OPEN

Clinical Evaluation of a Multiparametric Deep Learning Model for Glioblastoma Segmentation Using Heterogeneous Magnetic Resonance Imaging Data From Clinical Routine

Michael Perkuhn, MSc, MD,† Pantelis Stavrinou, MD,‡ Frank Thiele, MSc,*† Georgy Shakirin, PhD,*† Manoj Mohan, MSc,§ Dionysios Garmpis,* Christoph Kabbasch, MD,* and Jan Borggrefe, MD**

Objectives: The aims of this study were, first, to evaluate a deep learning–based, automatic glioblastoma (GB) tumor segmentation algorithm on clinical routine data from multiple centers and compare the results to a ground truth, manual expert segmentation, and second, to evaluate the quality of the segmentation results across heterogeneous acquisition protocols of routinely acquired clinical magnetic resonance imaging (MRI) examinations from multiple centers.

Materials and Methods: The data consisted of preoperative MRI scans (T1, T2, FLAIR, and contrast-enhanced [CE] T1) of 64 patients with an initial diagnosis of primary GB, which were acquired in 15 institutions with varying protocols. All images underwent preprocessing (coregistration, skull stripping, resampling to isotropic resolution, normalization) and were fed into an independently trained deep learning model based on DeepMedic, a multilayer, multiscale convolutional neural network for detection and segmentation of tumor compartments. Automatic segmentation results for the whole tumor, necrosis, and CE tumor were compared with manual segmentations.

Results: Whole tumor and CE tumor compartments were correctly detected in 100% of the cases; necrosis was correctly detected in 91% of the cases. A high segmentation accuracy comparable to interrater variability was achieved for the whole tumor (mean dice similarity coefficient [DSC], 0.86 ± 0.09) and CE tumor (DSC, 0.78 ± 0.15). The DSC for tumor necrosis was 0.62 ± 0.30. We have observed robust segmentation quality over heterogeneous image acquisition protocols, for example, there were no correlations between resolution and segmentation accuracy of the single tumor compartments. Furthermore, no relevant correlation was found between quality of automatic segmentation and volume of interest properties (surface-to-volume ratio and volume).

Conclusions: The proposed approach for automatic segmentation of GB proved to be robust on routine clinical data and showed on all tumor compartments a high automatic detection rate and a high accuracy, comparable to interrater variability. Further work on improvements of the segmentation accuracy for the necrosis compartments should be guided by the evaluation of the clinical relevance.

Therefore, we propose this approach as a suitable building block for automatic tumor segmentation to support radiologists or neurosurgeons in the preoperative reading of GB MRI images and characterization of primary GB.

Received for publication February 14, 2018; and accepted for publication, after revision, April 11, 2018.

From the *Department of Radiology, University Hospital Cologne, Cologne; †Clinical Applications Research, Philips Research, Aachen; ‡Department of Neurosurgery, University Hospital Cologne, Cologne, Germany; and §Data Science, Philips Healthcare, Bangalore, India.

Conflicts of interest and sources of funding: Michael Perkuhn, Frank Thiele, and Georgy Shakirin are employees of Philips Research, and Manoj Mohan is an employee of Philips Healthcare. For the remaining authors, none were declared.

Correspondence to: Michael Perkuhn, MSc, MD, Department of Radiology, University Hospital Cologne, Kerpener Str 62, 50937 Köln, Germany. E-mail: michael.perkuhn@uk-koeln.de.

Copyright © 2018 The Author(s). Published by Wolters Kluwer Health, Inc. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

ISSN: 0020-9996/18/5311–0647
DOI: 10.1097/RLI.0000000000000484

Key Words: glioblastoma, GB, MRI, tumor segmentation, machine learning, deep learning

(*Invest Radiol* 2018;53: 647–654)

Glioblastoma (GB) is the most frequent primary brain cancer.[1] The diffuse and highly invasive growth as well as the intratumor heterogeneity makes it the most lethal cancer of the central nervous system.[2] Despite the aggressive combination therapy with surgery followed by radiation plus concomitant and adjuvant chemotherapy, the median survival time is still only 15 to 17 months.[3] Magnetic resonance imaging (MRI) is commonly used to evaluate location, size, spread, edema, and the biological status of the tumor noninvasively.[4] Magnetic resonance imaging is part of the standard clinical workup for GB management for planning and follow-up of surgery, chemotherapy, and radiation.

Detection of the tumor and determination of location and extension of the different tumor compartments are important for surgery planning. It has been shown in large studies that the extent of the resection of tumor volume of 98% or more is associated with longer survival time.[5] Furthermore, Hammoud et al[6] showed that MRI features such as sign of little or no necrosis and lower tumor enhancement are associated with longer survival time. Determination of the tumor extent is also highly relevant to radiomics,[7] an emerging imaging-based method for extracting quantitative features from standard-of-care imaging to establish predictive models.[8] Furthermore, a better identification of the different biological areas of the tumor improves the precision of tissue targeting in biopsy.

In radiological reading, tumor and compartment segmentation is not done routinely but would offer important additional clinical value in the aforementioned approaches. Manual expert segmentation is still considered as the gold standard, but it is time-consuming and highly variable based on the level of expertise.[9] Consequently, there remains an unmet need for fully automatic, user-independent detection and segmentation tools, with the potential to become an integral part of the clinical reading workflow.

Over the last decade, significant progress has been made in computer-assisted and machine learning–based segmentation algorithms for the identification and segmentation of brain lesions. Several main categories of approaches can be identified. There are semiautomatic and fully automatic approaches of various complexity, from simple thresholding or region growing algorithms to comprehensive model-based, supervised, and unsupervised machine learning algorithms.[10,11] To compare the different approaches to brain tumor segmentation algorithms objectively, the Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) challenge has been organized since 2012.[12] In the latest BRATS benchmark for pretreatment segmentation of brain tumors, the algorithms evaluated demonstrated results comparable with interrater variability.[13]

Recently, deep learning–based methods have shown promising and clinically relevant results.[14–16] The deep learning techniques showing

a high model capacity and the ability to learn highly discriminative features often outperform hand designed feature sets. In particular, 2-dimensional and 3-dimensional (3D) convolutional neural networks (CNNs) show promising results on clinical imaging data.[17,18] However, studies about evaluations of these algorithms using routine clinical data are still lacking. We chose DeepMedic, a 3D CNN-based algorithm for evaluation.[18] This algorithm is fully automatic and achieved high scores on the BRATS data. Its application is based on the most routinely performed MRI examinations: T1-weighted (T1w), CE T1-weighted (CE T1w), T2-weighted (T2w), and FLAIR.

