



## Research article

## Comparison of perioperative automated versus manual two-dimensional tumor analysis in glioblastoma patients



Frauke Kellner-Weldon<sup>a,\*</sup>, Christoph Stippich<sup>b</sup>, Roland Wiest<sup>a</sup>, Vera Lehmann<sup>a</sup>, Raphael Meier<sup>c</sup>, Jürgen Beck<sup>d</sup>, Philippe Schucht<sup>d</sup>, Andreas Raabe<sup>d</sup>, Mauricio Reyes<sup>c</sup>, Andrea Bink<sup>b</sup>

<sup>a</sup> Support Center for Advanced Neuroimaging – Institute for Diagnostic and Interventional Neuroradiology, University Hospital Inselspital and University of Bern, Bern, Switzerland

<sup>b</sup> Department of Radiology, Division of Diagnostic and Interventional Neuroradiology, University Hospital, Basel, Switzerland

<sup>c</sup> Institute of Surgical Technology and Biomechanics, University of Bern, Bern, Switzerland,

<sup>d</sup> Department of Neurosurgery, University Hospital Inselspital and University of Bern, Bern, Switzerland

## ARTICLE INFO

## Keywords:

Computer assisted reading  
automated data analysis  
glioblastoma  
machine learning  
MRI

## ABSTRACT

**Objectives:** Current recommendations for the measurement of tumor size in glioblastoma continue to employ manually measured 2D product diameters of enhancing tumor. To overcome the rater dependent variability, this study aimed to evaluate the potential of automated 2D tumor analysis (ATA) compared to highly experienced rater teams in the workup of pre- and postoperative image interpretation in a routine clinical setting.

**Materials and methods:** From 92 patients with newly diagnosed GB and performed surgery, manual rating of the sum product diameter (SPD) of enhancing tumor on magnetic resonance imaging (MRI) contrast enhanced T1w was compared to automated machine learning-based tumor analysis using FLAIR, T1w, T2w and contrast enhanced T1w.

**Results:** Preoperative correlation of SPD between two rater teams (1 and 2) was  $r = 0.921$  ( $p < 0.0001$ ). Difference among the rater teams and ATA ( $p = 0.567$ ) was not statistically significant. Correlation between team 1 vs. automated tumor analysis and team 2 vs. automated tumor analysis was  $r = 0.922$  and  $r = 0.897$ , respectively ( $p < 0.0001$  for both). For postoperative evaluation interrater agreement between team 1 and 2 was moderate (Kappa 0.53). Manual consensus classified 46 patients as completely resected enhancing tumor. Automated tumor analysis agreed in 13/46 (28%) due to overestimation caused by hemorrhage and choroid plexus enhancement.

**Conclusions:** Automated 2D measurements can be promisingly translated into clinical trials in the preoperative evaluation. Immediate postoperative SPD evaluation for extent of resection is mainly influenced by postoperative blood depositions and poses challenges for human raters and ATA alike.

### 1. Introduction

Glioblastoma is the most common primary brain tumor in adults [1]. Complete resection of this tumor entity is currently not possible, due to its infiltrating growth [2]. Still, the largest possible extent of resection is the primary goal of surgery as it has been shown to improve overall survival [3,4]. Therefore, an accurate preoperative evaluation of the tumor is needed prior to surgery. Current recommendations for the measurement of tumor size (Response-Assessment Neuro-Oncology (RANO) working group) [5], continue to employ two-dimensional (2D) product diameters of enhancing tumor on MRI. To overcome the rater

dependent variability and the time consuming measurement [6–8], automated delineation methods that segment tumor as a three-dimensional (3D) volume have been developed [9,10].

Automated and manual tumor subcompartment delineation has shown comparable performance in terms of prognosis and correlation with Visually Accessible Rembrandt Images (VASARI) features [11].

The aim of this study was to evaluate the performance of automated tumor analysis (ATA) to identify complete vs. incomplete resections in comparison with human image interpretations and 2-D diameter measurements. Complete resection (CRET) was defined as absence of any contrast-enhancing tumor volume on ceT1w imaging after surgical

**Abbreviations:** GB, glioblastoma; MRI, magnetic resonance imaging; FLAIR, fluid attenuation inversion recovery; SPD, sum product of diameter; ATA, automated tumor analysis; ceT1w, contrast enhanced T1weighted

\* Corresponding author at: Institute for Diagnostic and Interventional Neuroradiology, Inselspital, Bern University Hospital, Freiburgstrasse 4, 3010, Bern, Switzerland.

E-mail address: [Frauke.Kellner-Weldon@insel.ch](mailto:Frauke.Kellner-Weldon@insel.ch) (F. Kellner-Weldon).

procedure.

Although 3-D measures are increasingly recognized as alternative surrogate markers for tumor progression, 2-D measures are still recommended for the routine clinical follow-up. The goal of this study was to evaluate i) the agreement between automatic and manual estimates of 2-D tumor measures in preoperative images and ii) if automated tools can detect the amount of resection if 2-D measures are applied.

## 2. Materials and methods

### 2.1. Study population

We retrospectively identified 92 patients who were diagnosed with a histologically proven GBM and who underwent resective surgery and immediate postoperative MRI after resection. All patients were examined between 2009 and 2013. Inclusion criteria were newly diagnosed, untreated and histologically confirmed GB (WHO IV), performed surgery, Karnofsky performance  $\geq 70$  and age  $> 18$  y. Exclusion criteria were prior other malignancy, biopsy performed previously on the GB, postoperative MRI later than 72 h. The study was approved by the Local Research Ethics Commission. All patients provided written informed consent.

