frontiers in Neuroscience

ORIGINAL RESEARCH
published: 05 November 2019
doi: 10.3389/fnins.2019.01182

Divide and Conquer: Stratifying Training Data by Tumor Grade Improves Deep Learning-Based Brain Tumor Segmentation

Michael Rebsamen[1,2], Urspeter Knecht[3], Mauricio Reyes[4], Roland Wiest[1], Raphael Meier[1†] and Richard McKinley[1†]*

[1] Support Center for Advanced Neuroimaging (SCAN), University Institute of Diagnostic and Interventional Neuroradiology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland, [2] Graduate School for Cellular and Biomedical Sciences, University of Bern, Bern, Switzerland, [3] Institute for Surgical Technology and Biomechanics, University of Bern, Bern, Switzerland, [4] Healthcare Imaging A.I. Lab, Insel Data Science Center, Inselspital, Bern University Hospital, Bern, Switzerland

It is a general assumption in deep learning that more training data leads to better performance, and that models will learn to generalize well across heterogeneous input data as long as that variety is represented in the training set. Segmentation of brain tumors is a well-investigated topic in medical image computing, owing primarily to the availability of a large publicly-available dataset arising from the long-running yearly Multimodal Brain Tumor Segmentation (BraTS) challenge. Research efforts and publications addressing this dataset focus predominantly on technical improvements of model architectures and less on properties of the underlying data. Using the dataset and the method ranked third in the BraTS 2018 challenge, we performed experiments to examine the impact of tumor type on segmentation performance. We propose to stratify the training dataset into high-grade glioma (HGG) and low-grade glioma (LGG) subjects and train two separate models. Although we observed only minor gains in overall mean dice scores by this stratification, examining case-wise rankings of individual subjects revealed statistically significant improvements. Compared to a baseline model trained on both HGG and LGG cases, two separately trained models led to better performance in 64.9% of cases ($p < 0.0001$) for the tumor core. An analysis of subjects which did not profit from stratified training revealed that cases were missegmented which had poor image quality, or which presented clinically particularly challenging cases (e.g., underrepresented subtypes such as IDH1-mutant tumors), underlining the importance of such latent variables in the context of tumor segmentation. In summary, we found that segmentation models trained on the BraTS 2018 dataset, stratified according to tumor type, lead to a significant increase in segmentation performance. Furthermore, we demonstrated that this gain in segmentation performance is evident in the case-wise ranking of individual subjects but not in summary statistics. We conclude that it may

OPEN ACCESS

Edited by:
*Bjoern Menze,
Technical University of
Munich, Germany*

Reviewed by:
*Roberto Viviani,
University of Innsbruck, Austria
Benedikt Wiestler,
Technical University of
Munich, Germany*

***Correspondence:**
*Michael Rebsamen
michael.rebsamen@insel.ch*

[†]These authors share senior authorship

Specialty section:
*This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience*

Received: *30 April 2019*
Accepted: *18 October 2019*
Published: *05 November 2019*

Citation:
*Rebsamen M, Knecht U, Reyes M,
Wiest R, Meier R and McKinley R
(2019) Divide and Conquer: Stratifying
Training Data by Tumor Grade
Improves Deep Learning-Based Brain
Tumor Segmentation.
Front. Neurosci. 13:1182.
doi: 10.3389/fnins.2019.01182*

be useful to consider the segmentation of brain tumors of different types or grades as separate tasks, rather than developing one tool to segment them all. Consequently, making this information available for the test data should be considered, potentially leading to a more clinically relevant BraTS competition.

Keywords: magnetic resonance imaging, brain tumors, automatic segmentation, deep learning, training strategy, data stratification

1. INTRODUCTION

Gliomas are primary brain tumors which arise from glial cells. According to the World Health Organization (WHO) classification of tumors of the central nervous system (CNS) (Louis et al., 2016), they can be grouped into different tumor grades based on the underlying histology and molecular characteristics. Increasing tumor grade indicates the increasing malignancy of the tumor. Glioma are managed depending on grade, with treatment strategies ranging from tumor resection followed by combined radio- and chemotherapy to a "watch and wait" approach (Stupp et al., 2005; Grier, 2006). Glioblastoma are the most aggressive type of glioma (WHO grade IV) and make up 45% of all gliomas (Ostrom et al., 2014). The prime imaging technique in brain tumor diagnostics is Magnetic Resonance Imaging (MRI) (Essig et al., 2012). Standard acquisition protocols used to perform initial diagnosis and treatment monitoring include T1-weighted, T1-weighted gadolinium-enhanced, T2-weighted, and T2-weighted with fluid attenuated inversion recovery (FLAIR) sequences (Wen et al., 2010; Ellingson et al., 2015). The typical radiological appearance of a glioblastoma features a disrupted blood-brain barrier causing ring-enhancing lesions with central necrosis and peritumoral edema. In contrast, low-grade astrocytic tumors exhibit typically no contrast enhancement and are missing central necrosis (Pierallini et al., 1997).

In the case of glioblastoma, recent studies led to the discovery of a profound genetic heterogeneity among, and even within, tumors (Verhaak et al., 2010; Sottoriva et al., 2013). It has been shown that the underlying genetic and molecular heterogeneity can be associated with variations in imaging phenotype such as changes in tumor compartment volumes (Lai et al., 2013; Grossmann et al., 2016), contrast enhancement (Carrillo et al., 2012; Treiber et al., 2018), radiomic signatures (Gevaert et al., 2014), and tumor location (Carrillo et al., 2012; Ellingson et al., 2012). The imaging appearance of glioblastoma can further be altered by treatment causing radiation necrosis (Mullins et al., 2005) and pseudoprogression and -response (Hygino da Cruz et al., 2011), respectively. As a consequence, a machine learning segmentation algorithm needs to be capable of generalizing across this heterogeneity of glioblastoma imaging phenotypes.

Brain tumor segmentation is a well-investigated topic with a vast amount of available methods and yearly organized MICCAI Brain Tumor Segmentation (BraTS) Challenges since the year 2012 (Menze et al., 2015; Bakas et al., 2017c), serving as a public platform for algorithm comparison. With the rise of deep learning, brain tumor segmentation methods experienced significant gains in performance (Bakas et al., 2018). One of the central promises of deep learning methods is that they can be fed with raw data and are capable of automatically uncovering the underlying representation relevant for the task at hand (e.g., segmentation) from that data (LeCun et al., 2015). As a consequence, the time-consuming and error-prone manual engineering of features traditionally used in machine learning has been rendered obsolete. Recently, it was shown for vision tasks that model performance increases logarithmically based on volume of training data (Sun et al., 2017). This aligns with the general notion that more training data leads to a better generalization of a machine learning algorithm. Within the context of BraTS Challenges, deep learning methods are usually trained *ad hoc* on all of the available data, disregarding underlying latent factors such as genetic characteristics or even tumor grades. Although the tumor type is available to the challenge participants for the training data, this information is withheld for the validation and test data. Since part of the BraTS dataset is coming from *The Cancer Imaging Archive* (TCIA) (Bakas et al., 2017a,b,c), additional relevant information such as e.g., patient's gender, mutation subtypes [Isocitrate dehydrogenase (IDH), 1p19q co-deletion] and methylation status of MGMT-promotor could potentially be added as well.

The metric of choice for algorithm comparison in biomedical image segmentation challenges is the Dice coefficient, which was used in 92% of the 383 segmentation tasks reported in Maier-Hein et al. (2018). Predominantly, the Dice coefficient is reported in terms of summary statistics (mean/median) over patient cases and model comparison is performed on the basis of such summary statistics (metric-based ranking). Recently, the BraTS Challenge adopted a case-based ranking scheme. While metric-based rankings lead to more robust rankings than case-based rankings (Maier-Hein et al., 2018), it can be argued that distinct performance differences for individual patients may be obfuscated.

We hypothesize that deep learning methods for brain tumor segmentation can be significantly improved by taking into account latent factors along with tumor image appearance during model training. The purpose of this study is to demonstrate the impact of including prior knowledge of a particular latent factor (tumor grade) on the performance of a recently published, top-ranked deep learning method (McKinley et al., 2019a). Furthermore, the impact is studied employing both a metric-based and case-based rank analysis.

