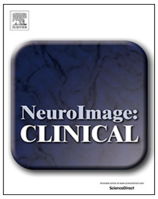
NeuroImage: Clinical 22 (2019) 101727

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl



Inter-rater agreement in glioma segmentations on longitudinal MRI

M. Visser[a,*], D.M.J. Müller[b,c], R.J.M. van Duijn[a], M. Smits[d], N. Verburg[b,c], E.J. Hendriks[a], R.J.A. Nabuurs[b,c], J.C.J. Bot[a], R.S. Eijgelaar[e], M. Witte[e], M.B. van Herk[f], F. Barkhof[a,g], P.C. de Witt Hamer[b], J.C. de Munck[a]

[a] *Department of Radiology and Nuclear Medicine, Amsterdam UMC, location VUmc, De Boelelaan 1117, 1081 HZ Amsterdam, the Netherlands*
[b] *Department of Neurosurgery, Amsterdam UMC, location VUmc, De Boelelaan 1117, 1081 HZ Amsterdam, the Netherlands*
[c] *Brain Tumor Center, Amsterdam UMC, location VUmc, De Boelelaan 1117, 1081 HZ Amsterdam, the Netherlands*
[d] *Department of Radiology and Nuclear Medicine, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, the Netherlands*
[e] *Department of Radiotherapy, The Netherlands Cancer Institute, Plesmanlaan 121, 1006 BE Amsterdam, the Netherlands*
[f] *Institute of Cancer Sciences, Manchester Cancer Research Centre, Division of Cancer Science, School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Sciences Centre, Manchester M13 9PL, United Kingdom*
[g] *Institutes of Neurology and Healthcare Engineering, University College London, Gower St, Bloomsbury, London WC1E 6BT, United Kingdom*

ARTICLE INFO

Keywords:
Inter-rater agreement
Manual segmentation
Glioma
Glioblastoma
Low-grade glioma
MRI

ABSTRACT

Background: Tumor segmentation of glioma on MRI is a technique to monitor, quantify and report disease progression. Manual MRI segmentation is the gold standard but very labor intensive. At present the quality of this gold standard is not known for different stages of the disease, and prior work has mainly focused on treatment-naive glioblastoma. In this paper we studied the inter-rater agreement of manual MRI segmentation of glioblastoma and WHO grade II-III glioma for novices and experts at three stages of disease. We also studied the impact of inter-observer variation on extent of resection and growth rate.

Methods: In 20 patients with WHO grade IV glioblastoma and 20 patients with WHO grade II-III glioma (defined as non-glioblastoma) both the enhancing and non-enhancing tumor elements were segmented on MRI, using specialized software, by four novices and four experts before surgery, after surgery and at time of tumor progression. We used the generalized conformity index (GCI) and the intra-class correlation coefficient (ICC) of tumor volume as main outcome measures for inter-rater agreement.

Results: For glioblastoma, segmentations by experts and novices were comparable. The inter-rater agreement of enhancing tumor elements was excellent before surgery (GCI 0.79, ICC 0.99) poor after surgery (GCI 0.32, ICC 0.92), and good at progression (GCI 0.65, ICC 0.91). For non-glioblastoma, the inter-rater agreement was generally higher between experts than between novices. The inter-rater agreement was excellent between experts before surgery (GCI 0.77, ICC 0.92), was reasonable after surgery (GCI 0.48, ICC 0.84), and good at progression (GCI 0.60, ICC 0.80). The inter-rater agreement was good between novices before surgery (GCI 0.66, ICC 0.73), was poor after surgery (GCI 0.33, ICC 0.55), and poor at progression (GCI 0.36, ICC 0.73). Further analysis showed that the lower inter-rater agreement of segmentation on postoperative MRI could only partly be explained by the smaller volumes and fragmentation of residual tumor. The median interquartile range of extent of resection between raters was 8.3% and of growth rate was 0.22 mm/year.

Conclusion: Manual tumor segmentations on MRI have reasonable agreement for use in spatial and volumetric analysis. Agreement in spatial overlap is of concern with segmentation after surgery for glioblastoma and with segmentation of non-glioblastoma by non-experts.

1. Introduction

Glioma is the most common primary brain tumor in adults (Crocetti et al., 2012; Ostrom et al., 2017). Gliomas are classified by histological type and malignancy grade (Louis et al., 2007). Despite surgical resection, radiotherapy and chemotherapy, the survival of glioma patients is limited, with a two-year survival of 15% for glioblastoma (WHO grade IV) and 85% for diffuse low-grade glioma and a ten-year survival of 2% and 58% respectively (Ostrom et al., 2017).

Although glioma segmentation on MRI is not generally considered

[*] Corresponding author at: Amsterdam UMC, location VUmc, Department of Neurosurgery, De Boelelaan 1117, 1081 HZ Amsterdam, the Netherlands.
E-mail address: m.visser2@vumc.nl (M. Visser).

https://doi.org/10.1016/j.nicl.2019.101727
Received 12 July 2018; Received in revised form 6 February 2019; Accepted 19 February 2019
Available online 22 February 2019

to be part of standard care, it is useful in clinical practice for documentation, prediction of survival, treatment planning, assessment of quality of care, and treatment response measurement. For diagnosis and surgical planning, several MRI sequences are typically applied to assess tumor location and extent (T1-, T2- and T2-FLAIR-weighted images) and the integrity of the blood brain barrier (T1-weighted images after administration of a gadolinium-based contrast agent). Tumors are segmented on pre- and postoperative MR scans for volumetric analysis and calculation of the extent of resection (EOR). The EOR is an important predictor of survival for gliomas (Brown et al., 2016; Lacroix et al., 2001; Sanai and Berger, 2018). Tumor segmentation is standard practice for planning of radiotherapy. During radiological follow-up the tumor volume is monitored, and timing of second line treatment is based on tumor growth. Quantitation of MRI tumor volumes has proven to be valuable for studying autonomous growth (Gui et al., 2018; Mandonnet et al., 2013; Mandonnet et al., 2008), quantification of the effects of (pharmacological) interventions (Ben Abdallah et al., 2018b; Mandonnet et al., 2010; Pallud et al., 2012a; Pallud et al., 2012b), and statistical maps of care (De Witt Hamer et al., 2013; Mandonnet et al., 2007) and disease mechanisms (Amelot et al., 2017; Ellingson et al., 2013; Wang et al., 2014).

These examples show that examination of glioma on MRI by human experts is important. Ideally, observer variation should be small. Theoretically, this variation could be reduced or eliminated by semi-automated or completely automated MRI segmentation algorithms, and such algorithms are being developed (Cordova et al., 2014; Gooya et al., 2012; Meier et al., 2016; Menze et al., 2015; Porz et al., 2016, 2014; Zaouche et al., 2018). To date the work on automatic segmentation has primarily focused on the segmentation of preoperative MR scans of patients with glioblastoma. However, the most recent BRATS tumor segmentation benchmarking challenges have put automatic detection of tumor volume change on follow-up MR scans on the agenda (Crimi et al., 2018, 2016).

Manual segmentation by experts is still considered to be the gold standard and therefore required for quantitative interpretation of MR images and for the validation of automated segmentation algorithms. Reproducibility of manual segmentations has been investigated previously by others (Ben Abdallah et al., 2018a, 2016; Bø et al., 2017; Cattaneo et al., 2005; Gutman et al., 2013; Huber et al., 2015; Kleesiek et al., 2016; Kubben et al., 2010; Provenzale et al., 2009; Provenzale and Mancini, 2012; Sorensen et al., 2001; Weltens et al., 2001). Most of this work was focused on manual segmentation of preoperative MRI in glioblastoma, although a few of these studies consider longitudinal data (Huber et al., 2015; Kleesiek et al., 2016; Kubben et al., 2010; Meier et al., 2016). Two studies (Ben Abdallah et al., 2016; Huber et al., 2015) have addressed the issue of required level of expertise, albeit for preoperative MRI. Both these studies indicated no significant influence of either clinical expertise, or the years of experience on the reproducibility of the segmentations. Since manual segmentation of 3D MRI is labor intensive, even when semi-automated methods are used, many studies are based on a limited number of included scans and raters. Finally, most studies address segmentation of glioblastoma and relatively few studies address lower grade gliomas, although in more recent studies lower grade glioma segmentation is being studied as well (Ben Abdallah et al., 2016; Bø et al., 2017).

In this study, we aim to establish the reproducibility of manual raters in the case of glioma segmentation on MRI, and the impact on extent of resection and growth rate measurements. We will therefore analyze the reproducibility of glioma segmentations at three MRI scan time points by eight raters with two levels of expertise for glioblastoma and non-glioblastomas.

2. Methods

2.1. Patients

Patients were randomly selected from a cohort treated at the Neurosurgical Center Amsterdam of the VU medical center (Amsterdam, The Netherlands) between 2009 and 2013 with standard T2-FLAIR-, T2-, T1-weighted images before and after contrast agent administration. All series were obtained at 3 time points: 1) preoperative, i.e. before first-time resective surgery, 2) postoperative, and 3) at disease progression. For interpretation of post-surgical ischemia, diffusion-weighted imaging on MRI after surgery was included as well. MR data from 20 patients with histopathologically confirmed WHO IV glioblastoma and from 20 patients with grade II-III glioma were included. All 20 gadolinium-enhancing gliomas had a histopathological diagnosis of glioblastoma WHO grade IV. Of the 20 non-enhancing gliomas, 12 were astrocytoma WHO grade II, four oligodendroglioma WHO grade II, three oligoastrocytoma WHO grade II, and one anaplastic astrocytoma WHO grade III, which we refer to as non-glioblastomas.

The preoperative MRI was made on average within one week before resection. The MRI after surgery was made within 72 h after resection for glioblastomas and on average at four months after resection for non-glioblastomas. The MRI at progression was the scan that demonstrated the first tumor progression according to tumor board meeting consensus.

