Neuro-Oncology Advances
6(1), vdae199, 2024 | https://doi.org/10.1093/noajnl/vdae199 | Advance Access date 16 November 2024

Deep learning-based postoperative glioblastoma segmentation and extent of resection evaluation: Development, external validation, and model comparison

Santiago Cepeda, Roberto Romero, Lidia Luque, Daniel García-Pérez, Guillermo Blasco, Luigi Tommaso Luppino, Samuel Kuttner, Olga Esteban-Sinovas, Ignacio Arrese, Ole Solheim, Live Eikenes, Anna Karlberg, Ángel Pérez-Núñez, Olivier Zanier, Carlo Serra, Victor E. Staartjes, Andrea Bianconi, Luca Francesco Rossi, Diego Garbossa, Trinidad Escudero, Roberto Hornero, and Rosario Sarabia

All author affiliations are listed at the end of the article

Corresponding Author: Santiago Cepeda MD, PhD, Department of Neurosurgery, Río Hortega University Hospital, Dulzaina 2, 47014 Valladolid, Spain (scepedac@saludcastillayleon.es).

Abstract

Background. The pursuit of automated methods to assess the extent of resection (EOR) in glioblastomas is challenging, requiring precise measurement of residual tumor volume. Many algorithms focus on preoperative scans, making them unsuitable for postoperative studies. Our objective was to develop a deep learning-based model for postoperative segmentation using magnetic resonance imaging (MRI). We also compared our model's performance with other available algorithms.

Methods. To develop the segmentation model, a training cohort from 3 research institutions and 3 public databases was used. Multiparametric MRI scans with ground truth labels for contrast-enhancing tumor (ET), edema, and surgical cavity, served as training data. The models were trained using MONAI and nnU-Net frameworks. Comparisons were made with currently available segmentation models using an external cohort from a research institution and a public database. Additionally, the model's ability to classify EOR was evaluated using the RANO-Resect classification system. To further validate our best-trained model, an additional independent cohort was used.

Results. The study included 586 scans: 395 for model training, 52 for model comparison, and 139 scans for independent validation. The nnU-Net framework produced the best model with median Dice scores of 0.81 for contrast ET, 0.77 for edema, and 0.81 for surgical cavities. Our best-trained model classified patients into maximal and submaximal resection categories with 96% accuracy in the model comparison dataset and 84% in the independent validation cohort.

Conclusions. Our nnU-Net-based model outperformed other algorithms in both segmentation and EOR classification tasks, providing a freely accessible tool with promising clinical applicability.

Key Points

- RH-GlioSeg-nnU-Net surpassed other algorithms with high Dice scores: contrast-enhancing tumor (0.81), edema (0.77), and surgical cavity (0.81).
- Our model automatically assessed the extent of resection according to the RANO-Resect classification with 85% accuracy.

Importance of the Study

The proposed model, RH-GlioSeg-nnU-Net, facilitates robust and reliable postoperative segmentation of glioblastomas, covering all tumor subregions and the surgical cavity. Additionally, we provide an automatic and standardized assessment of the extent of resection.

Glioblastoma, the most common malignant brain tumor, has a dismal prognosis with a median overall survival of approximately 15 months.[1] The extent of resection (EOR) is linked to survival, as recognized in various studies.[2,3] Classifying patients by EOR is crucial for therapy, prognosis, and clinical trial eligibility or stratification. Magnetic resonance imaging (MRI) is the preferred method for characterizing and monitoring these tumors. Specifically, postoperative MRI—recommended within 72 hours after surgery—is vital for estimating residual contrast-enhancing (CE) tumor volume, aiding in EOR assessment.[4]

Recently, the RANO resect group introduced a new classification system emphasizing prognostic implications.[3] Unlike, a previous publication,[2] which classified patients on the basis of relative tumor volume reduction, these new easy-to-use RANO categories stratify patients solely based on residual enhancing and non-enhancing tumor (ET) volumes. This approach offers more reliable stratification and potentially reduces technical effort by eliminating the need for preoperative volumetric analysis.

Automating the segmentation of residual tumors and assessing the EOR poses significant challenges for radiologists, especially in postoperative studies where hemorrhagic debris, ischemic changes, and artifacts are prevalent. The interrater agreement of manual tumor segmentation is excellent before surgery, but poor immediately after surgery and at progression. According to previous publications, the median interquartile range of EOR among raters is 8%.[5] Thus, a central review of images is often necessary in multicenter clinical trials, and comparisons between publications or centers in tumor registries are problematic. Additionally, precise and robust segmentation of the residual tumor and surgical cavity is crucial for optimal radiation treatment planning. As a result, there is a growing interest in methods to automate these tasks, as highlighted in recent publications.[6–16]

Our objective is to develop a comprehensive MRI image processing pipeline for segmenting tumor subregions in postoperative studies. To achieve this, we have explored 2 frameworks known for their robustness in medical image segmentation tasks: MONAI (https://monai.io/) and nnU-Net[17] (https://github.com/MIC-DKFZ/nnUNet). We aim to use convolutional neural networks available through these frameworks to segment the residual ET, the peritumoral region, and the postsurgical cavity. We hypothesize that effective training of a postoperative segmentation model requires diverse samples encompassing preoperative, early postoperative, and follow-up studies.

We aim to compare our model's performance with other pretrained, publicly available state-of-the-art tumor segmentation algorithms using an external validation cohort. In pursuit of a method suitable for longitudinal scans, we also intend to evaluate our model's applicability in preoperative scans.

Our main contribution lies in the development of a publicly accessible pipeline that integrates multiparametric MRI preprocessing with an automatic segmentation method, encompassing all tumor subregions, including the postoperative cavity. Additionally, we provide an automatic method for classifying EOR in glioblastoma patients according to the latest accepted categories from an oncological standpoint. Furthermore, to the best of our knowledge, there are no published comparisons of existing methods for segmenting postoperative scans in glioblastomas, a gap we aim to address through our study.

Methods

Dataset Description

The training dataset consisted of a multi-institutional cohort of patients who underwent surgery with a confirmed pathological diagnosis of IDH-wild-type glioblastoma according to the latest 2021 WHO Classification of Tumors of the Central Nervous System.[18] A total of 184 patients and 395 scans constituted the training cohort, distributed as follows: 57 patients from the Río Hortega University Hospital, Valladolid, Spain; 33 patients from St. Olavs University Hospital, Trondheim, Norway; 38 patients from The LUMIERE Dataset[19]; 30 patients from Burdenko's Glioblastoma Progression Dataset[20,21]; 21 patients from the 12 de Octubre University Hospital, Madrid, Spain; and 5 patients from the Ivy Glioblastoma Atlas Project (IvyGAP) dataset.[22,23] For each included patient, the following MRI sequences were employed: T1-weighted (T1w), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2w), and fluid-attenuated inversion recovery (FLAIR) images. Patients with inadequate image quality due to acquisition artifacts or missing MRI sequences were excluded from the study.

