

# Robust Deep Learning–based Segmentation of Glioblastoma on Routine Clinical MRI Scans Using Sparsified Training


Roelant S. Eijgelaar, MSc\* • Martin Visser, MSc\* • Dominique M. J. Müller, MSc • Frederik Barkhof, MD, PhD • Hugo Vrenken, PhD • Marcel van Herk, PhD • Lorenzo Bello, MD • Marco Conti Nibali, MD • Marco Rossi, MD • Tommaso Sciortino, MD • Mitchel S. Berger, MD • Shawn Hervey-Jumper, MD • Barbara Kiesel, MD • Georg Widhalm, MD • Julia Furtner, MD • Pierre A. J. T. Robe, MD, PhD • Emmanuel Mandonnet, MD, PhD • Philip C. De Witt Hamer, MD • Jan C. de Munck, PhD • Marnix G. Witte, PhD

From the Department of Radiation Oncology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands (R.S.E., M.v.H., M.G.W.); Department of Radiology and Nuclear Medicine, Amsterdam UMC, Location Vrije Universiteit Amsterdam, Amsterdam, the Netherlands (M.V., F.B., H.V., J.C.d.M.); Neurosurgical Center Amsterdam, Amsterdam UMC, Location Vrije Universiteit Amsterdam, Amsterdam, the Netherlands (D.M.J.M., P.C.D.W.H.); Institutes of Neurology & Healthcare Engineering, University College London, London, England (F.B.); Faculty of Biology, Medicine & Health, Division of Cancer Sciences, University of Manchester and Christie NHS Trust, Manchester, England (M.v.H.); Neurosurgical Oncology Unit, Department of Oncology and Hemato-Oncology, Università degli Studi di Milano, Humanitas Research Hospital, IRCCS, Milan, Italy (L.B., M.C.N., M.R., T.S.); Department of Neurologic Surgery, University of California–San Francisco, San Francisco, Calif (M.S.B., S.H.J.); Department of Neurosurgery, Medical University Vienna, Vienna, Austria (B.K., G.W.); Department of Biomedical Imaging and Image-guided Therapy, Medical University Vienna, Vienna, Austria (J.F.); Department of Neurology & Neurosurgery, University Medical Center Utrecht, Utrecht, the Netherlands (P.A.J.T.R.); and Department of Neurologic Surgery, Hôpital Lariboisière, Paris, France (E.M.). Received June 13, 2019; revision requested July 30; revision received April 10, 2020; accepted April 16. **Address correspondence to** M.G.W. (e-mail: [m.witte@nki.nl](mailto:m.witte@nki.nl)).

\*R.S.E. and M.V. contributed equally to this work.

This research is part of the program Innovative Medical Devices Initiative with project number 10-10400-96-14003, which is financed by the Netherlands Organization for Scientific Research (NWO). This research is also supported by a research grant from the Dutch Cancer Society (VU2014-7113). F.B. is supported by the NIHR-UCLH Biomedical Research Centre. M.v.H. is supported by the NIHR Manchester Biomedical Research Centre.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2020; 2(5):e190103 • <https://doi.org/10.1148/ryai.2020190103> • Content codes:  

**Purpose:** To improve the robustness of deep learning–based glioblastoma segmentation in a clinical setting with sparsified datasets.

**Materials and Methods:** In this retrospective study, preoperative T1-weighted, T2-weighted, T2-weighted fluid-attenuated inversion recovery, and postcontrast T1-weighted MRI from 117 patients (median age, 64 years; interquartile range [IQR], 55–73 years; 76 men) included within the Multimodal Brain Tumor Image Segmentation (BraTS) dataset plus a clinical dataset (2012–2013) with similar imaging modalities of 634 patients (median age, 59 years; IQR, 49–69 years; 382 men) with glioblastoma from six hospitals were used. Expert tumor delineations on the postcontrast images were available, but for various clinical datasets, one or more sequences were missing. The convolutional neural network, DeepMedic, was trained on combinations of complete and incomplete data with and without site-specific data. Sparsified training was introduced, which randomly simulated missing sequences during training. The effects of sparsified training and center-specific training were tested using Wilcoxon signed rank tests for paired measurements.

**Results:** A model trained exclusively on BraTS data reached a median Dice score of 0.81 for segmentation on BraTS test data but only 0.49 on the clinical data. Sparsified training improved performance (adjusted  $P < .05$ ), even when excluding test data with missing sequences, to median Dice score of 0.67. Inclusion of site-specific data during sparsified training led to higher model performance Dice scores greater than 0.8, on par with a model based on all complete and incomplete data. For the model using BraTS and clinical training data, inclusion of site-specific data or sparsified training was of no consequence.

**Conclusion:** Accurate and automatic segmentation of glioblastoma on clinical scans is feasible using a model based on large, heterogeneous, and partially incomplete datasets. Sparsified training may boost the performance of a smaller model based on public and site-specific data.

Supplemental material is available for this article.

Published under a CC BY 4.0 license.

Glioblastoma is the most common form of primary brain tumor in adults (1,2). Radiologic tumor evaluation is typically performed using CT or MRI with two-dimensional measures (3), but with advances in imaging and the need for more detailed tumor quantification, three-dimensional volumetric segmentation in MRI is recommended and is becoming more commonplace (4,5). Preoperative glioma segmentation at MRI can aid localized treatment planning and assessment of quality of care (6,7). Several MRI sequences

are acquired to assess tumor location, composition, and extent (8); however, manual tumor segmentation is a time-consuming process and subject to interrater variability (9).

Automated tumor segmentation is an active field of research. The 2017 Multimodal Brain Tumor Segmentation (BraTS) (10) challenge showed promising results with Dice scores of approximately 0.85 for the preoperative tumor core mostly using deep learning–based approaches (11). The NiftyNet (12) project offers standardized tools for deep

## Abbreviations

BraTS = Multimodal Brain Tumor Image Segmentation, FLAIR = fluid-attenuated inversion recovery, IQR = interquartile range, TCIA = The Cancer Imaging Archive

## Summary

Robust deep learning–based segmentation of glioblastoma on routine clinical data can be achieved using a large heterogeneous training dataset or using sparsified training on a combination of public and site-specific data.

## Key Points

- Institutional variations in MRI acquisition protocols, hardware, and software result in heterogeneous clinical image datasets that may lack specific sequences; the variability of input data may affect the training and performance of deep learning algorithms.
- Sparsified training in a dataset consisting of glioblastoma MRI scans improved the performance of a model based on public data to the level of performance of inclusion of center-specific training data, as well as reduced the influence of missing sequences.
- Models trained on large heterogeneous datasets with missing sequences did not require sparsified training or site-specific training data.

learning–based automatic segmentation. However, methods developed using high-quality homogeneous and complete research data may suffer from overfitting (13) and fail to achieve high segmentation quality in clinical scans with variable image quality and varying completeness of image sequences. Implementation of automatic segmentation in clinical practice is therefore still lacking.

