1-Gd and T2/FLAIR images. These models demonstrate excellent performance with limited input. Both the skull-stripping and tumor segmentation models work with one- or two-input MRI sequences, making them useful in cases where multi-parametric images are not available—especially in real-world clinical scenarios. These models help to address the issue of missing data, making it possible to include subjects for longitudinal assessment and monitoring treatment response, which would have otherwise been excluded.

Pediatric brain tumors (PBTs) are the most prevalent childhood cancers of the central nervous system (CNS), encompassing a wide range of histologies and survival rates,[1–4] and are one of the leading causes of cancer-related deaths in children, only secondary to leukemia.[5–7] The World Health Organization (WHO), in the fifth edition of its Classification of Tumors of the CNS (WHO CNS5), recognizes that PBTs possess distinct histological and molecular features.[8] Consequently, there are notable differences in neuroimaging characteristics between adults and PBT, including variations in brain structures, image signal intensity, skull formation, and tumor subregions.[9] These differences underscore the need for image processing and assessment tools tailored specifically to pediatric neuroimaging data.

Quantitative analysis of PBTs for response assessment requires accurately locating and delineating the tumorous region, a challenging and tedious task prone to inter-reader variability and lack of consensus.[10,11] While established automated preprocessing and tumor size measurement approaches exist for adult brain tumors,[12] and despite recent advances in developing pediatric-specific automated methods for tumor assessment,[13–18] there still remains a lack of comprehensive methods addressing the unique challenges of tumor assessment in pediatric patients.

Automated pediatric-specific approaches employing deep learning models, such as convolutional neural networks (CNNs), have been utilized for skull-stripping and segmentation of whole lesions or tumor subregions.[9,14–16,19] Skull-stripping, also referred to as brain extraction, is a crucial image preprocessing technique for isolating brain tissue from non-brain tissue in MRI. This step is vital for downstream neuroimaging analyses and plays an essential role in ensuring patient anonymization during data sharing. Various imaging analysis methodologies, including image intensity standardization for radiomic feature extraction, image registration, tumor segmentation, and the mapping of MRI to other imaging modalities, achieve higher accuracy when the images are skull-stripped.[20,21] Previous skull-stripping methods were either developed using MRI data from patients without brain tumors,[22] based on adult brain tumors,[20,23] or, although trained for PBTs—as in our previous study[16]—did not adequately cover deep-seated brain regions such as the sellar/suprasellar areas, leading to undersegmentation of tumors in these regions. In pediatrics, sellar and suprasellar tumors account for approximately 10% of all CNS tumors and encompass a diverse array of entities, each with unique histologic origins and radiological features.[24] These tumors often present with specific clinical and neuroimaging characteristics, necessitating tailored surgical interventions and therapeutic approaches.[24] Therefore, it is essential to accurately include these regions within the brain tissue for image processing tools to be generalizable across various PBT histologies. Furthermore, given that sellar/suprasellar tumors can distort the anatomy of the optic pathway, it is crucial to develop a tool that improves the extraction of brain tissue while preserving the sellar/suprasellar region.[25]

In our earlier work,[15,16] we developed multi-parametric tumor subregion segmentation models capable of efficiently predicting the whole tumor (WT) and different tumor subregions, including enhancing tumor (ET) core, non-enhancing tumor (NET) core, cystic component (CC), and peritumoral edema (ED), for a variety of PBTs using 4 standard MRI sequences: T1-weighted (T1), T1-weighted post-contrast enhanced (T1-Gd), T2-weighted (T2), and T2-weighted fluid-attenuated inversion recovery (FLAIR) images. However, in some instances, depending on the purpose of the imaging (eg, initial assessment versus follow-up imaging), not all 4 sequences are acquired at a given time point, or the images may be unusable due to artifacts or specific protocol settings. This lack of availability of multi-parametric scans, particularly in retrospective studies and longitudinal tumor response assessments, is especially pronounced in pediatric cases. The relatively low incidence of brain tumors in the pediatric population necessitates data collection from multiple sites, each with differing clinical protocols, leading to decreased harmonization of input data.

This inconsistency may result in the exclusion of subjects who otherwise meet eligibility criteria and could be included for model training and further analysis. Nevertheless, based on tumor histology, single-parametric scans can still provide valuable clinical information. For instance, in the context of diffuse midline glioma (DMG), delineating the WT based on T2 and/or FLAIR scans may be sufficient for longitudinal response assessment.[26] In the case of tumors with enhancing components, such as pediatric low-grade and high-grade gliomas (LGG and HGG, respectively), segmentation of the ET based on T1-Gd

images could be beneficial for evaluating tumor behavior and likely progression.[27,28] Although several studies have explored the efficacy of single versus multi-parametric MRI for developing segmentation pipelines in CNS lesions such as meningioma or vestibular schwannoma,[29,30] no established preprocessing pipelines or tools currently exist to address this problem in PBTs.

To address these unmet needs, we propose a generalizable pediatric preprocessing pipeline for enhanced automated brain tissue extraction (skull-stripping) and tumor segmentation. This pipeline complements our previously developed multi-parametric auto skull-stripping and tumor segmentation models. In this study, we trained 3D convolutional neural networks (CNNs) using a U-Net-based architecture (nnU-Net)[31] with multi- and single-parametric MRI sequences as inputs for the auto-segmentation tasks. We selected nnU-Net because it has been proven to outperform most available CNNs, especially in applications involving PBTs.[15,32,33]

We hypothesize that this pipeline can establish a standardized method for preprocessing pediatric brain MRI acquisitions. Furthermore, we expect that the proposed auto-segmentation models will demonstrate acceptable performance in segmenting either the WT or ET components, including sellar and suprasellar regions, even in subjects lacking multi-parametric MRI scans.

Methods

Data Description and Patient Cohort

This was a HIPAA-compliant, IRB-approved retrospective study of previously acquired data from the multi-institutional Children's Brain Tumor Network (CBTN) repository and BraTS-PEDs 2023.[34,35] Children's Brain Tumor Network is a biorepository (cbtn.org) that allows for the collection of specimens, longitudinal clinical and imaging data, and sharing of de-identified samples and data for future research. Written informed consent was obtained from the patients at the time of their enrollment. Subjects were included if they had the following 4 MR images obtained routinely for clinical evaluation of brain tumors: T1, T1-Gd, T2, and FLAIR. Additionally, only subjects who underwent minor surgical procedures that did not result in major neuroanatomical changes and had all 4 brain MR images mentioned above, were included. Subjects were excluded if they underwent surgical procedures that resulted in major changes to neuroanatomy.

Based on the availability of ground truth brain masks and tumor segmentations, 2 different subject cohorts were created as a part of this study, one for training the skull-stripping models and another for training the tumor segmentation models. A detailed description of the patient demographics and distribution of tumor histology for each cohort is provided (Figure 1).