The institutional review board at the VU medical center Amsterdam approved of this study (case nr. 2014.336), after which the data was gathered retrospectively from the clinical workflow. All patients provided written informed consent for use of their clinical data for medical research. The imaging was analyzed after anonymization in accordance with the Personal Data Protection Act.

2.2. MR-imaging

Imaging was performed on a variety of systems (Siemens, model Sonata or Avanto; GE medical systems, model Signa HDxt or DISCOVERY MR750; Toshiba, model Titan3T; Philips, model Panorama HFO or Ingenuity) with a field strength of 1 T (1% of all scans), 1.5 T (62% of all scans) or 3 T (37% of all scans). The standardized protocol included non-enhanced axial T1-weighted spin echo images [repetition time/echo time (TR/TE) 520–600/8–12 ms] with 5-mm slice thickness and axial T2-weighted turbo spin echo images (TR/TE 5190–8670/93–101 ms) with 5-mm slice thickness. Sagittal 3D turbo fluid-attenuated inversion-recovery (FLAIR) images [repetition time/echo time/inversion time (TR/TE/TI) 6500/355/2200 ms] with 1.3-mm slice thickness and axial single shot spin echo echo-planar diffusion-weighted (DWI) images (TR/TE 3400/122 ms) with 5-mm slice thickness were also derived. Diffusion gradients were applied along three orthogonal directions using b-values of 0, 500 and 1000 s/mm^2. Apparent diffusion coefficient (ADC) maps were calculated from the DWI images. Post-contrast (0.2 mmol/Kg) sagittal 3D T1-weighted MPRAGE gradient-echo (T1c) images (TR/TE/TI 2300-2700/5-4.5/950 ms) with 1- to 1.5-mm slice thickness were obtained.

All the DICOM images of pre-, postoperative MRI and at progression were loaded in the Elements environment (BrainLab™ GmBH, Feldkirchen, Germany) and were rigidly registered to the post-contrast T1-weighted MRI per time-point using the Image Fusion tool to facilitate visual comparison of scans. For non-glioblastomas, both immediate and late postoperative MRI were available to raters to discern regions of postsurgical diffusion restriction from residual tumor.

2.3. Manual segmentation

Four experts and four novices segmented each glioma at each timepoint as rater. The experts consisted of three neuro-radiologists (E1, E2 and E3) and one neurosurgeon (E4) with 8, 20, 18, and 20 years of clinical experience, respectively. The novices consisted of three neurosurgical residents (N1, N3 and N4), and one neuro-radiology resident (N2) with 1, 5, 3, and 3 years of clinical experience.

Raters were blinded for histopathological diagnosis and clinical follow-up of patients. Raters were asked to delineate both the non-enhancing and the contrast-enhancing tumor elements - if present - for all three MRI time points in each of 40 patients. To facilitate MRI interpretation, raters were acquainted with the VASARI-criteria (Visually AcceSAble Rembrandt Images, as proposed by The Cancer Imaging Archive (Clark et al., 2013)), but no segmentations rules were imposed. The raters were asked to segment the enhancing tumor elements on post contrast T1-weighted images and to include enclosed necrosis or cysts. Furthermore, they were requested to segment the non-enhancing tumor elements on T2/FLAIR-weighted images. A volume of zero was assigned when a rater determined absence of enhancing or non-enhancing elements.

Segmentations were made with the semi-automatic SmartBrush tool (Elements©, BrainLab™ GmBH, Feldkirchen, Germany) approved for use in clinical practice. Raters were instructed with the use of the software and practiced their skills with MRI sets for preoperative, postoperative and progression time points from two test patients, one contrast-enhancing case and one non-enhancing case. Afterwards they received feedback on their use of the software and on the requirements for the segmentations. From this point on no further feedback was provided. Raters were blinded for segmentations of the co-raters and received the 40 MRI sets in identical order. The order of the MRI sets was randomized to ensure mixing of glioma gradings.

2.4. Statistical analysis

First, we evaluated the agreement in the detection of any enhancing or non-enhancing tumor tissue between expert and novice raters using bar plots, as raters may not necessarily agree on tumor presence.

Second, we determined the inter-rater agreement in volume measurements derived from the segmentations using the intra-class correlation coefficient (ICC) (McGraw and Wong, 1996) and (Shrout and Fleiss, 1979). The specific ICC model used for this purpose is the ICC(A,1) from (McGraw and Wong, 1996) to quantify the inter-rater agreement on volume. ICC scores below 0.4 were considered as poor agreement, 0.4–0.6 as reasonable, 0.6–0.7 as good, and 0.7–1 as excellent (Bartko, 1991; Cicchetti, 1994).

Third, we determined the inter-rater agreement in spatial overlap using the generalized conformity index (GCI) (Kouwenhoven et al., 2009) that quantifies the spatial overlap among multiple spatial objects. This a mathematical generalization of the well-known Jaccard score, which quantifies the overlap of two volumes, as the ratio between the volume of the cross-section and the union of both volumes. When the segmented set by rater j is indicated as A_j and its volume by $\mathrm{Vol}(A_j)$, the GCI is expressed as:

$$GCI = \frac{\sum_{\mathrm{pairs}(i>j)} \mathrm{Vol}(A_i \cap A_j)}{\sum_{\mathrm{pairs}(i>j)} \mathrm{Vol}(A_i \cup A_j)} \tag{1}$$

where $\sum_{\mathrm{pairs}(i>j)} \ldots$ indicates summation over all combinations of unique pairs of raters. For two raters the GCI equals the Jaccard score, $\mathrm{GCI} = \mathrm{Vol}(A_1 \cap A_2)/\mathrm{Vol}(A_1 \cup A_2)$. The GCI was calculated separately for experts and novices, for each MRI time point of every patient. Raters who detected no tumor in a patient, i.e. a volume of zero, were omitted from the GCI calculation for that patient. A GCI of zero denotes no spatial overlap at all and a GCI of one denotes complete spatial overlap among raters. Scores of 0.7–1.0 are regarded as excellent (Bartko, 1991; Zijdenbos et al., 1994). The distributions of spatial overlap scores were visualized in scatter plots and boxplots. Differences in distributions between experts and novices were tested using the Fisher-Pitman permutation test (Ludbrook and Dudley, 1998).

Fourth, to evaluate when expert knowledge is required, we also determined the Jaccard indices between expert consensus and novice consensus segmentations. Majority voting over multiple raters is a well-established method to obtain a consensus segmentation that is a better ground truth than single rater's segmentation (Kittler et al., 1996). For these consensus segmentations, a voxel-wise majority vote of at least two of four raters was used.

Fifth, to evaluate the impact on clinical volumetric analysis, we calculated the extent of resection based on the pre- and postoperative MRI and the growth rate based on the postoperative and progression MRI for each rater. The extent of resection was based on volumes of enhancing elements for glioblastoma and on volumes of non-enhancing elements for non-glioblastoma:

$$EOR = \frac{V_{pre} - V_{post}}{V_{pre}} \times 100\% \tag{2}$$

where $\mathrm{V_{pre}}$ and $\mathrm{V_{post}}$ are the pre- and postoperative volumes of one rater.The growth rate was calculated as difference between the mean tumor diameters divided by the time-interval in years (Mandonnet et al., 2008), in which:

$$D_{mean} = (2 \times V)^{\frac{1}{3}} \tag{3}$$

where $\mathrm{D_{mean}}$ is the mean tumor diameter of the volume V of one rater. For the clinical volumetric analyses we used the interquartile range as measure of dispersion between the non-normal measurements of raters per case.

3. Results

3.1. Patient characteristics

Patients with glioblastoma had a mean age of 61.4 years (range 41.8–72.6) and consisted of 10 females and 10 males. Patients with non-glioblastoma had a mean age of 36.9 years (range 18.6–53.7) and consisted of 8 females and 12 males. The time between preoperative MRI and surgery was on average 7.8 days for glioblastoma and 53.6 days for non-glioblastoma. The time between surgery and the postoperative MRI was on average 1.2 days for glioblastoma and 4.11 months for non-glioblastoma. The time between surgery and the progressive MRI was on average 13.7 months (range: 5.6–30.7) for glioblastoma and 28.9 months (range 6.2–60.7) for non-glioblastoma. Enhancing and non-enhancing tumor were not treated as mutually exclusive by the raters, therefore overlap is present between the segmentations of enhancing and non-enhancing tumor. The average contrast-enhancing (with enclosed necrosis) tumor volume was 32.2 mL for glioblastoma and 0.8 mL for non-glioblastoma on the preoperative MRI, 2.7 and 0.0 mL on the postoperative MRI, and 24.2 and 5.2 mL on the progressive MRI. The average non-enhancing tumor volume was 88.6 mL for glioblastoma and 45.3 mL for non-glioblastoma on the preoperative MRI, 37.2 and 8.4 mL on the postoperative MRI, and 78.0 and 25.7 mL on the progressive MRI. The tumor was located in the left hemisphere in 8 patients with glioblastoma, and in 9 patients with non-glioblastoma. Detailed patient characteristics are presented in Table 1.

3.2. Tumor tissue detection

The number of raters that identified any tumor are plotted in Fig. 1. Zero raters would represent perfect agreement on absence of tumor, and four raters would represent perfect agreement on presence of tumor.

Experts and novices perfectly agreed on the presence of any enhancing tumor for glioblastoma and on any non-enhancing tumor for non-glioblastoma patients on preoperative MRIs. Few experts and even fewer novices detected enhancing tumor in non-glioblastoma patients preoperatively. In postoperative MRIs both experts and novices

Table 1
Patient characteristics.

