Title

Longitudinal Assessment of Posttreatment Diffuse Glioma Tissue Volumes with Three-dimensional Convolutional Neural Networks.

Journal

Radiology. Artificial intelligence, 4(5)

ISSN

2638-6100

Authors

Rudie, Jeffrey D
Calabrese, Evan
Saluja, Rachit
et al.

Publication Date

2022-09-01

DOI

10.1148/ryai.210243

Peer reviewed

Radiology: Artificial Intelligence

AI IN BRIEF

Longitudinal Assessment of Posttreatment Diffuse Glioma Tissue Volumes with Three-dimensional Convolutional Neural Networks

Jeffrey D. Rudie, MD, PhD • Evan Calabrese, MD, PhD • Rachit Saluja, MSE • David Weiss, MSE • John B. Colby, MD, PhD • Soonmee Cha, MD • Christopher P. Hess, MD, PhD • Andreas M. Rauschecker, MD, PhD • Leo P. Sugrue, MD, PhD • Javier E. Villanueva-Meyer, MD

From the Department of Radiology and Biomedical Imaging, University of California, San Francisco, 513 Parnassus Ave, Suite S-261D, Box 0628, San Francisco, CA 94143 (J.D.R., E.C., D.W., J.B.C., S.C., C.P.H., A.M.R., L.P.S., J.E.V.M.); and Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, Pa (R.S.). Received September 14, 2021; revision requested January 6, 2022; revision received May 17; accepted July 15. **Address correspondence to** J.D.R. (email: *Jeff.Rudie@gmail.com*).

Supported in part by an American Society of Neuroradiology Foundation Grant in Artificial Intelligence.

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

Radiology: Artificial Intelligence 2022; 4(5):e210243 • https://doi.org/10.1148/ryai.210243 • Content codes: AI MR OI NR

Neural networks were trained for segmentation and longitudinal assessment of posttreatment diffuse glioma. A retrospective cohort (from January 2018 to December 2019) of 298 patients with diffuse glioma (mean age, 52 years ± 14 [SD]; 177 men; 152 patients with glioblastoma, 72 patients with astrocytoma, and 74 patients with oligodendroglioma) who underwent two consecutive multimodal MRI examinations were randomly selected into training (n = 198) and testing (n = 100) samples. A posttreatment tumor segmentation three-dimensional nnU-Net convolutional neural network with multichannel inputs (T1, T2, and T1 postcontrast and fluid-attenuated inversion recovery [FLAIR]) was trained to segment three multiclass tissue types (peritumoral edematous, infiltrated, or treatment-changed tissue [ED]; active tumor or enhancing tissue [AT]; and necrotic core). Separate longitudinal change nnU-Nets were trained on registered and subtracted FLAIR and T1 postlongitudinal images to localize and better quantify and classify changes in ED and AT. Segmentation Dice scores, volume similarities, and 95th percentile Hausdorff distances ranged from 0.72 to 0.89, 0.90 to 0.96, and 2.5 to 3.6 mm, respectively. Accuracy rates of the posttreatment tumor segmentation and longitudinal change networks being able to classify longitudinal changes in ED and AT as increased, decreased, or unchanged were 76%–79% and 90%–91%, respectively. The accuracy levels of the longitudinal change networks were not significantly different from those of three neuroradiologists (accuracy, 90%–92%; κ, 0.58–0.63; $P > .05$). The results of this study support the potential clinical value of artificial intelligence–based automated longitudinal assessment of posttreatment diffuse glioma.

Supplemental material is available for this article.

MRI plays a central role in the evaluation of diffuse glioma disease progression and treatment response (1). Advanced artificial intelligence methods show promise for quantifying and assessing changes in tumor tissue volumes, which could improve diagnosis and treatment for patients with diffuse glioma (2).

The most common indication for glioma imaging is the longitudinal assessment of the disease burden after maximal safe resection, radiation, and chemotherapy (3). The U-Net convolutional neural network architecture (4,5) has excelled in tumor segmentation, with performance reaching human interrater reliability in the Multimodal Brain Tumor Segmentation (BraTS) challenges (5,6). However, the BraTS dataset is limited to a single time point: pretreatment brain MRI. The varied appearance of the brain after treatment, as well as the changes in subtle infiltrative tumor across imaging time points, make the accurate assessment of longitudinal changes in the tumor burden challenging. A method to quantitatively track the disease burden over time would be of great clinical value.

In this study, we trained neural networks to segment MR images of the brain after treatment in patients with all grades of diffuse glioma. In addition to training a network to segment posttreatment tumor tissues at individual time points, we trained separate longitudinal change networks to localize and quantify areas of changing tumor tissue types more precisely by including coregistered images from the two time points and subtraction images between the two time points. We then evaluated the performance of the networks and neuroradiologists to assess longitudinal changes in volumes of tumor tissue subregions.

Materials and Methods

Study Design and Patients

This study conducted at the University of California, San Francisco, Medical Center was in compliance with the Health Insurance Portability and Accountability Act and was approved by the institutional review board, with a waiver for written consent. A total of 298 patients (mean age, 52 years ± 14 [SD]; 177 men) with diffuse gliomas were retrospectively evaluated after excluding four initially selected patients with missing images (Table 1). Patients were identified from searches of institutional radiology archives (mPower Clinical Analytics; Nuance Communications) for MR images of the brain in consecutive discrete patients who underwent diffuse glioma posttreatment follow-up imaging between January 2018 and December

Abbreviations

AT = active tumor or enhancing tissue, BraTS = Multimodal Brain Tumor Segmentation Challenge, ED = peritumoral edematous, infiltrated, or treatment-changed tissue, FLAIR = fluid-attenuated inversion recovery, NCR = necrotic core, RANO = Response Assessment for Neuro-Oncology, 3D = three-dimensional

Summary

Three-dimensional U-Net convolutional neural networks segmented diffuse glioma tissue subregions on MR images with a high level of accuracy and assessed longitudinal changes at the level of neuroradiologists in patients undergoing routine posttreatment MRI.