The idea of leveraging prior information about tumor grades to improve segmentation has been presented as an extended abstract to the *International Conference on Medical Imaging with Deep Learning (MIDL)* along with preliminary results (Meier et al., 2019).

2. MATERIALS AND METHODS

2.1. Study Data

The study is based on publicly-available data of the BraTS 2018 Challenge (Menze et al., 2015; Bakas et al., 2017c). In particular, the training dataset was used, which includes 75 patients with low-grade glioma (LGG) and 210 patients with high-grade glioma (HGG). The imaging data encompasses four MR image sequences (T1-weighted, T1-weighted with contrast agent, T2-weighted, and T2-weighted FLAIR sequences), which are part of the consensus recommendations for a standardized brain tumor imaging protocol in clinical trials (Ellingson et al., 2015). The imaging data stem from 19 different institutions, which relied on different MR scanners and acquisition protocols. Manual segmentations of three tumor compartments were available: contrast-enhancing tumor, non-enhancing/necrosis combined, and edema. The regions which were considered for evaluation in the BraTS 2018 challenge as well as in the study at hand were: contrast-enhancing tumor, tumor core (all compartments except edema), and whole tumor (all compartments). More details on the preprocessing and the evolution of the BraTS dataset can be found in Bakas et al. (2017c).

2.2. Automatic Segmentation

The network architecture used for the automatic segmentation is equivalent to the model ranked third in the BraTS 2018 challenge (McKinley et al., 2019a). In brief, it is a U-net-style structure with densely connected blocks of dilated convolutions. The segmentation is performed slice-wise where the input data includes the two neighboring slices from below and above from all four image modalities (i.e., input dimension is $batch \times 4 \times 5 \times 192 \times 192$). The final segmentation is the result of ensembling the predictions from all three directions (sagittal, axial, and coronal).

In a pre-processing step, the data are first normalized to zero mean and unit variance. Data augmentation consists of a combination of randomly flipping the images along the midline and random rotations [$angle \sim U(-15, +15)$] around all principal axis. Additionally, the standardized voxel intensities are randomly shifted [$amount \sim N(0, 0.5)$] and scaled [$factor \sim N(1, 0.2)$].

The networks were trained with a *focal loss* function, *RMSprop* as optimizer with a cosine-annealing learning rate schedule, and a batch-size of two.

2.3. Stratified Model Training

Three different models were trained independently, each with a five-fold cross-validation: A baseline model with all available training data (number of samples $N = 285$), an HGG-only model ($N = 210$), and an LGG-only model ($N = 75$). Network architecture and hyperparameters were the same for all models which were trained on a *Nvidia GeForce GTX 1080 Ti* GPU with 11GB memory over 80 epochs. Qualitatively, the performance on the validation-set was saturating with no observable overfitting (see **Figure S1**).

2.4. Statistical Analysis

The statistical analysis was performed using *R* with the *stats* package version 3.5.1 (R Core Team, 2018). For comparison of spatial overlap of estimated tumor segmentations with manual ground truth data, the Dice coefficient was used. Segmentation performance in terms of Dice coefficient of the different deep learning models was summarized by descriptive statistics (median, interquartile range). Case-based rank analysis included computation of percentage of improved patient cases for given pairing of deep learning models. The stratified models were compared to the baseline by means of paired difference tests: differences between the cross-validated classifiers were examined on HGG cases only, on LGG cases only, and on the whole dataset (using the combined results of the stratified LGG and HGG classifiers). Non-parametric tests were employed due to the rank-based form of the data. The significance level of the analysis was set to α=0.05.

3. RESULTS

3.1. Quantitative Analysis

Summary statistics for the segmentation performance in terms of Dice coefficient are shown in **Figure 1**. The baseline model reached a median Dice of 0.841 (1.5 × IQR = 0.465–1.000) for the contrast enhancing compartment, 0.899 (0.554–1.000) for the core, and 0.920 (0.786–1.000) for the whole tumor. Comparable, the combined results from the separately trained HGG/LGG models were 0.838 (0.415–1.000) for contrast enhancing, 0.902 (0.584–1.000) for core, and 0.916 (0.800–1.000) for tumor.

The combined results of the two separately trained models showed an improvement for the segmentation of the tumor core in 64.9% ($p < 0.0001$) of the subjects compared to the baseline model (**Table 1**). No statistically significant changes were observed for the other compartments. This performance gain originates primarily from the HGG cases where 70.3% of the subjects showed an improved segmentation for the tumor core and 58.5% of the subjects also for the contrast enhancing compartment. From the 183 subjects that showed an improved segmentation of the core, 26 increased by a Dice of 0.1 or more. Conversely, from the 99 subjects with a declined performance, 21 decreased by a Dice of −0.1 or more (**Figure 2**).

3.2. Qualitative Analysis of Selected Cases

From **Table 1** it is evident that, especially for high-grade glioma, stratified training leads to improved segmentation performance. In order to further investigate this aspect, a visual review of selected cases was performed. To identify cases mostly affected by the stratified training, Dice coefficients between the segmentations of the two models (baseline vs. HGG) were calculated. Cases with a Dice agreement < 0.8 of the tumor core between the baseline and stratified models were selected for a qualitative manual inspection followed by a review with a board-certified neuroradiologist with more than 8 years of experience in brain tumor diagnostics. In order to render the visual review more systematic, we define three categories of causes for variability in tumor segmentation performance: 1. The input data generated by the imaging process, which is affected by the idiosyncrasies of the MR scanner, potential image artifacts and patient motion, and image preprocessing. 2. The manual ground truth segmentation. 3. The tumor

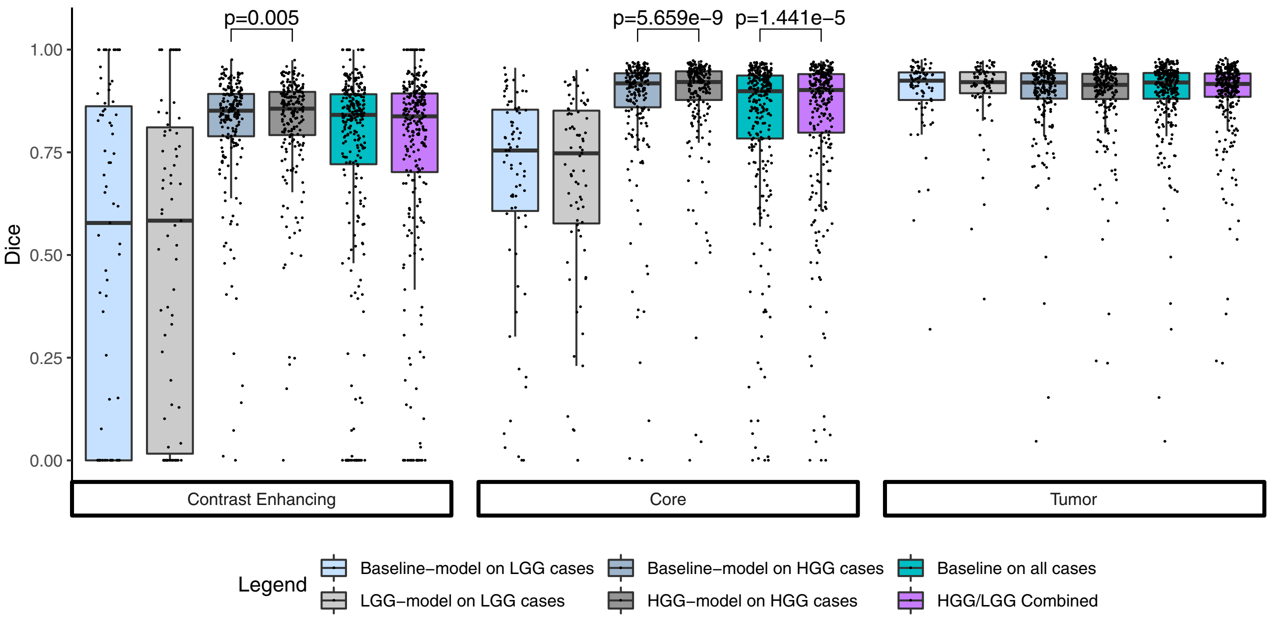

FIGURE 1 | Summary statistics for the segmentation of the three compartments by means of a Tukey boxplot. *p*-values indicate statistically significant ($p < 0.05$) improvements determined by a one-sided Wilcoxon signed rank test.

phenotype (e.g., IDH-mutant tumor, presence of intratumoral hemorrhage, or cystic components) which causes distinctively different image appearances.

