Title

Stepwise Transfer Learning for Expert-level Pediatric Brain Tumor MRI Segmentation in a Limited Data Scenario.

Journal

Radiology: Artificial Intelligence, 6(4)

Authors

Boyd, Aidan
Ye, Zezhong
Prabhu, Sanjay
et al.

Publication Date

2024-07-01

DOI

10.1148/ryai.230254

Peer reviewed

Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Stepwise Transfer Learning for Expert-level Pediatric Brain Tumor MRI Segmentation in a Limited Data Scenario

*Aidan Boyd, PhD** • *Zezhong Ye, PhD** • *Sanjay P. Prabhu, MBBS* • *Michael C. Tjong, MD* • *Yining Zha, BS* • *Anna Zapaishchykova, MS* • *Sridhar Vajapeyam, PhD* • *Paul J. Catalano, PhD* • *Hasaan Hayat, BS* • *Rishi Chopra, BS* • *Kevin X. Liu, MD, DPhil* • *Ali Nabavizadeh, MD* • *Adam C. Resnick, PhD* • *Sabine Mueller, MD, PhD* • *Daphne A. Haas-Kogan, MD, MBA* • *Hugo J. W. L. Aerts, PhD* • *Tina Y. Poussaint, MD* • *Benjamin H. Kann, MD*

From the Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, Mass (A.B., Z.Y., Y.Z., A.Z., H.H., R.C., H.J.W.L.A., B.H.K.); Department of Radiation Oncology (A.B., Z.Y., M.C.T., Y.Z., A.Z., H.H., R.C., K.X.L., D.A.H.K., H.J.W.L.A., B.H.K.) and Department of Radiology (H.J.W.L.A.), Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, 75 Francis St, Boston, MA 02115; Department of Radiology, Boston Children's Hospital, Harvard Medical School, Boston, Mass (S.P.P., S.V., T.Y.P.); Department of Biostatistics and Computational Biology, Harvard T.H. Chan School of Public Health, Boston, Mass (P.J.C.); Center for Data-Driven Discovery in Biomedicine (D3b) (A.N., A.C.R.) and Department of Neurosurgery (A.C.R.), Children's Hospital of Philadelphia, Philadelphia, Pa; Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pa (A.N.); Departments of Neurology, Pediatrics, and Neurologic Surgery, University of California, San Francisco, San Francisco, Calif (S.M.); and Department of Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, the Netherlands (H.J.W.L.A.). Received July 11, 2023; revision requested October 5; revision received May 6, 2024; accepted June 18. **Address correspondence to** B.H.K. (email: *Benjamin_Kann@dfci.harvard.edu*).

* A.B. and Z.Y. contributed equally to this work.

Supported in part by the National Institutes of Health (grant nos. U54CA274516, U24CA194354, U01CA190234, U01CA209414, R35CA22052, and K08DE030216), the National Cancer Institute Specialized Programs of Research Excellence (SPORE) grant (grant no. 2P50CA165962), the European Research Council (grant no. 866504), the Radiological Society of North America (grant no. RSCH2017), the Pediatric Low-Grade Astrocytoma Program at the Pediatric Brain Tumor Foundation, Botha-Chan Low Grade Glioma Consortium, and the William M. Wood Foundation. A.B. is supported by Brigham and Women's Hospital, Harvard Medical School, Dana-Farber Cancer Institute, and Boston Children's Hospital. S.V. is supported by Brigham and Women's Hospital. K.X.L. is supported by the National Institutes of Health Loan Repayment Program (grant no. L40CA264321).

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2024; 6(4):e230254 • https://doi.org/10.1148/ryai.230254 • Content codes: AI PD

Purpose: To develop, externally test, and evaluate clinical acceptability of a deep learning pediatric brain tumor segmentation model using stepwise transfer learning.

Materials and Methods: In this retrospective study, the authors leveraged two T2-weighted MRI datasets (May 2001 through December 2015) from a national brain tumor consortium (n = 184; median age, 7 years [range, 1–23 years]; 94 male patients) and a pediatric cancer center (n = 100; median age, 8 years [range, 1–19 years]; 47 male patients) to develop and evaluate deep learning neural networks for pediatric low-grade glioma segmentation using a stepwise transfer learning approach to maximize performance in a limited data scenario. The best model was externally tested on an independent test set and subjected to randomized blinded evaluation by three clinicians, wherein they assessed clinical acceptability of expert- and artificial intelligence (AI)–generated segmentations via 10-point Likert scales and Turing tests.

Results: The best AI model used in-domain stepwise transfer learning (median Dice score coefficient, 0.88 [IQR, 0.72–0.91] vs 0.812 [IQR, 0.56–0.89] for baseline model; P = .049). With external testing, the AI model yielded excellent accuracy using reference standards from three clinical experts (median Dice similarity coefficients: expert 1, 0.83 [IQR, 0.75–0.90]; expert 2, 0.81 [IQR, 0.70–0.89]; expert 3, 0.81 [IQR, 0.68–0.88]; mean accuracy, 0.82). For clinical benchmarking (n = 100 scans), experts rated AI-based segmentations higher on average compared with other experts (median Likert score, 9 [IQR, 7–9] vs 7 [IQR 7–9]) and rated more AI segmentations as clinically acceptable (80.2% vs 65.4%). Experts correctly predicted the origin of AI segmentations in an average of 26.0% of cases.

Conclusion: Stepwise transfer learning enabled expert-level automated pediatric brain tumor autosegmentation and volumetric measurement with a high level of clinical acceptability.

Supplemental material is available for this article.

Pediatric brain tumors are the most common solid malignancies in children, and of these, pediatric low-grade gliomas (pLGGs) are the most prevalent (1). pLGGs are heterogeneous in their molecular underpinnings, natural history, and aggressiveness, making management decisions challenging (2–4). Optimal risk stratification, response assessment, and surveillance for pLGG hinge on the ability to accurately localize, measure, and characterize brain tumors on MR images, which in turn relies on accurate tumor segmentation. Accurate tumor segmentation would enable real-time practical volumetric assessment as well as serve as an important localizer for radiomics analyses and image classification models.

Manual segmentation of pediatric brain tumors is time-consuming, labor intensive, and requires specialized expertise. Given these inherent limitations, there has been interest in developing autosegmentation tools for pediatric brain tumors (5–7). Advances in medical imaging techniques and computational methods have led to various approaches for brain tumor segmentation (7–9). Recently, deep learning for medical imaging has emerged, offering solutions to diverse clinical challenges (10–16). Deep learning–based

Abbreviations

AI = artificial intelligence, BraTS = Brain Tumor Segmentation dataset, CBTN = Children's Brain Tumor Network, DFCI/BCH = Dana-Farber Cancer Institute/Boston Children's Hospital, DSC = Dice score coefficient, FLAIR = fluid-attenuated inversion recovery, pLGG = pediatric low-grade glioma, RVD = relative volume difference

Summary

A deep learning MRI-based autosegmentation model for pediatric low-grade glioma that was developed and externally tested using a stepwise transfer learning approach demonstrated comparable performance and clinical acceptability with pediatric neuroradiologists and radiation oncologists.

Key Points

- Stepwise transfer learning demonstrated gains in performance for deep learning segmentation of pediatric low-grade glioma at T2-weighted MRI (median Dice score coefficient, 0.88 [IQR, 0.72–0.91]) compared with other methodologies and yielded segmentation performance comparable to human experts at external testing.
- For blinded clinical acceptability testing, the model received a higher average Likert score rating and a greater proportion of clinically acceptable segmentations compared with experts (transfer-encoder model [80.1%] vs average expert [65.4%]).
- Turing tests showed uniformly low ability of experts' ability to correctly identify the origin of transfer-encoder model segmentations as artificial intelligence generated versus human generated (mean accuracy, 26%).

