Comparison of 2D (RANO) and volumetric methods for assessment of recurrent glioblastoma treated with bevacizumab – a report from the BELOB trial.

Renske Gahrmann
Martin van den Bent
Bronno van der Holt
René Vernhout
Walter Taal
Maaike Vos
Jan Cees de Groot
Laurens Beerepoot
Jan Buter
Zwenneke Flach
Monique Hanse
Bas Jasperse
Marion Smits

NEURO ONCOL 2017;19(6)853-861.

ABSTRACT

Background. The current method for assessing progressive disease (PD) in glioblastoma is according to the Response Assessment in Neuro-Oncology (RANO) criteria. Bevacizumab-treated patients may show pseudo-response on post-contrast T1w, and a more infiltrative non-enhancing growth pattern on T2w/FLAIR-images. We investigated whether the RANO criteria remain the method of choice for assessing bevacizumab-treated recurrent glioblastoma when compared to various volumetric methods.

Methods. Patients with assessable MRI-data from the BELOB-trial (n=148) were included. Patients were treated with bevacizumab, lomustine, or both. At first and second radiological follow-up (6 and 12 weeks) PD was determined using the 2D RANO criteria and various volumetric methods based on enhancing tumor only and enhancing plus non-enhancing tumor. Differences in overall survival (OS) between PD and non-PD patients were assessed with the log-rank test and a cox-model. Hazard ratios (HR) and their 95% confidence-intervals were determined.

Results. For all patients together, all methods (except subtraction of non-enhancing from enhancing volume at first follow-up) showed significant differences in OS between PD and non-PD patients (p<.001). The largest risk-increase for death in case of PD at both first and second follow-up was found with the RANO criteria: HR = 2.81 (95% CI, 1.92-4.10) and HR = 2.80 (95% CI, 1.75-4.49) respectively. In the bevacizumab-treated patients, all methods assessed showed significant differences in OS between PD and non-PD patients. There were no significant differences between methods.

Conclusions. In the first 12 weeks, volumetric methods did not provide significant improvement over the RANO criteria as a post-treatment prognostic marker.

Keywords

Recurrent glioblastoma, bevacizumab, RANO, volumetry

Importance of study

Currently, the 2D RANO criteria are the method of choice for assessing progressive disease in glioblastoma. However, those treated with bevacizumab may show pseudo-response on post contrast T1w-images and non-enhancing tumor growth on T2w/FLAIR-images upon progression that may not optimally be captured with the 2D RANO criteria. We compared the 2D RANO criteria with various volumetric methods based on enhancement, subtraction and T2w/FLAIR abnormalities in this distinct patient group. The risk increase for death was determined based on the presence of progressive disease upon follow-up. The largest hazard ratios (HR) at first and second follow-up were found with the 2D RANO criteria: HR = 2.81 (95% CI, 1.92-4.10) and HR = 2.80 (95% CI, 1.75-4.49) respectively. We found no clear differences in prediction of overall survival between the 2D RANO and volumetric methods. The routine use of volumetric methods in clinical trials in recurrent glioblastoma is therefore not warranted.

INTRODUCTION

Glioblastoma is the most common malignant primary brain tumor, comprising 15% of all primary brain and central nervous system tumors and almost half of all primary malignant brain tumors. The 5-year survival rate is only about 5% despite treatment of newly diagnosed patients with surgery, radiotherapy and concomitant and adjuvant temozolomide[1,2].

In 2009, the United States Food and Drug Administration (FDA) approved bevacizumab (Avastin®, Genentech, San Francisco) for second-line treatment in patients with recurrent glioblastoma based on the observed response rates from phase 2 trials with bevacizumab and irinotecan[3,4]. Bevacizumab is a humanized monoclonal antibody directed against Vascular Endothelial Growth Factor (VEGF), which inhibits angiogenesis and normalizes abnormally permeable tumor blood vessels[5,6]. This may also lead to a decrease (or complete disappearance) of enhancing tumor on post-contrast T1w-images without actual changes in tumor size. Due to this phenomenon of pseudo-responses the conventional MacDonald assessment criteria[6], that rely primarily on the assessment of enhancing lesions, were no longer sufficient and have been replaced by the Response Assessment in Neuro-Oncology or RANO criteria[7]. These have become the standard in clinical neuro-oncology and include the assessment of T2w/FLAIR (non-enhancing) abnormalities in addition to enhancing lesions. Progressive disease (PD) is defined as a ≥25% increase in the sum of the products of perpendicular diameters of enhancing tumor, a significant increase in non-enhancing tumor, the appearance of new lesions, or clinical deterioration. It has been shown that adding non-enhancing abnormalities based on T2w/FLAIR imaging to the response assessment in recurrent glioblastoma treated with bevacizumab may lead to earlier detection of PD[8].

Despite the inclusion of T2w/FLAIR assessment into the RANO criteria, the main focus of response evaluation remains on enhancing tumor. The observed initial decrease of enhancing tumor and increase in T2w/FLAIR abnormalities at progression[9,10] in patients treated with bevacizumab suggests that more advanced methods of assessment, such as volumetry, might improve prediction of overall survival (OS). This is particularly relevant to the assessment of T2w/FLAIR abnormalities, since it is defined as a qualitative, and not a quantitative change in volume. Volumetric assessments are likely to increase precision of measurements of enhancement in irregularly shaped tumors, such as glioblastoma, and T2w/FLAIR abnormalities are likely to be more reliably assessed quantitatively with volumetry.

The occurrence of pseudo-response, i.e. the decrease of tumor enhancement due to vascular normalization rather than a true anti-tumor effect in glioblastoma patients treated with VEGF inhibitors, has now been well documented and questions arise whether further improvement of the RANO criteria is needed. Several studies have investigated the evaluation of tumor response using volumetric measures from both enhancing and non-enhancing recurrent glioblastoma treated with bevacizumab[9,11,12]. Boxerman et al[13] also directly compared 2D with volumetric methods in this specific patient group but did not assess the full RANO criteria. Ultimately, to be a reliable surrogate endpoint in phase I and II studies, response should reflect OS, the gold standard in oncology trials. Our aim was to determine whether in bevacizumab-treated recurrent glioblastoma volumetric methods are superior to the 2D RANO criteria in determining PD in association with OS.

For this purpose we used data from the BELOB-trial, which is the first randomized and properly controlled phase 2 trial in recurrent glioblastoma, comparing single-agent bevacizumab or lomustine with bevacizumab plus lomustine[14].

METHODS

Patients

A total of 148 eligible patients with first recurrence of glioblastoma were included in the BELOB-trial, a randomized controlled phase 2 trial in which patients received bevacizumab (Avastin®) (n=50), lomustine (n=46) or both (n=52). Patients were recruited between December 2009 and October 2011 from 5 University hospitals and 9 community hospitals in the Netherlands. Patients were at least 18 years of age and had received no prior treatment with anti-VEGF or nitrosoureas. All patients provided written informed consent according to national regulations. A more detailed description of the study and its findings can be found in Taal et al. 2014[14].

The primary endpoint of the BELOB trial was 9-month OS. Additional outcome measures were median progression-free survival (PFS), PFS at 6 and 12 months, median OS, OS at 6 and 12 months, and proportion of patients with objective response. In the current radiological analysis, OS was used as the endpoint.