Heterogeneity of the BraTS data has increased since 2017 by addition of glioblastoma data from eight institutions in The Cancer Imaging Archive (TCIA) (13) and manual ground truth segmentations (14). Previously proposed methods to address the issue of missing sequences generally focused on data imputation (15–17) or network adjustments (16–18), but such approaches have limited generalizability.

In this study, we determined the performance of the automatic glioblastoma segmentation tool DeepMedic (19,20) when using partially incomplete multi-institution clinical imaging data with a wide variability in imaging parameters. We investigated the effects on segmentation performance of center-specific training, missing sequences, and expanding the training data with a heterogeneous dataset. Furthermore, we propose a sparsified training protocol randomly nullifying secondary image sequences and show how this improves robustness in the case of missing data.

## Materials and Methods

Approval of this retrospective study protocol was obtained from institutional review boards, and informed consent from patients was obtained according to local regulations. The data were obtained and anonymized in accordance with the General Data Protection Regulation and Health Insurance Portability and Accountability Act.

### Public BraTS Patient Dataset

From the BraTS dataset updated in 2013 (7) and selected scans from the TCIA (11,12), preoperative baseline scans were se-

lected, resulting in a dataset of 117 patients (median age, 64 years; interquartile range [IQR], 55–73 years; 41 women, 76 men) with four MRI types present: pre- and postcontrast T1-weighted, T2-weighted, and T2-weighted fluid-attenuated inversion recovery (FLAIR). Most scans were obtained at a field strength of 1.5 T ( $n = 59$ ); the remaining used a 3-T scanner ( $n = 35$ ). The segmented tumor core, defined as the union of enhancing and nonenhancing tumor, and necrotic regions were used for analysis. For BraTS, the tumors were manually segmented; for TCIA, they were semiautomatically segmented using GLISTRBoost (<https://www.med.upenn.edu/sbia/glistrboost.html>) and manually corrected.

### Clinical Patient Dataset

Preoperative MR images of 634 adult patients (median age, 59 years; IQR, 49–69 years; 382 men, 244 women, eight unknown) with a histopathologic diagnosis of glioblastoma were collected (Table 1). These patients received consecutive surgical treatment between 2012 and 2013 at one of six multinational tertiary referral hospitals (referred to as hospitals 1–6). Patients either underwent a resection or a biopsy. All patients with at least a preoperative postcontrast T1-weighted scan were included, and no patients were excluded. These data were collected as part of the PICTURE project (<https://www.pictureproject.nl>) (7,9,21–23). The histopathologic diagnosis was determined according to the World Health Organization 2007 criteria (24). Scans were obtained using 21 different MRI scanner models, and most scans were obtained at a field strength of 1.5 T ( $n = 316$ ) or 3 T ( $n = 292$ ), with smaller numbers at 1 T ( $n = 22$ ) or 0.4 T ( $n = 1$ ). Scan resolutions varied, ranging between 0.5- and 6-mm slice thickness. Images had varying MRI parameters because of hospital-specific settings (eg, field of view) and scanner specifications (see Tables E1 and E2 [supplement] for more details). In 54% (345 of 634) of patients, one or more of the secondary sequences were missing (see Table 2).

Tumor segmentation following the Visually Accessible Rembrandt Images criteria (25) was performed by a single manual rater with 3 years of experience in neurosurgical residency (D.M.J.M.) under supervision of a neurosurgeon (P.C.D.W.H.) and neuroradiologist (F.B.) using the semiautomatic SmartBrush tool (BrainLab, Feldkirchen, Germany). Performance on preoperative glioblastoma segmentation of this rater was comparable with expert level (9). Tumor volume was defined as the union of the enhancing tumor and enclosed necrosis, which is comparable with tumor core segmentations from BraTS and TCIA glioblastoma with the exclusion of nonenhancing tumor.

Secondary sequences (T1-weighted, T2-weighted, and FLAIR) were rigidly registered to the postcontrast T1-weighted sequences and then rigidly registered to the MNI09a template (<http://nist.mni.mcgill.ca/?p=904>) and thus resampled to 1-mm isotropic voxels. Registrations were performed with the Advanced Normalization Tools (26). Bias field corrections were performed with N4 bias correction (27), and skull stripping was performed with a routine that relied on the Atropos (28) tool.

### Automatic Tumor Segmentation

Automatic segmentations were performed using the convolutional neural network DeepMedic (19,20) as implemented

**Table 1: Clinical Patient and Public BraTS Characteristics**

Characteristic	Clinical Patients ( $n = 634$ )	Public BraTS ( $n = 97$ )*
Age (y)	64 (55–72)	59 (49–69)
Sex (male/female/unknown)	382/244/8	76/41/0
Preoperative KPS	8 (7–9)	...
Overall survival (mo)	9.8 (3.9–19.5)	...
Enhancing tumor volume (mL)	31 (14–53)	40 (21–66)

Note.—Unless otherwise indicated, characteristics are shown as median, with interquartile range in parentheses. BraTS = Multimodal Brain Tumor Image Segmentation, KPS = Karnofsky performance status.

\* There were a total of 117 patients within the public BraTS dataset. BraTS characteristics could only be determined for 97 patients who were also included in the The Cancer Imaging Archive dataset. Enhancing tumor volume was calculated for all included BraTS patients.

**Table 2: Missing Scans in Clinical Dataset**

Hospital	Complete	T1w	T2w	FLAIR	T1w, T2w	T1w, FLAIR	T2w, FLAIR	T1w, T2w, FLAIR	Total
1	89	1	2	1	0	0	1	3	97
2	6	0	67	0	3	0	0	0	76
3	59	6	0	0	0	3	1	16	85
4	51	1	1	0	1	0	1	18	73
5	61	65	2	0	0	2	1	1	132
6	16	1	13	2	0	0	130	9	171
Total	282	74	85	3	4	5	134	47	634

Note.—FLAIR = T2-weighted fluid-attenuated inversion recovery, T1w = precontrast T1-weighted, T2w = T2-weighted.

in the NiftyNet (12) framework. DeepMedic consists of two pathways that are 11 layers deep and accepts three-dimensional patches as inputs. The patch size was set to  $57 \times 57 \times 57$  voxels with a downsample factor of three, resulting in an output of  $9 \times 9 \times 9$  voxels. Training minimized a Dice loss function (29) for 30 000 iterations by using the Adam (30) optimizer. The learning rate started at 0.001 and was divided by 2 every 5000 iterations up to iteration 15 000 and every 1500 iterations thereafter. The last iteration was used for inference on the test datasets. All models were trained and evaluated on a computer equipped with a single Tesla P100 GPU (3584 CUDA cores, 12 GB).

