Combined molecular subtyping, grading, and segmentation of glioma using multi-task deep learning

van der Voort, Sebastian R.; Incekara, Fatih; Wijnenga, Maarten M.J.; Kapsas, Georgios; Schouten, Joost W.; French, Pim J.; Niessen, Wiro J.; Smits, Marion; Klein, Stefan; More Authors

DOI
10.1093/neuonc/noac166

Publication date
2023

Document Version
Final published version

Published in
Neuro-oncology

Citation (APA)
van der Voort, S. R., Incekara, F., Wijnenga, M. M. J., Kapsas, G., Schouten, J. W., French, P. J., Niessen, W. J., Smits, M., Klein, S., & More Authors (2023). Combined molecular subtyping, grading, and segmentation of glioma using multi-task deep learning. *Neuro-oncology*, *25*(2), 279-289. https://doi.org/10.1093/neuonc/noac166

Neuro-Oncology
25(2), 279–289, 2023 | https://doi.org/10.1093/neuonc/noac166 | Advance Access date 5 July 2022

Combined molecular subtyping, grading, and segmentation of glioma using multi-task deep learning

Sebastian R. van der Voort[#], Fatih Incekara[#], Maarten M. J. Wijnenga, Georgios Kapsas, Renske Gahrmann, Joost W. Schouten, Rishi Nandoe Tewarie, Geert J. Lycklama, Philip C. De Witt Hamer, Roelant S. Eijgelaar, Pim J. French, Hendrikus J. Dubbink, Arnaud J. P. E. Vincent, Wiro J. Niessen, Martin J. van den Bent, Marion Smits[†], and Stefan Klein[†]

Biomedical Imaging Group Rotterdam, Department of Radiology and Nuclear Medicine, Erasmus MC University Medical Centre Rotterdam, Rotterdam, the Netherlands (S.R.v.d.V., W.J.N., S.K.); Department of Radiology and Nuclear Medicine, Erasmus MC University Medical Centre Rotterdam, Rotterdam, the Netherlands (F.I., G.K., R.G., M.S.); Department of Neurosurgery, Brain Tumor Center, Erasmus MC University Medical Centre Rotterdam, Rotterdam, the Netherlands (F.I., J.W.S., A.J.P.E.V.); Department of Neurology, Brain Tumor Center, Erasmus MC Cancer Institute, Rotterdam, the Netherlands (M.M.J.W., P.J.F., M.J.v.d.B.); Department of Neurosurgery, Haaglanden Medical Center, the Hague, the Netherlands (R.N.T.); Department of Radiology, Haaglanden Medical Center, the Hague, the Netherlands (G.J.L.); Department of Pathology, Brain Tumor Center at Erasmus MC Cancer Institute, Rotterdam, the Netherlands (H.J.D.); Imaging Physics, Faculty of Applied Sciences, Delft University of Technology, Delft, the Netherlands (W.J.N.); Department of Neurosurgery, Cancer Center Amsterdam, Brain Tumor Center, Amsterdam UMC, Amsterdam, Netherlands (P.C.D.W.H., R.S.E.)

[#]These authors contributed equally to this work.
[†]These authors contributed equally to this work.

Corresponding Author: Stefan Klein, PhD, Biomedical Imaging Group Rotterdam, Department of Radiology and Nuclear Medicine, Erasmus MC University Medical Centre Rotterdam, Dr. Molewaterplein 50/60, 3015GE, Rottterdam, The Netherlands (s.klein@erasmusmc.nl).

Abstract

Background. Accurate characterization of glioma is crucial for clinical decision making. A delineation of the tumor is also desirable in the initial decision stages but is time-consuming. Previously, deep learning methods have been developed that can either non-invasively predict the genetic or histological features of glioma, or that can automatically delineate the tumor, but not both tasks at the same time. Here, we present our method that can predict the molecular subtype and grade, while simultaneously providing a delineation of the tumor.

Methods. We developed a single multi-task convolutional neural network that uses the full 3D, structural, preoperative MRI scans to predict the IDH mutation status, the 1p/19q co-deletion status, and the grade of a tumor, while simultaneously segmenting the tumor. We trained our method using a patient cohort containing 1508 glioma patients from 16 institutes. We tested our method on an independent dataset of 240 patients from 13 different institutes.

Results. In the independent test set, we achieved an IDH-AUC of 0.90, an 1p/19q co-deletion AUC of 0.85, and a grade AUC of 0.81 (grade II/III/IV). For the tumor delineation, we achieved a mean whole tumor Dice score of 0.84.

Conclusions. We developed a method that non-invasively predicts multiple, clinically relevant features of glioma. Evaluation in an independent dataset shows that the method achieves a high performance and that it generalizes well to the broader clinical population. This first-of-its-kind method opens the door to more generalizable, instead of hyper-specialized, AI methods.

Key Points

- Deep learning predicts glioma subtype based on MRI.
- The method is fully automated as it automatically locates and segments the tumor.
- External validation on data of 240 patients from 13 centers shows good generalization.

Importance of the Study

We developed a method that predicts multiple genetic and histological features required for the subtyping of glioma, while simultaneously providing an automatic tumor segmentation. This multi-task approach obviates the need for multiple, single-task methods, simplifying clinical implementation. Our method was trained using a glioma dataset containing 1508 patients originating from 16 different institutes, where no patients were excluded based on the scan appearance or clinical characteristics. These minimal exclusion criteria make it straightforward for clinical experts to use our method. Performance evaluation on an independent dataset showed that our method generalizes well and that our method achieves a high performance. We also provide the code and the trained model, further allowing for the quick adoption of our method.