Scanning procedure

Patients underwent standardized MRI scanning at baseline and follow-up with approximately 6-week intervals, i.e. a follow-up scan was made after every treatment cycle. The standardized MRI protocol can be found in the ***Supplementary Files**.* Imaging was performed at 1.5T and 3.0T scanners. During the study, MRI quality assessment was performed. Unfortunately, some patients still had incomplete imaging datasets as specified in the results section.

PD assessment

The presence of PD was determined with the 2D RANO method (method 1) and 4 volumetric methods: total contrast-enhancing (CE) volume measured on T1w post-contrast images only (method 2), total CE volume measured on subtraction (post-contrast minus pre-contrast) images only (method 3), total CE volume (as measured on T1w post-contrast images) complemented with non-enhancing volume measured on FLAIR images (method 4), and total CE volume as measured on subtraction images complemented with non-enhancing volume (method 5).

RANO assessment to establish PD was performed centrally (pre- and post-contrast T1w- and T2w/FLAIR-images) by 2 independent reviewers (MB, MS). In case of disagreement, PD was decided by an adjudicator (BJ).

Volumes of enhancing areas and non-enhancing (FLAIR) abnormalities were measured on 3D T1w post-contrast and 3D FLAIR images, respectively, by a single rater (RG) using a semi-automated technique in BrainLab I-Plan 4.0 Cranial (BrainLab, Feldkirchen, Germany) software. This technique involves the manual placement of 'inclusion' and 'exclusion' points (***Figure 1***), after which an algorithm is used to render a volume of interest (VOI). All VOIs were visually checked in three directions and adjusted if needed.

In each scan, in case of more lesions all lesions were measured separately and summed for the current analysis to obtain a single volumetric measure of both the enhancing tumor and one of FLAIR abnormalities. Blood vessels, dura and necrotic areas were excluded. Enhancing areas were included in the FLAIR VOIs, because these areas are also hyperintense on FLAIR. The commonly T2w-hyperintese cortical

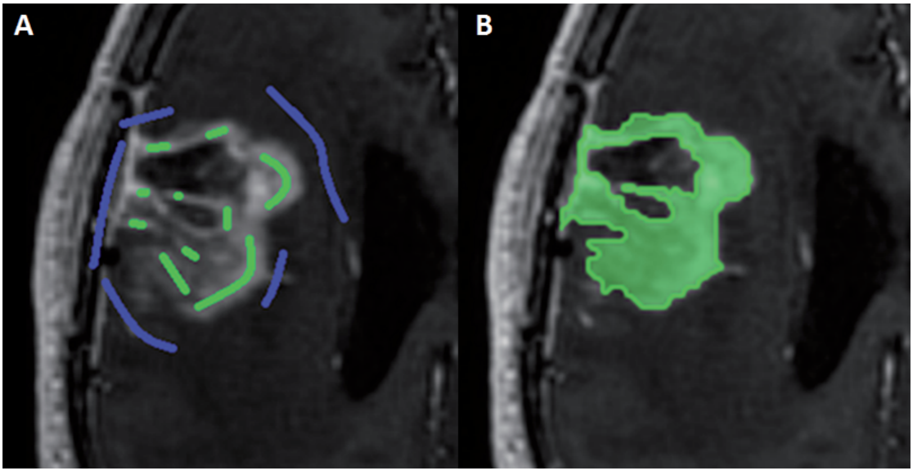

Figure 1. Example of segmentation in BrainLab I-Plan 4.0 Cranial using inclusion (green) and exclusion (blue) lines/points in the axial plane (**A**) and the resulting segmentation (**B**).

ribbon was excluded. Lesions clearly of vascular origin and periventricular apical capping (i.e. not continuous with FLAIR lesions and remaining unchanged in time) were excluded, as well as the septum pellucidem. FLAIR abnormalities when present in both left- and right hemispheres were measured separately when possible.

Subtraction images were created with FSL-FLIRT (FMRIB Software Library, Oxford, England) with custom scripts in AFNI (National Institute of Mental Health, Bethesda) created by Ellingson et al[15]. VOIs of resulting enhancing tumor areas were drawn manually in MRIcron (Chris Rorden, www.mrico.com, version 6.6.2013) by a single rater (RG). Necrotic areas, blood vessels and dura were excluded.

PD was defined according to the 2D RANO criteria (as described earlier) for method 1. For the volumetric methods (methods 2-5), PD was defined as ≥40% increase in enhancing/subtraction volume, which was the most commonly used threshold in previous literature[16-18], ≥25% increase of FLAIR volume[19], or the appearance of new lesions, whether enhancing or non-enhancing. The 40% threshold for volumetric assessment is based on an extrapolation from the ≥25% increase in the sum of perpendicular diameters of a 2D lesion to a 3D sphere-shape and the assumption that all increase in size is equal in every direction. Setting the threshold this high will ensure that only patients showing clear PD are categorized as such.

Statistical analysis

Patients were classified as to PD or non-PD according to each of the 5 methods at both first and second follow-up scans. For each of the follow-up moments separately, OS was determined from the date of the scan to death from any cause. Patients still alive at last contact were censored. Kaplan-Meier survival curves were drawn from each follow-up time points for all patients from all treatments groups, as well as for patients treated either with bevacizumab (with/without lomustine) or lomustine separately. A log-rank test was used to determine the difference in OS between the PD and non-PD patients as established by each of the methods. To determine whether there were significant differences between methods in predicting OS, hazard ratios (HR) and their corresponding 95% confidence intervals (CI) were determined by means of a cox regression analysis per method, using each method as a single covariate. The HR was determined for all patients together and for the different treatment groups per method. The overlap of the corresponding 95% CIs was determined and conclusions were drawn based on the extent of overlap. Statistical analysis was performed using SPSS Statistics for Windows (IBM Corp. 2012, Version21.0, Armonk, NY).

A post-hoc power analysis (G*Power[20]) was performed for all patients together and for the two different treatment groups per method at both first and second follow-up to evaluate the validity of the results.

RESULTS

Patients

148 patients were included in the BELOB-trial. At first follow-up 10 patients were excluded from the analysis due to a lack of follow-up and missing data required for 2D RANO assessment. Some further patients had missing 3D T1w pre- and/or post-contrast or 3D FLAIR images, resulting in the following numbers of patients available for analysis at first follow-up: n=138 for method 1, n=121 for method 2, n=86 for method 3, n=109 for method 4, and n=89 for method 5. At second follow-up, at which time patients with prior PD or death had dropped out, the following numbers of patients with adequate imaging were available for analysis: n=83 for method 1, n=78 for method 2, n=71 for method 3, n=71 for method 4, and n=67 for method 5.

Assessment of all patients from all treatment groups

Significant differences in OS (***Figures 2 and 3***) were found between all patients with PD and non-PD as determined with methods 1, 2, 4 and 5 at first follow-up, and with all 5 methods at second follow-up (***Table 1***). At first follow-up the highest risk increase for death was found for PD as determined with method 1, i.e. the RANO criteria, with an HR of 2.81 (95% CI, 1.92 – 4.10). In comparing the methods, the extent of overlap in CIs (>50%) indicates that there are no significant differences in HR. At second follow-up the highest risk increase was again found for PD determined with the RANO criteria (method 1), with an HR of 2.80 (95% CI, 1.75 – 4.49). Again, no significant differences were observed between methods as determined by the extent of overlap in CIs. For a graphic depiction of the direct comparison between the HRs and 95% CI we refer to the ***Supplementary Files (figure S1)***.