Key Points

- Convolutional neural networks trained on 396 multimodal MR images accurately segmented tissue subregions in posttreatment diffuse glioma, with median Dice scores ranging from 0.72 to 0.89 and volume similarities ranging from 0.90 to 0.96 in a held-out test set of 200 MR images.
- A network trained to segment individual posttreatment diffuse gliomas at two consecutive time points demonstrated rates of accuracy between 76% and 88% for automatically classifying longitudinal changes in tumor tissue subregions relative to rates achieved by neuroradiologists.
- Separate networks trained on registered and subtraction images from the two consecutive time points demonstrated rates of accuracy between 90% and 93% for classifying longitudinal changes, which was not significantly different from the accuracy rates achieved by neuroradiologists (90%–94%).

Keywords

MR Imaging, Neuro-Oncology, Neural Networks, CNS, Brain/Brain Stem, Segmentation, Quantification, Convolutional Neural Network (CNN)

2019. There were 152 patients with high-grade astrocytoma (glioblastoma), 74 patients with grade 2 or 3 oligodendroglioma, and 72 patients with grade 2 or 3 astrocytoma. Patients typically received standard-of-care therapy, which included maximal surgical resection combined with either radiation or radiation and chemotherapy, depending on the tumor grade. Images from two posttreatment time points were obtained for each patient (596 total images). One hundred ninety-eight patients (396 images) were randomly selected to be included in the training set, with the remaining 100 patients (200 images) being included in the test set.

The publicly available 2020 BraTS training dataset (7,8) was used to train an initial network for preliminary segmentations of the posttreatment MR images. The 2020 BraTS dataset consists of 369 preoperative patients with diffuse glioma. Manual expert segmentation of the BraTS dataset delineated three tumor subregions: necrotic core (NCR); active tumor or enhancing tissue (AT); and peritumoral edematous, infiltrative, or treatment-changed tissue (ED). The whole-tumor extent is defined as the union of all three distinct subregions (ED, AT, and NCR), and the tumor core is defined as the union of the AT and NCR.

Imaging Data Acquisition

The details of the BraTS imaging acquisition parameters, representing a heterogeneous multisite preoperative dataset, are found elsewhere (7,8). Patients from the University of California, San Francisco, Medical Center underwent a standardized brain tumor MRI protocol that used one of four 3.0-T GE Discovery 750 (GE Healthcare) imagers. The imaging protocol included three-dimensional (3D) precontrast T1-weighted images; 3D postcontrast T1-weighted images; 3D T2-weighted images; and 3D T2 fluid-attenuated inversion recovery (FLAIR) images. Typical imaging parameters are detailed in Appendix E1 (supplement).

Reference Standard Voxelwise Annotations

Reference standard voxelwise segmentations of tumor tissue subregions were generated by a neuroradiology attending physician (J.D.R., with 1 year of experience as a neuroradiology attending physician and 5 years of segmentation experience) by using an iterative process initially refined from a model trained on the BraTS 2020 dataset, which is detailed in Appendix E2 (supplement). The tumor tissue subregions of NCR, AT, and ED followed the BraTS segmentation guidelines, with the following modifications made related to the posttreatment nature of the images: Resection cavities were not included in any of the tissue subregions; gliosis and postradiation changes manifesting as T2 or FLAIR hyperintensity were also included in the ED class; and smooth linear thin enhancement underlying the craniotomy or in the resection cavity were not included in the AT class, but any enhancing tissue that could possibly reflect tumor was included.

Reference Standard Longitudinal Change Categorical Annotations

Annotations for the longitudinal tumor volume change categories (increased, decreased, and unchanged) for the AT and ED were based on the neuroradiologist's clinical assessment of significant changes in volume between time points, regardless of the magnitude of change, in contrast to Response Assessment for Neuro-Oncology (RANO) criteria (9), which require a 25% or more change in tumor volume to classify progression. For the training set, these change categories were created by a single neuroradiology attending physician with reference to the final radiology reports and unblinded review of the images (J.D.R.). For the test set, three board-certified academic neuroradiologists (L.P.S., J.E.V.M., and A.M.R., with 6 years, 5 years, and 1 year of post–neuroradiology fellowship experience, respectively) provided annotations for the longitudinal change categories and were blinded to the radiology reports or other clinical information. The final reference standard categorical change categories for the test set were determined by majority consensus from the three neuroradiologists' annotations.

Image Preprocessing

We implemented an automated image preprocessing pipeline similar to that of BraTS, which included intermodality registration, $1 \times 1 \times 1$ interpolation, skull stripping, and bias correction, as detailed in Appendix E3 (supplement). For the longitudinal change networks, additional preprocessing was performed, including registration and subtraction between time points, as detailed in Appendix E4 (supplement).