### 2.2. Automated image processing

We used the automatic brain tumor analysis software BraTumIA, which has been clinically evaluated for longitudinal tumor volumetry in previous studies. Within the framework of this study, we were interested in the automatic detection of enhancing tumor, which is part of the more extended volume analysis that is offered by the software. A detailed description of the software and its potential applications has been published previously [12,13]. In short, BraTumIA is a supervised machine learning based software that relies on expert annotated training data to learn the relationships between imaging features and tissue classification. It relies on multisequence MRI (T1weighted (w), contrast enhanced (ce) T1w, T2w, fluid attenuated inversion recovery (FLAIR)) to perform automated tumor analysis. It performs co-registration of the multisequence images, skull-stripping (i.e. brain extraction), and tissue classification. Beyond volumetric measurements, BraTumIA also provides measures of (2D) tumor diameters.

As human reference, four board-certified neuroradiologists with a mean of 13 years (range, 7–17y) of experience in neuroradiology, working in two different university hospitals, rated the imaging data in two teams (team 1 and team 2, each consisting of two raters from both hospitals) in a consensus fashion.

### 2.3. Manual annotations and quality control

Manual annotations were performed on contrast enhanced T1weighted images (ceT1w) after checking for confounding blood products on T1w images. As annotations we used sum of the products of

diameters (SPD) measures, and presence/absence of residual tumor. Tumor SPD measures (in  $\text{mm}^2$ ) were acquired according to the RANO recommendations, using a hospital picture archiving system (Sectra IDS7, Linköping, Sweden) by each team. Similarly, each team rated the presence/absence of residual tumor. In case of disagreements, a final joint consensus reading was achieved by team 3, which consisted of one rater from each team 1 and 2.

For automated analysis, BraTumIA was trained on an independent dataset of 54 pre- and post-operative cases, as described in [12]. Upon automated analysis of the segmented pre- and postoperative MRI data sets, two readers performed quality control of the results to see if (i) the data were inconclusively processed due to motion artefacts or (ii) if the skull stripping process had failed. If any of these conditions was found, the MRI dataset was removed.

### 2.4. Statistical methods

Statistical analysis was performed with IBM SPSS Statistics 21.ink and the R software package. Agreement on presence/absence of residual tumor between teams 1 and 2 was assessed using the Cohen's kappa statistic. A Kruskal-Wallis test was used to assess multiple differences between manual and automated preoperative mean values of tumor diameters. Paired differences were analyzed using a Wilcoxon signed-rank test. Correlation of results for mean SPD-values between raters and automated analysis was analyzed using Pearson's r. A value of  $p < 0.05$  was considered statistically significant.

## 3. Results

92 patients (47 male, 45 female; mean 62 y, range 60–80 y) met the inclusion criteria. Time between pre- and postoperative MRI was 7.5 days, (range 1–55, median 5.5 days) and between operation and postoperative MRI 1.15 days (range 0–2 days). All in-house examinations (pre-/postoperative) (n = 81/92) were performed on MR scanners of the same vendor (Siemens Medical Solutions, Erlangen, Germany- 1.5 T Avanto (n = 30/49), 1.5 T Aera (n = 12/12), 3.0 T Verio (n = 21/20) and 3.0 Trio (n = 18/11)). Of the remaining 11 preoperative MR images six were performed on Siemens scanners [1.5 T Avanto (n = 3), 1.5 T unknown scanner type (n = 2), 1.0 T unknown scanner type (n = 1)], three on Philips scanners [1.5 T Intera (n = 1), 1.5 T unknown scanner type (n = 1), 3.0 T unknown scanner type (n = 1)], and two on GE scanners [1.0 T unknown scanner type (n = 1), and 3.0 T Discovery MR750 (n = 1)]. The sequence parameters of the T1w, T2w, FLAIR and contrast-enhanced T1 w (ceT1) images are shown in Table 1. In-house patients received 0.1 mmol/kg gadolinium based contrast agent (Gd-DTPA). Data on contrast agent and dose for non in-house patients was not available. For an overview of processed patient data see Fig. 1.

### 3.1. Manual annotation of 2D diameters

Correlation (Pearson's R-value) of SPD values measured between

**Table 1**  
Pre- and postoperative MRI sequence parameters included for automated analysis (n = 60 patients).

| preoperative | postoperative | Sequence T1 non-enhanced | Sequence T1 contrast-enhanced | Sequence T2 weighted | Sequence FLAIR |
|--------------|---------------|--------------------------|-------------------------------|----------------------|----------------|
| 7            | 7             | T1 5 mm <sup>a</sup>     | T1C 5 mm                      | T2 5 mm              | FLAIR 4–5 mm   |
| 2            | 1             | T1 5 mm                  | T1C 5 mm                      | T2 3D 1 mm           | FLAIR 4–5 mm   |
| 25           | 28            | T1 5 mm                  | T1C 3D 1 mm                   | T2 5 mm              | FLAIR 4–5 mm   |
| 11           | 12            | T1 5 mm                  | T1C 3D 1 mm                   | T2 3D 1 mm           | FLAIR 4–5 mm   |
| 2            | 0             | T1 5 mm                  | T1C 3D 1 mm                   | T2 5 mm              | FLAIR 3D 1 mm  |
| 2            | 1             | T1 3D 1 mm               | T1C 3D 1 mm                   | T2 5 mm              | FLAIR 4–5 mm   |
| 11           | 10            | T1 3D 1 mm               | T1C 3D 1 mm                   | T2 3D 1 mm           | FLAIR 4–5 mm   |
| 0            | 1             | T1 3D 1 mm               | T1C 5 mm                      | T2 3D 1 mm           | FLAIR 4–5 mm   |

<sup>a</sup> cursive: spin echo sequences; all others gradient echo sequences.