In **Figure 3** the obtained Dice coefficients between the segmentation results of the HGG model for the tumor core and the ground truth were plotted against the Dice coefficients between the results of the HGG model and the segmentation of the baseline model, which was trained on all available data. We can broadly define four different territories in the scatterplot: The upper right corner which contains cases for which both models achieved high segmentation performance. If we move to the upper left corner, we encounter cases for which the HGG model achieved high segmentation performance with discrepancies when compared to the results of the baseline model. If we move from the upper right corner to the lower right corner, we encounter cases for which the HGG model agreed with the segmentation result of the baseline model but did not agree with the ground truth result. Finally, the lower left corner contains cases for which the segmentation results of the HGG model did neither agree with the ground truth nor with the segmentation of the baseline model. The corresponding scatterplots for the other two compartments can be found in **Figures S2, S3**. The identified outlier cases are listed in **Table 2** with the segmentation performance of the two models and an assessment category. Below we present the observations based on a visual review for a selection of the identified outliers. Visualizations for the remaining outliers can be found in **Figures S4–S14**.

Brats18_2013_21_1 (**Figure 4**). The baseline model provided superior performance for segmenting the tumor core in this HGG example. The lesion exhibits a large non-enhancing tumor mass (typically seen in LGG) and we speculate that the presence of LGG cases in the baseline model led to the improved tumor core segmentation performance when compared to the HGG model's result. The appearance of the tumor is further complicated by the presence of cystic components, which exhibit a homogeneous signal that is strongly hypointense in T1-weighted and hyperintense in T2-weighted images.

TABLE 1 | Ratio in % of better performing subjects compared to baseline.

	CE		Core		Tumor	
	% Subjects	*p*	% Subjects	*p*	% Subjects	*p*
LGG vs. Baseline	41.7	0.877	49.3	0.454	54.7	0.208
HGG vs. Baseline	**58.4**	**0.005**	**70.3**	**5.659e-09**	46.7	0.877
HGG/LGG vs. Baseline	54.6	0.127	**64.9**	**1.441e-05**	48.8	0.725

Statistical significance is determined by a one-sided Wilcoxon signed rank test. Bold numbers indicate statistically significant (p < 0.05) results. CE: contrast enhancing.

Brats18_2013_25_1 (**Figure 5**). Both models failed to segment the tumor core for this HGG case. The tumor core contains strongly hypointense areas in the T2-weighted and FLAIR images with corresponding heterogeneous signal intensity in the T1-weighted image. When considering the T1/T1c-weighted images, one can observe the presence of recruited blood vessels. This image appearance may indicate the presence of an intratumoral hemorrhage.

Brats18_CBICA_AXJ_1 (**Figure 6**). The segmentation of the core from the HGG model is closer to ground truth. The tumor was indicated to be an HGG. However, the provided ground truth segmentation seems to be missing part of the

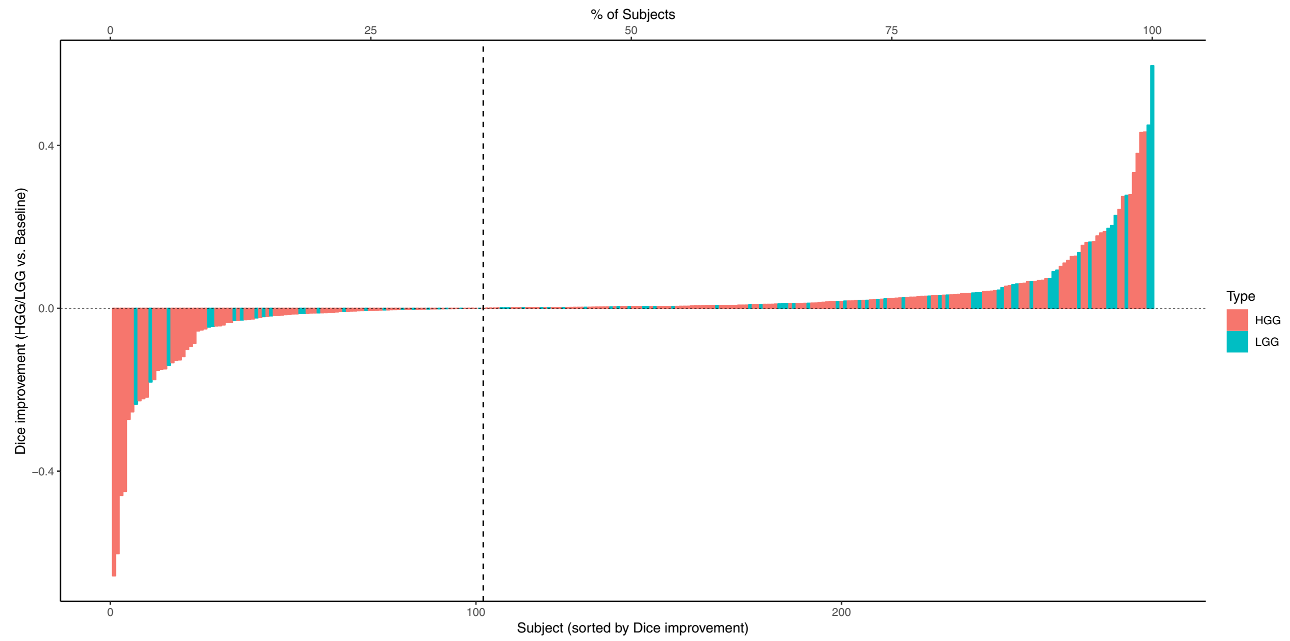

FIGURE 2 | Absolute change of Dice coefficients of the tumor core for each subject. Positive changes were observed for the subjects on the right side of the dashed vertical line.

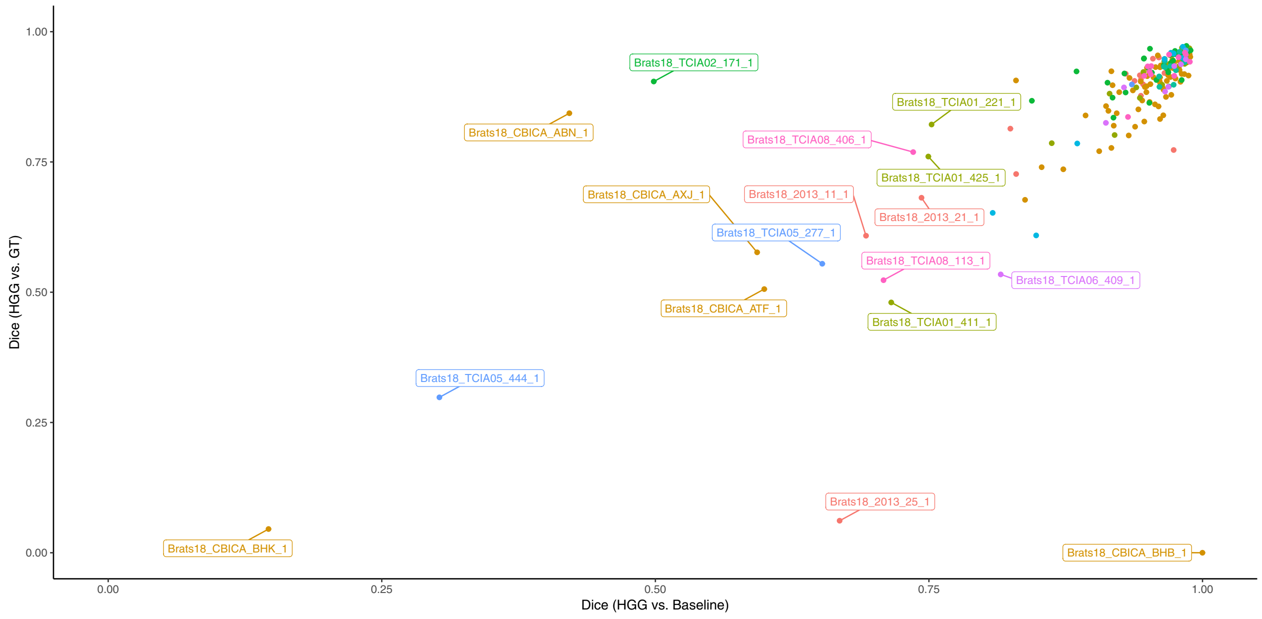

FIGURE 3 | Performance of the HGG-only model for the tumor core (y-axis) and agreement with the baseline model (x-axis). Subjects with a label were visually reviewed. Colors indicate the center (2013, CBICA, TCIA01-08).

tumor mass in the frontal lobe. Furthermore, we argue that a large part of the lesion corresponds to non-enhancing tumor rather than edema. We base this assumption on the heterogeneous appearance in the T2-weighted images and more importantly the strong cortical space-occupying effect together with a distortion of the gray/white matter junction. In contrast, edema would preserve the gray/white matter junction as well as the cortical ribbon and propagate along the white matter fiber tracts. A possible alternative for ground truth is shown in **Figure 6**.