The aims of this study were as follows:

1. To evaluate a state-of-the-art, fully automatic brain tumor segmentation compared with manual annotations by expert readers.
2. To evaluate variations of the segmentation results with highly heterogeneous clinical MRI examinations performed in multiple institutions, using different acquisition protocols and scanners from different vendors.

MATERIALS AND METHODS

Study Population

In this retrospective study, we included consecutive patients with newly diagnosed, supratentorial GBs, eligible for surgical resection, referred to and treated in our institution between 2010 and 2014. Patients with infratentorial and secondary GBs were excluded.

A total of 64 patients with biopsy-proven GB and completed preoperative MR examinations (T1w, T2w, FLAIR, and CE T1w) were reviewed.

Patients' characteristics are summarized in Table 1. The study was approved by the Local Research Ethics Commission with waved informed consent.

Magnetic Resonance Imaging

Magnetic resonance images were acquired on 8 different scanner types at a total of 15 institutions. Thirty-four of 64 MR examinations have been conducted at our institution. The remaining 30 MR examinations have been conducted at the referring 14 institutions. The majority of images were acquired at 1.5 T field strength. Details of the acquisitions are shown in Table 1. T1w, T2w, FLAIR, and CE T1w series were acquired according to standard clinical acquisition protocols for the different scanners. Slice thickness ranged from 1 to 8 mm in all patients, in-plane resolution ranged from 0.3 to 1 mm (Table 1).

Manual Segmentation

For evaluation of the deep learning model, ground truth tumor compartments were delineated manually. This was performed using a semiautomatic approach with subsequent manual editing (IntelliSpace Discovery; Philips Healthcare, Best, the Netherlands), both performed by a radiologist and a senior neuroradiologist in a consensus reading. In addition, the segmented volumes of interest (VOIs) were compared with the volumes segmented independently by neurosurgeons using the iPlan software (Brainlab GmbH, Feldkirchen, Germany). In case of discrepancies, the segmentation was reviewed until consensus was reached.

The procedure followed the BRATS challenge.[13] Volumes of interest were created for (*a*) the whole tumor, (*b*) contrast-enhancing tumor, and (*c*) tumor necrosis. The whole tumor VOI was segmented on the T2w and FLAIR sequences, including the contrast-enhancing and tumor necrosis compartment. Contrast-enhancing tumor and necrosis were delineated on the CE T1w series.

Workflow for Automatic Segmentation Pipeline

The overall workflow for automatic tumor segmentation was built as shown in Figure 1. Four series (T1w, T2w, FLAIR, and CE T1w) were first preprocessed, then automatic tumor segmentation was performed using the trained deep learning model, followed by postprocessing of the output VOIs. The whole workflow was performed completely automatically.

TABLE 1. Patients' and Imaging Characteristics

Patients Characteristics	
Median age (range), y	64 (28–86)
Sex (percentage)	39 Male (61%), 25 Female (39%)
Median survival (range), mo	15.6 (1.5–48.9)
No. involved institutions	15
MRI characteristics	
Scanner model	Philips Achieva, 7 (10.9%) Philips Gyroscan, 2 (3.1%) Philips Intera, 42 (65.6%) Philips Panorama, 1 (1.6%) Siemens Aera, 2 (3.1%) Siemens Avanto, 4 (6.3%) Siemens Espree, 2 (3.1%) Siemens Magnetom, 1(1.6%) Siemens SymphonyTim, 3 (4.7%)
Field strength	1 T (2), 1.5 T (57), 3 T (5)
Mean (range) pixel dimensions, mm^3	
Contrast-enhanced T1w	0.70 (0.36–1.00) × 0.70 (0.36–1.00) × 6.1 (1.0–8.0)
T1w	0.71 (0.36–1.00) × 0.71 (0.36–1.00) × 6.2 (1.0–8.0)
T2w	0.48 (0.30–1.00) × 0.48 (0.30–1.00) × 6.1 (2.0–8.0)
FLAIR	0.80 (0.45–0.98) × 0.80 (0.45–0.98) × 6.3 (3.0–8.0)

Mean pixel dimension was calculated as an average pixel size for each direction of the image coordinate system (x, y, z) for all subjects.

© 2018 Wolters Kluwer Health, Inc. All rights reserved.

