



Reliability of Semi-Automated Segmentations in Glioblastoma

T. Huber · G. Alber · S. Bette · T. Boeckh-Behrens ·
J. Gempt · F. Ringel · E. Alberts · C. Zimmer · J. S. Bauer

Received: 14 July 2015 / Accepted: 29 September 2015 / Published online: 21 October 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract

Purpose In glioblastoma, quantitative volumetric measurements of contrast-enhancing or fluid-attenuated inversion recovery (FLAIR) hyperintense tumor compartments are needed for an objective assessment of therapy response. The aim of this study was to evaluate the reliability of a semi-automated, region-growing segmentation tool for determining tumor volume in patients with glioblastoma among different users of the software.

Methods A total of 320 segmentations of tumor-associated FLAIR changes and contrast-enhancing tumor tissue were performed by different raters (neuroradiologists, medical students, and volunteers). All patients underwent high-resolution magnetic resonance imaging including a 3D-FLAIR and a 3D-MPRage sequence. Segmentations were done using a semi-automated, region-growing segmentation tool. Intra- and inter-rater-reliability were addressed by intra-class-correlation (ICC). Root-mean-square error (RMSE) was used to determine the precision error. Dice score was calculated to measure the overlap between segmentations.

Results Semi-automated segmentation showed a high ICC (>0.985) for all groups indicating an excellent intra- and

inter-rater-reliability. Significant smaller precision errors and higher Dice scores were observed for FLAIR segmentations compared with segmentations of contrast-enhancement. Single rater segmentations showed the lowest RMSE for FLAIR of 3.3 % (MPRage: 8.2 %). Both, single raters and neuroradiologists had the lowest precision error for longitudinal evaluation of FLAIR changes.

Conclusions Semi-automated volumetry of glioblastoma was reliably performed by all groups of raters, even without neuroradiologic expertise. Interestingly, segmentations of tumor-associated FLAIR changes were more reliable than segmentations of contrast enhancement. In longitudinal evaluations, an experienced rater can detect progressive FLAIR changes of less than 15 % reliably in a quantitative way which could help to detect progressive disease earlier.

Keywords Brain tumor segmentation · Glioblastoma · Semi-automated segmentation · Region-growing · Smartbrush · Reliability

Introduction

Glioblastoma is the most common primary malignant brain tumor with still a very poor prognosis despite modern therapeutic strategies like image-guided resection, chemo- and radiotherapy [1, 2]. Since a definite curative therapy is lacking, lengthening of overall survival remains the main goal in most cases [3]. Promising novel therapies like bevacizumab, which targets tumor angiogenesis, unfortunately failed to show a benefit in overall survival, nevertheless recent studies observed a prolonged progression-free survival [4]. So far, early detection and repeated resections, also of little recurrences, seem to prolong overall survival, even though studies are still controversially discussed [3, 5, 6].

T. Huber (✉) · G. Alber · S. Bette · T. Boeckh-Behrens ·
E. Alberts · C. Zimmer · J. S. Bauer
Department of Neuroradiology, Klinikum rechts der Isar,
Technical University of Munich,
Ismaninger Str. 22,
81675 Munich, Germany
e-mail: thomas-huber@tum.de

J. Gempt · F. Ringel
Department of Neurosurgery, Klinikum rechts der Isar,
Technical University of Munich,
Ismaninger Str. 22,
81675 Munich, Germany

Postoperative and therapy-related MR tumor monitoring plays a crucial role to determine response, stable or progressive disease and to distinguish between progression and therapeutic-induced pseudo-progression [7]. In 1990, Macdonald et al. introduced response criteria based on a two-dimensional measure of contrast enhancement on computed tomography scans which got later revised by the RANO (response assessment in neuro-oncology) working group and which are now widely applied for magnetic resonance imaging (MRI) [7, 8]. Even though RANO criteria provided a quantitative and objective attempt to address radiographic changes in the course of glioblastoma, there are still several limitations due to the two-dimensional evaluation, for example, irregular shaped tumors, multifocal lesions, or cystic components [7, 9]. A three-dimensional volumetric approach could overcome these limitations. With modern software solutions being available, it is likely that tumor volume, as a quantitative measure, could play an increasingly important role in the decision making process of neuroradiologists in postoperative tumor monitoring [9].

Different approaches have been made for assessing tumor volume including manual, automated, or semi-automated segmentation methods [10–12]. However, for interpreting volumetric results, it is important to benchmark the segmentation tool in terms of intra- and inter-rater reliability and precision error. Only changes in volume exceeding certain cutoffs, like the least significant change (LSC), should be regarded as significant. Thus, the purpose of this study was to evaluate the reliability of a commercially available semi-automatic segmentation tool in glioblastoma patients. Secondary objectives were to assess the user-dependence for different experience levels and to distinguish between the segmentation of fluid-attenuated inversion recovery (FLAIR) volume (FV) and contrast-enhancing volume (CEV).

Methods

Patients

A total of 320 segmentations of FLAIR and magnetization prepared rapid gradient echo (MPRage) sequences were done in five patients with glioblastoma (four male individuals, mean age at imaging 58.8 ± 12.0 years). All patients underwent tumor resection at the Department of Neurosurgery at our institution and received preoperative MRI and postoperative follow-up MRI at the Department of Neuroradiology between November 2013 and August 2014. The study was approved by the local ethics committee in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments [13]. Histopathological

analysis was done at the Department of Neuropathology and confirmed the diagnosis of glioblastoma in all cases.