Regarding the timing of the MRI studies, the training cohort included 181 early postoperative scans, defined as those conducted within the initial 72 hours following surgery, in accordance with current guidelines.[4,24,25] Additionally, the training cohort included 112 preoperative scans and 102 follow-up scans, where tumor recurrence was diagnosed based on the modified RANO criteria.[26]

The external validation cohort comprised 2 subsets of patients. The first subgroup (model comparison cohort) comprises 2 Spanish centers and one public dataset, the Quantitative Imaging Network Glioblastoma (QIN-GBM) Treatment Response dataset.[21,27,28] This dataset included 15 patients from La Princesa University Hospital, Madrid,

Spain, and 21 patients from Albacete University Hospital, Castilla-La Mancha, Spain. Patients from Spanish centers underwent early postoperative scans of glioblastomas treated with complete and partial resection. Patients from the QIN-GBM dataset have late postoperative scans, encompassing patients who were scanned after surgery but before the initiation of radiation therapy, with a range of 2–5 days between scans, and all patients underwent partial tumor resection.

The second subgroup (independent validation cohort) consisted of a retrospective cohort from Oslo University Hospital, as reported in a previous study.[16] This subset included 139 patients with early postoperative scans.

Finally, we utilized the online validation dataset BraTS'20 (https://ipp.cbica.upenn.edu/) to assess the model's performance on preoperative scans. This dataset comprises 125 patients, and detailed descriptions can be found in the associated publications.[29]

The distribution of time point scans and their characteristics are outlined in Table 1. The acquisition protocols for each of the sample centers are provided in Supplementary Table 1. The acquisition protocols did not fully adhere to the recommendations of a standardized brain tumor imaging protocol,[30] with the main difference being that in 3 of the 5 centers, the pre-contrast T1-weighted sequences were acquired in 2D.

The utilization of anonymous data was authorized by the Regional Committee for Medical and Health Research Ethics (REK), Norway, with approval numbers 2016/1791, 397012, and 2019/510, and the Research Ethics Committee (CEIm) at the Río Hortega University Hospital, Valladolid, Spain, with approval number 21-PI085.

Image Preprocessing

Multiparametric MRI scans were converted to NifTI format using dcm2niix (https://github.com/rordenlab/dcm2niix) and coregistered to the SRI24 anatomical atlas,[31] then resampled to 1mm isotropic voxel resolution using SimpleElastix.[32] Skull stripping was performed using SynthStrip,[33] followed by intensity normalization using the Z-score method. The processed images were set to dimensions of 240 × 240 × 155 voxels. The entire processing pipeline is available at https://github.com/smcch/Postoperative-Glioblastoma-Segmentation. For datasets sourced from public repositories, the processing pipeline was tailored to meet the specific requirements of each dataset, incorporating only the essential steps, if needed, for each case. Additionally, attention was given to the variations in labels among different algorithms, ensuring their comparability with those of the ground truth. The preprocessing requirements for each model included in the comparison were properly fulfilled.

Ground Truth Segmentation

All ground truth segmentations of the training dataset and model comparison cohort were conducted by 2 neurosurgeons [SC, and IA] with over 10 years of experience in neuroimaging of brain tumors. The 4 processed MRI sequences were available for segmentation using a resampled voxel resolution of 1 mm^3. ITK-SNAP software, version 4.0.1 (http://itksnap.org), was utilized for this task. Initially, semiautomatic segmentation was performed using the active contour tool and the clustering mode. Three labels were generated:

- Label 1—CE tumor: Residual tumor identified as T1ce hyperintense but T1w hypointense tissue, distinguishing it from hyperintense blood.
- Label 2—Edema/infiltration: Includes all peritumoral T2-FLAIR signal changes.
- Label 3—Surgical cavity: Encompasses hematic debris, hemostatic material, and air in the cavity.

Each label was subsequently manually corrected slice by slice. For preoperative studies, label 3 was assigned to necrosis. For follow-up studies, label 3 included both the surgical cavity and necrosis if both were identifiable. The segmentations were reviewed and approved by a neuroradiologist [TE] with over 15 years of experience. The approximate segmentation time for each patient was 35 minutes.

For the independent validation cohort, a combination of semiautomatic and deep learning-aided preliminary segmentation was used. Further refinement was performed by experts using ITK-SNAP. Processed MRI sequences were used for segmentation in some cases in the dataset, whereas the original resolution was used in others, as described in the related publication.[16] For this subset of patients, only ET labels were available.

MONAI Framework Training Description

We used the UNETR network architecture[34] within the MONAI framework, focusing on technical specifics to optimize performance. MRI volumes were resized to 128 × 128 × 64 voxels. The data augmentation pipeline included random flips, rotations, elastic deformations, and intensity adjustments. UNETR was configured with 4 input and 4 output channels (including background), a feature size of 32, a hidden size of 768, 12 attention heads, and a DiceFocal loss function. The dataset was partitioned into 5 folds for cross-validation, with each fold trained over 200 epochs. An ensemble evaluation of models from different folds was used to finalize segmentation predictions, utilizing a voting mechanism to improve accuracy. Postprocessing techniques or refinement of the predicted segmentations were not used. The model trained using this framework was named: the Río Hortega Glioblastoma Segmentation UNETR (RH-GlioSeg-UNETR).

nnU-Net Framework Training Description

We used the nnU-Net framework in its 3D full-resolution version, using a dataset partitioned into 5 folds for cross-validation, with each fold trained over 1000 epochs. The loss function combined Dice and cross-entropy. Data augmentation techniques such as rotations, scaling, Gaussian noise and blur, brightness and contrast adjustments,

Table 1. Dataset Distribution and Description of Segmentation Labels Across the Sample

Center/Dataset	Number of patients	Total number of scans	Preoperative				Postoperative					Follow-up			
			n	Volume (cm^3)			*n*	EOR (GTR/RT)	Volume (cm^3)			*n*	Volume (cm^3)		
				ET	ED	NEC			ET	ED	CAV		ET	ED	NEC/CAV
Training dataset															
All centers	184	395	112	20.39 (24.40)	56.89 (69.96)	9.71 (21.44)	181	135/43	2.55 (12.31)	29.60 (42.36)	16.89 (21.05)	102	6.72 (16.56)	37.5 (54.21)	8.78 (16.02)
Río Hortega University Hospital	57	162	57	25.52 (26.66)	62.54 (61.96)	9.23 (15.47)	57	37/17	0.77 (1.52)	29.60 (35.48)	19.60 (29.19)	48	9.57 (19.33)	54.77 (58.38)	10.09 (20.86)
12 de Octubre University Hospital	21	63	21	27.85 (25.43)	64.69 (65.35)	10.34 (22.85)	21	20/1	4.09 (0.00)	34.33 (54.13)	17.06 (22.63)	21	10.18 (21.99)	45.24 (56.05)	10.10 (15.18)
St Olav's University Hospital	33	87	29	16.56 (14.37)	24.78 (47.06)	8.67 (24.21)	30	30/0	-	18.15 (36.02)	10.38 (14.20)	28	2.06 (4.68)	15.03 (20.53)	6.09 (11.40)
LUMIERE	38	38	—	v	-	-	38	33/5	0.25 (0.26)	38.17 (38.35)	18.42 (20.67)	—	—	—	—
Burdenko-GBM-Progression	30	30	—	—	—	—	30	10/20	11.62 (18.98)	12.95 (36.37)	15.02 (23.88)	—	—	—	—
Ivy-GAP	5	15	5	30.40 (4.89)	75.22 (27.56)	21.55 (6.78)	5	5/0	-	39.28 (41.81)	22.80 (20.08)	5	12.25 (13.39)	55.60 (23.33)	8.55 (5.49)
Model comparison cohort															
All centers	52	52	—	—	—	—	52	23/29	6.08 (12.26)	23.38 (37.11)	20.75 (31.18)	—	—	—	—
Albacete University Hospital	21	21	—	—	—	—	21	11/10	2.42 (3.50)	23.64 (28.75)	31.23 (38.71)	—	—	—	—
La Princesa University Hospital	15	15	—	—	—	—	15	12/3	2.68 (1.61)	34.73 (38.32)	20.75 (31.18)	—	—	—	—
QIN-GBM Treatment Response	16	16	—	—	—	—	16	0/16	13.63 (25.61)	11.52 (38.82)	9.27 (16.75)	—	—	—	—
Independent validation cohort															
Oslo University Hospital	139	139	—	—	—	—	139	5/134	0.66 (2.47)	—	—	—	—	—	—

