Neuro-Oncology Advances

6(1), vdae192, 2024 | https://doi.org/10.1093/noajnl/vdae192 | Advance Access date 10 November 2024

Clinical application of machine-based deep learning in patients with radiologically presumed adult-type diffuse glioma grades 2 or 3

Tomás Gómez Vecchio, Alice Neimantaite, Erik Thurin, Julia Furtner, Ole Solheim, Johan Pallud, Mitchel Berger, Georg Widhalm, Jiri Bartek, Ida Häggström, Irene Y.H. Gu, and Asgeir Store Jakola

All author affiliations are listed at the end of the article

Corresponding Author: Asgeir Store Jakola, MD, PhD, Department of Clinical Neuroscience, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, 41345 Gothenburg, Sweden (jakola.asgeir@gu.se).

Abstract

Background. Radiologically presumed diffuse lower-grade glioma (dLGG) are typically non or minimal enhancing tumors, with hyperintensity in T2w-images. The aim of this study was to test the clinical usefulness of deep learning (DL) in *IDH* mutation prediction in patients with radiologically presumed dLGG.

Methods. Three hundred and fourteen patients were retrospectively recruited from 6 neurosurgical departments in Sweden, Norway, France, Austria, and the United States. Collected data included patients' age, sex, tumor molecular characteristics (*IDH*, and 1p19q), and routine preoperative radiological images. A clinical model was built using multivariable logistic regression with the variables age and tumor location. DL models were built using MRI data only, and 4 DL architectures used in glioma research. In the final validation test, the clinical model and the best DL model were scored on an external validation cohort with 155 patients from the Erasmus Glioma Dataset.

Results. The mean age in the recruited and external cohorts was 45.0 (SD 14.3) and 44.3 years (SD 14.6). The cohorts were rather similar, except for sex distribution (53.5% vs 64.5% males, *P*-value = .03) and *IDH* status (30.9% vs 12.9% *IDH* wild-type, *P*-value <.01). Overall, the area under the curve for the prediction of *IDH* mutations in the external validation cohort was 0.86, 0.82, and 0.87 for the clinical model, the DL model, and the model combining both models' probabilities.

Conclusions. In their current state, when these complex models were applied to our clinical scenario, they did not seem to provide a net gain compared to our baseline clinical model.

Key Points

- The clinical and the deep learning models showed good results predicting *IDH* status.
- Combined, these models showed good overall test and diagnostic properties.

Under current standards, a glioma diagnosis is confirmed with tissue analyses and classification according to the 2021 World Health Organization Classification of Tumors of the Central Nervous System (WHO-CNS).[1] However, before surgery, a glioma diagnosis is presumed from magnetic resonance imaging (MRI) findings.[1–6] Thus, the MRI evaluation is fundamental in decision making regarding treatment strategy. Paradoxically, benefits of surgical treatment strategies may depend on the diagnosis, which is not known before surgery. For instance, although safety is a major priority in decision making, surgically induced deficits are known to be associated with reduced survival in *isocitrate dehydrogenase (IDH)* wild-type glioblastoma.[7] Conversely, in *IDH*-mutated lower-grade gliomas, the survival benefit of extensive resections is much higher, especially in patients with astrocytoma.[8] These factors, combined with patient rehabilitation potential and estimated

Importance of the Study

Although there is a need to perform personalized glioma surgery, preoperatively, the histomolecular diagnosis is most often unknown for patients with non or minimal enhancing presumed glioma. Application of clinical models is not widespread and knowledge about their predictors may also vary among clinicians. In this scenario, predicting *IDH* mutations with high specificity would allow time for better patient counseling and surgery planning. This study shows that in their current state, the usefulness of deep learning in the radiologically presumed lower-grade glioma scenario is not clearly superior to conventional models based on clinical data. Efforts to generate models to answer this clinical question should be encouraged.

recovery times, may affect surgical aggressiveness in decision making. Thus, there is great interest in predicting the histomolecular diagnosis based upon the preoperative MRI.[9,10]