MR-Imaging

All patients underwent high-resolution MRI on a 3 T scanner (Achieva 3 T, Philips Medical Systems, The Netherlands) using an 8-channel or 16-channel phased array head coil. A 3D T2-weighted FLAIR sequence ($1.11 \times 1.11 \times 1.12$ mm³, TR/TE of 4800/306 ms) and a 3D T1-weighted MPRage sequence (isotropic resolution 1 mm³, TR/TE 9/4 ms) with and without contrast agent, aligned parallel to the anterior/posterior commissure lines were acquired. A contrast medium injection system was used (Spectris Solaris EP, Siemens Medical, Erlangen, Germany) for the administration of Magnograf® (MaRoTrast, Jena, Germany; 0.2 ml/kg body weight) as contrast agent.

Raters

Intra-rater reliability was assessed by one experienced single rater. Inter-rater reliability was assessed by three different groups of raters separately. The first group consisted of four nonexperienced raters without any medical background (volunteers) who got a brief introduction of 30 min into the software, principles of brain tumor segmentation and anatomy. For the second group, four medical students were chosen (medical students), who got the same introduction as the first group. Volunteers and medical students had stand-by supervision of an experienced rater while segmentations were done. The third group consisted of four experienced physicians working at the Department of Neuroradiology (neuroradiologists), who got a short software introduction only. The criteria of segmentation (see below) were told to all raters prior to the segmentation.

Semi-automated Volumetry

All 320 segmentations of FLAIR and MPRage sequences were done with the novel tool “smartbrush”/BrainLab Elements (BrainLab, Feldkirchen, Germany). Smartbrush is a semi-automated tool for segmentation, based on a region-growing algorithm, a standard technique in medical image processing [12, 14, 15]. First, a region-growing algorithm-aided, 2D-segmentation is manually drawn in the central part of the tumor which is then 3D-interpolated when feeding the algorithm with an additional 2D-segmentation in a perpendicular slice. Manual changes to the segmentation can be easily realized by adding or erasing certain regions of interest either with the help of the region-growing algorithm or completely manually. In each tumor both, FV and CEV were segmented separately. Segmentations of FV included all perifocal and tumor-associated FLAIR hyperin-

tensities. Surrounding nontumor related hyperintense spots (e.g., microangiopathy) were not included in FV as well as the resection cavity and cysts (Fig. 1). For CEV, contrast-enhancement only should be segmented, whereas bigger cysts, ventricular plexus, vessels or T1-weighted hyperintense blood residuals should not be segmented. Therefore, the MPRage sequence without contrast agent or the subtraction image of the MPRage pre and post contrast agent were displayed on a different screen of the same workstation that was used for segmentation. To reduce variability of volume averaging only segmentations with a total volume of $>0.5 \text{ cm}^3$ were considered measurable [7]. For assessment of intra-rater reliability, the images of each patient were segmented four times by the same rater at baseline and follow-up MRI with an interval of 1 week in between each segmentation approach. For determining inter-rater reliability, the images of each patient were segmented by all rat-

ers of a group at baseline and follow-up MRI. The relative change in volume between baseline and follow-up MRI was calculated as the quotient of follow-up volume and initial volume of segmentations and is displayed as a factor (f); $f > 1$ indicates an increase in volume, $f < 1$ shows a decrease in volume between the two different time points.

Statistical Analysis and Illustrations

Consistency among the segmentations of a single rater is referred to as intra-rater reliability. Consistency among the segmentations of different raters is termed inter-rater reliability. Both, intra- and inter-rater reliability were assessed by intra-class correlation (ICC) in a two-way mixed, consistency, average-measure approach [16, 17]. ICC estimates can range between 1 showing perfect agreement and 0 if only random agreement exists. Cutoffs for a qualitative rat-

