Published in final edited form as:

Eur Radiol. 2009 March ; 19(3): 599–609. doi:10.1007/s00330-008-1191-7.

Reliability of tumor volume estimation from MR images in patients with malignant glioma. Results from the American College of Radiology Imaging Network (ACRIN) 6662 Trial

Birgit B. Ertl-Wagner,

Institute of Clinical Radiology, University of Munich— Grosshadern Campus, Marchioninistrasse 15, 81377 Munich, Germany, e-mail: Birgit.Ertl-Wagner@med.uni-muenchen.de, Tel.: +49-89-70953620, Fax: +49-89-70958832

Jeffrey D. Blume,

Center for Statistical Sciences, Brown University, Providence, RI, USA

Donald Peck,

Department of Radiology, Henry Ford Hospital, Detroit, MI, USA

Jayaram K. Udupa,

Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA

Benjamin Herman,

Center for Statistical Sciences, Brown University, Providence, RI, USA

Anthony Levering,

American College of Radiology Imaging Network, Philadelphia, PA, USA

Ilona M. Schmalfuss, and

Department of Radiology, University of Florida-Gainesville, Gainesville, FL, USA

The members of the ACRIN 6662 study group

Abstract

Reliable assessment of tumor growth in malignant glioma poses a common problem both clinically and when studying novel therapeutic agents. We aimed to evaluate two software-systems in their ability to estimate volume change of tumor and/or edema on magnetic resonance (MR) images of malignant gliomas. Twenty patients with malignant glioma were included from different sites. Serial post-operative MR images were assessed with two software systems representative of the two fundamental segmentation methods, single-image fuzzy analysis (3DVIEWNIX-TV) and multi-spectral-image analysis (Eigentool), and with a manual method by 16 independent readers (eight MR-certified technologists, four neuroradiology fellows, four neuroradiologists). Enhancing tumor volume and tumor volume plus edema were assessed independently by each reader. Intraclass correlation coefficients (ICCs), variance components, and prediction intervals were estimated. There were no significant differences in the average tumor volume change over time between the software systems ($p > 0.05$). Both software systems were much more reliable and yielded smaller prediction intervals than manual measurements. No significant differences were observed between the volume changes determined by fellows/neuroradiologists or technologists. Semi-automated software systems are reliable tools to serve as outcome parameters in clinical studies and the basis for therapeutic decision-making for malignant gliomas, whereas manual measurements are less reliable and should not be the basis for clinical or research outcome studies.

Introduction

In the United States, approximately 17,000 patients are affected by primary brain tumors each year [1]. Approximately 60% of these are gliomas, with the majority belonging to the high-grade or WHO grade III and IV category. The median survival rate of a patient diagnosed with a high-grade glioma is still under 12 months. Cure rates are exceedingly low [2]. Despite marked advances in surgery, chemotherapy, radiation therapy, and diagnostic radiology, survival rates in patients with high-grade gliomas appear largely unaffected [3-5]. There is a pronounced need for new, effective therapeutic approaches to brain tumors. Magnetic resonance (MR) imaging is a decisive factor in the assessment of the efficacy of new therapeutic strategies and is, therefore, a paramount factor in the design of brain tumor therapy trials. MR imaging moreover plays a key role in the clinical evaluation of brain tumor patients both for the primary diagnostic assessment and for surgical planning as well as post-operative follow-up.

The quantitative assessment of the tumor volume [6-8] is required for measurement of tumor response to a therapy and may also serve as a surrogate end-point in randomized clinical trials. (e.g., phase II trials). In addition, a quantitative assessment of the tumor volume also serves as a guide for the clinician and the patient in therapeutic decision-making both in clinical and in study settings [8].

In the year 2000, guidelines to evaluate the response to treatment in solid tumors were published by a task force set up by the National Cancer Institute (NCI) of the United States, the European Organization for Research and Treatment of Cancer (EORTC) and the National Cancer Institute of Canada Clinical Trials Group: the Response Evaluation Criteria in Solid Tumors (RECIST) [6]. RECIST guidelines recommend a manual metric measurement of the longest diameter of the lesion by use of a ruler or calipers. If more than one lesion is present, the sum of the longest diameters of all target lesions should be calculated [6]. The RECIST criteria have the advantage of being an easy, fast and readily available tool in the evaluation of tumor response [7]. They, however, solely focus on a monodimensional measurement and do not take into account the three-dimensionality and the complex and heterogeneous structure of malignant gliomas. Moreover, manual measurements can be expected to be highly operator-dependent and may vary widely, depending on the reporting neuroradiologist. Thus, a reliable volumetric assessment of the tumor volume and of the volume of the surrounding T2 signal abnormality is desirable.