In mixed datasets containing both typical *IDH* wild-type glioblastoma and *IDH*-mutated grade 2 gliomas, the prediction of *IDH* status is very accurate.[11] However, since the vast majority of glioblastomas exhibit ring-like enhancement and radiological necrosis, traits in practice never seen in *IDH*-mutated grade 2 tumors, this is not as surprising. The main clinical problem is thus not to separate these radiologically very different entities (Figure 1), albeit to predict *IDH* status in similar looking tumors, for instance the non and minimal enhancing diffuse gliomas traditionally said to be radiologically presumed lower-grade gliomas (dLGG). Non and minimal enhancing diffuse gliomas, usually seen in preoperative contrast-enhanced T1-weighted (T1c), include no contrast enhancement, faint and patchy contrast enhancement, and focal or nodular contrast enhancement in the intra-tumoral area.[12–14]

In our recent population-based study, approximately one third of the radiologically presumed dLGG (including non and minimal enhancing tumors) held a different diagnosis than dLGG after histomolecular evaluation.[15] This phenomenon is also reflected in other case selections.[14,16] The idea that non-enhancing gliomas can be clearly separated is probably fueled because of specific features seen and used clinically as well. Features such as calcifications or the so-called T2-FLAIR mismatch provide fair hints of diagnosis.[17]

Although extensive research has been produced in the field of deep learning, recent reviews have shown that there is a misuse of poor reporting standards and a lack of validation in clinical settings.[11,18–20] Thus, it is still unclear what the net gain of complex models for radiological *IDH* status prediction is. Our aim was to test the clinical usefulness and to present the current level of performance using previously published well-performing deep learning architectures in radiologically presumed lower-grade glioma without significant contrast enhancement. We used multicenter cohorts for model building and an external validation cohort, with reporting according to recent checklists developed for this field.

Materials and Methods

Patient Inclusion

In this multicentric study including adult patients diagnosed with glioma, patients from 6 neurosurgical departments were selected for training and in-training validation of the predictive models (ie, modeling cohort). We selected patients from the Erasmus Glioma Database as our external validation cohort, serving as an unseen dataset.[21] The 2016 World Health Organization Classification of Tumors of the Central Nervous System (WHO-CNS) was used for this study.[22] The inclusion criteria were adult patients (age ≥ 18 years old) with a diagnosis of diffuse glioma grades 2 or 3 with availability of

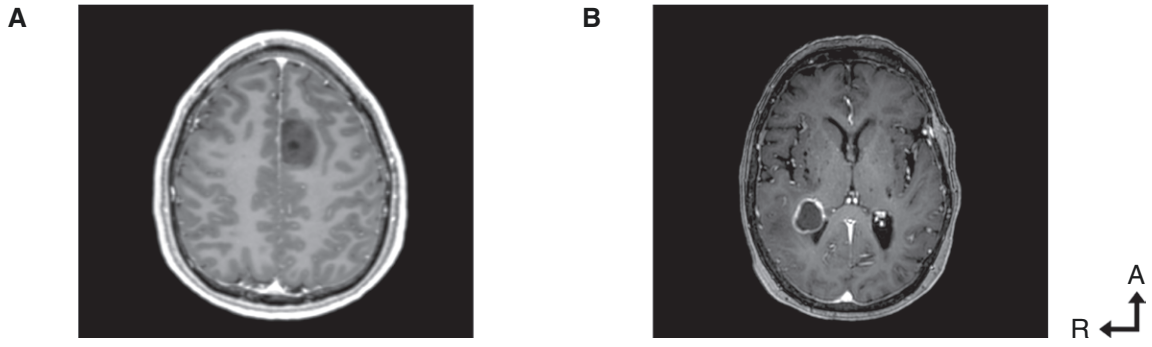

Figure 1. IDH-mutated diffuse glioma grade 2 compared to a grade 4 IDH wild-type glioblastoma. Axial images of 2 MRI, contrast-enhanced T1-weighted, showing a patient with a non-enhancing glioma (A) compared to another patient with a typical ring-like enhancing glioblastoma (B).

routine preoperative MRI images (T1c, T2w images, and FLAIR images). Patients with a diagnosis of diffuse glioma WHO-CNS 2016 grade 4, and patients lacking molecular data needed for the WHO-CNS 2016 classification were excluded. For a flow chart and a detailed list of excluded patients by center of origin, we refer to the Supplementary Material.