Image Preprocessing and Data Preparation

Data preparation for single- and multi-parametric model training included a series of preprocessing steps (Supplementary Figure 1). First, all images were reoriented to left-posterior-superior (LPS) coordinate system. Next, the T1-Gd image was coregistered to the SRI24-atlas space and subsequently the T1, T2, and FLAIR images were coregistered to the T1-Gd image. Images were resampled to 1 mm^3 isotropic resolution and the image dimensions were changed to 240 × 240 × 155, based on the anatomical SR124-atlas space.[36] Coregistration was performed using a greedy algorithm in the Cancer Imaging Phenomics Toolkit open-source software v.1.8.1 (CaPTk, https://www.cbica.upenn.edu/captk).[37,38]

A semiautomated process was used to create ground truth segmentation masks for model training. For the skull-stripping model, images were passed through an existing automated skull-stripping tool based on DeepMedic from CaPTk.[39] The resulting brain masks underwent manual modification to make any corrections and, importantly, to include the sellar/suprasellar regions within the brain masks. Similarly, to generate initial tumor segmentations, images were passed through a baseline automated tumor segmentation tool,[15] which segmented 4 various tumor subregions—ET, NET, CC, and ED—where present. For the baseline segmentations generated above, manual revisions were made by trained researchers using ITK-SNAP.[40] These revisions were then reviewed by one of the 3 practicing neuroradiologists (J.W. with 6 years, A.N. with 10 years, and A.V. with 16 years of clinical neuroradiology experience, respectively), and were iteratively corrected until final approval by one of the neuroradiologists. All manual annotators, including the neuroradiologists, had received prior training and participated in consensus sessions. From the finalized manual segmentations, WT segmentation masks were created by combining all tumor subregions into a single segmentation label. Separately, the ET subregion was extracted. These WT and ET masks were used as ground truth for model training and evaluation.

Model Training and Validation

We trained two 3D CNNs using nnU-Net for automated skull-stripping: one with multi-parametric input, and the other with single-parametric input (T1, T1-Gd, and T2 or FLAIR). We also trained 2 different 3D CNNs using nnU-Net for automated brain tumor segmentation: one using only T2 or FLAIR images as input for auto-segmentation of the WT region (without individual tumor subregions), and the other using T1-Gd along with either T2 or FLAIR images for auto-segmentation of the ET region (excluding other tumor subregions). nnU-Net v1 (https://github.com/MIC-DKFZ/nnUNet/tree/nnunetv1) with 5-fold cross validation was trained.[31] Training parameters were: initial learning rate = 0.0, stochastic gradient descent (SGD) with Nesterov momentum (μ = 0.99), and number of epochs = 1000 × 250 minibatches.

A total of 336 subjects were included in the skull-stripping model training, with 153 subjects in the training cohort and a withheld set of 183 subjects in the testing cohort. Four hundred eighty-nine subjects were included in tumor segmentation model training, with 364 subjects in the training cohort and a withheld set of 125 subjects in the testing cohort. A variety of patient demographics and brain tumor histology

Figure 1. Patient demographics for subjects used in model training and testing for the skull-stripping and brain tumor segmentation cohorts.
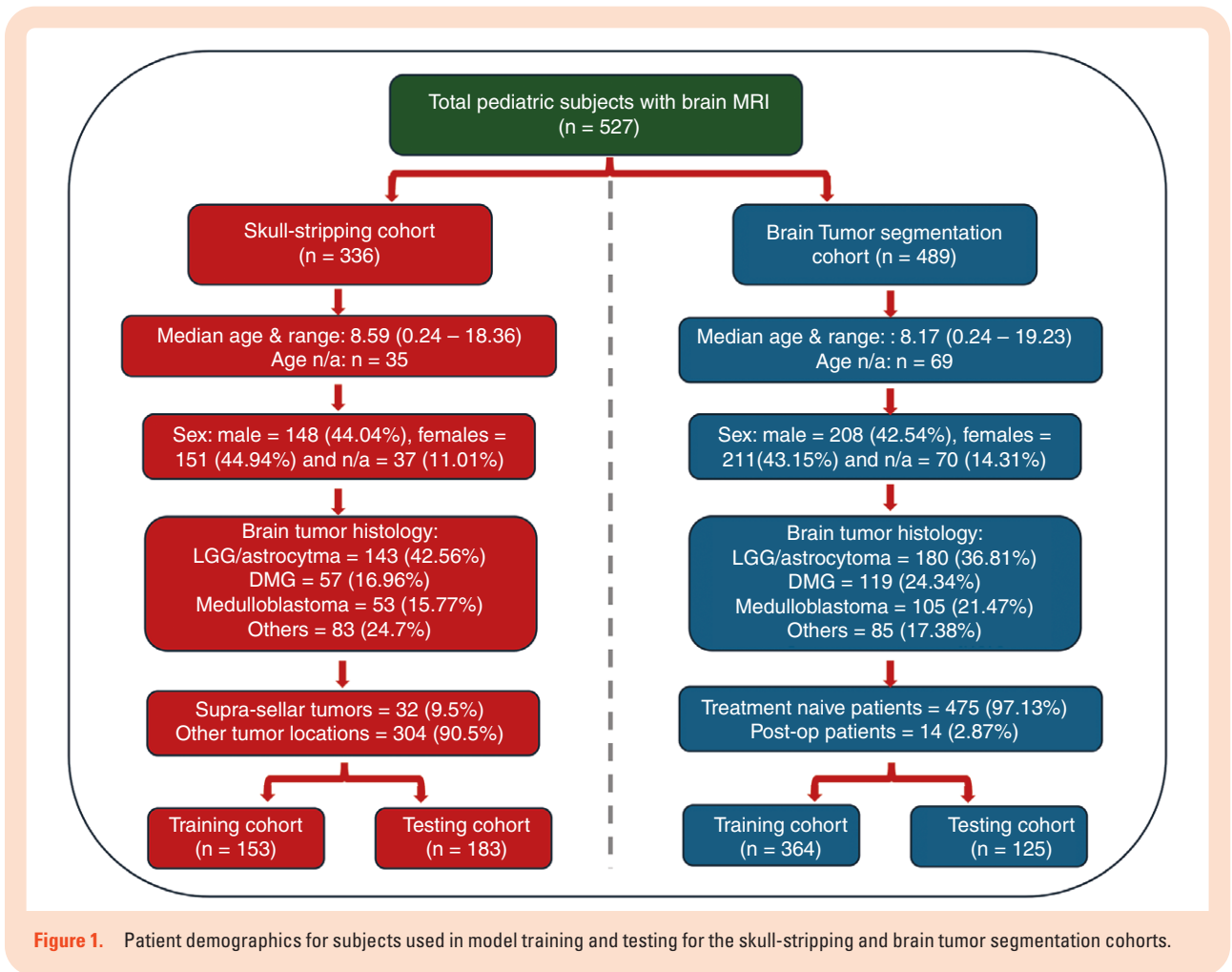
were included in this study in order to build robust and generalizable models (Figure 1). The entire image preprocessing, automatic skull-stripping, and tumor segmentation pipeline are presented in Supplementary Figure 1.