### Sparsified Training

Where scans were missing, empty (zero-filled) scans with the same resolution and orientation as the other scans were inserted. Sparsified training was implemented as an augmentation layer by randomly setting secondary sequences to zero with independent probabilities of 20%. This percentage approximated the frequency of missing sequences in the clinical dataset. The NiftyNet pipeline includes histogram normalization and whitening (12,31). These layers were adjusted to ensure that zero-filled volumes representing missing sequences were unaffected. Consequently, missing sequences in the original data and from the sparsified training augmentation layer were fed to the network as all zero matrices, which, as a result of the whitening layer, corresponds to the mean intensity of the available data.

### Tumor Segmentation Evaluation

Tumor segmentations were evaluated using Dice score, Hausdorff distance, and sensitivity metrics (32). Dice score is defined as:

$$\text{DICE} = \frac{2|S_m \cap S_a|}{|S_m| + |S_a|} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}},$$

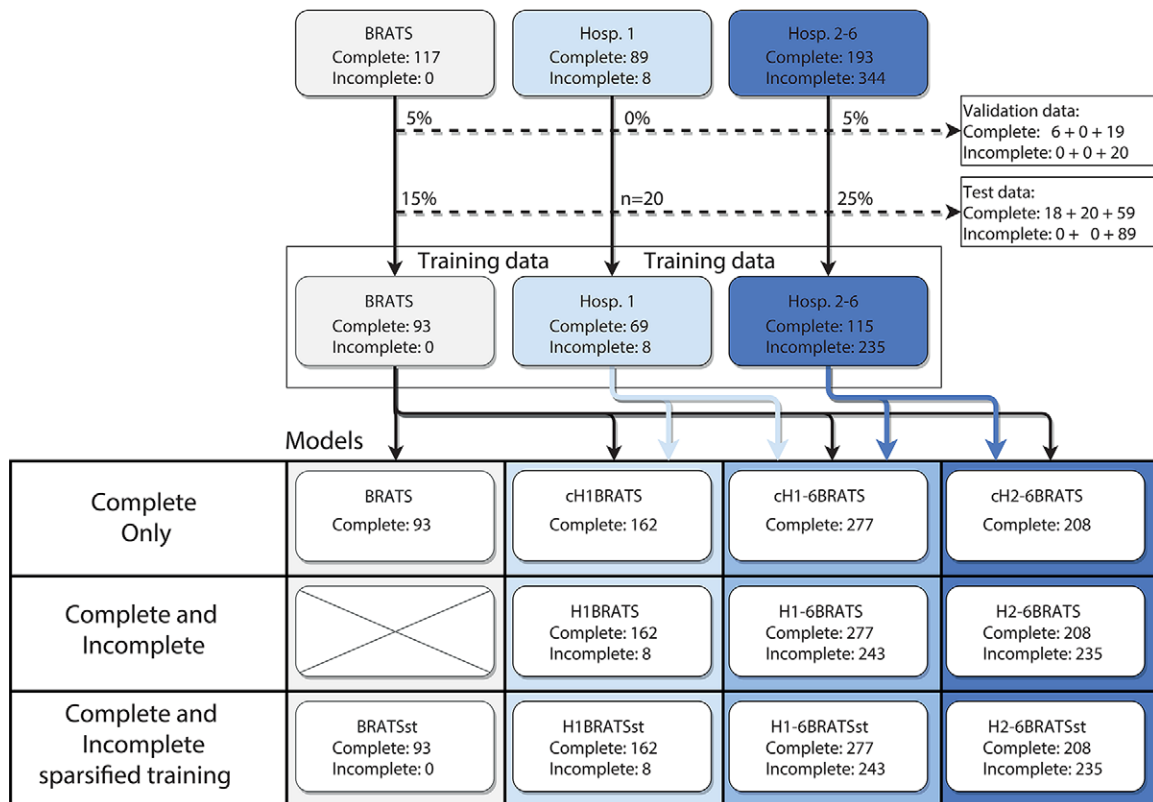
where  $S_m$  is the manually segmented tumor voxels,  $S_a$  is the automatically segmented tumor voxels, TP is the number of true-positive voxels, FP is the false-positive voxels, and FN is the false-negative voxels. Sensitivity follows as:

$$\text{SENS} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

The Hausdorff distance measures the maximum distance between borders and is therefore sensitive to evaluation direction and outliers. We therefore used the undirected 95th percentile Hausdorff distance:

$$d_{0.95H} = \max \left\{ 95 \text{ percentile } \min_{p_a \in P_a} d(p_a, P_m), 95 \text{ percentile } \min_{p_m \in P_m} d(p_m, P_a) \right\},$$

where  $P_m$  is the set of vertices describing the border of the manual segmentation,  $P_a$  is the set of vertices describing the border of the automatic segmentation, and  $d(p_a, p_m)$  is the distance between two vertices.



**Figure 1:** The clinical and public data were divided into three main groups: Multimodal Brain Tumor Image Segmentation (BraTS), hospital 1, and hospitals 2 to 6. The dashed arrows show the (fraction of) patients assigned to the validation and test data. The remaining scans were used to train a total of 11 distinct networks by varying all of the included training data, use of all available data, or only patients with complete imaging, and use of sparsified training. Each model is described by the abbreviation of the included data, preceded with a “c” if only patients with complete imaging were included, and followed by “st” if sparsified training was enabled. The 20 test patients from hospital 1 were previously studied in Visser et al (9).

## Experimental Design

The BraTS dataset was randomly split into training (80%,  $n = 93$ ), validation (5%,  $n = 6$ ), and testing (15%,  $n = 18$ ) sets. The clinical test set included 20 manually selected patients from hospital 1 with complete imaging for whom the interrater agreement of manual segmentations had been previously studied (<https://doi.org/10.17026/dans-zg9-nhrj>) (9). For hospitals 2 to 6, patients were randomly subdivided into training (70%,  $n_{\text{total}} = 350$ ), validation (5%,  $n_{\text{total}} = 39$ ), and testing (25%,  $n_{\text{total}} = 148$ ) where  $n_{\text{total}}$  is the total number of patients for hospitals 2 to 6. Combinations of training data from different datasets were used to train various models. Validation data were used to confirm convergence, and hyperparameters were then kept constant for all models. Test data were used to evaluate the performance of the different models.

A total of 11 networks were trained by varying the included hospitals, the inclusion of patients with incomplete imaging, and the use of sparsified training. BraTS patients were included in all models. Figure 1 explains the naming convention for the different models.