ET, residual enhancing tumor; ED, edema; NEC, necrosis; CAV, Surgical cavity; EOR, Extent of resection; GTR, Gross total resection defined as the absence of residual contrast-enhancing tumor; RT, residual tumor; *n*, number of scans. Volumes are expressed as the median and interquartile range.

low-resolution simulations, gamma correction, and mirroring were applied to enhance the robustness of the model. This setup was designed to achieve precise segmentation results through detailed feature extraction and extensive model training. Using this framework, no postprocessing techniques were applied to the predicted segmentation. The model trained using this framework was named: the Río Hortega Glioblastoma Segmentation UNETR (RH-GlioSeg-nnU-Net).

EOR Definition

Using the volumetric information, the EOR was defined according to the latest classification system proposed by the RANO resect group as follows[3]:

- **Class 1 (Supramaximal CE resection)**: No residual CE tumor plus ≤ 5 cm^3 of non-CE tumor.
- **Class 2 (Maximal CE resection)**: ≤ 1 cm^3 of residual CE tumor.
  - **Class 2A (Complete CE resection)**: No residual CE tumor plus > 5 cm^3 of non-CE tumor.
  - **Class 2B (Near total CE resection)**: ≤ 1 cm^3 of residual CE tumor.
- **Class 3 (Submaximal resection)**: > 1 cm^3 of residual CE tumor.
  - **Class 3A (Subtotal CE resection)**: ≤ 5 cm^3 of residual CE tumor.
  - **Class 3B (Partial CE resection)**: > 5 cm^3 of residual CE tumor.
- **Class 4 (Biopsy)**: No reduction in tumor volume.

Since only ET labels were available for the independent validation cohort, the EOR classification was only applicable to define the Class 2 (maximal resection) and Class 3 (submaximal resection) categories by the proposed 1 cm^3 threshold.

Evaluation Metrics

To assess the performance of the models for segmenting postoperative MRI scans, we employed the USE-Evaluator.[35] Traditional metrics often fail to capture the nuances of clinical datasets, especially when dealing with small residual tumor labels or cases with empty annotations, such as in patients who underwent gross total resection. USE-Evaluator includes volume-based metrics such as Volumetric Similarity, which assesses how closely the volumes of the predicted and reference regions match, and Absolute Volume Difference, which quantifies the difference between these volumes. Overlap metrics like the Dice Score and Intersection over Union measure the extent of overlap between the predicted and reference regions, with higher values indicating better alignment. Additionally, distance-based metrics such as the 95th percentile Hausdorff Distance and Average Symmetric Surface Distance (both measured in millimeters) evaluate the spatial differences between the surfaces of the 2 regions, where smaller distances indicate more accurate boundary delineation.

While traditional metrics for image segmentation return "NaN" or 0 values, when the model correctly predicts an empty mask, we used USE-Evaluator to set a volumetric threshold of 0.1 cm^3 below which the agreement between the reference annotation and prediction is automatically evaluated as an image-level classification task. This strict threshold has been adopted in line with similar studies, taking into account factors such as the size of the voxel, the minimum size interpretable by the human eye, and the necessity to differentiate residual tumor from small linear enhancements of pia matter in the walls of the surgical cavity and small blood vessels.[36]

To assess the models' ability to classify the EOR, we employed precision, recall, F1 score, the area under the curve (AUC), and accuracy. Precision measures the proportion of true positives among all positive predictions, indicating how often the model is correct when it predicts a positive outcome. Recall (or sensitivity) reflects the model's ability to identify true positives from all actual positives. The F1 score is the harmonic mean of precision and recall, providing a balanced measure when there is an uneven class distribution. The AUC of the receiver operating characteristic curve was used to assess the model's ability to distinguish between classes. Finally, accuracy represents the overall proportion of correct predictions.

Models Used for Comparison

The main automatic segmentation models currently available were used. They were the following: *DeepMedic* (https://github.com/deepmedic/deepmedic),[37] *HD-GLIO* (https://github.com/NeuroAI-HD/HD-GLIO),[38,39] *PICTURE nnU-Net* (https://gitlab.com/picture-production/picture-nnunet-package),[14,38,39] *DeepEOR*,[9] *Raidionics AGU-Net* (https://github.com/raidionics/Raidionics),[40,41] nnU-Net-CPS (https://github.com/lidialuq/resect-glio),[16] and Turin U-Net.[15] Detailed descriptions of the algorithms are available in related publications.

Computational Resources

For both training and evaluation of the models, a machine equipped with an Intel Core i7 processor, 64 GB of RAM, and a dedicated RTX 3090 24 GB GPU was utilized. The model based on the MONAI framework and nnU-Net was trained using Python 3.9 and PyTorch version 2.1.1 + cu121. For the Emory University and DeepEOR models, TensorFlow version 2.10.0 was employed. Raidionics AGU-Net was executed via its graphical interface on the Windows 10 operating system. PICTURE-nnU-Net, HD-GLIO, University of Turin, nnUnet-CPS, and DeepMedic were implemented in WSL Ubuntu version 20.04.4 LTS using Python 3.8, TensorFlow version 2.13.0, and PyTorch version 2.0.1.

Results

The training cohort consisted of 395 scans from 184 patients. Among the total scans, 112 were preoperative, 181 were early postoperative, and 102 were follow-up scans.

The model comparison cohort and the independent validation cohort consisted of 52 and 139 early postoperative scans, respectively. The median volume of residual ET in the early postoperative scans was 2.55 cm^3, 6.08 cm^3, and 0.66 cm^3 for the training, model comparison, and independent validation cohorts, respectively. Details of the volumes for the labels and the EOR distribution by dataset and center are presented in Table 1.

When the class distribution according to the RANO resect EOR system is analyzed, distinct patterns emerge across the datasets. In the training cohort, Class 2A (Complete CE resection) was predominant, comprising 61.9% of the scans. In the comparison validation cohort, Class 2A remained the most common class at 42.6%. In contrast, the independent validation cohort revealed Class 2B as the most frequent class at 54%. The detailed distributions are shown in Supplementary Figure 1.