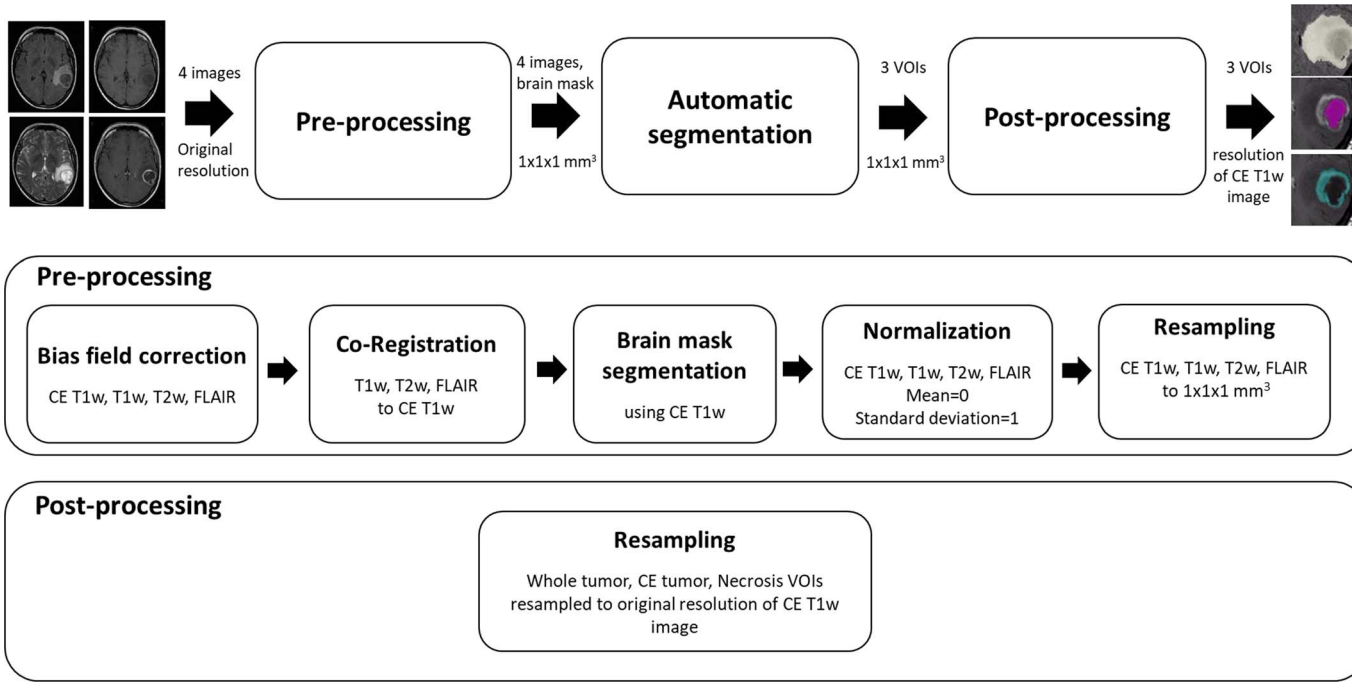

FIGURE 1. Overall workflow of the automatic brain tumor segmentation pipeline.

Image Processing

Magnetic resonance images were preprocessed with established tools (SPM8: Statistical Parametric Mapping software package version 8; Wellcome Trust Centre for Neuroimaging, London, United Kingdom; Intellispace Discovery; Philips Healthcare, Best, the Netherlands) before feeding into automatic segmentation. The preprocessing pipeline is shown in Figure 2. All 4 series were bias field corrected. Then, in a second step, the corrected T1w, T2w, and FLAIR images were coregistered to the

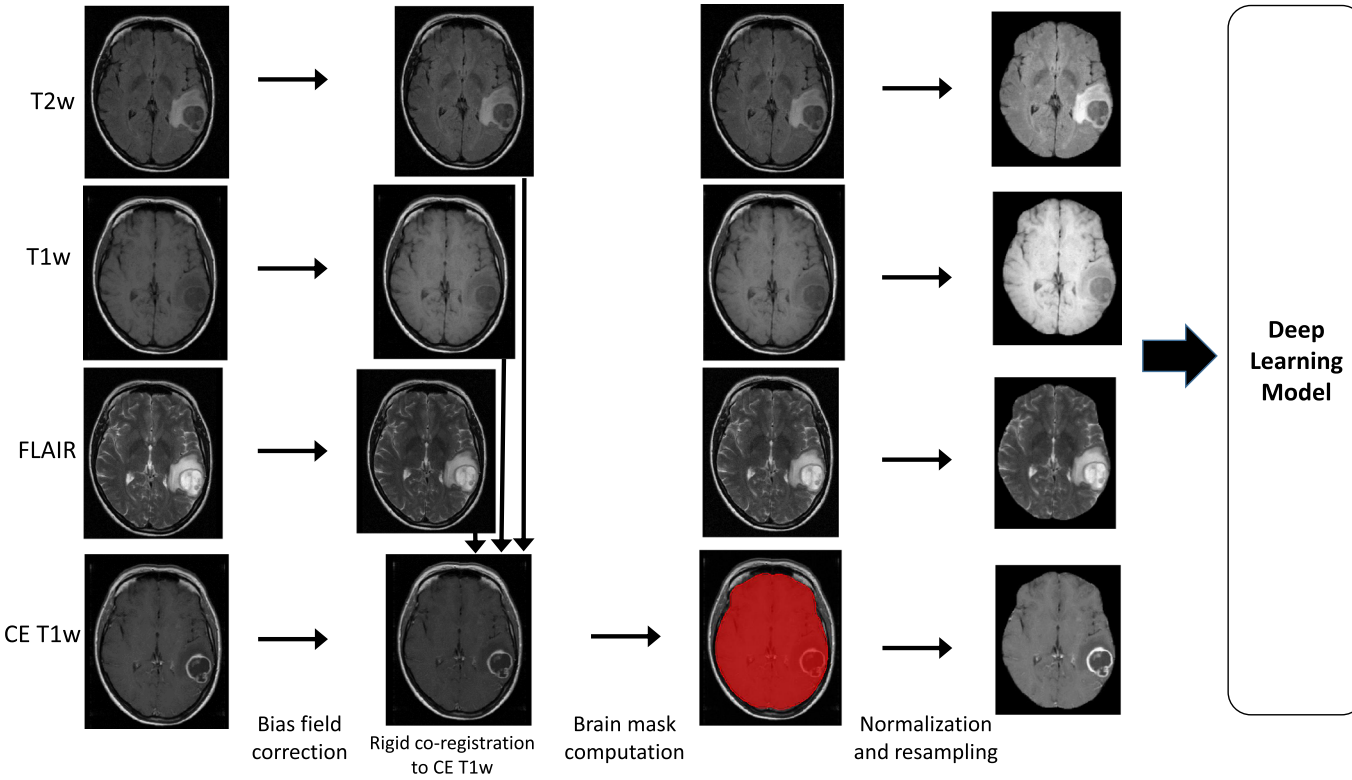

FIGURE 2. Image preprocessing pipeline: (1) Bias field correction was applied to all 4 sequences; (2) T2, T1, and FLAIR images were coregistered to T1 CE image; (3) Brain mask was computed on T1 CE image and propagated to the registered T2, T1, and FLAIR images; (4) All 4 images were scaled (zero mean and standard deviation of 1) and resampled to resolution of $1 \times 1 \times 1$ mm^3.