The study was approved by the Swedish Ethical Review Authority, nr. 702-18. For further details on inclusion and the collected data, we refer to the Supplementary Material.

Clinical Variables and Analytical Methods

Recorded clinical variables included age, sex, tumor volume, coordinates of the tumor center of gravity (tumor location), molecular assessment of *IDH* status, and 1p19q status (required only for *IDH*-mutated tumors). Molecular assessment was provided by each center according to their institutional practice.

Tumor delineations in the preoperative MRI for the modeling cohort were produced centrally using a semiautomated method that was previously reported (ie, image segmentation).[23] The Erasmus Glioma Database was provided with tumor delineations.[21] Due to the characteristics of the image processing method, in cases of multifocal tumors, only the main tumor component was preserved as the region of interest. For details on the segmentations of the region of interest, we refer to the Supplementary Material.

All images were processed similarly before the analysis, producing a multicentric train and in-training validation dataset and an equally processed external validation dataset. The final datasets consisted of 3 sets of axial images per patient, including 3 MRI modalities T1c, T2w, and FLAIR. To avoid information leakages (ie, leaking information into the training data from the external validation dataset), all images were separately processed strictly respecting patient and the origin center partition. We followed a standard image processing method previously reported.[24] For details on image processing including all the steps from image extraction from the respective hospital PACS (picture archiving and communication system) to the input of the artificial network, we refer to the Supplementary Material.

Statistics and Probability

Descriptive and comparative statistics were produced with IBM SPSS Statistics 28 (IBM Corp.). Central tendencies and variable distribution are expressed in means and standard deviations (SD). Data distribution was assessed with the Kolmogorov–Smirnov test. Comparison between groups were produced with the *t*-test or the Mann–Whitney test, and the Fisher's exact test, as appropriate. In contrast to the machine-learning regression model, the deep learning model require 2 datasets for model building (ie, train and in-training validation partitions). For comparison purposes, data about train and in-training validation partitions are grouped and presented together as the *modeling cohort*. The external validation dataset is presented as the *external validation cohort*.

Model Generation

A clinical-based machine-learning model was trained using multiple logistic regression. First, simple logistic regressions with *IDH* status as the dependent variable were performed using the available preoperative clinical variables of the modeling cohort as predictors (ie, age, sex, tumor volume, and the coordinates of the tumor center of gravity). Then, a result-driven multivariable logistic regression based upon the statistically significant predictors from the simple logistic regressions was produced. It also used *IDH* status as the dependent variable, and age together with the coordinates of the tumor center of gravity as predictors. All tests were 2-sided, the statistically significant level was set to P-value < .05. For details on the simple and multivariable logistic regressions, we refer to the Supplementary Material.

The deep learning models were trained only with the preoperative imaging data without access to the preoperative clinical variables. Based on the evidence from the available literature, 4 analytical frameworks (ie, reproducible deep learning architectures or artificial neural networks) were chosen: Convolutional Auto-Encoders, Residual Network, Dense Network, and Mobile Network.[25–28] For a short discussion of the selection process, we refer to the Supplementary Material. Using the modeling cohort, all frameworks were trained and internally validated respecting the same data partition (train and in-training validation) and the same set of hyperparameters. According to a preplanned search grid, a total of 208 models were built. Only the model with the best performance in the in-training validation data partition was selected for external validation and discussed in the study. For details on data partition, set of hyperparameters, choice of analytical frameworks, and metrics of performance for model selection, we refer to the Supplementary Material.

In short, the clinical-based machine-learning model (hereafter baseline clinical model) and the deep learning model were generated based exclusively on the modeling cohort and were scored using the external validation cohort. For a flow chart describing model generation and scoring workflow, we refer to the Supplementary Material. All the results disclosed in the following sections derived from scoring these models in the external validation cohort.