In the model comparison validation cohort, the top-performing model was based on the nn-U-Net framework (RH-GlioSeg-nnU-Net), which achieved median Dice scores of 0.81, 0.77, and 0.81 for the labels ET, edema, and surgical cavity, respectively. Supplementary Figure 2 provides an illustrative example of the predicted labels from each model included in the comparison.

A comprehensive comparison of the proposed algorithm's performance against other available algorithms is presented in Table 2 and Figure 1.

After grouping ET volumes by quartiles, we identified a direct relationship between residual ET volume and the Dice score. Patients with a residual tumor volume of less than 2.69 cm^3 presented lower Dice score values across all models compared to those with higher ET volumes. An illustration of the distribution of Dice score and ET volumes is provided in Supplementary Figure 3.

In the image-level classification task for the label ET, when a threshold of 0.1 cm^3 was used, the RH-GlioSeg-nnU-Net model achieved the highest performance, with an AUC of 0.98 and a precision of 0.93. Additional details and comparisons of the models are provided in Supplementary Table 2.

The comparative analysis of models for classifying the EOR using the RANO-resect system included 2 classification levels: a 2-class classification (maximal and submaximal CE resection) and a full 5-class classification (supramaximal, complete, near-total, subtotal, and partial CE resection). For the 2-class classification, the top 3 models were RH-GlioSeg-nnU-Net with an accuracy of 0.96, PICTURE nnU-Net with an accuracy of 0.92, and a tie between nnU-Net-CPS and HD-GLIO, both with an accuracy of 0.90. In the 5-class classification, the leading models were RH-GlioSeg-nnU-Net with an accuracy of 0.85, HD-GLIO with an accuracy of 0.79, and PICTURE nnU-Net with an accuracy of 0.64. The detailed results and additional metrics for all the models are comprehensively documented in Table 3 and Supplementary Figure 4a-b.

Table 2. Performance Evaluation Across the Model Comparison Validation Cohort

Label	Model	ASSD	DSC	HD 95	IoU	Precision	Sensitivity	VAD	VS
ET	DeepEOR	25.86 ± 3.34	0.23 ± 0.05	91.02 ± 3.75	0.13 ± 0.03	0.16 ± 0.03	0.81 ± 0.1	18.63 ± 4.16	0.35 ± 0.04
	DeepMedic	2.51 ± 0.57	0.65 ± 0.05	11.58 ± 5.12	0.49 ± 0.06	0.53 ± 0.07	**0.91 ± 0.03**	3.94 ± 1.06	0.77 ± 0.07
	HD-GLIO	1.07 ± 0.14	0.76 ± 0.02	**3.16 ± 0.66**	0.61 ± 0.03	0.80 ± 0.02	0.71 ± 0.05	1.37 ± 0.51	**0.98 ± 0.01**
	nnU-Net-CPS	1.87 ± 0.54	0.69 ± 0.05	9.85 ± 3.05	0.52 ± 0.05	0.83 ± 0.06	0.63 ± 0.03	**0.64 ± 0.19**	0.81 ± 0.03
	PICTURE-nnU-Net	1.50 ± 0.09	0.73 ± 0.02	5.24 ± 1.04	0.57 ± 0.03	0.75 ± 0.03	0.80 ± 0.05	1.38 ± 0.38	0.88 ± 0.05
	Raidionics AGU-Net	2.29 ± 0.49	0.65 ± 0.03	8.12 ± 4.31	0.48 ± 0.03	0.55 ± 0.04	0.89 ± 0.03	3.08 ± 1.06	0.87 ± 0.05
	RH-GlioSeg-nnU-Net	**0.95 ± 0.17**	**0.81 ± 0.04**	3.24 ± 0.78	**0.68 ± 0.05**	**0.86 ± 0.06**	0.80 ± 0.05	0.93 ± 0.21	0.95 ± 0.03
	RH-GlioSeg-UNETR	1.30 ± 0.51	0.73 ± 0.05	4.47 ± 4.1	0.57 ± 0.06	0.79 ± 0.04	0.72 ± 0.08	1.04 ± 0.42	0.96 ± 0.03
	Turin U-Net	27.52 ± 1.94	0.05 ± 0.03	86.54 ± 3.97	0.02 ± 0.01	0.03 ± 0.02	0.29 ± 0.1	67.97 ± 4.54	0.15 ± 0.06
ED	DeepEOR	4.44 ± 0.82	0.58 ± 0.06	20.23 ± 5.94	0.41 ± 0.06	0.52 ± 0.07	0.81 ± 0.03	23.04 ± 5.23	0.69 ± 0.05
	DeepMedic	2.60 ± 0.29	0.70 ± 0.04	13.45 ± 1.58	0.53 ± 0.05	0.57 ± 0.06	**0.90 ± 0.02**	14.12 ± 2.76	0.73 ± 0.05
	HD-GLIO	1.91 ± 0.52	0.70 ± 0.03	9.00 ± 1.95	0.54 ± 0.04	**0.79 ± 0.04**	0.71 ± 0.06	8.23 ± 1.45	0.82 ± 0.02
	PICTURE-nnU-Net	1.81 ± 0.34	0.74 ± 0.04	6.56 ± 1.68	0.59 ± 0.05	0.77 ± 0.05	0.75 ± 0.05	6.05 ± 1.73	**0.87 ± 0.02**
	RH-GlioSeg-nnU-Net	**1.46 ± 0.22**	**0.77 ± 0.02**	**5.05 ± 0.99**	0.62 ± 0.03	0.74 ± 0.06	0.89 ± 0.03	**4.93 ± 1.36**	0.86 ± 0.03
	RH-GlioSeg-UNETR	1.71 ± 0.21	**0.77 ± 0.02**	6.44 ± 1.21	**0.63 ± 0.03**	0.73 ± 0.05	0.87 ± 0.03	5.73 ± 1.59	0.85 ± 0.03
	Turin U-Net	23.82 ± 1.43	0.15 ± 0.03	81.30 ± 3.38	0.08 ± 0.02	0.08 ± 0.02	0.89 ± 0.02	208.52 ± 8.29	0.17 ± 0.04
CAV	PICTURE-nnU-Net	1.96 ± 0.24	0.77 ± 0.03	6.08 ± 0.74	0.63 ± 0.05	0.80 ± 0.04	**0.86 ± 0.01**	**4.50 ± 0.81**	**0.89 ± 0.04**
	RH-GlioSeg-nnU-Net	**1.42 ± 0.19**	**0.81 ± 0.04**	**4.36 ± 0.67**	**0.68 ± 0.05**	0.96 ± 0.01	0.71 ± 0.05	5.30 ± 0.89	0.87 ± 0.03
	RH-GlioSeg-UNETR	2.26 ± 0.28	0.75 ± 0.02	7.00 ± 1.14	0.59 ± 0.03	0.93 ± 0.01	0.63 ± 0.04	5.39 ± 1.47	0.84 ± 0.04
	Turin U-Net	8.77 ± 1.89	0.14 ± 0.04	19.82 ± 5.52	0.07 ± 0.03	**1.00 ± 0.01**	0.07 ± 0.03	19.90 ± 5.1	0.01 ± 0.02

The best-performing values are highlighted in bold.
ET, residual enhancing tumor; ED, edema; CAV, surgical cavity; ASSD, average symmetric surface distance; DSC, dice similarity coefficient; HD 95, Hausdorff distance 95th percentile; IoU, Intersection over union;VAD, volume absolute difference; VS, volumetric similarity; Values are expressed as median ± 95% confidence Interval (bootstrapped).

