scientific reports

Check for updates

OPEN

Implication of tumor morphology and MRI characteristics on the accuracy of automated versus human segmentation of GBM areas

Valeria Cerina[1,4]✉, Chiara Benedetta Rui[2,4], Andrea Di Cristofori[1,2,3], Davide Ferlito[2,4], Giorgio Carrabba[2,3,4], Carlo Giussani[2,3,4], Gianpaolo Basso[3,4,5,6] & Elisabetta De Bernardi[3,4,6]

An assessment scheme is proposed to evaluate GBM gross tumor core and T2-FLAIR hyper-intensity segmentations on preoperative multicentric MR images as a function of tumor morphology and MRI characteristics. 74 gross tumor core and T2-FLAIR hyper-intensity BraTS-Toolkit and DeepBraTumIA automatic segmentations, and 42 gross tumor core neurosurgeon manual segmentations were accordingly evaluated. Brats-Toolkit and DeepBraTumIA generally provide accurate segmentations, particularly for the most common round-shaped or well-demarked tumors, where: (1) gross tumor segmentation correctly includes necrosis and contrast enhanced tumor in 100% and 97.06% of cases (vs. 73.68% for manual segmentation) and wrongly includes healthy or non-tumor related tissues in 2.94% and 20.59% of cases (vs. 10.53% for manual segmentations); (2) T2-FLAIR hyper-intensity segmentations completely includes edema in 88.24% of cases for both software. MR image quality has little impact on the segmentation performance on these tumors. Conversely, on less common tumors with more complex tissue distribution and infiltrative behavior, manual segmentation works better than BraTS-Toolkit and DeepBraTumIA, and image quality has a larger impact on automatic segmentation performance. BraTS-Toolkit and DeepBraTumIA gross tumor segmentation properly includes necrosis and contrast enhanced areas in 50% and 37.50% of cases (vs. 66.67% for manual segmentation), all corresponding to higher image quality; T2-FLAIR hyper-intensity segmentation wrongly includes necrosis and contrast enhanced areas in 37.50% and 50% of cases.

Keywords Glioblastoma, Automatic segmentation, Surgical planning, BraTS-Toolkit, DeepBraTumIA, Manual segmentation, Segmentation assessment scheme

Glioblastoma (GBM) is the most prevalent type of malignant brain tumor[1,2] and it is biologically characterized by two regions of interest: the contrast enhancing (CE)[3] region and the peritumoral region[4,5]. The CE region is the main surgical target and encompasses necrotic areas and highly proliferating cells[6,7]. The peritumoral region is still an area of ongoing research[8–13], but it is believed to play a significant role in brain tumor progression[5,11,12,14]. The peritumoral zone (PBZ) is typically defined on Magnetic Resonance (MR) images as a T2-FLAIR hyperintense region consisting of edematous or infiltrated healthy brain[15–17].

The current standard of care for GBM is maximal safe resection followed by adjuvant chemo-radiation-therapy[6,18–20]. Supramarginal resection[21–27] (removal of both CE and non-enhancing regions) can be offered in few cases, according to the distribution of MR T2-FLAIR hyper-intensities[24,28,29]. The extent of tumor resection (EOR) is the most important prognostic factor, and it is related to the initial gross tumor volume (GTV). Additionally, the volume of the peritumoral T2-FLAIR hyper-intensity also influences the extent of supramarginal resection. As a result, the volume of brain tissue to be removed is an essential consideration in planning the surgical target and adjuvant radiotherapy, measuring the EOR, and counseling on what the neurosurgeon can offer or not to the patient. Conventionally, the CE region is manually segmented by

[1]PhD program in Neuroscience, School of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy. [2]Neurosurgery, Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy. [3]CENTRO STUDI DIPARTIMENTALE GBM-BI-TRACE (GlioBlastoMa-BIcocca-TRAnslational-CEnter), Milan, Italy. [4]School of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy. [5]Neuroradiology, Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy. [6]Bicocca Bioinformatics Biostatistics and Bioimaging B4 Center, School of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy. ✉email: valeria.cerina@unimib.it; v.cerina2@campus.unimib.it

neurosurgeons on preoperative CE T1-weigthed (T1ce) MRI scans, also for comparisons with post-operative scans. The peritumoral region is instead not always segmented in clinics, according to what is considered the treatment target and the size and extension of the PBZ.

Artificial intelligence (AI) is driving a technological revolution in medicine, helping clinicians to perform repetitive tasks more accurately and efficiently, limiting inter-observer variability. The development of AI-based automatic GBM segmentation has been actively addressed in literature since 2012[30–34]. A particularly interesting contribution came from the MICCAI - Multimodal Brain Tumor Image Segmentation Benchmark (BraTS) Challenge (http://www.braintumorsegmentation.org/), whose better performing deep learning algorithms have been collected into a freely available GBM segmentation tool, BraTS-Toolkit[35]. Other GBM segmentation initiatives are DeepBraTumIA (https://www.nitrc.org/projects/deepbratumia/) and the more recent Federated Tumor Segmentation (FeTS)[31]. Automatic segmentation tools could provide important support in the GBM contouring, helping in speeding up the process and in reducing the inter-observer variability. Another advantage of automatic segmentation tools is the possibility to exploit information contained in other scans besides T1ce, like T2 and T2-FLAIR. Lastly, not only the CE region (tumor core) but also the peritumoral T2-FLAIR hyper-intensity area can be easily segmented.

Automatic segmentation tools used in medical imaging are typically evaluated by computing the Dice Index or the Hausdorff Distance (HD) Vs a reference standard obtained as a consensus of expert operators' manual contours. The Dice Index is a measure of similarity or overlap between the segmented volume and the reference; the HD reflects the spatial distance between the segmented and reference surfaces. In the GBM context, Dice Index and HD are separately calculated for the CE and the peritumoral regions. These indexes have been widely used to compare and rank automatic segmentation strategies[30,36].

However, it's worth noticing that in a heterogeneous tumor like GBM, a similar Dice (or HD) score between a segmentation and the reference contour can correspond to different types of errors that have varying impacts on surgical resection planning. For example, the same Dice score may indicate over-inclusion or under-inclusion of tissues. Furthermore, the Dice Index and HD do not provide any information about the nature of the tissues being included or excluded. In a recent work, Kofler et al. showed that standard quantitative segmentation quality metrics correlate only moderately with experts' qualitative evaluations in glioma segmentation, particularly for the CE region[37]. To address this point, a more comprehensive evaluation of GBM segmentations is needed, especially for the benefit of neurosurgeons.

The first aim of our work is to propose a specific assessment scheme for evaluating the segmentation of GBM tumor core (CE and non-CE) and T2-FLAIR hyper-intensity. This scheme does not rely on a quantitative comparison with a reference, but on the visual evaluation of a senior neuroradiologist to determine if the correct tissues have been included or excluded from the segmentation.

The second objective of our work is to employ the proposed assessment scheme to evaluate the reliability of two state of the art automatic segmentation software in comparison to neurosurgeon manual segmentations, as a function of tumor morphology and MR image characteristics. The assessment scheme was applied to a set of preoperative GBM MR images from various centers both to tumor core manual segmentations provided by dedicated neurosurgeons and to tumor core and T2-FLAIR hyper-intensity segmentations automatically obtained with BraTS-Toolkit and DeepBraTumIA. Images were divided into groups, according to tumor morphology and MR sequence characteristics. The intention for this evaluation is to provide insight into the types of tumors and images where AI-based segmentation methods are currently applicable for use by neurosurgeons, as well as the circumstances in which further refinement by the neurosurgeon or segmentation software improvement are still required.