Metrics of Reporting and Performance Evaluation

The metrics of reporting included a confusion matrix of the actual and predicted positive (*IDH*-mutated) and negative (*IDH* wild-type) classes, together with the model accuracy, sensitivity, specificity, positive predicted value (PPV), negative predicted value (NPV), and area under the curve (AUC) of the prediction (ie, classification task). Except for the area under the curve that was calculated with SPSS based upon the probability of the positive class of the resulting prediction, all other metrics derive from the confusion matrix. Youden's J statistic was used to select the optimal predicted probability cut off. In addition, the default classification cutoff of 0.5 together with an optimal cutoff at specificity 0.90 were used to report the model metrics in the Supplementary Material.

Radiological Post Hoc Analysis

A radiological post hoc analysis was carried on as an additional evaluation method. All cases misclassified by the baseline clinical model, the deep learning model, or the combined model (ie, false positives and false negatives), were visually inspected by radiologists (E.T., J.F.S.). The inspection had 2 aims: first, to identify any irregularities in the images resulting from technical manipulation (image processing); second, to identify relevant clinical patterns in the images associated with the results of the prediction.

Standards of Reporting

In this study, we followed the recommendations of the Checklist for Artificial Intelligence in Medical Imaging, including aspects covered by the reproducibility and replicability checklist, and the checklist for transparent reporting of multivariable prediction models for individual prognosis or diagnosis.[29–31] These are available in the Supplementary Material together with the checklist for evaluation of radiomics research.[32]

Results

Patient Characteristics

A total of 1576 patients were screened for the inclusion, 570 patients were excluded due to a diagnosis other than diffuse glioma grades 2 or 3 (according to the WHO 2016 classification), 181 patients were excluded due to missing molecular data, and 356 patients were excluded due to lack of one or more of the necessary MRI sequences after image processing. Thus, a total of 469 patients were included in the study. The mean age was 44.8 years (SD 14.4), and 268 were males (57 %). In the study cohort, there were 352 patients with *IDH*-mutated tumors (75%), 173 patients with 1p19q co-deleted tumors (37%), and 333 patients with grade 2 tumors (71%). For details on tumor grade according to WHO 2016, contrast enhancement, volume, and location, we refer to Table 1.

Table 1. Patient and Tumor Characteristics (N = 469)

Variable	Study cohort
Age at surgery, mean (SD)	44.8 (14.4)
Male, No (%)	268 (57.1)
Molecular data, No (%):	
IDH[a] mutated	352 (75.1)
1p19q[b] co-deletion	173 (36.9)
WHO[c] 2016 tumor grade, No (%)	
Grade 2	333 (71.0)
Grade 3	136 (29.0)
Contrast enhancement in T1c images, No (%)	
Faint and patchy	108 (23)
Focal and/or nodular	46 (9.8)*
Coordinates of the tumor center of gravity in MNI[d] image space, mean (SD)	
x	88.3 (24.3)
y	134.1 (27.5)
z	95.7 (28.9)
Tumor volume in MNI space, mean (SD)	
Milliliters	78.5 (79.6)

[a]Isocitrate dehydrogenase.
[b]Short arm chromosome 1 and the long arm of chromosome 19.
[c]World Health Organization Classification of Tumors of the Central Nervous System.
[d]Montreal Neurological Institute.
*In 5 of the cases part of the tumor showed ring-like enhancement.

Data Partition and Cohort Comparison

Data partition was made at the patient level including 255 (81.2%) patients in the training partition and 59 (18.8%) patients in the in-training validation partition. The training partition of the modeling cohort include patients from Sahlgrenska University Hospital (227 cases), Karolinska University Hospital (20 cases), and St. Olav's University Hospital (8 cases). The in-training validation partition of the modeling cohort include patients exclusively from GHU Paris Psychiatrie & Neurosciences (12 cases), University of California (28 cases), and Medical University of Vienna (19 cases). Finally, the external validation cohort included only patients from The Erasmus Glioma Database (155 cases).

A comparison between patients in the modeling and external validation cohorts showed that there was no statistically significant difference between cohorts regarding age, tumor volume, or tumor location (measured as center of gravity). Sex, *IDH* status, and tumor grade were different between the modeling and external validation cohorts. A similar proportion of minimal enhancing tumors was found among cohorts, with proportionally less focal or nodular contrast enhancing tumors in the external validation cohort. For details, we refer to Table 2.

