

Intra-rater variability in low-grade glioma segmentation

Hans Kristian Bø^{1,2} · Ole Solheim^{3,4,5} · Asgeir Store Jakola^{3,6,7} · Kjell-Arne Kvistad¹ · Ingerid Reinertsen^{5,8} · Erik Magnus Berntsen^{1,2}

Received: 27 June 2016 / Accepted: 7 November 2016 / Published online: 11 November 2016
© Springer Science+Business Media New York 2016

Abstract Assessment of size and growth are key radiological factors in low-grade gliomas (LGGs), both for prognostication and treatment evaluation, but the reliability of LGG-segmentation is scarcely studied. With a diffuse and invasive growth pattern, usually without contrast enhancement, these tumors can be difficult to delineate. The aim of this study was to investigate the intra-observer variability in LGG-segmentation for a radiologist without prior segmentation experience. Pre-operative 3D FLAIR images of 23 LGGs were segmented three times in the software 3D Slicer. Tumor volumes were calculated, together with the absolute and relative difference between the segmentations.

To quantify the intra-rater variability, we used the Jaccard coefficient comparing both two (J2) and three (J3) segmentations as well as the Hausdorff Distance (HD). The variability measured with J2 improved significantly between the two last segmentations compared to the two first, going from 0.87 to 0.90 ($p=0.04$). Between the last two segmentations, larger tumors showed a tendency towards smaller relative volume difference ($p=0.07$), while tumors with well-defined borders had significantly less variability measured with both J2 ($p=0.04$) and HD ($p<0.01$). We found no significant relationship between variability and histological sub-types or Apparent Diffusion Coefficients (ADC). We found that the intra-rater variability can be considerable in serial LGG-segmentation, but the variability seems to decrease with experience and higher grade of border conspicuity. Our findings highlight that some criteria defining tumor borders and progression in 3D volumetric segmentation is needed, if moving from 2D to 3D assessment of size and growth of LGGs.

✉ Hans Kristian Bø
hans.kr.b@gmail.com

- ¹ Department of Radiology and Nuclear Medicine, St. Olavs University Hospital, P.O. Box 3250, Sluppen, 7006 Trondheim, Norway
- ² Department of Circulation and Medical Imaging, Faculty of Medicine, NTNU, Norwegian University of Science and Technology, Trondheim, Norway
- ³ Department of Neurosurgery, St. Olavs University Hospital, Trondheim, Norway
- ⁴ Department of Neuroscience, Faculty of Medicine, NTNU, Norwegian University of Science and Technology, Trondheim, Norway
- ⁵ National Competence Centre for Ultrasound and Image Guided Therapy, St. Olavs University Hospital, Trondheim, Norway
- ⁶ Department of Neurosurgery, Sahlgrenska University Hospital, Gothenburg, Sweden
- ⁷ Institute of Neuroscience and Physiology, Sahlgrenska Academy, Gothenburg, Sweden
- ⁸ SINTEF, Technology and Society, Department of Medical Technology, Trondheim, Norway

Keywords Glioma · Segmentation · Magnetic resonance imaging · Intra-rater variability

Introduction

Diffuse low-grade gliomas (LGGs) are World Health Organization (WHO) grade II tumors with an infiltrative growth pattern and constitute 10–15% of all primary brain tumors in adults [1–4]. They are initially slow growing and predominantly affect otherwise healthy young adults. Eventually, a malignant transformation into higher grade gliomas will occur. In recent years, a survival benefit from earlier surgical and oncological treatment in the low-grade

stage of the disease has been demonstrated, compared to a wait-and-see approach [2, 5, 6].

It is known that tumor size at diagnosis, extent of surgical resection and volume of the residual tumor are strong prognostic factors [7–17]. Several studies have also shown that the growth rate of the tumor is associated with risk of malignant transformation and overall prognosis [7, 18–21]. After first-line therapy, a significant re-growth seen from repeated Magnetic Resonance Imaging (MRI) examinations during follow-up often forms the basis for clinical decision making concerning reoperations or adjuvant treatment. Thus, size and growth rate are key radiological factors in LGG care, both for prognostication and for clinical decision making. Due to the infiltrative growth pattern and often subtle changes between MRI assessments, volumetric assessment is supposed to be the most sensitive method [7, 22, 23].

Manual segmentation by an experienced operator is considered the gold standard for volumetric segmentation of brain tumors, however this is a very time consuming procedure [24, 25]. Most previous studies have investigated inter-rater variability and compare manual to semi-automatic methods [21, 26–30], while few have explored the intra-rater variability in low-grade glioma segmentation. As only one radiologist is involved in most clinical studies and many clinical settings, knowledge of the intra-rater variability in LGG assessment is highly relevant.

In this study, we sought to address the intra-rater variability in manually verified semi-automatic segmentation of low-grade gliomas by performing serial segmentations of the same tumors, all done by one radiologist. We also sought to explore possible factors associated with variability.

Methods

Study population

Patients were included from an ongoing study on LGGs. Tumor borders were radiologically evaluated and classified as: (1) well-defined, (2) partially absent, or (3) absent. Well-defined indicates a discrete border between tumor and normal appearing brain, partially absent is a more diffuse border zone, but still possible to separate tumor and normal brain, and absent is an ill-defined border and infiltrative finger-like growth pattern.

Image acquisition

MRI images used for tumor segmentation were all pre-operative 3D Fluid Attenuated Inversion Recovery (FLAIR) images. Acquisitions were done on three different

MRI systems, thus with slightly different echo time, repetition time and inversion time (TE/TR/TI). Seventeen patients were examined using a Siemens Skyra 3.0T scanner (389/5000/1800 ms) with both slice-thickness and in-plane resolution of 1 mm. Four patients had their scan on a Siemens Avanto 1.5T scanner (333/6000/2200 ms or 474/6500/1800 ms) with both slice-thickness and in-plane resolution of 1 mm. Two patients were examined with a Philips Intera 3.0T scanner (350/8000/2400 ms) with slice-thickness 1.2, 0.6 mm overlap and in-plane resolution of 0.43 mm.

Segmentation procedure

For the segmentation procedure we used the open source software 3D Slicer 4.4.0 (<http://www.slicer.org>), which is a software platform for quantitative imaging, designed for use in cancer care [31]. 3D Slicer consists of a core platform with several standard modules and a graphical user interface.

For the segmentations in this study, we used the “Grow-Cut” region based segmentation algorithm in the built-in “Editor” module. First, the border of the tumor was manually marked on at least one slice in each of the three planes (transversal, coronal and sagittal). Then the area outside the tumor was marked with a different color. The “GrowCut” algorithm was then run, resulting in an image label superimposed on the MRI image [32]. This label was further edited to fit with the tumor borders, first with the “dilate” and “erode” functions and then manually. Tumor volume in mL was then obtained using the “Label statistics” extension.