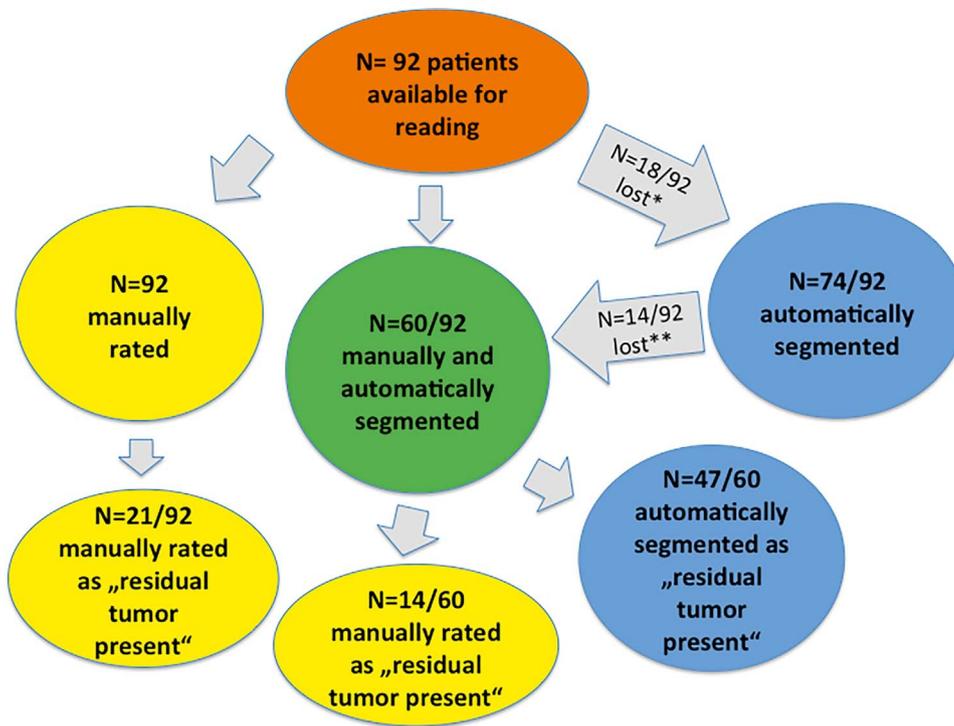


Fig. 1. Overview of available patient data for manual and automated analysis.

\*Not usable for automated analysis due to missing MRI data, motion artefacts, and difficulties of the program with skull stripping.

\*\*Not usable for automated segmentation due to gross faulty non-detection of enhancing tumor (13 patients) and due to missing manual data of one patient from one rater team and therefore no possible comparison.

team 1 and 2 was 0.921 ( $p < 0.0001$ ). When rating the postoperative images for residual tumor, team 2 omitted decision in 2 patients, due to uncertainty (diffuse signal changes at the resection cavity), leaving 90 patients for further analysis. There was moderate interrater agreement of team 1 and 2 for detection of postoperative residual tumor (Kappa 0.53). Twelve patients were initially rated to have residual tumor by both teams. In 13 patients, differing teams' results led to a consensus reading of these patients, which resulted in residual tumor in 9 of 13 patients rated concordantly by team 3. By consensus 21 out of 90 patients were evaluated to have residual tumor.

### 3.2. Quality control

In the 92 included cases, incomplete data sets led to exclusion of 11 patients. Motion artefacts were found in one patient. Skull stripping failed in 6 patients. These 18 exclusions led to a patient number to be further evaluated of  $n = 74$ . ATA was considered incomplete to a variable degree in 13 out of 74 patients preoperatively. As the predominant reason, we identified tumor that abutted the dura, which prohibited the appropriate delineation of the neighboring tissues. Detailed reasons for exclusion were: 4 × failure of multimodal sequence alignment, 1 × missegmentation of the whole brain, 8 × misdetection of tumor tissues. Of the 61 patients left, one patient had been manually rated only by team 1 and was therefore not available for comparison. Thus, overall 60 patients were included into the automated analysis.

### 3.3. Comparison of automated versus manual 2D annotations on preoperative data

For 60 patients the preoperative mean SPD of team 1, team 2, and the software were 1382 mm<sup>2</sup> (SD  $\pm$  776 mm<sup>2</sup>), 1486 mm<sup>2</sup> ( $\pm$  779 mm<sup>2</sup>), and 1330 mm<sup>2</sup> ( $\pm$  710 mm<sup>2</sup>), respectively. The absolute difference of the mean SPD between team 1 and team 2 was 104 mm<sup>2</sup>, automated versus team 1 was 52 mm<sup>2</sup> and automated versus team 2 was 156 mm<sup>2</sup> (Fig. 2). A Kruskal-Wallis test under  $\alpha = 0.05$  did not detect a statistically significant difference among the rater teams and ATA ( $p = 0.567$ ). Correlation (Pearson's R) between team 1 vs. ATA and team 2 vs. ATA was 0.922 and 0.897, respectively, with

$p < 0.0001$  for both. See Table 2.