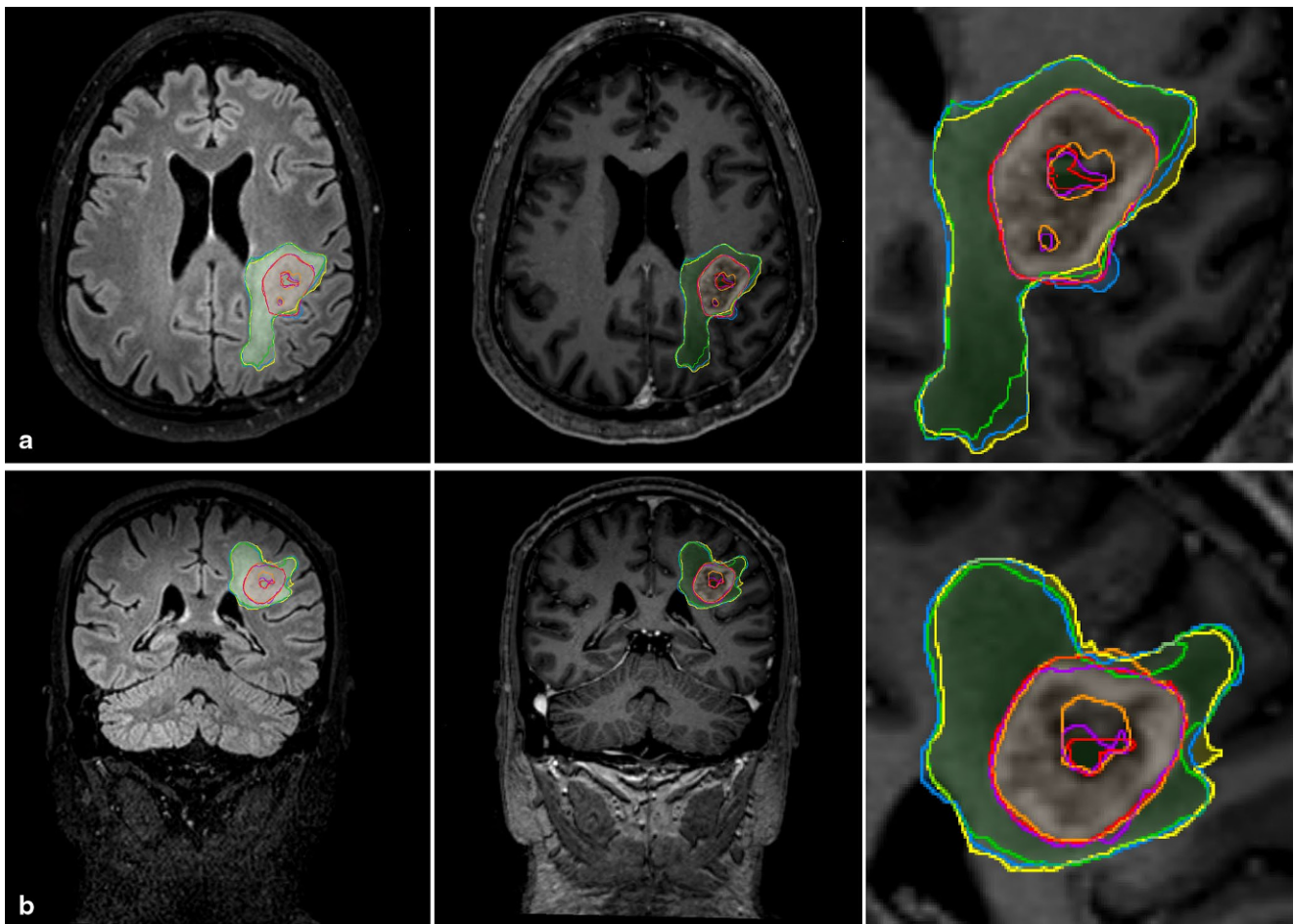


Fig. 1 Exemplary data set of patient 2 showing different delineations for fluid-attenuated inversion recovery volume (FV) (*outer rims*) and contrast-enhancing volume (CEV) (*inner rims*) of a central cystic/necrotic glioblastoma in the left parietal lobe at baseline magnetic resonance imaging. FLAIR images (*left*) and post contrast MPRage

images (*middle* and magnified section on the *right*) in axial **a** and coronal view **b**. Colors for FV: *yellow* (neuroradiologist), *green* (medical student), *blue* (volunteer); Colors for CEV: *orange* (neuroradiologist), *red* (medical student) and *purple* (volunteer)

ing of ICC are poor (<.4), fair (.40–.59), good (.60–.74), and excellent (.75–1.0) [18].

The coefficient of variation (CoV) was calculated as the quotient of standard deviation (SD) and *arithmetic* mean (x) of the different segmentations (1). Significance between CoV in FV and CEV was calculated by Wilcoxon signed-rank test for dependent samples. For precision error, the root-mean-square error (RMSE) was calculated as the root mean square of CoV (2) [19]. Knowing the RMSE, the LSC was calculated by multiplying the RMSE with the factor 2.77 [20], with changes exceeding the LSC considered as statistically significant on a 95% confidence interval. As a measure for the overlap between two segmentations (A, B) the Dice score was applied (3) [21]. A Dice score of 1 shows perfect agreement between two segmentations, a Dice score of 0 indicates no overlapping regions.

$$CoV_n = \frac{SD_n}{x} \times 100\% \quad (1)$$

$$RMSE = \sqrt{\sum_{n=1}^m CoV_n^2 / m} \quad (2)$$

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|} \quad (3)$$

Calculations of Wilcoxon signed-rank test and ICC were done with IBM SPSS Statistics, release 23.0 (IBM, Armonk, NY, USA). All other calculations were done with Excel (Microsoft, Redmond, USA). Illustrations were done with BrainLab Elements (BrainLab, Feldkirchen, Germany) and Power Point 2010 (Microsoft, Redmond, USA). Tables were drawn with Word 2010 (Microsoft, Redmond, USA) and IBM SPSS Statistics, release 23.0 (IBM, Armonk, NY, USA).

Results

Total tumor volumes are shown in Table 1 with the lowest volume for FV of 29.0 cm³ (± 3.3 cm³) and maximum FV of 144.4 cm³ (± 4.8 cm³). Minimal CEV was below 0.5 cm³ with a maximum of 26.2 cm³ (± 2.8 cm³).

Intra-rater Reliability

Intra-rater reliability was excellent with an ICC of 0.998 [Confidence interval (CI) 0.996–1.000] for single rater segmentations of FLAIR images and with an ICC of 0.990 [CI 0.974–0.998] for post contrast MPRage images [18, 22]. The precision error for segmentations of FV trended towards lower values than for segmentations of CEV

($p = .086$). Overall, RMSE for FV was 3.3% whereas RMSE for CEV was 8.2% (Table 2). The median Dice score, as a measure for the overlap between the segmentations, was .92 for FLAIR and .88 for contrast enhancement showing higher agreement for FLAIR segmentations (Fig. 2).