Glioma is the most common primary brain tumor, with its most common subtype glioblastoma, in particular, being one of the deadliest forms of cancer. Differences in survival and treatment response of glioma are attributed to their genetic and histological features, specifically the isocitrate dehydrogenase (IDH) mutation status, the 1p/19q co-deletion status, and the tumor grade.[1] Therefore, in 2016, the World Health Organization (WHO) updated its brain tumor classification, categorizing glioma based on these genetic and histological features.[2] In current clinical practice, these features are determined from tumor tissue. While this is not an issue in patients in whom the tumor can be resected, this is problematic when resection can not safely be performed. In these instances, surgical biopsy is performed with the sole purpose of obtaining tissue for diagnosis, which, although relatively safe, is not without risk.[3]Therefore, there has been an increasing interest in complementary non-invasive alternatives that can provide the genetic and histological information.[4]

Magnetic resonance imaging (MRI) has been proposed as a possible candidate because of its non-invasive nature and its current place in routine clinical care. Research has shown that certain MRI features, such as a heterogeneous tumor appearance, correlate with the genetic and histological features of glioma.[5] This notion has popularized, in addition to already popular applications such as tumor segmentation, the use of machine learning methods for the prediction of genetic and histological features, known as radiomics.[6] Although a plethora of such methods now exist, they have found little translation to the clinic.[6]

An often discussed challenge for the adoption of machine learning methods in clinical practice is the lack of standardization, resulting in heterogeneity of patient populations, imaging protocols, and scan quality.[7,8] Since machine learning methods are prone to overfitting, this heterogeneity questions the validity of such methods in a broader patient population.[8] Furthermore, it has been noted that most current research concerns narrow task-specific methods that lack the context between different related tasks, which might restrict the performance of these methods.[9]

An important technical limitation when using deep learning methods is the limited GPU memory, which restricts the size of models that can be trained.[10] This is a problem especially for clinical data, which is often 3D, requiring even more memory than the commonly used 2D networks. This further limits the size of these models resulting in shallower models and the use of patches of a scan instead of using the full 3D scan as an input, which limits the amount of context these methods can extract from the scans.

Here, we present a new method that addresses the above problems. Our method consists of a single, multi-task convolutional neural network (CNN) that can predict the IDH mutation status, the 1p/19q co-deletion status, and the grade (grade II/III/IV) of a tumor, while also simultaneously segmenting the tumor (to be precise, the T2w-hyperintense tumor region). Figure 1 illustrates this approach. To the best of our knowledge, this is the first method that provides all of this information at the same time, allowing clinical experts to derive the WHO subtype from the individually predicted genetic and histological features, while also allowing them to consider or disregard specific predictions as they deem fit. Exploiting the capabilities of the latest GPUs, optimizing our implementation to reduce the memory footprint, and using distributed multi-GPU training, we were able to train a model that uses the full 3D scan as an input. We trained our method using a patient cohort containing 1508 patients included from 16 different institutes. To ensure the broad applicability of our method, we used minimal inclusion criteria, only requiring the 4 most commonly used MRI sequences: pre- and post-contrast T1-weighted (T1w), T2-weighted (T2w), andT2-weighted fluid-attenuated inversion recovery (T2w-FLAIR).[11] No constraints were placed on the patients' clinical characteristics, such as the tumor grade, or the radiological characteristics of scans, such as the scan quality. In this way, our method could capture the heterogeneity that is naturally present in clinical data. We tested our method on an independent dataset of 240 patients from 13 different institutes, to evaluate the true generalizability of our method. Our results show that we can predict multiple clinical features of glioma from MRI scans in a diverse patient population.

Materials and Methods

Patient Population

Our dataset consisted of retrospectively collected data that formed a convenience series. All included patients were newly diagnosed, preoperative patients.

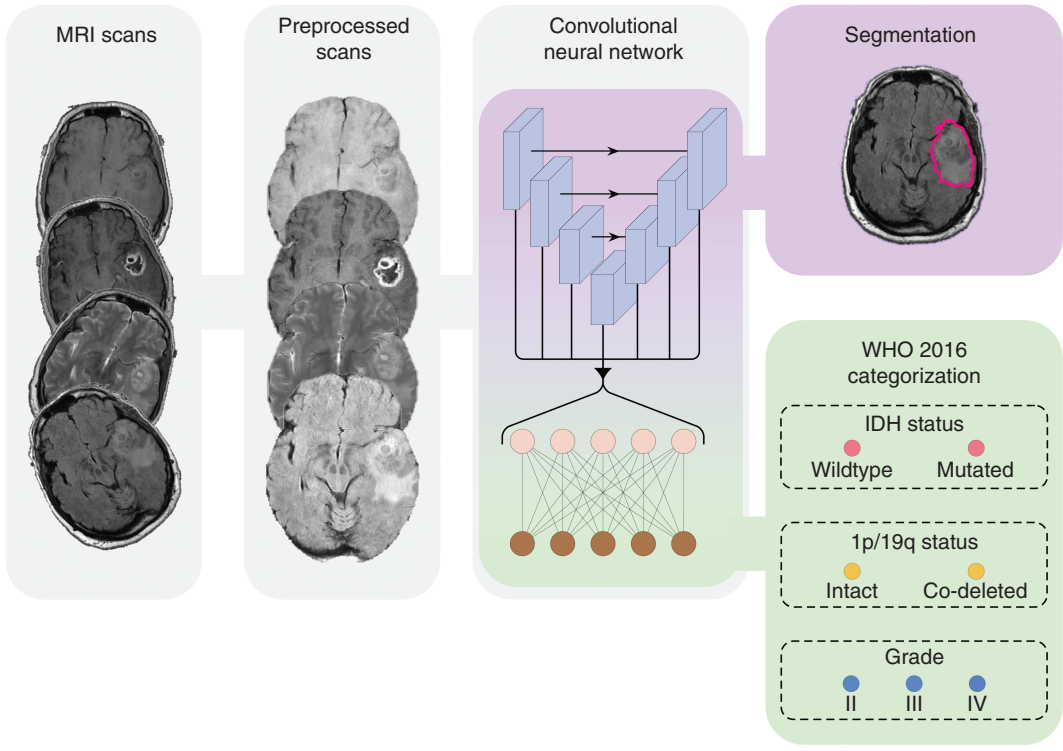

Fig. 1 Overview of our method. Pre- and post-contrast T1w, T2w, and T2w-FLAIR scans are used as an input. The scans are registered to an atlas, bias field corrected, skull stripped, and normalized before being passed through our convolutional neural network. One branch of the network segments the tumor, while at the same time the features are combined to predict the IDH status, 1p/19q status, and grade of the tumor.