Statistical Analysis

The performances of the different nnU-Net models with respect to the expert manual ground truth segmentations were evaluated using several evaluation metrics, including Dice score (Sørensen-Dice similarity coefficient), sensitivity, and 95% Hausdorff distance.

For skull-stripping, paired *t*-tests were used to determine any differences in the Dice scores between single- and multi-parametric skull-stripping models, whereas 2-sample *t*-tests were used to compare the differences in Dice scores between different PBT histologies and age ranges. The correspondence in performance between single- and multi-parametric skull-stripping models was evaluated using Pearson's correlation.

For tumor segmentation: paired *t*-tests were used to compare the performance between T2 or FLAIR inputs for WT segmentation, T1-Gd and T2 or FLAIR inputs for ET segmentation for the different histologies. Two-sample *t*-tests were used to compare the difference in ET volumes

for different histologies, using T1-Gd and T2 or T1-Gd and FLAIR inputs. The correlation between ET Dice scores and ET volumes was evaluated using Pearson's correlation.

Results

Skull-Stripping Model performance

Supplementary Table 1 shows the resulting Dice scores, sensitivity, and 95% Hausdorff distance for the multi- and single-parametric skull-stripping models for both the whole brain and when selecting only the slices containing sellar/suprasellar regions. For whole brain masks, the multi- and single-parametric models demonstrated similar performance, as indicated by the Dice scores (Supplementary Table 1; Figure 2A). A similar trend was observed for the sellar/suprasellar slices (Supplementary Table 1; Figure 2B), with the median Dice scores being slightly higher than those for whole brain masks (Supplementary Table 1). When comparing the performance of the proposed muti-parametric model against the DeepMedic model (which was not trained to include sellar/suprasellar regions) specifically for the sellar/suprasellar slices, the Dice scores (mean ± SD [median]) are similar
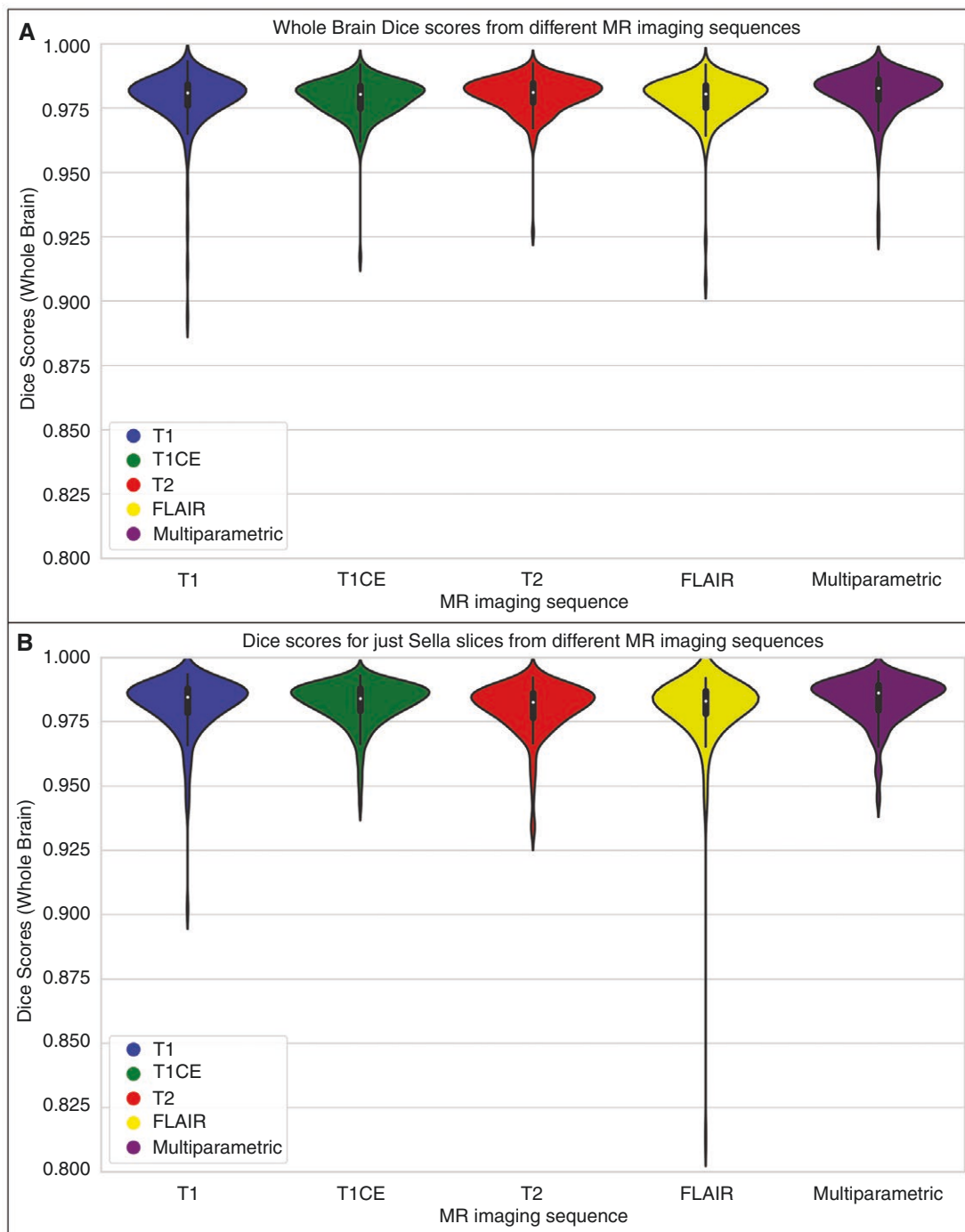
Figure 2. Violin plots showing the distribution of Dice scores for the single-parametric skull-stripping models (T1, T1-Gd, and T2 or FLAIR images) compared to the multi-parametric skull-stripping model for whole brain mask (A) and sellar/suprasellar slices (B).

(0.98 ± 0.01 [0.99] vs 0.98 ± 0.01 [0.98]). However, the proposed model showed higher sensitivity (0.98 ± 0.01[0.99] vs 0.97 ± 0.01 [0.98]) and lower Hausdorff distance (1.06 ± 0.32 [1] vs 1.18 ± 0.36 [1]). The differences in performance between the 2 models are less pronounced due to the smaller size of the sellar/suprasellar regions compared to the entire brain tissue.