Keywords

Stepwise Transfer Learning, Pediatric Brain Tumors, MRI Segmentation, Deep Learning

autosegmentation is a promising approach for accurate and efficient brain tumor segmentation, including pediatric tumors (5,17,18), though distinct challenges remain. With less than 2000 annual average cases (19), pLGGs are relatively rare tumors, and there are no publicly available datasets for training models. Most brain tumor segmentation algorithms have been developed for adult glioma, which are much more common and have large volumes of public and institutional data for training (20,21). In contrast, there has been only limited study of dedicated pediatric glioma segmentation, with a paucity of pLGG-specific models that rely on small single-institution datasets that have not been externally tested nor subjected to clinical testing (17,18). Human clinical evaluation of segmentation models is essential to benchmark performance to experts and determine their true level of performance and potential for clinical translation.

Recently, advances have been made in knowledge transfer learning (22) and self- (23) and semisupervision (24) as methods to improve deep learning performance in limited-data scenarios. These techniques have shown promise in improving medical image analysis algorithms. However, they can be challenging to implement and have not yet been applied to pediatric brain tumors. Pediatric brain tumors represent an ideal setting for applying these techniques, given the need to maximize performance with relatively scarce data. Here, we aim to bridge the translational gap for pediatric brain tumor segmentation algorithms and achieve clinically acceptable performance in a limited-data scenario by leveraging stepwise transfer learning, external testing, and randomized blinded human acceptability testing.

Materials and Methods

Study Design and Datasets

This study was conducted in accordance with the Declaration of Helsinki guidelines and following the approval of the local institutional review board. Waiver of consent was obtained from the institutional review board prior to research initiation due to use of public datasets and the retrospective nature of the study. This report adheres to the Checklist for Artificial Intelligence in Medical Imaging guidelines (25). Data from one national consortium (Children's Brain Tumor Network [CBTN]) and one high-volume academic institution (Dana-Farber Cancer Institute/Boston Children's Hospital [DFCI/BCH]) from May 2001 through December 2015 were included. We decided to develop a T2-weighted MRI segmentation model given that this is the most commonly used sequence used for pLGG volumetrics and was the most consistently available. Scans were subject to manual quality control, removing scans with substantial artifact and/or poor image quality. Patient inclusion criteria were the following: *(a)* 0–25 years of age, *(b)* histopathologically confirmed pLGG, and *(c)* availability of preoperative brain MRI with a T2-weighted imaging sequence. Spinal cord tumors were not considered for this study. Of the 212 scans available from the CBTN as of data abstraction, 184 (88%) were included in this study as a development dataset. One hundred pretreatment MRI acquisitions were selected at random from the DFCI/BCH pLGG clinic patients for validation studies. Adult glioma MRI acquisitions (n = 1251) with expert-generated segmentation masks were acquired from the 2021 Brain Tumor Segmentation Challenge (BraTS) (26–28) to support transfer learning. All BraTS patients (n = 1251) in this study have been analyzed in previous publications, yet none of these previous studies have focused on transfer learning for pLGG. A subset of CBTN patients (n = 140) was previously analyzed in another study (5). However, this previous study incorporated four MRI sequences, whereas our study focused solely on T2-weighted images.

MRI Preprocessing

MR images were converted from Digital Imaging and Communications in Medicine (DICOM) format to Neuroimaging Informatics Technology Initiative (NIfTI) format via rasterization packages using the dcm2nii package in Python v3.8 (Python Software Foundation). N4 bias field correction was adopted to correct the low-frequency intensity nonuniformity present on MR images using SimpleITK in Python v3.8. All scans were resampled to 1 × 1 × 1 mm^3 voxel size using linear interpolation and then coregistered to a pediatric MRI template sourced from the National Institutes of Health MRI Study of Normal Brain Development (NIHPD) Objective 1 Atlases (29) using rigid registration with SimpleITK. Lastly, brain extraction was performed for all the scans using the HD-BET package in Python v3.8 (30).

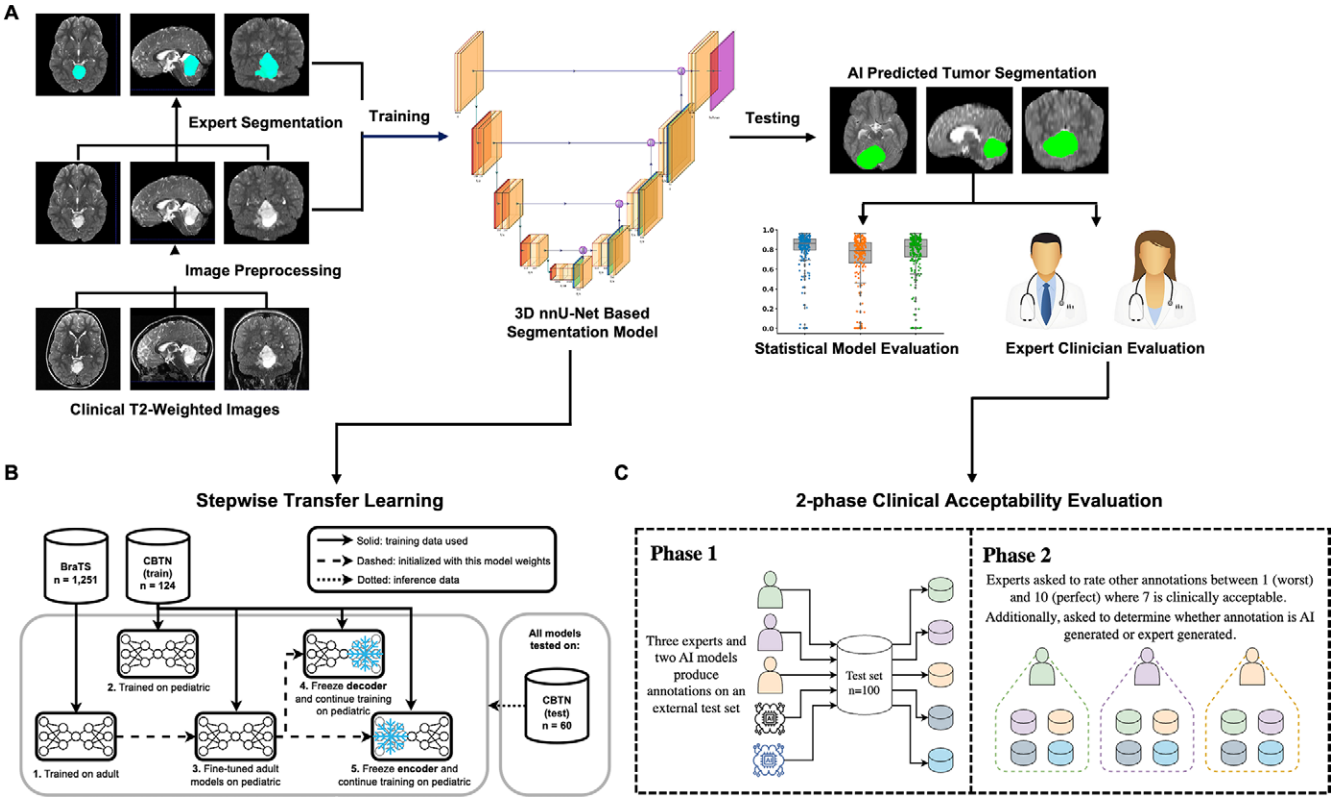

Figure 1: Schematic illustration of the study design. **(A)** An overview of the study workflow, including data preprocessing, expert segmentation of the tumors, model training and testing, and the model clinical acceptability evaluation. Statistical model evaluation includes the primary end point of Dice score coefficient (DSC) and the secondary end point of relative volume change. **(B)** A workflow showing the proposed in-domain stepwise transfer learning with detailed sequential steps involved in this approach. **(C)** A workflow detailing the two-phase clinical acceptability evaluation. AI = artificial intelligence, BraTS = Brain Tumor Segmentation dataset, CBTN = Children's Brain Tumor Network, 3D = three-dimensional.