The train set was collected from 4 in-house datasets and 5 publicly available datasets. In-house datasets were collected from 4 different institutes: Erasmus MC (EMC), Haaglanden Medical Center (HMC), Amsterdam UMC (AUMC),[12] and University Medical Center Utrecht (UMCU). Four of the five public datasets were collected from The Cancer Imaging Archive (TCIA)[13]: the Repository of Molecular Brain Neoplasia Data (REMBRANDT) collection,[14] the Clinical Proteomic Tumor Analysis Consortium Glioblastoma Multiforme (CPTAC-GBM) collection,[15] the Ivy Glioblastoma Atlas Project (Ivy GAP) collection,[16,17] and the Brain-Tumor-Progression collection.[18] The fifth dataset was the 2019 Brain Tumor Segmentation (BraTS) challenge dataset,[19–21] from which we excluded the patients that were also available in the TCGA-LGG and TCGA-GBM collections.[22,23]

For the internal datasets from the EMC and the HMC, manual segmentations were available, which were made by 4 different clinical experts, who received guidance from a senior neuroradiologist (>10 years of experience). For patients where segmentations from more than 1 observer were available, we randomly selected one of the segmentations to use in the train set. The segmentations from the AUMC data were made by a single observer (a senior neuroradiologist with >10 years of experience) of the study by Visser et al.[12] From the public datasets, only the BraTS dataset and the Brain-Tumor-Progression dataset provided manual segmentations. Segmentations of the BraTS dataset, as provided in the 2019 training and validation set were used. For the Brain Tumor-Progression dataset, the segmentations as provided in the TCIA data collection were used. This resulted in a varied dataset of segmentations, with segmentations from a variety of qualified different observers.

Patients were included if they were newly diagnosed with a glioma and when preoperative pre- and post-contrast T1w, T2w, and T2w-FLAIR scans were available; no further inclusion criteria were set. For example, patients were not excluded based on the radiological characteristics of the scan, such as low imaging quality or imaging artifacts, or the glioma's clinical characteristics such as the grade. If multiple scans of the same contrast type were available in a single scan session (eg, multiple T2w scans), the scan upon which the segmentation was made was selected. If no segmentation was available, or the segmentation was not made based on that scan contrast, the scan with the highest axial resolution was used, where a 3D acquisition was preferred over a 2D acquisition.

For the in-house data, genetic and histological data were available for the EMC, HMC, and UMCU datasets, which were obtained from the analysis of tumor tissue after biopsy or resection. Genetic and histological data of the public datasets were also available for the REMBRANDT, CPTAC-GBM, and Ivy GAP collections. Data for the REMBRANDT and CPTAC-GBM collections were collected from the clinical data available at the TCIA.[14,15] For the Ivy

GAP collection, the genetic and histological data were obtained from the Swedish Institute at https://ivygap.org/.

As a test set, we used the TCGA-LGG and TCGA-GBM collections from the TCIA.[22,23] Genetic and histological labels were obtained from the clinical data available at the TCIA, and from data published by Ceccarelli et al.[24] Segmentations were used as available from the TCIA, based on the 2018 BraTS challenge.[20,25,26] Tumors were manually segmented by 1-4 raters, following the same annotation protocol, and their ground truth annotations were approved by experienced neuroradiologists.[21] The inclusion criteria for the patients included in the BraTS challenge were the same as our inclusion criteria: the presence of a preoperative pre- and post-contrast T1w, T2w, and T2w FLAIR scan. Thus, patients from the TCGA-LGG and TCGA-GBM were included if a segmentation from the BraTS challenge was available. However, for 3 patients, we found that although they did have manual segmentations, they did not meet our inclusion requirements: TCGA-08-0509 and TCGA-08-0510 from TCGA-GBM because they did not have a pre-contrast T1w scan and TCGA-FG-7634 from TCGA-LGG because there was no post-contrast T1w scan.

Classification Model

We used a CNN to predict the genetic and histological features and to automatically segment the T2w-hyperintense region. This network was developed to facilitate the auxiliary tasks, and optimized in terms of memory efficiency to allow for the full 3D scan to be used as an input. The MRI scans were pre-processed (which included registration to an atlas, skull-stripping, and intensity normalization), and the pre-processed scans from the train set were used to develop and train the model. We used 15% of the train set (226 patients) as a validation set and selected the model parameters that achieved the best performance on this validation set as the optimal model parameters. Subsequently, the model was trained using these optimal model parameters and the full train set.

The performance of the model was then evaluated on the independent test set. This independent test set was not used during the development of the model, and the performance of our model in the test set was only evaluated one time, once the optimal model parameters were established using the train set.

We evaluated the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity for the genetic and histological features. For the AUC, we also evaluated the 95% confidence interval (CI) intervals using basic bootstrapping with 2500 iterations. To evaluate the segmentation performance we evaluated the Dice score, Hausdorff distance, and the volumetric similarity coefficient. The Dice score is a measure of the overlap between the ground truth and predicted segmentation (where 1 indicates perfect overlap). The Hausdorff distance is a measure of the agreement between the boundaries of the ground truth and predicted segmentation (lower is better). The volumetric similarity coefficient is a measure of the agreement in volume (where 1 indicates perfect agreement).

All the details of the pre-processing of the scans, the development and structure of our model, the evaluation of the model performance, and the post-processing of the results are given in Supplementary Appendix A.

Ethics Statement

The study was approved by the Medical Ethical Committee of Erasmus MC (METC-2016-419), who waived the need for written informed consent from the patients due to the retrospective nature of this study and the (emotional) burden that would result from contacting the patients or their relatives to obtain consent. The study was performed in accordance with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

STARD Statement

We have followed the "Standards for Reporting Diagnostic Accuracy Studies" (STARD) guidelines for the reporting of this study.[27] The filled-out checklist is presented in Supplementary Appendix B.

Results

Patient Characteristics

We included a total of 1748 patients in our study, 1508 as a train set and 240 as an independent test set. The patients in the train set originated from 9 different data collections and 16 different institutes, and the test set was collected from 2 different data collections and 13 different institutes. Table 1 provides a full overview of the patient characteristics in the train and test set, and Figure 2 shows the inclusion flowchart and the distribution of the patients over the different data collections in the train set and test set. Supplementary Appendix C provides an overview of the MRI acquisition parameters of the scans, which shows that scans were acquired from scanners of different vendors, using a wide variety of acquisition settings.