TABLE 2 | Performance of selected cases for the two models.

Subject	Assessment	Dice Baseline-model			Dice HGG-model		
		CE	Core	Tumor	CE	Core	Tumor
Brats18_2013_11_1	1	0.14	0.45	0.90	0.17	0.61	0.89
Brats18_2013_21_1	3	0.80	0.83	0.94	0.76	0.68	0.94
Brats18_2013_25_1	3	0.18	0.10	0.90	0.25	0.06	0.90
Brats18_CBICA_ABN_1	2	0.84	0.41	0.82	0.79	0.84	0.83
Brats18_CBICA_ATF_1	3	0.69	0.73	0.69	0.65	0.51	0.63
Brats18_CBICA_AXJ_1	2	0.79	0.35	0.90	0.79	0.58	0.90
Brats18_CBICA_BHB_1	2	0.00	0.00	0.15	0.00	0.00	0.24
Brats18_CBICA_BHK_1	2	0.01	0.00	0.05	0.25	0.05	0.24
Brats18_TCIA01_221_1	2	0.76	0.88	0.95	0.48	0.82	0.95
Brats18_TCIA01_411_1	1	0.07	0.24	0.71	0.23	0.48	0.64
Brats18_TCIA01_425_1	–	0.26	0.58	0.75	0.68	0.76	0.78
Brats18_TCIA02_171_1	2	0.89	0.47	0.95	0.89	0.90	0.95
Brats18_TCIA04_343_1	2	0.69	0.73	0.74	0.59	0.61	0.66
Brats18_TCIA05_277_1	3	0.42	0.37	0.85	0.56	0.55	0.90
Brats18_TCIA05_444_1	3	0.39	0.96	0.94	0.54	0.30	0.89
Brats18_TCIA06_409_1	–	0.52	0.53	0.89	0.50	0.53	0.86
Brats18_TCIA08_113_1	1	0.91	0.36	0.97	0.79	0.52	0.92
Brats18_TCIA08_406_1	1	0.65	0.63	0.88	0.68	0.77	0.90

Assessment after a qualitative review with a neuroradiologist. Assessment 1: Issue with input image quality, 2: Possible problem with ground truth, 3: Special phenotype, GT: ground truth, CE: contrast enhancing.

Brats18_CBICA_BHB_1 (**Figure 7**). Both models failed completely to segment the lesion for this HGG case. However, the provided ground truth segmentation seems to overestimate the presence of edema. While we agree on the whole tumor segmentation, we argue that the present T2-weighted hyperintensity indicates the presence of non-enhancing tumor rather than edema. Similarly to case Brats18_CBICA_AXJ_1 the gray/white matter junction is distorted. This is especially evident when considering the unaffected contralateral hemisphere. The poor segmentation performance of both models might be the result of an under-representation of training samples with such a subtle tumor core which is potentially ambiguously labeled in other cases as well.

Brats18_TCIA01_221_1 (**Figure 8**). The baseline model provided the better tumor core segmentation for this HGG case. However, when comparing the segmentation of the contrast-enhancing tumor of the HGG model, we argue that the ground truth segmentation slightly undersegments it. This is clearly visible for the enhancing rim next to the midline.

Brats18_TCIA01_425_1 (**Figure 9**). The baseline model underestimated the subtle contrast-enhancement of this HGG case. We can speculate that in the situation of subtle enhancements the baseline model was biased more toward segmenting a tumor core with small enhancing foci, whereas the HGG model was capable of delineating the full extent of the contrast-enhancement.

Brats18_TCIA05_444_1 (**Figure 10**). The baseline model provided a better segmentation than the HGG model for this case. The tumor was indicated to be an HGG. The location of the tumor in the frontal lobe and its appearance exhibiting focal contrast enhancements and a large non-enhancing tumor mass are suspicious of a potential IDH-mutant glioblastoma. This would imply that it initially emerged from an LGG (called secondary glioblastoma). Applying the LGG model to the case significantly outperforms the HGG model (**Figure 10**), which would support the hypothesis of a mutated LGG.

While the previous analysis of cases was to some extent speculation, we can nevertheless condense three main, factual observations from it: First, individual segmentation results are strongly affected by the composition of the segmentation model's training data. Second, depending on the underlying factors that caused a given image appearance and segmentation ground truth, a given subset of the training data can actually improve the segmentation result compared to a baseline trained on all data. Third, disagreement (or joint failure) among segmentation models trained on different subsets of training data (**Figure 3**) may actually help in the identification of these underlying factors. Among the manually reviewed 18 cases with a large deviation between the two models, we observed issues with the input images (4 cases), potentially arguable ground truth (7 cases), and special imaging phenotypes (5 cases). Arguable ground truth is often attributed to edema that could be labeled as tumor core instead. Edema typically propagates along white matter and spares cortical ribbons as well as deep gray matter structures (Pope et al., 2005), while non-enhancing tumor leads to a distortion of the gray/white matter junction [cf. BRATS18_CBICA_BHB_1 (**Figure 7**) FLAIR with the case presented in Figure 3 of Lasocki and Gaillard, 2019].

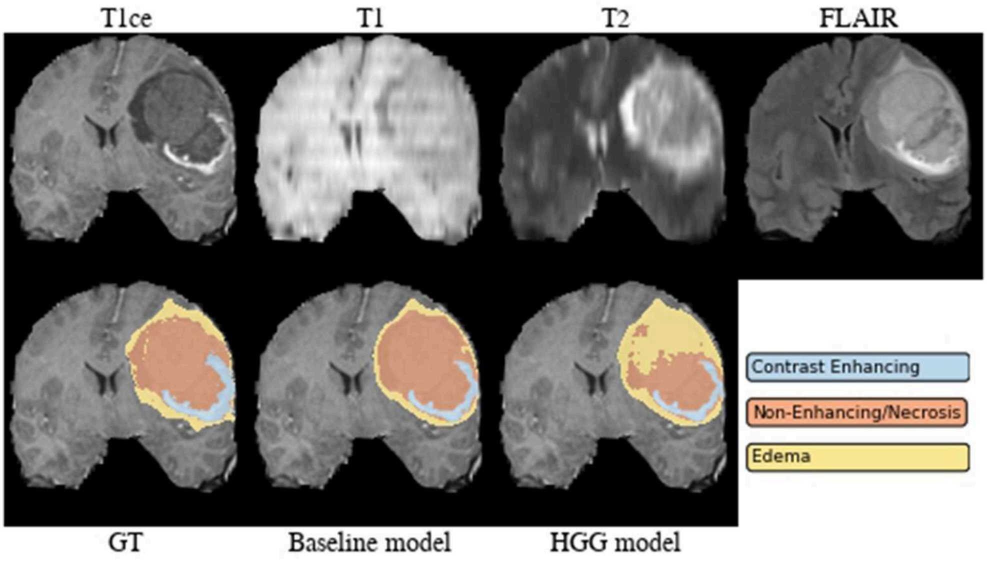

FIGURE 4 | Brats18_2013_21_1.