In the past, MR imaging-based tumor-volume assessment with manual segmentation was cumbersome and frequently unreliable. Previous reports have demonstrated rather large inter- and intra-rater variations with other segmentation techniques and have reported a case dependency for the accuracy of measurements [9]. However, the effects of reader expertise and previous surgical treatment upon the inter- and intra-rater variability with computer-assisted techniques have not yet been reported. It is necessary to evaluate the variation of segmentation systems before their application in brain tumor therapy trials becomes feasible.

In recent years, computer-assisted systems for the assessment of brain tumor volumes have been developed, in which the reader marks the area of interest, while the system segments the various imaging regions and computes the respective volume.

Although there are many computer-assisted methods that have been developed, this study utilizes two representative methods, Eigentool and 3DVIEWNIX-TV [10-16]. The two software systems were selected due to the following reasons. First, the methods implemented in the two systems are representative of two classes of techniques available for image

segmentation, particularly for delineating tumors in MR images. These two classes are single-image fuzzy analysis (3DVIEWNIX-TV) and multi-spectral-image analysis (Eigentool). Second, these techniques have been published and technically evaluated [17-19]. Third, the software systems have been used in tumor volume estimation and other volume measurement applications independently of this project [17-19].

Both systems minimize operator dependence on tumor-volume assessment. However, the extent of interobserver variability for the assessment of tumor volume remains to be demonstrated. Moreover, the extent of the influence of the reader's expertise and training level on the reproducibility remains unclear. Such knowledge will aid in the design of multi-center studies, as the relevant level of reader training can be chosen beforehand.

The overall goal of this study was to assess and explore the value of two computer-assisted methods for estimating the volumes of the contrast enhancing tissue and the hyperintensity seen on FLAIR images (i.e., the perifocal edema) in patients with recurrent malignant glioma in a multi-center multi-reader study.

Materials and methods

Imaging, case selection and stratification

All institutions submitting images underwent study-specific IRB approval. A copy of the IRB approval for each institution was filed at ACRIN headquarters. Moreover, the study protocol was approved by the Cancer Therapy Evaluation Program (CTEP) of the National Cancer Institute (NCI).

The following inclusion criteria had to be met in order to be eligible for the study:

- A histologically proven diagnosis of WHO grade III or IV glioma, including glioblastoma multiforme, anaplastic astrocytoma, anaplastic oligodendroglioma.
 - Two subsequent post-operative scans demonstrating tumor recurrence; these had to be acquired at the first and at the second follow up 3 months and 6 months post surgery, respectively.
 - All cases had to be post-operative cases, in which tumor tissue had been removed with a macroscopically complete R0 resection.
 - All cases had to demonstrate a tumor recurrence that was substantiated either by clinical follow-up, histological evidence, or FDG-PET.

Images were obtained on 1.5-T MR imaging systems. All studies included T1-weighted sequences before and after the intravenous administration of a gadolinium (Gd)-based contrast medium (TR $\approx$ 500 ms, TE = minimum full, slice thickness 3.0 mm, matrix 256 $\times$ 192 mm, field of view 200 $\times$ 200 mm; standard dosage of a Gd-based contrast medium applied as a bolus injection, i.e., 0.1 mmol (= 0.2 ml) Gd-DTPA per kg body weight, applied as a bolus injection), a FLAIR sequence (TR $\approx$ 10,000 ms, TE $\approx$ 150 ms, TI $\approx$ 2,200 ms, slice thickness 3.0 mm, matrix 256 $\times$ 192 mm, field of view 200 $\times$ 200 mm), and a T2-weighted sequence (TR $\approx$ 3,500 ms, TE $\approx$ 90 ms, slice thickness 3.0 mm, matrix 256 $\times$ 192 mm, field of view 200 $\times$ 200 mm), all obtained in an axial plane. No interslice gaps and no interleaved acquisitions were used.

Eligible cases were identified through the databases of the submitting sites and consecutive cases were selected from among those with complete inclusion criteria. An excess of 87 cases was submitted. Cases were evaluated by a panel of three board-certified neuroradiologists (advisory panel) regarding difficulty of assessment (i.e., presence or absence of necrosis and complexity of shape) and amount of change of tumor volume. The cases were graded on a scale

from 1 to 10 regarding the above-mentioned criteria. Eight cases were randomly selected within each of three strata formed by the Biostatistics Center (BC): the first with minimal change in tumor volume and considered simple, the second was designated moderate, and the third was designated difficult with a large change in tumor volume and complex shape.