© 2018 Wolters Kluwer Health, Inc. All rights reserved.

reference space defined by the CE T1w series. Consequently, the ground truth VOIs for contrast-enhancing tumor and necrosis were, by definition, in the same reference space. In the third step, the manual segmentation of the whole tumor VOI was aligned to the reference space using the 6-parameter transformation obtained from the FLAIR image. Then, a brain mask was computed (SPM8 "New Segmentation"[19]) and applied to obtain skull-stripped images. Finally, images were normalized to zero-mean and standard deviation of 1 and resampled to isotropic resolution of $1 \times 1 \times 1$ mm^3.[18] The image processing pipeline was executed fully automatically without user interaction. Processing results were visually checked for quality control.

For automatic segmentation of tumor compartments, a multiparametric deep learning model was applied to the preprocessed data.

Deep Learning Model

The deep learning model is based on the recently published DeepMedic architecture, which provided top scoring results on the BRATS data set.[18] The DeepMedic architecture was installed on a graphics processing unit server at our institution.

The model was trained on an independent data set available through the BRATS 2015 challenge. The training data consisted of 220 cases of GB with expert manual segmentations of the tumor compartments.[13] The 220 cases were split into 190 for training and 30 for validation during the training procedure. The data was preprocessed as described previously.

The DeepMedic architecture consists of a deep 3D CNN followed by a 3D fully connected network to remove false-positives. The 3D CNN includes 2 pathways that apply different image resolutions to capture characteristics of the tumor appearance at 2 different spatial ranges. Inputs to the 2 pathways are centered at the same image location, but for the second input, the image is down-sampled to a third of its original size. The model consists of an 11-layer architecture with kernels of size 3^3. The last layers of the 2 pathways have receptive fields of size 17^3 voxels. For inference, image segments of 45^3 voxels are fed into the model. Finally, a fully-connected conditional random field is applied, which has a smoothing effect.[18]

The deep learning model resulted in automatic segmentation of 4 tumor compartments (edema, contrast-enhancing tumor, necrosis, nonenhancing tumor). The whole tumor region was obtained as the union of all other segmented regions, as defined in the BRATS benchmark.[13]

Statistical Analysis

To evaluate automatic segmentation, the resulting VOIs were compared with the manual ground truth annotations. For the whole tumor, contrast-enhancing tumor, and necrosis, the VOIs were compared with respect to volume and voxel-wise accuracy. As usual, the accuracy was computed as overlap of ground truth segmentation (VOI_{gt}) and model segmentation (VOI_{model}) using the dice similarity coefficient (DSC)[20]:

$$\mathrm{DSC}(VOIgt, VOImodel) = \frac{2|VOIgt \cap VOImodel|}{|VOIgt| + |VOImodel|}$$

In addition, the sensitivity (true-positive rate) and positive predictive value (PPV) of the automatic segmentation was assessed on the voxel level using the following expressions[13]:

$$\mathrm{Sensitivity}(VOIgt, VOImodel) = \frac{|VOIgt \cap VOImodel|}{|VOIgt|}$$

$$\mathrm{PPV}(VOIgt, VOImodel) = \frac{|VOIgt \cap VOImodel|}{|VOImodel|}$$

We investigated possible dependencies of the algorithm accuracy on size and shape of the tumor and on image resolution. For that, DSCs were correlated (Pearson correlation) with volume, surface-to-volume ratio, and with lowest resolution for each subject. Surface-to-volume ratio was calculated using the pyradiomics package[21] on resampled ($1 \times 1 \times 1$ mm^3) ground truth VOIs.

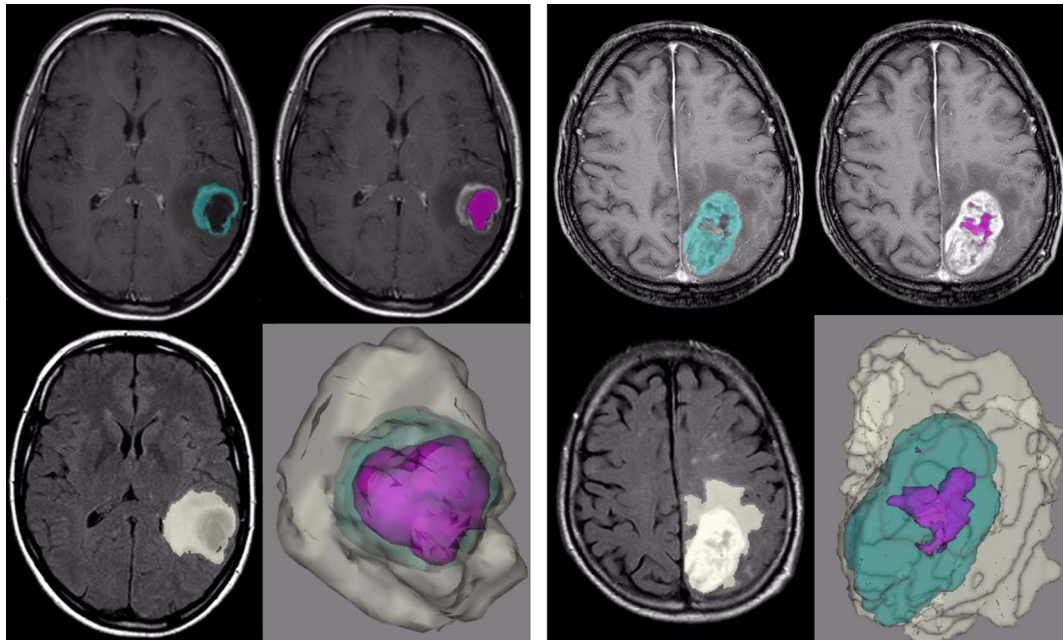

FIGURE 3. Two GB cases. Left, Subject with whole tumor (49 mL) and necrotic core (7 mL). CE tumor (11 mL) has a rim-like shape. Right, Subject with an irregular shape of the whole tumor (108 mL) and necrotic core (3 mL). CE tumor (26 mL) has round shape with the necrotic core inside. Top row, CE T1w with automatically segmented CE tumor (left, blue) and necrosis (right, magenta). Bottom row: FLAIR image with automatically segmented whole tumor (left, gray), 3D rendering (right).