The distribution of whole brain Dice scores using the multi- and single-parametric skull-stripping models for different brain tumor histologies—low-grade glioma (LGG), medulloblastoma, diffuse midline glioma (DMG), and Other Histologies including high-grade glioma—astrocytoma (HGG), ependymoma, and ganglioglioma—shows that both models performed similarly well (Supplementary Figure 2).
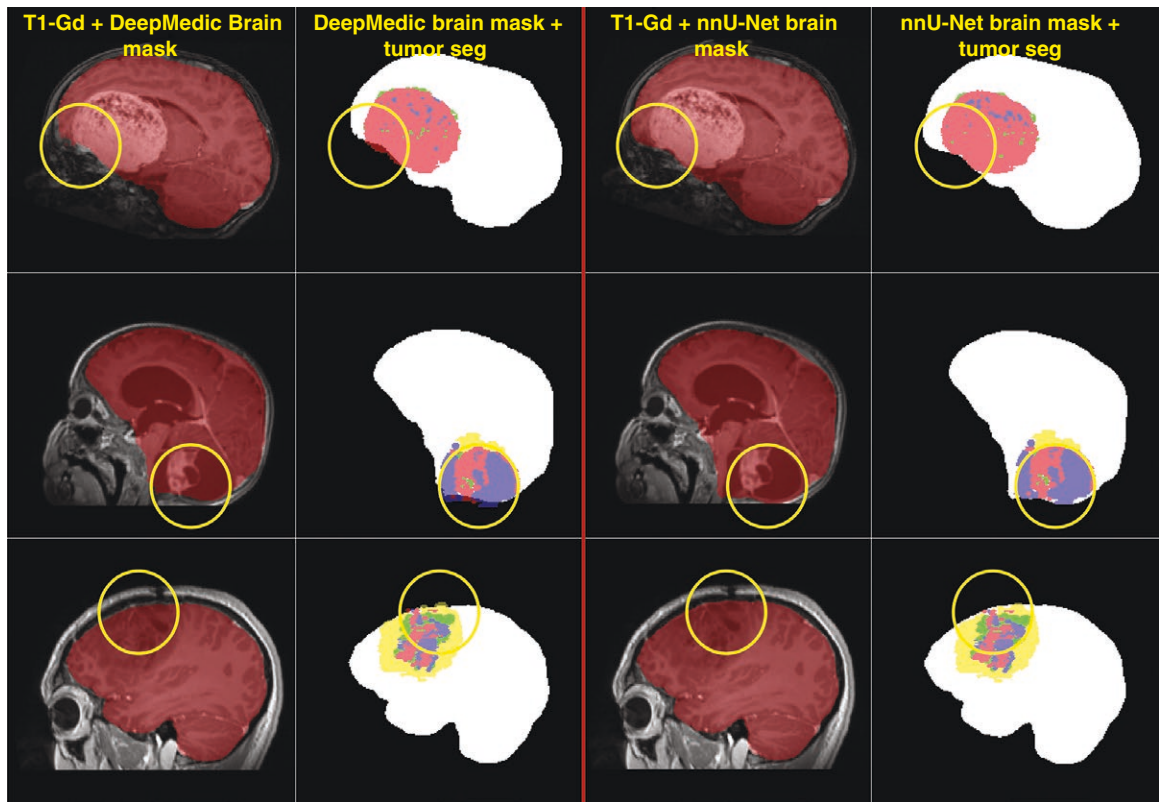
Figure 3. Examples of skull-stripping performance using the multi-parametric nnU-Net-based model compared to an earlier pediatric DeepMedic-based skull-stripping model. The nnU-Net-based model shows improved performance, successfully segmenting tumor regions that the DeepMedic-based model fails to include as part of the brain tissue.

Performance was largely affected by one DMG subject in the single-parametric model with T1 input, which had a whole brain Dice score of 0.89, leading to a slightly less dense group-level distribution.

Comparison of the Dice scores between the multi- and single-parametric models for whole brain and sellar slices showed a significant correlation ($P < .05$), demonstrating similar performance for skull-stripping in both whole brain and sellar slices (Supplementary Figures 3 and 4).

Whole brain Dice scores of the multi-parametric model showed no significant difference across different age ranges (0–3, 3–13, and 13–18 years) (all $P > .05$), suggesting the model's generalizability across all pediatric age groups (Supplementary Figure 5).

Supplementary Figure 6 depicts representative pre-processed brain MR images overlaid with ground truth brain masks along with predicted whole brain masks from single- and multi-parametric skull-stripping models. The successful performance of the multi-parametric model on craniopharyngioma and germinoma, which originate from the sellar/suprasellar regions, is also shown.

We further tested the impact of our proposed skull-stripping model on automated tumor segmentation by comparing skull-stripped images to nonskull-stripped images in data from 12 subjects with sellar/suprasellar tumors. Supplementary Table 2 shows the mean ± SD (median) values for Dice score, sensitivity, and 95%

Hausdorff distance metrics for WT segmentation in both skull-stripped and nonskull-stripped images. The results indicate no significant difference, based on the Wilcoxon signed-rank test ($P = .18$), suggesting no adverse impact of skull-stripping on a downstream task. Finally, a qualitive comparison of the nnU-Net-based skull-stripping model with the previously reported DeepMedic-based skull-stripping model, reveals that the nnU-Net-based skull-stripping model performs better at including sellar/suprasellar areas, frontal lobe, and brainstem from the brain tumor as part of the extracted brain tissue region (Figure 3).

Tumor Segmentation Model Performance

Table 1 shows the Dice scores, sensitivity, and 95% Hausdorff distance for (1) T2 or FLAIR model–WT region and (2) T1-Gd and either T2 or FLAIR model–ET region. The distribution of Dice scores from the T2 and FLAIR models showed no significant difference between various tumor histologies for both T2 and FLAIR images (Figure 4A). An example of preprocessed MRI sequences, images overlaid with the ground truth segmentation, and predicted segmentations for the T1-Gd and FLAIR model, T1-Gd and T2 model, and T2 or FLAIR model is shown (Figure 5A–C, respectively).