Baseline Clinical Model

The AUC of the baseline clinical model predicting *IDH*-mutated status based on clinical variables was 0.86 (accuracy 0.76, sensitivity 0.73, specificity 0.95, PPV 0.99, NPV 0.35). A confusion matrix including actual and predicted values of the baseline clinical model is disclosed in Table 3.

Analytical Frameworks

The best performing analytical framework was ResNet152.[26] The AUC of this deep learning model classifying *IDH*-mutated status based on preoperative MRI was 0.82 (accuracy 0.82, sensitivity 0.82, specificity 0.80, PPV 0.97, NPV 0.40). A confusion matrix including actual and predicted values of the deep learning model is disclosed in Table 3. The results from the grid search are reported in the Supplementary Material.

Table 2. Cohort Comparison of Patient and Tumor Characteristics (*N* = 469)

Variable	Modeling cohort *N* = 314	External validation cohort *N* = 155	*P*-value
Age at surgery, mean (SD)	45.0 (14.3)	44.3 (14.6)	.65
Male, No (%)	168 (53.5)	100 (64.5)	.03
Molecular data, No (%):			
IDH[a] mutated	217 (69.1)	135 (87.1)	<.01
1p19q[b] co-deletion	104 (33.1)	69 (44.5)	.07
WHO[c] 2016 tumor grade, No (%)			
Grade 2	206 (65.6)	127 (81.9)	<.01
Contrast enhancement in T1c images, No (%)			
Faint and patchy	64 (20.4)	44 (28.4)	.06
Focal and/or nodular	39 (12.4)[*]	7 (4.5)[**]	<.01
Coordinates of the tumor center of gravity in MNI[c] image space, mean (SD)			
x	87.2 (23.8)	90.5 (25.1)	.17
y	134.1 (28.1)	134.2 (26.1)	.98
z	96.2 (29.0)	94.7 (31.5)	.62
Tumor volume in MNI[d] space, mean (SD)			
Milliliters	75.2 (73.2)	85.0 (91.1)	.24

[a]Isocitrate dehydrogenase.
[b]Short arm chromosome 1 and the long arm of chromosome 19.
[c]World Health Organization Classification of Tumors of the Central Nervous System.
[d]Montreal Neurological Institute.
[*]In 3 of the cases, part of the tumor showed ring-like enhancement.
[**]In 2 of the cases, part of the tumor showed ring-like enhancement.

Table 3. Confusion Matrices. Results from the baseline clinical model, the deep learning model, and the combined model. All models were scored with the external validation cohort (*N* = 155)

Confusion matrix baseline clinical model			
		Actual condition	
		Positive	Negative
Predicted condition	Predicted positive	99	1
	Predicted negative	36	19
Confusion matrix deep learning model			
		Actual condition	
		Positive	Negative
Predicted condition	Predicted positive	111	4
	Predicted negative	24	16
Confusion matrix combined model			
		Actual condition	
		Positive	Negative
Predicted condition	Predicted positive	126	5
	Predicted negative	9	15

Combined Results

To combine the baseline clinical model and the deep learning model, the individual probabilities of both models were averaged (Figures 2 and 3). The AUC of the combined model classifying *IDH*-mutated status based on the averaged probabilities of the clinical and deep learning models was 0.87 (accuracy 0.91, sensitivity 0.93, specificity 0.75, PPV 0.96, NPV 0.63). A confusion matrix including actual and predicted values of the combined model is disclosed in Table 3. Additionally, a comparison of the results for all three model is shown in Table 4.

Radiological Post Hoc Analysis

A total of 50 cases were misclassified by any of the models. For details on the number of false positives and false negatives for each one of the models, we refer to Table 3. Upon traditional radiological interpretation, no irregularities product of technical manipulation were found. An inspection of the original MRI sequences for all 50 misclassified cases revealed the presence of artifacts in 17 (34%) cases when any of the MRI modalities were included.

Regarding the 4 false-positive cases predicted by the deep learning model, expert opinion coincided in that the radiological diagnosis indicated an *IDH*-mutated tumor.