Materials and methods

Dataset

An multicentric dataset of 143 GBM patients surgically treated from 2018 to 2023 by the Neurosurgery Unit at Fondazione IRCCS San Gerardo dei Tintori in Monza (Italy) was considered. The dataset included patients more than 18 years old that underwent surgery both biopsy or surgical resection. Clinical, demographic and radiological data were collected retrospectively including: age at diagnosis, sex, side and site of the tumor. Preoperative MR studies acquired with heterogeneous MRI protocols but including pre-contrast T1-weighted (T1), post-contrast T1 (T1ce), T2-weighted (T2) and T2-FLAIR (FLAIR) sequences free of artefacts were selected, for a total of 74 studies. Of the 74 patients, median age was 62 years (22–87), 41 (55%) patients were male. The tumor was located in the frontal lobe for 22 (30%) patients, in the temporal lobe for 18 (24%) patients, in the insular lobe for 6 (8,1%) patients, in the occipital lobe for 7 (9.5%) patients, in the parietal lobe for 11 (14.9%) patients, in the corpus callosum for 5 (6.8%) patients, in the thalamus for 3 (4%) patients and bilateral in 2 (2.7%) patients. Tumor was located in the right hemisphere in 40 cases (54%), in the left hemisphere in 27 cases (36%), along midline structures in 4 cases and multifocal in 3 cases.

Ethics

The study was performed in accordance with the ethical standards of the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The results of this paper are part of the study approved by the ethic committee of Comitato Etico Territoriale Lombardia 3 under the study GLIOMA_NEURO (protocol number 464–09/08/2023). All patients underwent diagnostic and therapeutic procedures approved for their specific disease and part of the current clinical practice. Each patient signed an informed consent form during the hospital recovery for use of clinical, histological and radiological data for research purposes according to the hospital policy. All the data collected during the study were completely anonymized after collection.

Tumor classification

MRI studies have been classified into five typologies depending on tumor characteristics and tissues distribution on images.

- Type A: Monocentric tumor where the tumor mass (necrosis and possible core non-enhancing) is all included into contrast-enhanced continuous margins.
- Type B: Irregular tumor bulk distribution; necrotic spots are linked by non-enhancing tumor areas.
- Type C: Monocentric tumor where the tumor mass extends also beyond contrast-enhancing continuous margins.
- Type D: Multiple well-defined tumor areas, consisting of necrosis, enhancing and non-enhancing tumor; these areas are surrounded by healthy/edematous tissues and can be located also in different hemispheres. Tumoral spots are considered as separate treatment targets, therefore only the surgical main target has been assessed.
- Type E: This tumor typology does not exhibit CE tumoral tissues or CE hyperintensity is faded and margins are not clearly defined. All tumoral tissues are considered as non-enhancing tumor.

Image quality subgrouping

MRI studies have been also divided into groups, according to MRI sequence characteristics and parameters. Two aspects were considered: magnetic field strength (3T or 1.5T) and voxel dimensions. If axial pixel size and spacing between adjacent slices were of the same order of magnitude and both less than 1.5 mm, the voxel was considered isotropic, otherwise it was considered anisotropic. Since T1ce sequences were always isotropic and T2 sequences always anisotropic (generally T2-TSE acquired on the axial plan), the division into groups relied on T1 and T2-FLAIR quality. Six groups were defined, enumerated from higher (Group 1) to lower (Group 6) image quality. On 3T images, T2-FLAIR is always isotropic, therefore no counterpart of the 1.5T Group 5 and Group 6 has been determined. From higher to lower quality, groups are defined as follows:

- Group 1: Magnetic field strength 3T; Isotropic sequences T1, T1ce, FLAIR; Anisotropic sequences T2.
- Group 2: Magnetic field strength 1.5T; Isotropic sequences T1, T1ce, FLAIR; Anisotropic sequences T2.
- Group 3: Magnetic field strength 3T; Isotropic sequences T1ce, FLAIR; Anisotropic sequences T1, T2.
- Group 4: Magnetic field strength 1.5T; Isotropic sequences T1ce, FLAIR; Anisotropic sequences T1, T2.
- Group 5: Magnetic field strength 1.5T; Isotropic sequences T1, T1ce; Anisotropic sequences FLAIR, T2.
- Group 6: Magnetic field strength 1.5T; Isotropic sequences T1ce; Anisotropic sequences T1, FLAIR, T2.

Manual segmentation—current clinical practice

42/74 studies were segmented manually, according to the current clinical practice. Tumor volume was assessed on preoperative MRI using BrainLab™ segmentation software (BrainLab™, Germany)[38]. Tumors were segmented manually without using the semiautomated tool[39]. All segmentations were made on T1ce sequence in three projections, and then refined on FLAIR and T1 sequences to include only the contrast enhanced tumor and the tumoral part, without including perilesional edema. All tumor volumes were expressed in cubic centimeters (cc). All the segmentations were performed in double-blind modality by a dedicated Neurosurgeon and then reviewed and corrected by a second dedicated one.

Automatic segmentation

All 74 studies were automatically segmented with BraTS-Toolkit and DeepBraTumIA, which both require as input four MRI sequences (T1ce, T1, T2 and T2-FLAIR scans).

BraTS-Toolkit segmentation

The BraTS-Toolkit consists of three analysis steps, i.e. BraTS-Preprocessor, BraTS-Segmentor and BraTS-Fusionator[35]. The Preprocessor provides image co-registration, conversion to BraTS space and defacing/skull stripping by means of HD-BET for GPU mode. BraTS-Segmentor represents an interface for the BraTS algorithmic repository[40] where all Docker images containing Deep Learning models and correspondent codes from BraTS challenges are stored. The Segmentor allows the user to run selected Docker images, producing multiple segmentations. The Fusionator module provides two algorithms, Majority Voting (MAV) and Selective and Iterative Method for Performance Level Estimation (SIMPLE), to combine segmentations obtained with Segmentor in a unique final segmentation.

Among the Docker images available at November 2023 (https://hub.docker.com/u/brats) we made a 3-step selection, excluding: (1) older releases of more recent Docker images; (2) Docker images not working on our workstation; (3) Docker images that applied to a group of 4 patients provided segmentations judged strongly incorrect by both the neuroradiologist and the neurosurgeon. 8 Docker images were selected to be used in the present study (Docker images created by F.Isensee, J. Haozhe, Y. Wang, Y. Yuan, X. Feng, Y. Zha, N. Nuechterlein). The SIMPLE method for segmentation fusion has been chosen (i.e. each segmentation is compared to the current consensus fusion and the resulting Dice overlap score represents the weight for MAV).

According to BraTS challenges since 2017 onwards[40] three segmented Regions of Interest (ROIs) are provided for each study:

- LABEL_1: this label comprises the necrotic core, or necrocyst, that resides within the enhancing rim of high grade gliomas and the non-enhancing gross abnormality (NET – Non Enhancing Tumor) that resides inside enhancing tumor margins or can be identified as clearly distinguishable from the surrounding vasogenic

edema on T2. Before BraTS-2017, NET was treated as a separate label (LABEL_3), but since its identification is challenging and could be overestimated by annotators, it was decided to combine it with the necrotic label.
- LABEL_2: this region comprises the hyper-intense T2-FLAIR tumor related area that includes edematous and invaded tissue, i.e. the NET present in the peritumoral area.
- LALEL_4: describes the T1ce enhancing regions that can be recognized within the gross tumor abnormality, but that does not comprise the necrotic center.