Algorithm Performance

In the test set, we achieved an AUC of 0.90 for the IDH mutation status prediction, an AUC of 0.85 for the 1p/19q co-deletion prediction, and an AUC of 0.81 for the grade prediction. The full results are shown in Table 2, with the corresponding receiver operating characteristic (ROC) curves in Figure 3. Table 2 also shows the results in (clinically relevant) subgroups of patients. This shows that we achieved an IDH-AUC of 0.81 in low-grade glioma (LGG) (grade II/III), an IDH-AUC of 0.64 in high-grade glioma (HGG) (grade IV), and a 1p/19q-AUC of 0.73 in LGG. When only predicting LGG vs HGG instead of predicting the individual grades, we achieved an AUC of 0.91. In Supplementary Appendix D, we provide confusion matrices for the IDH mutation status, 1p/19q co-deletion status, and grade predictions, as

Table 1. Patient Characteristics for the Train and Test Sets

	Train Set		Test Set	
	N	%	N	%
Total	1508	100.0	240	100.0
Sex				
Female	465	30.8	108	45.0
Male	786	52.1	130	54.2
Unknown	257	17.1	2	0.8
Age (years)				
<40	208	13.8	50	20.8
40-60	552	36.6	113	47.1
>60	608	40.3	75	31.3
Unknown	140	9.3	2	0.8
IDH status				
Mutated	226	15.0	88	36.7
Wildtype	440	29.2	129	53.7
Unknown	842	55.8	23	9.6
1p/19q co-deletion status				
Co-deleted	103	6.8	26	10.8
Intact	337	22.4	207	86.3
Unknown	1068	70.8	7	2.9
Grade				
II	230	15.3	47	19.6
III	114	7.6	59	24.6
IV	830	55.0	132	55.0
Unknown	334	22.1	2	0.8
WHO 2016 subtype				
Oligodendroglioma	96	6.4	26	10.8
Astrocytoma, IDH-mutated	98	6.5	57	23.7
Astrocytoma, IDH-wildtype	31	2.0	22	9.2
GBM, IDH-mutated	16	1.1	5	2.1
GBM, IDH-wildtype	331	21.9	106	44.2
Unknown	936	62.1	24	10.0
Segmentation				
Manual	716	47.5	240	100
Automatic	792	52.5	0	0

Abbreviations: GBM, glioblastoma; IDH, isocitrate dehydrogenase; WHO, World Health Organization.

well as a confusion matrix for the final WHO 2016 subtype, which shows that only one patient was predicted as a non-existing WHO 2016 subtype. In Supplementary Appendix E, we provide the individual predictions and ground truth labels for all patients in the test set to allow for the calculation of additional metrics.

For the automatic segmentation, we achieved a mean Dice score of 0.84, a mean Hausdorff distance of 18.9 mm, and a mean volumetric similarity coefficient of 0.90. Figure 4 shows boxplots of the Dice scores, Hausdorff distances, and volumetric similarity coefficients for the different patients in the test set. In Supplementary Appendix F, we show 5 patients that were randomly selected from both the TCGA-LGG and TCGA-GBM data collections, to demonstrate the automatic segmentations made by our method.

To evaluate the face validity and clinical benefit of our algorithm we compared the survival of patients based on the ground truth and predicted WHO 2016 subtypes, shown in Supplementary Appendix G. This comparison shows that there is no significant difference in the survival of patients between the true WHO 2016 subtype and the predicted subtype.

To assess the future-proofness of our method, we evaluated the performance of our method in predicting the

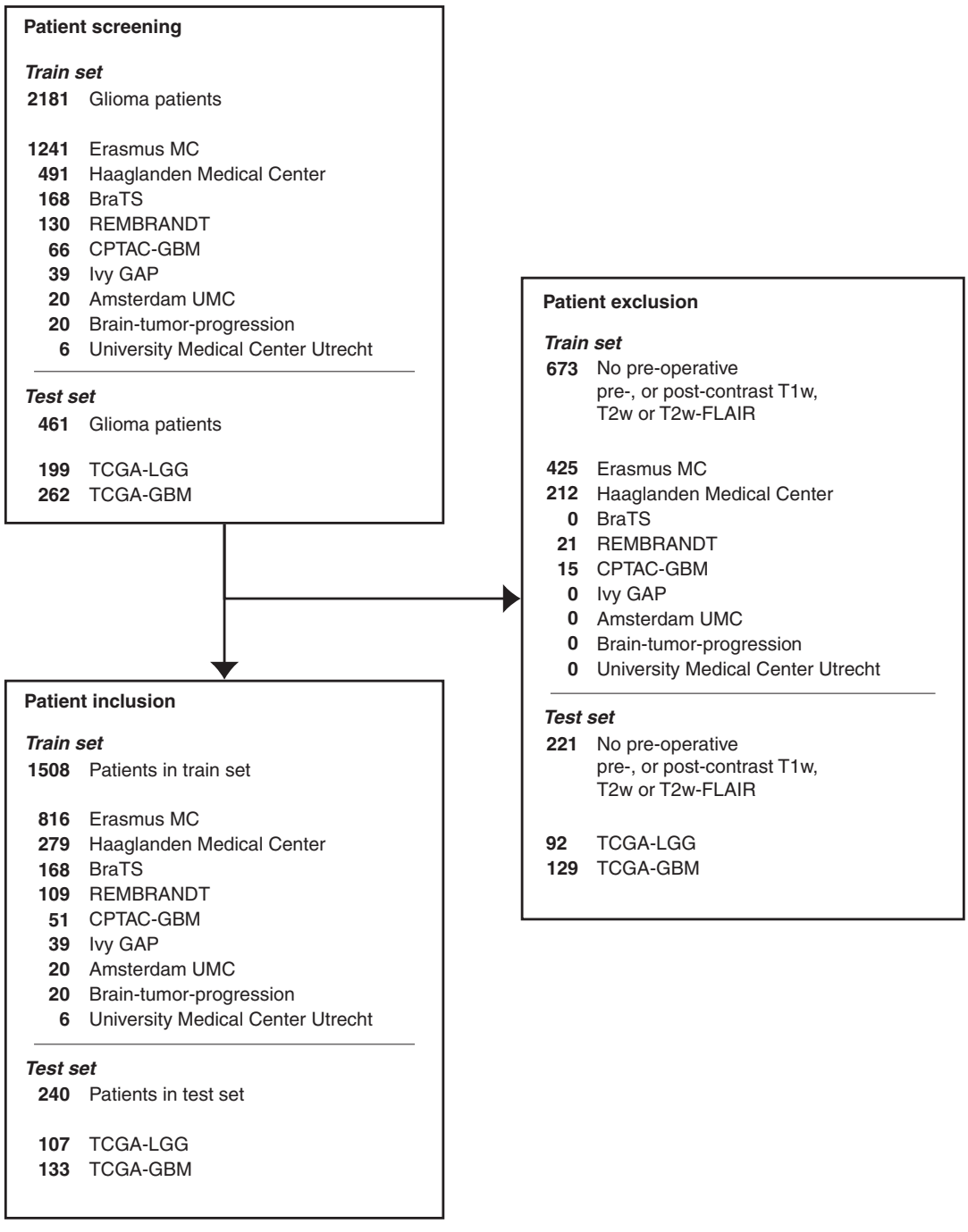

Fig. 2 Inclusion flowchart of the train set and test set.

subtypes according to the new WHO 2021 guidelines, see Supplementary Appendix H. These results show that the performance for the WHO 2021 subtypes is in fact better than for the WHO 2016 subtypes.