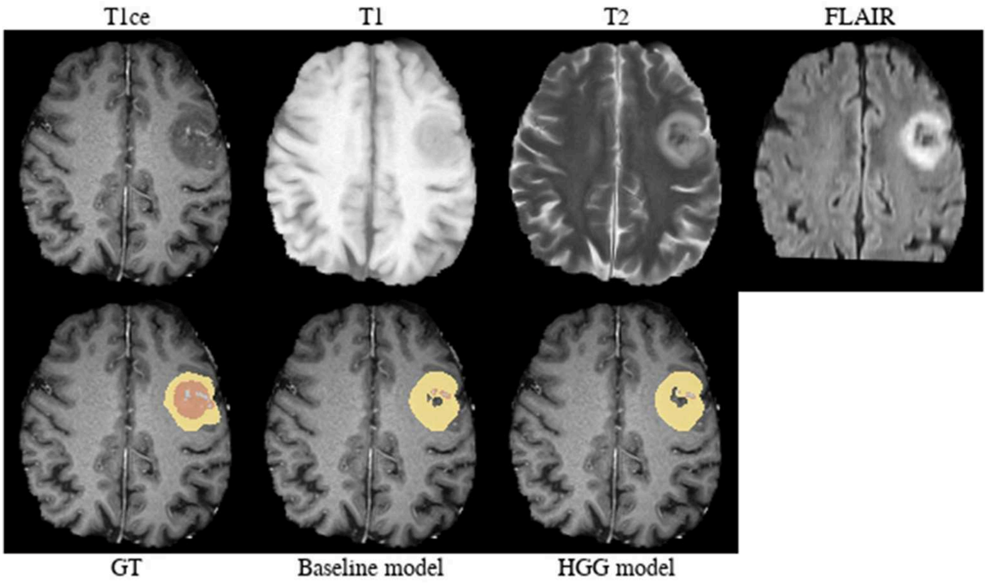

FIGURE 5 | Brats18_2013_25_1.

4. DISCUSSION

The title of the manuscript contains the phrase "Divide and Conquer," where "Divide" refers to the stratification of training data. Data stratification and subsequent model training was employed as a simple, straightforward technique to include prior knowledge. We have proposed two ways of how to use data stratification to "conquer" brain tumor segmentation: First, the targeted application of a specialized model (HGG model) to the respective data (HGG test case). Second, the utilization of disagreement among specialized models' outputs and ground truth segmentations to identify outliers and possible latent factors hampering generalization.

Implicitly adding prior information to the models by stratifying the data by tumor type (HGG and LGG) seems to be beneficial for the segmentation of the tumor core for high-grade glioma. Yet, the LGG-only model, which was trained with fewer samples ($N = 75$) compared to the baseline model ($N = 285$), showed no statistically significant deterioration of the segmentation performance. A statistically significant

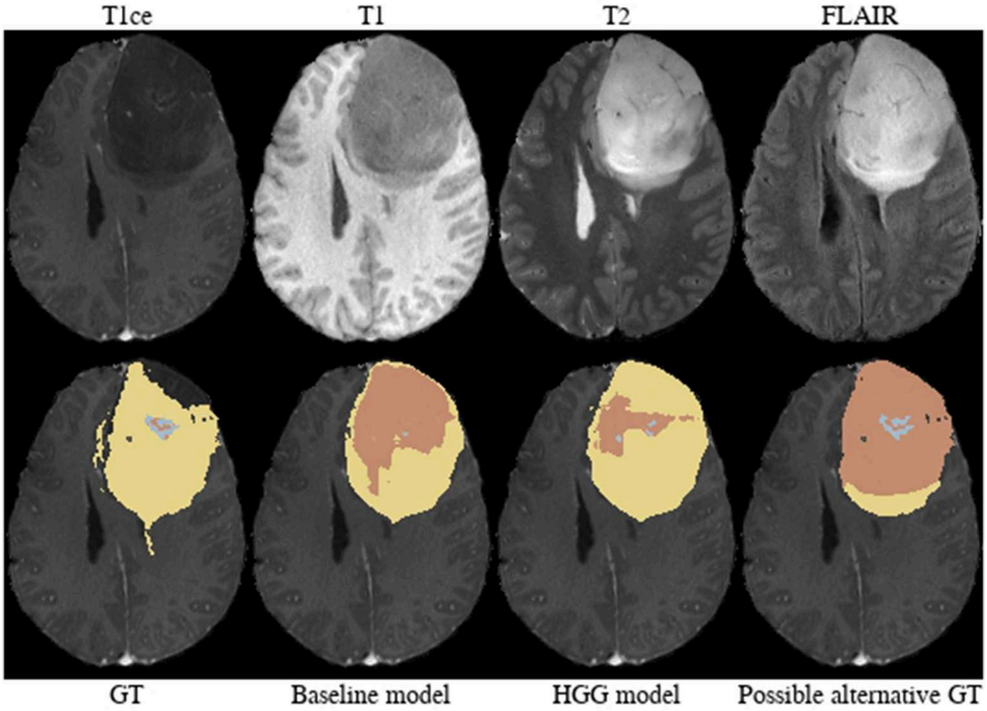

FIGURE 6 | Brats18_CBICA_AXJ_1.

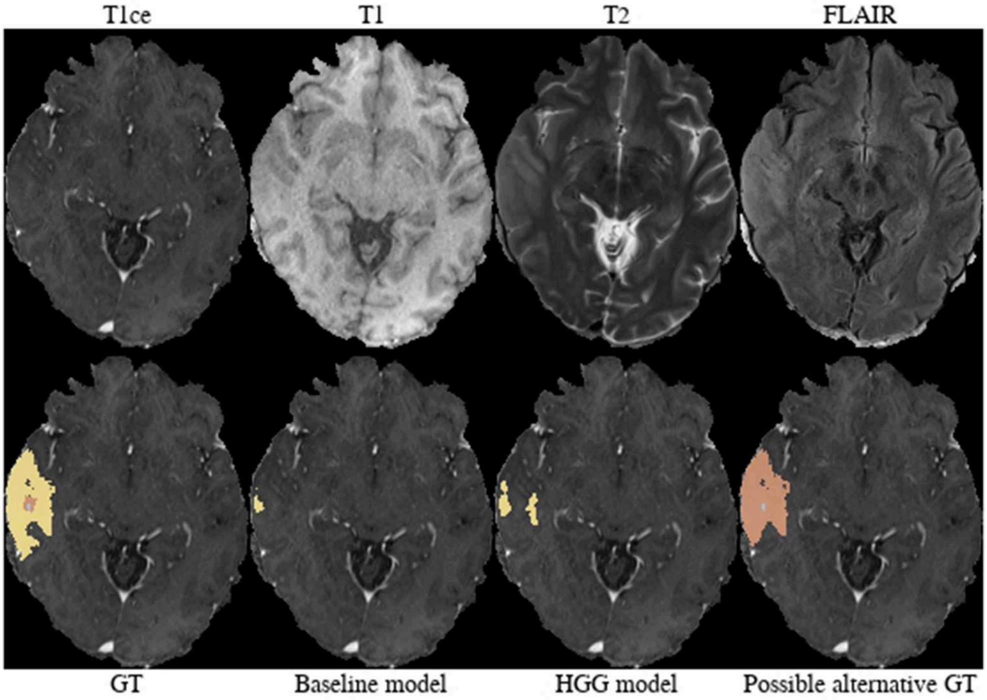

FIGURE 7 | Brats18_CBICA_BHB_1.

improvement in 64.9% of the subjects for the tumor core is accompanied by a non-significant improvement of 54.6% for contrast enhancing and non-significant decrease (only 48.8% better-ranked subjects) for the whole tumor. It has been shown in multiple studies (Asari et al., 1994; Wiestler et al., 2016; Hsieh et al., 2017) that HGG and LGG tend to exhibit different