Table 1. Performance Metrics for T2 or FLAIR Whole Tumor Segmentation Models, and T1-Gd and T2 or FLAIR Enhancing Tumor Segmentation Models

Model and region/metric	Dice score Mean ± SD (median)	Sensitivity Mean± SD (median)	95% Hausdorff distance Mean ± SD (median)
One-sequence models—WT			
FLAIR-only	0.84 ± 0.17 (0.90)	0.83 ± 0.18 (0.88)	8.09 ± 13.20 (3.32)
T2-only	0.82 ± 0.19 (0.89)	0.8 ± 0.21 (0.88)	8.24 ± 12.79 (3.61)
Two-sequence models—ET			
T1-Gd and FLAIR	0.65 ± 0.35 (0.79)	0.76 ± 0.26 (0.86)	6.41 ± 9.25 (3)
T1-Gd and T2	0.64 ± 0.36 (0.79)	0.76 ± 0.26 (0.85)	6.09 ± 9.33 (2.83)

Abbreviations: ET, enhancing tumor; FLAIR, fluid-attenuated inversion recovery; SD, standard deviation; WT, whole tumor.
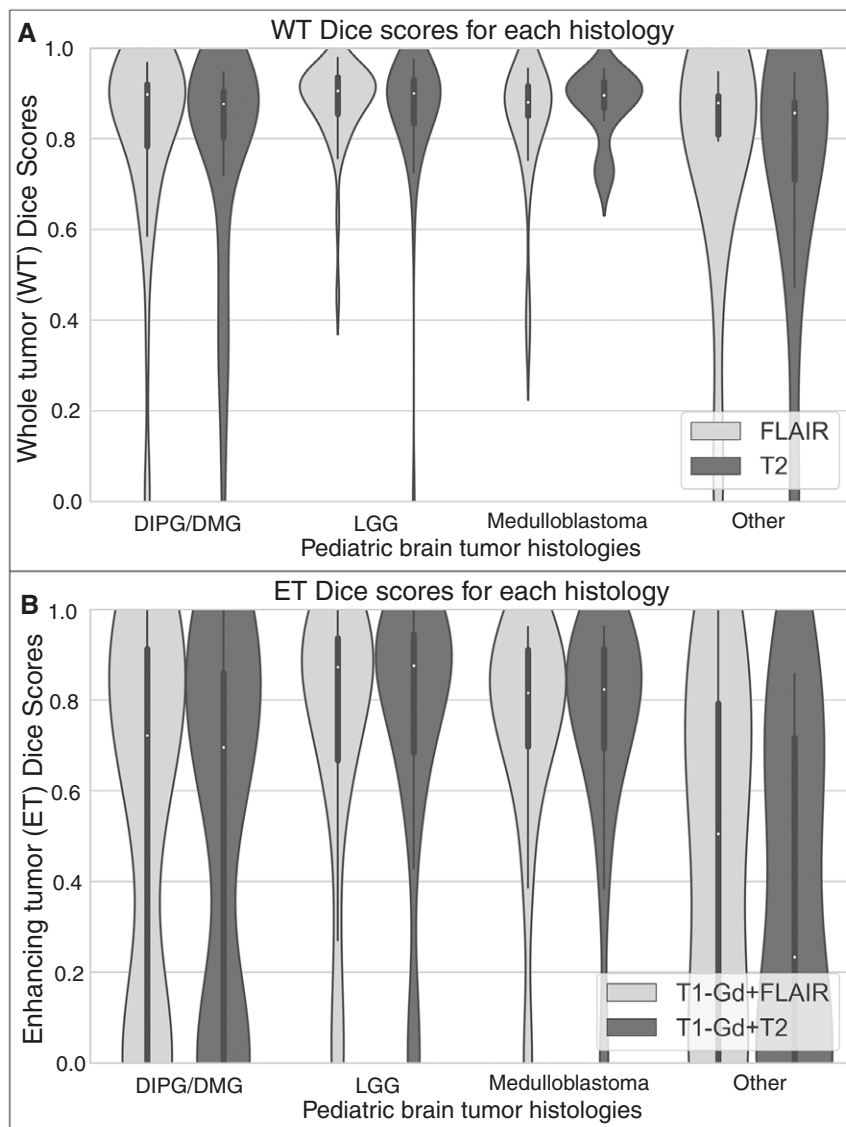

Figure 4. Violin plots showing the distribution of whole tumor (WT) Dice scores using the T2 or FLAIR tumor segmentation model for different histologies (A). There is no significant difference in the Dice scores between T2 and FLAIR inputs for different tumor histologies. Violin plots showing the distribution of enhancing tumor (ET) Dice scores using the T1-Gd and T2 or FLAIR tumor segmentation models for the different histologies (B). Enhancing tumor Dice scores from LGG and medulloblastoma patients were significantly higher than those from DIPG/DMG and other histologies (all $P < .0.5$).
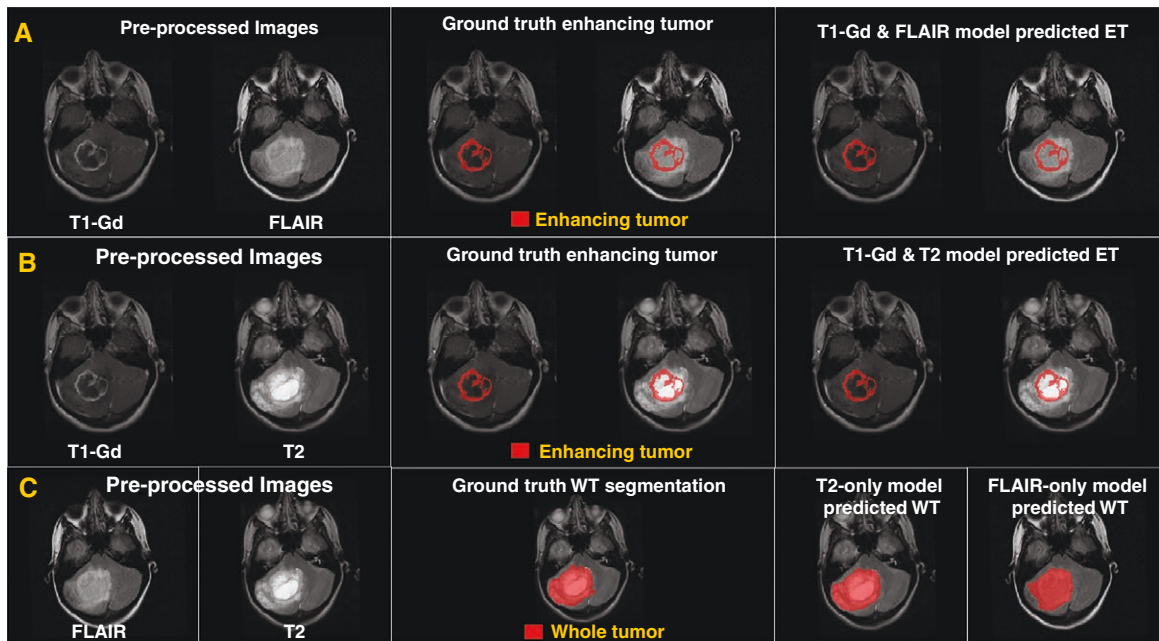
Figure 5. (A) and (B) Examples of brain MR images overlaid with ground truth segmentation labels and model-predicted enhancing tumor segmentation labels for T1-Gd and FLAIR, and T1-Gd and T2, respectively. (C) Results for the T2 or FLAIR model for segmenting whole tumor (WT).