Table 1: Patient Demographics, Diffuse Glioma Types, and Tumor Statistics

Demographic	Training	Testing	Total	*P* Value
No. of patients	198	100	298	…
Age (y)	51.6 ± 13.9	52.3 ± 14.1	51.9 ± 14.0	.60
No. of men	118 (60)	59 (59)	177 (59)	.92
Primary cancer type				
Glioblastoma (grade 4)	103 (52)	49 (49)	152 (51)	.63
Astrocytoma (grade 2 or 3)	48 (24)	24 (24)	72 (24)	.96
Oligodendroglioma (grade 2 or 3)	47 (24)	27 (27)	74 (25)	.54
Longitudinal change				
Unchanged ED	115 (58)	56 (56)	171 (57)	.40
Increased ED	66 (33)	32 (32)	98 (33)	.82
Decreased ED	15 (8)	12 (12)	27 (9)	.21
Unchanged AT	120 (61)	60 (60)	180 (60)	.92
Increased AT	64 (32)	31 (31)	95 (32)	.82
Decreased AT	16 (8)	9 (9)	25 (8)	.79
Tumor volumetric information				
Presence of AT	118 (60)	69 (69)	187 (63)	.29
Presence of NCR	66 (33)	40 (40)	106 (36)	.26
WT volume (cm^3)	44.2 ± 42.6	47.8 ± 47.1	45.4 ± 44.2	.37
ED volume (cm^3)	37.5 ± 34.2	39.0 ± 38.0	38.0 ± 35.0	.64
TC volume (cm^3)	11.5 ± 16.0	12.6 ± 17.5	11.9 ± 16.6	.54
AT volume (cm^3)	9.2 ± 12.2	9.7 ± 13.0	9.2 ± 12.2	.56
NCR volume (cm^3)	4.3 ± 6.3	5.0 ± 7.1	4.6 ± 6.6	.50

Note.—Data are presented as means ± SDs or as counts with percentages in parentheses. *P* values are for comparisons between the training and validation datasets and the test dataset and were obtained by using *t* tests or χ^2 tests. AT = active tumor or enhancing tissue, ED = peritumoral edematous, infiltrated, or treatment-changed tissue, NCR = necrotic core, TC = tumor core (AT + NCR), WT = whole tumor (ED + AT + NCR).

U-Net Convolutional Neural Network Architecture

For model training and inference, we used the default settings of nnU-Net (5), which is a self-configuring method that automatically performs preprocessing, network architecture, and hyperparameter tuning (5). Details of the nnU-Net implementation and training parameters are found in Appendix E5 (supplement). The architectures and example inputs and outputs of the posttreatment tumor segmentation network and the ED and AT longitudinal change networks are shown in Figure 1. The longitudinal change networks were developed to localize and better quantify and classify changes in ED and AT.

Performance Metrics and Statistical Analysis

Tissue segmentation performance of the whole tumor, ED, tumor core, AT, and NCR subregions in the test set were evaluated by using the Dice metric (10), volume similarities, and the 95th percentile Hausdorff distance, as detailed in Appendix E6 (supplement).

To classify the longitudinal change categories for the AT and ED subregions from the tumor segmentation model, we applied a threshold for the percent change and minimum volume of AT, which was determined by optimizing classification accuracy in the training set. For the ED and AT longitudinal change networks, a total volumetric change threshold was applied, which was also determined by optimizing classification accuracy in the training set.

Average neuroradiologist performance was calculated relative to the final reference standard annotations. Among the three neuroradiologists' annotations, Cohen κ statistics and interrater performance were also calculated. Accuracy rates of the tumor segmentation and longitudinal change networks for assessing longitudinal changes in ED and AT were compared with the accuracy rates demonstrated by neuroradiologists by using χ^2 tests. Statistical significance was defined as a *P* value less than .05.

Results

Tumor Tissue Volumes

Tumor volumetric information derived from manual segmentations of the training and test sets is shown in Table 1. A total of 63.1% (188 of 298) of patients had at least 0.02 cm^3 of AT (84.9% [129 of 152] of patients with glioblastoma, 47.2% [34 of 72] of patients with astrocytoma, and 35.1% [26 of 74] of patients with oligodendroglioma), and