All segmentations were done by one radiologist (H. K. B.) with 7 years of radiology experience, but without any prior experience in segmentation or 3D Slicer. When in doubt, tumor borders were discussed with an experienced neuroradiologist (K. A. K.) with 20 years of experience. All tumors were segmented three times; once before any tumors were segmented for the second time, and twice before any tumors were segmented for the third time. To minimize recall bias when segmenting for the second and third time, we made sure that at least 40 days passed between repeated segmentations. We also attempted to have the same time interval between the second and third segmentation as between the first and second.

Measures of agreement

The Jaccard coefficient and Hausdorff distance (HD) are widely used and validated measures of agreement in evaluation of segmentation procedures [33–36]. The Jaccard coefficient is an overlap index used to compare segmentations. If S_i represents a segmentation in a series of n

segmentations, each with volume $V(S_i)$, then the Jaccard coefficient is defined as:

$$J_n = \frac{n[V(\cap_i^n S_i)]}{\sum_{i=1}^n V(S_i)}$$

Jaccard coefficient with two segmentations, J2, is the same as Dice Similarity Coefficient (DSC), which can be shown to be a special case of the Kappa-statistic used for intra-rater agreement [34]. The Jaccard coefficient takes on values from 0 to 1, with 0 when there is no overlap and 1 when there is a perfect match between the segmented volumes. Interpretation is similar to the Kappa-statistic with a strong agreement with values 0.80–0.90 and almost perfect agreement with values above 0.90 [37]. We have included both J2, comparing each pair of two segmentations, and J3, comparing all three segmentations. HD is a measure of distance between two segmentations, defined as the greatest distance measured from each point on the surface of one segmented volume to the closest point on the surface of the other [36]. HD is especially sensitive to local surface variations.

Exploring factors associated with agreement

In an attempt to explore possible features associated with agreement we compared agreement in small vs. large tumors (dichotomized from median tumor volume), in various histopathological subtypes, in relation to mean ADC levels of the tumor (the smallest of the three tumor volumes was used) and in relation to border conspicuity.

Statistics

IBM SPSS Statistics, Version 23.0 (IBM Corp., Armonk, NY) was used for statistical analysis. Central tendencies are presented as mean (standard deviation [SD]) or median

(inter quartile range [IQR]) when skewed. Normality was assessed with histograms and tested with Shapiro–Wilk’s test. Differences in means were tested with two-tailed Student’s paired *t* test when normally distributed and with the two-tailed non-parametric Related-Samples Sign test when skewed. Furthermore, Jaccard coefficient and HD were calculated for exploring measures of segmentation variability. Differences in agreement were tested on sub-groups using Mann–Whitney *U* test when two groups and Kruskal–Wallis test when three groups. P values below 0.05 were considered significant, while p values between 0.05 and 0.10 were considered as trending towards significance [38].

Results

We included preoperative MRIs from 23 untreated patients (median age 41 years (range 18–49), 13 males), with histopathologically verified supratentorial WHO grade II gliomas, operated between 2011 and 2014 at our hospital. There were 10 (43%) oligodendrogliomas, 8 (35%) astrocytomas, 3 (13%) unspecified LGGs and 2 (9%) mixed astrocytomas. Localization was frontal in 12 (52%), insular in 6 (26%) and temporal in 5 (22%) patients. Border margins were well-defined in 12 (52%), partially absent in 8 (35%) and absent in 3 (13%) tumors. Four tumors (17%) had an eloquent localization (after Chang et al. [8]).

Mean time between segmentation cycle 1 and 2 was 144 days (range 43–201), and between cycle 2 and 3 it was 148 days (range 115–202) ($p=0.71$). Median tumor volume from segmentation cycle 1 was 26.4 mL (range 1.4–165.9), 27.6 mL (range 1.7–166.0) for the second cycle and 19.7 mL (range 1.4–163.0) for the third cycle. Comparison between segmentations 1 vs. 2, 2 vs. 3 and 1 vs. 3 are shown in Table 1. There was a median difference in tumor volume of –1.3 mL between the first and second segmentation, corresponding to a median relative difference of 14%

Table 1 Comparison of segmentation cycle 1 vs. 2 with segmentation cycle 2 vs. 3

	Cycle 1 vs. 2 A	Cycle 2 vs. 3 B	Cycle 1 vs. 3 C	p value A vs. B	p value A vs. C
Time between segmentations, mean (SD)	144 days (35)	148 days (22)	292 (27)	0.71	<0.01
Differences in volume, median (IQR)	–1.3 mL (–8.4 to 0.6)	–1.3 mL (–3.5 to –0.3)	–4.1 mL (–11.6 to –0.2)	0.68	<0.01
Relative differences in volume, median (IQR)	14% (5–28)	13% (2–19)	14% (2–41)	0.32	0.68
J2, median (IQR)	0.87 (0.79–0.91)	0.90 (0.83–0.93)	0.87 (0.71–0.93)	0.04	0.41
HD, median (IQR)	9.7 mm (5.9–14.9)	7.4 mm (4.7–9.9)	9.4 mm (5.2–19.1)	0.09	1.00

Median difference in volume and percent, Jaccard coefficient (J2) and Hausdorff Distance (HD)

SD standard deviation, IQR inter quartile range, mL milliliter, mm millimeter

(IQR 5–28), a median HD of 9.8 mm (IQR 5.9–14.9) and a median J2 of 0.87 (IQR 0.79–0.91). There was a median difference in tumor volume of -1.3 mL also between the second and third segmentation cycle, corresponding to a median relative difference of 13% (IQR 2–19), a median HD of 7.4 mm (IQR 4.7–9.9) and a median J2 of 0.90 (IQR 0.83–0.93). When comparing the first and last segmentation cycle, there was a median difference in tumor volume of -4.1 mL, corresponding to a median relative difference of 14% (IQR -11.6 to -0.2), a median HD of 9.4 mm (IQR 5.2–19.1) a median J2 of 0.87 (IQR 0.71–0.93). The difference in median tumor volume in subsequent segmentation cycles was not significantly different ($p=0.68$), but between the first and last segmentation cycle it was significant ($p=0.01$). Median Jaccard coefficient for all three segmentations (J3) was 0.82 (IQR 0.70–0.89). The absolute volume, HD, J2 and J3 for each tumor from each segmentation cycle are shown in Table 2, while a bar chart with the absolute tumor volume for all tumors from all three segmentations is shown in Fig. 1.