Reader selection and reading set-up

A total of 16 readers were recruited for the study; eight were radiology technologists, four neuroradiology fellows, and four staff neuroradiologists. All reading was conducted at the American College of Radiology Imaging Network (ACRIN) headquarters in Philadelphia, Pa. Two ACRIN representatives were trained on both Eigentool and 3DVIEWNIX-TV for the purposes of training the study readers in both techniques.

Each reader was offered five training cases prior to entering the evaluation stage of the study. With these five training cases, readers were instructed to practice the applications and to perform volume measurements both for the Gd-enhancing tumor volume and for the FLAIR hyperintensity under the surveillance of the two ACRIN representatives. Practice with the training cases was performed for all three modalities, i.e., for Eigentool, 3DVIEWNIX-TV and for manual measurements, by all readers. The readers entered the evaluation stage of the study only after they felt comfortable with the usage of all three modalities. The evaluation stage of the study consisted of 20 cases with two time-points evaluated with each system, resulting in a total of 120 dataset evaluations totaling a full week of readings per reader.

Image processing and volume determination

The volume of the enhancing tissue and the hyperintensity seen on FLAIR images was determined at each time-point for all cases with all modalities. To reduce heterogeneity and save time, all image sets from a single acquisition time-point were pre-registered by using 3DVIEWNIX-TV or Eigentool. For volumetric assessment with Eigentool, registration of all images (pre- and post-T1, T2, FLAIR) was required. For volumetric assessment with 3DVIEWNIX-TV, registration of only the pre- and post-contrast T1 images was necessary. Since 3DVIEWNIX-TV uses FLAIR images on their own, no registration is required for these images. The 3DVIEWNIX-TV software does not use T2 images in the analysis. Image processing for 3DVIEWNIX-TV (www.mipg.upenn.edu) was performed on a 600 MHz Pentium PC using the Linux operating system. For Eigentool, image processing was performed by using a SUN SPARC Workstation with the Solaris operating system (see www.radiologyresearch.org, a Windows XP version is also available). The complete steps used to segment tissue can be found on the respective software websites.

In addition to the volume determination with both software systems, the Gd enhancement and the FLAIR hyperintensity were measured manually. Since the volume of the lesion is being determined in this study, and the RECIST criteria only measure the longest diameter of the lesion, each reader identified the cross sectional image with the largest linear tumor dimensions on the FLAIR and on the contrast-enhanced T1-weighted images. The longest lesion diameter for each lesion component was measured in the two orthogonal directions. The number of sections and section spacing on which each lesion component was identified was also noted and from these values the volume of the Gd enhancement and FLAIR hyperintensity was determined based on an ellipsoid shape.

The time measured to perform the volume estimations included the uploading of the images for each modality.

Statistical considerations

The objective of the study was to assess the two systems for estimating volumes of brain tumors on MR images in patients with new, postoperative, and recurrent malignant gliomas. The primary endpoint of the study was the volume change of the Gd enhancement and FLAIR hyperintensity in milliliters (ml) for each case. The interclass correlation coefficient (ICC) was estimated for each lesion component volume and the change in these volumes over time. The ICC is defined as the ratio of variance components; specifically, the variance components due to cases divided by the sum of the variance components due to readers, cases, and error. The ICC was estimated under a standard regression model with normal errors containing a random effects term for readers and cases. Technically, this is a classic three-factor (cases, raters, and systems) linear model without replication at three-way interaction level, but the model is easily expanded into a five-factor model (experience and tumor size change added) to explore secondary aims. An analogous ICC with the variance component for readers in the numerator is also of interest and was computed.

Because ICCs are percentages of total variance, it is important to examine and assess the individual variance components themselves. Small components of variance due to readers indicate that the system is reliable over different readers. In addition, we estimated the length of a 95% prediction interval for the differences in volume measurements, which describes the variation of a prediction over the entire population of cases. For example, a prediction interval length of 20 ml, would indicate that the unknown volume measurements can appear as ± 10 ml off the true volume. This should not be confused with reliability measurements, which describe the repeatability of a specific case measurement by different readers.

The raters evaluated 20 case series, consisting of two time-points, under each system and using the manual method. The order in which images were read was randomized, separately for each system and reader, to avoid recall bias and reduce any learning effect. Note that an ICC of at least 0.60 is considered *substantial reliability*, an ICC of at least 0.75 or more is considered *excellent reliability*, and an ICC above 0.85 is considered *almost perfect reliability*. Sample size projections were based on achieving sufficiently high lower confidence bound on the estimate of the ICC [16]. Note that the study was sufficiently large to estimate the reliability in the change of tumor volumes, absolute measures of tumor volume, as well as estimation of reader specific reliability. Data were analyzed at the Center for Statistical Sciences at Brown University, which serves as the Biostatistics Center for all ACRIN trials. Data were prospectively monitored and cleaned in a collaborative effort with ACRIN data management. SAS v9.13 and Stata v7.0 were used to process data and facilitate statistical analyses.