We also evaluated the effect of training the model on different subsets of the dataset to investigate the effect of adding more data, see Supplementary Appendix I. These results show that for almost all cases the best results are obtained using the full train set to construct the model.

Model Interpretability

To provide insight into the behavior of our model we created saliency maps and visualized selected filter outputs of

Table 2. Evaluation Results of the Final Model on the Test Set

Prediction Task	AUC (95% CI)	Accuracy	Sensitivity	Specificity
All patients				
IDH	0.90 (0.85-0.95)	0.84	0.72	0.93
1p/19q	0.85 (0.77-0.92)	0.89	0.39	0.95
Grade (II/III/IV)	0.81 (0.76-0.85)	0.71	n/a	n/a
Grade II	0.91 (0.87-0.95)	0.86	0.75	0.89
Grade III	0.69 (0.62-0.77)	0.75	0.17	0.94
Grade IV	0.91 (0.87-0.94)	0.82	0.95	0.66
LGG vs HGG	0.91 (0.87-0.94)	0.84	0.72	0.93
LGG patients				
IDH	0.81 (0.72-0.89)	0.74	0.73	0.77
1p/19q	0.73 (0.60-0.83)	0.76	0.39	0.89
HGG patients				
IDH	0.64 (0.21-0.99)	0.94	0.40	0.96

Abbreviations: AUC, area under the receiver operating characteristic curve; CI, confidence interval; HGG, high-grade glioma; IDH, isocitrate dehydrogenase; LGG, low-grade glioma.

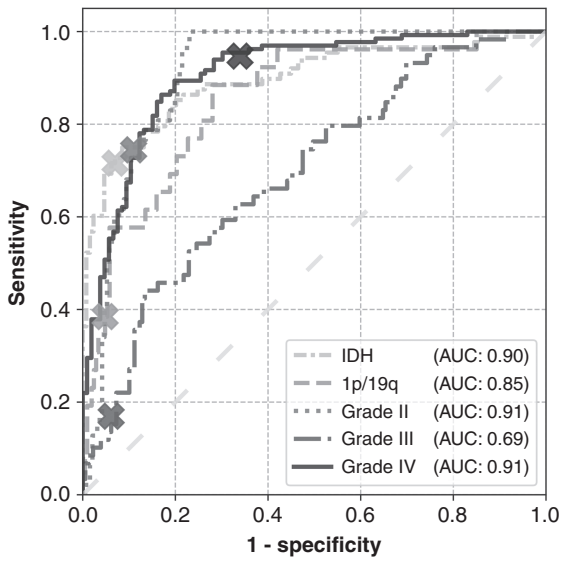

Fig. 3 Receiver operating characteristic (ROC) curves of the genetic and histological features are evaluated on the test set. The crosses indicate the location of the decision threshold for the reported accuracy, sensitivity, and specificity.

our networks in Supplementary Appendix J. These maps and visualizations show which parts of the scans contributed the most to the prediction. The saliency maps show that for a LGG patient the network focused on a bright rim in the T2w-FLAIR scan, whereas for the HGG patient the network focused on the enhancement in the post-contrast T1w scan. The filter output visualizations also show that the network focuses on the tumor, and these filters seem to recognize specific imaging features such as the contrast enhancement in the post-contrast T1-weighted scan and T2w-FLAIR brightness.

Model Robustness

By not excluding scans from our train set based on radiological characteristics, we were able to make our model robust to low scan quality, as can be seen in an example from the test set in Supplementary Figure F.4. Even though this example scan contained imaging artifacts, our method was able to properly segment the tumor (DICE score of 0.87), and correctly predicted the tumor as an IDH-wildtype, grade IV tumor.

We considered 2 examples of scans that were incorrectly predicted by our method, see Supplementary Appendix K. These 2 examples were chosen because our network assigned high prediction scores to the wrong classes for these cases. Supplementary Figure K.11a shows an example of an IDH-mutated, 1p/19q co-deleted, grade II glioma that was predicted as IDH-wildtype, 1p/19q intact, grade IV by our method. Our method's prediction was most likely caused by the hyperintensities in the post-contrast T1w scan being interpreted as contrast enhancement. Since these hyperintensities are also present in the pre-contrast T1w scan they are most likely calcifications, and the radiological appearance of this tumor is indicative of an oligodendroglioma. Supplementary Figure K.11b shows an example of an IDH-wildtype, 1p/19q intact, grade IV glioma that was predicted as an IDH-mutated, 1p/19q intact, grade III glioma by our method.

Finally, we evaluated the model performance when omitting 1 of the 4 MRI sequences during training and testing of the model to assess whether all sequences are (equally) necessary. These results are shown in Supplementary Appendix L and support the fact that our model trained using all 4 MRI sequences achieves the best overall performance.

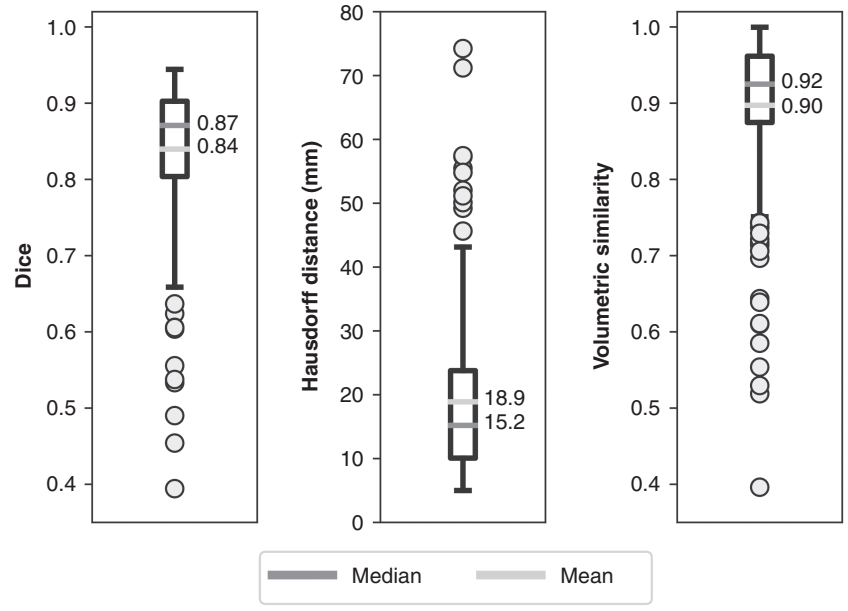

Fig. 4 Dice scores, Hausdorff distances, and volumetric similarity coefficients for all patients in the test set.