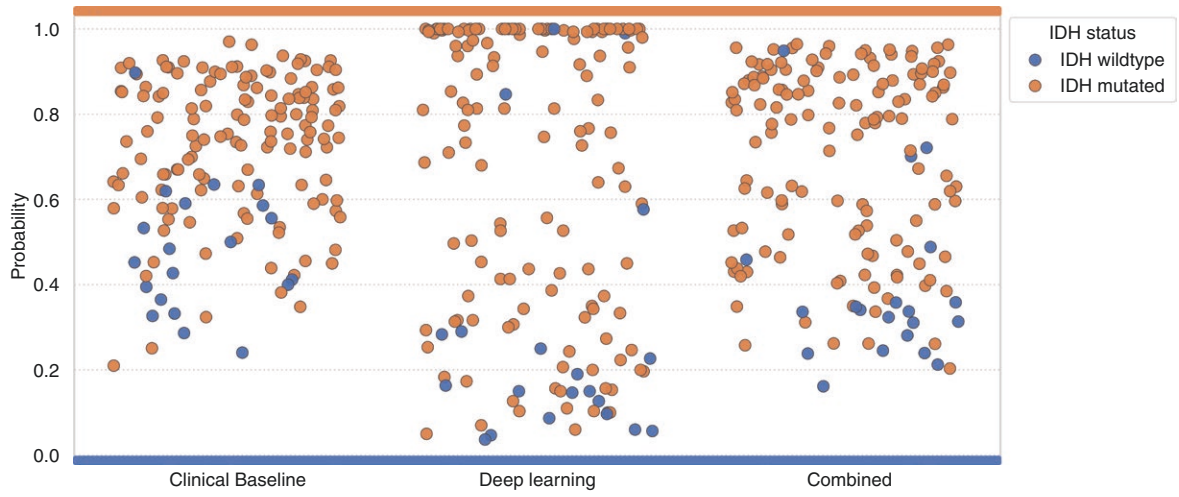

Figure 2. Scatter plot of the predicted and averaged probabilities. From left to right each column displays the predicted probabilities of each model (including the averaged combined model) for the 155 cases of the external test dataset. While the y-axis represents the predicted probability, the color markers represent the true IDH status.

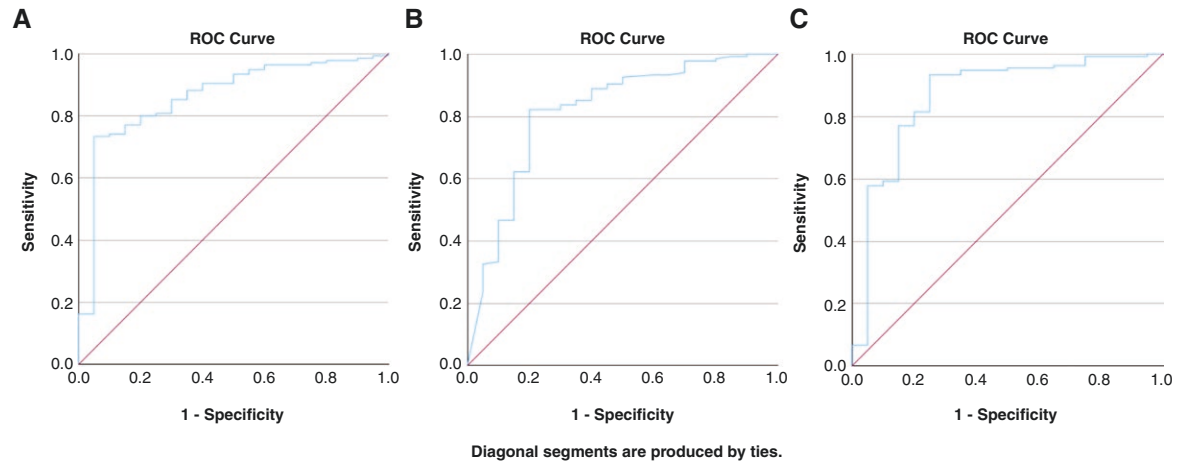

Figure 3. Receiver operating characteristic curves. From left to right each graph displays the receiver operating characteristic curves of each model (including the averaged combined model) on the external test dataset: baseline clinical model (A); the deep learning model (B); and the combined model (C).