Table 1. Hazard ratios of each of the methods for all patients at first and second follow-up (FU1 and FU2).

Method	HR (95% CI) FU1	p-value	HR (95% CI) FU2	p-value
1. 2D RANO	2.81 (1.92 – 4.10)	< .001	2.80 (1.75 – 4.49)	< .001
2. Contrast enhancing volume	1.99 (1.29 – 3.05)	.002	2.17 (1.27 – 3.71)	.004
3. Subtraction volume	1.80 (0.99 – 3.26)	.054	2.16 (1.27 – 3.68)	.005
4. Contrast enhancing + FLAIR volume	2.45 (1.60 – 3.74)	<.001	2.57 (1.53 – 4.31)	<.001
5. Subtraction + FLAIR volume	2.66 (1.61 – 4.38)	<.001	2.48 (1.47 – 4.18)	.001

Assessment per treatment group

The post-hoc power analysis showed that power in the lomustine-only patient group was well below 80%, precluding meaningful analysis. Within the bevacizumab-treated group, power was also insufficient for a meaningful analysis at first follow-up for method 3 and at second follow-up for methods 2, 3 and 5. Differences in

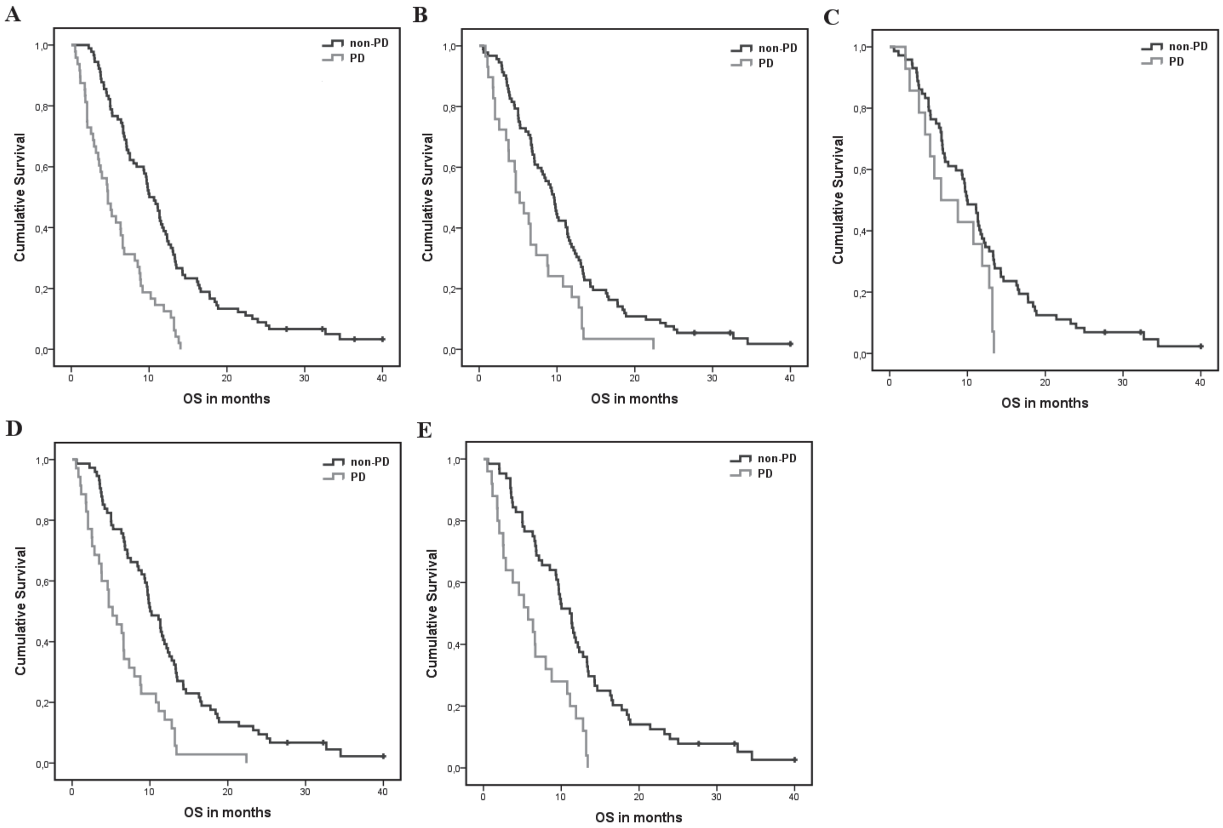

Figure 2. Kaplan-Meier curves of all progressive (PD) versus non-progressive (non-PD) patients for each of the methods at first follow-up. (**A**) 2D RANO, (**B**) Contrast enhancing volume, (**C**) Subtraction volume, (**D**) Contrast enhancing + FLAIR volume, and (**E**) Subtraction + FLAIR volume. OS = overall survival.