© 2018 Wolters Kluwer Health, Inc. All rights reserved.

RESULTS

Manual Segmentation

Fully automated image processing and tumor segmentation was completed for all 64 patients. After visual quality control, 2 patients were excluded due to GB location in the brain stem or incomplete coverage of the tumor area by T2w series.

Manual segmentation of the whole tumor and contrast-enhancing tumor compartments was completed for all patients. Necrosis compartments were observed in 58 of the 62 patients.

Automatic Detection, Localization, and Segmentation of the Tumor

The deep learning model automatically detected, localized, and segmented the whole tumor and contrast-enhancing tumor in all 62 patients.

Necrosis was automatically detected, localized, and segmented in 53 of the 58 cases with ground truth necrosis. For the 5 cases without automatic detection (false-negatives), the mean necrosis volume on manual segmentation was 3.3 ± 1.9 mL. Further analysis of the necrosis VOI was restricted to the 53 cases where necrosis was correctly detected. Absence of the necrosis was correctly detected in 3 of 4 cases (in 1 case, the algorithm segmented a necrotic core of 1.5 mL).

Example cases with automatic segmentations are shown in Figure 3.

Volumes and results are reported in Table 2. Significant correlations with Pearson $r > 0.8$ ($P < 0.0001$) were found between the volumes of automatic and manual segmentations (Table 2, Fig. 4). Absolute volumes of automatic and manual segmentations were comparable, that is, no bias was observed in automatic segmentation.

High voxel-wise overlap was obtained for the whole and contrast-enhancing tumor volumes (DSC of 0.86 and 0.78, respectively). For 53 patients with detected necrosis, an overlap with a DSC of 0.62 was observed. The high PPV of 0.89 indicates a low number of false-positive voxels. Altogether, it demonstrates that automatic segmentation tends to underestimate necrosis. This is consistent with the smaller volumes of the automatic necrosis segmentations.

Correlation With Image and VOI Properties

Image and tumor properties may influence quality of automatic segmentation. To evaluate the effect, we correlated DSC for different VOIs with image resolution, VOI surface-to-volume ratio, and VOI volume (Table 3). The correlation scatter plots are shown in Figure 5. For all VOIs, no strong correlation was found between DSC and image resolution, VOI surface-to-volume ratio, and VOI volume.

The processing time of the deep learning model including postprocessing was less than 5 minutes per subject using an NVIDIA Tesla P100 graphics processing unit.

DISCUSSION

This study evaluated fully automatic detection and segmentation of brain tumors based on a deep learning algorithm and compared the results to manual annotations by expert readers. We furthermore investigated if segmentation results vary across clinical MRI examinations from multiple institutions using different acquisition protocols and scanners from different vendors.

The whole tumor and CE tumor VOIs have been correctly detected and localized in all cases. The necrosis VOI was correctly detected and localized in 91% of the cases.

For the automatic segmentation, the algorithm we have chosen achieved top scoring results with the BRATS test data set as reported by Kamnitsas et al[18] (see Table 4, row 2; DSC range, 0.63–0.85). The automatic segmentation results we achieved with this algorithm in our study on clinical routine data appeared to be slightly better (Table 4, row 1; DSC range, 0.78–0.86). One limitation of the deep learning algorithm is the requirement that all 4 MR input series (CE T1, FLAIR, T1, T2) need to be present. If one of the series is not available, the proposed model cannot be applied.

It is of particular clinical importance that the automatic segmentation results are in the same range as the interrater variability (Table 4, row 4; DSC range, 0.74–0.85) reported by Menze et al.[13]

Furthermore, in the first pass of our manual segmentation procedure, we observed differences between neuroradiologists' and neurosurgeons' annotation in 45% of the cases for CE tumor VOIs and in 8% of the cases for the whole tumor VOIs, based on a volume discrepancy threshold of 30%.

Applying the same 30% volume threshold to the differences between the deep learning algorithm and the ground truth, we obtained discrepancies in 25% of the cases for CE tumor VOIs and in 8% of the cases for the whole tumor VOIs. This further supports our finding that the variability between automatic and ground truth segmentations is in the range of the variability of manual segmentations by expert readers.

To put this in perspective, we compare the results from our data with the selected deep learning algorithm to another high scoring algorithm on the BRATS data by Pereira et al.[22] Our results are slightly better for the CE tumor segmentation (Pereira et al: DSC, 0.75; DeepMedic on our clinical data: DSC, 0.78).

The second part of the study addressed the variations of the segmentation results across the acquired clinical MRI examinations. In contrast to other studies,[23,24] we did not use a standardized protocol across institutions. The data analyzed in our study was obtained using a wide range of MRI acquisition protocols (15 institutions, 8 MR scanner models, eg, slice thickness ranging from 1 to 8 mm). Within the wide range of scanner models, there is a predominance of Philips 1.5 T scanners in our data. However, due to the fact that these scanners (Intera, Achieva, and Gyroscan models) are located at 6 different institutions and heterogeneous image protocols have been applied on the scanners, the data has sufficient heterogeneity to be suitable as a proof point for wider

TABLE 2. Segmentation Results

	Whole Tumor	Contrast-Enhancing Tumor	Necrosis
Subjects	62	62	53
Volume manual, mL*	122.6 ± 69.2	27.6 ± 16.7	17.0 ± 14.8
Volume automatic, mL*	116.1 ± 66.9	26.0 ± 16.4	13.8 ± 14.1
Coefficient of correlation (r)	0.96	0.90	0.84
Dice similarity coefficient*	0.86 ± 0.09	0.78 ± 0.15	0.62 ± 0.30
Sensitivity*	0.84 ± 0.13	0.78 ± 0.20	0.57 ± 0.31
Positive predictive value*	0.90 ± 0.06	0.83 ± 0.09	0.89 ± 0.20

*Mean ± standard deviation.