DeepBraTumIA segmentation

DeepBraTumIA is a deep learning-based software tool for GBM automated segmentation (https://www.nitrc.org/projects/deepbratumia/). It incorporates an image quality control step (which was disabled to work on the entire dataset), an image registration and MNI normalization step, a skull stripping step, and finally a voxel-wise segmentation step that produces three labels as output: Necrotic label, CE label and Edema label. NET assignment is not explicitly described; generally, NET areas internal or proximal to the CE area are assigned to the necrotic core label, while NET areas distal from the CE are assigned to the Edema label.

Segmentation quality assessment

Manual and automatic segmentations were evaluated by a senior neuroradiologist, utilizing the ITK-snap software as a visualizer. Segmentations were presented randomly, mixing tumor typologies, image quality groups and segmentation types (manual or automatic). As the objective of this study is to assess GBM segmentations to assist neurosurgeons in surgical planning, BraTS Toolkit LABEL_1 and LABEL_4 (and similarly DeepBraTumIA Necrotic label and CE label) were fused into a unique label, called TUMOR_CORE_LABEL, corresponding to the gross Tumor Core, i.e. the surgical-treatment target usually segmented manually. BraTS Toolkit LABEL_2 and DeepBraTumIA Edema label were instead indicated as FLAIR_Hyperintensity_LABEL.

TUMOR_CORE_LABEL and FLAIR_Hyperintensity_LABEL were independently evaluated. Automatic TUMOR_CORE_LABEL and manual segmentation were assessed with the same evaluation scheme. In the segmentation quality evaluation, the presence/absence of four tissue components in ROIs was considered: Necrotic Core (Necrosis), T1ce CE regions, FLAIR Tumor Related region comprising NET and Edema, and finally NOT-Related tumor signal, e.g. non pathologic tissue, inflammatory damages, gliosis etc.

The evaluation scheme is composed of a 5-point quality scale, as well as label-specific inquiries regarding the proper inclusion/exclusion of various tissue components.

Five-point quality scale definition

The goodness of each segmented ROI in terms of margins definition accuracy and proper tissue inclusion has been preliminarily assessed through the 5-point scale (1 = Correctly overlapped to the correspondent region highlighted by the neuroradiologist; 2 = Not-perfectly overlapped but including all pathologic tissue; 3 = Miss of pathologic tissue inclusion, not compromising overall tumor core segmentation; 4 = Miss of pathologic tissue inclusion resulting in substantial of tumor core miss; 5 = Misclassification of normal tissue as tumor core, of edema as tumor core, of tumor core as edema). The final clinical goal of surgical treatment planning has been considered to assign each score.

Label-specific inquires

Specific inquires for a detailed quality assessment of TUMOR_CORE_LABEL and FLAIR_Hyperintensity_LABEL are presented in Table 1. Given that the NET assignment to segmentation labels may differ in the two automatic software, we decided to limit FLAIR_Hyperintensity_LABEL inquires to edema and to create a specific inquire for NET, evaluating the capacity of the whole segmentation (i.e. union of TUMOR_CORE_LABEL and FLAIR_Hyperintensity_LABEL) to encompass it, including intratumoral NET, peritumoral NET and NET areas distant from the core.

Results

Tumor classification

The overall dataset was classified into the defined 5 tumor typologies, according to tumoral tissue distribution on images (Type A: 34 studies; Type B: 8 studies; Type C: 20 studies; Type D: 9 studies; Type E: 3 studies). Figure 1 reports an example of each type, together with the corresponding BraTS-Toolkit segmentation.

TUMOR_CORE_LABEL		
COMPLETELY includes all Necrosis & CE	WRONGLY includes healthy or non-tumor related tissue	WRONGLY includes edema
FLAIR_Hyperintensity_LABEL – edema		
COMPLETELY includes all edema	WRONGLY includes healthy or non-tumor related tissue	WRONGLY includes Necrosis & CE
NET – INFILTRATED TUMOR		
NET present in the TUMOR_CORE_LABEL	NET present in the FLAIR_Hyperintensity_LABEL	NET COMPLETELY included in labels union

Table 1. Label-specific inquires for segmentation quality assessment specific for TUMOR CORE_LABEL (first row), FLAIR_Hyperintensity_LABEL edema (second row) and NET (third row).