MRI Review and Segmentation

For model development and initial training, tumors on all T2-weighted scans within the CBTN (n = 184) and 60 scans within the DFCI/BCH cohort were initially segmented by a board-certified radiation oncologist (B.H.K.) to serve as primary reference standard segmentations. To determine segmentation variation across clinical experts of different specializations, a board-certified radiation oncologist specializing in central nervous system tumors (M.J.T.) and a board-certified pediatric neuroradiologist (S.P.P.) independently annotated 100 scans from the DFCI/BCH external dataset for clinical acceptability testing. During a file saving procedure, unexpectedly, nine and 11 scans, respectively, were found to have corrupt data. Thus, 91 and 89 scans, respectively, were included for these two annotators in the analysis. All the annotators were instructed to segment all areas of T2-weighted signal abnormality concerning for tumor involvement, including areas of peritumoral T2-weighted hyperintensity if suspicious for tumor involvement. Segmentations were performed and saved in NIfTI format using ITK-SNAP v4.0 *(http://www.itksnap.org)* with three-dimensional axial, sagittal, and coronal views (Fig S1).

Deep Learning Approach: Stepwise In-Domain Transfer Learning

The baseline model used the nnU-Net architecture (31) with built-in ensembling for training and inference. The nnUNet architecture was used due to the proven performance on many medical imaging tasks as well as its general-purpose nature. Early stopping was implemented, ending training if there was no improvement on the validation set for 50 epochs, with a maximum of 1000 epochs. All other training parameters include learning rate, batch size, data augmentations, and loss function, followed the default settings of nnU-Net (31). Models were trained with the nnU-Net built-in fivefold cross-validation procedure and with the final model consisting of an ensemble of the five folds. Prior to training, we randomly split the CBTN development set (n = 184) into a training set (n = 124; 67%) to be used for the fivefold cross-validation procedure and a blinded internal hold-out test set (n = 60; 33%). We investigated five model frameworks as specified below, and each approach was evaluated and compared via performance on the CBTN hold-out test set (n = 60) (Fig 1A). The top-performing model was selected for further testing using the external test set from DFCI/BCH (n = 100).

Adult brain tumor model: BraTS model.— To determine if an adult glioma model could perform well in the pediatric setting out of the box, we trained the nnU-Net framework using the BraTS 2021 training set (n = 1251) and the internal hold-out set from CBTN (n = 60).

Training from scratch: scratch model.— To investigate the hypothesis that a model trained on pediatric brain tumors would

outperform a model trained on adult tumors on a pediatric test set, an nnU-Net model was trained from scratch using the CBTN training set (n = 124), which consisted of limited pediatric data and required less training data compared with the BraTS model. As specified above, we conducted a standard five-fold cross-validation procedure using 20% of the training data (n = 24) as validation for each cross-validation fold, with the final model representing an ensemble of the five folds.

In-domain transfer learning from adults: transfer model.— We hypothesized that fine-tuning a BraTS model with pediatric data (ie, in-domain transfer learning) may improve performance. For this experiment, the nnU-Net was pretrained with the BraTS model weights (n = 1251) and then fine-tuned using additional training on the CBTN data (n = 124) following the same procedure as the scratch model.

Stepwise transfer learning.— We hypothesized that freezing specific model parameters during fine-tuning could further enhance convergence and reduce overfitting in this limited data scenario by reducing the number of parameters for optimization. To test this, starting with the transfer model checkpoint, we further fine-tuned the model while freezing either the encoder block (transfer-encoder model) or the decoder block (transfer-decoder model) (Fig 1B).

Model Evaluation and Statistical Analysis

The primary performance end point was the Dice score coefficient (DSC) (32), with a DSC value greater than 0.80 indicating suitability for further clinical testing (9). Median DSC values were compared between models using Wilcoxon rank sum tests on the CBTN test set. The highest performing model on the CBTN set was externally tested on the DFCI/BCH dataset. Furthermore, the relative volume difference (RVD) was calculated by dividing the volume difference between the predicted and reference standard tumor volumes by the reference standard tumor volume. Other secondary end points also included aggregated DSC and intraclass coefficient for tumor volumes.

Randomized Blinded Clinical Acceptability Testing and Interexpert Variability

While DSC is an important quantitative measure of segmentation performance, positioning the algorithm for real-world use requires clinical validation and benchmarking (33,34). Pairwise interexpert variability between the three annotators as well as two models (BraTS model and transfer-encoder model) was evaluated with DSC to determine if model performance was comparable to interexpert variability. To assess the clinical utility of the artificial intelligence (AI) models, the three experts conducted a blinded segmentation rating and acceptability study (Fig 1C) as follows: For each of the 100 DFCI/BCH cases, each expert was presented with three different segmentations overlaid with each other (Appendix S1). The three segmentations consisted of one to two expert segmentations (from the other two annotators) and one to two AI-generated segmentations (selected from the BraTS model and/or the transfer-encoder model) selected at random (n = 300 total segmentations per reviewer). The expert raters were blinded to the origin of the segmentations. The order and color of the segmentations displayed for each scan was randomized to reduce bias. The ratings were carried out using SegmentationReview (35). Experts were given written instructions and asked to rate each of the three segmentations on a scale from 1 (worst) to 10 (perfect), with a 7 being defined as a clinically acceptable threshold for volumetric assessment (Fig S2). For each segmentation, raters were also asked to guess whether the segmentation was AI generated (ie, a Turing test). For comparing paired measurements while appropriately addressing the dependency between measurements on the same cases, we analyzed the mean interrater group DSC values using the Wilcoxon signed rank test. Expert rating scores underwent analysis using the Kruskal-Wallis test, followed by Mann-Whitney U tests as post hoc tests. Additionally, the clinical acceptability and Turing test results were subjected to analysis using the χ^2 test. To adjust for multiple group analysis, we employed the Benjamini-Hochberg correction to correct P values. Two-sided tests with a significance level of $P < .05$ were considered statistically significant. Statistical analyses were conducted using the Statsmodels 0.14.1 and the SciPy 1.11.0 packages in Python 3.8.

Data and Code Availability

BraTS data including raw MR images may be requested from The Cancer Image Archive *(https://www.med.upenn.edu/cbica/brats/)*. Although raw MRI data cannot be shared, all measured results to replicate the statistical analysis are shared at the GitHub web page *(https://github.com/AIM-KannLab/pLGG_Segmentation)*. Furthermore, we include test samples from a publicly available dataset with deep learning and expert reader annotations. The code of the deep learning system as well as the trained model and statistical analysis are publicly available.

Results

Patient Characteristics

The pLGG cohort included 284 patients from two cohorts, 184 patients in the CBTN development set and 100 patients in the DFCI/BCH external test set (Table 1). In the CBTN cohort, the median age was 7 years (range, 1–23 years), with 84 (45.7%) female and 94 (51.1%) male patients. In the DFCI/BCH cohort, the median age was 8 years (range, 1–19 years), with 53 (53%) female and 47 (47%) male patients. All patients had pathologically diagnosed grade I or II low-grade glioma, with various histologic subtypes and tumor locations. MRI scan parameters across datasets are found in Table S3 and S4.

In-Domain Stepwise Transfer Learning Improves Segmentation Performance

The BraTS model exhibited high performance on adult data but significantly declined when applied to pediatric data (median DSC, 0.93 [IQR, 0.89–0.95] to median DSC, 0.81 [IQR,

Table 1: Patient Demographics

Parameter	Patient Cohorts	
	CBTN (n = 184)	DFCI/BCH (n = 100)
Age (y)		
Median (range)	7 (1–23)	8 (1–19)
Sex		
Female	84 (45.7)	53 (53)
Male	94 (51.1)	47 (47)
Unknown	6 (3.2)	0 (0)
Race or ethnicity		
Non-Hispanic White	118 (64.1)	68 (68)
African American/Black	24 (13.0)	5 (5)
Hispanic/Latinx	17 (9.2)	4 (4)
Asian American/Asian	3 (1.6)	4 (4)
Other/unknown	22 (12)	19 (19)
Histologic diagnosis		
Pilocytic astrocytoma	58 (31.5)	34 (34)
Pilomyxoid astrocytoma	10 (5.4)	0 (0)
Juvenile pilocytic astrocytoma	0 (0)	13 (13)
Ganglioglioma	1 (0.5)	11 (11)
Oligodendroglioma	1 (0.5)	0 (0)
Diffuse astrocytoma	9 (4.9)	3 (3)
Fibrillary astrocytoma	13 (7.1)	0 (0)
Optic pathway glioma	0 (0)	3 (3)
Other low-grade glioma or astrocytoma	92 (50.0)	36 (36)
Primary tumor location		
Posterior fossa	48 (26.1)	28 (28)
Temporal lobe	13 (7.1)	18 (18)
Frontal lobe	7 (3.8)	2 (2)
Cerebellum	0 (0)	18 (18)
Suprasellar	27 (14.7)	6 (6)
Optic pathway	27 (14.7)	3 (3)
Brainstem	23 (12.5)	3 (3)
Thalamus	7 (3.8)	2 (2)
Others	32 (17.4)	20 (20)

Note.—Unless otherwise noted, data are reported as numbers of patients with percentages in parentheses. CBTN = Children's Brain Tumor Network, DFCI/BCH = Dana-Farber Cancer Institute/Boston Children's Hospital.