Subgroup analyses were performed (Table 3). There was a tendency towards smaller relative volume variability

($p=0.07$) and a significantly higher J2 ($p<0.01$) in larger tumors in segmentation cycle 2 vs. 3. Tumors with well-defined border showed less variability compared to tumors with partially absent and absent border in segmentation cycle 2 vs. 3, with significantly smaller difference in median absolute volume ($p=0.04$), smaller HD ($p<0.01$) and higher J2 ($p=0.04$). Comparing histopathological subtypes, astrocytomas were significantly smaller than oligodendrogliomas in all segmentation cycles ($p\leq 0.04$), but there was no difference in HD, J2 or J3 between the histopathological subgroups. There was no significant difference in tumor volume, HD, J2 or J3, between tumors with low or high ADC-values.

Discussion

In this study we found a better overlap agreement when the same LGGs were repeatedly segmented, with significantly increased J2 between the two last segmentations compared to the two first. We interpret this as a decreased intra-rater variability with increasing experience, which again could

Table 2 Volumes in milliliter (mL) for each tumor and each segmentation cycle

Case No.	Volume cycle 1 (mL)	Volume cycle 2 (mL)	Volume cycle 3 (mL)	HD 1 vs. 2 (mm)	HD 2 vs. 3 (mm)	HD 1 vs. 3 (mm)	J2 1 vs. 2	J2 2 vs. 3	J2 1 vs. 3	J3 1 vs. 2 vs. 3
27	1.4	1.7	1.6	3.3	2.0	3.2	0.89	0.91	0.89	0.85
4	2.0	2.6	1.4	9.9	9.9	3.7	0.82	0.69	0.78	0.66
15	2.1	2.1	1.9	2.0	2.5	2.2	0.91	0.89	0.89	0.85
6	4.9	3.6	4.2	1.3	5.0	9.4	0.82	0.88	0.85	0.80
16	9.3	10.3	9.1	4.6	4.5	2.2	0.90	0.92	0.93	0.88
18	11.3	7.9	6.4	9.8	3.8	10.6	0.78	0.85	0.70	0.70
3	12.1	13.6	11.8	11.0	6.7	9.8	0.86	0.89	0.87	0.82
19	15.1	14.0	11.0	5.9	8.9	9.1	0.87	0.85	0.82	0.78
2	18.0	14.3	6.4	14.9	36.7	46.9	0.73	0.55	0.39	0.36
8	19.9	14.4	11.8	9.0	11.5	14.0	0.79	0.82	0.71	0.68
30	25.2	28.7	28.7	4.5	5.1	5.2	0.85	0.95	0.91	0.89
20	26.4	7.1	5.8	34.3	9.1	32.3	0.40	0.83	0.35	0.40
24	32.1	32.8	32.1	3.7	3.6	3.5	0.96	0.96	0.95	0.93
9	33.9	29.3	28.5	11.6	6.1	12.4	0.89	0.93	0.89	0.86
23	34.4	34.8	30.0	14.0	12.1	8.6	0.91	0.90	0.89	0.85
17	47.6	43.1	47.2	7.3	4.7	5.7	0.93	0.94	0.95	0.91
13	47.9	28.9	25.3	19.0	7.7	21.4	0.72	0.90	0.67	0.70
22	48.0	27.6	19.7	19.6	9.4	21.5	0.71	0.82	0.58	0.61
25	57.7	60.7	59.9	15.4	12.6	7.4	0.92	0.93	0.93	0.89
7	58.1	43.0	39.0	21.2	9.3	22.0	0.83	0.92	0.79	0.81
10	89.4	77.2	56.5	8.6	18.8	19.1	0.90	0.80	0.75	0.71
26	164.2	155.8	155.8	8.6	5.2	9.6	0.93	0.96	0.93	0.92
14	165.9	166.0	163.1	6.2	7.4	6.1	0.96	0.96	0.96	0.94

Hausdorff Distance (HD) in millimeter (mm) and Jaccard coefficient (J2) between segmentation cycle 1 and 2, segmentation cycle 2 and 3 and segmentation cycle 1 and 3. Jaccard coefficient (J3) for segmentations 1, 2 and 3

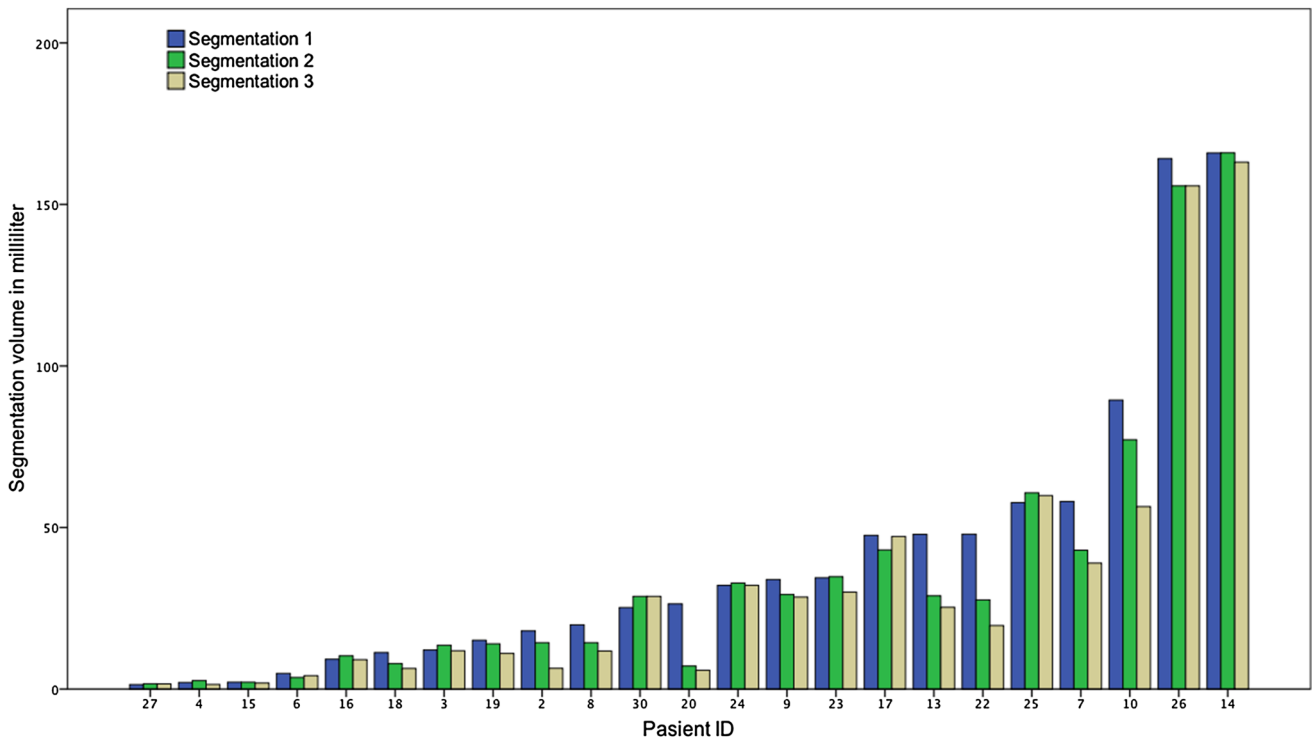


Fig. 1 Bar chart with volume in mL for all tumors and all segmentations, ordered by volume in segmentation 1. Segmentation 1 in blue, segmentation 2 in green and segmentation 3 in yellow