The performance on clinical data of a network based on publicly available BraTS data was evaluated. These results show the combined effects of imaging heterogeneity, missing scans, and other issues related to overfitting and domain adaptation. The effects of missing MRI sequences were further investigated by simulating all possible combinations of missing secondary

sequences in the test dataset of 20 patients. Sparsified training was explored as a strategy to mitigate the impact of missing sequences. The impact of image heterogeneity between institutes was estimated by evaluating networks trained with and without hospital 1 data on the set of 20 patients, thus assessing the need for hospital-specific data in the training set.

## Statistical Analyses

Dice scores comparing automatic to manual segmentations of all model pairs with and without sparsified training, as well as all model pairs with and without hospital 1-specific training data, were compared using Wilcoxon signed rank tests for paired measurements (using R, version 3.6.1, <https://www.r-project.org/>) (33) for complete imaging and all combinations of simulated missing scans in the set of 20 patients from hospital 1. A conservative multiple testing correction was applied, calculating adjusted  $P$  values using a single Bonferroni correction on all ( $n = 72$ ) calculated  $P$  values. Adjusted  $P$  values less than .05 were considered significant.

## Results

### Segmentation Performance on the Clinical Patient Dataset

The BraTS trained model achieved a median Dice score of 0.81 on BraTS test data, which was comparable with scores cited



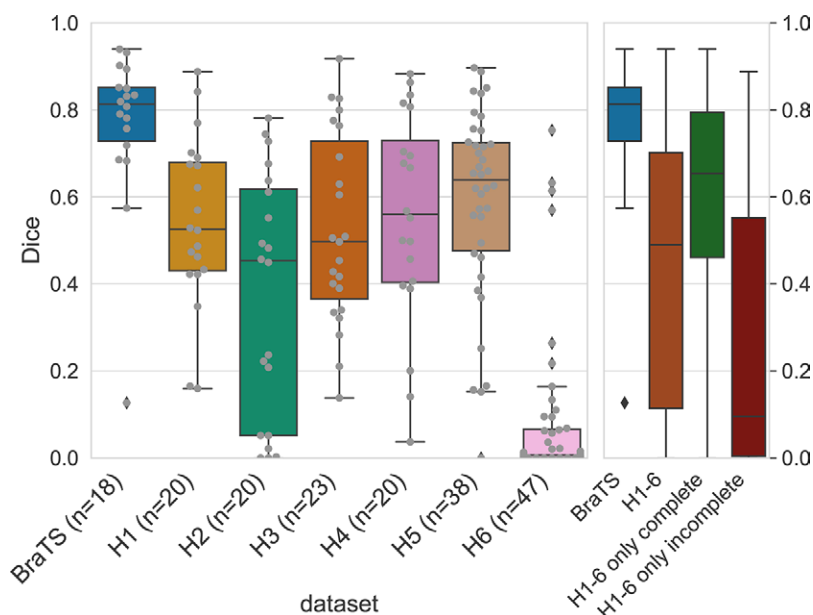
in the literature (11,20,34). Performance on test datasets of the clinical data (hospitals 1–6) was substantially lower, with an overall median Dice score of 0.49 (Fig 2). Dice scores were similar (medians around 0.55) between hospitals 1, 3, 4, and 5 but highly variable in hospital 2 and very low (median: 0.01) in

hospital 6. Results for the 95th percentile Hausdorff distance and sensitivity can be found in Figure E1 (supplement).

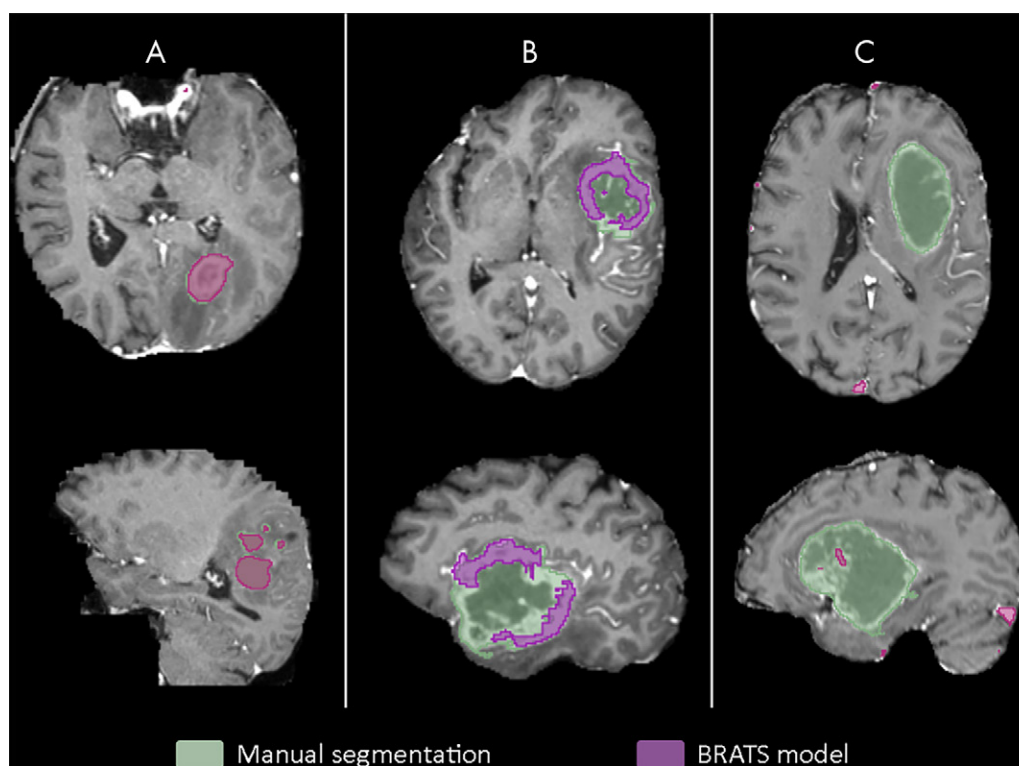
### Segmentation Performance with Missing MRI Sequences

In hospital 2 the T2-weighted sequences were missing, and in hospital 6, both the T2-weighted and FLAIR sequences for a large fraction of patients were missing. The subset of all test patients missing one or more sequences had a median Dice score of 0.095 (IQR, 0.005–0.55), and for the patients with all sequences available, this was 0.65 (IQR, 0.46–0.79). Sensitivity and  $d_{95H}$  are shown in Figure E2 (supplement). Figure 3 shows examples of the segmentation result of the BraTS model for a BraTS test patient, a clinical patient with all secondary sequences available, and a clinical patient who had missing T2-weighted and FLAIR sequences.