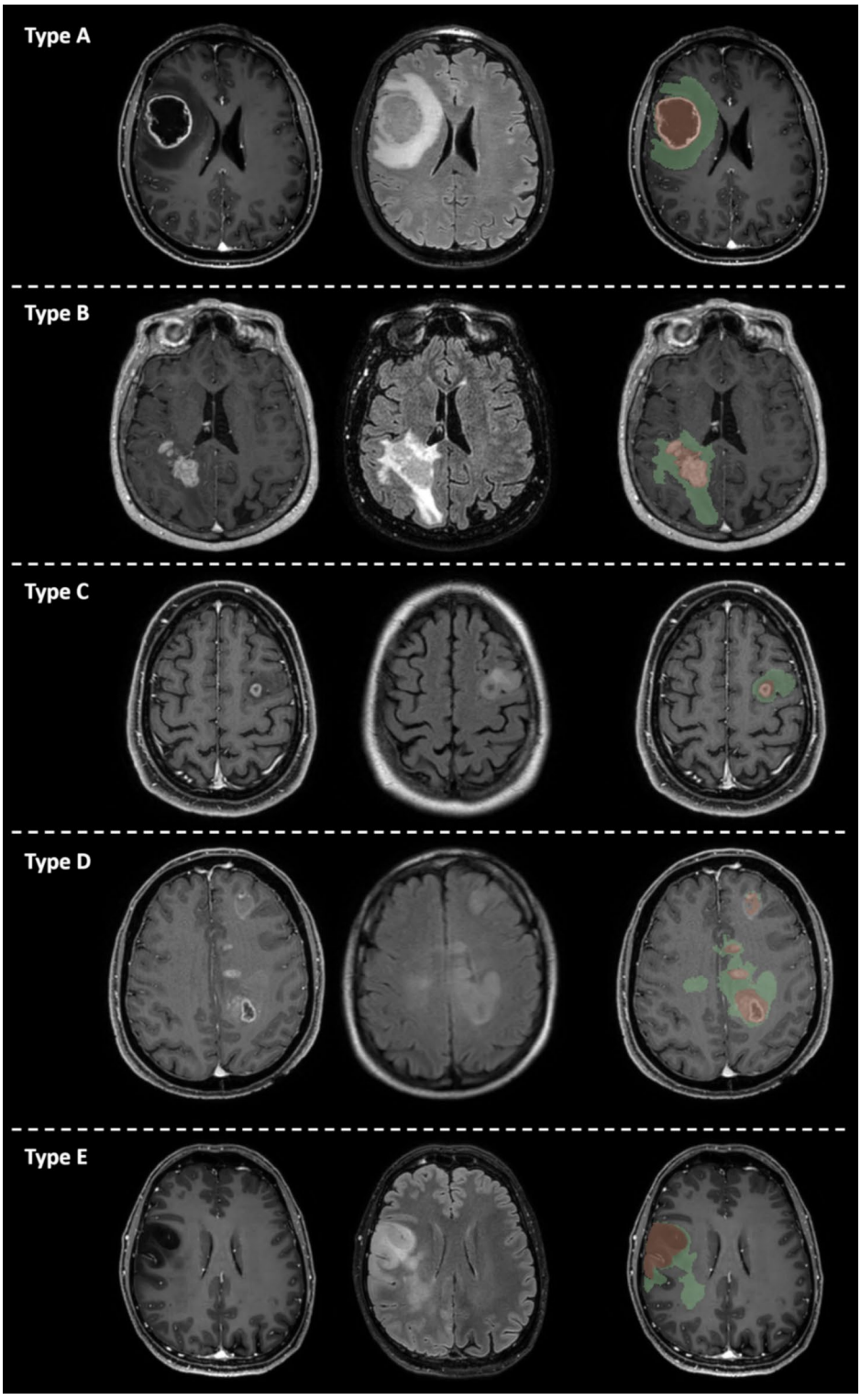

Fig. 1. Each tumor typology is showed: T1ce (left), T2-FLAIR (middle) and BraTS segmentation (right). TUMOR_CORE_LABEL is showed in red color and FLAIR_Hyperintensity_LABEL in green color. For each typology, an axial slice is reported, aiming at showing each particular tumor distribution.

Image quality subgrouping

As to the subdivision of MRI studies into image quality groups, we obtained the following distributions, showing a non-homogeneity of MRI protocols in the considered dataset:

- Group 1 ($n = 12$): 7 Type A, 0 Type B, 3 Type C, 1 Type D, 1 Type E.

- Group 2 ($n=24$): 10 Type A, 4 Type B, 7 Type C, 2 Type D, 1 Type E.
- Group 3 ($n=4$): 1 Type A, 0 Type B, 2 Type C, 1 Type D, 0 Type E.
- Group 4 ($n=15$): 8 Type A, 2 Type B, 4 Type C, 1 Type D, 0 Type E.
- Group 5 ($n=4$): 2 Type A, 0 Type B, 2 Type C, 0 Type D, 0 Type E.
- Group 6 ($n=15$): 6 Type A, 2 Type B, 2 Type C, 4 Type D, 1 Type E.

Automatic segmentation assessment

Figure 2 reports the results obtained applying the 5-point quality scale to the BraTS-Toolkit and DeepBraTumIA segmentations of the 74 studies. The distribution of the five scores is represented per tumor typology, on the left panel for TUMOR_CORE_LABEL, on the right panel for FLAIR_Hyperintensity_LABEL. Type A and Type C obtained the highest number of perfect overlap (score = 1).

To represent the 5-point quality scale results as a function of image quality, a cut-off score was set to 3, considering scores 1–3 as good. Segmentation goodness is reported in Fig. 3. Numbers on cells refer to the number of good segmentations / the number of studies assigned to the cell. Cell colors refer to the good segmentation's percentage. Black cells do not include any study. Almost all groups report a high percentage of good segmentations. No substantial dependence of segmentation goodness on image quality results for the two most common typologies (Type A and Type C); less common typologies (Type B and Type E) show lower goodness percentages.

Table 2 reports the answers to label specific inquires, as a function of tumor typology for TUMOR_CORE_LABEL, FLAIR_Hyperintensity_LABEL edema and NET. Results are presented as percentages with respect to the total number of patients belonging to the cell.

Manual segmentation assessment

The evaluation scale proposed for the TUMOR_CORE_LABEL has been applied also to the manual segmentation on the subgroup of 42 patients. Figure 4 shows score occurrences per tumor typology, reporting also BraTS-Toolkit and DeepBraTumIA TUMOR_CORE_LABEL results on the same patient's subgroup, for comparison. Manual segmentation and automatic segmentation perform similarly on Type C, differently on the other types, particularly on the most common Type A, where BraTS-Toolkit surpasses manual segmentation and on the less common Types B and E, where manual segmentation performs better.

Figure 5 reports the percentage of studies achieving a good (1–3) image quality score for each combination of tumor typology and image quality.

Table 3 Reports the answers to label specific inquires for manual segmentation, as a function of tumor typology.

Discussion

The automatic segmentation of GBM MR images could play a fundamental role in the clinical work-up by aiding in the identification of tumoral areas to plan surgery and residual tumor to plan radiation therapy treatment. Moreover, an accurate and observer-independent segmentation is essential in research, as tumor volumes, peritumoral area volumes and extent of resection are fundamental factors for population stratification in every

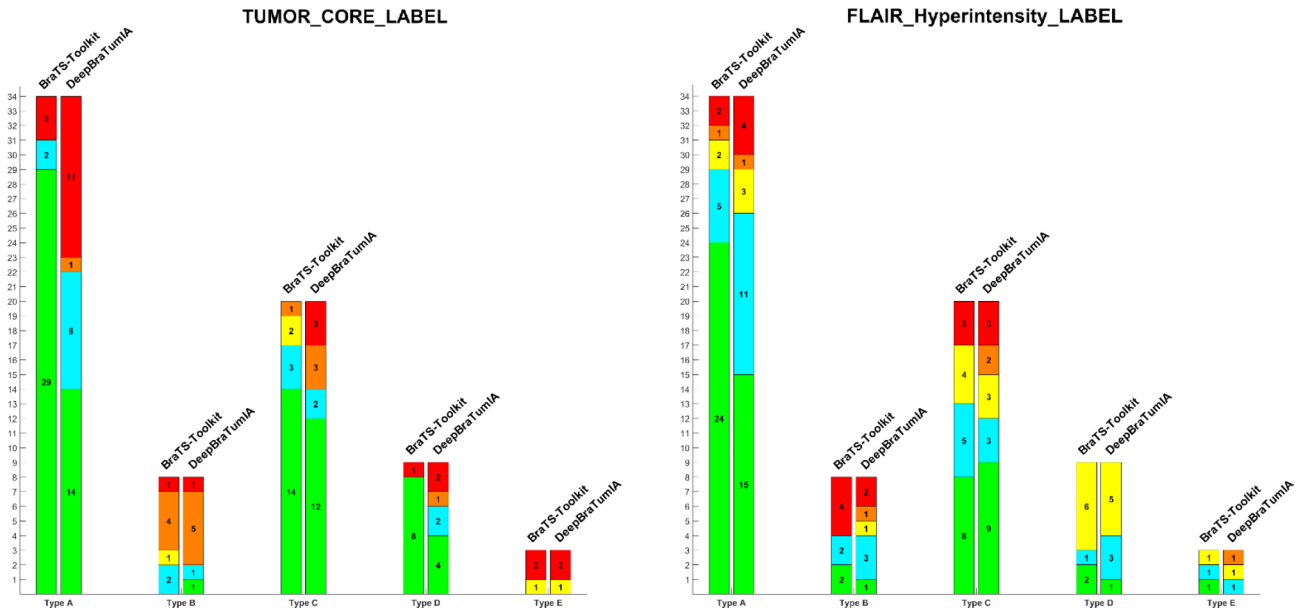

Fig. 2. Specific score occurrence for TUMOR_CORE_LABEL (left plot) and FLAIR_Hyperintensity_LABEL (right plot). Bars refer to tumor typologies. Colors refers to evaluation scale scores: 1 (green), 2 (light blue), 3 (yellow), 4 (orange), 5 (red).