None of the 4 false-positive cases expressed contrast enhancement in T1c images, and one of them showed signs of T2-FLAIR mismatch. As for the 24 false-negative cases predicted by the deep learning model, 12 (50%) of the false-negative cases extended over any of the temporal lobes. In addition, 6 (25%) of the 24 false-negative cases expressed faint or nodular contrast enhancement in T1c images.

Discussion

Under the premise that *radiologically presumed lower-grade glioma* (grades 2 and 3 with non or minimal enhancement) encompasses different tumor entities based on their *IDH* mutation status, we tested the usefulness of deep learning models in this clinical scenario by comparing their best prediction results with the predictions results of the baseline clinical model. Interestingly, the baseline clinical model showed good performance in predicting *IDH* status based solely upon patients' age and voxel-based tumor location. Furthermore, when the models were combined the results showed good overall test and diagnostic properties.

Compared to the baseline clinical model, the deep learning model showed lower specificity and higher sensitivity in predicting *IDH*-mutated tumors. Showing that, both models have the potential to answer different clinical questions based upon the selected operating point

Table 4. Comparative Results Table of the Baseline, Deep Learning, and Combined Models Scored with the External Validation Cohort (*N* = 155)

Model	Acc.[a]	Sen.	Spec.	PPV	NPV	AUC
Baseline clinical model	0.76	0.73	0.95	0.99	0.35	0.86
Deep learning model	0.82	0.82	0.80	0.97	0.40	0.82
Combined model	0.91	0.93	0.75	0.96	0.63	0.87

[a]All values were rounded to 2 decimals. **Abbreviations:** accuracy (acc.), sensitivity (sen.), specificity (spec.), positive predicted value (PPV), negative predicted value (NPV), area under the curve (AUC). Optimal predicted probability cutoff values from the coordinates of the AUC tables.

(ie, prediction threshold). Furthermore, when both models were combined, we noticed a slight improvement in the AUC together with better overall test and diagnostic properties. Our findings were also sustained by the qualitative radiological post hoc analysis, showing that no irregularities were found that could have undermined the performances of the models. In the proposed scenario, when aiming to preoperatively identify *IDH*-mutated tumors, the deep learning model failed to deliver additional gain over the baseline clinical model, where we found the baseline clinical model to deliver more specific results than both the deep learning and the combined model.

We are aware of 6 clinical studies applying deep learning in a similar setting (ie, the identification of *IDH* status in routine preoperative MRI of patients with diffuse glioma grades 2 or 3).[33–38] However, only one of the studies validated their findings with an external cohort.[34] The others applied different methods of data partition using unseeing cases to validate their findings.[33,35–38] Altogether, the studies without external validation reported accuracies ranging from 0.73 to 0.94, claiming also high sensitivity and specificity. In contrast, and in line with our results, the only study disclosing external validation results reported an AUC of 0.81 (accuracy 0.74, sensitivity 0.73, specificity 0.77).[34] Our findings are in line with the literature, additionally showing that when compared with a baseline clinical model the clinical gain provided by these complex models at present seems to be modest. Still, application of clinical models is not widespread and knowledge about their predictors may also vary among clinicians. It might therefore prove more practical to implement a DL model based on automated MRI analyses.

Recent studies focusing on predicting *IDH* status applying deep learning in mixed glioma cohorts have shown promising results.[34,39–43] Like in our study, these approaches also use data collections in addition to single or multicenter cohorts, validating their findings with external datasets. Unfortunately, most of the patients in these studies have WHO-CNS 2016 grade 4 tumors and presumably a high proportion of contrast enhancing tumors. Altogether, these studies reported AUC ranging from 0.79 to impressive 0.97. Only one of the studies included a clinical baseline consisting of a logistic regression based on patients' age and the presence of contrast enhancement.[41] Although not directly answering to the *presumed diffuse lower-grade glioma scenario*, the additional clinical gain provided by these models remains unknown, specially knowing that conventional radiological assessment already showed good results identifying adult-type diffuse glioblastoma.[2]