The distribution of ET Dice scores using theT1-Gd withT2 or FLAIR inputs for different PBT histologies showed that ET Dice scores for DMG are significantly lower than those for LGG and medulloblastoma (Figure 4B). Two sample t-tests indicated significant differences in ET Dice scores between DMG and LGG ($P = .004$ for T1-Gd and T2, $P = .018$ for T1-Gd and FLAIR), between DMG and medulloblastoma ($P = .02$ for T1-Gd and T2, $P = .037$ for T1-Gd and FLAIR). Additionally, significant differences were found between tumors categorized as "Other Histologies" and LGG ($P = .0003$ for T1-Gd and T2, $P = .011$ for T1-Gd and FLAIR) and "Other Histologies" and medulloblastoma ($P = .0008$ for T1-Gd and T2, $P = .011$ for T1-Gd and FLAIR). High-grade gliomas and medulloblastoma subjects had significantly higher ET volumes than DMG and "Other Histologies," indicating the model's superior performance in histologies where the ET region is more prevalent and tumor size is larger (Supplementary Figure 7A, B). A scatter plot comparing ET Dice scores with ground truth ET volumes showed a low but significant correlation ($r = 0.33$, $P < .05$) (Supplementary Figure 7C).

Discussion

In this study, we aimed to develop generalizable, pediatric-specific automated methods for skull-stripping that includes the sellar/suprasellar regions and an automated tumor segmentation approach for use in limited data contexts. Our study demonstrated excellent results for skull-stripping for both the multi-parametric and single-parametric models, with similar performance on the whole brain mask and sellar/suprasellar slices. The brain tumor segmentation models presented in this study build on our previous work, enabling the evaluation of imaging sessions of PBTs without multi-parametric MRI acquisitions. Additionally, these models allow for the segmentation of the nonenhancing component by subtracting ET from WT in cases where T1-Gd and either T2 or FLAIR images are available.

This study utilized the most comprehensive pediatric brain MRI dataset for training skull-stripping and tumor auto-segmentation models to date, based on the diversity of tumor histologies included in both training and testing. The dataset was multi-institutional, encompassing a wide range of MRI scanner field strengths and manufacturers (see Supplementary Table 3). It included a variety of brain tumor types, with the major ones being LGG, medulloblastomas, and DMG, as well as other types such as HGG, ependymomas, germinomas, craniopharyngiomas, and other rare tumors.

Our automated tumor segmentation models using only T2 or FLAIR sequences achieved high accuracy in WT region segmentation (median Dice scores: 0.9 for FLAIR-only input and 0.89 for T2-only input). The results for ET region segmentation using combinations of T1-Gd with T2 or FLAIR sequences were more moderate (median Dice scores: 0.79 for both T1-Gd & T2 and T1-Gd & FLAIR models).

Comparing our results to the BraTS-PEDs 2023 challenge,[35] our WT and ET segmentation models performed better than the top-performing models. For WT segmentation using our T2 or FLAIR models compared to the top-performing BraTS-PEDS 2023 teams, Dice score was 0.84 ± 0.17 (0.90) versus 0.84 ± 0.16 (0.87), sensitivity was

0.83 ± 0.18 (0.88) versus 0.8 ± 0.09 (0.82), and 95% Hausdorff distance was 8.09 ± 13.20 (3.32) versus 18.05 ± 62.77 (4.30). Similarly, for ET segmentation using our T1-Gd with T2 or FLAIR models compared to the top-performing BraTS-PEDS 2023 teams, Dice score was 0.65 ± 0.35 (0.79) versus 0.65 ± 0.32 (0.74), sensitivity was 0.76 ± 0.26 (0.86) versus 0.7 ± 0.18 (0.74), and 95% Hausdorff distance was 6.41 ± 9.25 (3) versus 43.89 ± 108.59 (3.67). These results collectively highlight that while our models used limited imaging sets for segmentation of WT and ET regions, as opposed to the methods submitted to the BraTS-PEDs 2023 that use multi-parametric MRI, they achieve better performance.

While numerous studies have focused on skull-stripping in adult brain MRI scans, only a few have addressed this task using pediatric brain MRI scans. Kazerooni et al. demonstrated Dice scores of 0.98 using the DeepMedic architecture for skull-stripping with multi-parametric MR images, which is comparable to our results. However, their study used a smaller dataset of 21 cases for inference, did not include single-parametric models, and did not encompass the sellar/suprasellar regions in the brain masks.[16] Kim et al. reported whole brain Dice scores in the range of 0.79–0.8 using VUNO Med-DeepBrain for subjects with SCN1A mutations ($n = 21$) and healthy subjects ($n = 42$) in a multi-institutional and multi-scanner dataset. However, their study did not include any subjects with brain tumors and used a significantly smaller dataset compared to ours.[19] Chen et al. used ANUBEX based on nnU-Net for skull-stripping in neonates and compared it against 5 other deep learning models, achieving Dice scores in the range of 0.92–0.96.[41] However, their model only used T1 images as input and was tested on a small withheld dataset of 39 subjects. In contrast, our study demonstrated excellent skull-stripping results using both multi- and single-parametric models, outperforming previously published studies. Additionally, our study showed accurate skull-stripping performance across different pediatric age groups (0–3, 3–13, and 13–18 years), highlighting its robustness to the structural and signal intensity changes in the skull due to child development.

While Dice scores for skull-stripping with sellar/suprasellar region inclusion do not directly inform on the performance of downstream analyses, a comparison between the DeepMedic skull-stripping model (not trained to include sellar/suprasellar regions) and our proposed nnU-Net-based model indicates that the latter is qualitatively better at including brain tumor regions from the sellar/suprasellar areas, frontal lobe, and brainstem as part of the extracted brain tissue. Furthermore, using either skull-stripped images or non skull-stripped images as input to the automated tumor segmentation model did not impact its performance for tumors located in the sellar/suprasellar regions.

A few studies in the literature have investigated the use of T2 or FLAIR images for WT segmentation in pediatric populations. Boyd et al. trained and compared multiple stepwise transfer learning models based on nnU-Net for WT segmentation using T2 images.[14] The best-performing transfer-encoder model had a median Dice score of 0.877 for the internal validation set, whereas the external validation set had a median Dice score of 0.833. These median Dice scores are lower compared with the median WT Dice

score of 0.89 using T2 inputs for the tumor segmentation model presented in our study. Additionally, the transfer-encoder stepwise transfer learning model was only trained on LGG cases, whereas our multi-institutional dataset was larger and trained on a wider range of PBT histologies. Furthermore, Vafaeikia et al. trained a 2-step U-Net-based deep learning model for WT segmentation using just FLAIR images.[42] They reported a mean Dice score of 0.795, which is lower than the mean Dice score of 0.84 reported in our study for WT segmentation using just FLAIR images. The 2 step U-Net model was trained only on LGG patients and included data from a single institution, compared to the wider range of PBT histologies and multi-institutional dataset included in the present study. To the best of our knowledge, there is only one study in the literature that explored the use of one- or two-input MRI sequences for segmenting the ET region in the case of PBT. Peng et al. used T1-Gd and T2 images to train a U-Net model for ET segmentation from 638 preoperative PBT patients.[43] They reported mean and median Dice scores of 0.724 and 0.843, respectively. In comparison, our proposed model, using the same inputs, achieved mean and median Dice scores of 0.64 and 0.79, respectively. While the U-Net model was trained on a larger dataset and demonstrated better results, it is important to note that our proposed model works with T1-Gd and T2 or FLAIR inputs

Our study had a few limitations that are important to note. In subjects with smaller ET subregion volumes, even slight inaccuracies in model prediction can push the Dice scores to extreme values (close or equal to 0) due to the low number of voxels being compared. This disproportionately penalizes model performance and biases subsequent statistical analysis. Moreover, weak enhancement and poor-quality T1-Gd scans can complicate the segmentation of the ET region.[10] Despite these limitations, we believe that the proposed ET segmentation model will help reduce the burden of manual segmentation by providing a reliable initial prediction of the ET region, which can then be reviewed and modified, by experienced radiologists.