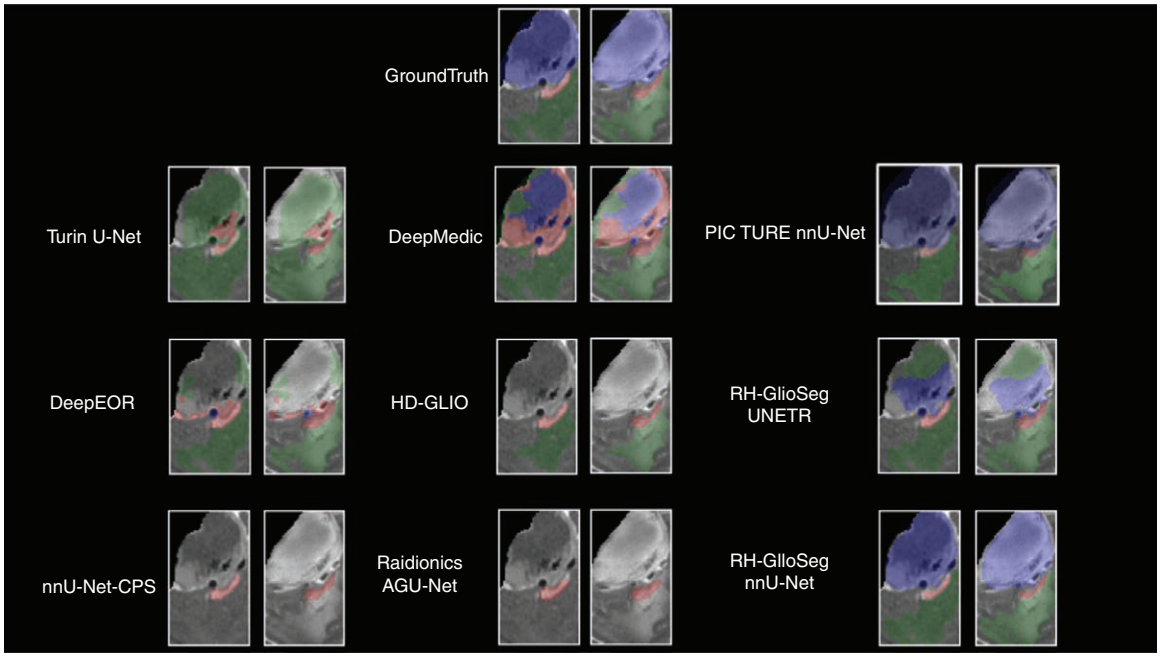

Figure 1. A descriptive example of the segmentations predicted by the models included in the comparison. The segmentations include the following labels: residual enhancing tumor, edema, and surgical cavity. The predicted labels are shown as a pair of images overlaid on (left) T1ce and (right) T2w. The visual distinction between the labels is consistent across the images for clarity.

Additionally, several examples of RH-GlioSeg-nnU-Net predictions in the model comparison cohort are shown in Figure 2.

Furthermore, our model with the highest Dice score performance, RH-GlioSeg-nnU-Net, was used to evaluate an independent validation cohort. The median Dice score for the ET label was 0.48, and the model achieved an image-level classification AUC of 0.98. For the 2-class classification of EOR, our model yielded an accuracy of 0.84. The detailed results are shown in Table 4 and Supplementary Figure 4b-c.

Finally, the RH-GlioSeg-nnU-Net model attained the highest overall overlap metrics and was selected to assess its performance on preoperative MR images using the BRATS 2020 validation dataset via the online platform. The mean Dice scores obtained were 0.78, 0.88, and 0.72 for the ET, whole tumor, and tumor core labels, respectively. The detailed evaluation results of the preoperative scans are provided in Supplementary Table 3.

Discussion

In this study, we compiled 6 datasets from collaborative research institutions and 4 datasets from publicly online available data sources encompassing pre- and postoperative multiparametric MRI studies. Our dataset boasts diversity, stemming from multiple sources, and varying categories of EOR in postoperative studies. Leveraging a robust convolutional neural network architecture, we trained a model of notable reliability.

Postoperative segmentation of glioblastomas presents a significant challenge, primarily due to the difficulty in accurately identifying residual ETs, especially when dealing with small volumes. The extensive variability observed in postoperative studies further complicates the standardization of methodologies. Variations in surgical techniques often result in patients exhibiting diverse EORs, despite undergoing surgery for glioblastoma in similar locations. Consequently, cases may vary from those with resections tightly confined to the enhancing component to those employing more aggressive strategies, such as supramarginal resections or lobectomies. These differences manifest notably in terms of the size of the surgical cavity and the deformation of the surrounding parenchyma. Additionally, the meticulousness of hemostasis significantly influences postsurgical outcomes, leading to clean cavities in some cases and the presence of blood debris, air, and hemostatic material in others.

Training a model to accurately segment residual tumors, especially small volumes, poses additional challenges, particularly in reliably predicting the "absence" of residual tumors. A model that excels at tumor segmentation may not necessarily be precise in identifying cases where no residual tumor exists, as it might tend to over-segment these regions.

In our dataset, the Class 2A (Complete CE resection) category predominated, this fact stands in contrast to other datasets where the proportion is typically reversed. Given these circumstances, our hypothesis was that the postoperative segmentation model would derive significant benefits from learning the characteristics of the tumor both preoperatively and in follow-up studies where tumor recurrence is detected.

Table 3. Classification Performance of Extent of Resection Across Model Comparison Cohort

EOR 2 Classes: maximal vs. submaximal CE resection					
Model	Precision	Recall	F1	Accuracy	AUC
RH-GlioSeg-nnU-Net	**0.97**	**0.96**	**0.96**	**0.96**	**0.96**
RH-GlioSeg-UNETR	0.89	0.89	0.88	0.89	0.89
PICTURE nnU-Net	0.92	0.92	0.92	0.92	0.92
HD-GLIO	0.91	0.91	0.90	0.90	0.91
DeepEOR	0.745	0.519	0.365	0.50	0.52
DeepMedic	0.81	0.72	0.69	0.71	0.72
Raidionics AGU-Net	0.89	0.89	0.88	0.89	0.89
Turin U-Net	0.24	0.50	0.33	0.48	0.50
nnU-Net-CPS	0.92	0.91	0.90	0.90	0.91
EOR 5 classes: Supramaximal, Complete, Near-total, Subtotal and Partial CE resection					
Model	**Precision**	**Recall**	**F1**	**Accuracy**	**AUC**
RH-GlioSeg-nnU-Net	**0.85**	**0.80**	**0.80**	**0.85**	**0.88**
RH-GlioSeg-UNETR	0.68	0.63	0.61	0.62	0.77
PICTURE nnU-Net	0.74	0.63	0.60	0.64	0.77
HD-GLIO	0.70	0.72	0.70	0.79	0.83
DeepEOR	0.10	0.22	0.13	0.31	0.52
DeepMedic	0.60	0.41	0.38	0.46	0.64
Turin U-Net	0.06	0.20	0.09	0.29	0.50

EOR, extent of resection; CE, contrast enhancing; AUC, area under the curve.