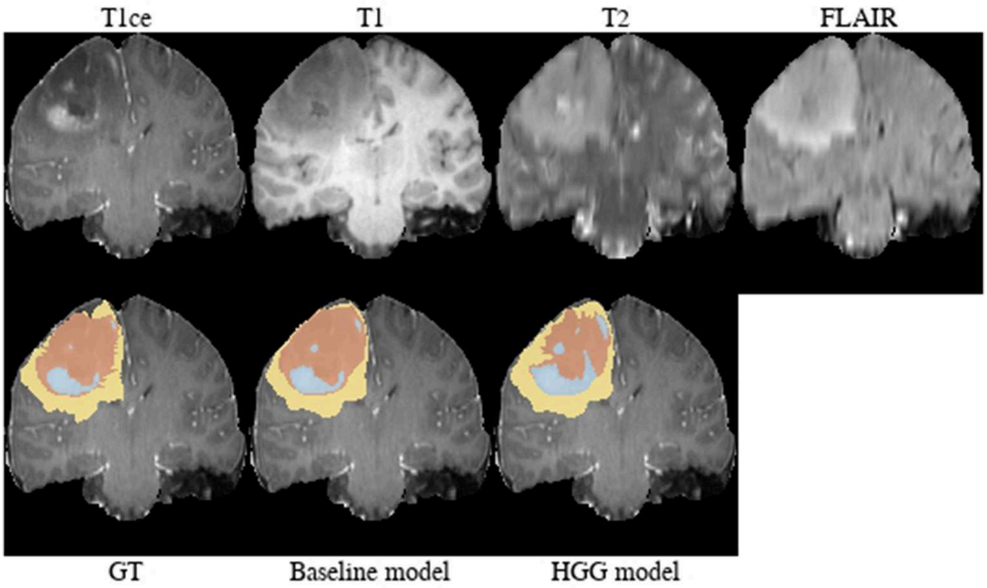

FIGURE 8 | Brats18_TCIA01_221_1.

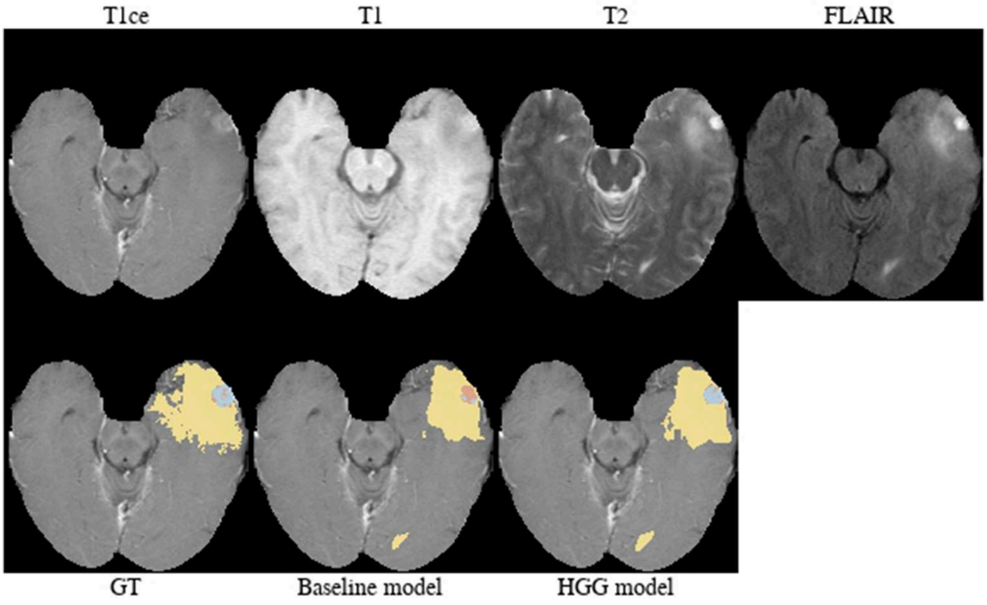

FIGURE 9 | Brats18_TCIA01_425_1.

qualitative and quantitative imaging features in structural MRI, involving heterogeneity of contrast enhancement, cystic components, intratumoral hemorrhage, and necrosis, which in context of tumor segmentation affect the definition of the tumor core greatly. Therefore, the stratification of the training data into HGG and LGG yields subsets with more homogeneous and consistent definitions of the tumor core. However, we presented also exceptions [e.g., BRATS18_2013_21_1 (**Figure 4**) in section 3.2] which actually profit from training data of opposite tumor grade.

In addition to improving segmentation performance, deep learning models trained on stratified data can be used to drive exploration of the training data. In section 3.2 we demonstrated that the disagreement between such models in relation to the ground truth data can assist in the identification of latent factors (e.g., imaging phenotypes) which may pose significant challenges in a deep learning model's capability to generalize across the complete problem domain. We argue that especially in a pathology as complex as brain cancer, the identification of such latent factors and their proper treatment in a deep learning

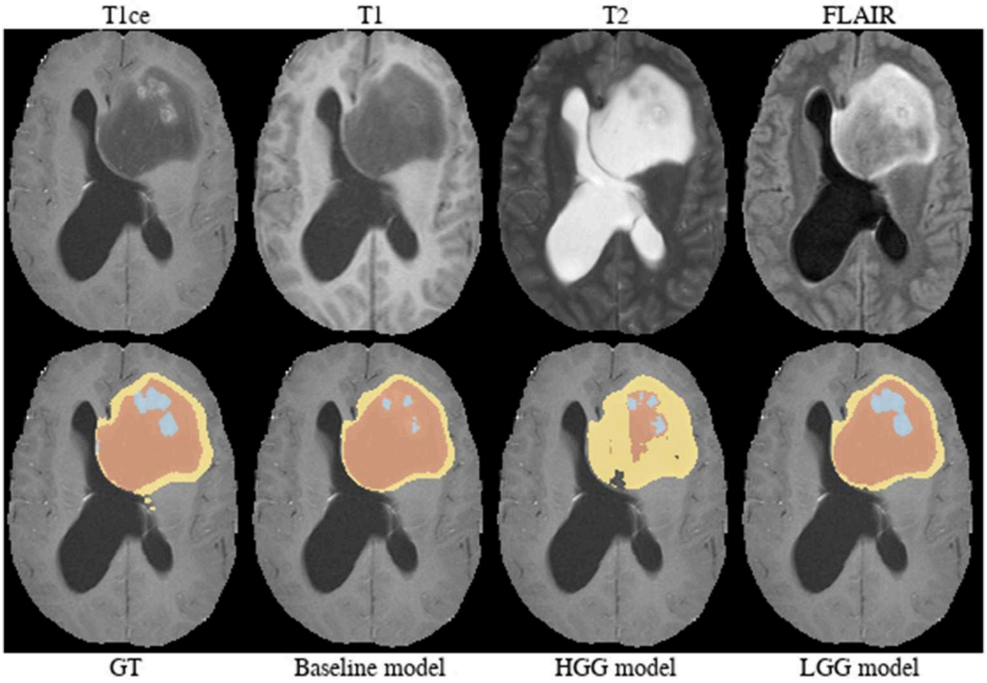

FIGURE 10 | Brats18_TCIA05_444_1.

model is of utmost importance to guarantee robust segmentation performance that satisfies clinical needs. In section 4.1 we provide propositions on how latent factors such as the tumor type could be treated in deep learning segmentation models beyond simple data stratification.

Our results demonstrated the potential of summary statistics (e.g., mean or median) to obfuscate significant differences between distributions of segmentation performance measures (e.g., Dice coefficient). These significant differences can be revealed through the calculation of a case-based ranking. Furthermore, case-based ranking enables the straightforward application of nonparametric statistics to detect significant differences with the advantage of more limited assumptions regarding the distribution of the data when compared to parametric statistics, and robustness to outliers. Case-based ranking also follows the narrative of precision medicine in which the identification of subpopulations of patients, who benefit from a medical intervention, based on experimental observations is central. It enables a more fine-grained analysis on the level of the patient and potentially an identification of patient subpopulations relevant for the task at hand.

Previously, Pereira et al. (2016) trained on data stratified into HGG and LGG. They employed two different Convolutional Neural Network architectures for patch-wise segmentation of HGG and LGG. In contrast, we hypothesized and demonstrated that a mere stratification of the training data into HGG and LGG without any changes to architectures or hyperparameters can lead to improved segmentation performance. Furthermore, their focus was on an ablation study of methodological components with respect to their two grade-specific architectures and their results were based on the BraTS 2013 Leaderboard dataset (21 HGG, 4 LGG cases) and BraTS 2013 Challenge dataset (10 HGG cases).

4.1. Outlook

With the rise of precision medicine and tailored therapies, the consideration of patient-specific information (e.g., genetics) becomes ubiquitous (Giardino et al., 2017). Leveraging data from multiple sources remains a challenge for the next generation imaging technologies (Kim et al., 2016), potentially requiring to rethink the *one size fits all* concept. For automatic brain tumor segmentation, various architectural and conceptual changes are imaginable beyond simple data stratification strategies.