In Figure 4, the effect of a particular missing sequence, or combination of missing sequences, is shown for the BraTS model using the 20 test patients from hospital 1. Although a missing pre-contrast T1-weighted sequence did not reduce performance (median Dice score: 0.66), missing T2-weighted (0.30) and especially missing FLAIR (0.13) sequences led to lower median Dice scores. A number of interactions can be observed. For example, missing pre-contrast T1-weighted and FLAIR (0.41) outperforms missing FLAIR



**Figure 2:** Dice scores of a model trained on publicly available Multimodal Brain Tumor Image Segmentation (BraTS) data, evaluated on a retrospective test cohort from six hospitals and BraTS. Gray bullets indicate individual scans. The right panel shows both the pooled results for hospitals 1 to 6 and the subsets of patients with complete and incomplete data.



**Figure 3:** Segmentations generated by the Multimodal Brain Tumor Image Segmentation (BraTS) model overlaid on the manual segmentations for, A, BraTS test patient (Dice score, 0.85), B, clinical patient with complete secondary imaging (Dice score, 0.43), and, C, patient with missing T2-weighted and T2-weighted fluid-attenuated inversion recovery images (Dice score, 0.001).

(0.13), and no secondary sequences (0.16) appears better than to use precontrast T1-weighted only (0.03). This same trend was observed in the other institutes (Figure E3 [supplement]).

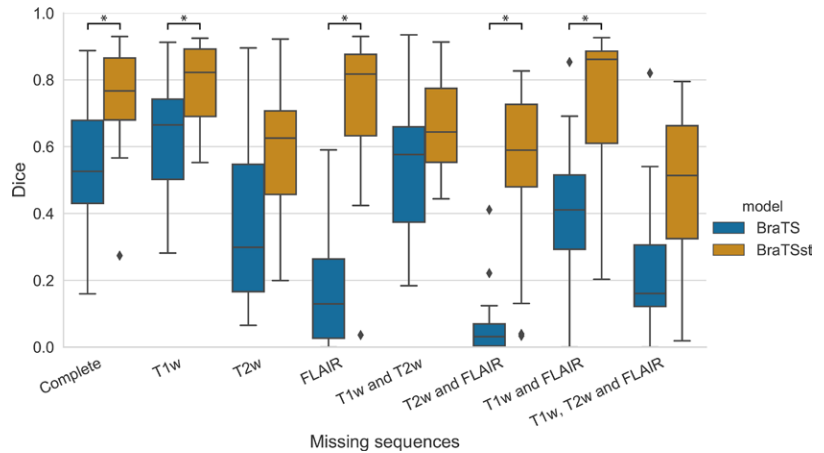
Sparsified training in the BraTS<sub>st</sub> model substantially improved performance for all combinations of missing sequences (Figure 4), with significant adjusted *P* values for missing FLAIR and/or precontrast T1-weighted sequences, as well as missing FLAIR and T2-weighted (adjusted *P* values of all model comparisons can be found in Tables E3 and E4 [supplement]). An especially large improvement was seen in the case of missing FLAIR. Also, complete test data of the BraTS<sub>st</sub> model (median Dice score: 0.77) had a higher score than that of the BraTS model (0.53), although it was statistically nonsignificant (adjusted *P* = .095). Results showed to be robust to the level of sparsity (see Figure E4 [supplement]).

### Image Heterogeneity in Training Sets Improved Segmentation

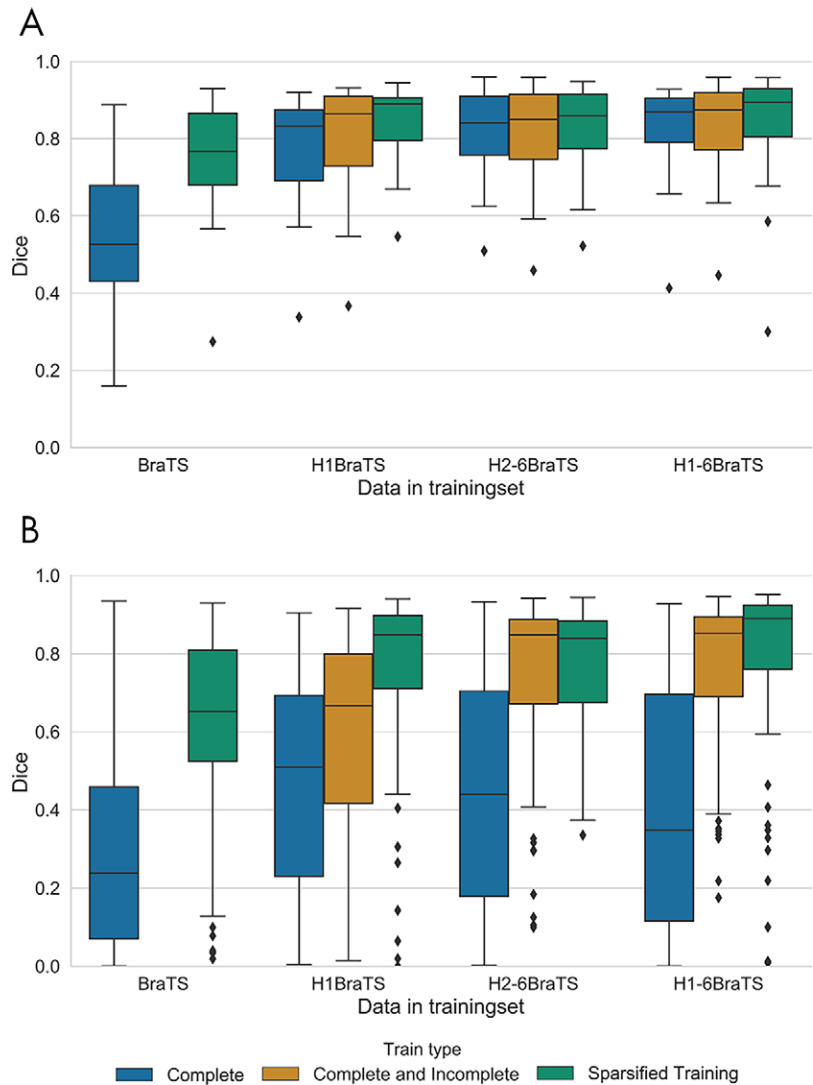
Evaluation results for models trained with the inclusion of clinical data are shown in Figure 5. When performing inference using all available scans of the 20 patients who were left out of hospital 1 (Fig 5, A), each model incorporating clinical data outperformed the BraTS model.

Omission of patients from hospital 1 during training did not significantly reduce the performance relative to the corresponding models using all available patients, unlike the models only trained on BraTS data (adjusted *P* < .05 for complete imaging [median Dice score difference: -0.07], only postcontrast T1-weighted available [-0.57], missing FLAIR [-0.50], and missing T2-weighted [-0.35]). Sparsified training neither significantly improved nor reduced the performance of models that included data from hospitals 2 to 6 (H<sub>2-6</sub>BraTS and H<sub>1-6</sub>BraTS). See Tables E3 and E4 (supplement) for tables with more detailed results, including median differences, 95% confidence intervals, and nonadjusted *P* values.