© 2018 Wolters Kluwer Health, Inc. All rights reserved.

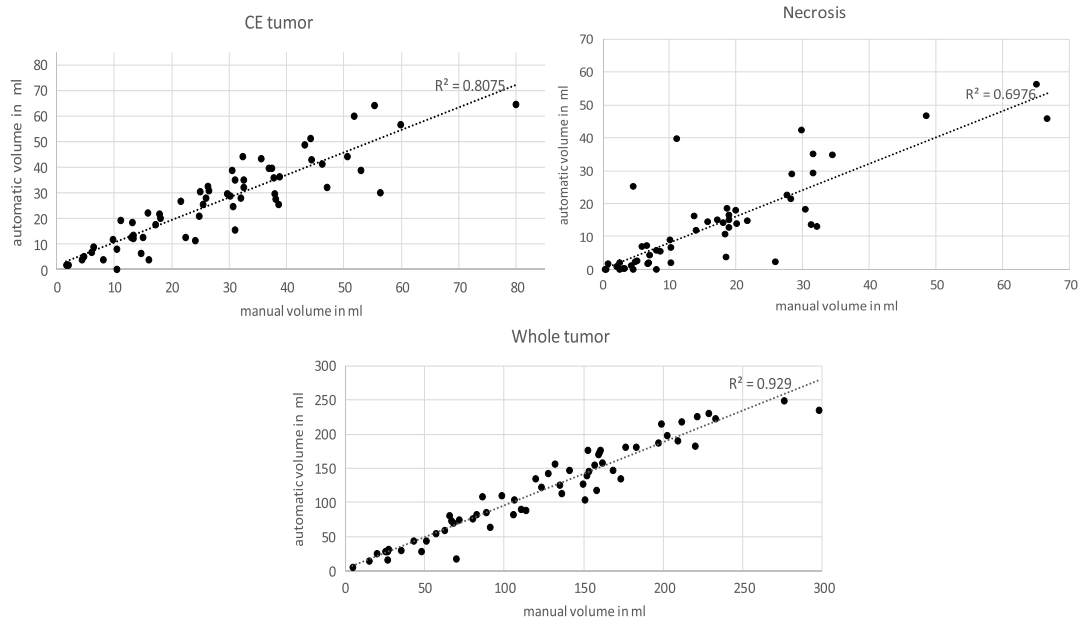

FIGURE 4. Scatter plots of correlations between volumes of automatic and manual segmentations for CE tumor (top left), necrosis (top right), and whole tumor (bottom).

applicability of the deep learning algorithm. Furthermore, the deep learning algorithm was trained on a completely independent, heterogeneous data set (BRATS data), based on GE, Siemens, and Philips scanners.

Independent from the variations in the imaging protocols, we have observed high DSC for the automatic compared with the ground truth segmentation. We have found no correlation between the DSC and slice thickness. Furthermore, no relevant correlation was found between quality of automatic segmentation and VOI properties (surface-to-volume ratio and volume).

These observations are clinically relevant because they show that the reported results of the selected deep learning algorithm, trained on the BRATS data, are reproducible on heterogeneous data sets acquired in clinical routine. Furthermore, the detection and segmentation results were not affected by variations in the imaging protocols and variations in the tumor shape and size.

It should be noted that we have not analyzed images without the presence of GB in this study. Thus, the sensitivity and specificity of the detection part of the algorithm has not been evaluated, which should be the subject of future work.

We see the potential of the deep learning algorithms, as evaluated in this work, in automatically analyzing images from primary GB. The analysis would take place in the background and before the images are read by the radiologist. At the time of the reading, the radiologist would be able to review the segmentation results, which could aid in the decision-making process. For this kind of workflow integration, it is necessary to provide an automatic and precise segmentation of the different tumor areas. In the primary GB setting, important clinical questions are as follows[25]:

- The selection of the area for maximum safe resection of the tumor to improve overall survival, while at the same time reducing the patients' functionality as little as possible.
- Identifying different compartments of the tumor, for example, relevant biopsy targets.
- Identifying relevant prognostic markers.

The segmentation of tumor compartments (whole tumor, CE tumor, and necrosis) evaluated in this work is important to address these questions properly.

The whole tumor VOI includes edema, the CE tumor, and the necrosis. The DSC on our clinical data and the BRATS test data is around 0.85 with a high PPV of 0.9, showing that only a low number of voxels are misclassified by the algorithm as normal tissue. Furthermore, the DSC is equal to the interrater DSC. In the clinical reading, the whole tumor VOI would be used to, for example, determine the extent of the edema. In the research setting, the edema part of the whole tumor VOI could be used to further evaluate potential tumor invasion in this area via, for example, diffusion tensor analysis.[26] Compared with the time-consuming and user-dependent manual segmentation, the results of the automatic segmentations suggest a high potential for integration in the radiology reading workflow.

TABLE 3. Correlation Coefficient *r* for DSC for Image Resolution, VOI Surface-to-Volume Ratio, and VOI Volume

	Lowest Resolution (*r*)	Surface-to-Volume Ratio (*r*)	Volume (*r*)
DSC (whole tumor)	−0.133 ($P = 0.30$)	−0.419 ($P < 0.01$)	0.218 ($P = 0.09$)
DSC (CE tumor)	−0.068 ($P = 0.60$)	−0.421 ($P < 0.01$)	0.180 ($P = 0.16$)
DSC (necrosis)	−0.022 ($P = 0.85$)	−0.622 ($P < 0.01$)	0.492 ($P < 0.01$)

Lowest resolution was defined for each subject as largest slice thickness of the 4 MR sequences.