Glioblastoma							Non-glioblastoma						
Pat	Path	Sex	Age	T1	T2	T3	Pat	Path	Sex	Age	T1	T2	T3
1	*GB*	F	67,1	13	0	415	21	*A2*	F	53,7	26	91	1746
2	*GB*	F	72,1	6	1	229	22	*O2*	M	44,7	1	111	1033
3	*GB*	M	65,3	2	0	920	23	*A2*	F	23,1	67	111	188
4	*GB*	F	66,1	2	1	310	24	*A2*	M	30,1	53	77	861
5	*GB*	M	66,7	1	1	474	25	*A2*	M	18,6	1	184	1477
6	*GB*	F	64,0	15	1	274	26	*A2*	F	21,8	9	92	1538
7	*GB*	M	45,4	4	3	591	27	*O2*	M	52,6	67	143	1595
8	*GB*	M	52,8	9	3	255	28	*A2*	F	35,5	46	108	1820
9	*GB*	M	61,3	7	0	279	29	*A2*	M	30,8	255	102	686
10	*GB*	M	70,5	2	0	184	30	*A2*	F	28,6	111	127	207
11	*GB*	M	75,5	1	1	188	31	*OA2*	M	34,8	109	1	191
12	*GB*	F	66,2	8	2	540	32	*A2*	F	48,2	2	99	573
13	*GB*	M	71,6	10	1	825	33	*A2*	M	29,1	12	101	1438
14	*GB*	M	55,1	2	3	770	34	*A2*	M	23.0	60	145	965
15	*GB*	F	42,3	5	1	732	35	*A2*	M	41,6	1	90	903
16	*GB*	M	73,0	18	1	329	36	*OA2*	F	39,5	1	61	183
17	*GB*	F	47,2	3	1	168	37	*A3*	M	52,8	8	170	306
18	*GB*	F	41,8	21	1	267	38	*A2*	M	37,8	52	161	186
19	*GB*	F	72,6	8	1	204	39	*O2*	F	44,3	68	380	402
20	*GB*	M	51,4	19	2	278	40	*OA2*	M	46,7	126	112	1038

T1: time of preoperative scans (days before surgery), T2: time of postoperative scans (days after surgery), T3: time of progression, GB: glioblastoma, A2: astrocytoma grade II, O2: Oligodendroglioma grade II, OA2: oligoastrocytoma grade II, A3: anaplastic astrocytoma grade III.

considerably disagreed on the presence of enhancing tumor in glioblastoma patients. Experts more frequently agreed perfectly on enhancing tumor presence than novices; novices more frequently agreed perfectly on enhancing tumor absence in postoperative MRIs. Experts generally agreed on tumor presence in non-glioblastoma patients postoperatively, whereas novices disagreed in one third of these

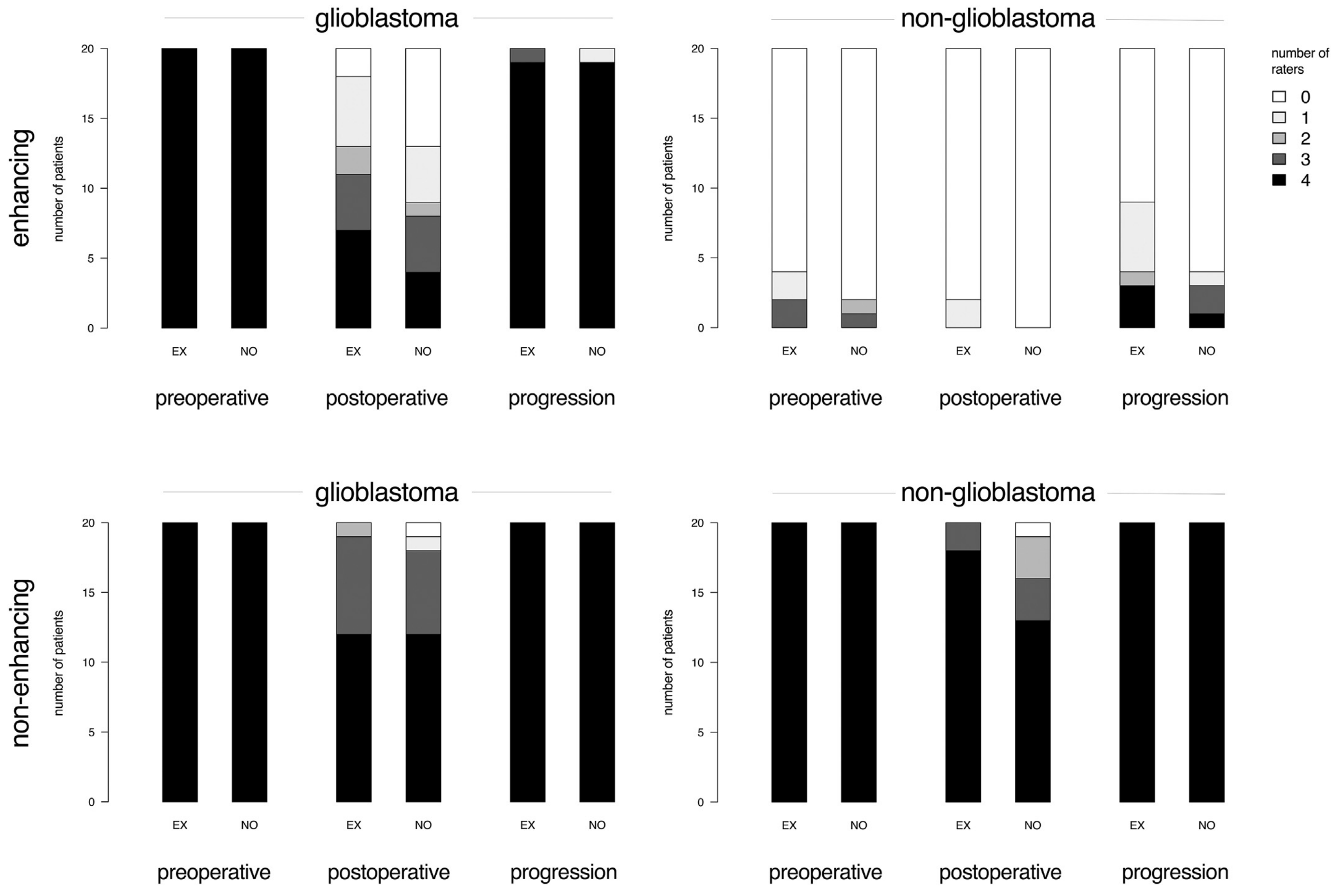


Fig. 1. Bar plots of the number of patients with corresponding number of expert (EX) and novice (NO) raters detecting any enhancing tumor and any non-enhancing tumor for glioblastoma and non-glioblastoma in MRIs preoperative, postoperative and at progression.

Table 2
Intra-class coefficient with 95% confidence intervals for experts and novices.

Histology group	Contrast	Rater	Preoperative	Postoperative	Progression
GB	Enhancing	Experts	0.99 (0.98–1.00)	0.92 (0.85–0.97)	0.91 (0.82–0.96)
GB	Enhancing	Novices	0.98 (0.96–1.00)	0.60 (0.39–0.78)	0.97 (0.95–0.99)
GB	Non-enhancing	Experts	0.61 (0.41–0.79)	0.25 (0.05–0.52)	0.53 (0.24–0.76)
GB	Non-enhancing	Novices	0.55 (0.24–0.78)	0.15 (0.00–0.38)	0.40 (0.09–0.67)
Non-GB	Enhancing	Experts	0.28 (0.07–0.55)	*	1.00 (1.00–1.00)
Non-GB	Enhancing	Novices	0.57 (0.35–0.77)	*	0.66 (0.47–0.83)
Non-GB	Non-enhancing	Experts	0.92 (0.81–0.97)	0.84 (0.70–0.93)	0.80 (0.65–0.91)
Non-GB	Non-enhancing	Novices	0.73 (0.40–0.89)	0.55 (0.32–0.76)	0.73 (0.46–0.88)

GB: glioblastoma.

* No enhancing elements were identified for non-glioblastomas in the postoperative MRI, with the exception of 2 disjoint residual volumes each by a different rater.

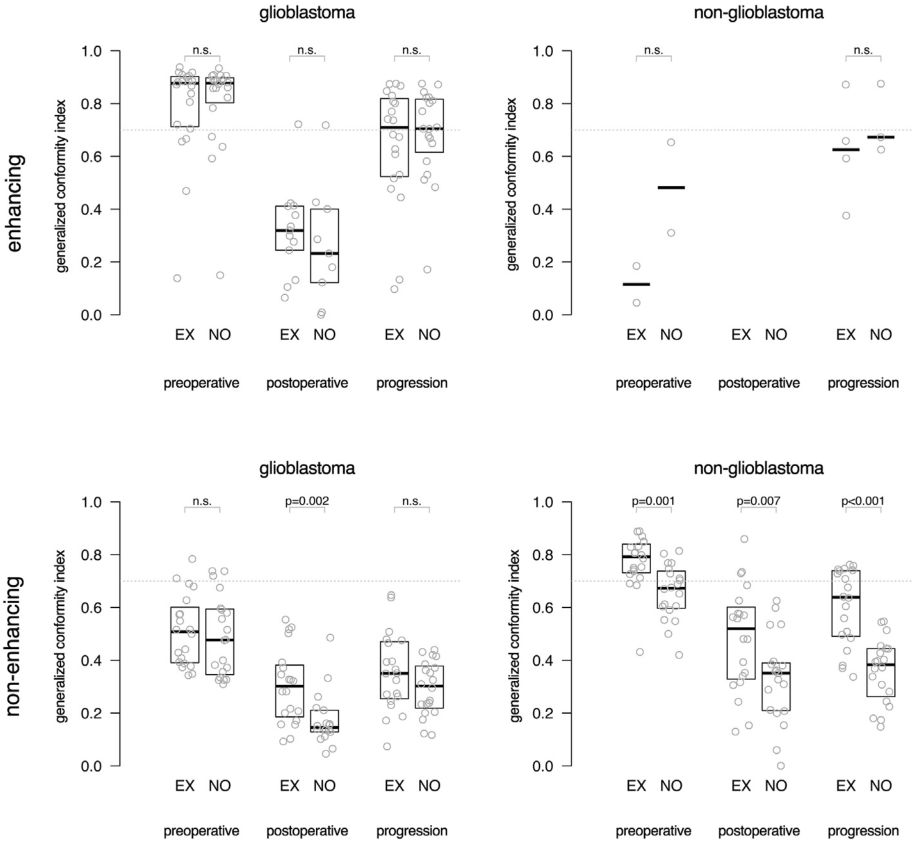


Fig. 2. Box plots of the spatial overlap among experts (EX) and novices (NO) measured as generalized conformity index for enhancing tumor and non-enhancing tumor segmentations of 20 glioblastoma and 20 non-glioblastoma patients in MRIs taken at preoperative, postoperative and progression time points. Each dot represents the agreement among raters for one patient's MRI. Indices above 0.7 are considered excellent. The median of measurements and interquartile distances are plotted as boxes, which were omitted when fewer than five data points were present. Few data points were available for enhancing tumor segmentations in non-glioblastoma, because the generalized conformity index could not be calculated when fewer than two observers detected tumor.