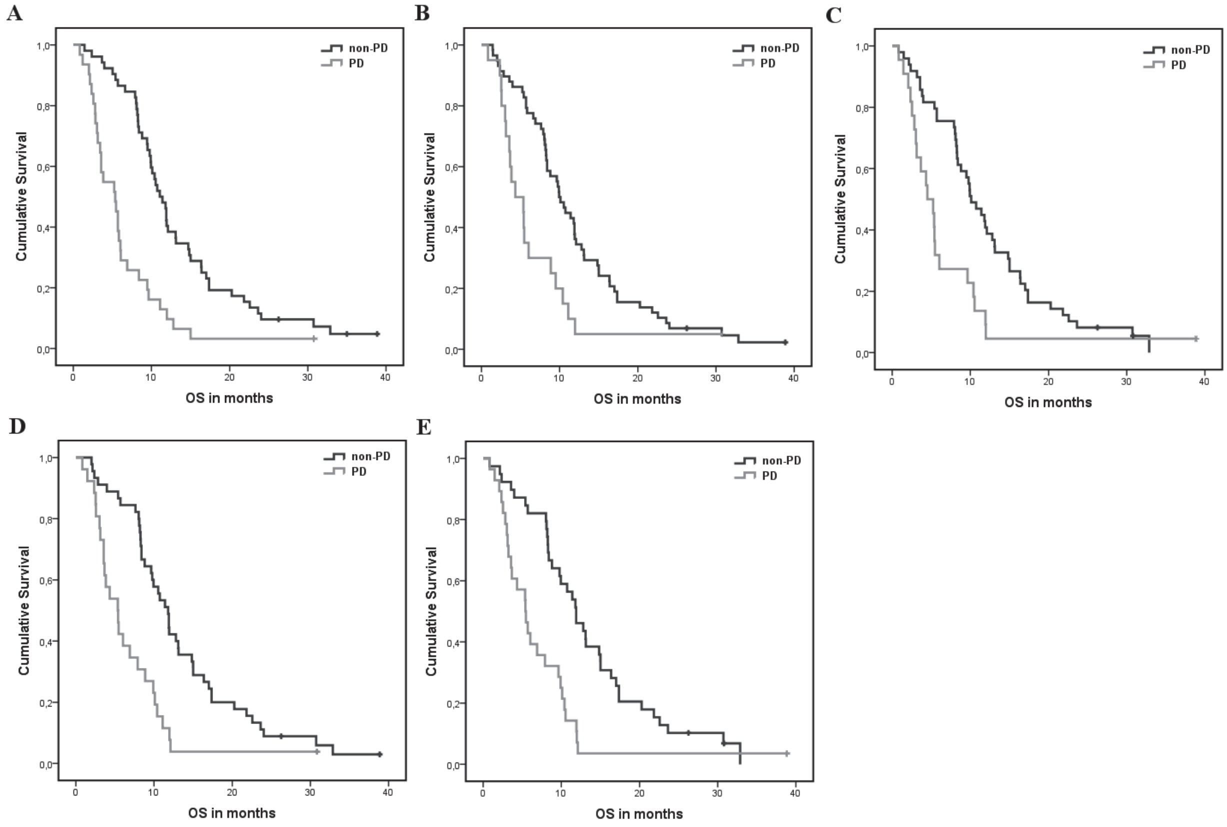

Figure 3. Kaplan-Meier curves of all progressive (PD) versus non-progressive (non-PD) patients for each of the methods at second follow-up. (**A**) 2D RANO, (**B**) Contrast enhancing volume, (**C**) Subtraction volume, (**D**) Contrast enhancing + FLAIR volume, and (**E**) Subtraction + FLAIR volume. OS = overall survival.

power between methods can be attributed to a difference in the number of patients available for analysis per method as well as differences in the number of patients classified as PD and non-PD.

In the bevacizumab-treated group significant differences in OS (***Figure 4***) were found between patients with PD and non-PD as determined with methods 1, 2, 4 and 5 at first follow-up, and with methods 1 and 4 at second follow-up (***Table 2***). The highest risk increase at first follow-up was found for PD as determined with method 2 (i.e. CE volume), with an HR of 7.21 (95% CI, 3.20 – 16.22). This was however not significantly different from other methods. At second follow-up only the results from methods 1 (i.e. RANO criteria) and 4 (i.e. combined analysis of CE and FLAIR volume) were deemed reliable (i.e. power >80%), with HRs of 2.44 (95% CI, 1.46 – 4.08) and 2.08 (95% CI, 1.16 – 3.73) respectively. These were not significantly different from each other. For a graphic depiction of the direct comparison between the HRs and 95% CI we refer to the ***Supplementary Files (Figure S1)**.*

Concordance between methods

Concordance rates between method 1 (2D RANO) and the volumetric methods (2-5) at first follow-up in all patients together and in the bevacizumab-treated group varied between 81.3% and 87.4%. For a more detailed report and individual examples we refer to the ***Supplementary Files (table S1 and Figures S2 and S3)**.* There are several underlying reasons for discrepancies between methods in individual cases. The main reasons were: a) PD based on the increase of FLAIR abnormalities, which was not picked up using methods 2 and 3 (enhancing/subtraction volume), b) PD determined in 2D RANO based on mixed response, not picked up with volumetric methods because change in total (non-) enhancing volume was analyzed, c) lesions that did not reach the threshold for PD in volumetry (≥40% increase), while having reached the threshold for PD in 2D RANO (≥25% increase), and d) significant increase in FLAIR volume in volumetry (≥25% increase), not scored as significant in 2D RANO (no threshold).

Table 2. Hazard ratios of methods with sufficient power (>80%) for bevacizumab-treated patients only at first and second follow-up (FU1 and FU2).

Method	HR (95% CI) FU1	p-value	HR (95% CI) FU2	p-value
1. 2D RANO	5.53 (3.12 – 9.80)	<.001	2.44 (1.46 – 4.08)	.001
2. Contrast enhancing volume	7.21 (3.20 – 16.22)	<.001		
4. Contrast enhancing + FLAIR volume	5.94 (3.06 – 11.53)	<.001	2.35 (1.32 – 4.19)	.004
5. Subtraction + FLAIR volume	5.63 (2.66 – 11.93)	<.001		

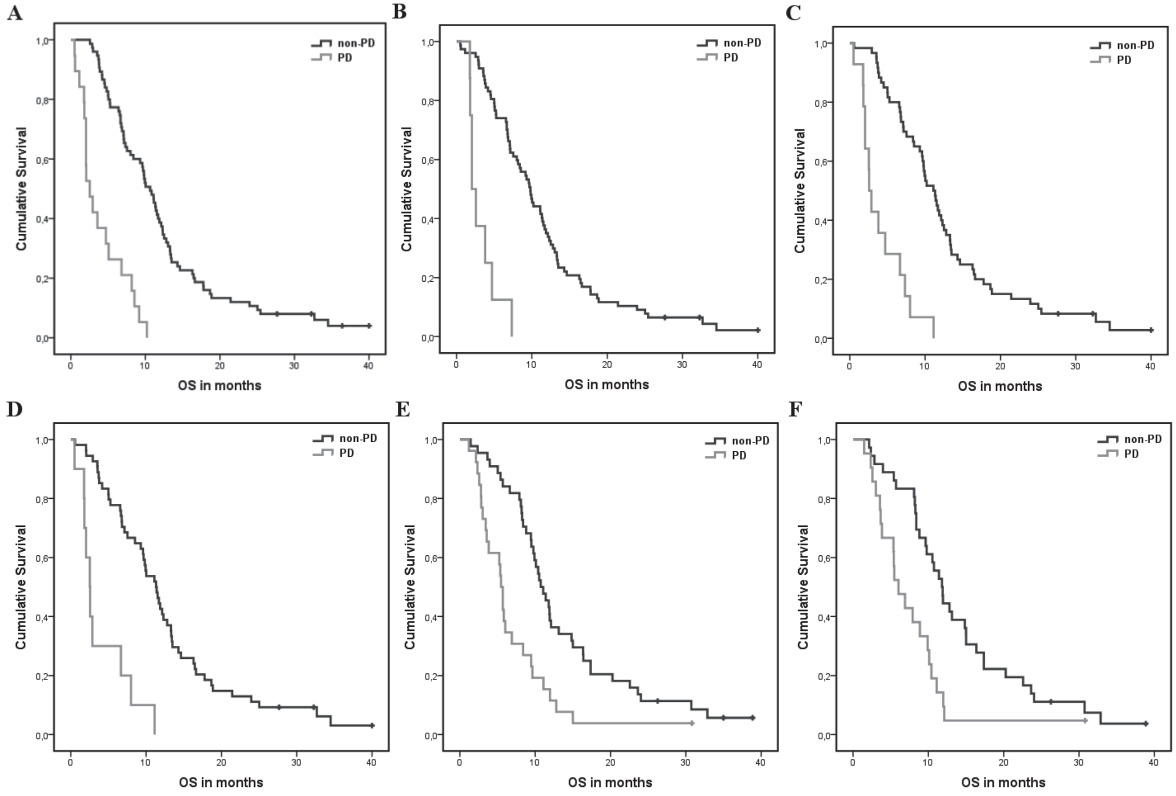

Figure 4. Kaplan-Meier curves of progressive (PD) versus non-progressive (non-PD) bevacizumab-treated patients only for methods with sufficient power (>80%) at first follow-up. (**A**) 2D RANO, (**B**) Contrast enhancing volume, (**C**) Contrast enhancing + FLAIR volume, and (**D**) Subtraction + FLAIR volume. And at second follow-up **E**) 2D RANO, and **F**) Contrast enhancing + FLAIR volume. OS = overall survival.