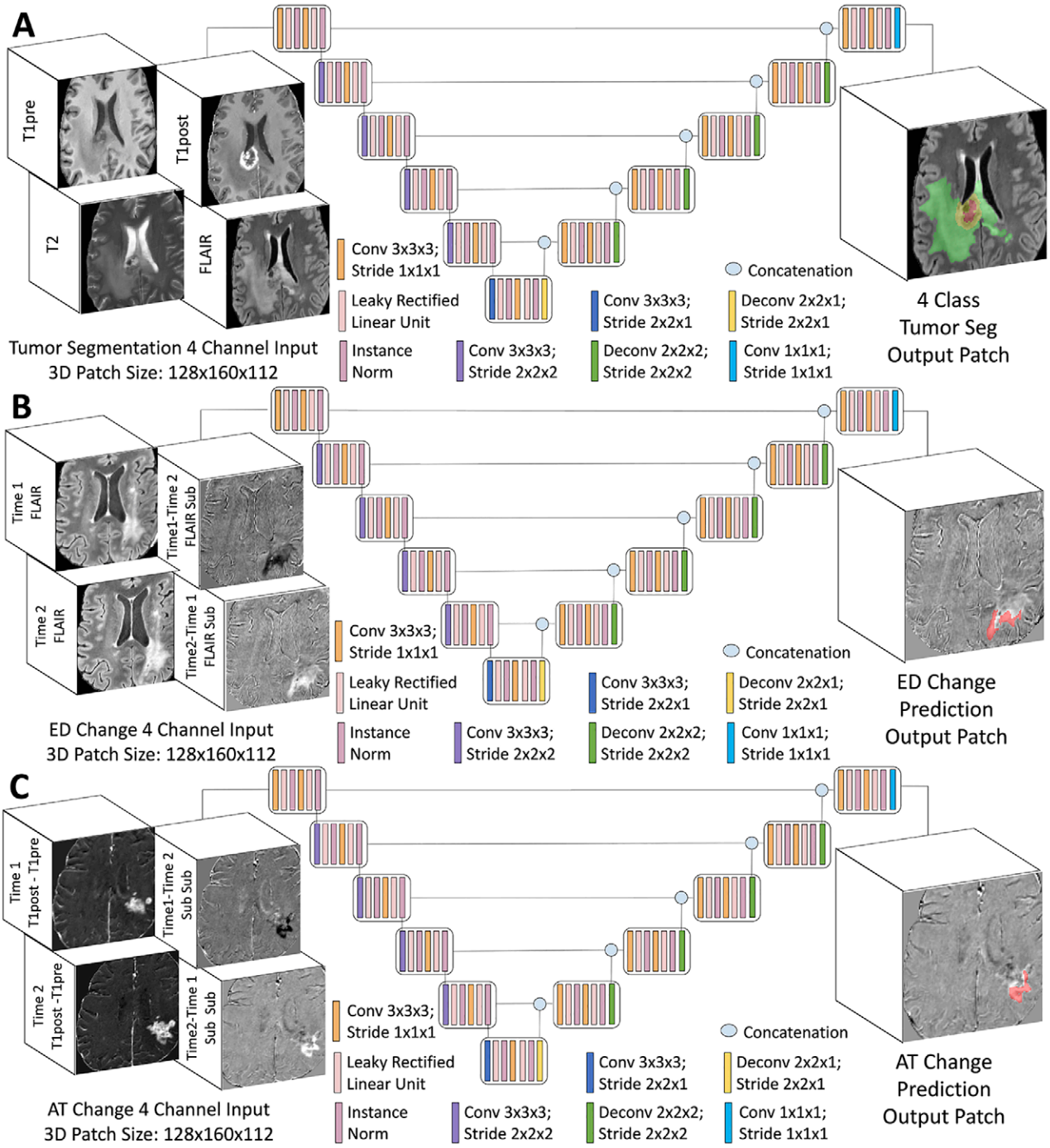

Figure 1: Three-dimensional (3D) nnU-Net neural networks used for the longitudinal assessment of diffuse glioma. **(A)** Posttreatment tumor segmentation network with T1-weighted, postcontrast ($T1_{post}$) images; fluid-attenuated inversion recovery (FLAIR) images; and T2-weighted images used as four-channel inputs and four-class outputs, consisting of background, active tumor or enhancing tissue (AT), necrotic core (NCR), and peritumoral edematous, infiltrated, or treatment-changed tissue (ED). **(B)** ED longitudinal change network with time 1 FLAIR images, time 2 FLAIR images, and time 1 minus time 2 and time 2 minus time 1 subtraction images as four-channel inputs and two-class outputs (background, increase in ED). **(C)** AT longitudinal change network with time 1 $T1_{post}$ minus T1-weighted precontrast ($T1_{pre}$) subtraction images; time 2 $T1_{post}$ minus $T1_{pre}$ subtraction images; time 1 minus time 2 subtraction images; and time 2 minus time 1 subtraction images as four-channel inputs and two-class outputs (background, increase in AT). Conv = convolution, Deconv = deconvolution, Norm = normalization, Seg = segmentation, Sub = subtraction.

35.6% (106 of 298) of patients had at least 0.02 cm^3 of NCR (56.6% [86 of 152] of patients with glioblastoma, 19.4% [14 of 72] of patients with astrocytoma, and 8.1% [six of 74] of patients with oligodendroglioma).

Segmentation Performance

The Dice scores, volume similarities, and 95th percentile Hausdorff distance metrics for the tissue subregions in the test set for the tumor segmentation and longitudinal change net-

Table 2: Test Set Segmentation Performance Metrics

	Dice Score		Volume Similarity		95th Percentile Hausdorff Distance (mm)	
Tumor Volume	Mean ± SD	Median	Mean ± SD	Median	Mean ± SD	Median
Posttreatment tumor segmentation network						
WT	0.86 ± 0.10	0.89 (0.84–0.93)	0.94 ± 0.10	0.96 (0.92–0.98)	6.9 ± 10.0	3.3 (1.7–7.1.0)
ED	0.85 ± 0.11	0.88 (0.83–0.92)	0.94 ± 0.09	0.96 (0.92–0.99)	6.6 ± 10.1	3.0 (1.4–6.7)
TC	0.71 ± 0.27	0.82 (0.55–0.92)	0.82 ± 0.25	0.95 (0.74–0.98)	8.6 ± 14.6	10.4 (1.4–8.3)
AT	0.71 ± 0.26	0.82 (0.55–0.92)	0.83 ± 0.25	0.96 (0.80–0.99)	8.2 ± 14.7	10.4 (1.0–7.9)
NCR	0.65 ± 0.29	0.72 (0.49–0.88)	0.80 ± 0.26	0.90 (0.74–0.97)	5.9 ± 8.1	10.4 (1.4–6.0)
Longitudinal change network						
ED change	0.73 ± 0.25	0.83 (0.64–0.88)	0.84 ± 0.27	0.94 (0.85–0.98)	10.3 ± 11.6	5.7 (2.0–15.1)
AT change	0.60 ± 0.26	0.67 (0.45–0.81)	0.73 ± 0.27	0.86 (0.68–0.92)	14.2 ± 16.9	5.4 (2.5–19.4)