By completely separating the data, each of the individual models has fewer data available for training, although with the benefit of a less heterogeneous domain (only one tumor type). Instead of implicitly adding the prior information of the tumor to the data by stratification, an alternative approach could be to explicitly add this information as input to the network. Particularly the first layers of the network might be less susceptible to the tumor type as filters for representation learning could share commonalities between both domains. By adding the information directly to the input layer or injecting it into the latent feature space might allow the network to intrinsically adapt the segmentation output according to the given tumor type.

A different approach would be to regard the problem of segmenting high-grade and low-grade glioma as a multiple-source adaptation problem. In this setting, the goal is to effectively combine base learners trained on multiple source domains in order to perform a prediction on a target domain, which can be any mixture of the source domains. In our case, the source domains would be subclasses of gliomas: either high-grade and low-grade data, or potentially a more fine-grained subdivision (e.g., WHO grade or classification). The target domain constitutes of a mix of different glioma cases. Recently, a number of theoretic and algorithmic contributions were made in the area of multiple-source adaptation (Hoffman et al., 2018; Zhao et al., 2018), which could be applied in the scenario of learning from multiple disease entities such as brain tumor types or grades.

The clinical importance of brain tumor segmentation for quantitative image analysis will only grow in the near future. Recently, various segmentation methods have been proposed which are capable of accurately delineating brain tumor compartments longitudinally (Weizman et al., 2014; Meier et al., 2016), perform assessment of treatment response (Huber et al., 2017; Kickingereder et al., 2019), are used for the purpose of radiomic analysis (Bakas et al., 2017c), and for performing planning of radiation therapy (Sharp et al., 2014; Herrmann et al., 2018; Agn et al., 2019; Lipkova et al., 2019). It is, therefore, necessary to provide automatic segmentation methods which are capable of robustly generalizing across different types or grades of brain tumors. Our methodology of training deep learning models on stratified training data is a straightforward approach to potentially improve the segmentation performance of already existing learning-based methods with regards to different tumor types.

In the light of our results and the trend toward precision medicine, we encourage challenge organizers to make information on the tumor type or grade available as additional input data, allowing teams to incorporate such prior information into their models.

4.2. Limitations

The evaluation is based solely on the BraTS training dataset (using cross-validation). Results for the official validation set are unknown since the required tumor type is not available for these data. Indeed we acknowledge that the tumor grade is usually not yet available on the first admission. However, we think automatic segmentation models will probably be employed first for retrospective studies, to assess the extent of resection in patients undergoing surgery (Meier et al., 2017), or to assess tumor progression postoperatively (Kickingereder et al., 2019) where tumor grades are usually known. First attempts have been made to classify tumor grades from MRI (Decuyper and Van Holen, 2019), which would allow identification of the correct model from imaging only. Alternatively, one might run such a segmentation algorithm twice: first for a rough identification of the tumor compartments and based on the result (e.g., presence of CE, ratio of compartment volumes, or manual review of the intermediate results by an expert) apply the specific model to get a refined segmentation.

The benefit of stratifying the training data has been shown with the model ranked third in the BraTS 2018 challenge (McKinley et al., 2019a). This particular model was chosen, as it was a top-ranked method in the most recent BraTS challenge (2018) that achieved its results using only a standard GPU and data from the BRATS challenge. The method ranked first (Myronenko, 2019) depended on a GPU with 32 GB of memory (to which most research groups do not have access), while the second-ranked method (Isensee et al., 2018) was co-trained with additional data (not including information about tumor grades). To what extent the proposed approach generalizes to other architectures remains an open question. Other models might suffer more from the reduction of training samples due to the stratification. The proposed architecture is known to be robust to fewer training samples (McKinley et al., 2019b).

5. CONCLUSION

Implicitly adding prior knowledge by dividing data into distinct domains can improve the performance of deep learning-based segmentation methods and compensate for the smaller number of samples available for training a model. The tumor grade has shown to be an important latent factor in the segmentation of gliomas. Comparing the performance of models by case-based ranking statistics may reveal significant differences that are otherwise concealed in summary statistics such as the mean Dice coefficient.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://braintumorsegmentation.org.

ETHICS STATEMENT

The study is based on a publicly available dataset from the Multimodal Brain Tumor Segmentation Challenge 2018 (BraTS, http://braintumorsegmentation.org/).

AUTHOR CONTRIBUTIONS

MReb, RM, RMc, RW, and MRey: design of experiments. MReb: perform experiments, data analysis. UK and RM: analysis of selected cases. MReb, RM, and RMc: results interpretation. MReb and RM: manuscript drafting, RMc: manuscript revision. All authors reviewed and approved the final version of the manuscript.

FUNDING

This work was supported by the Swiss Personalized Health Network (SPHN, project number 2018DRI10), the Swiss National Science Foundation (grant number 169607), and the Swiss Cancer League (grant number KFS-3979-08-2016).

ACKNOWLEDGMENTS

Calculations were performed on UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2019.01182/full#supplementary-material

REFERENCES

Agn, M., Af Rosenschöld, P. M., Puonti, O., Lundemann, M. J., Mancini, L., Papadaki, A., et al. (2019). A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning. *Med. Image Anal.* 54, 220–237. doi: 10.1016/j.media.2019.03.005

Asari, S., Makabe, T., Katayama, S., Itoh, T., Tsuchida, S., and Ohmoto, T. (1994). Assessment of the pathological grade of astrocytic gliomas using an MRI score. *Neuroradiology* 36, 308–310. doi: 10.1007/BF00593267

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017a). Segmentation labels for the pre-operative scans of the TCGA-GBM collection. *The Cancer Imaging Archive*. doi: 10.7937/k9/tcia.2017.klxwjj1q

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). Segmentation labels for the pre-operative scans of the TCGA-LGG collection. *The Cancer Imaging Archive*. doi: 10.7937/k9/tcia.2017.gjq7r0ef

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017c). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *CoRR* abs/1811.02629.

Carrillo, J., Lai, A., Nghiemphu, P., Kim, H., Phillips, H., Kharbanda, S., et al. (2012). Relationship between tumor enhancement, edema, IDH1 mutational status, MGMT promoter methylation, and survival in glioblastoma. *Am. J. Neuroradiol.* 33, 1349–1355. doi: 10.3174/ajnr.A2950

Decuyper, M., and Van Holen, R. (2019). "Fully automatic binary glioma grading based on pre-therapy MRI using 3D convolutional neural networks," *Presented at the International Conference on Medical Imaging with Deep Learning, MIDL 2019* (London).

Ellingson, B. M., Bendszus, M., Boxerman, J., Barboriak, D., Erickson, B. J., Smits, M., et al. (2015). Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. *Neuro Oncol.* 17, 1188–1198. doi: 10.1093/neuonc/nov095

Ellingson, B. M., Cloughesy, T. F., Pope, W. B., Zaw, T. M., Phillips, H., Lalezari, S., et al. (2012). Anatomic localization of o6-methylguanine DNA methyltransferase (MGMT) promoter methylated and unmethylated tumors: a radiographic study in 358 *de novo* human glioblastomas. *NeuroImage* 59, 908–916. doi: 10.1016/j.neuroimage.2011.09.076

Essig, M., Anzalone, N., Combs, S., Dörfler, A., Lee, S.-K., Picozzi, P., et al. (2012). MR imaging of neoplastic central nervous system lesions: review and recommendations for current practice. *Am. J. Neuroradiol.* 33, 803–817. doi: 10.3174/ajnr.A2640

Gevaert, O., Mitchell, L. A., Achrol, A. S., Xu, J., Echegaray, S., Steinberg, G. K., et al. (2014). Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features. *Radiology* 273, 168–174. doi: 10.1148/radiol.14131731

Giardino, A., Gupta, S., Olson, E., Sepulveda, K., Lenchik, L., Ivanidze, J., et al. (2017). Role of imaging in the era of precision medicine. *Acad. Radiol.* 24, 639–649. doi: 10.1016/j.acra.2016.11.021

Grier, J. T. (2006). Low-grade gliomas in adults. *Oncologist* 11, 681–693. doi: 10.1634/theoncologist.11-6-681

Grossmann, P., Gutman, D. A., Dunn, W. D., Holder, C. A., and Aerts, H. J. W. L. (2016). Imaging-genomics reveals driving pathways of MRI derived volumetric tumor phenotype features in glioblastoma. *BMC Cancer* 16:611. doi: 10.1186/s12885-016-2659-5

Herrmann, E., Ermiş, E., Meier, R., Blatti-Moreno, M., Knecht, U., Aebersold, D., et al. (2018). Fully automated segmentation of the brain resection cavity for radiation target volume definition in glioblastoma patients. *Int. J. Radiat. Oncol. Biol. Phys.* 102:S194. doi: 10.1016/j.ijrobp.2018.07.087

Hoffman, J., Mohri, M., and Zhang, N. (2018). "Algorithms and theory for multiple-source adaptation," in *Advances in Neural Information Processing Systems 31*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montreal, QC: Curran Associates, Inc.), 8246–8256.