Inter-rater Reliability

Among the different groups of raters ICC for FV was 0.996 [CI 0.989–0.999] for neuroradiologists, 0.994 [CI 0.985–0.998] for medical students and 0.996 [CI 0.990–0.999] for volunteers. ICC for CEV was 0.985 [CI 0.956–0.996] for neuroradiologists, 0.988 [CI 0.965–0.997] for medical students and 0.991 [CI 0.975–0.998] for volunteers, each indicating an excellent inter-rater reliability [18, 22]. Excellent inter-rater reliability shows that there is high agreement of segmented volume in each group and only minimal measurement failure which does not substantially decrease statistical power [17]. Again, the precision error for segmentations of FV compared with segmentations of CEV showed a trend towards lower values for medical students ($p = .066$) or was significantly lower for the group of neuroradiologists ($p = .011$) and volunteers ($p = .011$). RMSE for FV was 9.2% for the group of neuroradiologists whereas RMSE for CEV was 16.7%. RMSE in the group of medical students was 8.5% for FV versus 13.6% for CEV. Volunteers showed the lowest RMSE among all groups of raters for single time point segmentations: RMSE was 7.3% for FV while RMSE for CEV was 10.8%. However, precision error was significantly higher among all inter-rater groups compared with single rater segmentations ($p = .009$ for FV and $p = .036$ for CEV, data not shown in Table 2). Again, the overlap between segmentations of FLAIR was better in every group compared with the overlap of CEV segmentations with a median Dice score for neuroradiologists of .90 for FV and .83 for CEV. Median Dice scores for medical students were .88 for FV and .84 for CEV and for the group of volunteers .88 for FV and .84 for CEV. Visual comparison of segmentations between the groups revealed differences in ambiguous cases (Fig. 3).

Longitudinal Evaluation (Table 3)

The relative change in volume between baseline and follow-up MRI was calculated for each rater in each group. Here, single rater segmentations showed the lowest RMSE for FV of 5.2%. Remarkably, RMSE of neuroradiologists was only 7.5% for FLAIR segmentations, showing an improvement compared with RMSE for single segmentations in this group (9.2%). For medical students and volunteers, RMSE for FV was 10.1% each. RMSE for segmentations of contrast enhancement was higher for every group of raters compared with single segmentations: 12.7% for single rater,

Table 1 Mean volumes for FLAIR volume and contrast-enhancing volume as well as the relative change of volume between baseline and follow-up segmentations for all raters. The relative change (rel. change) of volume is displayed as a factor (f), with $f > 1$ indicating an increase in volume whereas $f < 1$ shows a decrease in volume

	Patient #	FLAIR			MPRage		
		FV (cm ³)	±SD (cm ³)	CoV (%)	CEV (cm ³)	±SD (cm ³)	CoV (%)
Baseline	1	118.5	8.5	7.2	14.6	1.7	11.6
	2	32.2	3.7	11.6	9.9	1.3	12.8
	3	29.0	3.3	11.2	3.6	0.9	26.6
	4	137.0	11.1	8.1	26.2	2.8	10.6
	5	59.3	4.2	7.1	7.9	0.7	8.5
Follow-up	1	144.4	4.8	3.3	22.0	2.6	11.7
	2	125.5	2.0	1.6	23.8	2.0	8.4
	3	35.8	4.0	11.3	2.8	0.3	12.3
	4	70.4	8.8	12.6	<0.5	-	-
	5	78.2	3.1	3.9	10.8	1.2	11.3
	Patient #	Rel. change in FV (f)	±SD	CoV (%)	Rel. change in CEV (f)	±SD	CoV (%)
Rel. change	1	1.22	0.07	5.60	1.51	0.21	13.61
	2	3.95	0.49	12.51	2.45	0.37	15.31
	3	1.24	0.09	7.07	0.81	0.18	22.56
	4	0.52	0.08	16.24	-	-	-
	5	1.32	0.10	7.33	1.37	0.13	9.49

FV FLAIR volume, CEV contrast enhancing volume, SD standard deviation, CoV coefficient of variation (quotient of standard deviation and arithmetic mean × 100)

Table 2 Root-mean-square error for all groups of raters for FLAIR segmentations (FV) and segmentations of contrast-enhancing tumor tissue (CEV) in baseline and follow-up magnetic resonance imaging. The number of segmentations in each category (n) is displayed as well as the p value for Wilcoxon signed-rank test between FV and CEV computed for each group independently

All segmentations	FV		CEV		p
	RMSE %	n	RMSE %	n	
Single rater	3.3	40	8.2	36	.086
Neuroradiologists	9.2	40	16.7	36	.011
Medical students	8.5	40	13.6	36	.066
Volunteers	7.3	40	10.8	36	.011
All raters	7.4	160	12.7	144	<.001

16.9% for neuroradiologists, 15.3% for medical students and 12.9% for volunteers.