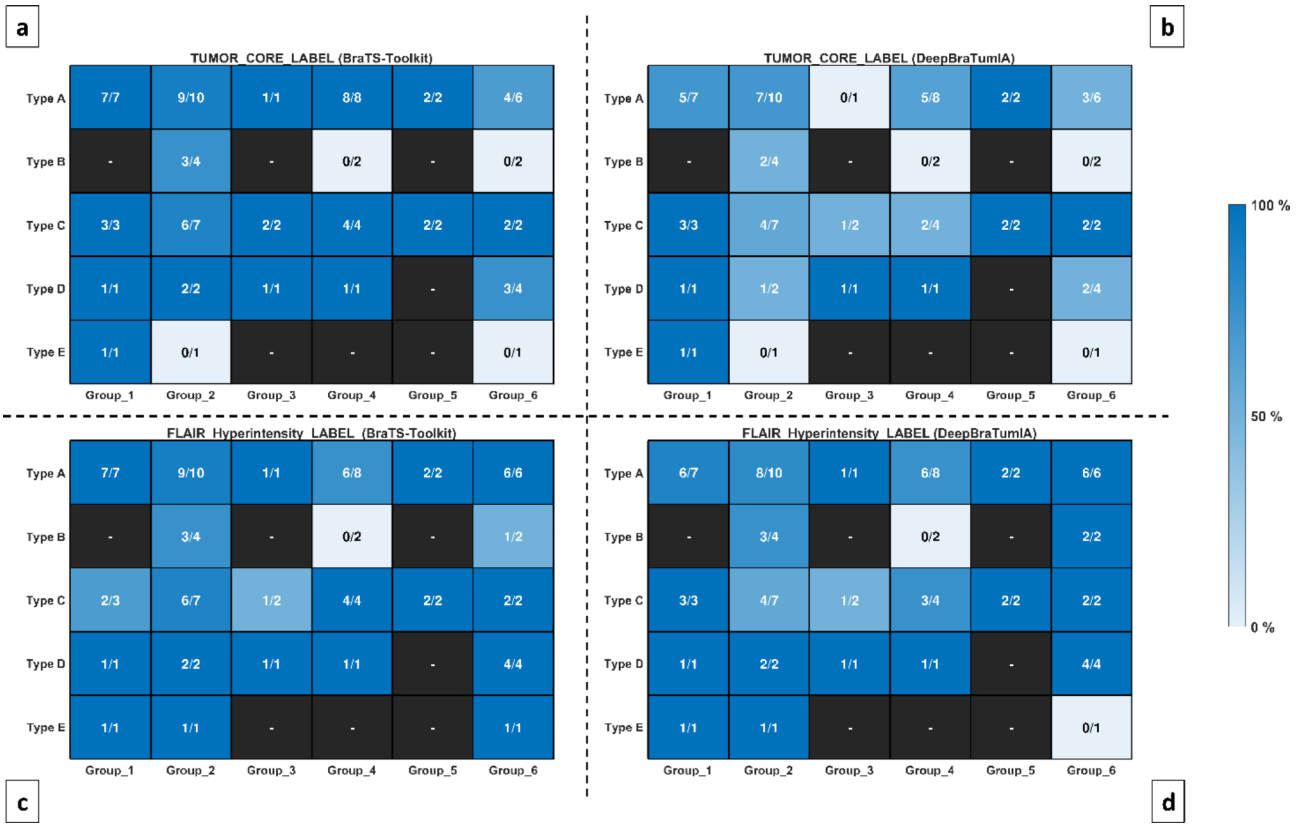

Fig. 3. Four tables showing the segmentation quality for TUMOR_CORE_LABEL obtained with BraTS-Toolkit (**a**), TUMOR_CORE_LABEL obtained with DeepBraTumIA (**b**); FLAIR_Hyperintensity_LABEL obtained with BraTS-Toolkit (**c**), FLAIR_Hyperintensity_LABEL obtained with DeepBraTumIA (**d**). In each table, for each tumor typology (rows) and image quality group (columns), patients with good segmentation according to the selected cut-off score (= 3) are reported against the total patients belonging to that typology/group. Cell colors refer to the good segmentation's percentage. Black cells do not include any patient.

		TUMOR_CORE_LABEL			FLAIR_Hyperintensity_LABEL - edema			NET – INFILTRATED TUMOR		
		COMPLETELY includes all Necrosis & CE	WRONGLY includes healthy or non-tumor related tissue	WRONGLY includes edema	COMPLETELY includes all edema	WRONGLY includes healthy or non-tumor related tissue	WRONGLY includes Necrosis & CE	NET present in the TUMOR CORE LABEL	NET present in the FLAIR Hyperintensity LABEL	NET COMPLETELY included in labels union
Type A ($n = 34$)	BraTS Toolkit	100.00%	2.94%	5.88%	88.24%	5.88%	0.00%	5/34	0/34	100.00%
	DeepBraTumIA	97.06%	20.59%	0.00%	88.24%	5.88%	5.88%	5/34	0/34	100.00%
Type B ($n = 8$)	BraTS Toolkit	50.00%	12.50%	0.00%	100.00%	25.00%	37.50%	4/8	6/8	100.00%
	DeepBraTumIA	37.50%	12.50%	0.00%	87.50%	0.00%	50.00%	4/8	7/8	50.00%
Type C ($n = 20$)	BraTS Toolkit	85.00%	0.00%	0.00%	95.00%	5.00%	10.00%	5/20	20/20	80.00%
	DeepBraTumIA	85.00%	10.00%	5.00%	80.00%	0.00%	15.00%	9/20	18/20	65.00%
Type D ($n = 9$)	BraTS Toolkit	100.00%	11.11%	0.00%	88.89%	0.00%	0.00%	2/9	9/9	22.22%
	DeepBraTumIA	88.89%	11.11%	0.00%	100.00%	0.00%	0.00%	3/9	8/9	11.11%
Type E ($n = 3$)	BraTS Toolkit	100.00%	66.67%	33.33%	100.00%	0.00%	0.00%	2/3	3/3	100.00%
	DeepBraTumIA	100.00%	33.33%	0.00%	100.00%	0.00%	0.00%	3/3	3/3	66.67%

Table 2. Answers to specific inquires for TUMOR_CORE_LABEL and FLAIR_Hyperintensity_LABEL for automatic segmentation. Results are expressed as percentage to the total number of patients belonging to the cell. The non enhancing tumor (NET - infiltrated tumor) compartment inclusion/exclusion is separately described, as some NET can be present and included in both labels.