Discussion

We have developed a method that can predict the IDH mutation status, 1p/19q co-deletion status, and grade of glioma, while simultaneously providing the tumor (T2w-hyperintense region) segmentation, based on preoperative MRI scans of newly diagnosed patients.

Our method can be beneficial for patients for whom resection cannot be safely performed, and in whom biopsy is performed solely for obtaining tissue for diagnosis. In these instances, our method can provide a non-invasive alternative. Furthermore, our method can act as an additional check of the histopathological results in case of uncertainty such as (suspected) sampling bias.

In an independent test set, which contained data from 13 different institutes, we demonstrated that our method predicts the genetic and histological features with good overall performance; we achieved an AUC of 0.90 for the IDH mutation status prediction, an AUC of 0.85 for the 1p/19q co-deletion prediction, an AUC of 0.81 for the grade prediction, and a mean whole tumor Dice score of 0.84. This performance on unseen data that was only used during the final evaluation of the algorithm, and that was purposefully not used to guide any decisions regarding the method design, shows the true generalizability of our method. Using the latest GPU capabilities we were able to train a large model, which uses the full 3D scan as input. Furthermore, by using a large cohort of patients originating from multiple institutes, we were able to make our method robust to the heterogeneity that is naturally present in clinical imaging data, such that it generalizes for broad application in clinical practice. To simplify the use of our method, we provide a Docker container that contains all the required software tools and data, and which performs the pre-processing steps, inference of the model, and the post-processing steps. In this way, our model can be applied to new data with minimal setup.

By using a multi-task network, our method could learn the context between different features. For example, IDH-wildtype and 1p/19q co-deletion are mutually exclusive.[28] If two separate methods had been used, one to predict the IDH status and one to predict the 1p/19q co-deletion status, an IDH-wildtype glioma might be predicted to be 1p/19q co-deleted, which does not stroke with the clinical reality. Since our method learns both of these genetic features simultaneously, it correctly learned not to predict 1p/19q co-deletion in tumors that were IDH-wildtype; there was only one patient for whom our algorithm predicted a tumor to be both IDH-wildtype and 1p/19q co-deleted. Furthermore, by predicting the genetic and histological features individually, instead of only predicting the WHO 2016 subtype, it is possible to adopt updated guidelines such as the new WHO 2021 guidelines.[29]

Some previous studies also used multi-task networks to predict the genetic and histological features of glioma.[30–32] Tang et al[30] used a multi-task network that predicts multiple genetic features, as well as the overall survival of glioblastoma. Since their method only works for glioblastoma patients, the tumor grade must be known in advance, complicating the use of their method in the preoperative setting when tumor grade is not yet known. Furthermore, their method requires a tumor segmentation prior to the application of their method, which is a time-consuming, expert task. In a study by Xue et al,[31] a multi-task network was

used, with a structure similar to the one proposed in this paper, to segment the tumor and predict the grade (LGG or HGG) and IDH mutation status. However, they do not predict the 1p/19q co-deletion status needed for the WHO 2016 subtyping. Lastly, Decuyper et al[32] used a multi-task network that predicts the IDH mutation and 1p/19q co-deletion status, and the tumor grade (LGG or HGG). Their method requires a tumor segmentation as input, which they obtain from a U-Net that is applied earlier in their pipeline; thus, their method requires two networks instead of the single network we use in our method. These differences aside, the most important limitation of each of these studies is the lack of an independent test set for evaluating their results. It is now considered essential that an independent test set is used, to prevent an overly optimistic estimate of a method's performance.[7] Thus, our study improves on this previous work by providing a single network that combines the different tasks, being trained on a more extensive and diverse dataset, not requiring a tumor segmentation as an input, providing all information needed for the WHO 2016 subtyping, and, crucially, by being evaluated in an independent test set.

An important genetic feature that is not predicted by our method is the O^6-methylguanine-methyltransferase (MGMT) methylation status. Although the MGMT methylation status is not part of the WHO 2016 subtyping, it is part of clinical management guidelines and is an important prognostic marker in glioblastoma.[2] In the initial stages of this study, we attempted to predict the MGMT methylation status; however, the performance of this prediction was poor. Furthermore, the methylation cutoff level, which is used to determine whether a tumor is MGMT methylated, shows a wide variety between institutes, leading to inconsistent results.[33] We therefore opted not to include the MGMT prediction at all, rather than to provide a poor prediction of an unsharply defined parameter. Although some methods attempted to predict the MGMT status, with varying degrees of success, there is still an ongoing discussion on the validity of MRI features of the MGMT status.[30,34]

Our method does not provide an uncertainty estimate for its predictions. Providing uncertainty estimates for CNN predictions is currently an important research topic, with new methods being developed which could provide an uncertainty estimate.[35] Unfortunately, there are still some challenges to overcome before these methods can be widely applied, and it is unclear how accurate these uncertainty estimates are.

In this study, we have only considered scans containing glioma pathology. However, this still requires a clinical expert to (correctly) identify the pathology in the scan before the scan can be analyzed by our method. Although automatic methods exist that can differentiate between different pathologies,[36] in clinical practice patients who come in are usually already suspected of having a glioma, and these lesions generally do not present a diagnostic dilemma for experienced neuroradiologists.