patients. At progression, experts and novices generally agreed on the presence of any enhancing tumor in glioblastoma and perfectly agreed on any non-enhancing tumor in non-glioblastoma patients. Experts more frequently identified enhancing tumor in non-glioblastoma patients at progression than novices. All experts and novices identified non-enhancing tumor in all glioblastoma and non-glioblastoma patients.

3.3. ICC of tumor volume

The ICCs of tumor volumes are shown in Table 2. Agreement in volume measurements among experts is excellent at all three time points for enhancing tumor elements in glioblastoma patients and excellent for non-enhancing tumor elements in non-glioblastoma patients (ICC ≥ 0.8). In contrast, the non-enhancing elements in glioblastoma patients have poor to fair agreement for both experts and novices. The agreement among experts is generally better than among novices.

3.4. Spatial overlap

Results for spatial agreement are represented as box-plots of the GCI between raters in Fig. 2, demonstrating that experts generally achieve a higher agreement in spatial overlap than the novices. For non-enhancing tumor segmentations of glioblastoma on postoperative MRI, experts had a significantly higher spatial overlap than novices with a median GCI of 0.30 versus 0.15 (p = .002). For non-enhancing tumor segmentations of non-glioblastoma at all MRI time points, experts had a significantly higher spatial agreement than novices with a median GCI of 0.79 versus 0.67 (p = .001) on preoperative MRI, 0.52 versus 0.35 (p = .007) on postoperative MRI and 0.64 versus 0.38 (p < .001) at progression.

The spatial agreement was invariably highest for preoperative segmentations and lowest for postoperative segmentations.

Agreement on enhancing tumor in glioblastoma was excellent among both experts and novices on preoperative MRI and at progression. Spatial agreement was lowest for enhancing tumor in glioblastoma on postoperative MRI, whereas this was affected by a substantial inter-

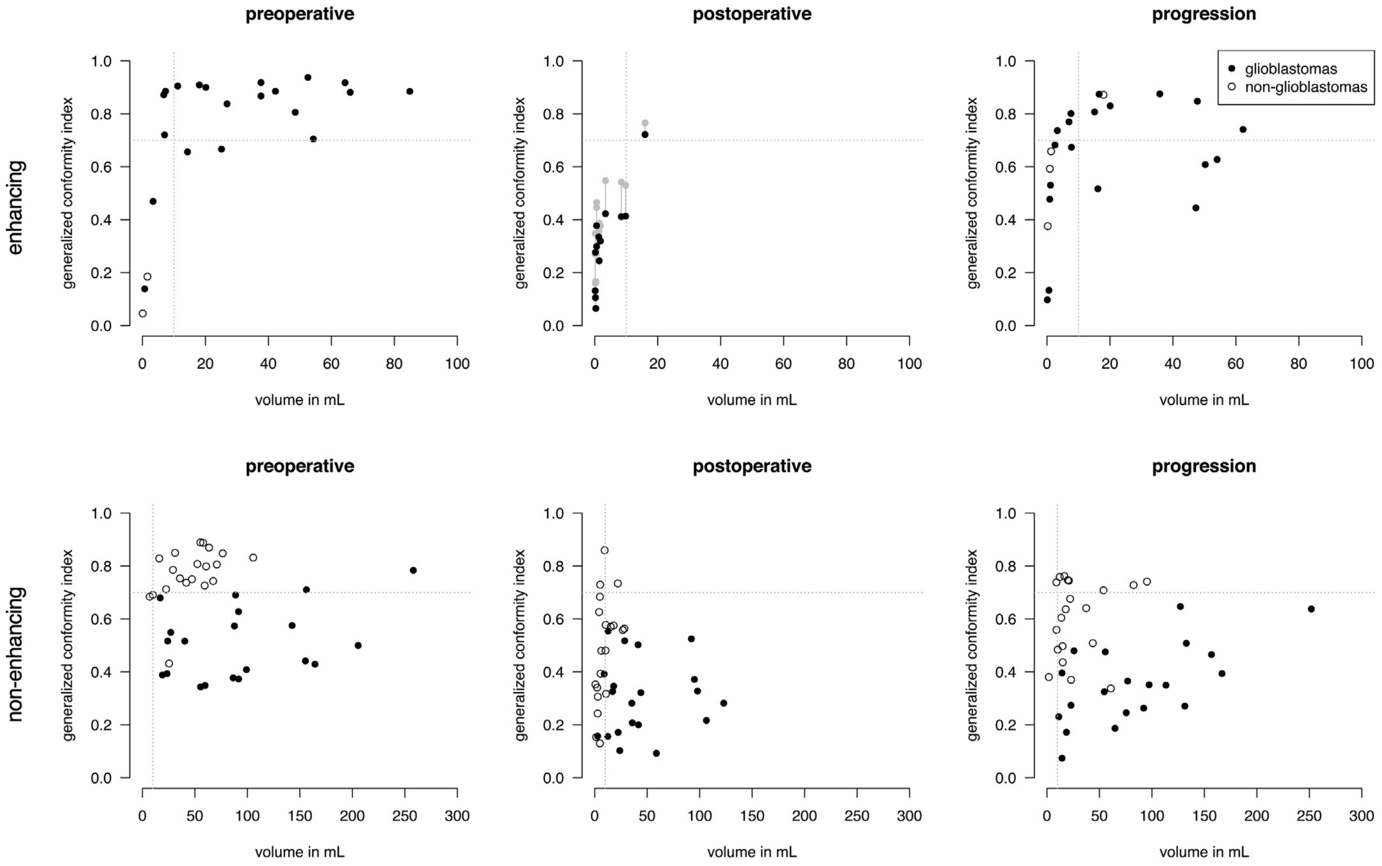


Fig. 3. Spatial overlap agreement as generalized conformity index versus tumor volume (average over experts) of enhancing tumor (A) and non-enhancing tumor (B) segmentations for glioblastomas and non-glioblastomas at subsequent MRI timings. Each dot represents the agreement of spatial overlap among experts on one patient's MRI. For enhancing tumor at postoperative phase it is shown that spatial overlap increases after artificial dilation of segmentation (grey dots), however not to the level of progression segmentation of the same volume.

observer disagreement on the presence of any enhancing tumor. Agreement on non-enhancing tumor was excellent among experts segmenting non-glioblastoma, and lowest among novices segmenting non-enhancing tumor for glioblastoma.

Spatial overlap agreement was generally higher for enhancing tumor in glioblastoma than for non-enhancing tumor in non-glioblastoma at all MRI timings.

To explore potential causes of the low spatial overlap agreement of postoperative enhancing tumor in glioblastoma patients, we hypothesized that lower object volumes and higher level of fragmentation may contribute to this. The scatter plots in Fig. 3A confirm that in particular enhancing tumor volumes smaller than 10 mL in glioblastoma come with a strikingly lower agreement. As tumor volumes on postoperative MRI are typically smaller than 10 mL, this may partly explain the low agreement. A similar small volume effect was observed in non-enhancing tumor segmentations of non-glioblastomas in Fig. 3B.

To take this one step further, we artificially dilated the enhancing tumor segmentations of glioblastomas with a 10 mm spherical structure element and recalculated the overlap of the dilated volumes (grey symbols, middle panel Fig. 3A). Although the overlap increases, it is still lower than undilated object volumes of similar size. Therefore, the lower agreement could not be fully explained by a small volume effect.

In addition, we compared the fragmentation of the tumor segmentations by calculating the number of connected components for patients with an enhancing tumor volume smaller than 10 mL. The average number of fragments was 2.14 ± 1.35 (SD) on postoperative MRI and 1.92 ± 1.96 at progression. Therefore, fragmentation of tumor segmentations did not fully explain the lower agreement on postoperative MRI either.

3.5. Majority voting consensus

Subsequently the spatial overlap agreement was determined between each rater's segmentations and the majority vote for experts and novices combined (Fig. 4). The plots shown in Fig. 4 show a similar trend as the group-wise analysis shown in Fig. 2. Again, the highest agreement was observed on preoperative MRIs, followed by MRIs at the time of progression, and lowest agreement for postoperative MRIs. The comparison against the majority vote allowed for scrutiny on the individual level, showing for the non-glioblastoma patients that one novice (N4) performed at a level similar to that of the experts. We also compared the majority vote for experts and for novices (Fig. 5) which shows that the novice consensus is comparable to the expert consensus for enhancing tumor on preoperative MRI and at progression for glioblastoma and for non-enhancing tumor on preoperative MRI for non-glioblastoma. Novice consensus shows only moderate agreement with expert consensus for enhancing tumor on postoperative MRI for glioblastoma and for non-enhancing tumor on postoperative MRI and at progression for non-glioblastoma.