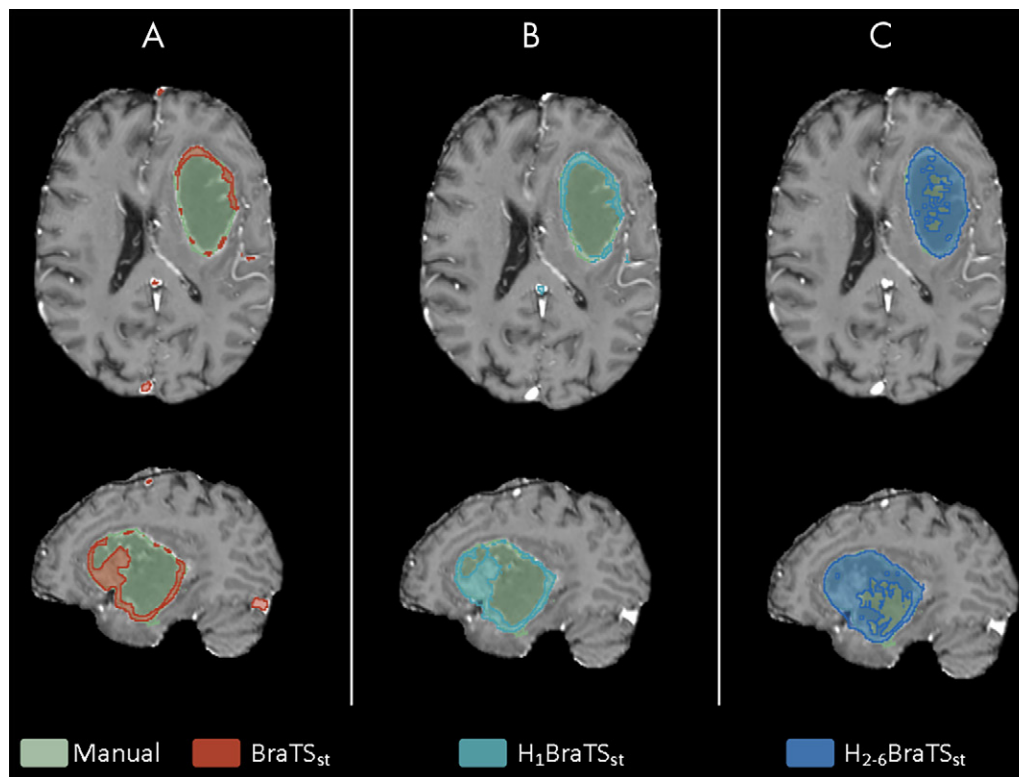
For simulated missing sequences (Fig 5, B), Dice scores for models trained exclusively on complete image sets were reduced. The inclusion of patients with incomplete imaging during training restored the performance for the H<sub>1-6</sub>BraTS and H<sub>2-6</sub>BraTS models, but only partially for the H<sub>1</sub>BraTS model (hospital 1 had only eight patients with incomplete imaging). Sparsified training restored the performance of the H<sub>1</sub>BraTS<sub>st</sub> model to the same level of performance (median Dice scores ≥ 0.84) as the models including the other hospitals. Omission of institution-specific data from hospital 1 did not significantly reduce



**Figure 4:** Dice performances of the Multimodal Brain Tumor Image Segmentation (BraTS) model with sparsified training (orange) and without sparsified training (blue). Missing data were simulated by artificially setting secondary images to zero during inference for the 20 test patients from hospital 1. Significant differences (adjusted *P* < .05) are indicated by an asterisk. FLAIR = fluid-attenuated inversion recovery, T1w = T1-weighted, T2w = T2-weighted.



**Figure 5:** Median Dice scores of all models in Figure 1 on, A, 20 test patients from hospital 1 with complete imaging and, B, for all combinations of simulated missing scans. BraTS = Multimodal Brain Tumor Image Segmentation.



**Figure 6:** Segmentation results for the, A, BraTSst (Dice score, 0.41), B, H<sub>1</sub>BraTSst (Dice score, 0.52), and, C, H<sub>2-6</sub>BraTSst (Dice score, 0.82) models for the patient with missing T2-weighted and T2-weighted fluid-attenuated inversion recovery from Figure 3, C. BraTS = Multimodal Brain Tumor Image Segmentation.

performance for the H<sub>1-6</sub>BraTS<sub>st</sub> model (adjusted  $P \geq .08$ ). Figure 6 shows the segmentation results of the BraTS<sub>st</sub>, H<sub>1</sub>BraTS<sub>st</sub>, and H<sub>2-6</sub>BraTS<sub>st</sub> models for the incomplete test patient of Figure 3. Performance (Dice, sensitivity, and  $d_{105H}$ ) of the H<sub>2-6</sub>BraTS<sub>st</sub> model is shown in more detail for individual test datasets in Figure E1 (supplement).

## Discussion

We have shown the feasibility of obtaining accurate automatic glioblastoma segmentations using deep learning without the need for hospital-specific training data, even in the case of missing secondary sequences. In clinical practice, missing sequences is a likely occurrence; for a small majority of clinical scans collected for this study, one or more of the secondary sequences were missing. A sparsified training strategy improved a model based on public data for use both on complete and incomplete clinical datasets and was able to bring a model based on single-institute data (plus those publicly available) to the level of a much larger multi-institute model.

We have shown that the models trained with the largest dataset (eg, H<sub>2-6</sub>BraTS<sub>st</sub>) reach median Dice scores (approximately 0.85 both on complete and incomplete data) that are comparable with the top-performing algorithms from the BraTS challenge and the results in Perkuhn et al (35), yet still below the excellent interrater agreement of 0.94 in the BraTS data (10) and of 0.93 in a subset of the dataset described in this article (9).

In this study, we have used the DeepMedic implementation in NiftyNet. We have chosen to use DeepMedic because

it was shown to be one of the top performers of the BraTS 2016 challenge, and several studies have reported Dice scores of DeepMedic (eg, as reference value) trained and evaluated on the BraTS data (mix of various gliomas). Results ranged between 0.72 and 0.83 (20), and submissions were from the BraTS proceedings (11,34).

Incomplete imaging was shown to greatly reduce performance in the unadjusted DeepMedic network. Even though introducing sparsified training and clinical data to the training dataset improved segmentation performance for incomplete data, the best performance was achieved if all sequences were available. Systematic use of all sequences in clinical practice is preferred (as advocated in Freyschlag et al [36]) and would help overcome this issue. Despite the efforts toward standardization, robustness to missing sequences remains valuable for analyzing real-world cohorts, flexibility to changes to standardized protocols, and bringing the benefits of automated segmentation to as many patients as possible.