Discussion

Quantitative volumetric reports of contrast-enhancing or FLAIR hyperintense tumor compartments are needed for an objective evaluation of stable or progressive disease in patients with glioblastoma [23]. This study showed that semi-automated segmentations of glioblastoma can be done reliably by different groups of raters with a commercial software solution based on a region-growing algorithm. Single rater segmentations showed the lowest precision error. In all groups, segmentations of FV showed lower precision errors compared with segmentations of CEV. In the longitudinal

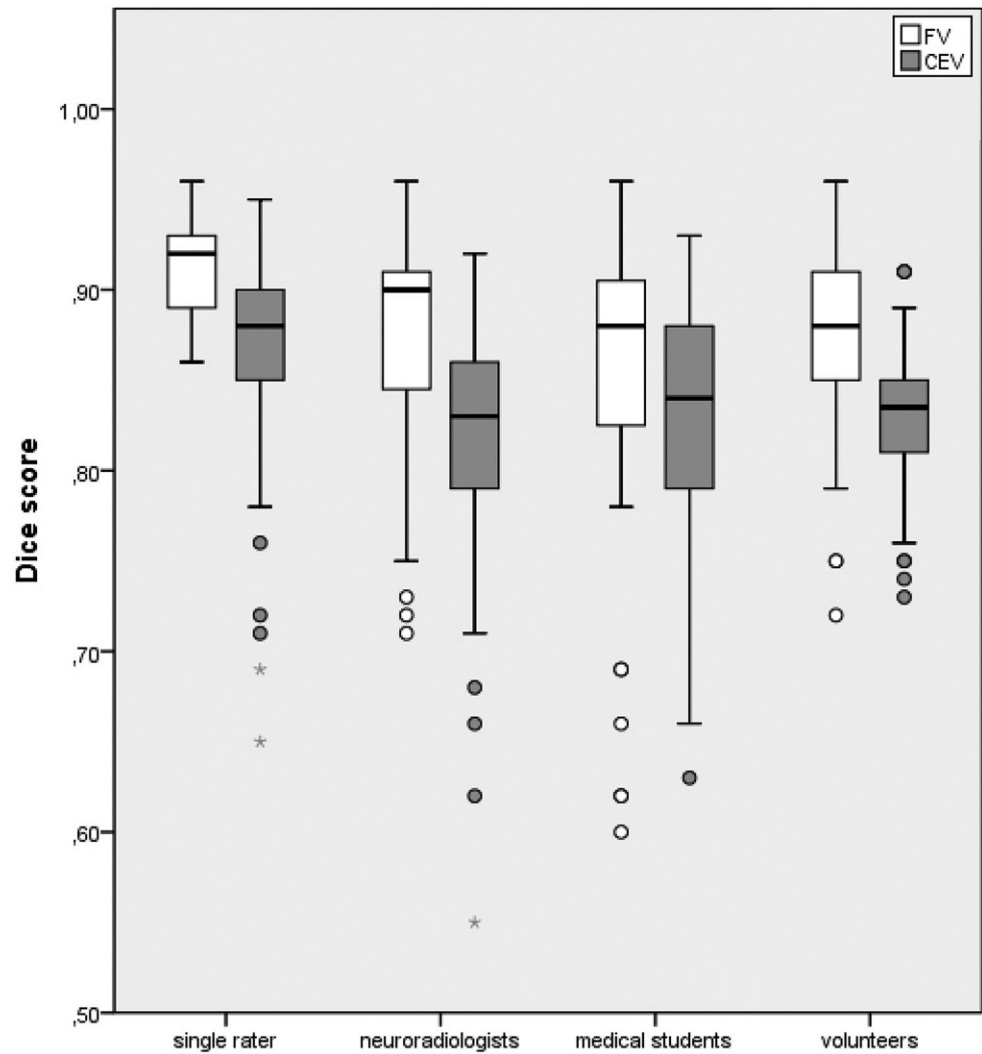
Table 3 Root-mean-square error of relative change of volume between baseline and follow-up magnetic resonance imaging for FLAIR volume and contrast-enhancing volume computed for each group of raters. The number of data sets in each category is displayed (n)

Relative change	FV		CEV	
	RMSE %	n	RMSE %	n
Single rater	5.2	20	12.7	16
Neuroradiologists	7.5	20	16.9	16
Medical students	10.1	20	15.3	16
Volunteers	10.1	20	12.9	16
All raters	8.5	80	14.6	64

evaluation, the relative change in volume between baseline and follow-up MRI showed the lowest precision error for single rater segmentations and for the group of neuroradiologists in FLAIR images.

RANO response criteria replaced the Macdonald criteria to assess changes in postoperative MR imaging of glioblastoma and are currently considered the state of the art [7, 8]. RANO response criteria define progressive disease as an increase of the perpendicular diameter product of more than 25% in postcontrast T1w-sequences [7]. Assuming an isotropic tumor growth, this two-dimensional increase of 25% in diameter product would result in a 39.8% increase of tumor volume (Fig. 4). In our volumetric approach, the lowest RMSE for segmentations of contrast enhancement was 8.2% for single raters. For interpretation of the RMSE, the LSC should be addressed. The LSC defines the minimal change which can be regarded as a significant change. A RMSE of 8.2% translates into a LSC of 22.7%, meaning

Fig. 2 Boxplots for Dice scores as a measure of the overlap between the segmentations for fluid-attenuated inversion recovery (FV) and contrast enhancement (CEV) for the different groups of raters. A Dice score of 1 indicates perfect agreement. Median values are indicated by black horizontal bars