Results

The average time to train on a software package until the operator felt comfortable with its usage was 138.6 ± 49.6 min for Eigentool and 126.7 ± 46.5 min for 3DVIEWNIX-TV. The mean time to complete a case, i.e., to estimate the volume of enhancing tumor and of tumor plus edema for a single time-point, was 14.85 ± 0.72 min for Eigentool for MR technologists and 14.67 ± 1.42 min for staff neuroradiologists and neuroradiology fellows (i.e., Staff/Fellows). For 3DVIEWNIX-TV, the mean time to complete a case was 18.88 ± 1.78 min for MR technologists and 14.68 ± 0.71 min for Staff/ Fellows. Operating times to estimate tumor volumes for a case overall were not significantly different between the two software systems ($p > 0.05$). The mean time to complete a case using the manual method was 8.94 ± 1.48 min, with no significant differences between MR technologists (8.87 ± 1.11 min) and Staff/Fellows (9.01 ± 2.74 min).

The actual volume change determined for the Gd enhancement and FLAIR hyperintensity for all cases and each reader is presented as box plots in Figs. 1 and 2. Overall there were no

significant differences in the average volume change determined by the two software systems ($p>0.05$) for both components of the lesion. On the other hand, significant differences were observed between the results from the software systems and the manual measurements both for the volume of the FLAIR hyperintensity and for the volume of the Gd-enhancing part of the lesion ($p<0.001$).

The variances and ICCs of the estimated volume change are given in Tables 1 and 2. The total variance was significantly larger for manual measurements when compared with either software system ($p<0.01$); i.e., the variance for the manual measurements were from three-times to an order of magnitude larger than that from either of the two software systems. But note that the proportion of variance attributable to the different readers is comparatively small for all methods used.

The length of 95% prediction intervals, which are a direct function of the total variance, corresponded well with the amounts of variance described above (see Figs. 1, 2 and Table 3). Prediction intervals were significantly lower for both software systems when compared with manual measurements ($p<0.01$). For instance, for estimating the volume differences over time, in our population, the length of the 95% prediction interval was about 185 ml for the Gd-enhancing part of the lesion and 312 ml for the hyperintensity in the FLAIR sequence when using the manual method. The 95% prediction interval lengths were 52 ml and 69 ml for the Gd-enhancing part of the lesion and 180 ml and 100 ml for the hyperintensity in the FLAIR sequence for 3DVIEWNIX-TV and Eigentool, respectively.

The volume estimates were not significantly different between Staff/Fellows and MR technologists for either software system ($p>0.05$). This held true for all levels of difficulty of the cases. In Figs. 1 and 2 there appear to be differences in the volume change measured by the two software systems for several cases (see case 14 in Fig. 1 and cases 5, 14, 18 and 19 in Fig. 2). In all but one of these cases (case 5), the cases were designated as difficult ("hard"). Note that even though there is an apparent difference in volume change estimates between the software systems seen in the figures, overall these differences were not statistically significantly. (Fig. 3).

No significant differences were found between the lengths of the 95% prediction intervals relating to the level of professional expertise of the readers for 3DVIEWNIX-TV for the volumes of either Gd enhancement or FLAIR hyperintensity ($p>0.05$, see Table 4). For Eigentool, the length of the 95% prediction interval was longer for Staff/Fellows compared with MR technologists for the volumes of Gd enhancement, while there were no significant differences for the volumes of FLAIR hyperintensity ($p>0.05$). Similarly, the sum of variances was not significantly different between level of professional expertise of the readers for 3DVIEWNIX-TV and for the volumes of FLAIR hyperintensity for Eigentool ($p>0.05$), while it was higher for Staff/Fellows when estimating the volumes of Gd enhancement for Eigentool ($p<0.01$). Also note that the sum of variances for 3DVIEWNIX-TV is larger on average for the FLAIR hyperintensity compared with Eigentool. The ICC was not significantly different between the different levels of professional expertise for either software system for estimating the volumes of Gd enhancement ($p>0.05$). Although they were higher for 3DVIEWNIX-TV when estimating the volume of FLAIR hyperintensity compared with Eigentool or the manual method. As seen in the analysis of the combined data, the results using the manual method were less reliable and the prediction intervals substantially larger compared with the software systems for all readers regardless of expertise and case difficulty.

Discussion

In this multi-center/multi-reader study, we aimed to investigate the value of two software systems in estimating volumes of the enhancing portion of malignant brain tumors and of peritumoral edema in order to assess their potential to serve as the basis for evaluating new therapeutic strategies and also as a valid clinical tool in the follow-up of patients. As it was impossible to conceive a study design in which an “absolute”—or even a relative—truth could serve as a reference standard, we evaluated parameters of reliability as the primary aim of our studies.