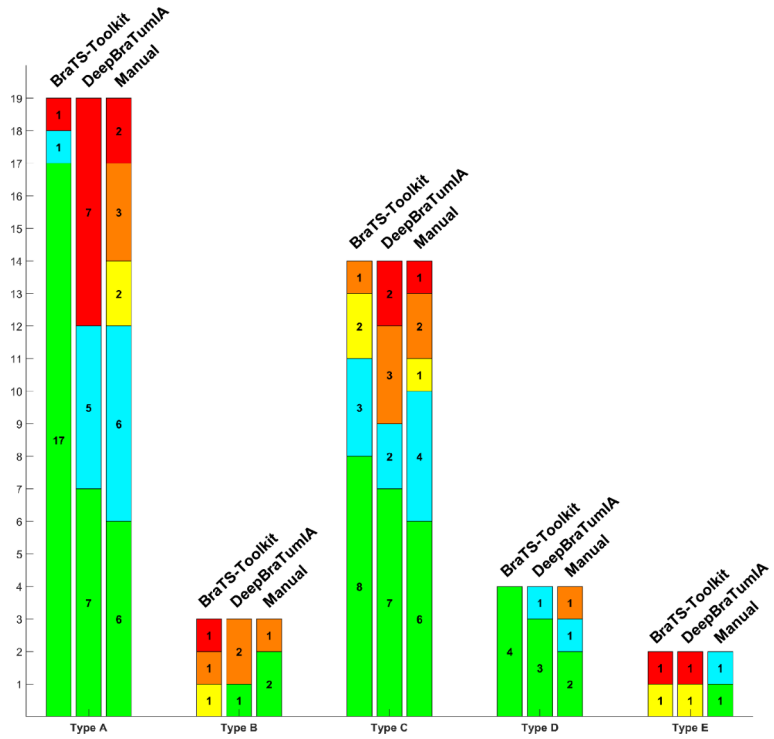

Fig. 4. Score occurrence on the subgroup of 42 manually segmented studies. BraTS-Toolkit TUMOR_CORE_LABEL (left bar), DeepBraTumIA TUMOR_CORE_LABEL (central bar) and manual segmentation (right bar) are compared the five tumor typologies. Colors refers to evaluation scale scores: 1 (green), 2 (light blue), 3 (yellow), 4 (orange), 5 (red).

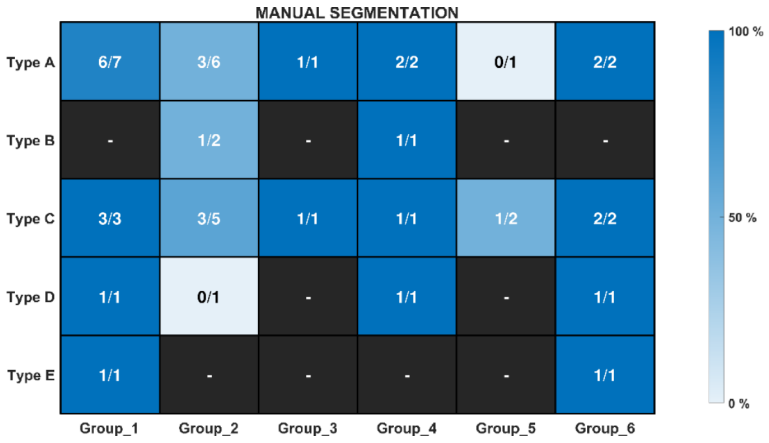

Fig. 5. For each tumor typology (rows) and image quality group (columns), patients with good manual segmentation according to the selected cut-off score (= 3) are reported against the total patients belonging to that typology/group. Cell colors refers to the good segmentation's percentage. Black cells do not include any patient.

neuro-oncological study. In this context, a clinically proven and valid automatic segmentation software could accelerate surgical planning and facilitate standardized data sharing among centers.

The first aim of this study was to define an assessment scheme to evaluate the quality of GBM tumor core and T2-FLAIR hyperintensity contours, for their potential use in neuro-oncological treatment planning. GBM core segmentation is typically performed for surgical planning and measuring the extent of resection. The surgical management of the peritumoral FLAIR hyperintensity area, that can include vasogenic edema as well as infiltrative non-enhancing tumor, is a topic of ongoing debate[6]. The availability of reliable contours of this area could lead to the development of new strategies, such as dividing it into its components (edema and infiltrative peritumoral areas)[41] for expanding resection margins (FLAIRectomy).

	Manual segmentation			
	COMPLETELY includes all Necrosis & CE	WRONGLY includes healthy or non-tumor related tissue	WRONGLY includes edema	NET present in the Tumor Core
Type A ($n=19$)	73.68%	10.53%	0.00%	1/19
Type B ($n=3$)	66.67%	0.00%	0.00%	1/3
Type C ($n=14$)	78.57%	7.14%	0.00%	5/14
Type D ($n=4$)	75.00%	0.00%	0.00%	1/4
Type E ($n=2$)	100.00%	0.00%	50.00%	2/2

Table 3. Answers to specific inquires for manual segmentation of the tumor core. Results are expressed as percentage on the total number of subjects per tumor typology. The Non Enhancing Tumor (NET - Infiltrated tumor) compartment inclusion is considered as a separate feature.

Validating automatic segmentation software is essential for transitioning it to clinical practice. Conventional quantitative indices, such as DICE and HD, are reliable for comparing software to determine the most accurate one and to prevent inter-observer errors. However, they do not provide information on the relevance of including or excluding different tissue types in the segmented regions. A study specifically focused on GBM segmentation showed a Pearson correlation between expert assessment and DICE equal to 0.36 for enhancing tumor, 0.37 for tumor core, 0.37 for necrosis, 0.44 for edema and 0.5 for whole tumor[37]. To address these aspects, we propose a segmentation assessment scheme that does not rely on a comparison with a reference contour but on an expert neuroradiologist's *a posteriori* visual analysis to assess how the various GBM tissue components are correctly included or not in the segmentation and if the segmented area wrongly includes healthy tissues. The assessment scheme consists of a 5-point quality score supported by label-specific inquiries for gross tumor core and T2-FLAIR hyper intensity on tissue inclusion-exclusion. The neuroradiologist must answer yes or no to the label-specific inquires, so in themselves the inquires do not generate a score if applied to a single MRI study. Scores can be obtained by applying the inquires to groups of patients and evaluating the percentage of positive or negative evaluations on the whole sample. We believe that such an assessment will provide the neurosurgeon with more comprehensive insights into the usability of automatic segmentation tools in resection planning compared to the Dice index. Our score may be useful for translating automatic segmentation software into clinical practice, offering a more targeted tool for setting-up or testing software accuracy, which is essential for effective application in clinics.

The proposed assessment scheme was then applied to evaluate BraTS-Toolkit and DeepBraTumIA automatic segmentations and to compare them with manual segmentation performed in double-blind modality by a dedicated Neurosurgeon and then reviewed and corrected by a second dedicated one on a multicentric preoperative patient cohort. Specifically, 74 tumor cores and T2-FLAIR hyper intensity BraTS-Toolkit and DeepBraTumIA automatic segmentations, and a subgroup of 42 tumor core manual segmentations were evaluated. MRI studies were classified into 5 tumor typologies and into 6 MR image parameter groups with the specific aim to assess segmentation accuracy as a function of tumor morphology and image quality.

The most evident result of this work is that "bulky" round shaped or well demarked tumors can be segmented with a very high precision by automatic tools, particularly by BraTS-Toolkit, while tumors with a less defined shape and with an infiltrative behavior are segmented with lower precision in favor of better manual performances. More specifically, on the most common tumors (Type A: monocentric tumors with contrast enhanced margins including the bulk tumor; Type C: monocentric tumors with bulk surpassing the contrast enhanced margins; Type D: multiple bulk tumor areas, each surrounded by healthy/edematous tissue), BraTS-Toolkit works very well, even better than manual segmentation, and nearly independently of image quality (i.e. magnetic field intensity and voxel dimension). On the most common Type A tumors, the union of LABEL_1 and LABEL_4 (here defined as TUMOR_CORE_LABEL) completely includes necrosis and contrast enhanced tumor in 100% of cases, better than the manual segmentation, which performs correctly in 73.68% of cases. The wrong inclusion of healthy or non-tumor related tissues happens in 2.94% of cases for BraTS-Toolkit and in 10.53% of cases for manual segmentations. BraTS-Toolkit succeeds in completely including edema in LABEL_2 (here defined as FLAIR_Hyperintensity_LABEL) in 88.24% of cases and in completely segmenting NET in 100% of cases. DeepBraTumIA performs similarly to BraTS-Toolkit on most of the label specific inquires. However, the TUMOR_CORE_LABEL wrongly includes healthy or non-tumor related tissues in 20.59% of Type A tumors and in 10% of Type C tumors.