indicate more confidence in tumor border interpretation. There was a non-significant decrease of median difference in tumor volume between these cycles of 1.3 mL, corresponding to a median relative difference of 13–14%. The variability is however better demonstrated when comparing segmentation cycles 1 vs. 3 to segmentation cycles 1 vs. 2, where there was a significant decrease in median tumor volume of 4.1 mL ($p < 0.01$). Although only trending towards significance between the second and third segmentation cycle, the variability measured as relative difference in percent seems larger in smaller lesions ($p = 0.07$), while tumors with well-defined borders showed a significantly smaller variability measured in absolute volume ($p = 0.04$), HD ($p < 0.01$) and J2 ($p = 0.04$). Intra-rater agreement was not associated with histopathological subgroups, and we did not find a clear association between ADC-values and variability of volume measures. This study demonstrates that intra-rater variability of the gold standard volume assessment can be substantial and should be accounted for. Thus, some criteria defining tumor borders and progression are needed if moving from 2D to 3D volume assessment of LGGs.

In clinical situations, growth is often based on so-called “eye-balling” or unsystematic measures of tumor diameters. In clinical trials, tumor size has classically been measured as the product of two orthogonal diameters, measured

on the axial slice with the largest diameter [39]. In the follow-up criteria from the Response Assessment in Neuro-Oncology (RANO) group, bi-diametric measurements are set as the standard method for response evaluation, mostly due to limited availability of volumetric measurements [40]. As LGGs have an irregular slow growth, it is commonly accepted that 3D volumetric measurements easier will catch subtle changes between examinations, although there is a lack of studies comparing 2D and 3D measurements [7, 22, 41, 42]. However, the accuracy of volume or growth measurement is presumably not only dependent on the choice of method, but may also be limited by the operator that has to draw the line between tumor and normal brain in diffusely infiltrating tumors. LGGs usually show no contrast enhancement and segmentation has to rely on the inherent contrast properties of the tissue, which can be very close to normal brain tissue.

Much work has been put into characterization of LGGs in order to determine prognostic predictors, emphasizing the importance of volumetric assessment of the tumors. Pallud et al. have in several studies shown that the radiological growth rate of the tumor can predict malignant transformation [7, 18, 19]. Their work is supported by others, using growth rates to predict transformation and patient outcome within 6 and 12 months [20, 21]. Two studies describe a semi-automatic strategy for quantifying

Table 3 Median tumor volume, absolute and relative volume differences and Hausdorff Distance (HD) for segmentation cycles 1 vs. 2 and cycles 2 vs. 3, Jaccard coefficient for segmentation cycles 1 vs. 2, cycles 2 vs. 3 and cycles 1 vs. 2 vs. 3 in sub-groups, grouped by size (dichotomized by median tumor volume), histopathology, ADC-values (dichotomized by median ADC-value) and border conspicuity

	Small size n = 11	Large size n = 12	p value	Astrocytoma n = 8	Oligodendroglioma n = 10	p value	Low ADC n = 11	High ADC n = 11	p value	Border absent n = 3	Border partially absent n = 8	Border well-defined n = 12	p value
Volume (mL)													
Cycle 1	11.3	47.9	<0.01	17.5	41.2	0.04	19.9	33.9	0.70	19.9	47.8	20.1	0.19
Cycle 2	7.9	38.9	<0.01	12.2	32.0	0.04	13.6	29.3	0.48	14.4	31.8	12.1	0.31
Cycle 3	6.4	35.6	<0.01	10.0	29.3	0.03	11.8	28.7	0.52	11.7	27.7	10.1	0.35
Absolute difference in volume (mL)													
1 vs. 2	-1.1	-4.6	0.46	-2.3	-4.1	0.97	-3.4	-0.1	0.53	-5.5	-4.8	0.0	0.26
2 vs. 3	-1.3	-1.9	0.71	-1.3	-2.6	0.24	-1.3	-1.2	0.87	-7.9	-2.6	-1.0	0.04
Relative difference in volume (%)													
1 vs. 2	-7.1	-7.3	0.85	-8.3	-9.4	0.83	-13.7	0.0	0.22	-20.4	-14.3	0.0	0.27
2 vs. 3	-18.1	-2.4	0.07	-14.8	-10.8	0.83	-12.6	-2.7	0.82	-26.8	-10.8	-7.1	0.10
HD (mm)													
1 vs. 2	9.8	10.1	0.49	6.6	12.8	0.07	17.5	9.9	0.45	9.0	14.7	6.1	0.02
2 vs. 3	6.7	7.5	0.45	4.6	8.5	0.03	9.1	9.4	0.62	18.8	8.5	4.9	<0.01
Jaccard coefficient													
1 vs. 2	0.82	0.91	0.05	0.88	0.87	0.97	0.86	0.89	0.41	0.79	0.84	0.89	0.53
2 vs. 3	0.85	0.93	<0.01	0.88	0.91	0.76	0.89	0.92	0.21	0.80	0.90	0.92	0.04
1 vs. 2 vs. 3	0.80	0.86	0.21	0.82	0.84	0.90	0.80	0.85	0.27	0.68	0.82	0.86	0.11

ADC apparent diffusion coefficient, mL milliliter, mm millimeter

tumor growth using grey level recognition in T2/FLAIR images, and both methods are based on prior manual expert segmentation [27, 30]. In the first, Angelini et al. found a high volume segmentation variability, both in the baseline segmentation and in the follow-up segmentations, because they are all based on manual tracing. In the other Weizman et al. looked at optic pathway and thalamic gliomas and showed quite good correlation between manual and semi-automatic segmentation volumes, but have not calculated intra- or inter-rater variability.

Several prior studies have measured inter-rater variability, but few focused on intra-rater variability. Both Kaus et al. and Akkus et al. found comparable levels of intra-rater variability between manual and semi-automatic segmentation of LGGs, but Jaccard coefficient was not calculated [26, 28]. In both studies intra-rater variability was lower than inter-rater variability. Zou et al. found a highly variable inter-rater DSC from 0.49 to 0.97 in LGG segmentation, comparable with our J2 values ranging from 0.40 to 0.96 [35]. With such high variability, experience in tumor border evaluation and general brain MRI interpretation will be highly important to minimize this factor, aiming for more consistent measurements, especially in a follow-up setting with growth evaluation.