### 3.4. Comparison of automated versus manual 2D annotations on postoperative data

Of the 60 patients included for ATA, the manual consensus reading yielded 46 patients classified as “completely resected”, and 14 as “partially resected”. ATA agreed in 13/46 (28%) with CRET (SPD value of zero). For the remaining 33/46 (72%) it yielded a non-zero SPD, disagreeing with the manual consensus rating. The mean SPD for these cases was 387 mm<sup>2</sup> ( $\pm$  389). For the 14 patients classified as “partially resected” by manual consensus rating a mean SPD of 100 mm<sup>2</sup> ( $\pm$  96) and 361 mm<sup>2</sup> ( $\pm$  258) was measured by manual and ATA, respectively (Fig. 3). See Table 2.

### 3.5. Qualitative data analysis

In order to better understand the disagreement of ATA with manual consensus rating on postoperative imaging data, a qualitative post-hoc analysis was performed for all automatically segmented cases ( $n = 74$ ). The analysis identified several confounding factors for equivocal tissue classification: The choroid plexus, blood vessels and blood products, with the latter forming the predominant source of error. An example of a preoperative segmentation and misclassified blood products in the respective postoperative image is shown in Fig. 4. Fig. 5 shows comparison of manual and automated 2-D measurements of enhancing tumor pre- and postoperatively for a patient with manually identified residual enhancing tumor and additional misclassification of scattered hyperintensities.

## 4. Discussion

Machine learning has gained much attention in computer-assisted radiology, particularly because it provides an automatic way to generalize human knowledge obtained from training data to future unknown test data. Translation from machine learning to clinical practice is a stepwise process and ATA has already shown to be successfully supporting radiologists and clinicians by providing accurate measures

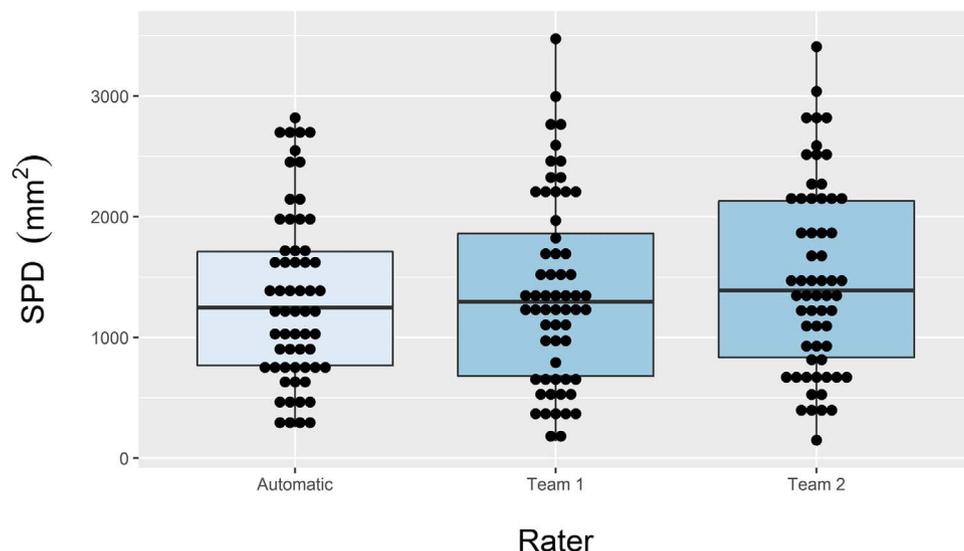


Fig. 2. Comparison of automated (light color) and manual (dark color) sum product diameter (SPD) for n = 60 patients.

**Table 2**  
Pre- and postoperative size of enhancing tumor sum product diameter (SPD) from automated tumor analysis (ATA) and expert rating (Manual).

|        |                      | N  | Minimum SPD/mm <sup>2</sup> | Maximum SPD/mm <sup>2</sup> | Mean SPD/mm <sup>2</sup> | Standard Deviation SPD/mm <sup>2</sup> |
|--------|----------------------|----|-----------------------------|-----------------------------|--------------------------|--|
| ATA    | preoperative         | 60 | 254                         | 2819                        | 1330                     | 710                                    |
|        | postoperative        | 47 | 59                          | 1644                        | 297                      | 354                                    |
| Manual | preoperative         | 60 | 140                         | 3475                        | 1382                     | 776                                    |
|        | team 1 preoperative  | 60 | 148                         | 3409                        | 1486                     | 779                                    |
|        | team 2 postoperative | 21 | 20                          | 742                         | 131                      | 166                                    |
|        | team 3               |    |                             |                             |                          |  |

of 3D tumor extensions in GB preoperatively and within longitudinal studies in single-center studies [12].

Although 3D-derived tumor volume measurements have been requested repeatedly by neurosurgeons and oncologists, 2D-based tumor measurements are still the basic method of choice in assessment of tumor progression and response, respectively. In this study, we investigated the potential of automated 2D measures to quantify residual tumor. The comparison between the MR-evaluation of pre- and

postoperative SPD measures of 92 patients with newly diagnosed GB by two rater teams and ATA showed no significant difference for the preoperative data. The postoperative evaluation highlighted common challenges in the assessment of the extent of surgical resections, namely the uncertainty of human rating and the susceptibility of automated analysis to image artifacts. We identified a moderate interrater agreement between human raters, even when acting in mixed pairs of experts from different centers (Kappa 0.53). The potential strength of human rating is the ability to identify hemorrhage and calcifications more safely than automated analysis. This was reflected in our study by a poor agreement (28%) between the manual (46) and automated detection (13) of complete resection.