There are several approaches to describe the reliability of a system. One of these is the ICC, which is a proportional coefficient that is influenced by the variability due to cases, the variability due to readers and the variability due to residual or “random” error [20,21]. The higher the proportion of total variability due to cases, the higher the ICC becomes. Conversely, the higher the proportion of the variability attributed to readers or to residual error is, the lower the ICC becomes. Therefore, in order to correctly interpret our results, it is important to examine the relative proportions of total variability due to each variance component. The total variances were considerably and significantly lower for both software systems compared with manual measurements, which indicates that the nature of the software systems is much less variable than manual measurements across the clinical population. In fact, the total amount of variability is up to tenfold higher for the manual measurements. This implies that these software systems can be used to reliably measure lesion volume changes in future clinical or research studies, much more so than manual measurements.

Another parameter aiding in the evaluation and interpretation of variability of volume measurements is the assessment of prediction intervals. These intervals indicate 95% of the observed volumes that may appear from an unknown volume. While the mean of the intervals may be different (depending on a specific case or clinical indication), their length will remain constant. Hence, their length is key for understanding the case-to-case variability in this population. Compared with the software methods, the manual measurements had prediction intervals that were twice as large, suggesting that using manual measurements would require a larger volume change to be present to reliably measure the change.

If, on the other hand, the desired measure is that the volume change determined by different readers at multiple sites is always the same, the prediction interval is not relevant, but instead the reader variance is important. This study clearly demonstrates that using semi-automated software systems the reader variance is extremely small and the prediction intervals are smaller than when using manual measurements, thus making them more reliable overall.

Our study demonstrated comparable results with MR technologists and Staff/Fellows as operators. The 95% prediction intervals were somewhat larger for the Staff/ Fellow for the Gd-enhancing volume with Eigentool. This can be partially explained by the principles of segmentation of the software systems we used in the study. Eigentool segments the signal from tissue into portions of “pure” and partial volume regions. The resultant segmented images are displayed as gray-scale images, where the pixel intensity is proportional to the partial volume (see Fig. 4a). Variations in slice thickness and orientation as they occur in multi-center studies are therefore expected to result in varying partial volume portions. Therefore, the gray-scale presentation of the partial volume segmentation is expected to vary to a certain degree the interpretation between different readers when performing the final lesion designation step in the processing. On the other hand, 3DVIEWNIX-TV, while handling varying partial volume effects and tissue heterogeneity by using fuzzy topological principles, displays the final segmentation results in a binary fashion, making the final lesion designation step simpler. This would explain the coinciding behavior of the relative percent variability from Eigentool and

the manual measurements, since they both use gray-scale images in defining the regions compared with 3DVIEWNIX-TV, which utilizes a binary image for the final review.

There are limitations of our study that need to be taken into account when interpreting the data. The major limitation of any study assessing the value of volumetric estimations of brain tumors is the lack of a reference standard. It is impossible to obtain an “absolute truth” of human brain tumor volumes in an in-vivo situation. Even if surgery is performed directly after MR imaging, it is not feasible to resect the tumor in a fashion that would allow a sensible volumetric assessment. Moreover, the surrounding hyperintensity in the FLAIR or T2-weighted images usually cannot be completely resected. Histopathological studies have, however, demonstrated this zone of abnormal signal intensity to contain viable tumor cells and to therefore also play a critical role in the prognosis of the patient [22-23]. The volumetric assessment of a malignant brain tumor should, therefore, not only include the zone of enhancement after the intravenous injection of a Gd-based contrast medium, but also the volume of the FLAIR hyperintensity. Moreover, even the correlation with an autopsy cannot be expected to provide reliable results due to the very rapid growth dynamic of malignant gliomas [24-25]. As it was impossible to design a study, in which an “absolute”—or even a surrogate—truth could serve as a reference standard, we decided to evaluate parameters of reliability as the primary aim of our studies.

In addition, the sample size of our population was relatively small. This was, however, counterbalanced by a large number of readers and data stratification from a larger sample. Inclusion of a larger sample into the reading phase of the study would have been prohibitive due to the extensive reading times. Sample size projections had been performed and the study was sufficiently large to estimate the reliability in the change of tumor volumes, absolute measures of tumor volume, as well as estimation of reader specific reliability.