As seen in the present study with a median difference in absolute volume between the first and third segmentation of 4.1 mL, intra-rater variability of manual volume segmentation should not be underestimated. This may be an argument for automatic methods of volume assessments. However, algorithms for automatic segmentation of LGGs have so far been disappointing, and in validation of automatic methods the substantial intra-rater variability of the current gold standard based on manual methods should be kept in mind [33]. For detecting progression or treatment responses in individual patients, a low inter- and intra-rater variability (i.e. reliability) might be more important than agreement with manual methods (i.e. validity). After all, the true volume of any glioma is always larger than depicted with any current imaging modality. In example, Pallud et al. found tumor cells 20 mm from the margins of such FLAIR abnormalities, while Zetterling et al. found IDH1-positive tumor cells up to 14 mm from FLAIR abnormalities [4, 43].

Part of the variation in tumor segmentation presented in this paper could be because of variation in manual initialization of the “GrowCut” algorithm. Manual initialization of semi-automatic segmentation algorithms is an important yet ill addressed topic in the literature. As the tumors in this study have been segmented three times each with different manual initialisations, this study implicitly addresses the question of initialisation for this particular algorithm.

As mentioned, it seems to be an association between relative size and variability. We did not find this association within histopathological subtypes, where we could

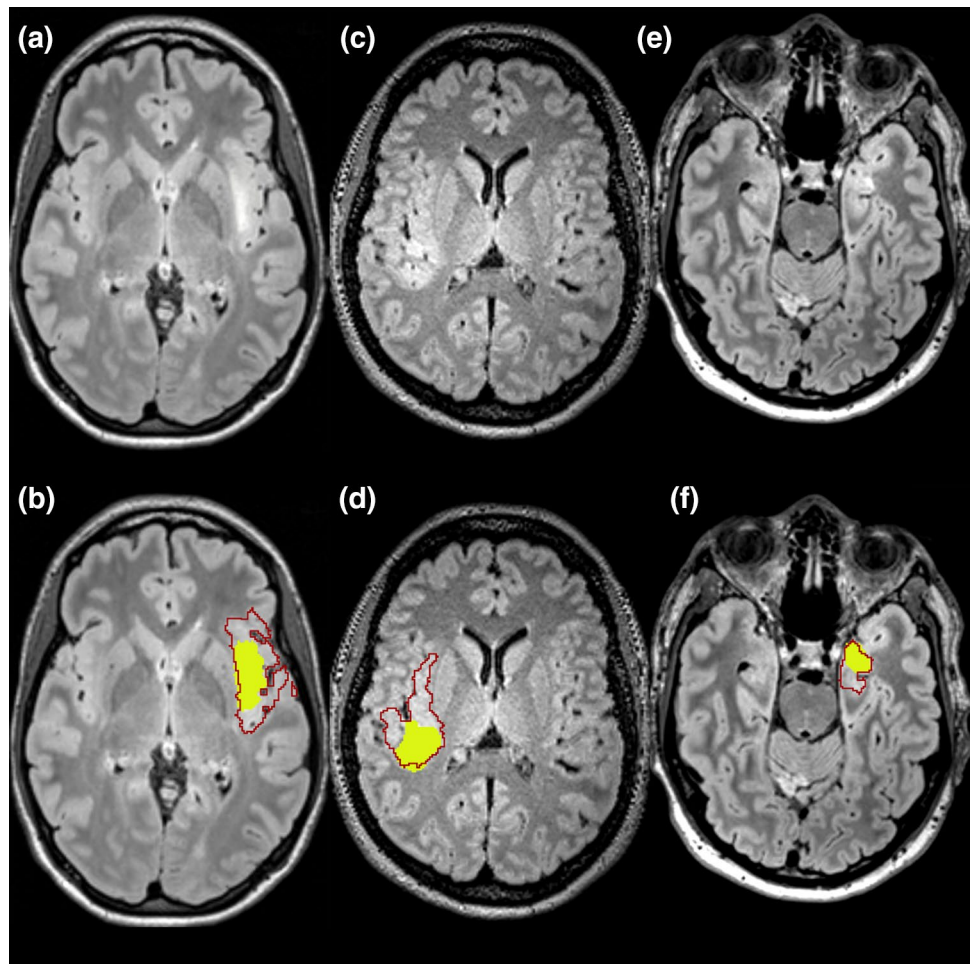
expect that variability would be larger in percent in the often smaller astrocytomas and larger in volume in the often larger oligodendrogliomas. Furthermore, since ADC values are associated with tumor cell density [44], we hypothesized higher relative variability in astrocytomas (that usually are less cell-dense than oligodendrogliomas [1]) with repeated volume assessment. Although we did not find an association between variability and ADC values, this could be a result of the fairly small sample size. Also, with the often high intratumoral heterogeneity, an ADC value from the tumor core instead of the entire tumor might be more representative.

Image acquisition makes the basis for further tumor evaluation and segmentation, and technical parameters need to be optimized. A 2D FLAIR acquisition gives higher image contrast than 3D FLAIR, both between grey and white matter and between lesion and white matter [45]. On the other hand, standard 2D FLAIR sequences typically have thicker slices of 3–5 mm to get sufficient signal-to-noise ratio, as well as interslice gaps. A significant difference between segmentations on 1 mm slices vs. 5 mm slices has been shown [46]. Thus, using interpolation to calculate segmentation volumes, volume estimates are less accurate, especially in small tumors [41, 46]. Therefore, a 3D sequence with isotropic voxels and no interslice gap will usually give a more accurate volume estimation and should possibly be part of standard tumor evaluation. Also, magnetic field strength influence image contrast, with higher contrast-to-noise ratio in FLAIR images acquired on 3.0T compared to 1.5T. This leads to small differences in lesion volume and should be taken into consideration [47–50].

Assessing relative volume differences there are some extreme values, namely 73% (case 20) in segmentation cycles 1 vs. 2 and 55% (case 2) and 45% (case 4) in segmentation cycles 2 vs. 3 (Fig. 2). Case 20 is a left insular glioma considered to have well-defined borders. It was the second tumor segmented in the very beginning, and with more experience the tumor borders were interpreted quite different and more consistently the second and third time, showing a smaller difference of 19%. Case 2 resides in the right insula and case 4 in the left medial temporal lobe. Both were classified to have “absent” tumor borders, making them difficult to delineate consistently. In addition, case 4 is a rather small tumor, with a mean volume of only 2 mL.

Our study has limitations. We do not have a gold standard to compare our segmentation results with, which makes it difficult to evaluate whether we are closer to the biological truth or just in more agreement with ourselves. In general, with a small number of participants, it is difficult to draw conclusions from such sub-group analysis.