Our method shows good overall performance, but there are noticeable performance differences between tumor categories. For example, when our method predicts a tumor as an IDH-wildtype glioblastoma, it is correct almost all of the time. On the other hand, it has some difficulty differentiating IDH-mutated, 1p/19q co-deleted LGG from other LGG. The sensitivity for the prediction of grade III glioma was low, which might be caused by the lack of a central pathology review. Because of this, there were differences in molecular testing and histological analysis, and it is known that distinguishing between grade II and grade III has a poor observer reliability.[37] This difference in the predictive performance between different subgroups can also be explained by the imbalance in data. For example, there were much fewer 1p/19q co-deleted LGG than IDH-wildtype HGG. Although we compensated for this imbalance during the training of the model, this cannot compete with the availability of actual data and the resulting increase of data diversity in certain subgroups. Thus, even though our method can be relevant for certain subgroups, our method's performance still needs to be improved to ensure relevancy for the full patient population.

In future work, we aim to increase the performance of our method by including perfusion-weighted imaging (PWI) and diffusion-weighted imaging (DWI) since there has been an increasing amount of evidence that these physiological imaging modalities contain additional information that correlates with the tumor's genetic features and aggressiveness.[38] They were not included in this study since PWI and, to a lesser extent, DWI is not as ingrained in the clinical imaging routine as the structural scans used in this work.[11] Thus, including these modalities would limit our method's clinical applicability and substantially reduce the number of patients in the train and test set. However, PWI and DWI are increasingly becoming more commonplace, which will allow including these in future research and which might improve performance.

To further increase the performance of our model, we have made the trained model publicly available, and invite other researchers to continue training the model with their data. In this way, the model will be presented with an even larger variety of data, which can further improve the predictive performance.

In conclusion, we have developed a non-invasive method that can predict the IDH mutation status, 1p/19q co-deletion status, and grade of glioma, while at the same time automatically segmenting the tumor, based on preoperative MRI scans with high overall performance. Although the performance of our method might need to be improved before it will find widespread clinical acceptance, we believe that this research is an important step forward in the field of radiomics. Predicting multiple clinical features simultaneously steps away from the conventional single-task methods and is more in line with the clinical practice where multiple clinical features are considered simultaneously and may even be related. Furthermore, by not limiting the patient population used to develop our method to a selection based on clinical or radiological characteristics, we alleviate the need for a priori (expert) knowledge, which may not always be available. Although steps still have to be taken before radiomics will find its way into the clinic, especially in terms of performance, our work provides a crucial step forward by resolving some of the hurdles of clinical implementation now and paving the way for a full transition in the future.

Supplementary Material

Supplementary material is available at *Neuro-Oncology* online.

Keywords

deep learning | glioma | multi-task | radiomics | segmentation

Funding

S.R.v.d.V and F.I. acknowledge funding by the Dutch Cancer Society (KWF project number EMCR 2015-7859). This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 952103.

Unpublished material. Bakas S, Reyes M, Jakab A, et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. Unpublished 2019.

Acknowledgments

Data used in this publication were generated by the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The results published here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Conflict of interest statement. W.J.N. is the founder, scientific lead, and shareholder of Quantib BV. H.J.D. has the following interest that is not related to the current work: grants, personal fees, and non-financial support from AstraZeneca, personal fees from AbbVie, personal fees from Janssen, personal fees from Pfizer, personal fees from PGDx, personal fees from MSD, personal fees from Lilly. M.S. received honoraria for independent trial review from Parexel Ltd (paid to the institution, no direct relation with the presented work) and speaker fees from GE Healthcare (paid to the institution, no direct relation to the presented work). All other authors declare no conflict of interest.

Authorship statement. Contrive the study and design the experiments: S.R.v.d.V., F.I., W.J.N., M.S., and S.K. Include patients in the different studies.: F.I., M.M.J.W., J.W.S., R.N.T., G.J.L., P.C.D.W.H., R.S.E., A.J.P.E.V., M.J.v.d.B., and M.S. Collect the data: S.R.v.d.V., F.I., M.M.J.W., G.K., R.G., J.W.S., R.N.T., G.J.L., P.C.D.W.H., R.S.E., P.J.F., H.J.D., A.J.P.E.V., M.J.v.d.B., and M.S. Perform experiments: S.R.v.d.V. Interpret results and verify data: S.R.v.d.V., F.I., M.S., and S.K. Create initial draft of the paper: S.R.v.d.V., F.I., M.J.v.d.B., M.S., and S.K. Revise the paper: M.M.J.W., G.K., R.G., J.W.S., R.N.T., G.J.L., P.C.D.W.H., R.S.E., P.J.F., H.J.D., A.J.P.E.V., W.J.N., and M.J.v.d.B.

Data Availability

An overview of the patients included from the public datasets used in the training and testing of the algorithm, and their ground truth label is available in Supplementary Appendix M. The data from the public datasets are available in TCIA under the following DOIs:

- REMBRANDT: 10.7937/K9/TCIA.2015.588OZUZB
- CPTAC-GBM: 10.7937/k9/tcia.2018.3rje41q1
- Ivy GAP: 10.7937/K9/TCIA.2016.XLwaN6nL
- Brain-Tumor-Progression: 10.7937/K9/TCIA.2018.15quzvnb
- TCGA-LGG: 10.7937/K9/TCIA.2016.L4LTD3TK
- TCGA-GBM: 10.7937/K9/TCIA.2016.RNYFUYE9

The segmentations for the public datasets from the TCIA are available under the following DOIs:

- Brain-Tumor-Progression: 10.7937/10.7937/K9/TCIA.2018.15quzvnb
- TCGA-LGG: 10.7937/K9/TCIA.2017.GJQ7R0EF
- TCGA-GBM: 10.7937/K9/TCIA.2017.KLXWJJ1Q

Data from the BraTS are available at http://braintumorsegmentation.org/. Data of 774 patients from the Erasmus MC dataset are available at https://xnat.bmia.nl/REST/projects/egd and are described in a separate data publication.[39]

The prediction results and data used to evaluate the model performance are available as a supplementary file to this article. The trained model is available at https://github.com/Svdvoort/PrognosAIs_glioma. The code used in this paper is available on GitHub under an Apache 2 license at https://github.com/Svdvoort/PrognosAIs_glioma. This code includes the full pipeline from the registration of the patients to the final post-processing of the predictions.