0.56–0.89]; $P < .001$) (Fig 2; Table 2). Volumetric assessment accuracy also decreased significantly on pediatric data (median RVD, 5.2% [IQR, 2.4%–11.9%] to median RVD, 19.2% [IQR, 10.9%–68.2%]; $P < .001$). All scratch and transfer models surpassed the BraTS model in terms of DSC, with the transfer-encoder model exhibiting a significant difference ($P = .049$), achieving a DSC value of 0.88 [IQR, 0.72–0.91] compared with the DSC value of the BraTS model of 0.81 [IQR, 0.56–0.89]. Among the five models, the transfer-encoder model demonstrated the highest DSC values (median DSC, 0.88 [IQR, 0.72–0.91]; aggregated DSC, 0.84), alongside the RVD value (median RVD, 10.9% [IQR, 3.2%–31.0%]), as well as the lowest rate of failed or empty segmentations (6.7% [four of 60]) (Table 2). Representative cases in Figure 3 highlight gains in accuracy of transfer-encoder compared with other models. Given improvements across several metrics (Table 2), we selected the transfer-encoder model for further testing.

External Testing of Stepwise Transfer Learning

On an external testing set with expert segmentations (n = 60 scans from DFCI/BCH), the transfer-encoder model achieved a median DSC value of 0.83 (IQR, 0.74–0.90) and median RVD of 16.1% (IQR, 5.8%–39.3%) as compared with manual segmentations. We performed failure analysis for cases with a DSC value less than 0.6 and found six cases in total. The failures were caused by the following factors: *(a)* tumor located in ventricle (Fig S3E; n = 1), *(b)* large cystic area in brain (Fig S3A, S3C, S3D; n = 3), *(c)* empty segmentation from poor image quality due to respacing for large section thickness (Fig S3B; n = 1), and (*d*) under-segmentations for large heterogeneous tumor lesion (Fig S3F; n = 1).

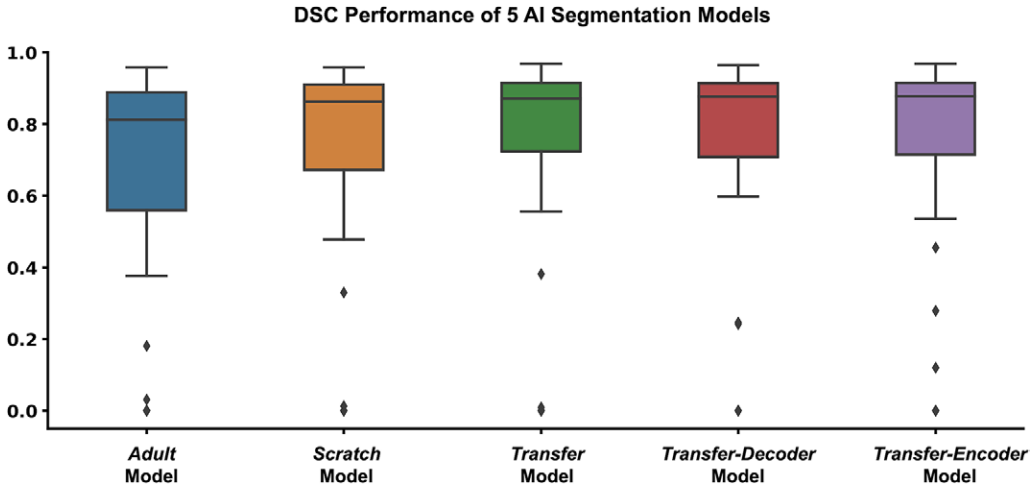

Figure 2: Graph shows the comparative performance of deep learning training methodologies on the internal test set (*n* = 60). Among the five different segmentation models assessed, the methods using stepwise transfer learning (transfer-decoder and transfer-encoder) had the highest segmentation accuracy. Additionally, the transfer-encoder model generated the fewest segmentations with a Dice score coefficient (DSC) of 0 (*n* = 4; 6.7%) (indicating a complete segmentation miss). Conversely, the brain tumor segmentation (Brain Tumor Segmentation dataset [BraTS]) model exhibited the highest number of segmentations with a DSC value of 0 (*n* = 11; 18.3%). The transfer-encoder model demonstrated the highest median DSC value (0.88 [IQR, 0.72–0.91]) and was selected for further investigation. The BraTS model, trained only on adult glioma, demonstrated the lowest median DSC (0.81 [IQR, 0.56–0.89]). AI = artificial intelligence.

Table 2: Model Performance on Internal Test Set from the Children Brain Tumor Consortium

Model	Median DSC	Aggregated DSC	Median RVD (%)	ICC (95% CI)	Percentage of Cases with a DSC Value of 0
BraTS model	0.81 (0.56–0.89)	0.73	19.2 (10.9–68.2)	0.74 (0.59–0.83)	15.0 (9/60)
Scratch model	0.86 (0.67–0.91)	0.82	9.8 (4.0–40.7)	0.80 (0.69–0.88)	10.0 (6/60)
Transfer model	0.87 (0.72–0.91)	0.82	12.4 (3.6–33.4)	0.80 (0.69–0.88)	8.3 (5/60)
Transfer-decoder model	0.88 (0.71–0.91)	0.83	12.6 (5.2–28.9)	0.83 (0.74–0.90)	8.3 (5/60)
Transfer-encoder model	0.88 (0.72–0.91)	0.84	10.9 (3.2–31.0)	0.83 (0.73–0.89)	6.7 (4/60)

Note.—Unless otherwise noted, values are medians with IQRs in parentheses or percentages with numerators and denominators in parentheses. BraTS = Brain Tumor Segmentation dataset, DSC = Dice score coefficient, RVD = relative volume difference, ICC = intraclass correlation coefficient.

Clinical Acceptability Testing of Stepwise Transfer Learning Model

Interannotator DSC results between experts and AI models were calculated using segmentations from expert 1, expert 2, and expert 3 as the reference standard on the external test set (Fig 4A; *n* = 60). Based on expert 1, the transfer-encoder model exhibited no significant differences in DSC values compared with both expert 2 and expert 3 (Fig 4B; transfer-encoder model vs expert 2: median, 0.83 [IQR, 0.75–0.90] vs 0.87 [IQR, 0.79–0.90]; *P* = .11) (transfer-encoder model vs expert 3: median, 0.83 [IQR, 0.75–0.90] vs 0.850 [IQR, 0.81–0.90]; *P* = .09). Notably, transfer-encoder, expert 2, and expert 3 all demonstrated significantly higher DSC values than BraTS (*P* = .02, *P* = .001, *P* = .005, respectively). With expert 2 (Fig 4C) and expert 3 (Fig 4D) as reference standards, transfer-encoder exhibited significantly lower DSC values compared with the interexpert variability but displayed significantly higher DSC values compared with the BraTS model.