A promising finding of our study was that both manual and automated approaches performed similarly regarding preoperative SPD measures, which is in line with previous work done by Porz et al. [13] and has been confirmed in a second study [14]. Contrary to previous analyses, the dataset under investigation, are derived from a clinical setting with a clinical protocol of a tertiary hospital that includes outpatient imaging with an overall less rigid imaging protocol. In addition, this study compared manual and automatic SPD, as recommended by RANO, instead of volumetric measurement. The reconfirmed preoperative agreement suggests that automated analysis is useful for processing large data sets where manual delineation is prohibitive –

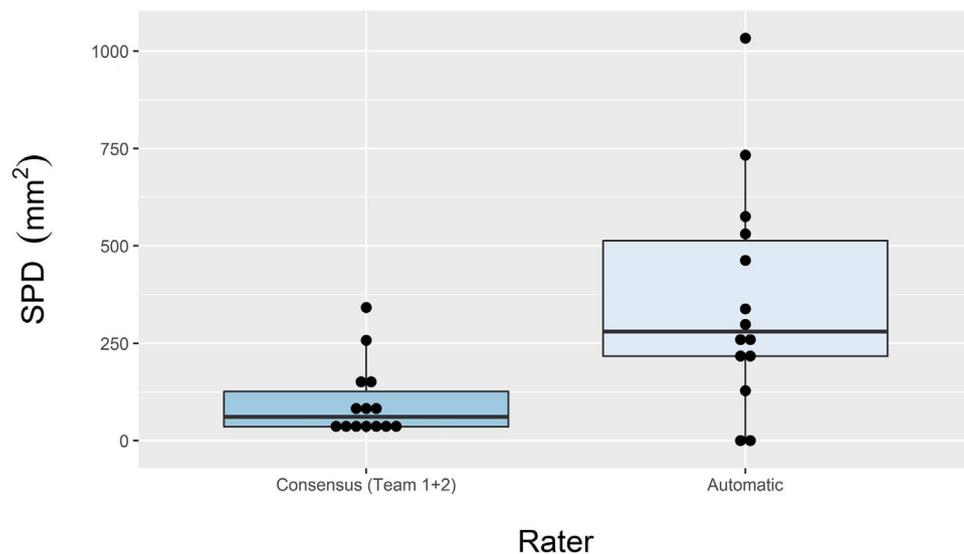


Fig. 3. Comparison of automatic (light color) and manual (dark color) postoperative annotations in N = 14 patients.

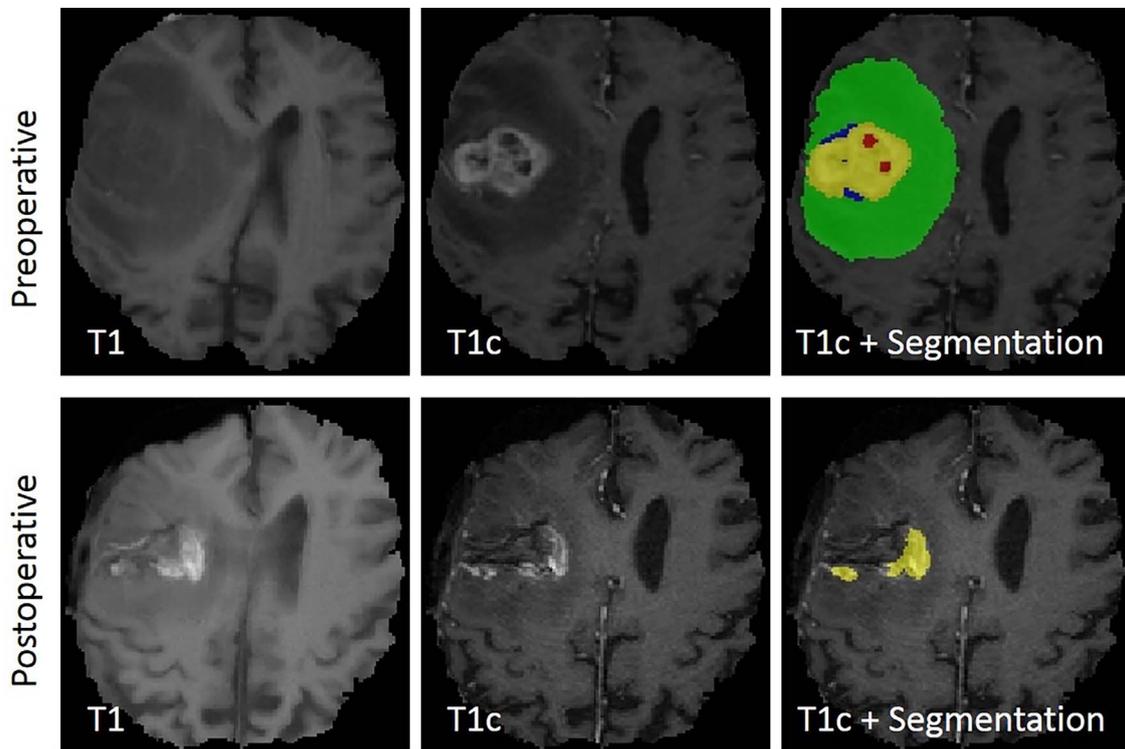


Fig. 4. Example segmentation for a pre- (upper row) and postoperative image (lower row) of the same patient. The preoperative segmentation includes edema (green), necrosis (red), enhancing (yellow) and non-enhancing tumor (blue). The postoperative segmentation shows blood products that were wrongly identified as blood products by the algorithm. From left to right: T1-weighted image, gadolinium-enhanced T1-weighted image, delineation of automated method.