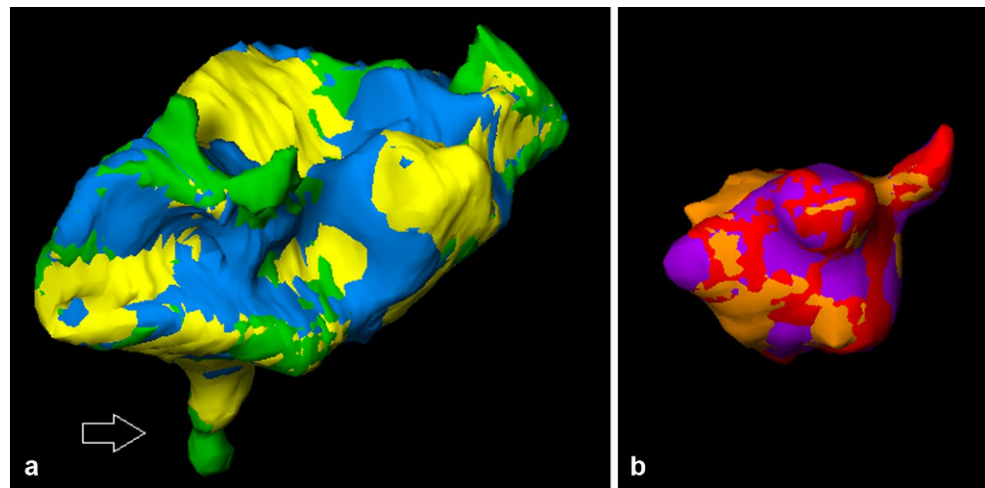


that only changes in volume exceeding this value should be considered significant. Assuming a valid transition of two-dimensional RANO criteria to our three-dimensional volumetric approach, our results show that RANOs' cutoff value for contrast-enhancing lesions (39.8%) can be easily met by single-rater segmentations of CEV. For different raters, the LSC for CEV ranged between 30.0–46.3% indicating that the cutoff value of 39.8% for detection of disease progression cannot always be met. Instead of segmentations of CEV, our results show that FLAIR segmentations can be done more precisely. For single rater segmentations of FLAIR sequences a RMSE of 3.3% (LSC=9.1%) shows that changes in FLAIR volume can be detected more than 4 times as precise as recommended by RANO for CEV (39.8%). Even in the longitudinal evaluation, FLAIR segmentations showed the lowest precision error with a RMSE of only 5.2% (LSC=14.4%). This LSC corresponds to an increase in the perpendicular diameter product of only 9.4%, indicating that 3D-FLAIR segmentations can reliably

detect changes of more than half the size as currently recommended by the 2D-RANO approach for CEV.

Configuration of contrast enhancement shows a high variety among glioblastoma leading to certain limitations of 2D-RANO criteria, that is, in multifocal lesions, enhancing parts of resection cavities or cystic tumors [9]. Our results show that 3D-segmentation can be done reliably and with a precision error lower than RANO criteria supporting the hypothesis that volumetric approaches could be superior to area-based measurements in determining tumor size in those cases [9]. There are growing doubts if contrast enhancement is a valid surrogate marker for detection of progressive disease in glioblastoma, since any disturbance of the blood–brain barrier can lead to gadolinium enhancement (e.g., postoperative scarring, radiation, ischemia) [7]. Considering this and the high-precision error for CEV segmentations, tumor-associated FLAIR changes might be more suited for semi-automated detection of progressive disease. Obviously, most of the above-mentioned disturbances of the blood–brain barrier can also affect FLAIR changes but

Fig. 3 Rendered 3D-volumes of different raters in patient 2 in fluid-attenuated inversion recovery (FLAIR) images **a** and post contrast MPRage images **b**. Wallerian degeneration (*white arrow*) was inconsistently segmented in FLAIR sequence among the three different raters (colors for FLAIR volume: *yellow*=neuroradiologist, *green*=medical student, *blue*=volunteer; colors for contrast-enhancing volume: *orange*=neuroradiologist, *red*=medical student, *purple*=volunteer). Objects are not scaled



with a precision error being more than two times lower for FLAIR segmentations of single raters, tumor growth may be detected earlier.

Among the three different groups of raters, volunteers showed the lowest precision error for single segmentations. This may be due to the user-friendly interface, the 3D-interpolation of the smartbrush, as recently described by another group [11], and due to the fact that volunteers had a permanent stand-by supervisor. As this experienced supervisor helped in doubtful and ambiguous cases, these results may be partly interpreted as a mixture of intra- and inter-rater reliability. Interestingly, when looking at the relative change in volume between baseline and follow-up MRI for each rater, neuroradiologists showed a lower RMSE than for single segmentations, suggesting that neuroradiologists kept their individual ways of segmenting, whereas the other groups did not. However, for an objective and quantitative evaluation of volumetric data, brain tumor segmentations should not be influenced by individual ways of image interpretation.

Previous studies showed that semi-automated segmentation techniques allow a reliable and fast volumetric assessment of CEV in glioblastoma [23, 24]. In these studies, only one rater did the semi-automated segmentations and compared the results to other segmentation techniques. However, our approach aimed to assess the reliability of semi-automated volumetry itself for different groups of users. In the above mentioned studies only CEV was delineated and segmentations of FLAIR hyperintense tumor parts were not addressed. We showed that especially FLAIR changes can be delineated excellent with a semi-automated approach. Porz et al. compared an automated segmentation method to manual segmentations of two experts in terms of Dice scores [25]. Interestingly, this study did include FLAIR images and showed the highest agreement of manual segmentations for experts when FLAIR hyperintense edema was included in the segmentations, emphasizing our

results. Automatic segmentation, which is considered more objective but less accurate than semi-automated volumetry, was inferior [25]. Additionally, Porz et al. also reported the highest variation among segmentations for CEV, again emphasizing our findings.