For clinical acceptability testing, the overall rating scores for segmentations for transfer-encoder (median, 9 [IQR, 7–9]) were higher than those for BraTS (median, 8 [IQR, 6–9]), expert 1 (median, 7 [IQR, 6–9]), expert 2 (median, 8 [IQR, 7–9]), and expert 3 (median, 6 [IQR, 5–8]). Based on individual experts, the mean segmentation quality scores for the transfer-encoder model were consistently higher than or comparable to those of the experts and notably higher than those of the BraTS model (Fig 5A). For instance, according to expert 2, the transfer-encoder model (median, 8 [IQR, 6–9]) exhibited significantly higher mean scores compared with expert 1 (median, 6 [IQR, 5–7]; *P* = .003), expert 3 (median, 5 [IQR, 5–6]; *P* < .001), and the BraTS model (median, 7 [IQR, 5–8]; *P* < .001). Overall, transfer-encoder exhibited a higher proportion of clinically acceptable (rating score ≥ 7) segmentations (80.1% [182 of 227]) compared with BraTS (72.1% [165 of 229]) and individual experts (expert 1, 68.3% [82 of 120]; expert 2, 78.7% [122 of 155]; expert 3, 49.3% [66 of 134]). Based on the individual expert, these percentages for transfer-encoder were comparable

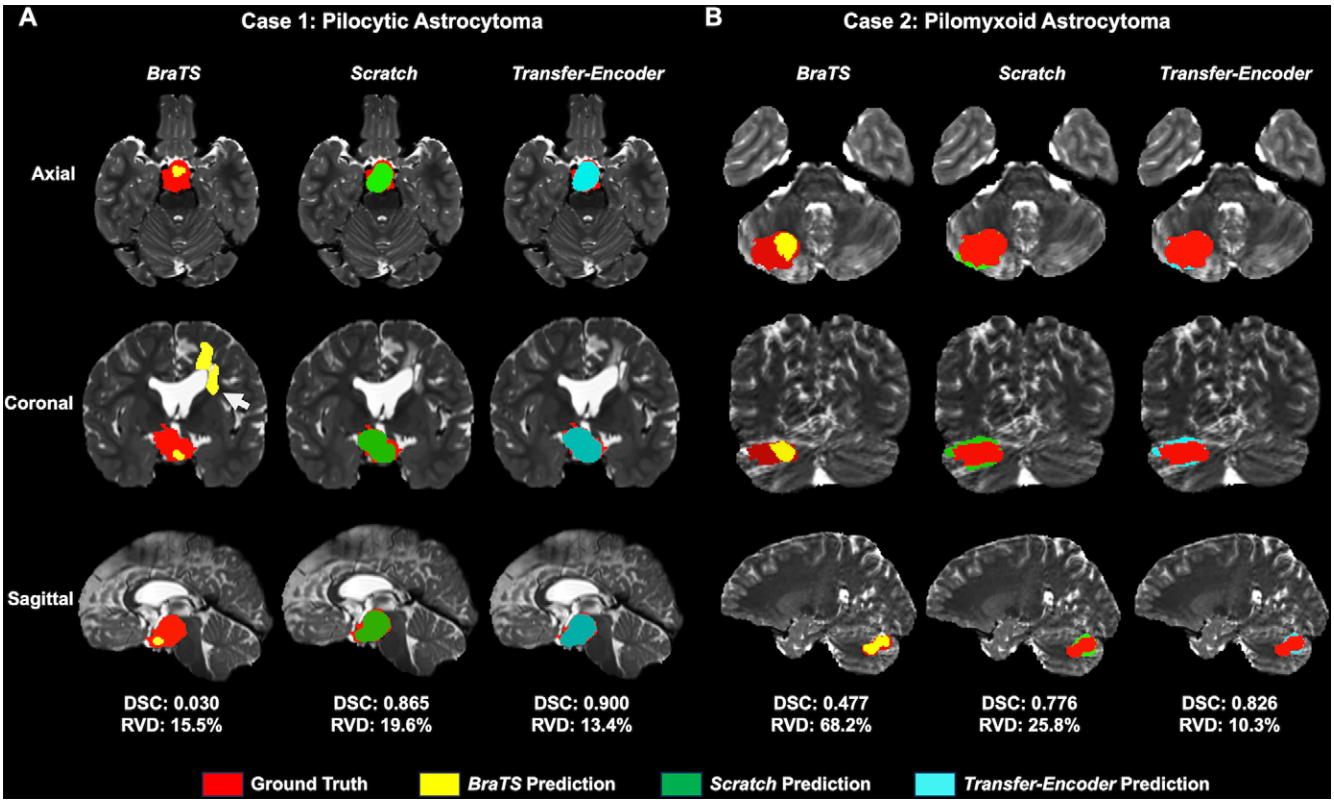

Figure 3: Expert tumor segmentations and AI-predicted tumor segmentations on T2-weighted images by the Brain Tumor Segmentation dataset (BraTS), scratch, and transfer-encoder models from two representative cases. **(A)** In the case of pilocytic astrocytoma from an 8-year-old boy, the transfer-encoder model exhibited excellent tumor segmentation performance, achieving a Dice score coefficient (DSC) value of 0.90 and a volume difference of 13.4%, surpassing the results of BraTS model (DSC, 0.03; volume difference, 15.5%) and the scratch model (DSC, 0.87; volume difference, 19.6%). Notably, the BraTS model exhibited a small volume difference but a very low DSC value due to false-positive predictions on the vasogenic edema (arrow). **(B)** In another representative case of pilomyxoid astrocytoma from a 17-year-old adolescent boy, the transfer-encoder model also demonstrated a higher DSC value (0.83) and lower volume difference (10.3%) compared with the BraTS model (DSC, 0.477; volume difference, 68.2%) and the scratch model (DSC, 0.78; volume difference, 25.8%). All *P* values were corrected to adjust for multiple group analysis with the Benjamini-Hochberg correction. RVD = relative volume difference.

to or higher than those from other experts. For example, from expert 1, transfer-encoder (82.9% [58 of 70]) demonstrated slightly higher percentages of clinically acceptable segmentations compared with expert 2 (82.4% [61 of 74]; $P > .99$) and expert 3 (79.4% [50 of 63]; $P > .99$). However, according to expert 2, the transfer-encoder model exhibited a higher percentage of clinically acceptable segmentations (72.0% [54 of 75]) compared with those from expert 1 (48.3% [29 of 60]; $P = .009$), expert 3 (21.6%, [16 of 71]; $P < .001$), and the BraTS model (56.8% [46 of 81]; $P = .07$) (Fig 5B). Furthermore, results from the Turing test revealed a consistent challenge for experts in distinguishing AI-generated from expert-generated segmentations. Specifically, expert 1, 2, and 3 identified transfer-encoder segmentations as AI generated in only 22.9% (16 of 70), 36% (27 of 75), and 19.5% (16 of 82) of scans, respectively, which are notably lower than those from the experts and the BraTS model (Fig 5C). Conversely, a large percentage of experts' segmentations were judged to be AI generated by all three experts. For instance, expert 2 rated 88.3% (53 of 60) of segmentations from expert 1 and 90.1% (64 of 71) of segmentations from expert 3 as AI generated, while expert 3 rated 41.7% (25 of 60) of segmentations from expert 1 and 39.5% (32 of 81) of segmentations from expert 2 as AI generated. The median Likert score and clinical acceptability rate was 9 (IQR, 7–9) and 80.2%, respectively, for transfer-encoder segmentations and 7 (IQR, 7–9) and 65.4%, respectively, for other experts' segmentation.