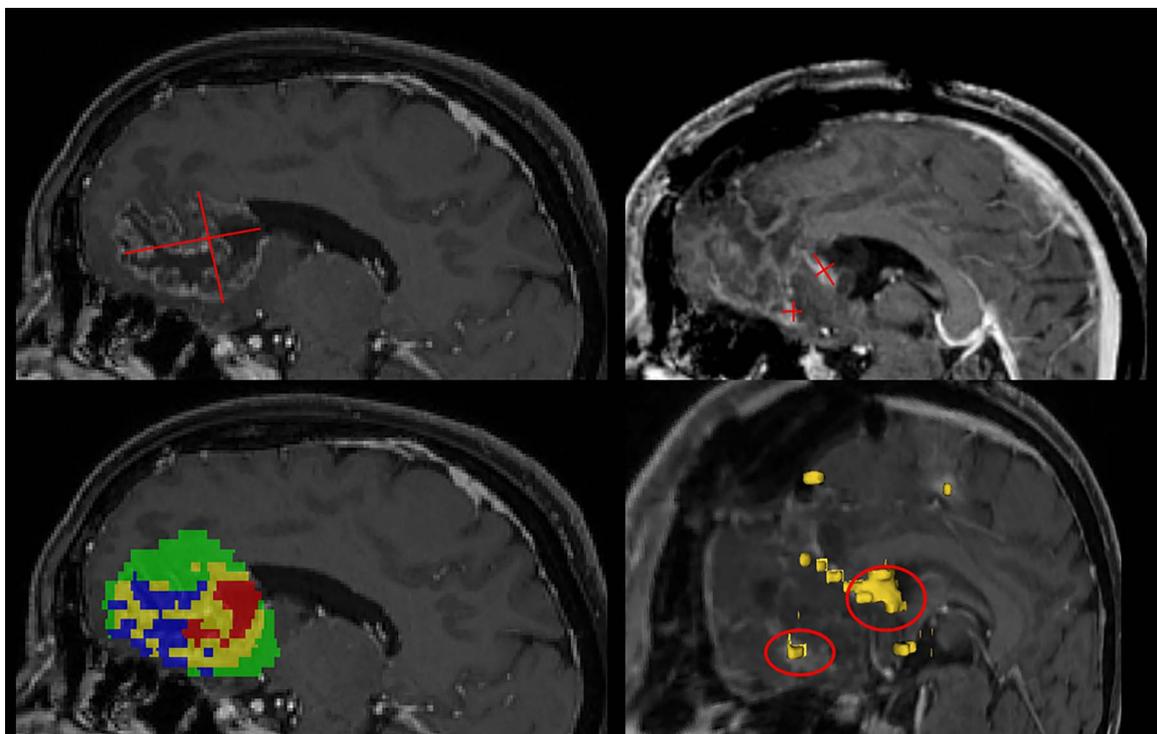


Fig. 5. Example patient (contrast enhanced T1 images) with pre- (left column) and postoperative images (right column). Images in upper row show the manual annotation of two perpendicular diameters of enhancing tumor preoperatively (left) on 2D sagittal images. Postoperatively (right) measurements include perpendicular diameters of enhancing residual tumor in two locations. Preoperative segmentation (lower row, left) includes edema (green), necrosis (red), enhancing (yellow) and non-enhancing tumor (blue). Postoperative segmentation (lower row, right 3D overlay) shows the two locations (red circle), identified as residual tumor in manual annotation, as well as scattered small hyperintensities, misclassified as residual tumor (yellow).

under the precondition that the automated results need to be checked by radiologists.

From 92 patients approximately one third of the patients did not have workable imaging data for automated analysis. In 11 patients the standard protocol did not include all required sequences. In 4 patients imaging did not cover the tumor completely in all planes. These excluded data maybe considered a limitation of this study, but in a proper sense they indicate the importance of developing imaging sequences and analysis tools in a joint and clinically-relevant fashion in order to optimize the extraction of radiological information.

For the preoperative data, automated analysis revealed drawbacks in some patients whose tumor was located near the skull and abutted the dura. In these cases ATA could not completely separate signal intensity of the dura and connective tissue of the skull from tumor tissue. This yielded grossly false over- or underestimation of tumor volumes, yet a mistake that can be easily detected by the expert reader. These drawbacks show that manual control for correct segmentation is currently a precondition for ATA. In addition, training data should be selected accordingly to improve recognition of tumor versus dura and connective tissue.