DISCUSSION

We investigated whether PD, as determined with various MRI-based methods, was associated with OS, the gold standard in oncology studies, in order to identify the optimal method for radiological treatment response assessment, in particular in the context of anti-angiogenic treatment of recurrent glioblastoma. Currently, the 2D RANO criteria are the established method of choice for determining PD in studies on glioblastoma. We found that in patients with recurrent glioblastoma treated with bevacizumab determining PD with volumetric methods, with or without subtraction, did not provide significant improvement as a post-treatment prognostic marker at 6 and 12 weeks follow-up. The volumetric methods assessed were based on enhancing lesions only (methods 2 and 3) or on the combined analyses of enhancing and FLAIR lesions (methods 4 and 5). The volumetric methods were also not significantly different from each other. In the past, several studies on the added value of volumetry in recurrent glioblastoma in determining OS have been performed[11, 21,22], most notably by Boxerman et al.[13], in which a direct comparison is made between 2D and volumetric methods. We analyzed all patients together and the bevacizumab-treated patients separately. Both these analyses failed to show significant differences between methods. The lomustine-only group comprised one-third of all patients, but because this group was too small to draw reliable conclusions from separate analysis, its influence on the analysis of all patients together is unclear. This is however of less importance, since our main goal was to look at the added value of volumetry (and subtraction) in bevacizumab-treated patients. A larger dataset would be required to draw definitive conclusions about lomustine treated patients.

Comparisons between 2D (and other linear) and volumetric measures have been investigated extensively in a whole range of different tumors[23-25], including glioma, with contradictory results. Dempsey et al. compared 1D, 2D and volumetric measurements of enhancing volume of high-grade gliomas and found only volumetry to be predictive for OS[26]. Shah et al. on the other hand compared linear measures with volumetry and found these comparable when correlated with median progression-free survival (mPFS), but found linear methods superior when correlated with OS[24]. Galanis et al. compared 1D, 2D, area, and volume-measures of enhancing (and non-enhancing) newly diagnosed gliomas and found with a time-dependent Cox model that PD measured by all four methods was predictive of OS in enhancing tumors[27]. Boxerman et al. compared 2D and volumetric measurements of enhancement and the added value of non-enhancing volume in recurrent glioblastoma treated with bevacizumab plus a chemotherapeutic agent and found that 2D and volumetric measurements were equally good at predicting OS based on PD at 8 and 16 weeks[13]. Our results are in accordance with these findings. The main differences between their and our studies are that we compared the current golden standard, i.e. the 2D RANO criteria, with various volumetric methods including subtraction, and that our study had three different treatment arms. It must be noted that the 2D (RANO) assessment of our study was performed by raters with extensive clinical experience, i.e. a neuro-oncologist and two neuro-radiologists. This may have had a positive influence on the performance of the 2D method in predicting OS. Semi-automated volumetric assessment was performed by a single rater. Previous studies have however shown inter- and intra-rater variability to be lower in computer-assisted (volumetric) methods when compared to (non-automated) diameter methods[17,28].

Aside from an early pseudo-response, bevacizumab-treated glioblastoma patients have also shown an increase of T2w/FLAIR abnormalities at progression[9,29]. This non-enhancing tumor progression is likely due to the cooption of blood vessels by tumor cells, along which the tumor infiltrates to more distant areas[30]. The 2D RANO criteria adopted a subjective approach for assessing T2w/FLAIR abnormalities. Volumetric methods assessing T2w/FLAIR are potentially better suited to quantify these abnormalities, especially in view of the often complex shape of glioblastoma. We assessed the added value of volumetric FLAIR measures to the enhancement (and subtraction) measures and found no differences with the 2D RANO criteria. We did not assess FLAIR volume separately, because previous research suggests that volumetric FLAIR measures alone are insufficient for predicting OS. Ellingson et al. for instance found that FLAIR volume (initial, residual, and change) was not predictive of either PFS or OS[11]. Similarly, Schaub et al. determined that PD on FLAIR measures alone was not predictive of OS[22]. Huang et al. found a correlation between post-treatment T2w/FLAIR volume and OS, but this correlation disappeared when corrected for enhancing volume[20].

In addition to measuring enhancing tumor volume on T1w post-contrast and non-enhancing tumor volume on FLAIR images, we measured enhancing tumor volume on subtraction images. Patients treated with bevacizumab may develop T1 hyperintense lesions in a previously enhancing area with a reported prevalence between 20 and 80%[15,31-33]. These lesions likely represent calcifications[31]. The presence of T1 hyperintensities might hinder assessment of residual enhancement. Also, enhancement may become more vague, interfering with measurement. Ellingson et al. found that subtraction images improved visualization, tumor volume quantification and prediction of OS in patients with recurrent glioblastoma treated with bevacizumab in comparison to T1w post-contrast images[15]. We found that prediction of OS by assessing PD on subtraction (method 3) and subtraction plus FLAIR volumes (method 5) was similar to that with methods not using subtraction techniques. It must be noted that the number of scans suitable for this analysis was limited due to the multicenter study designs with hospitals using scanners from different vendors, with varying data quality and protocol violations. This was especially problematic for subtraction methods, because matching 3D pre- and post-contrast T1w images were commonly not available. Unfortunately, this meant that not all subtraction-based methods could be reliably assessed in the bevacizumab-treated patient group. These technical shortcomings unfortunately resemble the real life conditions under which clinical trials are being conducted. The recently published standardized MRI brain tumor imaging protocol should overcome some of these technical difficulties[34].

There are several important considerations when designing novel ways to assess outcome in phase I and phase II trials. First and foremost, they must correlate with the final endpoint used in phase III trials, they must be reliable, and they must be feasible in large multicenter trials. A new method must bring either a clear benefit in being more accurate, or it must bring an advantage in being simpler but with a similar precision. Volumetric assessment is more precise in irregularly shaped tumors, but labor intensive, time-consuming and more complex, also in the case of semi-automated techniques. It is also not readily available in many institutions. Our data suggest volumetric assessment is not better as compared to classical 2D assessment. Therefore, there seems to be no rationale at this point to start using volumetric assessment in trials on recurrent glioblastoma.

There are some limitations to consider in this retrospective study. As previously mentioned, rater-experience may have positively influenced the 2D RANO assessment. The interrater variability for assessment according to the 2D RANO from this trial will be reported separately. It must also be noted that the volumetry measurements were performed by a single observer. Secondly, our data was analyzed at first and second follow-up and presence of PD was determined at these time points only. Later follow-up data was not included in this study, because there were not enough patients that had not yet shown PD left to obtain reliable results.

Our future efforts are aimed at exploring more advanced imaging techniques, such as diffusion imaging-derived apparent diffusion coefficient (ADC) as well as relative cerebral blood volume (rCBV) values from enhancing and T2w/FLAIR lesions in bevacizumab-treated patients. Additionally, in future exploratory analyses different volumetric thresholds for determining PD and a combination of enhancing volume plus qualitative FLAIR assessment will be investigated to provide more insight in the added value of volumetry in this patient group and that of FLAIR in particular.

We conclude that the current, widely used and easily applicable, 2D RANO criteria remain valid for response evaluation in patients with recurrent glioblastoma treated with bevacizumab. Volumetric and subtraction evaluation methods failed to yield a superior correlation with OS in the first 12 weeks in this patient group. Our data therefore do not support their routine use in clinical trials.