Note.—Dice scores, volume similarities, and 95th percentile Hausdorff distances are shown as means ± SDs, and medians are shown with 25%–75% IQRs in parentheses. ED and AT change performance metrics include increases and decreases. AT = active tumor or enhancing tissue, ED = peritumoral edematous, infiltrated, or treatment-changed tissue, NCR = necrotic core, TC = tumor core (AT + NCR), WT = whole tumor (ED + AT + NCR).

works are shown in Table 2. Median Dice scores for the tumor segmentation network ranged from 0.72 to 0.89 compared with 0.43 to 0.70 for a model trained only on the BraTS 2020 data (Table E1 [supplement]). Example test cases from the tumor segmentation network are shown in Figure 2. Example test cases for the ED and AT longitudinal change networks are shown in Figure 3. Correlations and Bland-Altman plots between manual and predicted tumor subregion volumes and correlations between tumor subregion volumes and Dice scores are shown in Figure E1 (supplement) and are detailed in Appendix E7 (supplement).

Classification of Categorical Changes in Tumor Volumes

Test cases were automatically classified into increased, decreased, or unchanged longitudinal change categories by using thresholds established by optimizing training data classification performance accuracy. For the tumor segmentation network, thresholds were set at 15% and −15% for increased or decreased ED and at 15% and −20% for increased or decreased AT (with a minimum change of tissue of 0.5 cm^3 required for AT). The tumor segmentation network achieved accuracies of 76% (ED) and 79% (AT) for predicting the three categorical change categories and achieved accuracies of 79% (ED) and 88% (AT) for predicting two categorical change categories (increased vs not increased). The full performance metrics, including the sensitivity, specificity, negative predictive value, positive predictive value, F1 score, and accuracy metrics, are shown in Table 3. For the longitudinal change networks, net change thresholds were set at 0.2 cm^3 and −0.5 cm^3 for increased or decreased ED volume and at 0.1 cm^3 and −0.25 cm^3 for increased or decreased AT volume. The longitudinal change networks were 91% and 90% accurate at predicting the three categorical change categories for ED and AT, respectively, and were 93% and 92% accurate at predicting the increased versus not increased categories, respectively. Higher longitudinal classification performance was associated with larger changes in tumor volumes and smaller tumors, as detailed in Appendix E8 (supplement).

Neuroradiologist Interrater Reliability and Performance

The average neuroradiologist achieved accuracy rates of 90.3% and 91.7% for the three-class changes in ED and AT, respectively, and achieved accuracy rates of 93.7% and 94.0% for the two-class changes in ED and AT, respectively; the full performance statistics are shown in Table 3. Neuroradiologist performance was significantly higher than tumor segmentation network performance for all tasks ($P < .05$; Table 3), but there was no evidence of a difference compared with longitudinal change network performance ($P > .05$).

The average interrater reliability values of the three neuroradiologists for ED and AT were 81% and 84%, respectively, for the three-class longitudinal assessment task, with Cohen κ values of 0.58 and 0.63, respectively (Table 3). For the two-class longitudinal assessment task, the interrater reliability values were 87% and 88%, respectively (Cohen κ, 0.71 and 0.72), which are considered to indicate moderate agreement (11).

Discussion

Artificial intelligence methods designed for the longitudinal assessment of glioma can play a critical role in improving the accuracy and efficiency of assessments of the change in the tumor burden in both routine practice and clinical trials. In this study, we trained state-of-the-art, 3D U-Net neural networks for longitudinal assessment of posttreatment diffuse glioma MR images. The networks performed with high segmentation accuracy across glioma tissue subregions, and they classified categorical changes in volumes of tumor tissue types at the level of neuroradiologists.

Although most prior work has evaluated methods for segmentation of preoperative diffuse glioma tissue subregions,