In summary, we were able to demonstrate in our multi-center multi-reader study that the reliability parameters of both software-based tumor-volume estimations were far superior to manual measurements. We, therefore, conclude that semi-automated software systems similar to those used in this study are reliable tools to serve as outcome parameters in clinical studies and as a basis for therapeutic decision-making in patients with malignant glioma, whereas manual measurements are less reliable and should not be the sole basis for clinical or research outcome studies. In addition, since MR technologists and Staff/ Fellow neuroradiologists demonstrated comparable results, the way is paved for a more economical and time-efficient study design, especially in large multi-center trials. While the need for neuroradiologists to supervise image analysis will certainly not be obviated, the time-consuming volume analysis can be reliably performed by trained MR technologists, thus reducing costs and time constraints.

Acknowledgements

This work was supported by the American College of Radiology Imaging Network (grant number CA80098-05). We would like to acknowledge Sharlene Snowdon for excellent technical assistance. In addition we would like to sincerely acknowledge all our readers and the members of the consensus panel, namely Suresh Patel, M.D., Jeffrey Kochan, M.D., Yair Safriel, M.D., Gagandeep Mangat, M.D., Jeff Bennett, M.D., Eric Lopez Del Valle, M.D., John Corrigan, M.D., Varun Nitroo, M.D., Chad Holder, M.D., Phil Kousourbris, M.D., Scott G. Hudson, R.T., Lisa M. Desiderio, R.T., Terence Slyman, R.T., Mary Ellen Bentham, R.T., Tony Festa, R.T., Glenn Ferrick, R.T., Maureen Perachio, R.T., Tom Hodge, R.T.

References

1. Landis SH, Murray T, Bolden S, Wingo PA. Cancer statistics. CA Cancer J Clin 1999;49:8–31. [PubMed: 10200775]

2. Curran WJ Jr, Scott CB, Horton J, et al. Recursive partitioning analysis of prognostic factors in three Radiation Therapy Oncology Group malignant glioma trials. *J Natl Cancer Inst* 1993;85:704–710. [PubMed: 8478956]
3. Walker MD, Alexander E Jr, Hunt WE, et al. Evaluation of BCNU and/or radiotherapy in the treatment of anaplastic gliomas. A cooperative clinical trial. *J Neurosurg* 1978;49:333–343. [PubMed: 355604]
4. Walker MD, Green SB, Byar DP, et al. Randomized comparisons of radiotherapy and nitrosoureas for the treatment of malignant glioma after surgery. *N Engl J Med* 1980;303:1323–1329. [PubMed: 7001230]
5. Fine HA, Dear KB, Loeffler JS, Black PM, Canellos GP. Meta-analysis of radiation therapy with and without adjuvant chemotherapy for malignant gliomas in adults. *Cancer* 1993;71:2585–2597. [PubMed: 8453582]
6. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *J Natl Cancer Inst* 2000;92:205–216. [PubMed: 10655437]
7. Husband JE, Schwartz LH, Spencer J, et al. Evaluation of the response to treatment of solid tumours—a consensus statement of the International Cancer Imaging Society. *Br J Cancer* 2004;90:2256–2260. [PubMed: 15150551]
8. Therasse P. Measuring the clinical response. What does it mean? *Eur J Cancer* 2002;38:1817–1823. [PubMed: 12204662]
9. Vaidyanathan M, Clarke LP, Velthuisen RP, et al. Comparison of supervised MRI segmentation methods for tumor volume determination during therapy. *Magn Reson Imaging* 1995;13:719–728. [PubMed: 8569446]
10. Peck DJ, Windham JP, Soltanian-Zadeh H, Roebuck JR. A fast and accurate algorithm for volume determination in MRI. *Med Phys* 1992;19:599–605. [PubMed: 1508098]
11. Peck DJ, Windham JP, Emery LL, Soltanian-Zadeh H, Hearshen DO, Mikkelsen T. Cerebral tumor volume calculations using planimetric and eigenimage analysis. *Med Phys* 1996;23:2035–2042. [PubMed: 8994168]
12. Jacobs MA, Knight RA, Soltanian-Zadeh H, et al. Unsupervised segmentation of multiparameter MRI in experimental cerebral ischemia with comparison to T2, diffusion, and ADC MRI parameters and histopathological validation. *J Magn Reson Imaging* 2000;11:425–437. [PubMed: 10767072]
13. Soltanian-Zadeh H, Peck DJ, Windham JP, Mikkelsen T. Brain tumor segmentation and characterization by pattern analysis of multispectral NMR images. *NMR Biomed* 1998;11:201–208. [PubMed: 9719574]
14. Udupa JK, Wei L, Samarasekera S, Miki Y, van Buchem MA, Grossman RI. Multiple sclerosis lesion quantification using fuzzy-connectedness principles. *IEEE Trans Med Imaging* 1997;16:598–609. [PubMed: 9368115]
15. Udupa JK, Herman GT. Medical image reconstruction, processing, visualization, and analysis: the MIPG perspective. Medical Image Processing Group. *IEEE Trans Med Imaging* 2002;21:281–295. [PubMed: 12022617]
16. Moonis G, Liu J, Udupa JK, Hackney DB. Estimation of tumor volume with fuzzy-connectedness segmentation of MR images. *AJNR Am J Neuroradiol* 2002;23:356–363. [PubMed: 11900999]
17. Udupa JK, Nyul LG, Ge Y, Grossman RI. Multiprotocol MR image segmentation in multiple sclerosis: experience with over 1,000 studies. *Acad Radiol* 2001;8:1116–1126. [PubMed: 11721811]
18. Nyul LG, Udupa JK. MR image analysis in multiple sclerosis. *Neuroimaging Clin N Am* 2000;10:799–816. [PubMed: 11359726]
19. Jacobs MA, Knight RA, Windham JP, et al. Identification of cerebral ischemic lesions in rat using Eigenimage filtered magnetic resonance imaging. *Brain Res* 1999;837:83–94. [PubMed: 10433991]
20. Zou G, Donner A. Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. *Biometrics* 2004;60:807–811. [PubMed: 15339305]
21. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. *J Biomed Inform* 2002;35:99–110. [PubMed: 12474424]