To our knowledge, no other children's hospitals currently incorporate automated tumor segmentation into their clinical workflow. However, for the BraTS-PEDs 2024 challenge, we utilized our pretrained automated tumor segmentation model to generate initial segmentations. These were subsequently reviewed and refined by experienced radiologists. The feedback from the radiologists was positive, indicating that the automated segmentations significantly reduced the manual revision workload. As part of the upcoming BraTS-PEDs 2025 challenge (now recognized as a MICCAI Lighthouse Challenge), we will thoroughly evaluate inter- and intra-reader variability when starting from automated segmentations versus manual segmentation without preprocessing. The findings from this evaluation will be shared with the community.

Future work will include training a separate tumor segmentation model for postoperative subjects, perhaps with an additional label for the resected tumor. Incorporating intraorbital tumors and fine-tuning the model with data from specific histologies, such as DMG, will improve its performance in segmenting tumor subregions, beneficial for many applications. Additionally, we plan to apply our tumor segmentation models to clinical trial studies that

monitor tumor response to treatment, to further demonstrate their generalizability. Although this study is retrospective, we are actively working on validating our models with a prospective cohort to establish a reliable framework for routine clinical application.

In summary, this study presents enhanced skull-stripping and tumor segmentation models that are more generalizable across various PBT histologies and adaptable to limited MRI sequence availability. The proposed skull-stripping models can support applications such as synthesizing missing MRI sequences using generative adversarial networks[44] and extracting radiomic features, both of which depend on accurate and comprehensive brain tissue segmentation, including the entirety of tumor.

The single-parametric skull-stripping models, as well as one-input (T2 or FLAIR) WT segmentation and 2-input (T1-Gd and T2 or FLAIR) ET segmentation, enable the inclusion of cases with incomplete multi-parametric image sets in the limited data context. These advancements facilitate more extensive clinical translation and improved assessment of PBTs.

Supplementary material

Supplementary material is available online at *Neuro-Oncology Advances* (https://academic.oup.com/noa).

Keywords

Children's Brain Tumor Network | deep learning | magnetic resonance imaging | pediatric brain tumor segmentation | skull-stripping

Conflict of interest statement

P.B.S. has served as an advisor for REGENXBIO and Latus Bio. J.B.W. has received grant support from the National Institutes of Health and U.S. Department of Defense. All other coauthors have no conflicts of interest to disclose.

Authorship Statement

Conceptualization: A.F.K., A.N., and A.V.; Resources: P.B.S., A.C.R., A.N., and A.F.K.; Data curation: D.B.G., N.K., A.M.F., A.G., N.K., W.T., S.H., H.A., and K.V.; Writing—original draft preparation: D.B.G., N.K., and A.F.K.; Writing—review and editing: A.N., J.B.W., A.M.F., A.F.K., D.B.G., N.K., and A.V.; Supervision: A.F.K. and A.N.; Funding acquisition: A.F.K., P.B.S., and A.C.R.

Data Availability

All image processing tools that were used in this study are freely available for public use (CaPTk, https://www.cbica.upenn.edu/captk; ITK-SNAP, https://www.itksnap.org). The pediatric preprocessing and segmentation pipeline, along with pretrained nnU-Net skull-stripping and tumor segmentation models are publicly available online at (https://github.com/d3b-center/peds-brain-seg-pipeline-public). The stand-alone skull-stripping models are publicly available at (https://github.com/d3b-center/peds-brain-auto-skull-strip). Additionally, the data used for training/testing the models can be made available by the corresponding author upon reasonable request.

Affiliations

Center for Data-Driven Discovery in Biomedicine (D³b), The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA (D.B.G., Nastaran Khalili, A.M.F., A.G., Neda Khalili, W.T., S.H., H.A., K.V., P.B.S., A.R., A.V., A.N., A.F.K.); Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA (H.A., J.B.W., A.V., A.N., A.F.K.); Department of Neurosurgery, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA (P.B.S., A.R., A.F.K.); Department of Neurosurgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA (P.B.S., A.F.K.); Division of Radiology, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA (A.V.)

References

1. Hossain MJ, Xiao W, Tayeb M, Khan S. Epidemiology and prognostic factors of pediatric brain tumor survival in the US: evidence from four decades of population data. *Cancer Epidemiol.* 2021;72:101942.
2. Siegel DA, Li J, Ding H, et al. Racial and ethnic differences in survival of pediatric patients with brain and central nervous system cancer in the United States. *Pediatr Blood Cancer.* 2019;66(2):e27501.
3. Claus EB, Black PM. Survival rates and patterns of care for patients diagnosed with supratentorial low-grade gliomas. *Cancer.* 2006;106(6):1358–1363.
4. Girardi F, Matz M, Stiller C, et al. Global survival trends for brain tumors, by histology: analysis of individual records for 67,776 children diagnosed in 61 countries during 2000-2014 (CONCORD-3). *Neuro-Oncol.* 2023;25(3):580–592.
5. Ostrom QT, Price M, Ryan K, et al. CBTRUS statistical report: pediatric brain tumor foundation childhood and adolescent primary brain and other central nervous system tumors diagnosed in the United States in 2014–2018. *Neuro-Oncol.* 2022;24(Supplement_3):iii1–iii38.
6. Udaka YT, Packer RJ. Pediatric brain tumors. *Neurol Clin.* 2018;36(3):533–556.

7. Frühwald MC, Rutkowski S. Tumors of the central nervous system in children and adolescents. *Deutsches Ärzteblatt International.* 2011;108(22):390–397.

8. 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(8):1231–1251.

9. Simarro J, Meyer MI, Van Eyndhoven S, et al. A deep learning model for brain segmentation across pediatric and adult populations. *Sci Rep.* 2024;14(1):11735.