Fig. 2 Three tumors with extreme differences. Case 20 without (a) and with (b) segmentation label, segmentation 1 outlined in red and segmentation 2 in yellow. Case 2 without (c) and with (d) segmentation label, segmentation 2 outlined in red and segmentation 3 in yellow. Case 4 without (e) and with (f) segmentation label, segmentation 2 outlined in red and segmentation 3 in yellow



Conclusion

Our study demonstrates that intra-rater variability can be considerable in LGG volume segmentation, with significant volume difference between segmentation cycles. We did however find a decreased intra-rater variability with repeated segmentations measured with DSC, suggestive of an effect of experience. With no exact gold standard for comparison, it can be difficult to point out what makes this effect, whether it is actually getting closer to the true volume, or if the operator is only reaching a better internal agreement with him-/herself, or a combination of the two. Furthermore, we found that there was a significantly lower variability in segmentation of LGGs with well-defined tumor borders. This study shows that some criteria defining tumor borders and progression in 3D volumetric segmentation is needed if moving from 2D to 3D volume assessment of LGGs.

Acknowledgements A.S.J. holds a grant from The Norwegian Cancer Society for glioma research.

Compliance with ethical standards

Conflict of interest O.S. and E.M.B. have fundings from the Regional brain tumor registry of central Norway funded by the liaison committee of St. Olavs University Hospital and NTNU. The authors declare that they have no conflict of interest.

Ethical approval The study has been approved by the Regional Ethical Committee for Health Region Mid-Norway (ref 2014/1674).

References

- Sanai N, Chang S, Berger MS (2011) Low-grade gliomas in adults. *J Neurosurg*. doi:[10.3171/2011.7.jns10238](https://doi.org/10.3171/2011.7.jns10238)
- Buckner JC, Shaw EG, Pugh SL, Chakravarti A, Gilbert MR, Barger GR, Coons S, Ricci P, Bullard D, Brown PD, Stelzer K, Brachman D, Suh JH, Schultz CJ, Bahary JP, Fisher BJ, Kim H, Murtha AD, Bell EH, Won M, Mehta MP, Curran WJ Jr (2016) Radiation plus Procarbazine, CCNU, and Vincristine in Low-Grade Glioma. *N Engl J Med* 374(14):1344–1355. doi:[10.1056/NEJMoa1500925](https://doi.org/10.1056/NEJMoa1500925)
- Pallud J, Capelle L, Taillandier L, Badoual M, Duffau H, Mandonnet E (2013) The silent phase of diffuse low-grade