3.6. Clinical volumetric analysis: extent of resection and growth rate

The variation in extent of resection and growth rate between raters is plotted in Fig. 6. The agreement between raters on the extent of resection of glioblastoma is excellent with a median interquartile range of 1.2% and below 10% in 18 (90%) of 20 cases. At higher extents of resection the variation between raters is lower. For non-glioblastoma, the agreement between raters on the extent of resection is less than glioblastoma but still reasonable with a median interquartile range of

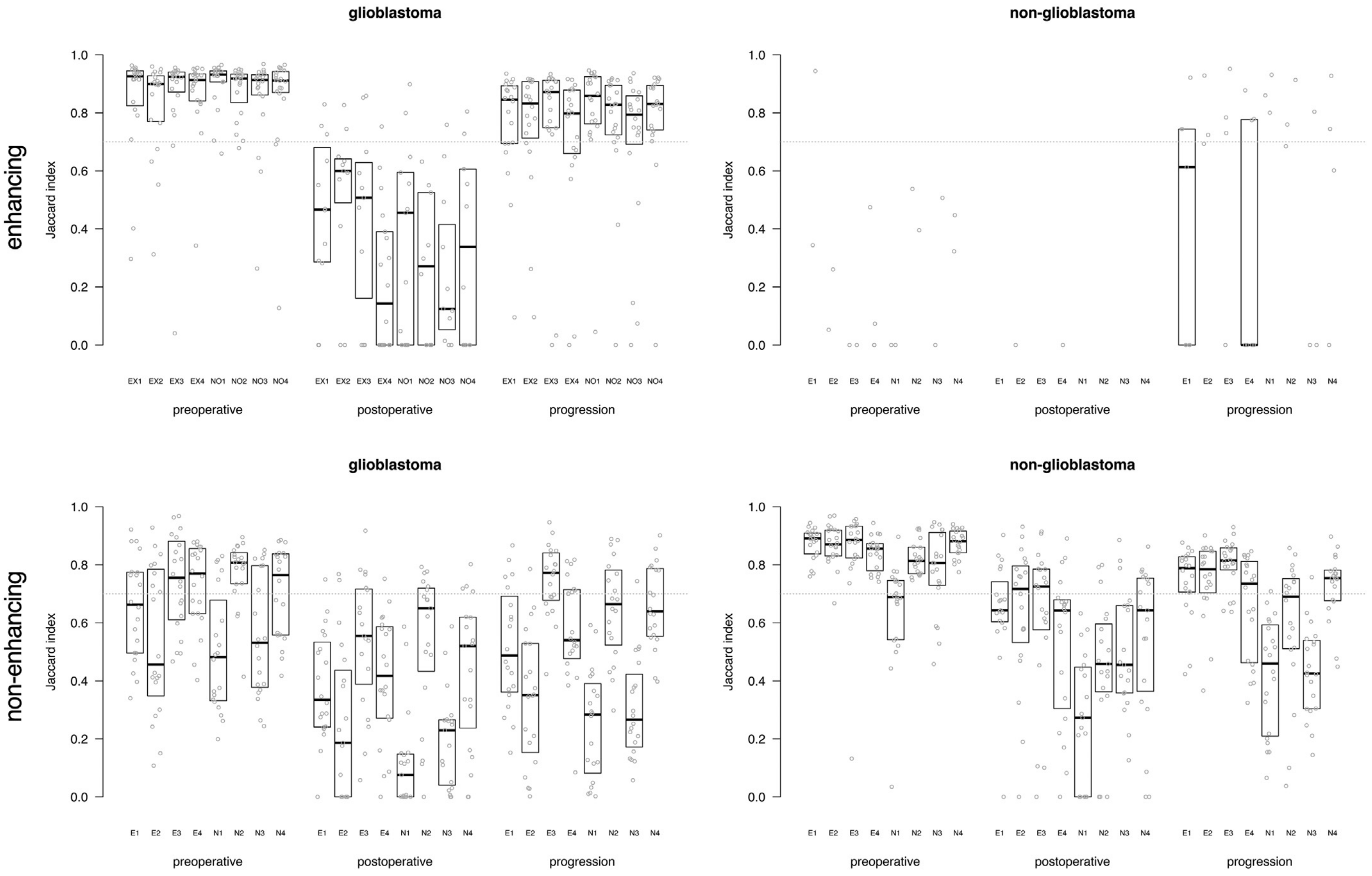


Fig. 4. Box plots of agreement between majority vote of all eight raters and each of the individual raters, as Jaccard index for enhancing tumor and non-enhancing tumor segmentations in glioblastoma and non-glioblastoma at the three MRI time points. Each dot represents the agreement between the consensus and the individual rater for one patient's segmentation. The first four subplots represent the experts, the second four refer to the novices. The median of measurements and interquartile distances are plotted as boxes, which were omitted when fewer than five data points were measured.

8.3% and below 10% in 10 (50%) of 20 cases. A correlation between extent of resection and variation between raters seems absent.

The agreement between raters on the growth rate of glioblastoma is quite high with a median interquartile range of 0.42 mm/y and below 1 mm/y in 16 (80%) of 20 cases. The agreement on growth rate is not correlated with growth rates. For non-glioblastoma, the agreement on growth rate was higher than for glioblastoma with a median interquartile range of 0.22 mm/y and below 1 mm/y in 18 (90%) of 20 cases. At lower growth rates the variation between raters is lower.

4. Discussion

In this study we present a comprehensive and systematic analysis of inter-rater agreement in glioma segmentations addressing glioblastoma and non-glioblastoma, at different stages of disease, and comparing experts, with extensive clinical experience, and novices, with limited training. Our main findings are that (1) the agreement on presence and overlap of preoperative tumor segmentations was high and of postoperative tumor segmentations was low, (2) experts demonstrated higher levels of agreement than novices, in particular for non-enhancing tumor segmentations in non-glioblastoma and (3) the agreement on enhancing tumor in non-glioblastoma and on non-enhancing tumor in glioblastoma was very low.

The inter-rater agreement on postoperative MRI is problematic. Raters disagree considerably on tumor presence, experts and novices alike, and even more so for enhancing tumor in glioblastoma than for non-enhancing tumor in non-glioblastoma. A possible explanation is that MRIs made a few days after glioblastoma surgery suffer from surgical artefacts, such as blood clots, luxury perfusion of post resection ischemia or contusion, distortion of tissue and blood vessels. Misinterpretation of these surgical artefacts may be diminished by subtraction of the T1-weighted MRI before contrast from the T1-weighted MRI after contrast. Many of these artefacts have resolved in the months after non-glioblastoma resection, which explains the higher agreement between raters in this patient population. This time to postoperative MRI is not available in patients with glioblastoma because radiotherapy, inducing further treatment artefacts, usually follows shortly. Segmentation for non-enhancing tumor in glioblastomas on postoperative MRI has a low inter-rater agreement and is deemed to be ill-defined as a ground truth due to poor spatial overlap and volume agreement. The main reason is that some raters attempted to distinguish non-enhancing tumor portions from pure edema in glioblastoma within T2/FLAIR hyper-intense regions, whereas others considered all hyper-intensity to be tumor. A clear instruction to include all hyper-intensity may improve the agreement. Common reasons for disagreement of enhancing portions consisted of small linear enhancement at the border of the resection cavities, which was considered to be sulcal vasculature or gliosis by some raters and residual tumor by others. Furthermore, some raters identified small multifocal enhancing nodules at distance from the resection cavity that were overlooked or considered normal vasculature by others, which resulted in poor volume overlap. In non-enhancing tumor segmentations of lower-grade glioma, novices typically identified tumor in the uncus adjacent to the tumor on T2/FLAIR-weighted MRI, which contained intensities similar to the contralateral uncus according to experts. Similarly novices included the hyper-intensity of the cortex adjacent to the sulci, where experts restricted their segmentation from sulcus to sulcus.

The inter-rater agreement on MRI at progression was slightly lower

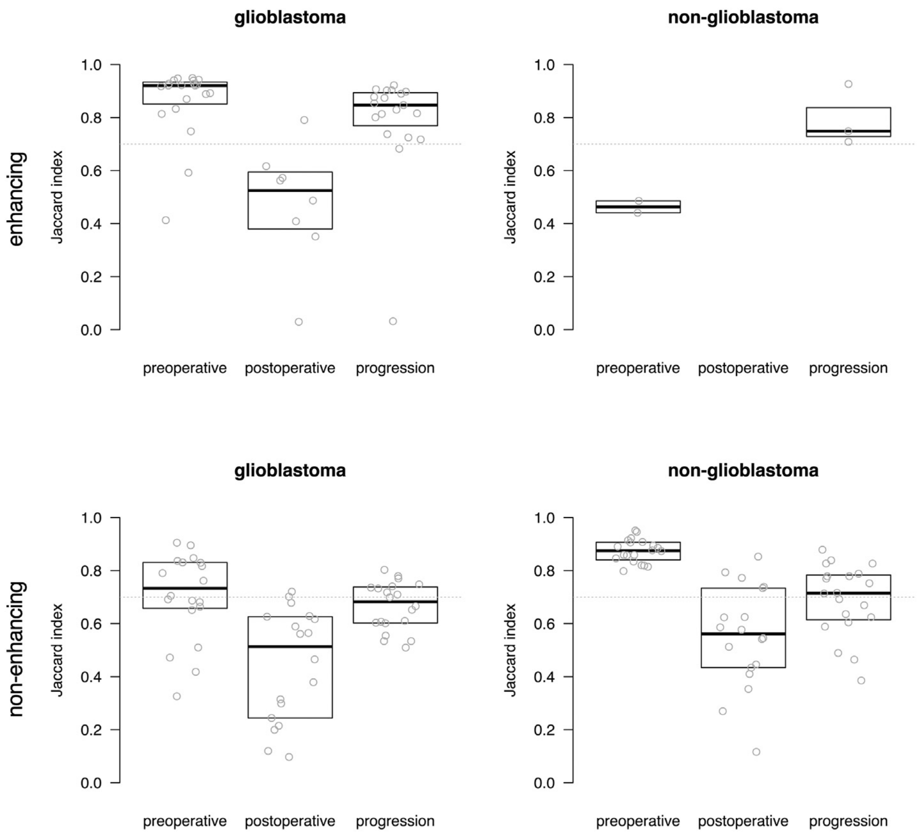


Fig. 5. Boxplots of agreement between rater and majority vote consensus of experts and novices combined measured as Jaccard index for enhancing and non-enhancing tumor segmentations in glioblastoma and non-glioblastoma at three MRI timings. Each dot represents the agreement between a rater's segmentations and the majority vote consensus of all raters for one patient's segmentation. Indices above 0.7 are considered excellent. The median of measurements and interquartile distances are plotted as boxes, which were omitted when fewer than five data points were measured.