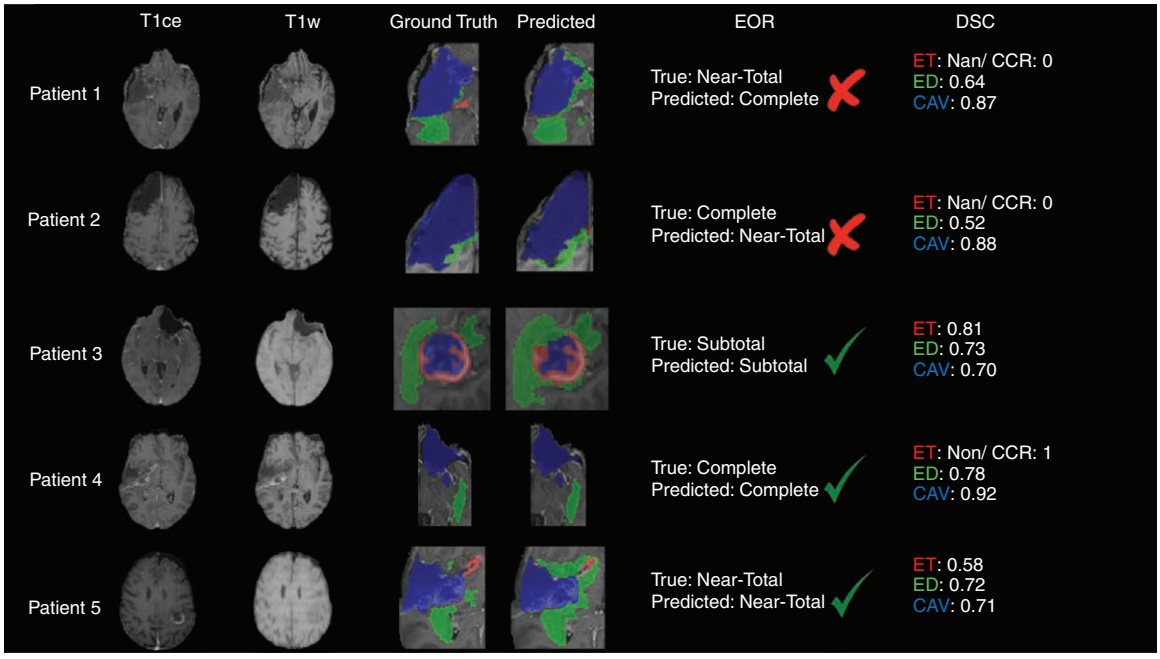

Figure 2. Examples of predictions made by the RH-GlioSeg-nnU-Net model. The classification status of the patient's resection extension (EOR) is indicated as either correct or incorrect. The ground truth and predicted segmentations are overlaid on T1 contrast-enhancing (T1ce) and T1 weighted (T1w) images to facilitate differentiation between blood remnants and residual enhancing tumors. The last column shows Dice Similarity Coefficient values for each label: enhancing tumor (ET), edema (ED), and surgical cavity (CAV). For cases with an empty label, the result is expressed as a classification task using the Correct Classification Rate (CCR).

Table 4. RH-GlioSeg-nnUnet Performance on Independent Validation Cohort

	Metric	ET
Segmentation task	ASSD	2.46 ± 0.30
	DSC	0.48 ± 0.04
	HD 95	11.13 ± 1.51
	IoU	0.32 ± 0.04
	Precision	0.55 ± 0.05
	Recall	0.53 ± 0.06
	VAD	0.61 ± 0.10
	VS	0.49 ± 0.06
Image-level classification task *	CCR	0.90 ± 0.03
	AUC	0.98 ± 0.01
	Precision	0.87 ± 0.02
	Sensitivity	0.95 ± 0.02
	Specificity	0.88 ± 0.04

ET, enhancing tumor; ASSD, average symmetric surface distance; DSC, dice similarity coefficient; HD 95, Hausdorff distance 95th percentile; IoU, Intersection over Union; VAD, volume absolute difference. VS, volumetric similarity. CCR, Correct classification rate. AUC, Area under the Curve. Values are expressed in median ± 95% Confidence Interval (bootstrapped). * 0.1 cm^3 threshold.

The automation of surgical cavity segmentation has potential applications in radiotherapy treatment planning, as shown by several studies.[6–8,42] However, few models provide comprehensive labeling of all relevant structures—such as edema, residual tumor, and surgical cavity—specifically in postoperative studies,[6,12,13,15] and many of these models are not publicly available. Our proposed solution addresses this gap by including all relevant subregions and demonstrating strong performance in detecting and estimating the volume of residual ET. With expert supervision, this approach could also save time in contouring treatment volumes.

By proposing this comparison, our aim was not to address criticism but rather to highlight strengths and glean insights from alternative approaches and strategies for a shared problem. Importantly, methodological comparisons among the models may not be feasible because of differences in their architectures, preprocessing and postprocessing pipelines, or the diverse datasets used for training. In addition, some of the models included in the comparison are not specifically designed for postoperative scans.[9,37] Notably, some models only include the possibility of segmenting the residual ET.[14,16,41] Therefore, our aim is not to benchmark them against each other but rather to provide a practical perspective on their performance in a clinical setting.

In terms of architectures and frameworks, we trained 2 models using the same dataset and employed an internal validation strategy with k-folds. However, the performance metrics are consistently higher when nnU-Net is used compared to UNETR. Despite both being 3D fully convolutional architectures and employing similar data augmentation strategies, it appears that a more complex architecture such as UNETR does not offer significant advantages over U-Net in this specific task.[43] Furthermore, all the models that achieved the highest scores in segmentation and EOR classification tasks were built upon the U-Net architecture.

To the best of our knowledge, our model offers for the first time an automatic way to classify EOR according to the latest system proposed by the RANO resect group.[3] From a neuro-oncological perspective, the key is to be able to categorize patients into maximal and submaximal surgical resection groups because of differences in terms of survival.[3]

Despite being trained primarily on early postoperative studies and follow-up data, our model demonstrates a robust ability to generalize to external preoperative datasets, such as the BraTS 2020 external validation cohort. This capability, coupled with its proven performance on postoperative data, underscores the model's potential versatility as a tool for segmenting glioblastoma throughout the entire treatment course.

The limitation of our model lies in the inherent challenge of accurately segmenting postsurgical studies while encompassing all relevant regions. While manual and semiautomatic segmentation serves as standards for training and evaluation, it is essential to acknowledge the variability between observers, which introduces a bias that is difficult to eliminate.

Our model has been publicly released to encourage further analysis, but most importantly to be tested in other clinical settings to prove its reproducibility and effectiveness. Importantly, these models are not intended to replace human observers but rather to increase their efficiency and improve diagnostic precision.

Finally, we firmly believe that only by adhering to an open science policy can the limitations in generating these types of computer-aided methods be overcome. Therefore, initiatives such as Federated Learning for Postoperative Segmentation of Treated Glioblastoma (FL-PoST; https://fets-ai.github.io/FL-PoST/) and the Brain Tumor Segmentation (BraTS) Challenge: Glioma Segmentation on Post-treatment MRI[44] should be expanded to facilitate the translation of knowledge into clinical practice.