The largest performance improvement that resulted from introducing sparsified training to the BraTS model was observed with missing FLAIR images, indicating that the information used by the model from the FLAIR could also be extracted from a combination of other sequences. Next to the variations in imaging protocols, variations in preprocessing (ie, the registration algorithm, atlas, and skull stripping algorithm) may have contributed to the heterogeneity between the public patient data and the clinical dataset. The largest performance differences could, however, be attributed to missing sequences. The performance

improvement from sparsified training in the BRaTS model for complete test data might be attributed to a general regularizing effect of sparsified training, reducing overfitting.

The DeepMedic implementation in NiftyNet (12) has minor differences compared with the original implementation (20), which used the now-deprecated Theano backend. However, we expect that these differences have a relatively minor impact on performance. The NiftyNet implementation does not include dropout in the final fully connected layers and uses a Dice-based loss function and Adam as an optimizer as opposed to RMSProp with Nesterov momentum in the original DeepMedic implementation. NiftyNet allows various datasets and augmentation options to be easily combined, and it facilitates publication of trained models ([https://niftynet.readthedocs.io/en/dev/model\\_zoo.html](https://niftynet.readthedocs.io/en/dev/model_zoo.html)).

Results of this study focused on the test data of hospital 1. A “leave one hospital out” cross-validation could provide more detailed results but was practically infeasible because of the computational costs involved.

The high  $d_{495H}$  (Figure E2 [supplement]) for some patients indicated that, for these patients, some false-positive regions were found relatively distant to the boundary of the tumor. This may limit the usability of these models for some applications. Improvements could be made by further postprocessing, inclusion of a distance component to the loss-function of the convolutional neural network, or further hyperparameter tuning (eg, sparsity frequency, numbers of layers and feature maps, number of iterations or learning rate). Improvements using a fully connected conditional random field (37) or the use of ensembles (38) were previously explored but were left out in our analyses because of the added complexity and training and inference times.

Our results showed that introducing sparsified training and a large heterogeneous training dataset improved the robustness of automatic segmentation of glioblastoma on routine clinical MRI. Improved robustness of automatic segmentation will bring the application of these algorithms a step closer to clinical practice.

**Acknowledgments:** This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative and the Translational Research IT (TraIT) project, an initiative from the Center for Translational Molecular Medicine (CTMM). We thank BrainLab for generously providing us with the SmartBrush software, a segmentation software, as a contribution in kind to this study. The authors declare that they have no competing interests.

**Author contributions:** Guarantors of integrity of entire study, R.S.E., M.V., F.B., M.C.N., B.K., P.A.J.T.R., M.G.W.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, R.S.E., M.V., F.B., M.v.H., L.B., M.C.N., S.H.J., B.K., G.W., J.C.d.M., M.G.W.; clinical studies, F.B., L.B., M.C.N., M.R., T.S., S.H.J., B.K., G.W., J.F., P.A.J.T.R., P.C.D.W.H.; experimental studies, R.S.E., L.B., M.C.N., M.G.W.; statistical analysis, R.S.E., M.V., M.v.H., L.B., M.C.N., P.A.J.T.R., J.C.d.M., M.G.W.; and manuscript editing, R.S.E., M.V., D.M.J.M., F.B., H.V., M.v.H., L.B., M.C.N., M.S.B., S.H.J., B.K., G.W., J.F., P.A.J.T.R., P.C.D.W.H., J.C.d.M., M.G.W.

**Disclosures of Conflicts of Interest:** R.S.E. disclosed no relevant relationships. M.V. Activities related to the present article: institution receives grant from Netherlands Organization for Scientific Research (NOW) (project number 10-10400-96-14003); institution receives grant from Dutch Cancer Society (VU2014-7113). Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. D.M.J.M.

disclosed no relevant relationships. F.B. Activities related to the present article: institution receives grant from Netherlands Organization for Scientific Research (NOW) for PICTURE project. Activities not related to the present article: author paid as board member of Roche, Bayer, and Merck (DSMB and Steering Committees); author is consultant for IXICO; institution receives grants from EU-H2020, UKMSS, NWO, MRC, HHR-BRCUCLH; author receives royalties from Springer books; institution paid for educational presentations by Biogen (PML educational website). Other relationships: disclosed no relevant relationships. H.V. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: institution is consultant for Merck (multiple sclerosis brain imaging consulting); institution receives grants from Teva, Novartis, Merck (research grants for multiple sclerosis studies). Other relationships: disclosed no relevant relationships. M.v.H. Activities related to the present article: institution receives grant from Dutch Cancer Society. Activities not related to the present article: institution receives grants from MRC Proximity to Discovery, MRC Studentship (PhD student); author paid for lectures by ESTRO (Honorarium -Sept 2018, ESTRO [ATP Meeting] €500 Honorarium -Feb 2019, ESTRO [IGRT Meeting] €500); author receives travel accommodations from IPFM, UKIO, AMC, AIFM, IGT Network, ICR, ESTRO, Elekta, NKI (conference and meeting expenses such as travel, accommodations, etc. IPFM July 2019, UKIO June 2019, July 2018, AMC May 2019, Jan 2019, BIR April 2019, March 2019, AIFM March 2019, IGT Network March 2019, Nov 2018, ICR March 2019, ESTRO Feb 2019, Sept 2018 Elekta June 2018, NKI July 2018). Other relationships: disclosed no relevant relationships. L.B. disclosed no relevant relationships. M.C.N. disclosed no relevant relationships. M.R. disclosed no relevant relationships. T.S. disclosed no relevant relationships. M.S.B. disclosed no relevant relationships. S.H.J. disclosed no relevant relationships. B.K. disclosed no relevant relationships. G.W. disclosed no relevant relationships. J.F. disclosed no relevant relationships. P.A.J.T.R. disclosed no relevant relationships. E.M. disclosed no relevant relationships. P.C.D.W.H. Activities related to the present article: institution receives grant from Dutch Cancer Society (VU2014-7113); BrainLab provided SmartBrush software; ZonMW (This research is part of the program Innovative Medical Devices Initiative with project number 10-10400-96-14003, which is financed by the Netherlands Organization for Scientific Research). Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. J.C.d.M. Activities related to the present article: institution received grant from ZonMW (ZonMW paid a research grant to VUmc to cover salaries of PhD students involved. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. M.G.W. disclosed no relevant relationships.