REFERENCES

1. Ostrom QTG, Gittleman H, Fulop J, et al. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2008-2012. Neuro Oncol. 2015;17(suppl 4):iv1-iv62.
2. Stupp RH, Hegi ME, Mason WP, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomized phase III study: 5-year analysis of the EORTC-NCIC trial. Lancet Oncol. 2009;10(5):459-466.
3. Friedman HS, Prados MD, Wen PY, et al. Bevacizumab alone and in combination with irinotecan in recurrent glioblastoma. J Clin Oncol. 2009;27(28):4733-4740.
4. Kreisl TN, Kim L, Moore K, et al. Phase II trial of single-agent bevacizumab followed by bevacizumab plus irinotecan at tumor progression in recurrent glioblastoma. J Clin Oncol. 2009;27(5):740-745.
5. Jain RK. Normalization of tumor vasculature: an emerging concept in antiangiogenic therapy. Science. 2005;307(5706):58-62.
6. Macdonald DR, Cascino TL, Schold SC Jr, et al. Response criteria for phase II studies of supratentorial malignant glioma. J Clin Oncol. 1990;8(7):1277-1280.
7. Wen PY, Macdonald DR, Reardon DA, et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. J Clin Oncol. 2010;28(11):1963-1972.
8. Gállego Pérez-Larraya J, Lahutte M, Petirena G, et al. Response assessment in recurrent glioblastoma treated with irinotecan-bevacizumab: comparative analysis of the Macdonald, RECIST, RANO, and RECIST + F criteria. Neuro Oncol. 2012;14(5):667-673.
9. Norden AD, Young GS, Setayesh K, et al. Bevacizumab for recurrent malignant gliomas: efficacy, toxicity, and patterns of recurrence. Neurology. 2008;70(10):779-787.
10. Iwamoto FM, Abrey LE, Beal K, et al. Patterns of relapse and prognosis after bevacizumab failure in recurrent glioblastoma. Neurology. 2009;73(15):1200-1206.
11. Ellingson BM, Cloughesy TF, Lai A, et al. Quantitative volumetric analysis of conventional MRI response in recurrent glioblastoma treated with bevacizumab. Neuro Oncol. 2011;13(4):401-409.
12. Ananthnarayan SB, Bahng J, Roring J, et al. Time course of imaging changes of GBM during extended bevacizumab treatment. J Neurooncol. 2008;88(3):339-347.
13. Boxerman JL, Zhang Z, Safriel Y et al. Early post-bevacizumab progression on contrast-enhanced MRI as a prognostic marker for overall survival in recurrent glioblastoma: results from the ACRIN 6677/RTOG 0625 Central Reader Study. Neuro-Oncology. 2013;15(7):945-954.
14. Taal W, Oosterkamp HM, Walenkamp AM, et al. Single-agent bevacizumab or lomustine versus a combination of bevacizumab plus lomustine in patients with recurrent glioblastoma (BELOB trial): a randomized controlled phase 2 trial. Lancet Oncol. 2014;15(9):943-953.
15. Ellingson BM, Kim HJ, Woodworth DC, et al. Recurrent glioblastoma treated with bevacizumab: contrast-enhanced T1-weighted subtraction maps improve tumor delineation and aid prediction of survival in a multicenter clinical trial. Radiology. 2014;271(1):200-210.
16. Pichler J, Pachinger C, Pelz M, et al. MRI assessment of relapsed glioblastoma during treatment with bevacizumab: volumetric measurement of enhanced and FLAIR lesions for evaluation of response and progression - a pilot study. Eur J Radiol. 2013;82(5):240-245.
17. Chow DS, Qi J, Miloushev VZ, et al. Semiautomated volumetric measurement on postcontrast MR imaging for analysis of recurrent and residual disease in glioblastoma multiforme. AJNR Am J Neuroradiol. 2014;35(3):498-503.

18. Wang MY, Cheng JL, Han YH, et al. Measurement of tumor size in adult glioblastoma: classical cross-sectional criteria on 2D MRI or volumetric criteria on high resolution 3D MRI? Eur J Radiol. 2012;81(9):2370-2374.

19. Gerstner ER, Chen PJ, Wen PY, et al. Infiltrative patterns of glioblastoma spread detected via diffusion MRI after treatment with cediranib. Neuro Oncol. 2010;12(5):466-472.

20. Faul F, Erdfelder E, Lang AG, et al. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39(2):175-191.

21. Huang RY, Rahman R, Hamdan A, et al. Recurrent glioblastoma: volumetric assessment and stratification of patient survival with early posttreatment magnetic resonance imaging in patients treated with bevacizumab. Cancer. 2013;119(19):3479-3488.

22. Schaub C, Greschus S, Seifert M, et al. FLAIR-only progression in bevacizumab-treated relapsing glioblastoma does not predict short survival. Oncology. 2013;85(3):191-195.

23. Tran LN, Brown MS, Goldin JG, et al. Comparison of treatment response classifications between unidimensional, bidimensional, and volumetric measurements of metastatic lung lesions on chest computed tomography. Acad Radiol. 2004;11(12):1355-1360.

24. Shah GD, Kesari S, Xu R, et al. Comparison of linear and volumetric criteria in assessing tumor response in adult high-grade gliomas. Neuro Oncol. 2006;8(1):38-46.

25. Sohaib SA, Turner B, Hanson JA, et al. CT assessment of tumour response to treatment: comparison of linear, cross-sectional and volumetric measures of tumour size. Br J Radiol. 2000;73(875): 1178-1184.

26. Dempsey MF, Condon BR, Hadley DM. Measurement of tumor "size" in recurrent malignant glioma: 1D, 2D, or 3D? AJNR Am J Neuroradiol. 2005;26(4):770-776.

27. Galanis E, Buckner JC, Maurer MJ, et al. Validation of neuroradiologic response assessment in gliomas: measurement by RECIST, two-dimensional computer-assisted tumor area, and computer-assisted tumor volume methods. Neuro Oncol. 2006;8(2):156-165.

28. Sorensen AG, Patel S, Harmath C, et al. Comparison of diameter and perimeter methods for tumor volume calculation. J Clin Oncol. 2001;19(2):551-557.

29. Zuniga RM, Torcuator R, Jain R, et al. Efficacy, safety and patterns of response and recurrence in patients recurrent high-grade gliomas treated with bevacizumab plus irinotecan. J Neurooncol. 2009;91(3):329-336.

30. Rubenstein JL, Kim J, Ozawa T, et al. Anti-VEGF antibody treatment of glioblastoma prolongs survival but results in increased vascular cooption. Neoplasia. 2000;2(4):306-314.

31. Bähr O, Hattingen E, Rieger J, et al. Bevacizumab-induced tumor calcifications as a surrogate marker of outcome in patients with glioblastoma. Neuro Oncol. 2011;13(9): 1020-1029.

32. Bähr O, Harter PN, Weise LM, et al. Sustained focal antitumor activity of bevacizumab in recurrent glioblastoma. Neurology. 2014;83(3):227-234.

33. Mong S, Ellingson BM, Nghiemphu PL, et al. Persistent diffusion-restricted lesions in bevacizumab-treated malignant gliomas are associated with improved survival compared with matched controls. AJNR Am J Neuroradiol. 2012;33(9):1763-1770.