Discussion

In this study, we developed, externally tested, and clinically benchmarked a deep learning pipeline using stepwise transfer learning for automated expert-level pLGG segmentation and volumetric measurement. Stepwise transfer learning yielded high segmentation performance (DSC, 0.88; IQR, 0.72–0.91) that was comparable to interexpert agreement for clinical evaluation. Additionally, experts rated the clinical acceptability of the stepwise transfer learning–based model on par or higher than other experts' segmentations. Turing tests also indicated uniformly low ability of experts' ability to correctly identify the origin of transfer-encoder model segmentations as AI generated versus human generated. Accurate tumor autosegmentation models could be useful for risk stratification, monitoring tumor progression, assessing treatment response, and surgical approach (5). However, tumor autosegmentation models have had limited traction in use for pediatric tumors due to very sparse available training data. We leveraged a strategy of in-domain stepwise transfer learning to demonstrate measurable gains in segmentation accuracy and clinical acceptability that was on par with clinician performance. To our knowledge, this

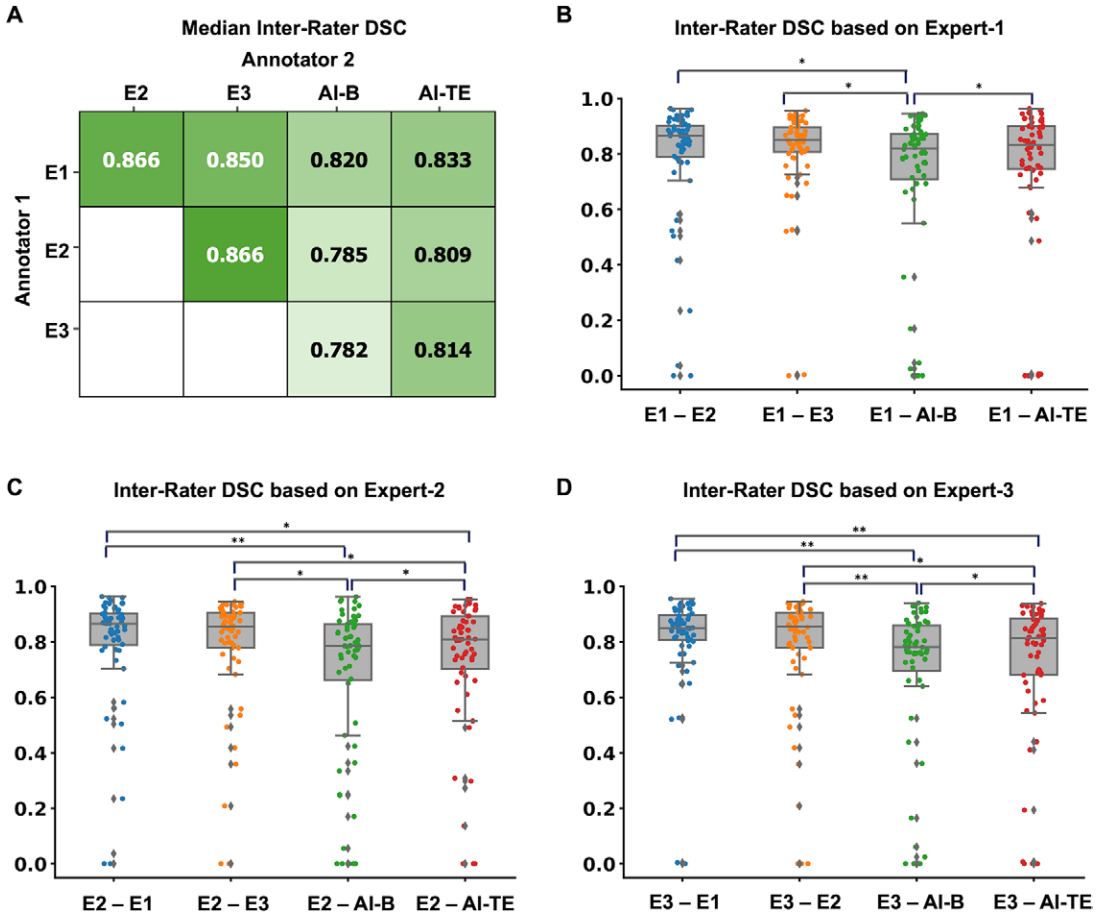

Figure 4: The interrater Dice score coefficient (DSC) values among three clinical experts and two artificial intelligence (AI) models using an external test dataset comprising 60 cases. Overall, the transfer-encoder model exhibits higher DSC values compared with the Brain Tumor Segmentation dataset [BraTS] model but lower DSC values compared with the interrater agreement among different experts. **(A)** The distribution plots depict the median interrater DSC values calculated from the segmentations of three experts and two AI models. **(B)** Boxplots display the interrater DSCs for segmentations from two experts and two AI models, with expert 1's segmentations serving as the reference standard. **(C)** Boxplots illustrate interrater DSC values using expert 2's (E2) segmentations as the reference standard. **(D)** Boxplots demonstrate interrater DSC values using expert 3's (E3) segmentations as the reference standard. E1 = expert 1, AI-B = BraTS model, AI-TE = transfer-encoder model. All *P* values were corrected to adjust for multiple group analysis with the Benjamini-Hochberg correction. Points outside the whiskers represent outliers. * = $P < .05$, ** = $P < .001$.

is the first study to use stepwise transfer learning in this context and to evaluate clinical acceptability of autosegmentation tools. The rigorous clinical benchmarking studies with three blinded experts suggest that this approach nears a performance ceiling for pLGG segmentation (ie, output is comparable and indistinguishable to human experts).

The current state-of-the-art approaches for automated brain tumor segmentation rely on deep learning. However, most available autosegmentation tools have been specifically developed and trained for adult brain cancers, particularly glioblastoma (8,20,21). In this work, we find that tools such as these do not effectively generalize to pediatric brain tumors. Performance degradation may stem from the distinctive heterogeneous imaging appearance and types of pediatric brain tumors compared with adult brain tumors, as well as the anatomic differences resulting from the ongoing brain development in children. Several studies have proposed various deep learning solutions to address the segmentation of pediatric brain tumors, achieving DSC values ranging between 0.68 and 0.88 (5,17,36,37). However, the clinical acceptability of these approaches was not validated or benchmarked against adult models. To date, only one study has proposed an algorithm specifically for pLGG, achieving a DSC value of 0.77 (18). This study utilized fluid-attenuated inversion recovery (FLAIR) images from 311 patients from a single institution. The proposed model employed deep multitask learning, incorporating a genetic alteration classifier of a tumor as an auxiliary task to the main segmentation network (18). However, this model was trained on a limited number of MRI scans from a single institute and lacked external testing and clinical testing. In the present study, our stepwise transfer learning model achieved better model performance compared with previous work (18). Improved performance may be due to sequential knowledge transfer, first from the adult setting, and then the pediatric setting. Additionally, freezing the encoder or decoder in the final fine-tuning step enabled optimization of a smaller parameter space which may have mitigated overfitting given the limited amount of data. Training on a sufficient quantity