than the inter-rater agreement on preoperative MRI, and higher than on postoperative MRI, which is in agreement with the relative volumes. The agreement in spatial overlap for non-glioblastoma segmentation found in this study, with a GCI of 0.60 among experts, is in agreement with that found by others (Gui et al., 2018) based on two experts segmenting two MRIs. Novices can replace experts in segmentations of enhancing tumor in glioblastoma on MRI at progression. Nevertheless, experts seem to be required for non-enhancing tumor segmentation in non-glioblastomas on MRI at progression. MRIs at progression of non-glioblastomas are difficult to interpret because these suffer from artefacts from radiation therapy that cannot be discerned from disease progression (Tensaouti et al., 2017).

The combination of results from experts and novices may incorrectly overlook performance of individual raters and therefore be an oversimplification. Interestingly, the comparison of individual raters with the consensus of all raters shows that one novice (the last in Fig. 4) seems to provide segmentations of similar quality as experts.

For glioblastomas, the spatial overlap agreement between raters was high on preoperative MRI, which is not surprising due to the unambiguous distinction of contrast enhancing tumor to non-enhancing surrounding tissue. At progression the contrast becomes more ambiguous due to treatment effects such as pseudo-progression or radiation induced necrosis (Tensaouti et al., 2017). The contrast becomes even more ambiguous on postoperative MRI with small fragmented residual tumor in the presence of surgical artefacts. Of note is that despite the lack of spatial overlap agreement, the volume ICC scores in glioblastoma are high, particularly among experts, in contrast to findings by others (Kubben et al., 2010). Perhaps this discrepancy is due to the agreement on absence of residual tumor in several of our patients, whereas in the previously published study (Kubben et al., 2010) all 8 patients had postoperative residue. Our data support that, despite low agreement in spatial overlap, the agreement in volume measurements is reasonable, which is commonly used for determining the extent of resection.

The impact of inter-rater disagreement on common clinical volumetric analyses such as the extent of resection and the growth rate appear to be limited. The extent of resection calculations for glioblastoma justifies use of exact percentages by a single rater for cohort reports. Extent of resection calculations for non-glioblastoma are subject to more variation, and therefore would likely be better represented by categories of near-complete, subtotal and partial resections, for instance. Furthermore, the growth rate calculation agreements justify use of exact growth rates by a single rater, even more so for non-glioblastoma than glioblastoma.

An important aspect that impact scores like Dice and Jaccard (of which the GCI is an extension) is the effect of small volumes, which biases these scores to be lower as volume decreases. Distance measures are considered less susceptible to this small volume bias (e.g. (Dubuisson and Jain, 1994; Steenbakkers et al., 2005)), but require correlated surfaces to establish a distance measure and this is undefined in case of multiple tumor fragments, as is common for glioma segmentations. Possible causes for the poor to moderate spatial overlap agreement as described by the GCI for postoperative data include the relative small volumes and tumor fragmentation. However, we showed that the enhanced tumor segmentations at progression of glioblastoma patients have similar fragmentation but were associated with a higher spatial overlap. Even when the postoperative segmentations were artificially dilated to reduce the volume effect the overlaps stayed well below those of the results at progression. Therefore, we conclude that segmentations on postoperative imaging are more complex than those at progression.

This study has some limitations. We have used a commercial semi-automatic segmentation tool, which may not be available to other users. We have selected this tool, because it is time-efficient and intuitive and is in common use in clinical settings for the treatment of patients with brain tumors. Furthermore, we adopted the VASARI-criteria for radiological definitions of glioblastoma, which are based on standard T1- and T2-weighted sequences. These standard sequences are

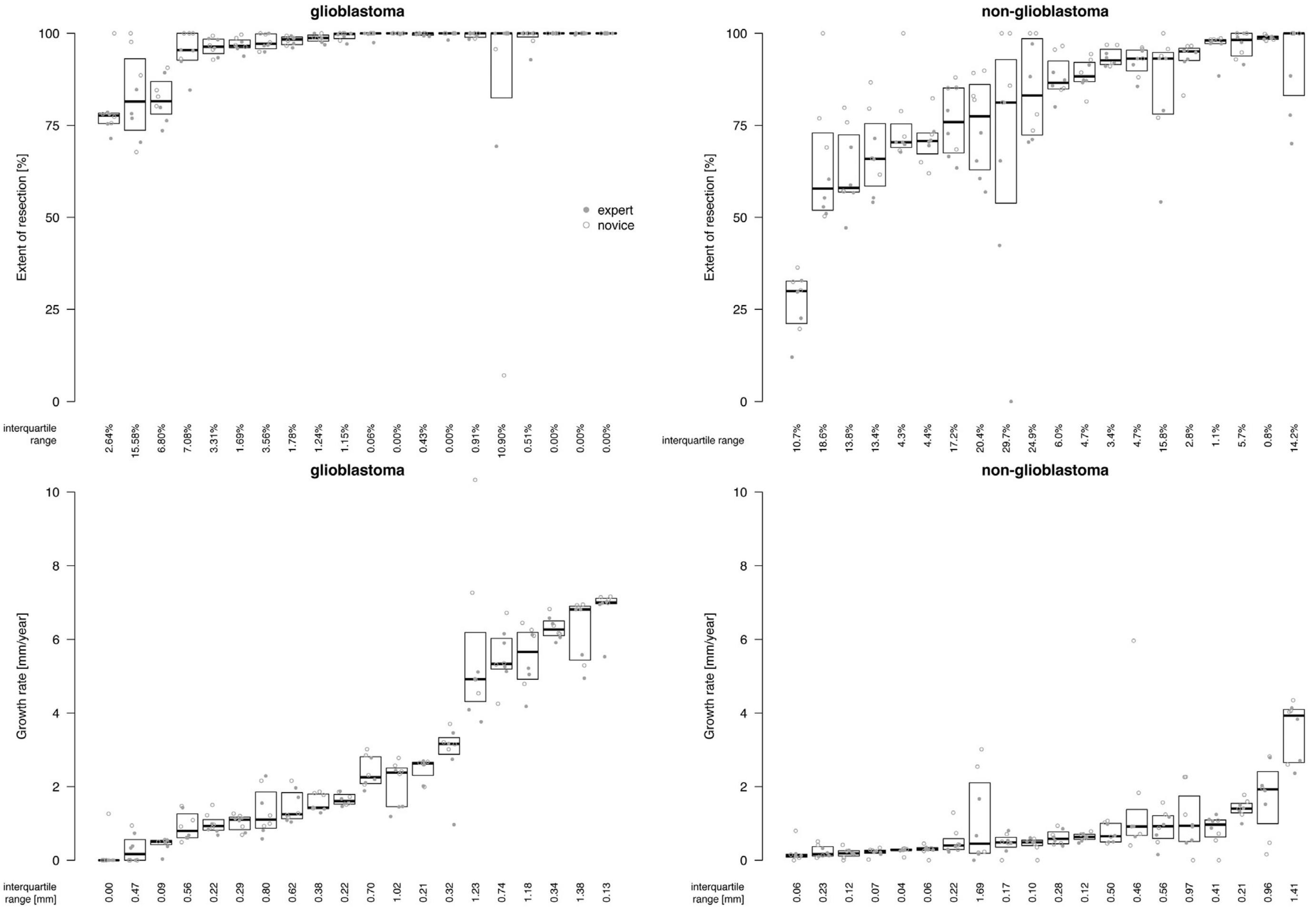


Fig. 6. The variation in extent of resection and growth rate for glioblastoma and non-glioblastoma between eight raters per patient. In each plot patients are sorted by median extent of resection and growth rate, respectively. Each dot represents the calculation for one patient of one rater. Experts and novices are labelled according to the legend. The median of measurements and interquartile distances are plotted as boxes. The quartile coefficients of dispersion are plotted below the boxplots.

Table 3
An overview of previous studies on inter-rater agreement.

Authors	Year	Low grade			High grade			#Exp	#Nov	Context
		Pre	Post	Prog	Pre	Post	Prog			
Weltens et al., 2001	2001				4			6	3	Added value of MRI to CT for segmentation.
Cattaneo et al., 2005	2005				7				5*	idem
Provenzale et al., 2009	2009						22**	8		Reproducibility of 2D tumor dimensions.
Kubben et al., 2010	2010				8	8		2	1	Manual PreOp/PostOp glioblastoma segmentation
Gooya et al., 2012	2012				10				2[a]	GLISTR
Provenzale and Mancini, 2012	2012				5		5[b]	3	4	Reproducibility of 2D tumor dimensions.
Cordova et al., 2014	2014				37	37		1[e]	2	Semi-automatic segmentation.
Porz et al., 2014	2014				25			1[c]	1[c]	BraTumIA
Menze et al., 2015	2015	14			51			4		BRATS
Huber et al., 2015	2015				5	5		4	8	Evaluation of inter-rater variability
Ben Abdallah et al., 2016	2016	9		3[b]				13		Idem
Porz et al., 2016	2016				19			4		BraTumIA
Kleesiek et al., 2016	2016				15		15	2		Semi-automatic segmentation
Meier et al., 2016[d]	2016				14	14	14	1	1	BraTumIA (longitudinal)
Bø et al., 2017	2017	23						1		Intra-rater assessment
Zaouche et al., 2018	2018	4						2		Semi-automatic segmentation
Gui et al., 2018	2018			4				2		Quantification of progression
This Study	2018	20	20	20	20	20	20	4	4	Evaluation of inter-rater variability

[a] Unspecified type of rater.
[b] Moment after surgery not specified.
[c] Supervised by expert neuro-radiologist.
[d] This study has multiple longitudinal moments after postoperative.
[e] Expert used as ground truth, novices test semi-automated method.

known to have poor performance to distinguish tumor infiltration from normal brain (Verburg et al. 2017). Perhaps better performance can be expected from (combinations of) advanced imaging, which should then be used to improve tumor segmentation.