10. Familiar AM, Fathi Kazerooni A, Vossough A, et al. Towards consistency in pediatric brain tumor measurements: challenges, solutions, and the role of artificial intelligence-based segmentation. *Neuro-Oncology.* 2024;26(9):1557–1571.

11. Veiga-Canuto D, Cerdà-Alberich L, Sangüesa Nebot C, et al. Comparative multicentric evaluation of inter-observer variability in manual and automatic segmentation of neuroblastic tumors in magnetic resonance images. *Cancers.* 2022;14(15):3648.

12. Khalighi S, Reddy K, Midya A, et al. Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *npj Precis Oncol.* 2024;8(1):80.

13. Liu Z, Tong L, Chen L, et al. Deep learning based brain tumor segmentation: a survey. *Complex & Intelligent Systems.* 2022;9:1001–1026.

14. Boyd A, Ye Z, Prabhu S, et al. Stepwise transfer learning for expert-level pediatric brain tumor MRI segmentation in a limited data scenario. *Radiol Artif Intell.* 2023;6(4):e230254.

15. Vossough A, Khalili N, Familiar AM, et al. Training and comparison of nnu-net and deepmedic methods for autosegmentation of pediatric brain tumors. *Am J Neuroradiol.* 2024;45(8):1081–1089.

16. Fathi Kazerooni A, 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):vdad027.

17. Liu X, Jiang Z, Roth H, et al. From adult to pediatric: deep learning-based automatic segmentation of rare pediatric brain tumors. 2023;12465(05).

18. Aggarwal M, Tiwari AK, Sarathi MP, Bijalwan A. An early detection and segmentation of brain tumor using deep neural network. *BMC Med Inform Decis Mak.* 2023;23(1):78.

19. Kim M-J, Hong EP, Yum M-S, et al. Deep learning-based, fully automated, pediatric brain segmentation. *Sci Rep.* 2024;14(1):4344.

20. Hoopes A, Mora JS, Dalca AV, Fischl B, Hoffmann M. SynthStrip: skull-stripping for any brain image. *Neuroimage.* 2022;260:119474.

21. Kalavathi P, Prasath VB. Methods on skull stripping of MRI head scan images-a review. *J Digit Imaging.* 2016;29(3):365–379.

22. Fatima A, Shahid AR, Raza B, Madni TM, Janjua UI. State-of-the-art traditional to the machine- and deep-learning-based skull stripping techniques, models, and algorithms. *J Digit Imaging.* 2020;33(6):1443–1464.

23. Hwang H, Rehman HZU, Lee S. 3D U-Net for skull stripping in brain MRI. *Appl Sci.* 2019;9(3):569.

24. Maia R, Miranda A, Geraldo AF, et al. Neuroimaging of pediatric tumors of the sellar region—A review in light of the 2021 WHO classification of tumors of the central nervous system. *Front Pediatr.* 2023;11:1162654.

25. Al-Bader D, Hasan A, Behbehani R. Sellar masses: diagnosis and treatment. *Front Ophthalmol.* 2022;2:970580.

26. Gilligan LA, DeWire-Schottmiller MD, Fouladi M, DeBlank P, Leach JL. Tumor response assessment in diffuse intrinsic pontine glioma: comparison of semiautomated volumetric, semiautomated linear, and manual linear tumor measurement strategies. *AJNR Am J Neuroradiol.* 2020;41(5):866–873.

27. Fangusaro J, Witt O, Hernáiz Driever P, et al. Response assessment in paediatric low-grade glioma: recommendations from the Response Assessment in Pediatric Neuro-Oncology (RAPNO) working group. *Lancet Oncol.* 2020;21(6):e305–e316.

28. 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.

29. Lee W-K, Yang H-C, Lee C-C, et al. Lesion delineation framework for vestibular schwannoma, meningioma and brain metastasis for gamma knife radiosurgery using stereotactic magnetic resonance images. *Comput Methods Programs Biomed.* 2023;229:107311.

30. Lee W-K, Wu C-C, Lee C-C, et al. Combining analysis of multi-parametric MR images into a convolutional neural network: precise target delineation for vestibular schwannoma treatment planning. *Artif Intell Med.* 2020;107:101911.

31. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods.* 2020;18(2):203–211.

32. Kharaji M, Abbasi H, Orouskhani Y, Shomalzadeh M, Kazemi F, Orouskhani M. Brain tumor segmentation with advanced nnU-Net: pediatrics and adults tumors. *Neurosci Informat.* 2024;4(2):100156.

33. Boer M, Kos TM, Fick T, et al. NnU-Net versus mesh growing algorithm as a tool for the robust and timely segmentation of neurosurgical 3D images in contrast-enhanced T1 MRI scans. *Acta Neurochir.* 2024;166(1):92.

34. Lilly JV, Rokita JL, Mason JL, et al. The children's brain tumor network (CBTN) - Accelerating research in pediatric central nervous system tumors through collaboration and open science. *Neoplasia (New York, N.Y.).* 2023;35:100846.

35. Kazerooni AF, Khalili N, Liu X, et al. BraTS-PEDs: results of the Multi-Consortium International Pediatric Brain Tumor Segmentation Challenge 2023. *arXiv preprint arXi.v* 2024:2407.08855.

36. Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A. The SRI24 multi-channel atlas of normal adult human brain structure. *Hum Brain Mapp.* 2010;31(5):798–819.

37. Davatzikos C, Rathore S, Bakas S, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *J. Med. Imaging (Bellingham, Wash.).* 2018;5(1):011018.

38. Pati S, Singh A, Rathore S, et al. The Cancer Imaging Phenomics Toolkit (CaPTk): technical overview. *BrainLes (Workshop).* 2020;11993:380–394.

39. Thakur SP, Doshi J, Pati S, et al. Skull-stripping of glioblastoma MRI scans using 3D deep learning. *BrainLes (Workshop).* 2019;11992:57–68.

40. Yushkevich PA, Gao Y, Gerig G. ITK-SNAP: an interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: Conference proceedings:... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference. 2016;2016.

41. Chen JV, Li Y, Tang F, et al. Automated neonatal nnU-Net brain MRI extractor trained on a large multi-institutional dataset. *Sci Rep.* 2024;14:4583.

42. Vafaeikia P, Wagner MW, Hawkins C, et al. MRI-based end-to-end pediatric low-grade glioma segmentation and classification. *Can Assoc Radiol J.* 2024;75(1):153–160.

43. Peng J, Kim DD, Patel JB, et al. Deep learning-based automatic tumor burden assessment of pediatric high-grade gliomas, medulloblastomas, and other leptomeningeal seeding tumors. *Neuro-Oncol.* 2021;23(12):2124–2124.

44. Chrysochoou D, Familiar A, Gandhi D, et al. IMG-08. Synthesizing missing MRI sequences in pediatric brain tumors using generative adversarial networks; towards improved volumetric tumor assessment. *Neuro-Oncol.* 2024;26(supplement 4).