An important finding of our study was that automated postoperative SPD measures revealed a trend towards overestimation. Postoperatively, the source of error for overestimation was due to misassignment of tumor to other tissues. The automated evaluation found residual tumor in 72% (33/46 patients) of the datasets, which were assigned “completely resected” by manual expert reading. The major confounder incorrectly segmented and attributed to tumor were blood products around the resection cavity. Second to postoperative hemorrhage was erroneous detection of vessels and choroid plexus. For the latter two, we claim that this does not hamper postoperative monitoring tumor response, because this bias is likely constant in the pre- and postoperative setting. The third entity incorrectly segmented and attributed to tumor were blood products around the resection cavity. Clinically, using thresholds, such as 98% extent of resection of GB volume, has proven useful to find significant survival advantage [15]. In imaging, currently, lesions with one diameter smaller than 10 mm are defined as non-measurable by the latest RANO criteria. Similarly to the clinical approach of thresholding, we suggest the introduction of threshold values for 2-D-based tumor size measurement that go beyond the current recommendations for non-measurable lesions by considering computer-generated 2D measures, which are able to better capture tumor extent independent of anatomic-plane, and can avoid an otherwise non-measurable lesion caused by the given orientation, as reported by Reuter et al. [16]. As computer-assisted tumor volumetry progresses, we expect it to bring new solutions to handle and better monitor lesions currently defined as non-measurable. This might also further improve uncertainties resulting from low-levels of agreement about residual enhancing tumor among human raters. It reflects clinical practice, where subtle differences between hyperintense signal changes on non-enhanced T1 w and ceT1 w are not easy to recognize and decision making between blood, residual tumor and vessel induced signal changes is still – even for experienced neuroradiologists motivated to find consensus – challenging.

Training the software so far has mostly relied on the human rater to define the different tumor compartments. However, for the difficult task of detecting and assessing tumor residuals, the recent study [17] has shown that the percentage agreement between BraTumIA and the human raters is comparable to the agreement between the human raters themselves, showing a good consistency level of the automated system. In the future, training of the ATA system will ideally incorporate such parameters as morbidity and survival outcome and tumor omics, in addition to an array of further possible imaging sequences (e.g. susceptibility-, diffusion-, and perfusion-imaging).

A limitation of the study is the inhomogeneity of sequence acquisition, due to the retrospective nature of the study design. Nearly 20% of the patients' data sets were not available for ATA mostly reasons

unrelated to the requirements of the ATA program. Current efforts to standardize brain tumor imaging protocols [18] will certainly increase the performance of ATA.

## 5. Conclusions

This study provides a stand-alone comparison between a fully automatic software tool and two expert rater groups from two university hospitals in a clinical setting which shows no significant differences between the preoperative automatic and manual results of GB assessment. Maximum diameter and SPD measures following the current RANO criteria of GB can be promisingly translated into clinical trials in preoperative evaluation.

Immediate postoperative SPD evaluation for extent of resection without a priori thresholding are influenced by the amount of postoperative blood depositions and reactive enhancement of residual structures neighboring the tumor bed after resection. Thus, such studies may benefit from a complementation by 3D volumetric criteria that incorporate a threshold for an extent of resection or residual tumor volume or by similarly adjusted 2D measures that have to be correlated with clinical outcome in further studies.

## Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

## Compliance with ethical standards

The authors declare that they have no conflict of interest. All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

## Funding and role of the funding source

This work was supported by the Bernische Krebsliga [grant number 158783] and received funding from the European Union's Seventh Framework Programme for research technological development and demonstration [grant number 600841]. The funding source was not involved in the study design, the collection, analysis or interpretation of data, in the writing of the report nor in the decision to submit the article for publication.

## Acknowledgment

We thank Lorena Hulliger for coordinating the reading and the picture archiving team of our local department for making imaging data available.

## References

- [1] F.B. Furnari, T. Fenton, R.M. Bachoo, A. Mukasa, J.M. Stommel, A. Stegh, W.C. Hahn, K.L. Ligon, D.N. Louis, C. Brennan, L. Chin, R.A. Depinho, W.K. Cavenee, Malignant astrocytic glioma: genetics, biology, and paths to treatment, *Genes Dev.* (2007) 2683–2710, <http://dx.doi.org/10.1101/gad.1596707>.
- [2] J.H. Rees, J.G. Smirniotopoulos, R.V. Jones, K. Wong, Glioblastoma multiforme: radiologic-pathologic correlation, *Radiographics* 16 (1996) 1413, <http://dx.doi.org/10.1148/radiographics.16.6.8946545>.
- [3] D. Kuhnt, A. Becker, O. Ganslandt, M. Bauer, M. Buchfelder, C. Nimsky, Correlation of the extent of tumor volume resection and patient survival in surgery of glioblastoma multiforme with high-field intraoperative MRI guidance, *Neuro Oncol.* 13 (2011) 1339–1348.
- [4] U. Pichlmeier, A. Bink, G. Schackert, W. Stummer, Resection and survival in glioblastoma multiforme: an RTOG recursive partitioning analysis of ALA study patients, *Neuro Oncol.* 10 (6) (2008) 1025–1034, <http://dx.doi.org/10.1215/15228517-2008-052>.