Our study is an extension of the current literature, summarized in Table 3, which often focuses on glioblastoma with manual segmentations on preoperative MRI as reference to evaluate novel (semi-) automatic tumor segmentation algorithms. In the recent literature, more and more expert segmentations are made publically available (e.g. BRATS data (Menze et al., 2015) and (Bakas et al., 2017)) and are being used for the validation of (semi-) automatic algorithms (e.g. (Zaouche et al., 2018)). However, such data sets are of limited value when each segmentation results from a single rater and the inter-rater variability is unknown.

Experts generally have higher agreement than novices, suggesting that expert segmentations are better than those of novices in particular for non-enhancing tumor segmentations in non-glioblastomas, although novices have similar agreement for enhancing tumor segmentations of glioblastomas on preoperative MRI and at progression. Our results indicate that preoperative tumor segmentation is done reliably by novices and experts. For other applications of tumor segmentation, such as assessment of quality of care, treatment response measurement, and evaluation of progression, segmentations are less reliable and sensitivity analysis of different raters would be needed. In practice, it is not realistic to obtain consensus segmentations from multiple experts. A promising future strategy could be to use standardized fully automated tumor segmentation algorithms which is probably more reproducible than manual segmentations, but which may be inaccurate as well. To determine the accuracy of segmentations, ground truth histopathological correlation of tumor presence would be required.

Acknowledgements

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative and the Translational Research IT (TraIT) project, an initiative from the Center for Translational Molecular Medicine (CTMM).

Source of funding

This research is part of the PICTURE project supported by the program Innovative Medical Devices Initiative (IMDI) with project number 10-10400-96-14003, which is financed by the Netherlands Organization for Scientific Research (NWO). This research is also supported by a research grant from the Dutch Cancer Society (VU2014-7113). Furthermore, F. Barkhof is supported by the National Institute for Health Research (NIHR) of the University College London Hospitals (UCLH) Biomedical Research Centre. BrainLab® has provided us with their proprietary neuro-navigational software suite Elements, which comprises the SmartBrush functionality.

Conflicts of interest

The authors of this paper have not conflict of interest to report.

References

Amelot, A., Deroulers, C., Badoual, M., Polivka, M., Adle-Biassette, H., Houdart, E., Carpentier, A.F., Froelich, S., Mandonnet, E., 2017. Surgical decision making from image-based biophysical modeling of glioblastoma: not ready for primetime. Neurosurgery 80, 793–799. https://doi.org/10.1093/neuros/nyw186.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing the cancer genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data 4, 1–13. https://doi.org/10.1038/sdata.2017.117.

Bartko, J.J., 1991. Measurement and reliability: statistical thinking considerations. Schizophr. Bull. 17, 483–489. https://doi.org/10.1093/schbul/17.3.483.

Ben Abdallah, M., Blonski, M., Wantz-Mezieres, S., Gaudeau, Y., Taillandier, L., Moureaux, J.-M., 2016. Statistical evaluation of manual segmentation of a diffuse low-grade glioma MRI dataset. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 4403–4406. https://doi.org/10.1109/EMBC.2016.7591703.

Ben Abdallah, M., Blonski, M., Wantz-Mézières, S., Gaudeau, Y., Taillandier, L., Moureaux, J.-M., 2018a. Relevance of two manual tumour volume estimation methods for diffuse low-grade gliomas. Health. Technol. Lett. 5, 13–17. https://doi.org/10.1049/htl.2017.0013.

Ben Abdallah, M., Blonski, M., Wantz-Mezieres, S., Gaudeau, Y., Taillandier, L., Moureaux, J.-M., Darlix, A., Menjot de Champfleur, N., Duffau, H., 2018b. Data-driven predictive models of diffuse low-grade gliomas under chemotherapy. IEEE J. Biomed. Heal. Informatics 1. https://doi.org/10.1109/JBHI.2018.2834159.

Bø, H.K., Solheim, O., Jakola, A.S., Kvistad, K.A., Reinertsen, I., Berntsen, E.M., 2017. Intra-rater variability in low-grade glioma segmentation. J. Neuro-Oncol. 131, 393–402. https://doi.org/10.1007/s11060-016-2312-9.

Brown, T.J., Brennan, M.C., Li, M., Church, E.W., Brandmeir, N.J., Rakszawski, K.L., Patel, A.S., Rizk, E.B., Suki, D., Sawaya, R., Glantz, M., 2016. Association of the extent of resection with survival in glioblastoma: a systematic review and meta-analysis. JAMA Oncol. https://doi.org/10.1001/jamaoncol.2016.1373.

Cattaneo, G.M., Reni, M., Rizzo, G., Castellone, P., Ceresoli, G.L., Cozzarini, C., Ferreri, a.J.M., Passoni, P., Calandrino, R., 2005. Target delineation in post-operative radiotherapy of brain gliomas: interobserver variability and impact of image registration of MR(pre-operative) images on treatment planning CT scans. Radiother. Oncol. 75, 217–223. https://doi.org/10.1016/j.radonc.2005.03.012.

Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6, 284–290. https://doi.org/10.1037/1040-3590.6.4.284.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F., 2013. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging 26, 1045–1057. https://doi.org/10.1007/s10278-013-9622-7.

Cordova, J.S., Schreibmann, E., Hadjipanayis, C.G., Guo, Y., Shu, H.-K.G., Shim, H., Holder, C.A., 2014. Quantitative tumor segmentation for evaluation of extent of glioblastoma resection to facilitate multisite clinical trials. Transl. Oncol. 7, 40–W5. https://doi.org/10.1593/tlo.13835.

Crimi, A., Menze, B., Maier, O., Reyes, M., Hutchison, D., 2016. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Lecture Notes in Computer Science. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-55524-9.

Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M., 2018. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Lecture Notes in Computer Science. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-75238-9.

Crocetti, E., Trama, A., Stiller, C., Caldarella, A., Soffietti, R., Jaal, J., Weber, D.C., Ricardi, U., Slowinski, J., Brandes, A., 2012. Epidemiology of glial and non-glial brain tumours in Europe. Eur. J. Cancer 48, 1532–1542. https://doi.org/10.1016/j.ejca.2011.12.013.

De Witt Hamer, P.C., Hendriks, E.J., Mandonnet, E., Barkhof, F., Zwinderman, A.H., Duffau, H., 2013. Resection probability maps for quality assessment of glioma surgery without brain location Bias. PLoS One 8. https://doi.org/10.1371/journal.pone.0073353.

Dubuisson, M., Jain, A.K., 1994. A modified Hausdorff distance for object matching. Int. Conf. Pattern Recognit. 566–568. https://doi.org/10.1109/ICPR.1994.576361.

Ellingson, B.M., Lai, A., Harris, R.J., Selfridge, J.M., Yong, W.H., Das, K., Pope, W.B., Nghiemphu, P.L., Vinters, H.V., Liau, L.M., Mischel, P.S., Cloughesy, T.F., 2013. Probabilistic radiographic atlas of glioblastoma phenotypes. Am. J. Neuroradiol. 34, 533–540. https://doi.org/10.3174/ajnr.A3253.

Gooya, A., Pohl, K.M., Bilello, M., Cirillo, L., Biros, G., Melhem, E.R., Davatzikos, C., 2012. GLISTR: glioma image segmentation and registration. IEEE Trans. Med. Imaging 31, 1941–1954. https://doi.org/10.1109/TMI.2012.2210558.

Gui, C., Kosteniuk, S.E., Lau, J.C., Megyesi, J.F., 2018. Tumor growth dynamics in serially-imaged low-grade glioma patients. J. Neuro-Oncol. 1 (9). https://doi.org/10.1007/s11060-018-2857-x.

Gutman, D. a, Cooper, L. a D., Hwang, S.N., Holder, C. a, Gao, J., Aurora, T.D., Dunn, W.D., Scarpace, L., Mikkelsen, T., Jain, R., Wintermark, M., Jilwan, M., Raghavan, P., Huang, E., Clifford, R.J., Mongkolwat, P., Kleper, V., Freymann, J., Kirby, J., Zinn, P.O., Moreno, C.S., Jaffe, C., Colen, R., Rubin, D.L., Saltz, J., Flanders, A., Brat, D.J., 2013. MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. Radiology 267, 560–569. https://doi.org/10.1148/radiol.13120118.

Huber, T., Alber, G., Bette, S., Boeckh-Behrens, T., Gempt, J., Ringel, F., Alberts, E., Zimmer, C., Bauer, J.S., 2015. Reliability of semi-automated segmentations in Glioblastoma. Clin. Neuroradiol. https://doi.org/10.1007/s00062-015-0471-2.

Kittler, J., Hater, M., Duin, R.P.W., 1996. Combining classifiers. Proc. - Int. Conf. Pattern Recognit. 2, 897–901. https://doi.org/10.1109/ICPR.1996.547205.