DSC indicates dice similarity coefficient; CE, contrast-enhanced; VOI, volume of interest.

© 2018 Wolters Kluwer Health, Inc. All rights reserved.

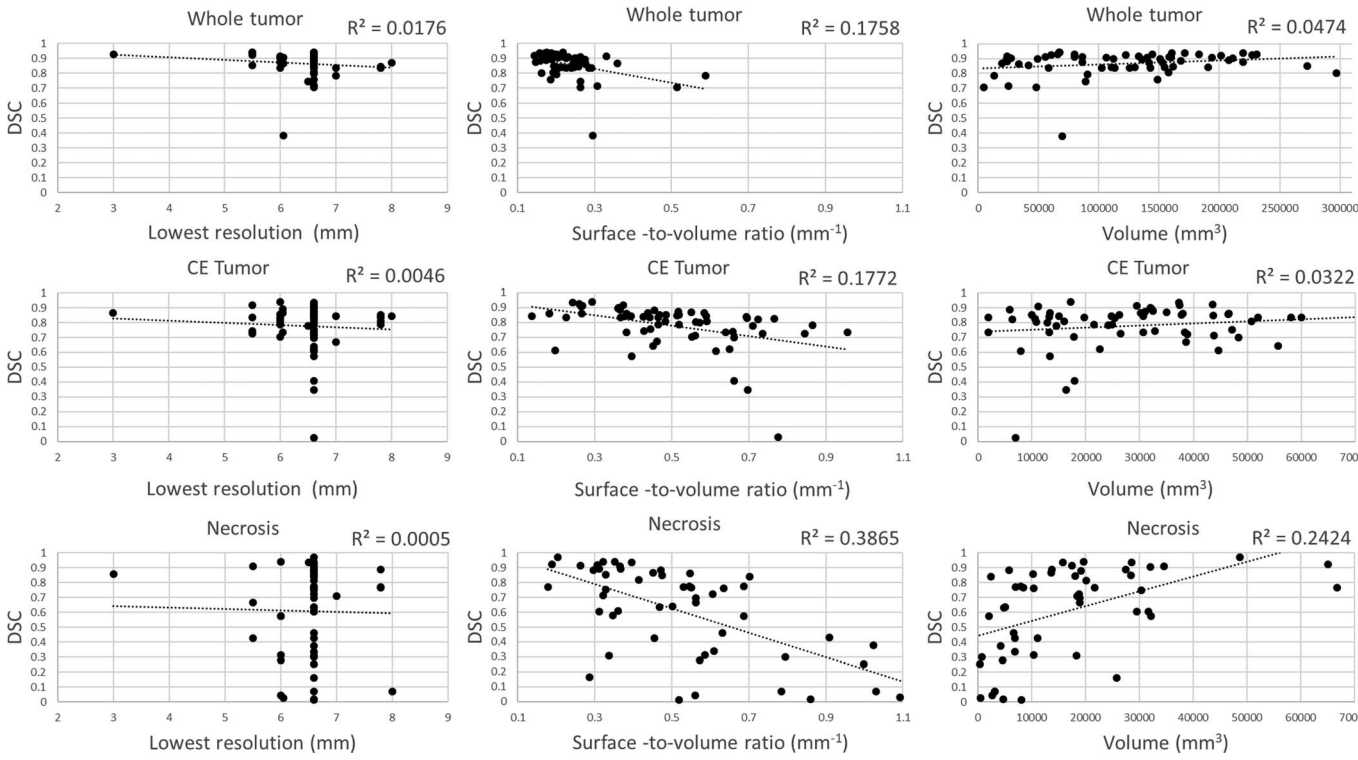

FIGURE 5. Scatter plots of correlations between segmentation accuracy (DSC) with lowest resolution, surface-to-volume ratio, volume for whole tumor, CE tumor, and necrosis VOIs.

The CE tumor compartment is the area of the tumor with contrast agent accumulation, hyperintense on a postcontrast, T1w image. Biologically, this area reflects the part of the tumor with leaky and poorly constructed vessel. The DSC in our clinical data (0.78) was higher than on the BRATS test data (DSC, 0.63). Similar to the whole tumor VOI, the PPV value of 0.83 shows that a relatively low number of voxels is misclassified as non-CE tumor.

In the clinical reading, the CE tumor compartment is important to determine the resection boarder and to identify relevant biopsy targets. In the research setting, this VOI is important, for example, in multiparametric analysis of potential pseudoprogression after radiation therapy as part of the longitudinal tracking.

Taking into account that the CE tumor compartments are more fragmented and complex in shape, they will take longer to segment manually than the whole tumor VOI. Therefore, the results will be even more user dependent. The automatically generated and user-independent CE tumor VOI segmentation with a DSC of 0.78 further supports the potential for the integration in the radiology workflow.

To our knowledge, no data are currently available on interrater variability or on comparison of automatic and manual segmentations for necrosis. For automatic segmentation of necrosis, a DSC of 0.62 was achieved, which is a reasonable result for small and heterogeneous volumes of the necrotic compartment (eg, Fig. 3) in this study cohort (17.0 ± 14.8 mL). Furthermore, as for the other compartments, the algorithms achieved a high PPV (0.89 ± 0.20) for necrosis, showing that only a low number of voxels are misclassified by the algorithm as nonnecrosis. Manual segmentation of tumor necrosis would be the most challenging part because these areas are sometimes very small, heterogeneous, and scattered. Furthermore, the required accuracy for the segmentation of the necrosis needs more discussion and clinical evaluation.

The results of the study show that the proposed algorithm for automatic detection of the primary GB tumor and the segmentation of the different tumor compartments has the potential to reproducibly and fully automatically support the clinical reading and preoperative planning.