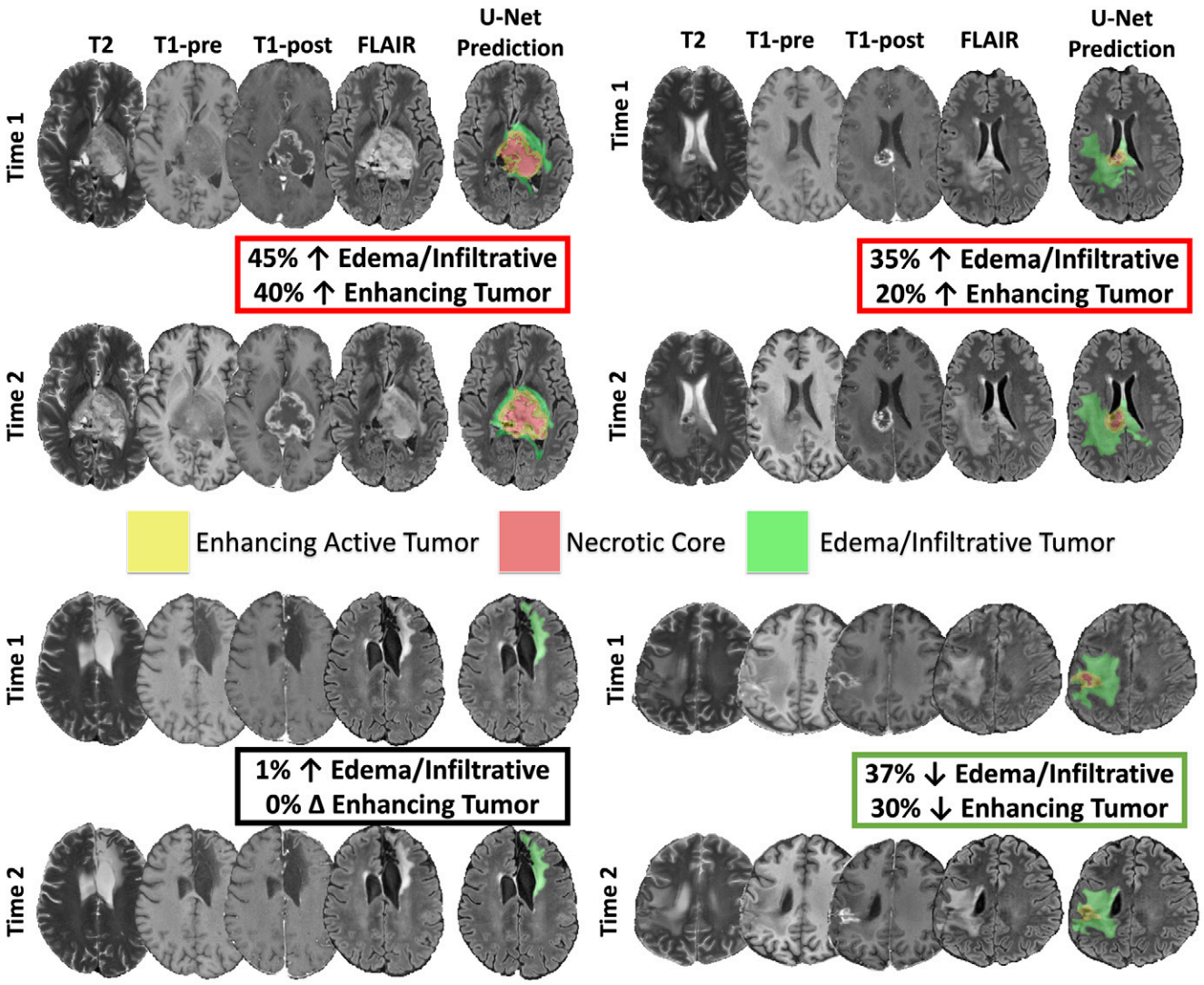

Figure 2: Examples of tumor segmentation network–predicted segmentations. Four example cases segmented with two time points with example axial T2-weighted images, T1-weighted precontrast images (T1-pre), T1-weighted postcontrast images (T1-post), fluid-attenuated inversion recovery (FLAIR) images, and example tumor tissue class segmentations overlaid on the FLAIR image (green = peritumoral edematous, infiltrated, or treatment-changed tissue [ED]; yellow = active tumor or enhancing tissue [AT]; red = necrotic tumor core). The total percent change in the ED and AT subregions are shown between the two time points for each of the four sets of cases.

typically relying on the BraTS (7,8) dataset, two recent studies have performed automated segmentation of posttreatment glioma (12,13). These studies focused on posttreatment glioma segmentation in the context of the RANO criteria, achieving high Dice scores and good correlations with manual RANO measurements. In contrast, we explicitly designed a system for routine radiologic longitudinal assessment, which incorporated consecutive time points into a model that could classify and localize relatively small changes in tumor volumes. Multiple adaptations, including custom skull stripping, revised expert annotation protocols, and a longitudinal input, were required to develop a robust algorithm. Just as radiologists directly compare images from different time points, we created longitudinal change networks to precisely localize and quantify changes in ED and AT across time points. Ultimately, this approach achieved 90%–93% accuracy in the longitudinal classification of changes in tumor subregion volumes, which was not significantly different from the accuracy rates achieved by three neuroradiologists. Although the posttreatment tumor segmentation network had excellent segmentation metrics, similar to prior posttreatment glioma segmentation studies (12,13), and interrater reliability (14,15), the network's accuracy for classifying changes in longitudinal tumor subregions was only 76%–88%. This was significantly lower than neuroradiologist accuracy, suggesting that an approach that does not directly incorporate longitudinal time points could have more limited clinical use. Cases misassigned by the tumor segmentation network often showed subtle changes in tumor tissue volumes within a large tumor or technical differences between images, which was less impactful on the longitudinal change networks.