22. Tovi M, Hartman M, Lilja A, Ericsson A. MR imaging in cerebral gliomas. Tissue component analysis in correlation with histopathology of whole-brain specimens. *Acta Radiol* 1994;35:495–505. [PubMed: 8086262]
23. Dean BL, Drayer BP, Bird CR, et al. Gliomas: classification with MR imaging. *Radiology* 1990;174:411–415. [PubMed: 2153310]
24. Lacroix M, Abi-Said D, Fournier DR, et al. A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. *J Neurosurg* 2001;95:190–198. [PubMed: 11780887]
25. Hobbs SK, Shi G, Homer R, Harsh G, Atlas SW, Bednarski MD. Magnetic resonance image-guided proteomics of human glioblastoma multiforme. *J Magn Reson Imaging* 2003;18:530–536. [PubMed: 14579395]

Fig. 1.

Box plot of the Gd-enhancement volume change for each case and each reader using 3DVIEWNIX-TV (orange), Eigentool (green) and manual (blue) measurements. The colored box is known as the interquartile range (IQR) and represents observations lying between the 25th and 75th percentiles of the data with the median (50th percentile) drawn as a light blue line within the box. The upper (lower) fence is drawn to the maximum (or minimum) observation within ± 1.5 IQR, with points outside the fences plotted as outlier points. The box plot describes the distribution of values and, as such, a longer colored box indicates greater variability

Fig. 2.

Plot of the FLAIR hyperintensity volume change for each case and each reader using 3DVIEWNIX-TV (orange), Eigentool (green) and manual (blue) measurements. The colored box is known as the interquartile range (IQR) and represents observations lying between the 25th and 75th percentiles of the data with the median (50th percentile) drawn as a light blue line within the box. The upper (lower) fence is drawn to the maximum (or minimum) observation within ± 1.5 IQR, with points outside the fences plotted as outlier points. The box plot describes the distribution of values and, as such, a longer colored box indicates greater variability

Fig. 3.

Screen capture using the 3DVIEWNIX-TV software. **a** A display of a slice of the FLAIR image with an overlaid *rectangular box* encompassing the tumor. **b** The segmented FLAIR hyperintense region overlaid on a display of the FLAIR image slice for operator verification (*right*). For reference, the part of the slice in the box is also shown (*left*). **c** A display of a slice of T1 (*upper left*), T1e (*upper right*), the difference (T1e-T1) image (*lower left*), and the segmented enhancing tumor (*lower right*)

Fig. 4.

Screen capture using the Eigentool software. **a** On the left the original MRI for one location is shown. The images are: a) FLAIR, b) pre-Gd T1 weighted image, c) post-Gd T1 weighted image, and d) T2 weighted image. On the right are the segmented results for this location: e) Gd enhancement, f) FLAIR hyperintensity, g) normal tissue, and h) CSF. **b** Combined segmentation color-map images. On the left are the segmented results for this location similar to **a**; i.e., a) Gd enhancement, b) FLAIR hyperintensity, c) CSF, and d) normal tissue. On the right is the combined segmentation color-map image for the same location. Note the different colors within the color-map image corresponding to the different classifications of partial volume tissue. Also shown is the ROI defining the area to be included in the volume calculation.