Hsieh, K. L.-C., Tsai, R.-J., Teng, Y.-C., and Lo, C.-M. (2017). Effect of a computer-aided diagnosis system on radiologists' performance in grading gliomas with MRI. *PLoS ONE* 12:e0171342. doi: 10.1371/journal.pone.0171342

Huber, T., Alber, G., Bette, S., Kaesmacher, J., Boeckh-Behrens, T., Gempt, J., et al. (2017). Progressive disease in glioblastoma: benefits and limitations of semi-automated volumetry. *PLoS ONE* 12:e0173112. doi: 10.1371/journal.pone.0173112

Hygino da Cruz, L., Rodriguez, I., Domingues, R., Gasparetto, E., and Sorensen, A. (2011). Pseudoprogression and pseudoresponse: imaging challenges in the assessment of posttreatment glioma. *Am. J. Neuroradiol.* 32, 1978–1985. doi: 10.3174/ajnr.A2397

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). "No new-net," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 234–244.

Kickingereder, P., Isensee, F., Tursunova, I., Petersen, J., Neuberger, U., Bonekamp, D., et al. (2019). Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* 20, 728–740. doi: 10.1016/S1470-2045(19)30098-1

Kim, M. M., Parolia, A., Dunphy, M. P., and Venneti, S. (2016). Non-invasive metabolic imaging of brain tumours in the era of precision medicine. *Nat. Rev. Clin. Oncol.* 13, 725–739. doi: 10.1038/nrclinonc.2016.108

Lai, A., Eskin, A., Ellingson, B. M., Phillips, H. S., Nghiemphu, P. L., Chowdhury, R., et al. (2013). Identifying the mesenchymal molecular subtype of glioblastoma using quantitative volumetric analysis of anatomic magnetic resonance images. *Neuro Oncol.* 15, 626–634. doi: 10.1093/neuonc/not008

Lasocki, A., and Gaillard, F. (2019). Non-contrast-enhancing tumor: a new frontier in glioblastoma research. *Am. J. Neuroradiol.* 40, 758–765. doi: 10.3174/ajnr.A6025

LeCun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lipkova, J., Angelikopoulos, P., Wu, S., Alberts, E., Wiestler, B., Diehl, C., et al. (2019). Personalized radiotherapy design for glioblastoma: integrating mathematical tumor models, multimodal scans and bayesian inference. *IEEE Trans. Med. Imaging* 38, 1875–1884. doi: 10.1109/TMI.2019.2902044

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., et al. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* 9:5217. doi: 10.1038/s41467-018-07619-7

McKinley, R., Meier, R., and Wiest, R. (2019a). "Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer International Publishing), 456–465.

McKinley, R., Rebsamen, M., Meier, R., Reyes, M., Rummel, C., and Wiest, R. (2019b). Few-shot brain segmentation from weakly labeled data with deep heteroscedastic multi-task networks. *arXiv [preprint]. arXiv:1904.02436*.

Meier, R., Knecht, U., Loosli, T., Bauer, S., Slotboom, J., Wiest, R., et al. (2016). Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor volumetry. *Sci. Rep.* 6:23376. doi: 10.1038/srep23376

Meier, R., Porz, N., Knecht, U., Loosli, T., Schucht, P., Beck, J., et al. (2017). Automatic estimation of extent of resection and residual tumor volume of patients with glioblastoma. *J. Neurosurg.* 127, 798–806. doi: 10.3171/2016.9.JNS16146

Meier, R., Rebsamen, M., Knecht, U., Reyes, M., Wiest, R., and McKinley, R. (2019). "Stratify or inject: two simple training strategies to improve brain tumor segmentation," in *Presented at the International Conference on Medical Imaging with Deep Learning, MIDL 2019* (London).

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Mullins, M. E., Barest, G. D., Schaefer, P. W., Hochberg, F. H., Gonzalez, R. G., and Lev, M. H. (2005). Radiation necrosis versus glioma recurrence: conventional MR imaging clues to diagnosis. *Am. J. Neuroradiol.* 26, 1967–1972.

Myronenko, A. (2019). "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 311–320.

Ostrom, Q. T., Bauchet, L., Davis, F. G., Deltour, I., Fisher, J. L., Langer, C. E., et al. (2014). The epidemiology of glioma in adults: a state of the science review. *Neuro Oncol.* 16, 896–913. doi: 10.1093/neuonc/nou087

Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging* 35, 1240–1251. doi: 10.1109/TMI.2016.2538465

Pierallini, A., Bonamini, M., Bozzao, A., Pantano, P., Stefano, D. D., Ferone, E., et al. (1997). Supratentorial diffuse astrocytic tumours: proposal of an MRI classification. *Eur. Radiol.* 7, 395–399. doi: 10.1007/s003300050173

Pope, W. B., Sayre, J., Perlina, A., Villablanca, J. P., Mischel, P. S., and Cloughesy, T. F. (2005). MR imaging correlates of survival in patients with high-grade gliomas. *Am. J. Neuroradiol.* 26, 2466–2474.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Sharp, G., Fritscher, K. D., Pekar, V., Peroni, M., Shusharina, N., Veeraraghavan, H., et al. (2014). Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Med. Phys.* 41:050902. doi: 10.1118/1.4871620

Sottoriva, A., Spiteri, I., Piccirillo, S. G. M., Touloumis, A., Collins, V. P., Marioni, J. C., et al. (2013). Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl. Acad. Sci. U.S.A.* 110, 4009–4014. doi: 10.1073/pnas.1219747110

Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J., et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New Engl. J. Med.* 352, 987–996. doi: 10.1056/NEJMoa043330

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). "Revisiting unreasonable effectiveness of data in deep learning era," in *IEEE International Conference on Computer Vision, ICCV 2017* (Venice), 843–852.

Treiber, J. M., Steed, T. C., Brandel, M. G., Patel, K. S., Dale, A. M., Carter, B. S., et al. (2018). Molecular physiology of contrast enhancement in glioblastomas: an analysis of the cancer imaging archive (TCIA). *J. Clin. Neurosci.* 55, 86–92. doi: 10.1016/j.jocn.2018.06.018

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Weizman, L., Sira, L. B., Joskowicz, L., Rubin, D. L., Yeom, K. W., Constantini, S., et al. (2014). Semiautomatic segmentation and follow-up of multicomponent low-grade tumors in longitudinal brain MRI studies. *Med. Phys.* 41:052303. doi: 10.1118/1.4871040

Wen, P. Y., MacDonald, D. R., Reardon, D. A., Cloughesy, T. F., Sorensen, A. G., Galanis, E., et al. (2010). Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J. Clin. Oncol.* 28, 1963–1972. doi: 10.1200/JCO.2009.26.3541

Wiestler, B., Kluge, A., Lukas, M., Gempt, J., Ringel, F., Schlegel, J., et al. (2016). Multiparametric MRI-based differentiation of WHO grade II/III glioma and WHO grade IV glioblastoma. *Sci. Rep.* 6:35142. doi: 10.1038/srep35142

Zhao, H., Zhang, S., Wu, G., Moura, J. M. F., Costeira, J. P., and Gordon, G. J. (2018). "Adversarial multiple source domain adaptation," in *Advances in Neural Information Processing Systems 31*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc.), 8559–8570.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.