On less common and more complex tissue distribution tumors (Type B: irregular bulk distribution; Type E: absence of contrast enhanced area), manual segmentation works better than automatic segmentation, probably due to the limited quantity of tumors of this type present in automatic segmentation software training sets. Furthermore, image quality seems to have a larger influence on the automatic segmentation performance on these tumors. On Type B tumors, BraTS-Toolkit and DeepBraTumIA TUMOR_CORE_LABELs properly include necrosis and contrast enhanced areas in 50% and 37.50% of cases respectively, which all correspond to a higher image quality. Necrosis and contrast enhanced areas are conversely wrongly included in FLAIR_

Hyperintensity_LABEL in 37.50% of cases for BraTS-Toolkit and in 50% of cases for DeepBraTumIA. On all Type E tumors, both automatic software provide an accurate FLAIR_Hyperintensity_LABEL and a TUMOR_CORE_LABEL that completely includes necrosis and CE. However, BraTS-Toolkit wrongly includes healthy tissues in 66.67% and edema in 33.33% of cases. Moreover, on Type E some T2 hyper-intensities related to leukoaraiosis were segmented as pieces of tumor. DeepBraTumIA better excludes healthy tissues and edema from the TUMOR_CORE_LABEL but succeeds in completely segmenting NET in 66.7% of cases.

This study presents several limitations. The primary limitation is the reliance on a single expert for ratings and on a single BrainLab™ manual segmentation. This constraint inherently limits the ability to assess inter-rater variability in both segmentation quality assessment and manual segmentation. As to inter-rater variability in GBM manual segmentation, previous literature has reported an inter-rater median Dice of 0.83 for CE[42] and >0.85 for the whole tumor core[43]. To the best of our knowledge, no studies have assessed the manual segmentation inter-rater variability in terms of qualitative assessment by an external expert rater. In this study, it was not possible to obtain manual contours from a second operator. The evaluated contours were those created during clinical activity for surgical planning. These were initially created by a dedicated neurosurgeon and subsequently reviewed and adjusted by a second dedicated neurosurgeon, as per clinical practice. We think that an evaluation scheme such as the one proposed in this study may facilitate the implementation of inter-rater variability assessments in terms of qualitative evaluations by external expert raters.

Regarding inter-rater variability in segmentation quality assessment, the literature is still limited. From the analyses conducted by Kofler et al.[37], who requested 15 expert operators to evaluate a manual segmentation, 2 automatic segmentations, and the consensus between multiple automatic segmentations on a 1–5 Likert scale, it can be inferred that the variability in the evaluation of manual segmentation is nearly comparable to that in the evaluation of automatic segmentation. The variability appears to be lower when evaluating the consensus between multiple automatic segmentations, which should correspond more closely to the truth. In this study, it was not feasible to have the evaluation of segmentations performed by a second expert operator. The entire segmentation quality assessment was performed by one senior neuroradiologist. To enhance results reliability, subsequent revisions and discussions of all the evaluations were carried out in collaboration with the senior neurosurgeon. However, as this study evaluated a consensus between multiple segmentations (SIMPLE consensus between 8 state of the art Docker images in Brats Toolkit) and a manual segmentation produced by the agreement between two operators, we may argue, based on the aforementioned observations, that the inter-variability around the two evaluations could be similar. These aspects will obviously require verification in future studies.

The second significant limitation of the study is the sample numerosity, which is at the lower bound of acceptability. In particular, the manual segmentation assessment was conducted on a subgroup of 42 patients, corresponding to 57% of the study population. This subgroup comprised studies that were previously manually segmented as part of the neurosurgery clinical workflow. Given the significant time required by manual segmentation, we opted not to request the neurosurgeon to perform additional segmentations specifically for this study. Instead, we utilized the previously segmented sample to draw consistent conclusions. The findings derived from tumor type and image quality groups with smaller sample sizes necessitate confirmation through larger-scale studies. Nevertheless, we posit that the observations made in the larger groups may be considered valid.

Lastly, this study focused exclusively on pre-operative data, as the primary objective was surgical planning. We are conscious of the significance of automatic segmentation in post-operative images for both radiotherapy planning and the development of prognostic recurrence models. Future investigations will extend the focus to include these types of images as well.

Data availability

The datasets generated and/or analyzed during the ongoing GLIOMA_NEURO (protocol number 464 – 09/08/2023) for the current study are available upon reasonable request directly to the corresponding author (Valeria Cerina - valeria.cerina@unimib.it).

Received: 27 August 2024; Accepted: 2 January 2025
Published online: 16 January 2025

References

1. Schaff, L. R. & Mellinghoff, I. K. Glioblastoma and other primary brain malignancies in adults: A review. *JAMA* **329**, 574–587 (2023).
2. Bale, T. A. & Rosenblum, M. K. The 2021 WHO classification of tumors of the Central Nervous System: An update on pediatric low-grade gliomas and glioneuronal tumors. *Brain Pathol.* **32** (2022).
3. Mattei, L., Prada, F., Marchetti, M., Gaviani, P. & DiMeco, F. Differentiating brain radionecrosis from tumour recurrence: A role for contrast-enhanced ultrasound? *Acta Neurochir. (Wien)*. **159**, 2405–2408 (2017).
4. Schoenegger, K. et al. Peritumoral edema on MRI at initial diagnosis: An independent prognostic factor for glioblastoma? *Eur. J. Neurol.* **16**, 874–878 (2009).
5. Long, H. et al. MRI radiomic features of peritumoral edema may predict the recurrence sites of glioblastoma multiforme. *Front. Oncol.* **12**, (2023).
6. Molinaro, A. M. et al. Association of maximal extent of resection of contrast-enhanced and non-contrast-enhanced tumor with survival within molecular subgroups of patients with newly diagnosed glioblastoma. *JAMA Oncol.* **6**, 495–503 (2020).
7. Pessina, F. et al. Maximize surgical resection beyond contrast-enhancing boundaries in newly diagnosed glioblastoma multiforme: Is it useful and safe? A single institution retrospective experience. *J. Neurooncol.* **135**, 129–139 (2023).
8. Choi, Y. et al. Analysis of heterogeneity of peritumoral T2 hyperintensity in patients with pretreatment glioblastoma: Prognostic value of MRI-based radiomics. *Eur. J. Radiol.* **120**, 108642 (2019).
9. D'Alessio, A., Proietti, G., Sica, G. & Scicchitano, B. M. Pathological and molecular features of Glioblastoma and its Peritumoral tissue. *Cancers (Basel)* **11** (2019).