Note that this ROI is also displayed on the segmented images and it does not include fat, skin etc. from outside the brain

Table 1
Summary of the variance components contributing to the total variance when estimating volume differences in the FLAIR hyperintensity over time with the respective methods (*LCB* lower confidence bound, *UCB* upper confidence bound)

3DVIEWNIX-TV						
	Variance component	95% LCB	95% UCB	% total (ICC)	95% LCB	95% UCB
Cases	2,009.08	1,895.21	2,128.23	0.95	0.92	0.98
Readers	0.00	0.00	19.28	0.00	0.00	0.01
Error	102.82	38.95	156.53	0.05	0.02	0.07
Total	2111.89	—	—	—	—	—
Eigentool						
	Variance component	95% LCB	95% UCB	% total (ICC)	95% LCB	95% UCB
Cases	226.53	175.53	344.45	0.35	0.26	0.57
Readers	1.66	0.00	79.44	0.00	0.00	0.10
Error	422.40	206.33	564.02	0.65	0.42	0.69
Total	650.60	—	—	—	—	—
Manual						
	Variance component	95% LCB	95% UCB	% total (ICC)	95% LCB	95% UCB
Cases	2,782.05	1,655.03	4,790.47	0.44	0.33	0.61
Readers	0.00	0.00	425.42	0.00	0.00	0.06
Error	3,564.82	2,125.08	4,455.45	0.56	0.37	0.66
Total	6,346.87	—	—	—	—	—

Table 2
Summary of the variance components contributing to the total variance when estimating volume differences of the Gd-enhancing lesion over time with the respective methods

3DVIEWNIX-TV						
	Variance component	95% LCB	95% UCB	% total (ICC)	95% LCB	95% UCB
Cases	89.08	74.76	118.47	0.51	0.40	0.71
Readers	0.00	0.00	10.22	0.00	0.00	0.06
Error	87.05	38.58	129.27	0.49	0.28	0.59
Total	176.13	—	—	—	—	—
Eigentool						
	Variance component	95% LCB	95% UCB	% total (ICC)	95% LCB	95% UCB
Cases	76.84	26.59	196.72	0.25	0.12	0.53
Readers	0.00	0.00	43.45	0.00	0.00	0.09
Error	232.64	73.63	379.43	0.75	0.44	0.85
Total	309.47	—	—	—	—	—
Manual						
	Variance component	95% LCB	95% UCB	% total (ICC)	95% LCB	95% UCB
Cases	397.98	283.37	1,118.21	0.18	0.14	0.39
Readers	0.00	0.00	383.66	0.00	0.00	0.10
Error	1,825.26	617.52	3,225.44	0.82	0.58	0.84
Total	2,223.23	—	—	—	—	—

Table 3
Summary of the prediction intervals for estimating volume differences of the FLAIR hyperintensity and of the Gd-enhancing lesion over time with the respective methods

	FLAIR volume			Gd-enhancing volume		
	Length	95% LCB	95% UCB	Length	95% LCB	95% UCB
3DVIEWNIX-TV	180.14	112.94	228.36	52.02	41.97	60.43
Eigentool	99.99	85.67	112.50	68.96	60.94	76.14
Manual	312.30	258.92	357.80	184.83	166.21	201.75

Table 4

95% prediction intervals, sum of variances and ICCs depending on the level of professional expertise of the reader (*CI* confidence interval)

	3DVIEWNIX-TV		Eigentool		Manual	
	Staff/Fellows	Technologists	Staff/Fellows	Technologists	Staff/Fellows	Technologists
95% prediction interval (95% CI)						
Gd	51 (40, 61)	53 (42, 61)	86 (73, 98)	46 (39, 51)	211 (185, 235)	153 (134, 171)
FLAIR	181 (111, 231)	179 (115, 226)	104 (86,119)	96 (80, 109)	250 (202, 290)	364 (294, 422)
Sum of variances (95% CI)						
Gd	172 (105, 238)	181 (117, 245)	485 (349, 622)	135 (102, 169)	2,915 (2,223, 3,607)	1,534 (1,161, 1,908)
FLAIR	2,132 (798, 3,467)	2,092 (854, 3,328)	697 (478,917)	597 (418, 775)	4,064 (2,656, 5,473)	8,617 (5,620, 11615)
ICC (95% CI)						
Gd	0.53 (0.42, 0.65)	0.465 (0.31, 0.62)	0.29 (0.10, 0.48)	0.187 (0.02, 0.35)	0.145 (0.00, 0.30)	0.170 (0.07, 0.27)
FLAIR	0.98 (0.97, 0.99)	0.920 (0.88, 0.96)	0.37 (0.25, 0.50)	0.335 (0.17, 0.50)	0.445 (0.32, 0.58)	0.45 (0.31, 0.59)