Kleesiek, J., Petersen, J., Döring, M., Maier-Hein, K., Köthe, U., Wick, W., Hamprecht, F.A., Bendszus, M., Biller, A., 2016. Virtual Raters for reproducible and objective assessments in radiology. Sci. Rep. 6, 25007. https://doi.org/10.1038/srep25007.

Kouwenhoven, E., Giezen, M., Struikmans, H., 2009. Measuring the similarity of target volume delineations independent of the number of observers. Phys. Med. Biol. 54, 2863–2873. https://doi.org/10.1088/0031-9155/54/9/018.

Kubben, P.L., Postma, A.a., Kessels, A.G.H., van Overbeeke, J.J., van Santbrink, H., 2010. Intraobserver and interobserver agreement in volumetric assessment of glioblastoma multiforme resection. Neurosurgery 67, 1329–1334. https://doi.org/10.1227/NEU.0b013e3181efbb08.

Lacroix, M., Abi-Said, D., Fourney, D.R., Gokaslan, Z.L., Shi, W., DeMonte, F., Lang, F.F., McCutcheon, I.E., Hassenbusch, S.J., Holland, E., Hess, K., Michael, C., Miller, D., Sawaya, R., 2001. A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. J. Neurosurg. 95, 190–198. https://doi.org/10.3171/jns.2001.95.2.0190.

Louis, D.N., Ohgaki, H., Wiestler, O.D., Cavenee, W.K., Burger, P.C., Jouvet, A., Scheithauer, B.W., Kleihues, P., 2007. The 2007 WHO classification of tumours of the central nervous system. Acta Neuropathol. 114, 97–109. https://doi.org/10.1007/s00401-007-0243-4.

Ludbrook, J., Dudley, H., 1998. Why permutation tests are superior to t and F tests in biomedical research. Am. Stat. 52, 127–132. https://doi.org/10.1080/00031305.1998.10480551.

Mandonnet, E., Jbabdi, S., Taillandier, L., Galanaud, D., Benali, H., Capelle, L., Duffau, H., 2007. Preoperative estimation of residual volume for WHO grade II glioma resected with intraoperative functional mapping. Neuro. Oncol. 9, 63–69. https://doi.org/10.1215/15228517-2006-015.

Mandonnet, E., Pallud, J., Clatz, O., Taillandier, L., Konukoglu, E., Duffau, H., Capelle, L., 2008. Computational modeling of the WHO grade II glioma dynamics: principles and applications to management paradigm. Neurosurg. Rev. 31, 263–268. https://doi.org/10.1007/s10143-008-0128-6.

Mandonnet, E., Pallud, J., Fontaine, D., Taillandier, L., Bauchet, L., Peruzzi, P., Guyotat, J., Bernier, V., Baron, M.H., Duffau, H., Capelle, L., 2010. Inter- and intrapatients comparison of WHO grade II glioma kinetics before and after surgical resection. Neurosurg. Rev. 33, 91–95. https://doi.org/10.1007/s10143-009-0229-x.

Mandonnet, E., Wait, S., Choi, L., Teo, C., 2013. The importance of measuring the velocity of diameter expansion on MRI in upfront management of suspected WHO grade II glioma - case report. Neurochirurgie 59, 89–92. https://doi.org/10.1016/j.neuchi.2013.02.005.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1, 30–46. https://doi.org/10.1037/1082-989X.1.1.30.

Meier, R., Knecht, U., Loosli, T., Bauer, S., Slotboom, J., Wiest, R., Reyes, M., 2016. Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor Volumetry. Nat. Sci. Rep. 1–11. https://doi.org/10.1038/srep23376. In press.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A.M., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H.H., Demiralp, C., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A.J.A., Meier, R., Pereira, S.S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34, 1993–2024. https://doi.org/10.1109/TMI.2014.2377694.

Ostrom, Q.T., Gittleman, H., Liao, P., Vecchione-Koval, T., Wolinsky, Y., Kruchko, C., Barnholtz-Sloan, J.S., 2017. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014. Neuro-Oncology 19, v1–v88. https://doi.org/10.1093/neuonc/nox158.

Pallud, J., Llitjos, J.-F., Dhermain, F., Varlet, P., Dezamis, E., Devaux, B., Souillard-Scemama, R., Sanai, N., Koziak, M., Page, P., Schlienger, M., Daumas-Duport, C., Meder, J.-F., Oppenheim, C., Roux, F.-X., 2012a. Dynamic imaging response following radiation therapy predicts long-term outcomes for diffuse low-grade gliomas. Neuro-Oncology 14, 496–505. https://doi.org/10.1093/neuonc/nos069.

Pallud, J., Taillandier, L., Capelle, L., Fontaine, D., Peyre, M., Ducray, F., Duffau, H., Mandonnet, E., 2012b. Quantitative morphological magnetic resonance imaging follow-up of low-grade glioma: a plea for systematic measurement of growth rates. Neurosurgery 71, 729–739. https://doi.org/10.1227/NEU.0b013e31826213de.

Porz, N., Bauer, S., Pica, A., Schucht, P., Beck, J., Verma, R.K., Slotboom, J., Reyes, M., Wiest, R., 2014. Multi-modal glioblastoma segmentation: man versus machine. PLoS One 9, e96873. https://doi.org/10.1371/journal.pone.0096873.

Porz, N., Habegger, S., Meier, R., Verma, R., Jilch, A., Fichtner, J., Knecht, U., Radina, C., Schucht, P., Beck, J., Raabe, A., Slotboom, J., Reyes, M., Wiest, R., 2016. Fully automated enhanced tumor compartmentalization: man vs. machine reloaded. PLoS One 11, e0165302. https://doi.org/10.1371/journal.pone.0165302.

Provenzale, J.M., Mancini, M.C., 2012. Assessment of intra-observer variability in measurement of high-grade brain tumors. J. Neuro-Oncol. 108, 477–483. https://doi.org/10.1007/s11060-012-0843-2.

Provenzale, J.M., Ison, C., DeLong, D., 2009. Bidimensional measurements in brain tumors: assessment of interobserver variability. Am. J. Roentgenol. 193, 515–522. https://doi.org/10.2214/AJR.09.2615.

Sanai, N., Berger, M.S., 2018. Surgical oncology for gliomas: the state of the art. Nat. Rev. Clin. Oncol. 15, 112–125. https://doi.org/10.1038/nrclinonc.2017.171.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428. https://doi.org/10.1037/0033-2909.86.2.420.

Sorensen, A.G., Patel, S., Harmath, C., Bridges, S., Synnott, J., Sievers, A., Yoon, Y.-H., Lee, E.J., Yang, M.C., Lewis, R.F., Harris, G.J., Lev, M., Schaefer, P.W., Buchbinder, B.R., Barest, G., Yamada, K., Ponzo, J., Kwon, H.Y., Gemmete, J., Farkas, J., Tievsky, A.L., Ziegler, R.B., Salhus, M.R.C., Weisskoff, R., 2001. Comparison of diameter and perimeter methods for tumor volume calculation. J. Clin. Oncol. 19, 551–557. https://doi.org/10.1200/JCO.2001.19.2.551.

Steenbakkers, R.J.H.M., Duppen, J.C., Fitton, I., Deurloo, K.E.I., Zijp, L., Uitterhoeve, A.L.J., Rodrigus, P.T.R., Kramer, G.W.P., Bussink, J., De Jaeger, K., Belderbos, J.S.A., Hart, A.A.M., Nowak, P.J.C.M., van Herk, M., Rasch, C.R.N., 2005. Observer variation in target volume delineation of lung cancer related to radiation oncologist-computer interaction: a “big brother” evaluation. Radiother. Oncol. 77, 182–190. https://doi.org/10.1016/j.radonc.2005.09.017.

Tensaouti, F., Khalifa, J., Lusque, A., Plas, B., Lotterie, J.A., Berry, I., Laprie, A., Cohen-Jonathan Moyal, E., Lubrano, V., 2017. Response assessment in neuro-oncology criteria, contrast enhancement and perfusion MRI for assessing progression in glioblastoma. Neuroradiology 1013–1020. https://doi.org/10.1007/s00234-017-1899-7.

Verburg, N., Hoefnagels, F.W.A., Barkhof, F., Boellaard, R., Goldman, S., Guo, J., Heimans, J.J., Hoekstra, O.S., Jain, R., Kinoshita, M., Pouwels, P.J.W., Price, S.J., Reijneveld, J.C., Stadlbauer, A., Vandertop, W.P., Wesseling, P., Zwinderman, A.H., De Witt Hamer, P.C., 2017. Diagnostic Accuracy of Neuroimaging to Delineate Diffuse Gliomas within the Brain: A Meta-Analysis. Am J Neuroradiol 38, 1884–1891. https://doi.org/10.3174/ajnr.A5368.

Wang, Y., Qian, T., You, G., Peng, X., Chen, C., You, Y., Yao, K., Wu, C., Ma, J., Sha, Z., Wang, S., Jiang, T., 2014. Localizing seizure-susceptible brain regions associated with low-grade gliomas using voxel-based lesion-symptom mapping. Neuro-Oncology 17, 282–288. https://doi.org/10.1093/neuonc/nou130.

Weltens, C., Menten, J., Feron, M., Bellon, E., Demaerel, P., Maes, F., Van den Bogaert, W., van der Schueren, E., 2001. Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging. Radiother. Oncol. 60, 49–59. https://doi.org/10.1016/S0167-8140(01)00371-1.

Zaouche, R., Belaid, A., Aloui, S., Solaiman, B., Lecornu, L., Ben Salem, D., Tliba, S., 2018. Semi-automatic method for low-grade gliomas segmentation in magnetic resonance imaging. IRBM 39, 116–128. https://doi.org/10.1016/j.irbm.2018.01.004.

Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C., 1994. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans. Med. Imaging 13, 716–724. https://doi.org/10.1109/42.363096.