Limitations of this study included the fact that all patients were imaged with a single type of imager at a single site and that the voxelwise annotations were generated by a single expert. The algorithm should ideally be trained on heterogeneous

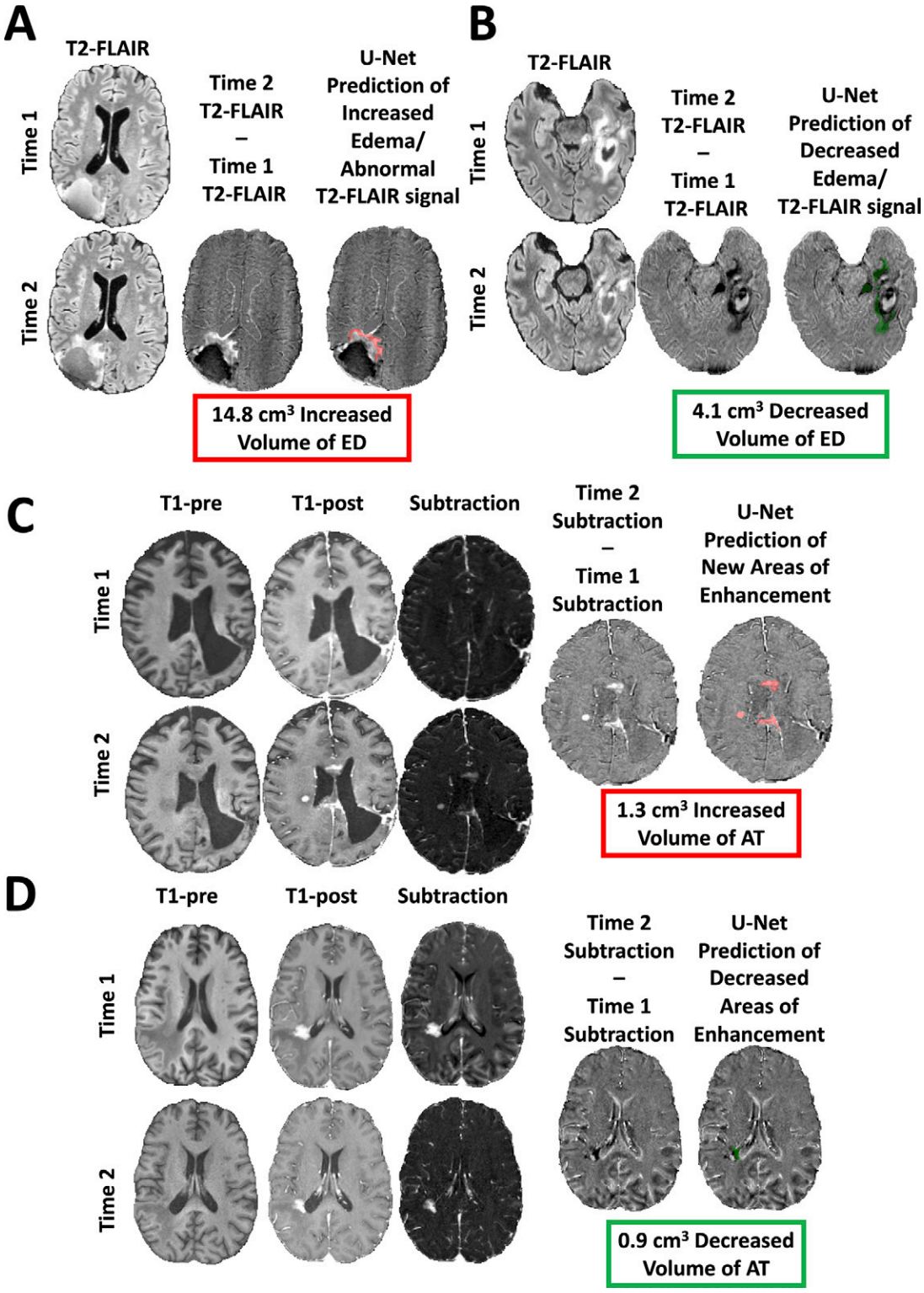

Figure 3: Examples of longitudinal change network–predicted segmentations. **(A, B)** Examples of fluid-attenuated inversion recovery (FLAIR) longitudinal change network–predicted segmentations. Two example cases with example time 1 FLAIR images, time 2 FLAIR images, and time 2 minus time 1 FLAIR subtraction images and an example of predicted segmentations (red = increasing, green = decreasing) overlaid on the time 2 minus time 1 FLAIR images. The total predicted changes in the volume of peritumoral edematous, infiltrated, or treatment-changed tissue (ED) between time points are displayed for each case. **(C, D)** Example enhancement longitudinal change network–predicted segmentations. Two example cases with example time 1 and time 2 T1-weighted precontrast (T1-pre) images, T1-weighted postcontrast (T1-post) images, time 1 and time 2 T1-post minus T1-pre subtraction images, time 2 subtraction minus time 1 subtraction images, and predicted segmentations (red = increasing, green = decreasing) overlaid on the time 2 subtraction minus time 1 subtraction images. The total predicted changes in volume of active tumor or enhancing tissue (AT) between time points are displayed for each case.

Table 3: Longitudinal Classification Task Performance Metrics

Parameter	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1 (%)	Accuracy (%)	Interrater (%)	Cohen κ	χ^2 Test *P* Value
Three classes (increased, decreased, unchanged)									
ED									
Tumor segmentation network	73	79	73	79	73	76	…	…	<.001
Longitudinal change network	91	91	89	93	85	91	…	…	.84
Attending neuroradiologists	91 ± 2	90 ± 3	88 ± 2	93 ± 2	89 ± 1	90 ± 1	81 ± 1	0.58 ± 0.01	…
AT									
Tumor segmentation network	79	79	71	86	75	79	…	…	<.001
Longitudinal change network	88	92	88	92	88	90	…	…	.61
Attending neuroradiologists	93 ± 7	91 ± 5	86 ± 7	96 ± 4	90 ± 6	92 ± 5	84 ± 6	0.63 ± 0.13	…
Two classes (increased, not increased)									
ED									
Tumor segmentation network	72	82	66	87	86	79	…	…	<.001
Longitudinal change network	94	93	86	97	90	93	…	…	.81
Attending neuroradiologists	94 ± 1	94 ± 2	92 ± 2	95 ± 1	93 ± 1	94 ± 1	87 ± 1	0.71 ± 0.01	…
AT									
Tumor segmentation network	77	93	83	90	80	88	…	…	.05
Longitudinal change network	87	94	87	94	87	92	…	…	.48
Attending neuroradiologists	94 ± 6	94.1 ± 5	91 ± 8	96 ± 3	92 ± 6	94 ± 4	88 ± 4	0.72 ± 0.11	…