- [5] P.Y. Wen, D.R. Macdonald, D.A. Reardon, T.F. Cloughesy, A.G. Sorensen, E. Galanis, J. DeGroot, W. Wick, M.R. Gilbert, A.B. Lassman, C. Tsien, T. Mikkelsen, E.T. Wong, M.C. Chamberlain, R. Stupp, K.R. Lamborn, M.A. Vogelbaum, M.J. Van Den Bent, S.M. Chang, Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group, *J. Clin. Oncol.* 28 (2010) 1963–1972, <http://dx.doi.org/10.1200/JCO.2009.26.3541>.
- [6] B.A.G. Sorensen, S. Patel, C. Harmath, S. Bridges, J. Synnott, A. Sievers, Y. Yoon, E.J. Lee, M.C. Yang, R.F. Lewis, G.J. Harris, M. Lev, P.W. Schaefer, B.R. Buchbinder, G. Barest, K. Yamada, J. Ponzio, H.Y. Kwon, J. Gemmete, J. Farkas, A.L. Tievsky, R.B. Ziegler, M.R.C. Salhus, R. Weisskoff, Comparison of diameter and perimeter methods for tumor volume calculation, *J. Clin. Oncol.* 19 (2017) 551–557.
- [7] M.A. Deeley, A. Chen, R. Datterli, J.H. Noble, A.J. Cmelak, E.F. Donnelly, A.W. Malcolm, L. Moretti, J. Jaboin, K. Niermann, E.S. Yang, D.S. Yu, F. Yei, T. Koyama, G.X. Ding, B.M. Dawant, Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study, *Phys. Med. Biol.* 56 (2011) 4557–4577, <http://dx.doi.org/10.1088/0031-9155/56/14/021>.
- [8] J.M. Provenzale, M.C. Mancini, Assessment of intra-observer variability in measurement of high-grade brain tumors, *J. Neurooncol.* 108 (2012) 477–483, <http://dx.doi.org/10.1007/s11060-012-0843-2>.
- [9] J. Egger, T. Kapur, A. Fedorov, S. Pieper, J.V. Miller, H. Veeraraghavan, B. Freisleben, A.J. Golby, C. Nimsky, R. Kikinis, GBM volumetry using the 3D slicer medical image computing platform, *Sci. Rep.* (2013) 1–7, <http://dx.doi.org/10.1038/srep01364>.
- [10] S. Bauer, R. Wiest, L.-P. Nolte, M. Reyes, A survey of MRI-based medical image analysis for brain tumor studies, *Phys. Med. Biol.* 58 (2013) R97–R129, <http://dx.doi.org/10.1088/0031-9155/58/13/R97>.
- [11] E. Rios Velazquez, R. Meier, W.D. Dunn, B. Alexander, R. Wiest, S. Bauer, D.A. Gutman, M. Reyes, H.J.W.L. Aerts, Fully automatic GBM segmentation in the TCGA-GBM dataset: prognosis and correlation with VASARI features, *Sci. Rep.* 5 (2015) 16822, <http://dx.doi.org/10.1038/srep16822>.
- [12] R. Meier, U. Knecht, T. Loosli, S. Bauer, J. Slotboom, R. Wiest, M. Reyes, Clinical evaluation of a fully- automatic segmentation method for longitudinal brain tumor volumetry, *Sci Rep.* (2016) 1–11, <http://dx.doi.org/10.1038/srep23376>.
- [13] N. Porz, S. Bauer, A. Pica, P. Schucht, J. Beck, R.K. Verma, J. Slotboom, M. Reyes, R. Wiest, Multi-modal glioblastoma segmentation: man versus machine, *PLoS One* 9 (2014) 1–9, <http://dx.doi.org/10.1371/journal.pone.0096873>.
- [14] N. Porz, S. Habegger, R. Meier, R. Verma, A. Jilch, J. Fichtner, U. Knecht, C. Radina, P. Schucht, J. Beck, A. Raabe, J. Slotboom, M. Reyes, R. Wiest, Fully automated enhanced tumor compartmentalization: man vs. machine reloaded, *PLoS One* 11 (2016) e0165302, <http://dx.doi.org/10.1371/journal.pone.0165302>.
- [15] M. Lacroix, D. Abi-Said, D.R. Fournay, Z.L. Gokaslan, W. Shi, F. DeMonte, F.F. Lang, I.E. McCutcheon, S.J. Hassenbusch, E. Holland, K. Hess, C. Michael, D. Miller, R. Sawaya, A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival, *J. Neurosurg.* 95 (2001) 190–198, <http://dx.doi.org/10.3171/jns.2001.95.2.0190>.
- [16] M. Reuter, E.R. Gerstner, O. Rapalino, T.T. Batchelor, B. Rosen, B. Fischl, Impact of MRI head placement on glioma response assessment, *J. Neurooncol.* 118 (2014) 123–129, <http://dx.doi.org/10.1007/s11060-014-1403-8>.
- [17] R. Meier, N. Porz, U. Knecht, T. Loosli, P. Schucht, J. Beck, J. Slotboom, R. Wiest, M. Reyes, Automatic estimation of extent of resection and residual tumor volume of patients with glioblastoma, *J. Neurosurg.* (2017) 1–9, <http://dx.doi.org/10.3171/2016.9.jns16146>.
- [18] B.M. Ellingson, M. Bendszus, J. Boxerman, D. Barboriak, B.J. Erickson, M. Smits, S.J. Nelson, E. Gerstner, B. Alexander, G. Goldmacher, W. Wick, M. Vogelbaum, M. Weller, E. Galanis, J. Kalpathy-Cramer, L. Shankar, P. Jacobs, W.B. Pope, D. Yang, C. Chung, M.V. Knopp, S. Cha, M.J. van den Bent, S. Chang, W.K. Yung, T.F. Cloughesy, P.Y. Wen, M.R. Gilbert, Jumpstarting Brain Tumor Drug Development Coalition Imaging Standardization Steering Committee, Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials, *Neuro Oncol.* 17 (9) (2015) 1188–1198, <http://dx.doi.org/10.1093/neuonc/nov095>.