34. Ellingson BM, Bendszus M, Boxerman J, et al. Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. Neuro Oncol. 2015;17(9):1188-1198.

SUPPLEMENTARY FILES

MRI protocol.

The standardized MRI protocol consisted of the following sequences with full brain coverage:

- Pre-contrast three-dimensional (3D) T1-weighted (T1w) IR FSPGR images with slice thickness and in-plane resolution ≤1mm.
- Transverse diffusion weighted image (DWI) with slice thickness 3mm (no gap), an in-plane resolution of 2mm, and b=0 and b=1000 s/mm^2.
- 3D T2-weighted (T2w) fluid attenuation inversion recovery (FLAIR) with slice thickness and in-plane resolution ≤1mm and fat-saturation (if possible).
- Transverse 2D T2w image with slice thickness ≤3mm and in-plane resolution ≤1mm.
- Post-contrast 3D T1w IR FSPGR images with slice thickness and in-plane resolution ≤1mm.
- Selected sites also performed dynamic susceptibility contrast perfusion imaging: transverse gradient echo (GE) echo planar imaging (EPI) perfusion with slice thickness ≤5mm (no gap), an in-plane resolution of ≤3mm, repetition time <2000ms, 50 phases (100s) with a scan delay of 20s after injection of 10mmol Gd at 4-5ml/s followed by 25ml NaCl. A contrast preload bolus of 0.05 Gd mmol/kg body weight was given prior to the 3D FLAIR acquisition. In sites not performing perfusion imaging, the full contrast bolus was given prior to the 2D T2w image acquisition.

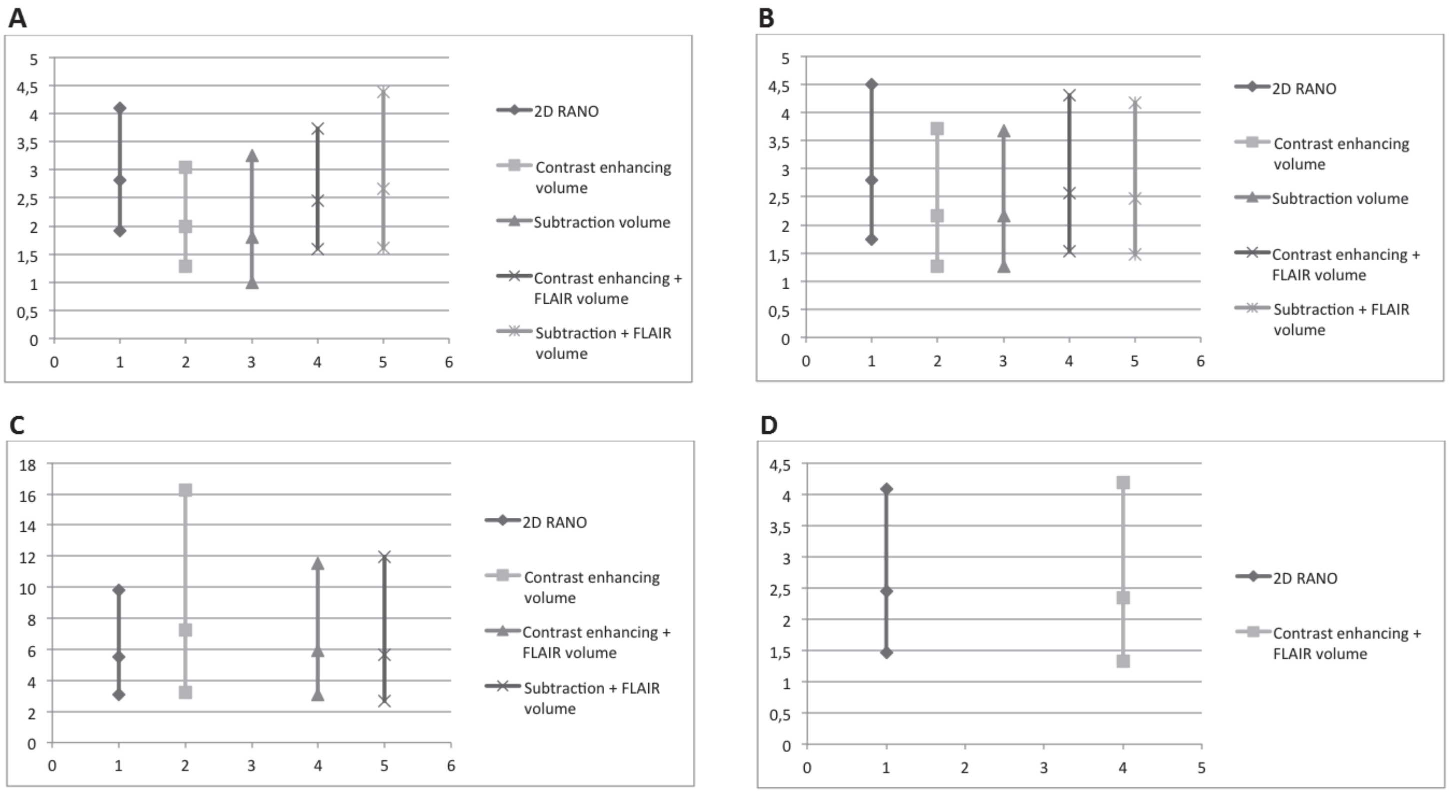

Figure S1. All patients together, hazard ratios and their 95% confidence intervals at first (A) and second (B) follow-up (as depicted in Table 1 in the manuscript). Bevacizumab-treated patients only, hazard ratios and their 95% confidence intervals at first (C) and second (B) follow-up (as depicted in Table 2 in the manuscript).