Our study highlights the value of using a diverse multi-institutional dataset from longitudinal patient studies in conjunction with the robust nnU-Net framework, which achieves excellent performance in segmentation and EOR classification tasks. By comparing a wide range of openly available models, we provide a comprehensive guide for users to select the best model for their specific needs, ultimately bringing automatic glioblastoma segmentation closer to widespread clinical application.

Supplementary material

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

Keywords

deep learning | glioblastomas | neural network | postoperative | segmentation

Funding

This work was partially funded by a grant awarded by the "Instituto Carlos III, Proyectos I-D-i, Acción Estratégica en Salud 2022," under the project titled "Prediction of tumor recurrence in glioblastomas using magnetic resonance imaging, machine learning, and transcriptomic analysis: A supratotal resection guided by artificial intelligence," reference PI22/01680.

Acknowledgments

The authors sincerely appreciate the collaboration of all contributors to the development of the segmentation models included in this comparison. We acknowledge their technical support and willingness to share the source code of their publications. Special thanks are extended to David Bouget from the Department of Health Research, SINTEF Digital, Trondheim, Norway; Roelant S. Eijgelaar from the Neurosurgical Center Amsterdam, Amsterdam UMC, Vrije Universiteit, Netherlands.

Conflicts of interest statement

All the authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or nonfinancial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Autorship statement

Conceptualization: Santiago Cepeda. Methodology: Santiago Cepeda, Roberto Romero, Luigi Tommaso Luppino, Samuel Kuttner, Olga Esteban-Sinovas, Ignacio Arrese, Olivier Zanier, and Trinidad Escudero. Software: Santiago Cepeda, Roberto Romero, Luigi Tommaso Luppino, Samuel Kuttner, Olivier Zanier, Andrea Bianconi, Luca Francesco Rossi, and Trinidad Escudero. Formal analysis: Santiago Cepeda, Roberto Romero, Luigi Tommaso Luppino, Samuel Kuttner, Ignacio Arrese, and Trinidad Escudero. Data curation: Santiago Cepeda, Daniel García-Pérez, Guillermo Blasco, and Ignacio Arrese. Validation: Lidia Luque, Daniel García-Pérez, Guillermo Blasco, Ole Solheim, Live Eikenes, Anna Karlberg, Ángel Pérez-Núñez, Carlo Serra, Victor E. Staartjes, Diego Garbossa. Writing—original draft, review, and editing: Santiago Cepeda, Lidia Luque, Luigi Tommaso Luppino, Samuel Kuttner, Ole Solheim, Roberto Hornero, and Rosario Sarabia. Investigation: Olga Esteban-Sinovas. Supervision: Roberto Hornero and Rosario Sarabia. Funding acquisition: Santiago Cepeda.

Data availability

The images used in this study are derived from a private dataset and are available upon request from the corresponding author. Additionally, part of the code used for this research is publicly accessible in the following repository:https://github.com/smcch/Postoperative-Glioblastoma-Segmentation.

Affiliations

Department of Neurosurgery, Río Hortega University Hospital, Valladolid, Spain (S.C., O.E.-S., I.A., R.S.); Biomedical Engineering Group, Universidad de Valladolid, Valladolid, Spain (R.R., R.H.); Center for Biomedical Research in Network of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Valladolid, Spain (R.R., R.H.); Computational Radiology and Artificial Intelligence (CRAI), Department of Physics and Computational Radiology, Clinic for Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway (L.L.); Department of Physics, University of Oslo, Oslo, Norway (L.L.); Department of Physics and Computational Radiology, Clinic for Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway (L.L.); Department of Neurosurgery, Albacete University Hospital, Albacete, Spain (D.G.-P.); Department of Neurosurgery, La Princesa University Hospital, Madrid, Spain (G.B.); Department of Physics and Technology, UiT The Arctic University of Norway, Tromsø, Norway (L.T.L., S.K.); The PET Imaging Center, University Hospital of North Norway, Tromsø, Norway (S.K.); Department of Neurosurgery, St. Olavs University Hospital, Trondheim, Norway (O.S.); Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway (O.S.); Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway (L.E., A.K.); Department of Radiology and Nuclear

Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway (A.K.); Department of Neurosurgery, 12 de Octubre University Hospital (i + 12), Madrid, Spain (A.P.-N.); Department of Surgery, School of Medicine, Complutense University, Madrid, Spain (A.P.-N.); Instituto de Investigación Sanitaria, 12 de Octubre University Hospital (i + 12), Madrid, Spain (A.P.-N.); Machine Intelligence in Clinical Neuroscience & Microsurgical Neuroanatomy (MICN) Laboratory, Department of Neurosurgery, Clinical Neuroscience Center, University Hospital Zürich, University of Zürich, Zürich, Switzerland (O.Z., C.S., V.E.S.); Division of Neurosurgery, Ospedale Policlinico San Martino, IRCCS for Oncology and Neurosciences, Genoa, Italy (A.B.); Neurosurgery Unit, Department of Neuroscience "Rita Levi Montalcini," University of Turin, Turin, Italy (A.B., D.G.); Department of Informatics, Polytechnic University of Turin, Turin, Italy (L.F.R.); Department of Radiology, Río Hortega University Hospital, Valladolid, Spain (T.E.); Institute for Research in Mathematics (IMUVA), University of Valladolid, Valladolid, Spain (R.H.)