The results are reproducible compared with former studies on different data and show a stable performance on a wide variety of clinical scanners and protocols. This reduces the risk described by some authors that spatial-temporal changes due to new MR machines or protocols will affect the performance of the algorithm.[25]

TABLE 4. Comparison of Results From Different Segmentation Approaches With the Selected Algorithm

		DSC		PPV		Sensitivity	
		Whole Tumor	CE Tumor	Whole Tumor	CE Tumor	Whole Tumor	CE Tumor
1	DeepMedic (our data set)	0.86	0.78	0.90	0.83	0.84	0.78
2	DeepMedic (BRATS 2015 test data)[18]	0.85	0.63	0.85	0.63	0.88	0.66
3	Pereira et al (BRATS 2015 test data)[22]	0.87	0.75				
4	Interrater BRATS mean[13]	0.85	0.74				

DSC indicates dice similarity coefficient; PPV, positive predictive value; CE, contrast-enhanced; BRATS, Multimodal Brain Tumor Segmentation Challenge.

© 2018 Wolters Kluwer Health, Inc. All rights reserved.

To reduce a possible bias in our ground truth generation, we combined semiautomatic approaches with a manual, consensus-based, repeated annotation by experts from radiology and neurosurgery. Next steps in research include using the automatic segmentation in the postsurgery setting to automatically detect and determine a residual tumor volume, which would influence the patient prognosis. A further step would be to apply the approach for automatic VOI generation in longitudinal tumor tracking to enable a multiparametric analysis in the case of, for example, pseudoresponse in targeted therapy. For these clinical questions, using the VOIs as input for a radiomics—[27,28] or in combination with genetic markers for a radiogenomics-analysis,[29,30] could be of further research interest.

To drive clinical acceptance of automatic segmentation in routine reading further validation of the clinical applicability of the algorithm is needed. Seeing the stability on the heterogeneous data a next step could be to pool the data from different centers for a multicenter trial to further prove and validate the stability and reproducibility of the algorithm results.

REFERENCES

1. Thakkar JP, Dolecek TA, Horbinski C, et al. Epidemiologic and molecular prognostic review of glioblastoma. *Cancer Epidemiol Biomarkers Prev*. 2014;23:1985–1996.
2. Omuro A. Glioblastoma and other malignant gliomas. *JAMA*. 2013;310:1842.
3. Stupp R, Mason WP, Van Den Bent MJ, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med*. 2005;35210:987–996.
4. Osborn A, Salzman K, Jhaveri M, et al. Diagnostic imaging: brain E-book. 2015.
5. Lacroix M, Abi-Said D, Fourney DR, et al. A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. *J Neurosurg*. 2001;95:190–198.
6. Hammoud MA, Sawaya R, Shi W, et al. Prognostic significance of preoperative MRI scans in glioblastoma multiforme. *J Neurooncol*. 1996;27:65–73.
7. Narang S, Lehrer M, Yang D, et al. Radiomics in glioblastoma: current status, challenges and potential opportunities. *Transl Cancer Res*. 2016;5:383–397.
8. Yip SS, Aerts HJ. Applications and limitations of radiomics. *Phys Med Biol*. 2016; 61:R150–R166.
9. Deeley MA, Chen A, Datteri R, et al. Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. *Phys Med Biol*. 2011;56:4557–4577.
10. Gordillo N, Montseny E, Sobrevilla P. State of the art survey on MRI brain tumor segmentation. *Magn Reson Imaging*. 2013;31:1426–1438.
11. Bauer S, Wiest R, Nolte LP, et al. A survey of MRI-based medical image analysis for brain tumor studies. *Phys Med Biol*. 2013;58:R97–R129.
12. Anon. Available at: http://braintumorsegmentation.org/.
13. Menze B. The multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;20:1878–1891.
14. Isin A, Direkogul C, Sah M. ScienceDirect review of MRI-based brain tumor image segmentation using deep learning methods. *Procedia Comput Sci*. 2016;102: 317–324.
15. Cicero M, Bilbily A, Colak E, et al. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Invest Radiol*. 2017;52:281–287.
16. Becker AS, Marcon M, Ghafoor S, et al. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Invest Radiol*. 2017;52:434–440.
17. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88.
18. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. 2017;36:61–78.
19. Ashburner J, Friston KJ. Unified segmentation. *Neuroimage*. 2005;26:839–851.
20. Crum WR, Camara O, Hill DL. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Trans Med Imaging*. 2006;25: 1451–1461.
21. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77:e104–e107.
22. Pereira S, Pinto A, Alves V, et al. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans Med Imaging*. 2016;35:1240–1251.
23. Huber T, Alber G, Bette S, et al. Progressive disease in glioblastoma: benefits and limitations of semi-automated volumetry. *PLoS One*. 2017;12:e0173112.
24. Porz N, Bauer S, Pica A, et al. Multi-modal glioblastoma segmentation: man versus machine. *PLoS One*. 2014;9:e96873.
25. Fuster-Garcia E, García-Gómez JM, De Angelis E, et al. Use case II: imaging biomarkers and new trends for integrated glioblastoma management. In: *Imaging Biomarkers*. Cham: Springer International Publishing; 2017:181–194.
26. Price SJ, Jena R, Burnet NG, et al. Improved delineation of glioma margins and regions of infiltration with the use of diffusion tensor imaging: an image-guided biopsy study. *AJNR Am J Neuroradiol*. 2006;27:1969–1974.
27. Ingrisch M, Schneider MJ, Nörenberg D, et al. Radiomic analysis reveals prognostic information in T1-weighted baseline magnetic resonance imaging in patients with glioblastoma. *Invest Radiol*. 2017;52:360–366.
28. Hainc N, Stippich C, Stieltjes B, et al. Experimental texture analysis in glioblastoma: a methodological study. *Invest Radiol*. 2017;52:367–373.
29. Gutman DA, Dunn WD, Grossmann P, et al. Somatic mutations associated with MRI-derived volumetric features in glioblastoma. *Neuroradiology*. 2015; 57:1227–1237.
30. Panth KM, Leijenaar RT, Carvalho S, et al. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. *Radiother Oncol*. 2015;116:462–466.

© 2018 Wolters Kluwer Health, Inc. All rights reserved.