- gliomas. Is it when we missed the action? *Acta Neurochir (Wien)* 155(12):2237–2242. doi:[10.1007/s00701-013-1886-7](https://doi.org/10.1007/s00701-013-1886-7)
4. Pallud J, Fontaine D, Duffau H, Mandonnet E, Sanai N, Taillandier L, Peruzzi P, Guillevin R, Bauchet L, Bernier V, Baron MH, Guyotat J, Capelle L (2010) Natural history of incidental World Health Organization grade II gliomas. *Ann Neurol* 68(5):727–733. doi:[10.1002/ana.22106](https://doi.org/10.1002/ana.22106)
 5. Lima GL, Zanello M, Mandonnet E, Taillandier L, Pallud J, Duffau H (2015) Incidental diffuse low-grade gliomas: from early detection to preventive neuro-oncological surgery. *Neurosurg Rev*. doi:[10.1007/s10143-015-0675-6](https://doi.org/10.1007/s10143-015-0675-6)
 6. Jakola AS, Myrmet KS, Kloster R, Torp SH, Lindal S, Unsgard G, Solheim O (2012) Comparison of a strategy favoring early surgical resection vs a strategy favoring watchful waiting in low-grade gliomas. *JAMA* 308(18):1881–1888. doi:[10.1001/jama.2012.12807](https://doi.org/10.1001/jama.2012.12807)
 7. Duffau H, Taillandier L (2015) New concepts in the management of diffuse low-grade glioma: Proposal of a multistage and individualized therapeutic approach. *Neuro. Oncol* 17(3):332–342. doi:[10.1093/neuonc/nou153](https://doi.org/10.1093/neuonc/nou153)
 8. Chang EF, Smith JS, Chang SM, Lamborn KR, Prados MD, Butowski N, Barbaro NM, Parsa AT, Berger MS, McDermott MM (2008) Preoperative prognostic classification system for hemispheric low-grade gliomas in adults. *J Neurosurg* 109(5):817–824. doi:[10.3171/JNS/2008/109/11/0817](https://doi.org/10.3171/JNS/2008/109/11/0817)
 9. Pignatti F, van den Bent M, Curran D, Debruyne C, Sylvester R, Therasse P, Afra D, Cornu P, Bolla M, Vecht C, Karim AB, European Organization for R, Treatment of Cancer Brain Tumor Cooperative G, European Organization for R, Treatment of Cancer Radiotherapy Cooperative G (2002) Prognostic factors for survival in adult patients with cerebral low-grade glioma. *J Clin Oncol* 20(8):2076–2084
 10. Ahmadi R, Dictus C, Hartmann C, Zurn O, Edler L, Hartmann M, Combs S, Herold-Mende C, Wirtz CR, Unterberg A (2009) Long-term outcome and survival of surgically treated supratentorial low-grade glioma in adult patients. *Acta Neurochir (Wien)* 151(11):1359–1365. doi:[10.1007/s00701-009-0435-x](https://doi.org/10.1007/s00701-009-0435-x)
 11. Capelle L, Fontaine D, Mandonnet E, Taillandier L, Golmard JL, Bauchet L, Pallud J, Peruzzi P, Baron MH, Kujas M, Guyotat J, Guillevin R, Frenay M, Taillibert S, Colin P, Rigau V, Vandebos F, Pinelli C, Duffau H, French Reseau d'Etude des G (2013) Spontaneous and therapeutic prognostic factors in adult hemispheric World Health Organization Grade II gliomas: a series of 1097 cases: clinical article. *J Neurosurg* 118(6):1157–1168. doi:[10.3171/2013.1.JNS121](https://doi.org/10.3171/2013.1.JNS121)
 12. Chaichana KL, McGirt MJ, Latta J, Olivi A, Quinones-Hinojosa A (2010) Recurrence and malignant degeneration after resection of adult hemispheric low-grade gliomas. *J Neurosurg* 112(1):10–17. doi:[10.3171/2008.10.JNS08608](https://doi.org/10.3171/2008.10.JNS08608)
 13. Claus EB, Horlacher A, Hsu L, Schwartz RB, Dello-Iacono D, Talos F, Jolesz FA, Black PM (2005) Survival rates in patients with low-grade glioma after intraoperative magnetic resonance image guidance. *Cancer* 103(6):1227–1233. doi:[10.1002/cncr.20867](https://doi.org/10.1002/cncr.20867)
 14. Ius T, Isola M, Budai R, Pauletto G, Tomasino B, Fadiga L, Skrap M (2012) Low-grade glioma surgery in eloquent areas: volumetric analysis of extent of resection and its impact on overall survival. A single-institution experience in 190 patients: clinical article. *J Neurosurg* 117(6):1039–1052. doi:[10.3171/2012.8.JNS12393](https://doi.org/10.3171/2012.8.JNS12393)
 15. McGirt MJ, Chaichana KL, Attenello FJ, Weingart JD, Than K, Burger PC, Olivi A, Brem H, Quinones-Hinojosa A (2008) Extent of surgical resection is independently associated with survival in patients with hemispheric infiltrating low-grade gliomas. *Neurosurgery* 63(4):700–707. doi:[10.1227/01.NEU.0000325729.41085.73](https://doi.org/10.1227/01.NEU.0000325729.41085.73)
 16. Sanai N, Berger MS (2009) Operative techniques for gliomas and the value of extent of resection. *Neurotheraphy* 6(3):478–486. doi:[10.1016/j.nurt.2009.04.005](https://doi.org/10.1016/j.nurt.2009.04.005)
 17. Smith JS, Chang EF, Lamborn KR, Chang SM, Prados MD, Cha S, Tihan T, Vandenberg S, McDermott MW, Berger MS (2008) Role of extent of resection in the long-term outcome of low-grade hemispheric gliomas. *J Clin Oncol* 26(8):1338–1345. doi:[10.1200/JCO.2007.13.9337](https://doi.org/10.1200/JCO.2007.13.9337)
 18. Pallud J, Blonski M, Mandonnet E, Audureau E, Fontaine D, Sanai N, Bauchet L, Peruzzi P, Frenay M, Colin P, Guillevin R, Bernier V, Baron MH, Guyotat J, Duffau H, Taillandier L, Capelle L (2013) Velocity of tumor spontaneous expansion predicts long-term outcomes for diffuse low-grade gliomas. *Neuro Oncol* 15(5):595–606. doi:[10.1093/neuonc/nos331](https://doi.org/10.1093/neuonc/nos331)
 19. Pallud J, Taillandier L, Capelle L, Fontaine D, Peyre M, Ducray F, Duffau H, Mandonnet E (2012) Quantitative morphological magnetic resonance imaging follow-up of low-grade glioma: a plea for systematic measurement of growth rates. *Neurosurgery* 71(3):729–739 (discussion 739–740). doi:[10.1227/NEU.0b013e31826213de](https://doi.org/10.1227/NEU.0b013e31826213de)
 20. Brasil Caseiras G, Ciccarello O, Altmann DR, Benton CE, Tozer DJ, Tofts PS, Yousry TA, Rees J, Waldman AD, Jäger HR (2009) Low-grade gliomas: six-month tumor growth predicts patient outcome better than admission tumor volume, relative cerebral blood volume, and apparent diffusion coefficient. *Radiology* 253(2):505–512. doi:[10.1148/radiol.2532081623](https://doi.org/10.1148/radiol.2532081623)
 21. Rees J, Watt H, Jäger HR, Benton C, Tozer D, Tofts P, Waldman A (2009) Volumes and growth rates of untreated adult low-grade gliomas indicate risk of early malignant transformation. *Eur J Radiol* 72(1):54–64. doi:[10.1016/j.ejrad.2008.06.013](https://doi.org/10.1016/j.ejrad.2008.06.013)
 22. Jakola AS, Moen KG, Solheim O, Kvistad KA (2013) “No growth” on serial MRI scans of a low grade glioma? *Acta Neurochir (Wien)* 155(12):2243–2244. doi:[10.1007/s00701-013-1914-7](https://doi.org/10.1007/s00701-013-1914-7)
 23. Mandonnet E, Pallud J, Fontaine D, Taillandier L, Bauchet L, Peruzzi P, Guyotat J, Bernier V, Baron MH, Duffau H, Capelle L (2010) Inter- and inpatients comparison of WHO grade II glioma kinetics before and after surgical resection. *Neurosurg Res* 33(1):91–96. doi:[10.1007/s10143-009-0229-x](https://doi.org/10.1007/s10143-009-0229-x)
 24. Bauer S, Wiest R, Nolte LP, Reyes M (2013) A survey of MRI-based medical image analysis for brain tumor studies. *Phys Med Biol* 58(13):R97–R129. doi:[10.1088/0031-9155/58/13/R97](https://doi.org/10.1088/0031-9155/58/13/R97)
 25. Porz N, Bauer S, Pica A, Schuch P, Beck J, Verma RK, Slotboom J, Reyes M, Wiest R (2014) Multi-modal glioblastoma segmentation: man versus machine. *PLoS One* 9(5):e96873. doi:[10.1371/journal.pone.0096873](https://doi.org/10.1371/journal.pone.0096873)
 26. Akkus Z, Sedlar J, Coufalova L, Korfiatis P, Kline TL, Warner JD, Agrawal J, Erickson BJ (2015) Semi-automated segmentation of pre-operative low grade gliomas in magnetic resonance imaging. *Cancer Imaging* 15:12. doi:[10.1186/s40644-015-0047-z](https://doi.org/10.1186/s40644-015-0047-z)
 27. Angelini ED, Delon J, Bah AB, Capelle L, Mandonnet E (2012) Differential MRI analysis for quantification of low grade glioma growth. *Med Image Anal* 16(1):114–126. doi:[10.1016/j.media.2011.05.014](https://doi.org/10.1016/j.media.2011.05.014)
 28. Kaus MR, Warfield SK, Nabavi A, Black PM, Jolesz FA, Kikinis R (2001) Automated segmentation of MR images of brain tumors. *Radiology* 218(2):586–591. doi:[10.1148/radiology.218.2.r01fe44586](https://doi.org/10.1148/radiology.218.2.r01fe44586)
 29. Mazzara GP, Velthuisen RP, Pearlman JL, Greenberg HM, Wagner H (2004) Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation. *Int J Radiat Oncol Biol Phys* 59(1):300–312. doi:[10.1016/j.ijrobp.2004.01.026](https://doi.org/10.1016/j.ijrobp.2004.01.026)
 30. Weizman L, Sira LB, Joskowicz L, Rubin DL, Yeom KW, Constantini S, Shofty B, Bashat DB (2014) Semiautomatic segmentation and follow-up of multicomponent low-grade tumors