References

1. Koshy M, Villano JL, Dolecek TA, et al. Improved survival time trends for glioblastoma using the SEER 17 population-based registries. *J Neurooncol*. 2012;107(1):207–212.
2. Karschnia P, Vogelbaum MA, Van Den Bent M, et al. Evidence-based recommendations on categories for extent of resection in diffuse glioma. *Eur J Cancer*. 2021;149(May):23–33.
3. Karschnia P, Young JS, Dono A, et al. Prognostic validation of a new classification system for extent of resection in glioblastoma: A report of the RANO *resect* group. *Neuro Oncol*. 2023;25(5):940–954.
4. Rykkje AM, Larsen VA, Skjøth-Rasmussen J, et al. Timing of Early Postoperative MRI following Primary Glioblastoma Surgery—A Retrospective Study of Contrast Enhancements in 311 Patients. *Diagnostics*. 2023;13(4):795.
5. Visser M, Müller DMJ, Van Duijn RJM, et al. Inter-rater agreement in glioma segmentations on longitudinal MRI. *NeuroImage: Clinical*. 2019;22:101727.
6. Ramesh KK, Xu KM, Trivedi AG, et al. A fully automated post-surgical brain tumor segmentation model for radiation treatment planning and longitudinal tracking. *Cancers (Basel)*. 2023;15(15):3956.
7. Breto AL, Cullison K, Zacharaki EI, et al. A deep learning approach for automatic segmentation during daily MRI-Linac Radiotherapy of Glioblastoma. *Cancers (Basel)*. 2023;15(21):5241.
8. Ermiş E, Jungo A, Poel R, et al. Fully automated brain resection cavity delineation for radiation target volume definition in glioblastoma patients using deep learning. *Radiat Oncol*. 2020;15(1):100.
9. Zanier O, Da Mutten R, Vieli M, et al. DeepEOR: automated perioperative volumetric assessment of variable grade gliomas using deep learning. *Acta Neurochir*. 2022;165(2):555–566.
10. Ghaffari M, Samarasinghe G, Jameson M, et al. Automated post-operative brain tumour segmentation: A deep learning model based on transfer learning from pre-operative images. *Magn Reson Imaging*. 2022;86(Feb):28–36.
11. Abayazeed AH, Abbassy A, Müeller M, et al. NS-HGlio: A generalizable and repeatable HGG segmentation and volumetric measurement AI algorithm for the longitudinal MRI assessment to inform RANO in trials and clinics. *Neurooncol. Adv.*. 2023;5(1):vdac184.
12. Nalepa J, Kotowski K, Machura B, et al. Deep learning automates bidimensional and volumetric tumor burden measurement from MRI in pre- and post-operative glioblastoma patients. *Comput Biol Med*. 2023;154(Mar):106603.
13. Lotan E, Zhang B, Dogra S, et al. Development and practical implementation of a deep learning–based pipeline for automated pre- and postoperative glioma segmentation. *AJNR Am J Neuroradiol*. 2022;43(1):24–32.
14. Helland RH, Ferles A, Pedersen A, et al. Segmentation of glioblastomas in early post-operative multi-modal MRI with deep neural networks. *Sci Rep*. 2023;13(1):18897.
15. Bianconi A, Rossi LF, Bonada M, et al. Deep learning-based algorithm for postoperative glioblastoma MRI segmentation: A promising new tool for tumor burden assessment. *Brain Inform*. 2023;10(1):26.
16. Luque L, Skogen K, MacIntosh BJ, et al. Standardized evaluation of the extent of resection in glioblastoma with automated early post-operative segmentation. *Front Radiol*. 2024;4:1357341.
17. 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.
18. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: A summary. *Neuro Oncol*. 2021;23(8):1231–1251.
19. Suter Y, Knecht U, Valenzuela W, et al. The LUMIERE Longitudinal Glioblastoma MRI with expert RANO evaluation. *Sci Data*. 2022;9(1):768.
20. Zolotova SV, Golanov AV, Pronin IN, et al. *Burdenko's Glioblastoma Progression Dataset (Burdenko-GBM-Progression)*. 2023. (Version 1) [Data set]. The Cancer Imaging Archive. doi: 10.7937/E1QP-D183
21. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging*. 2013;26(6):1045–1057.
22. Shah N, Feng X, Lankerovich M, Puchalski RB, Keogh B. *Data from Ivy Glioblastoma Atlas Project (IvyGAP) [Data set]. The Cancer Imaging Archive*. doi: 10.7937/K9/TCIA.2016.XLwaN6nL.
23. Puchalski RB, Shah N, Miller J, et al. An anatomic transcriptional atlas of human glioblastoma. *Science*. 2018;360(6389):660–663.
24. Stupp R, Brada M, van den Bent MJ, Tonn JC, Pentheroudakis G. ESMO Guidelines Working Group. High-grade glioma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2014;25(Suppl 3):iiiiii93–iiiii101.
25. Nabors LB, Portnow J, Ammirati M, et al. NCCN Guidelines Insights: Central Nervous System Cancers, Version 1.2017. *J Natl Compr Canc Netw*. 2017;15(11):1331–1345.
26. Ellingson BM, Wen PY, Cloughesy TF. Modified criteria for radiographic response assessment in glioblastoma clinical trials. *Neurotherapeutics*. 2017;14(2):307–320.
27. Mamonov AB, Kalpathy-Cramer J. *Data From QIN GBM Treatment Response*. 2016. The Cancer Imaging Archive. doi: 10.7937/k9/tcia.2016.nQF4gpn2
28. Prah MA, Stufflebeam SM, Paulson ES, et al. Repeatability of standardized and normalized relative CBV in patients with newly diagnosed glioblastoma. *AJNR Am J Neuroradiol*. 2015;36(9):1654–1661.
29. Bakas S, Akbari H, Sotiras A, et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*. 2017;4(Sep):170117.
30. Ellingson BM, Bendszus M, Boxerman J, et al; Jumpstarting Brain Tumor Drug Development Coalition Imaging Standardization Steering Committee. Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials. *Neuro Oncol*. 2015;17(9):1188–1198.
31. Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A. The SRI24 multichannel atlas of normal adult human brain structure. *Hum Brain Mapp*. 2010;31(5):798–819.

32. Marstal K, Berendsen F, Staring M, Klein S. SimpleElastix: A User-Friendly, Multi-lingual Library for Medical Image Registration. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Las Vegas, NV: IEEE; 2016:574–582.
33. Hoopes A, Mora JS, Dalca AV, Fischl B, Hoffmann MS. skull-stripping for any brain image. *Neuroimage*. 2022;260(Oct):119474.
34. Hatamizadeh A, Tang Y, Nath V, et al. UNETR: Transformers for 3D Medical Image Segmentation. 2021. http://arxiv.org/abs/2103.10504. Accessed December 29, 2023.
35. Ostmeier S, Axelrod B, Isensee F, et al. USE-Evaluator: Performance metrics for medical image segmentation models supervised by uncertain, small or empty reference annotations in neuroimaging. *Med Image Anal*. 2023;90(Dec):102927.
36. Stummer W, Pichlmeier U, Meinel T, et al; ALA-Glioma Study Group. Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: A randomised controlled multicentre phase III trial. *Lancet Oncol*. 2006;7(5):392–401.
37. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. 2017;36(Feb):61–78.
38. 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.
39. Isensee F, Jäger PF, Kohl SAA, Petersen J, Maier-Hein KH. Automated design of deep learning methods for biomedical image segmentation. *Nat Methods*. 2021;18(2):203–211.
40. Bouget D, Pedersen A, Jakola AS, et al. Preoperative brain tumor imaging: models and software for segmentation and standardized reporting. *Front Neurol*. 2022;13(Jul):932219.
41. Bouget D, Alsinan D, Gaitan V, et al. Raidionics: An open software for pre- and postoperative central nervous system tumor segmentation and standardized reporting. *Sci Rep*. 2023;13(1):15570.
42. Canalini L, Klein J, Pedrosa De Barros N, Sima DM, Miller D, Hahn HK. Comparison of different automatic solutions for resection cavity segmentation in postoperative MRI volumes including longitudinal acquisitions. In: Linte CA, Siewerdsen JH, eds. *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*. SPIE; 2021:71. doi: 10.1117/12.2580889
43. Isensee F, Wald T, Ulrich C, et al. nU-Net revisited: a call for rigorous validation in 3D medical image segmentation. In: Linguraru MG, Dou Q, Feragen A, et al., eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. vol 15009. Springer, Cham: Lecture Notes in Computer Science; 2024. doi: 10.1007/978-3-031-72114-4_47
44. de Verdier MC, Saluja R, Gagnon L, et al. *The 2024 Brain Tumor Segmentation (BraTS) Challenge: Glioma Segmentation on Post-treatment MRI*. Published online 2024. doi: 10.48550/arXiv.2405.18368