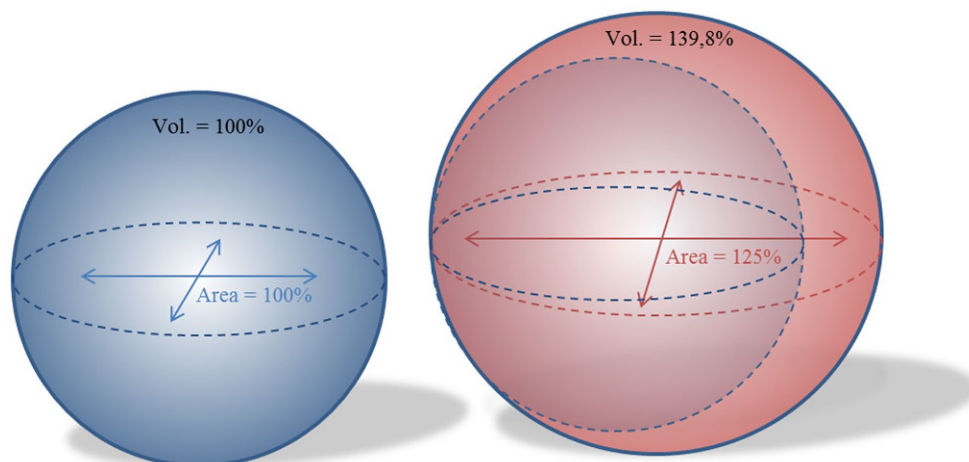
Radbruch et al. [24] showed earlier that about 10% of MRI scans in glioblastoma show exclusive progress of FLAIR changes which was often later followed by a progress of CEV. They proposed a threshold for progressive FLAIR changes of 15%. The results of our study showed that this FLAIR threshold could be reliably detected by our volumetric approach for single raters (LSC=14.4%). Using semi-automated segmentations of FLAIR changes, progressive disease in glioblastoma might be detected earlier in a reliable and quantitative way which could be easily implemented in the clinical routine.

Conclusions

Semi-automated delineation tumor volume with a commercial region-growing algorithm can be done easily and reliably by all groups of raters in patients with glioblastoma, even without neuroradiologic expertise. Segmentations of tumor-associated FLAIR changes were consistently more precise than segmentations of contrast enhancement with the best results in case of a single rater. Precision of experienced neuroradiologists outperformed the nonexperienced groups only in the longitudinal evaluation of FLAIR changes. Here, a single experienced rater could detect progressive FLAIR changes of less than 15% reliably in a quantitative way which could help to detect progressive disease earlier and more precise as currently recommended by RANO for contrast enhancement.

Acknowledgments We thank the group of medical students and volunteers for taking part in this study.

Fig. 4 Volume and area of a ball-shaped tumor configuration. An increase of 25% in tumor area (defined as perpendicular diameter product in the response assessment in neurooncology criteria) leads to an increase in volume from 100 to 139.8% for a ball-shaped tumor configuration



Ethical Standards This study was approved by the local ethics committee at the Klinikum rechts der Isar of the Technical University of Munich, Germany, in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments [13].

Conflict of Interest Brainlab (Feldkirchen, Germany) provided the segmentation software used in this study for research purpose.