## References

- Ostrom QT, Gittleman H, Liao P, et al. CBTRUS Statistical Report: primary brain and other central nervous system tumors diagnosed in the United States in 2010-2014. *Neuro Oncol* 2017;19(suppl 5):v1-v88.
- Crocetti E, Trama A, Stiller C, et al. Epidemiology of glial and non-glial brain tumours in Europe. *Eur J Cancer* 2012;48(10):1532-1542.
- Macdonald DR, Cascino TL, Schold SC Jr, Cairncross JG. Response criteria for phase II studies of supratentorial malignant glioma. *J Clin Oncol* 1990;8(7):1277-1280.
- Wen PY, Macdonald DR, Reardon DA, et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J Clin Oncol* 2010;28(11):1963-1972.
- Chang SM, Wen PY, Vogelbaum MA, Macdonald DR, van den Bent MJ. Response Assessment in Neuro-Oncology (RANO): more than imaging criteria for malignant glioma. *Neurooncol Pract* 2015;2(4):205-209.
- De Witt Hamer PC, Hendriks EJ, Mandonnet E, Barkhof F, Zwinderman AH, Duffau H. Resection probability maps for quality assessment of glioma surgery without brain location bias. *PLoS One* 2013;8(9):e73353.
- Müller DMJ, Robe PAJT, Eijgelhaar RS, et al. Comparing glioblastoma surgery decisions between teams using brain maps of tumor locations, biopsies, and resections. *JCO Clin Cancer Inform* 2019;3(3):1-12.
- Stupp R, Brada M, van den Bent MJ, Tonn JC, Pentheroudakis G; ESMO Guidelines Working Group. High-grade glioma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2014;25(suppl 3):iii93-iii101.
- Visser M, Müller DMJ, van Duijn RJM, et al. Inter-rater agreement in glioma segmentations on longitudinal MRI. *Neuroimage Clin* 2019;22:101727.
- Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging* 2015;34(10):1993-2024.
- Crimi A, Bakas S, Kuijff H, Menze B, Reyes M, eds. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. *Proceedings of BrainLes 2017*. Vol 10670. Cham, Switzerland: Springer, 2018; 149-489.



12. Gibson E, Li W, Sudre C, et al. NiftyNet: a deep-learning platform for medical imaging. *Comput Methods Programs Biomed* 2018;158:113–122.
13. AlBadawy EA, Saha A, Mazurowski MA. Deep learning for segmentation of brain tumors: impact of cross-institutional training and testing. *Med Phys* 2018;45(3):1150–1158.
14. Bakas S, Akbari H, Sotiras A, et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data* 2017;4:170117.
15. Dar SU, Yurt M, Karacan L, Erdem A, Erdem E, Cukur T. Image synthesis in multi-contrast MRI with conditional generative adversarial networks. *IEEE Trans Med Imaging* 2019;38(10):2375–2388.
16. Iglesias JE, Konukoglu E, Zikic D, Glocker B, Van Leemput K, Fischl B. Is synthesizing MRI contrast useful for inter-modality analysis? *Med Image Comput Comput Assist Interv* 2013;16(Pt 1):631–638.
17. Pan Y, Liu M, Lian C, Zhou T, Xia Y, Shen D. Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, eds. *International conference on medical image computing and computer-assisted intervention – MICCAI 2018*. Vol 11072. Cham, Switzerland: Springer, 2018; 455–463.
18. Havaii M, Guizard N, Chapados N, Bengio Y. HeMIS: hetero-modal image segmentation. In: Ourselin S, Joskowicz L, Sabuncu M, Unal G, Wells W, eds. *International conference on medical image computing and computer-assisted intervention – MICCAI 2016*. Vol 9901. Cham, Switzerland: Springer, 2016; 469–477.
19. 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.
20. Kamnitsas K, Ferrante E, Parisot S, et al. DeepMedic for brain tumor segmentation. In: Crimi A, Menze B, Maier O, Reyes M, Winzeck S, Handels H, eds. *Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Proceedings of BrainLes 2016*. Vol 10154. Cham, Switzerland: Springer, 2016; 138–149.
21. Eijgelaar RS, Bruynzeel AME, Lagerwaard FJ, et al. Earliest radiological progression in glioblastoma by multidisciplinary consensus review. *J Neurooncol* 2018;139(3):591–598.
22. Eijgelaar R, De Witt Hamer PC, Peeters CFW, Barkhof F, van Herk M, Witte MG. Voxelwise statistical methods to localize practice variation in brain tumor surgery. *PLoS One* 2019;14(9):e0222939.
23. Müller DMJ, Robe PA, Ardon H, et al. Quantifying eloquent locations for glioblastoma surgery using resection probability maps. *J Neurosurg* 2020. Published online April 3, 2020. Accessed May 15, 2020.
24. Louis DN, Ohgaki H, Wiestler OD, et al. The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol (Berl)* 2007;114(2):97–109.
25. VASARI Research Project. <https://wiki.cancerimagingarchive.net/display/Public/VASARI+Research+Project>. Accessed February 27, 2019.
26. Avants BB, Tustison N, Song G. Advanced Normalization Tools (ANTS). *Insight J* 2009;1–35. <http://hdl.handle.net/10380/3113>.
27. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging* 2010;29(6):1310–1320.
28. Avants BB, Tustison NJ, Wu J, Cook PA, Gee JC. An open source multi-variate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics* 2011;9(4):381–400.
29. Milletari F, Navab N, Ahmadi S. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. Presented at the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, Calif, October 25–28, 2016.
30. Kingma DP, Ba J. Adam: a method for stochastic optimization. *ArXiv 1412.6980* [preprint] <https://arxiv.org/abs/1412.6980>. Posted December 22, 2014. Accessed December 7, 2019.
31. Nyúl LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. *IEEE Trans Med Imaging* 2000;19(2):143–150.
32. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging* 2015;15(1):29.
33. R Core Team. R: a language and environment for statistical computing. Vienna, Austria, 2019. <https://www.r-project.org/>. Accessed December 7, 2019.
34. Bakas S, Menze B, Reyes M, et al. Pre-conference Proceedings of the International Multimodal Brain Tumor Segmentation (BraTS) Challenge 2017. 2017.
35. Perkuhn M, Stavrinou P, Thiele F, et al. Clinical Evaluation of a Multi-parametric Deep Learning Model for Glioblastoma Segmentation Using Heterogeneous Magnetic Resonance Imaging Data From Clinical Routine. *Invest Radiol* 2018;53(11):647–654.
36. Freyschlag CF, Krieg SM, Kerschbaumer J, et al. Imaging practice in low-grade gliomas among European specialized centers and proposal for a minimum core of imaging. *J Neurooncol* 2018;139(3):699–711.
37. Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with gaussian edge potentials. *ArXiv 1210.5644* [preprint] <http://arxiv.org/abs/1210.5644>. Posted October 20, 2012. Accessed November 10, 2019.
38. Kamnitsas K, Bai W, Ferrante E, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Crimi A, Bakas S, Kuijf H, Menze B, Reyes M, eds. *Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Proceedings of BrainLes 2017*. Vol 10670. Cham, Switzerland: Springer, 2018; 450–462.