Note.—Data are presented as percentages or as means ± SDs. The sensitivity, specificity, NPV, PPV, F1 score, and accuracy are shown for the ED and AT subregions for the tumor segmentation and longitudinal change networks for either two or three longitudinal change classes. The interrater accuracy and Cohen κ values are shown for the three attending neuroradiologists. χ^2 test *P* values are shown for comparisons between the networks and attending neuroradiologists. AT = active tumor or enhancing tissue, ED = peritumoral edematous, infiltrated, or treatment-changed tissue, NPV = negative predictive value, PPV = positive predictive value.

multisite data if the goal were to implement it broadly. Finally, this algorithm was designed to quantify changes in ED and AT volumes. Assessing whether these changes indicate true progression of disease or treatment-related changes, including pseudoprogression, remains a persistent challenge that might be addressed by incorporating relevant clinical information into future models.

Future directions also include testing whether such algorithms, when integrated into clinical systems, improve workflow efficiency and accuracy by automatically generating tumor volumes and regions of change that can be included in radiology reports. Future versions of this system could also be applied to different pathologic conditions and incorporate additional imaging modalities and clinical information. In conclusion, the results of the current study support the potential clinical value of artificial intelligence methods for longitudinal evaluation of posttreatment diffuse gliomas.

Author contributions: Guarantors of integrity of entire study, **J.D.R., J.E.V.M.**; 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, **J.D.R., J.B.C., S.C., A.M.R., J.E.V.M.**; clinical studies, **J.D.R., E.C., C.P.H., L.P.S., J.E.V.M.**; experimental studies, **J.D.R., R.S., D.W., J.B.C., A.M.R.**; statistical analysis, **J.D.R., D.W., J.B.C., L.P.S.**; and manuscript editing, all authors

Disclosures of conflicts of interest: **J.D.R.** American Society of Neuroradiology Foundation Grant in Artificial Intelligence funding for this work, including salary/time and computing resources; *Radiology: Artificial Intelligence* trainee editorial board alum. **E.C.** No relevant relationships. **R.S.** Consulting fees from Galileo CDS. **D.W.** Consulting fees from Galileo CDS from March 2019 to December 2020. **J.B.C.** No relevant relationships. **S.C.** No relevant relationships. **C.P.H.** Consulting fees from GE Healthcare; support for research travel from Siemens Healthineers; consulting fees from Data Safety Monitoring Board of uniQuire and Data Safety Monitoring Board of Focused Ultrasound Foundation; associate editor for *Radiology*; editorial board of *Academic Radiology*; Medical Advisory Board of PHACE Foundation. **A.M.R.** *Radiology: Artificial Intelligence* trainee editorial board alum. **L.P.S.** No relevant relationships. **J.E.V.M.** Grant from GE Healthcare, paid to UCSF; consulting fees from GE Healthcare, paid to author.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424. [Published correction appears in CA Cancer J Clin. 2020 Jul;70(4):313.]
2. Rudie JD, Rauschecker AM, Bryan RN, Davatzikos C, Mohan S. Emerging applications of artificial intelligence in neuro-oncology. Radiology 2019;290(3):607–618.
3. Gilbert MR, Wang M, Aldape KD, et al. Dose-dense temozolomide for newly diagnosed glioblastoma: a randomized phase III clinical trial. J Clin Oncol 2013;31(32):4085–4091.

4. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, eds. Medical image computing and computer-assisted intervention – MICCAI 2015. Lecture notes in computer science. Vol 9351. Cham, Switzerland: Springer, 2015; 234–241.
5. 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.
6. Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T, eds Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. BrainLes 2018. Lecture Notes in Computer Science. Vol 11384. Cham, Switzerland: Springer, 2019; 311–320.
7. 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. Accessed December 2019.
8. 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.
9. Wen PY, Chang SM, Van den Bent MJ, Vogelbaum MA, Macdonald DR, Lee EQ. Response Assessment in Neuro-Oncology clinical trials. J Clin Oncol 2017;35(21):2439–2449.
10. Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26(3):297–302.
11. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276–282.
12. Kickingereder P, Isensee F, Tursunova I, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. Lancet Oncol 2019;20(5):728–740.
13. Chang K, Beers AL, Bai HX, et al. Automatic assessment of glioma burden: a deep learning algorithm for fully automated volumetric and bidimensional measurement. Neuro Oncol 2019;21(11):1412–1422.
14. Rudie JD, Weiss DA, Saluja R, et al. Multi-disease segmentation of gliomas and white matter hyperintensities in the BraTS data using a 3D convolutional neural network. Front Comput Neurosci 2019;13:84.
15. Duong MT, Rudie JD, Wang J, et al. Convolutional neural network for automated FLAIR lesion segmentation on clinical brain MR imaging. AJNR Am J Neuroradiol 2019;40(8):1282–1290.