References

- Stupp R, Hegi ME, Mason WP, van den Bent MJ, Taphoorn MJ, Janzer RC, Ludwin SK, Allgeier A, Fisher B, Belanger K, Hau P, Brandes AA, Gijtenbeek J, Marosi C, Vecht CJ, Mokhtari K, Wesseling P, Villa S, Eisenhauer E, Gorlia T, Weller M, Lacombe D, Cairncross JG, Mirimanoff RO. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol*. 2009;10:459–66.
- Stupp R, Mason W, van den Bent MJ, Weller M, Fisher BM, Taphoorn MJB, Belanger K, Brandes AA, Marosi C, Bogdahn U, Curschmann J, Janzer RC, Ludwin SK, Gorlia T, Allgeier A, Lacombe D, Cairncross G, Eisenhauer E, Mirimanoff RO. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *Cancer Radiother*. 2005;9(3):196–7.
- Sughrue ME, Sheehan T, Bonney PA, Maurer AJ, Teo C. Aggressive repeat surgery for focally recurrent primary glioblastoma: outcomes and theoretical framework. *Neurosurg Focus*. 2015;38:E11.
- Chinot OL, Wick W, Mason W, Henriksson R, Saran F, Nishikawa R, Carpentier AF, Hoang-Xuan K, Kavan P, Cernea D, Brandes AA, Hilton M, Abrey L, Cloughesy T. Bevacizumab plus radiotherapy-temozolomide for newly diagnosed glioblastoma. *N Engl J Med*. 2014;370:709–22.
- Woernle CM, Péus D, Hofer S, Rushing EJ, Held U, Bozinov O, Krayenbühl N, Weller M, Regli L. Efficacy of surgery and further treatment of progressive glioblastoma. *World Neurosurg*. 2015;84:301–7.
- Franceschi E, Bartolotti M, Tosoni A, Bartolini S, Sturiale C, Fioravanti A, Pozzati E, Galzio R, Talacchi A, Volpin L, Morandi L, Danieli D, Ermani M, Brandes AA. The effect of re-operation on survival in patients with recurrent glioblastoma. *Anticancer Res*. 2015;35:1743–8.
- Wen PY, Macdonald DR, Reardon DA, Cloughesy TF, Sorensen AG, Galanis E, DeGroot J, Wick W, Gilbert MR, Lassman AB, Tsien C, Mikkelsen T, Wong ET, Chamberlain MC, Stupp R, Lamborn KR, Vogelbaum MA, van den Bent MJ, Chang SM. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J Clin Oncol*. 2010;28(11):1963–72.
- Macdonald DR, Cascino TL, Schold SC, Cairncross JG. Response criteria for phase II studies of supratentorial malignant glioma. *J Clin Oncol*. 1990;8:1277–80.
- Sorensen AG, Batchelor TT, Wen PY, Zhang W-T, Jain RK. Response criteria for glioma. *Nat Clin Pract Oncol*. 2008;5:634–44.
- Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, 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, Iftekharuddin 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. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34(10):1993–2024.
- Rana M, Modrow D, Keuchel J, Chui C, Rana M, Wagner M, Gellrich N-C. Development and evaluation of an automatic tumor segmentation tool: a comparison between automatic, semi-automatic and manual segmentation of mandibular odontogenic cysts and tumors. *J Craniomaxillofac Surg*. 2015;43:355–9.
- Gordillo N, Montseny E, Sobrevilla P. State of the art survey on MRI brain tumor segmentation. *Magn Reson Imaging*. 2013;31:1426–38.
- WMA General Assembly. World Medical Association Declaration of Helsinki. 1964.
- Hojjatolamlami SA, Kittler J. Region growing: a new approach. *IEEE Trans Image Process*. 1998;7:1079–84.
- Hojjatolamlami SA, Kruggel F. Segmentation of large brain lesions. *IEEE Trans Med Imaging*. 2001;20:666–9.
- McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996;1:30–46.
- Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol*. 2012;8:23–34.

18. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess.* 1994;6:284–90.
19. Glüer CC, Blake G, Lu Y, Blunt BA, Jergas M, Genant HK. Accurate assessment of precision errors: how to measure the reproducibility of bone densitometry techniques. *Osteoporos Int.* 1995;5:262–70.
20. Shepherd JA, Lu Y. A generalized least significant change for individuals measured on different DXA systems. *J Clin Densitom.* 2007;10:249–58.
21. Dice LR. Measures of the amount of ecologic association between species. *Ecology.* 1945;26:297–302.
22. Galluzzi L, Aaronson SA, Abrams J, Alnemri ES, Andrews DW, Baehrecke EH, Bazan NG, Blagosklonny MV, Blomgren K, Bornier C, Bredesen DE, Brenner C, Castedo M, Cidlowski JA, Ciechanover A, Cohen GM, De Laurenzi V, De Maria R, Deshmukh M, Dynlacht BD, El-Deiry WS, Flavell RA, Fulda S, Garrido C, Golstein P, Gougeon M-L, Green DR, Gronemeyer H, Hajnóczky G, Hardwick JM, Hengartner MO, Ichijo H, Jäättelä M, Kepp O, Kimchi A, Klionsky DJ, Knight RA, Kornbluth S, Kumar S, Levine B, Lipton SA, Lugli E, Madeo F, Malomi W, Marine J-CW, Martin SJ, Medema JP, Mehlen P, Melino G, Moll UM, Morselli E, Nagata S, Nicholson DW, Nicotera P, Nuñez G, Oren M, Penninger J, Pervaiz S, Peter ME, Piacentini M, Prehn JHM, Puthalakath H, Rabinovich GA, Rizzuto R, Rodrigues CMP, Rubinsztein DC, Rudel T, Scorrano L, Simon H-U, Steller H, Tschopp J, Tsujimoto Y, Vandenabeele P, Vitale I, Vousden KH, Youle RJ, Yuan J, Zhivotovsky B, Kroemer G. Guidelines for the use and interpretation of assays for monitoring cell death in higher eukaryotes. *Cell Death Differ.* 2009;16:1093–107.
23. Chow DS, Qi J, Guo X, Miloushev VZ, Iwamoto FM, Bruce JN, Lassman AB, Schwartz LH, Lignelli A, Zhao B, Filippi CG. Semi-automated volumetric measurement on postcontrast MR imaging for analysis of recurrent and residual disease in glioblastoma multiforme. *AJNR Am J Neuroradiol.* 2014;35:498–503.
24. Radbruch A, Lutz K, Wiestler B, Bäumer P, Heiland S, Wick W, Bendszus M. Relevance of T2 signal changes in the assessment of progression of glioblastoma according to the Response Assessment in Neurooncology criteria. *Neuro Oncol.* 2012;14:222–9.
25. Porz N, Bauer S, Pica A, Schucht P, Beck J, Verma RK, Slotboom J, Reyes M, Wiest R. Multi-modal glioblastoma segmentation: man versus machine. *PLoS One.* 2014;9(5):e96873.