- in longitudinal brain MRI studies. *Med Phys* 41(5):052303. doi:[10.1118/1.4871040](https://doi.org/10.1118/1.4871040)
31. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, Buatti J, Aylward S, Miller JV, Pieper S, Kikinis R (2012) 3D Slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging* 30(9):1323–1341. doi:[10.1016/j.mri.2012.05.001](https://doi.org/10.1016/j.mri.2012.05.001)
 32. Egger J, Kapur T, Fedorov A, Pieper S, Miller JV, Veeraraghavan H, Freisleben B, Golby AJ, Nimsky C, Kikinis R (2013) GBM volumetry using the 3D Slicer medical image computing platform. *Sci Rep* 3:1364. doi:[10.1038/srep01364](https://doi.org/10.1038/srep01364)
 33. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren J, Porz N, Slotboom J, Wiest R, Lanczi L, Gerstner E, Weber MA, Arbel T, Avants BB, Ayache N, Buendia P, Collins DL, Cordier N, Corso JJ, Criminisi A, Das T, Delingette H, Demiralp C, Durst CR, Dojat M, Doyle S, Festa J, Forbes F, Geremia E, Glocker B, Golland P, Guo X, Hamamci A, Iftekharrudin KM, Jena R, John NM, Konukoglu E, Lashkari D, Mariz JA, Meier R, Pereira S, Precup D, Price SJ, Raviv TR, Reza SM, Ryan M, Sarikaya D, Schwartz L, Shin HC, Shotton J, Silva CA, Sousa N, Subbanna NK, Szekely G, Taylor TJ, Thomas OM, Tustison NJ, Unal G, Vasseur F, Wintermark M, Ye DH, 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(10):1993–2024. doi:[10.1109/TMI.2014.2377694](https://doi.org/10.1109/TMI.2014.2377694)
 34. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC (1994) Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Trans Med Imaging* 13(4):716–724. doi:[10.1109/42.363096](https://doi.org/10.1109/42.363096)
 35. Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ, Wells WM 3rd, Jolesz FA, Kikinis R (2004) Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol* 11(2):178–189
 36. Crum WR, Camara O, Hill DL (2006) Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Trans Med Imaging* 25(11):1451–1461. doi:[10.1109/TMI.2006.880587](https://doi.org/10.1109/TMI.2006.880587)
 37. McHugh ML (2012) Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 22(3):276–282
 38. Rosner BA (2011) *Fundamentals of biostatistics*. 7th edn. Brooks/Cole. Cengage Learning, Boston
 39. Macdonald DR, Cascino TL, Schold SC Jr, Cairncross JG (1990) Response criteria for phase II studies of supratentorial malignant glioma. *J Clin Oncol* 8(7):1277–1280
 40. van den Bent MJ, Wefel JS, Schiff D, Taphoorn MJ, Jaeckle K, Junck L, Armstrong T, Choucair A, Waldman AD, Gorlia T, Chamberlain M, Baumert BG, Vogelbaum MA, Macdonald DR, Reardon DA, Wen PY, Chang SM, Jacobs AH (2011) Response assessment in neuro-oncology (a report of the RANO group): assessment of outcome in trials of diffuse low-grade gliomas. *Lancet Oncol* 12(6):583–593. doi:[10.1016/S1470-2045\(11\)70057-2](https://doi.org/10.1016/S1470-2045(11)70057-2)
 41. Schmitt P, Mandonnet E, Perdreau A, Angelini ED (2013) Effects of slice thickness and head rotation when measuring glioma sizes on MRI: in support of volume segmentation versus two largest diameters methods. *J Neurooncol* 112(2):165–172. doi:[10.1007/s11060-013-1051-4](https://doi.org/10.1007/s11060-013-1051-4)
 42. Sorensen AG, Patel S, Harmath C, Bridges S, Synnott J, Sievers A, Yoon YH, Lee EJ, Yang MC, Lewis RF, Harris GJ, Lev M, Schaefer PW, Buchbinder BR, Barest G, Yamada K, Ponzio J, Kwon HY, Gemmete J, Farkas J, Tievsky AL, Ziegler RB, Salhus MR, Weisskoff R (2001) Comparison of diameter and perimeter methods for tumor volume calculation. *J Clin Oncol* 19(2):551–557
 43. Zetterling M, Roodakker KR, Berntsson SG, Edqvist PH, Latini F, Landtblom AM, Ponten F, Alafuzoff I, Larsson EM, Smits A (2016) Extension of diffuse low-grade gliomas beyond radiological borders as shown by the coregistration of histopathological and magnetic resonance imaging data. *J Neurosurg*. doi:[10.3171/2015.10.jns15583](https://doi.org/10.3171/2015.10.jns15583)
 44. Chen L, Liu M, Bao J, Xia Y, Zhang J, Zhang L, Huang X, Wang J (2013) The correlation between apparent diffusion coefficient and tumor cellularity in patients: a meta-analysis. *PLoS One* 8(11):e79008. doi:[10.1371/journal.pone.0079008](https://doi.org/10.1371/journal.pone.0079008)
 45. Tschampa HJ, Urbach H, Malter M, Surges R, Greschus S, Gieseke J (2015) Magnetic resonance imaging of focal cortical dysplasia: comparison of 3D and 2D fluid attenuated inversion recovery sequences at 3 T. *Epilepsy Res* 116:8–14. doi:[10.1016/j.eplepsyres.2015.07.004](https://doi.org/10.1016/j.eplepsyres.2015.07.004)
 46. Stensjoen AL, Solheim O, Kvistad KA, Haberg AK, Salvesen O, Berntsen EM (2015) Growth dynamics of untreated glioblastomas in vivo. *Neuro Oncol* 17(10):1402–1411. doi:[10.1093/neuonc/nov029](https://doi.org/10.1093/neuonc/nov029)
 47. Tselikas L, Souillard-Scemama R, Naggara O, Mellerio C, Varlet P, Dezamis E, Domont J, Dhermain F, Devaux B, Chretien F, Meder JF, Pallud J, Oppenheim C (2015) Imaging of gliomas at 1.5 and 3T—A comparative study. *Neuro Onco* 17(6):895–900. doi:[10.1093/neuonc/nou332](https://doi.org/10.1093/neuonc/nou332)
 48. Neema M, Guss ZD, Stankiewicz JM, Arora A, Healy BC, Bakshi R (2009) Normal findings on brain fluid-attenuated inversion recovery MR images at 3T. *AJNR Am J Neuroradiol* 30(5):911–916. doi:[10.3174/ajnr.A1514](https://doi.org/10.3174/ajnr.A1514)
 49. Kamada K, Kakeda S, Ohnari N, Moriya J, Sato T, Korogi Y (2008) Signal intensity of motor and sensory cortices on T2-weighted and FLAIR images: intraindividual comparison of 1.5T and 3T MRI. *Eur Radiol* 18(12):2949–2955. doi:[10.1007/s00330-008-1069-8](https://doi.org/10.1007/s00330-008-1069-8)
 50. Guarnaschelli JN, Vagal AS, McKenzie JT, McPherson CM, Warnick RE, Batra V, Breneman JC, Lamba MA (2014) Target definition for malignant gliomas: no difference in radiation treatment volumes between 1.5T and 3T magnetic resonance imaging. *Pract Radiat Oncol* 4(5):e195–e201. doi:[10.1016/j.prro.2013.11.003](https://doi.org/10.1016/j.prro.2013.11.003)