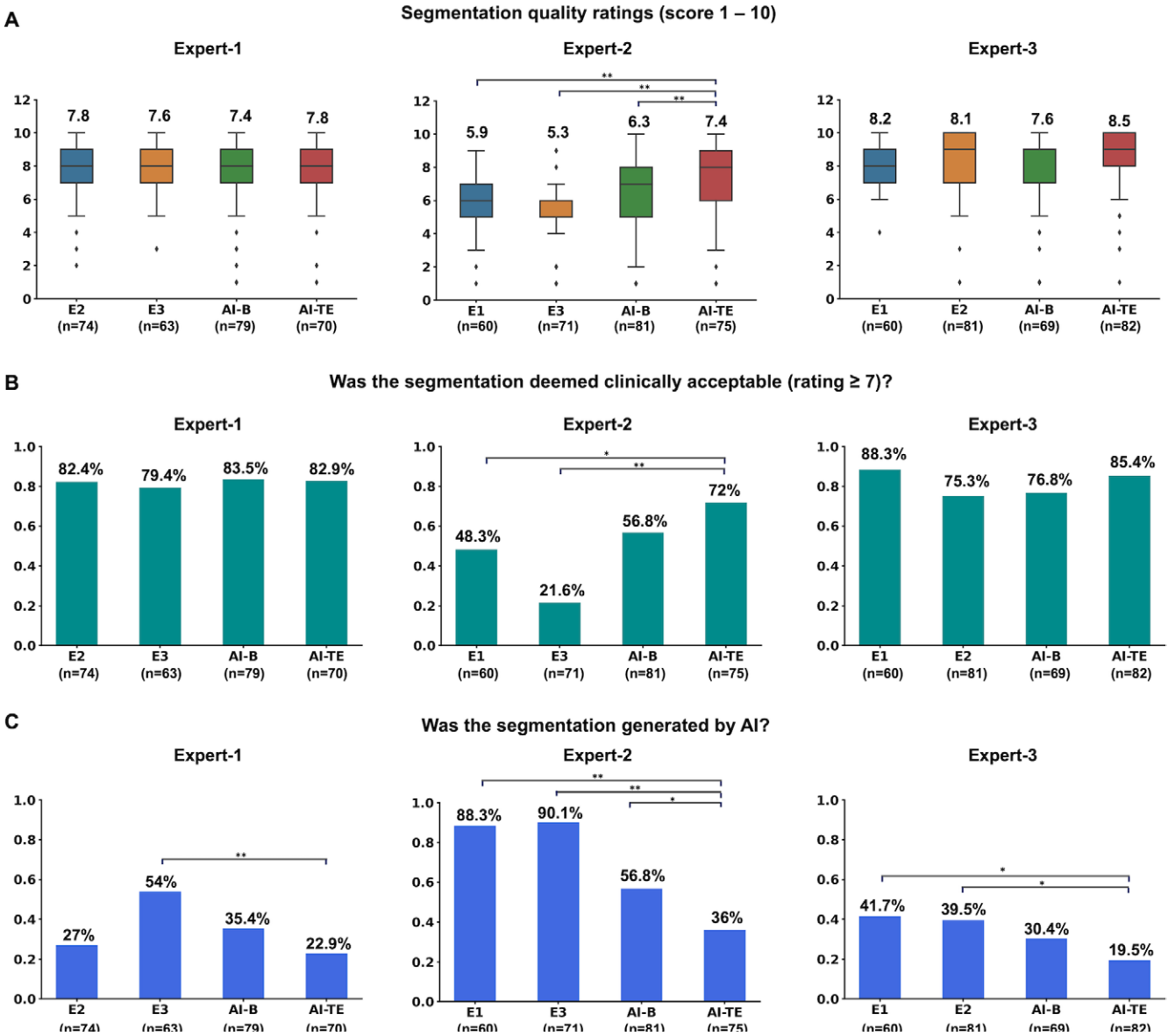

Figure 5: Clinical acceptability testing on segmentations. Three human experts were invited to rate randomized artificial intelligence (AI)–generated (Brain Tumor Segmentation dataset [BraTS] model or transfer-encoder model) or expert-generated segmentations while blinded to segmentation origin. **(A)** The mean segmentation rating scores for the three experts and two AI models, as individually evaluated by expert 1 (E1), expert 2 (E2), and expert 3 (E3). Boxplots are used to compare and group the rating scores for each expert and AI model, along with the corresponding *P* values from statistical tests for group comparisons. According to all three experts, the transfer-encoder model (AI-TE) model consistently shows better or comparable rating scores compared with experts and the BraTS (AI-B) model. **(B)** The percentage of segmentations that were considered acceptable, with a rating score of 7 or above, is summarized for each expert and model. **(C)** The accuracy of determining whether annotations were AI generated is shown for each expert and model, indicating how well the experts were able to distinguish between AI-generated and expert-generated segmentations. All *P* values were corrected to adjust for multiple group analysis with the Benjamini-Hochberg correction. * = $P < .05$, ** = $P < .001$.

of pediatric data from scratch may obviate the need for transfer learning, but what represents "sufficient" is yet to be defined for pLGGs and current datasets remain relatively small.

U-Net architectures, including the widely validated nnU-Net, have become the de facto standard for medical imaging segmentation challenges in recent years (31). While newer, more advanced algorithms have emerged, such as three-dimensional attention U-Net (38), hybrid models like M-Net (39), and Swin U-Net transformers (40). These models feature more sophisticated neural network architectures that typically demand larger datasets for training and are susceptible to overfitting when applied to small datasets. Thus, these models have not shown superiority to nnU-Net in most typical medical imaging challenges (31,41). The more acute problem in medical image segmentation, particularly for pediatric brain tumors, is limited data. Thus, we focused our strategy on ways to improve the performance of nnU-Net in a data-limited scenario, specifically via investigation of various stepwise transfer learning approaches. With this data-centric approach, that additionally leverages multi-institutional data, external testing, and rigorous blinded clinical evaluation, we were able to demonstrate that an nnU-Net–based algorithm can achieve performance on par, or perhaps exceeding that of human experts in pediatric glioma segmentation.

While statistical metrics like the DSC and RVD offer valuable insights into a model's overall segmentation performance, it is important to acknowledge their limitations in providing a comprehensive evaluation of a model's utility (6). To ensure a thorough evaluation and facilitate the clinical translation of our model, we conducted a rigorous clinical acceptability evaluation and validation process involving three expert clinicians. The involvement of expert clinicians provides valuable feedback and insights, accelerating the translation of the model into clinical practice. In our study, we went a step further by conducting blinded, segmentation rating, acceptability, and Turing tests involving the three expert clinicians. Notably, all experts performed worse than random chance (50%) in predicting the origin of the transfer-learned model segmentations, suggesting that this model passes the Turing test. Additionally, this clinical evaluation captures additional nuance regarding clinical utility of different deep learning approaches. For instance, we found that a stepwise transfer learning approach, beyond small increases in DSC, also generated more clinically acceptable segmentations that were indistinguishable from expert segmentations than an adult glioma–based model. To our knowledge, this is the first brain tumor segmentation study to incorporate such a clinical evaluation, which is critical in positioning a model for clinical translation.

There are several limitations of this study. First, the study was retrospective in nature, and selection of scans for inclusion in this study, while performed a priori and based solely on availability and scan quality, may introduce bias. Second, the model utilizes only T2-weighted images, as these were the most commonly available for all the patients included in our analysis. Both T2-weighted and T2 FLAIR sequences are considered the best sequences for tracking low-grade glioma (42), yet T2 FLAIR images were not available for many patients with pLGGs in our study, particularly in the CBTN cohort. Consequently, differentiation between vasogenic edema from tumor may have been more challenging for annotators, though there was still high interannotator segmentation agreement. Despite this limitation, our study demonstrated that stepwise transfer learning is a powerful approach to improve deep learning performance in a data-limited scenario. We encourage other investigators to use our framework in a multiparametric setting. A possible advantage of a T2-weighted–only model is that it may be more widely applicable for volumetric assessment in situations where multiparametric and contrast-enhanced scans are unavailable. Furthermore, it is important to note that our study focused solely on whole tumor segmentation and not the segmentation of specific tumor subregions. Consequently, the clinical utility of our findings may be restricted in certain cases, such as when change in cystic component is not relevant to the clinical response assessment. Finally, it is notable that the algorithm did fail on some cases, and while we identified certain factors that were associated with failures, it is difficult to predict with certainty why a failure occurred owing to the black box nature of deep learning algorithms. Therefore, it is important for the model output to undergo a clinical review prior to use in clinical decision-making.

In conclusion, we developed, externally tested, and clinically benchmarked an automated deep learning pipeline using in-domain stepwise transfer learning that enables expert-level MRI segmentation of pLGGs. With blinded evaluation, the model demonstrated clinically acceptable performance that was higher on average than clinical experts. Prospective and longitudinal evaluation of the pipeline is planned to determine the algorithm's potential for integration into the clinical care of children with low-grade glioma.

Author contributions: Guarantors of integrity of entire study, **Y.Z.**, **D.A.H.K.**, **B.H.K.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **A.B.**, **Z.Y.**, **D.A.H.K.**, **T.Y.P.**, **B.H.K.**; clinical studies, **A.B.**, **S.P.P.**, **M.C.T.**, **H.H.**, **T.P.**, **B.H.K.**; experimental studies, **A.B.**, **Z.Y.**, **S.P.P.**, **M.C.T.**, **H.H.**, **A.R.**, **H.J.W.L.A.**; statistical analysis, **Z.Y.**, **Y.Z.**, **P.J.C.**, **R.C.**, **A.C.R.**, **H.J.W.L.A.**; and manuscript editing, **A.B.**, **Z.Y.**, **S.P.P.**, **M.C.T.**, **A.Z.**, **S.V.**, **H.H.**, **R.C.**, **K.X.L.**, **A.N.**, **A.C.R.**, **S.M.**, **D.A.H.K.**, **H.J.W.L.A.**, **T.Y.P.**, **B.H.K.**

Disclosures of conflicts of interest: **A.B.** No relevant relationships. **Z.Y.** No relevant relationships. **S.P.P.** Direct payments for as expert medical witness for various attorneys across the United States. **M.C.T.** No relevant relationships. **Y.Z.** No relevant relationships. **A.Z.** No relevant relationships. **S.V.** No relevant relationships. **P.J.C.** No relevant relationships. **H.H.** No relevant relationships. **R.C.** No relevant relationships. **K.X.L.** No relevant relationships. **A.N.** No relevant relationships. **A.C.R.** No relevant relationships. **S.M.** No relevant relationships. **D.A.H.K.** No relevant relationships. **H.J.W.L.A.** Grants from the National Institutes of Health and the European Union to institution. **T.Y.P.** No relevant relationships. **B.H.K.** No relevant relationships.