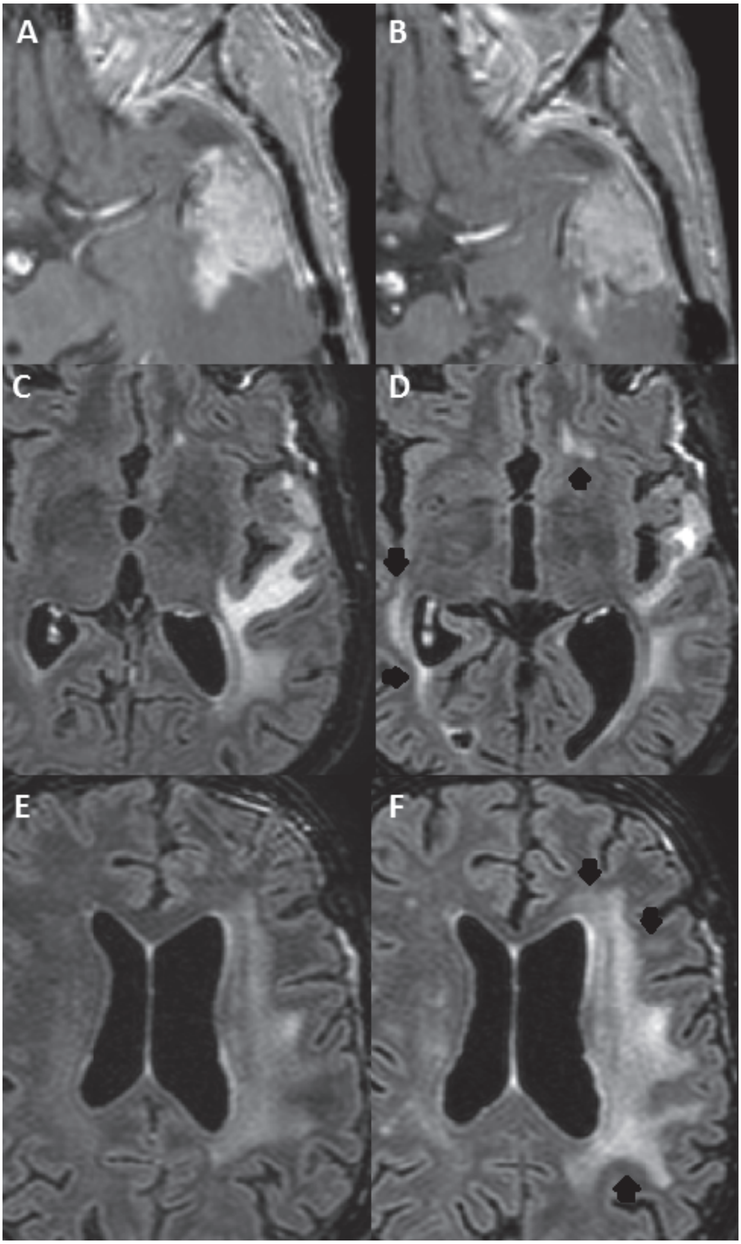

Figure S2. Discrepant case 1. Patient treated with bevacizumab and lomustine, scored as no progressive disease (PD) with 2D RANO (method 1) and contrast enhancing volume (method 2), but as PD with contrast-enhancing plus FLAIR volume (method 4). Enhancing lesion in left temporal lobe at baseline (**A**) and first follow-up (**B**) shows decreasing enhancement (decrease in volume of 13.23%) and increasing non-enhancing/FLAIR area at baseline (**C** and **E**) and first follow-up (**D** and **F**) at two different levels (increase in volume of 30.94%). The threshold for PD based on T2w/FLAIR volume increase is set at 25%. Black arrows indicate increasing FLAIR abnormalities.

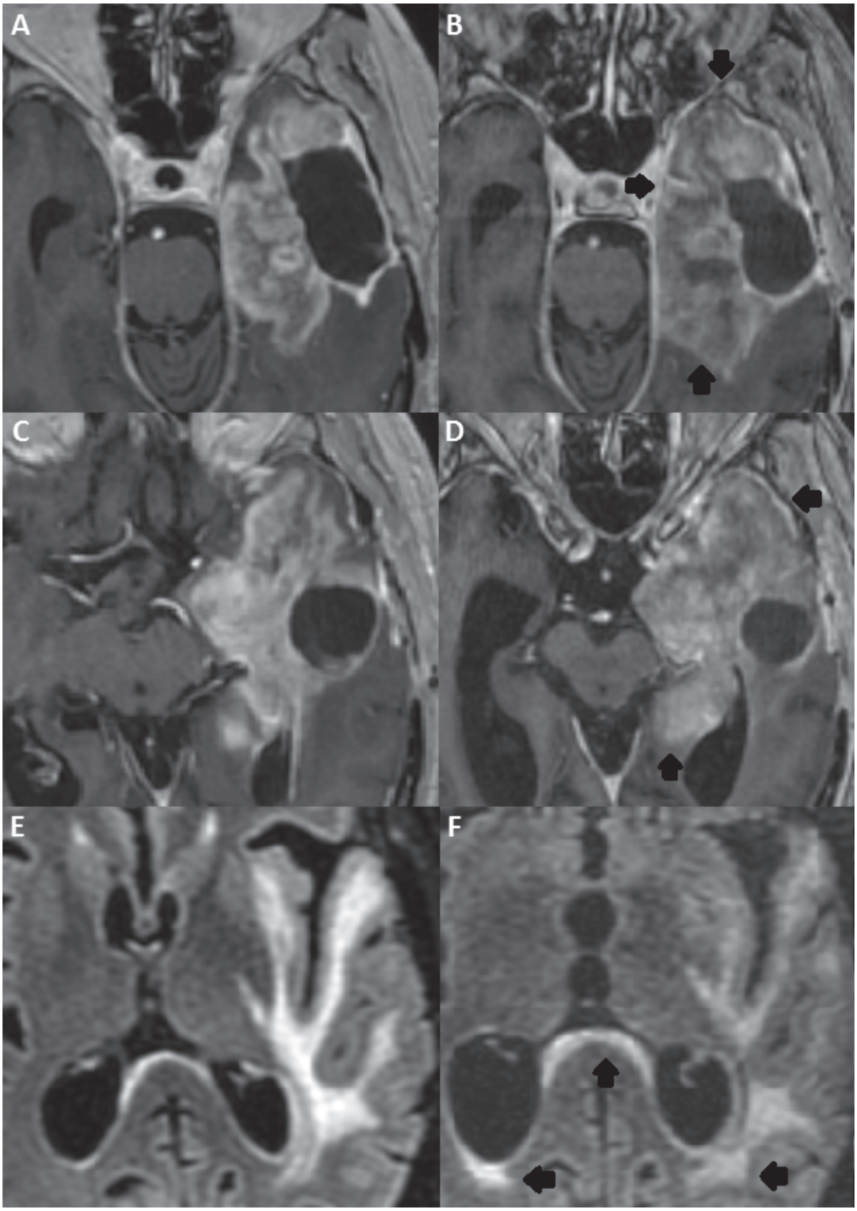

Figure S3. Discrepant case 2. Patient treated with lomustine, scored as progressive disease (PD) with 2D RANO (method 1), but as non PD with contrast enhancing volume (method 2) and contrast-enhancing plus FLAIR volume (method 4). Enhancing lesion in the left temporal lobe at baseline (A) and first follow-up (B) with limited increase in volume (increase 3.39%) and non-enhancing/FLAIR area (increase 6.0%) at baseline (C and E) and first follow-up (D and F) at two different levels. Note the differences in angulation between the different time points and the poor quality due to motion artifacts in the scans made at first follow-up. Black arrows indicate increasing enhancement and FLAIR abnormalities.

Table S1. Concordances between method 1 (2D RANO) and the volumetric methods, i.e. method 2 (contrast enhancing volume), method 3 (subtraction volume), method 4 (contrast enhancing + FLAIR volume), and method 5 (subtraction + FLAIR volume) at first follow-up in all patients together and in the bevacizumab-treated groups. In case of non-concordance, the presence of progressive disease (PD) was more frequently observed with method 1 than with the volumetric methods (2-5).

First follow-up	Comparison	Total number of patients	Number of concordant cases	PD established with method1	PD established with method 2/3/4/5
All patients	Method 1 vs 2	121	101 (83.5%)	16 (13.2%)	4 (3.3%)
	Method 1 vs 3	86	76 (87.4%)	9 (10.3%)	1 (1.1%)
	Method 1 vs 4	109	91 (83.5%)	10 (9.2%)	8 (7.3%)
	Method 1 vs 5	89	75 (84.3%)	9 (10.1%)	5 (5.6%)
Bevacizumab treated patients	Method 1 vs 2	85	70 (82.4%)	12 (14.1%)	3 (3.5%)
	Method 1 vs 4	74	61 (82.4%)	6 (8.1%)	7 (9.5%)
	Method 1 vs 5	64	52 (81.3%)	7 (10.9%)	5 (7.8%)