10. Blystad, I. et al. Quantitative MRI for analysis of peritumoral edema in malignant gliomas. *PLoS One* **12**, (2017).
11. Giambra, M. et al. The peritumoral brain zone in glioblastoma: where we are and where we are going. *J. Neurosci. Res.* **101**, 199–216 (2023).
12. Rathore, S. et al. Radiomic signature of infiltration in peritumoral edema predicts subsequent recurrence in glioblastoma: implications for personalized radiotherapy planning. *J. Med. Imaging.* **5**, 021219 (2018).
13. Lemée, J. M. et al. Characterizing the peritumoral brain zone in glioblastoma: A multidisciplinary analysis. *J. Neurooncol.* **122**, 53–61 (2015).
14. Xing, Z. et al. Predicting glioblastoma recurrence using multiparametric MR imaging of non-enhancing peritumoral regions at baseline. *Heliyon* **10** (2024).
15. Auer, T. A. et al. T2 mapping of the peritumoral infiltration zone of glioblastoma andanaplastic astrocytoma. *Neuroradiol. J.* **34**, 392 (2021).
16. Zakharova, N. E. et al. Perifocal Zone of Brain Gliomas: application of Diffusion Kurtosis and Perfusion MRI values for Tumor Invasion Border determination. *Cancers (Basel)* **15** (2023).
17. Salas-Gallardo, G. A., Lorea-Hernández, J. J., Robles-Gómez, Á. A. & Del Campo, C. M. C. & Peña-Ortega, F. Morphological differentiation of peritumoral brain zone microglia. *PLoS One* **19** (2024).
18. Stupp, R. et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl. J. Med.* **352**, 987–996 (2005).
19. Wen, P. Y. et al. Updated response assessment criteria for high-grade gliomas: Response assessment in neuro-oncology working group. *J. Clin. Oncol.* **28**, 1963–1972 (2010).
20. Sarkar, S. et al. Long-term outcomes following maximal safe resection in a contemporary series of childhood craniopharyngiomas. *Acta Neurochir. (Wien).* **163**, 499–509 (2021).
21. Giussani, C. et al. Perilesional resection technique of glioblastoma: intraoperative ultrasound and histological findings of the resection borders in a single center experience. *J. Neurooncol.* **161**, 625–632 (2023).
22. Al-Holou, W. N. et al. Perilesional Resection of Glioblastoma is independently Associated with Improved outcomes. *Neurosurgery* **86**, 112–121 (2020).
23. Guerrini, F., Roca, E. & Spena, G. Supramarginal Resection for Glioblastoma: it is time to set boundaries! A critical review on a hot topic. *Brain Sci.* **12** (2022).
24. Certo, F. et al. FLAIRectomy in Supramarginal Resection of Glioblastoma correlates with clinical outcome and survival analysis: A prospective, single Institution, Case Series. *Oper. Neurosurg. (Hagerstown Md).* **20**, 151–163 (2021).
25. Vivas-Buitrago, T. et al. Influence of Supramarginal Resection on Survival outcomes after Gross Total Resection in IDH-Wildtype Glioblastoma. *J. Neurosurg.* **136**, 1 (2022).
26. Khalafallah, A. M. et al. A crowdsourced consensus on supratotal resection versus gross total resection for anatomically distinct primary glioblastoma. *Neurosurgery* **89**, 712–719 (2021).
27. Sanai, N., Polley, M. Y., McDermott, M. W., Parsa, A. T. & Berger, M. S. An extent of resection threshold for newly diagnosed glioblastomas. *J. Neurosurg.* **115**, 3–8 (2011).
28. Aziz, P. A. et al. Supratotal Resection: An emerging Concept of Glioblastoma Multiforme surgery-systematic review and Meta-analysis. *World Neurosurg.* **179**, e46–e55 (2023).
29. De Leeuw, C. N. & Vogelbaum, M. A. Supratotal resection in glioma: a systematic review. *Neuro Oncol.* **21**, 179–188 (2019).
30. Menze, B. H. et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging.* **34**, 1993 (2015).
31. Pati, S. et al. Federated learning enables big data for rare cancer boundary detection. *Nat. Commun.* **13**, 1–17 (2022). (2022).
32. Fyllingen, E. H., Stensjøen, A. L., Berntsen, E. M., Solheim, O. & Reinertsen, I. Glioblastoma segmentation: Comparison of three different Software packages. *PLoS One.* **11**, e0164891 (2016).
33. Zhang, G., Zhou, J., He, G. & Zhu, H. Deep fusion of multi-modal features for brain tumor image segmentation. *Heliyon* **9**, e19266 (2023).
34. Smith, T. R. et al. Radiomics and machine learning in brain tumors and their habitat: A systematic review. (2023). https://doi.org/10.3390/cancers15153845
35. Kofler, F. et al. BraTS Toolkit: translating BraTS brain tumor segmentation algorithms into clinical and scientific practice. *Front. Neurosci.* **14**, 501835 (2020).
36. Battalapalli, D., Rao, B. V. V. S. N. P., Yogeeswari, P., Kesavadas, C. & Rajagopalan, V. An optimal brain tumor segmentation algorithm for clinical MRI dataset with low resolution and non-contiguous slices. *BMC Med. Imaging* **22**, (2022).
37. Kofler, F. et al. Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient. *Mach. Learn. Biomed. Imaging.* **2**, 27–71 (2023).
38. Huber, T. et al. Progressive disease in glioblastoma: benefits and limitations of semi-automated volumetry. *PLoS One* **12** (2017).
39. Yan, J. L. et al. Multimodal MRI characteristics of the glioblastoma infiltration beyond contrast enhancement. *Ther. Adv. Neurol. Disord* **12** (2019).
40. Bakas, S. et al. Identifying the best machine learning algorithms for Brain Tumor Segmentation, Progression Assessment, and overall survival prediction in the BRATS Challenge. *Sandra Gonzlez-Vill* **124** (2019).
41. Ruffle, J. K., Mohinta, S., Gray, R., Hyare, H. & Nachev, P. Brain tumour segmentation with incomplete imaging data graphical Abstract. *Brain Commun.* **5**, (2023).
42. Huber, T. et al. Reliability of semi-automated segmentations in Glioblastoma. *Clin. Neuroradiol.* **27**, 153–161 (2017).
43. Pati, S. et al. Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the Ivy Glioblastoma Atlas Project (Ivy GAP) dataset. *Med. Phys.* **47**, 6039 (2020).

Acknowledgements

This research was partially supported by the grant: Italian MUR Dipartimenti di Eccellenza 2023–2027 (l. 232/2016, art. 1, commi 314–337). We are grateful to Fondazione Tecnomed (https://fondazionetecnomed.it/).

Author contributions

V.C. : Conceptualization, Methodology, Software, Formal analysis, Data curation, Validation, Writing – original draft, Writing – review & editing, Visualization. C.B.R. and D.F. : Investigation, Formal analysis, Data Curation, Writing – review & editing. A.D.C. : Investigation, Resources, Data Curation, Writing – review & editing. C.G. and G.C. : Resources, Supervision, Writing – review & editing. G.B. : Conceptualization, Methodology, Formal analysis, Resources, Investigation, Supervision, Writing – review & editing, Project administration.E.D.B. : Conceptualization, Methodology, Software, Supervision, Writing – original draft, Writing – review & editing, Visualization, Project administration.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

.

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval
number: protocol number 464–09/08/2023, Comitato Etico Territoriale Lombardia 3.

Additional information
Correspondence and requests for materials should be addressed to V.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.