References

1. 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(3 Suppl 3):iii1–iii38.
2. Ryall S, Tabori U, Hawkins C. Pediatric low-grade glioma in the era of molecular diagnostics. Acta Neuropathol Commun 2020;8(1):30.
3. Manoharan N, Liu KX, Mueller S, Haas-Kogan DA, Bandopadhayay P. Pediatric low-grade glioma: Targeted therapeutics and clinical trials in the molecular era. Neoplasia 2023;36:100857.
4. Bandopadhayay P, Bergthold G, London WB, et al. Long-term outcome of 4,040 children diagnosed with pediatric low-grade gliomas: an analysis of the Surveillance Epidemiology and End Results (SEER) database. Pediatr Blood Cancer 2014;61(7):1173–1179.
5. 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.
6. Zhang S, Edwards A, Wang S, Patay Z, Bag AK, Scoggins MA. A Prior Knowledge Based Tumor and Tumoral Subregion Segmentation Tool for Pediatric Brain Tumors. arXiv 2109.14775 [preprint] https://arxiv.org/abs/2109.14775. Posted September 30, 2021. Accessed May 1, 2023.
7. Wels M, Carneiro G, Aplas A, Huber M, Hornegger J, Comaniciu D. A Discriminative Model-Constrained Graph Cuts Approach to Fully Automated Pediatric Brain Tumor Segmentation in 3-D MRI. In: Metaxas D, Axel L, Fichtinger G, Székely G, eds. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008. MICCAI 2008. Lecture Notes in Computer Science, vol 5241. Springer, 2008; 67–75.
8. Jyothi P, Singh AR. Deep learning models and traditional automated techniques for brain tumor segmentation in MRI: a review. Artif Intell Rev 2022;56(4):2923–2969.
9. Liu Z, Tong L, Chen L, et al. Deep learning based brain tumor segmentation: a survey. Complex Intell Systems 2023;9(1):1001–1026.
10. Kann BH, Likitlersuang J, Bontempi D, et al. Screening for extranodal extension in HPV-associated oropharyngeal carcinoma: evaluation of a CT-based deep learning algorithm in patient data from a multicentre, randomised de-escalation trial. Lancet Digit Health 2023;5(6):e360–e369.
11. Kazmierski M, Welch M, Kim S, et al. Multi-institutional Prognostic Modeling in Head and Neck Cancer: Evaluating Impact and Generalizability of Deep Learning and Radiomics. Cancer Res Commun 2023;3(6):1140–1151.
12. Ye Z, Qian JM, Hosny A, et al. Deep Learning-based Detection of Intravenous Contrast Enhancement on CT Scans. Radiol Artif Intell 2022;4(3):e210285.

13. Hosny A, Bitterman DS, Guthier CV, et al. Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. Lancet Digit Health 2022;4(9):e657–e666.
14. Jain A, Huang J, Ravipati Y, et al. Head and Neck Primary Tumor and Lymph Node Auto-segmentation for PET/CT Scans. In: Andrearczyk V, Oreiller V, Hatt M, Depeursinge A, eds. Head and Neck Tumor Segmentation and Outcome Prediction. HECKTOR 2022. Lecture Notes in Computer Science, vol 13626. Springer, 2023; 61–69.
15. Ye Z, Saraf A, Ravipati Y, et al. Development and Validation of an Automated Image-Based Deep Learning Platform for Sarcopenia Assessment in Head and Neck Cancer. JAMA Netw Open 2023;6(8):e2328280–e2328280.
16. Tak D, Ye Z, Zapaischykova A, et al. Noninvasive Molecular Subtyping of Pediatric Low-Grade Glioma with Self-Supervised Transfer Learning. Radiol Artif Intell 2024;6(3):e230333.
17. Nalepa J, Adamski S, Kotowski K, et al. Segmenting pediatric optic pathway gliomas from MRI using deep learning. Comput Biol Med 2022;142:105237.
18. Vafaeikia P, Wagner MW, Hawkins C, Tabori U, Ertl-Wagner BB, Khalvati F. Improving the Segmentation of Pediatric Low-Grade Gliomas Through Multitask Learning. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2022; 2119–2122.
19. Ostrom QT, Price M, Neff C, et al. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2015–2019. Neuro Oncol 2022;24(Suppl 5):v1–v95.
20. Ghaffari M, Sowmya A, Oliver R. Automated Brain Tumor Segmentation Using Multimodal Brain Scans: A Survey Based on Models Submitted to the BraTS 2012-2018 Challenges. IEEE Rev Biomed Eng 2020;13:156–168.
21. Biratu ES, Schwenker F, Ayano YM, Debelee TG. A survey of brain tumor segmentation and classification algorithms. J Imaging 2021;7(9):179.
22. Zhuang F, Qi Z, Duan K, et al. A Comprehensive Survey on Transfer Learning. Proc IEEE 2021;109(1):43–76.
23. Jaiswal A, Babu AR, Zadeh MZ, Banerjee D, Makedon F. A Survey on Contrastive Self-Supervised Learning. Technologies (Basel) 2021;9(1):2.
24. van Engelen JE, Hoos HH. A survey on semi-supervised learning. Mach Learn 2020;109(2):373–440.
25. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell 2020;2(2):e200029.
26. 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. arXiv 1811.02629 [preprint] https://arxiv.org/abs/1811.02629. Posted November 5, 2018. May 1, 2023.
27. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans Med Imaging 2015;34(10):1993–2024.
28. Baid U, Ghodasara S, Bilello M, et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. arXiv 2107.02314 [preprint] https://arxiv.org/abs/2107.02314. Posted July 5, 2021. Accessed May 1, 2023.
29. Fonov V, Evans AC, Botteron K, et al. Unbiased average age-appropriate atlases for pediatric studies. Neuroimage 2011;54(1):313–327.
30. Isensee F, Schell M, Pflueger I, et al. Automated brain extraction of multisequence MRI using artificial neural networks. Hum Brain Mapp 2019;40(17):4952–4964.
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 2021;18(2):203–211.
32. Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology 1945;26(3):297–302.
33. Kann BH, Hosny A, Aerts HJWL. Artificial intelligence for clinical oncology. Cancer Cell 2021;39(7):916–927.
34. Hosny A, Bitterman DS, Guthier CV, et al. Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. Lancet Digit Health 2022;4(9):e657–e666.
35. Zapaishchykova A, Tak D, Boyd A, Ye Z, Aerts HJWL, Kann BH. SegmentationReview: A Slicer3D extension for fast review of AI-generated segmentations. Softw Impacts 2023;17:100536.
36. Artzi M, Gershov S, Ben-Sira L, et al. Automatic segmentation, classification, and follow-up of optic pathway gliomas using deep learning and fuzzy c-means clustering based on MRI. Med Phys 2020;47(11):5693–5701.
37. 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 2022;24(2):289–299.
38. Oktay O, Schlemper J, Le Folgoc L, et al. Attention U-Net: Learning Where to Look for the Pancreas. arXiv 1804.03999 [preprint] https://arxiv.org/abs/1804.03999. Posted April 11, 2018. Accessed May 1, 2023.
39. Mehta R, Sivaswamy J. M-net: A Convolutional Neural Network for deep brain structure segmentation. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, 2017; 437–440.
40. Cao H, Wang Y, Chen J, et al. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In: Karlinsky L, Michaeli T, Nishino K, eds. Computer Vision – ECCV 2022 Workshops. ECCV 2022. Lecture Notes in Computer Science, vol 13803. Springer, 2023; 205–218.
41. Isensee F, Ulrich C, Wald T, Maier-Hein KH. Extending nnU-Net Is All You Need. Informatik aktuell 2023:12–17.
42. 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.