References

1. Dubbink HJ, Atmodimedjo PN, Kros JM, et al. Molecular classification of anaplastic oligodendroglioma using next-generation sequencing: a report of the prospective randomized EORTC Brain Tumor Group 26951 phase III trial. *Neuro Oncol.* 2015;18:388–400.
2. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 2016;131:803–820.
3. Chen C-C, Hsu P-W, Wu T-WE, et al. Stereotactic brain biopsy: single center retrospective analysis of complications. *Clin Neurol Neurosurg.* 2009;111:835–839.
4. Zhou M, Scott J, Chaudhury B, et al. Radiomics in brain tumor: image assessment, quantitative feature descriptors, and machine-learning approaches. *AJNR Am J Neuroradiol.* 2018;39:208–216.
5. Smits M. Imaging of oligodendroglioma. *Br J Radiol.* 2016;89:20150857.
6. Gore S, Chougule T, Jagtap J, Saini J, Ingalhalikar M. A review of radiomics and deep predictive modeling in glioma characterization. *Acad Radiol.* 2020;28(11):1599–1621.
7. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology.* 2016;278:563–577.

8. Thrall JH, Li X, Li Q, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *J Am Coll Radiol*. 2018;15:504–508.
9. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18:500–510.
10. Kopuklu O, Kose N, Gunduz A, Rigoll G. Resource efficient 3D convolutional neural networks. Paper presented at: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW); Seoul, South Korea; October 27–28, 2019:1910–1919.
11. Thust SC, Heiland S, Falini A, et al. Glioma imaging in Europe: a survey of 220 centres and recommendations for best clinical practice. *Eur Radiol*. 2018;28:3306–3317.
12. Visser M, Müller DMJ, van Duijn RJM, et al. Inter-rater agreement in glioma segmentations on longitudinal MRI. *Neuroimage Clin*. 2019;22:101727.
13. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013;26:1045–1057.
14. Scarpace L, Flanders AE, Jain R, Mikkelsen T, Andrews DW. *Data from REMBRANDT. Version 1*. The Cancer Imaging Archive. 2015. http://doi.org/10.7937/K9/TCIA.2015.588OZUZB. Accessed December 1, 2016.
15. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). *Radiology Data from the Clinical Proteomic Tumor Analysis Consortium Glioblastoma Multiforme CPTAC-GBM collection. Version 6*. The Cancer Imaging Archive. 2019 [Data Set]. https://doi.org/10.7937/k9/tcia.2018.3rje41q1
16. Shah N, Feng X, Lankerovich M, Puchalski RB, Keogh B. *Data from Ivy GAP. Version 1*. The Cancer Imaging Archive. 2016. https://doi.org/10.7937/K9/TCIA.2016.XLwaN6nL
17. Puchalski RB, Shah N, Miller J, et al. An anatomic transcriptional atlas of human glioblastoma. *Science*. 2018;360:660–663.
18. Schmainda K, Prah M. *Data from Brain-Tumor-Progression. Version 1*. 2018. https://doi.org/10.7937/K9/TCIA.2018.15quzvnb
19. Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34:1993–2024.
20. 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.
21. Bakas S, Reyes M, Jakab A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv*, doi:10.48550/arXiv.1811.02629, April 23, 2019, preprint: not peer reviewed.
22. Pedano N, Flanders AE, Scarpace L, et al. *Radiology data from The Cancer Genome Atlas Low Grade Glioma [TCGA-LGG] collection*. Version 2. 2016. https://doi.org/10.7937/K9/TCIA.2016.L4LTD3TK
23. Scarpace L, Mikkelsen T, Cha S, et al. *Radiology data from The Cancer Genome Atlas glioblastoma multiforme [TCGA-GBM] collection*. Version 3. 2015. https://doi.org/10.7937/K9/TCIA.2016.RNYFUYE9
24. Ceccarelli M, Barthel FP, Malta TM, et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*. 2016;164:550–563.
25. Bakas S, Akbari H, Sotiras A, et al. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection [Data Set]. The Cancer Imaging Archive. Version 1. 2017. https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF
26. Bakas S, Akbari H, Sotiras A, et al. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection [Data Set]. The Cancer Imaging Archive. Version 1. 2017. https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
27. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology*. 2015;277:826–832.
28. Labussière M, Idbaih A, Wang X, et al. All the 1p19q codeleted gliomas are mutated on IDH1 or IDH2. *Neurology*. 2010;74:1886–1890.
29. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro Oncol*. 2021;23:1231–1251.
30. Tang Z, Xu Y, Jiao Z, et al. Pre-operative overall survival time prediction for glioblastoma patients using deep learning on both imaging phenotype and genotype. In: Shen D, Liu T, Peters TM, et al., eds. *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*. Cham, Switzerland: Springer; 2019:415–422.
31. Xue Z, Xin B, Wang D, Wang X. Radiomics-enhanced multi-task neural network for non-invasive glioma subtyping and segmentation. In: Mohyud-Din H, Rathore S, eds. *RNO-AI 2019: Radiomics and Radiogenomics in Neuro-oncology*. Cham, Switzerland: Springer; 2020:81–90.
32. Decuyper M, Bonte S, Deblaere K, Van Holen R. Automated MRI based pipeline for segmentation and prediction of grade, IDH mutation and 1p19q co-deletion in glioma. *Comput Med Imaging Graph*. 2021;88:101831.
33. Malmström A, Lysiak M, Kristensen BW, Hovey E, Henriksson R, Soderkvist P. Do we really know who has an MGMT methylated glioma? Results of an international survey regarding use of MGMT analyses for glioma. *Neurooncol Pract*. 2020;7:68–76.
34. Mikkelsen VE, Dai HY, Stensjøen AL, et al. MGMT promoter methylation status is not related to histological or radiological features in IDH wild-type glioblastomas. *J Neuropathol Exp Neurol*. 2020;79:855–862.
35. Jungo A, Balsiger F, Reyes M. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Front Neurosci*. 2020;14:282.
36. Chen C, Ou X, Wang J, Guo W, Ma X. Radiomics-based machine learning in differentiation between glioblastoma and metastatic brain tumors. *Front Oncol*. 2019;9:806.
37. van den Bent MJ. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta Neuropathol*. 2010;120:297–304.
38. Park JE, Kim HS, Jo Y, et al. Radiomics prognostication model in glioblastoma using diffusion- and perfusion-weighted MRI. *Sci Rep*. 2020;10:4250.
39. van der Voort SR, Incekara F, Wijnenga MMJ, et al. The Erasmus Glioma Database (EGD): structural MRI scans, WHO 2016 subtypes, and segmentations of 774 patients with glioma. *Data Brief*. 2021;37:107191.