We believe that due to the generalized lack of essential metrics, poor reporting, and inclusion of mixed glioma cohorts, current deep learning implementations fail to provide a unified radiological solution in glioma tumor sub-classification. Our results indicate that DL models may facilitate sensitivity tuned inquiries showing that deep learning approaches, when applied to MRI in patients with *radiologically presumed lower-grade glioma*, seem to be able to produce accurate and stable predictions of *IDH* status. While DL models still have promising potential to predict *IDH* status, in our proposed scenario these complex models do not seem to provide a net gain when compared to baseline clinical models.

Although our study followed comprehensive reporting standards, it has some relevant limitations. Due to the restrictions imposed by the ethical review board, were we unable to include patients with diagnoses other than glioma. Based on our clinical experience, in clinical cohorts of radiologically presumed adult-type diffuse glioma grades 2 or 3, there is generally a small proportion of patients with a differential diagnosis (eg, gangliglioma and dysembryoplastic neuroepithelial tumors). In addition, our cohort lacks identification of CDKN2A/B homozygous deletions, a key molecular marker for the identification of *IDH*-mutant grade 4 tumors that might occur in a rather small proportion of our cohort. Due to the strict registration process applied to the images in the study, a considerable proportion of the modeling cohort was considered as missing key MRI sequences. This is presented as a limitation given the unavoidable dropout. However, it also represents a strength since images with irregularities from technical manipulation were discarded. Due to the characteristics of the data, model generation and the comparison of the predicted results were carried on an unbalance dataset as is expected in presumed dLGG. To compensate for these class imbalances, different loss functions with focal hyperparameters were used to train the deep learning models. In addition, the predicted results were presented in confusion matrices including the actual patient count. Also, the inclusion of a radiological post hoc analysis allowed us to qualitatively assess our findings. Acknowledging that the DL models might be further improved, for instance, by including the clinical variables as an additional input, for comparability purposes, the models were implemented respecting their original architectures.

Considering these limitations, our study is the first to approach applying deep learning to an entire cohort consisting exclusively of patients with adult-type diffuse glioma grades

2 or 3, typically demonstrating non or minimal contrast enhancement in which the final validation test was produced with an external cohort. Also, our study presents a large cohort of patients with adult-type diffuse glioma grades 2 or 3 subject to a deep learning *IDH* status prediction, where the addition of a clinical baseline model provides new insights over the advantages and disadvantages of the tested method.

Conclusions

The compared models showed moderate to good test and diagnostic properties. However, in their current state, the usefulness of deep learning in the *radiologically presumed lower-grade glioma* scenario is not clearly superior to conventional models based on clinical data. Efforts to generate models to answer clinically relevant questions should be encouraged.

Supplementary material

Supplementary material is available online at *Neuro-Oncology Advances* (https://academic.oup.com/noa).

Keywords

deep learning | glioma | grade 2 | grade 3 | isocitrate dehydrogenase | magnetic resonance imaging

Funding

Swedish state under the agreement between the Swedish Government and the county councils (ALFGBG-965622); Swedish Research Council (2017-00944).

Conflict of interest statement

None of the authors have commercial interests to disclose.

Author contributions

Conceptualization and design: T.G.V., A.N., A.S.J. Data collection: T.G.V., A.N., O.S., J.P., M.B., G.W., J.B., A.S.J. Data analysis and interpretation: T.G.V., A.N., J.F., E.T., I.H., A.S.J. Visualization (figures and tables): T.G.V. First draft of the manuscript: T.G.V. Supervision: I.Y.H.G., A.S.J. Material support/financing: A.S.J. Revising the manuscript: All authors. Final approval of the manuscript: All authors.

Data availability

The raw data and developed model weights supporting the conclusion of this article will be made available by the authors upon reasonable request.

Affiliations

Department of Clinical Neuroscience, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden (T.G.V., A.N., E.T., A.S.J.); Department of Radiology, Sahlgrenska University Hospital, Gothenburg, Sweden (E.T.); Medical Image Analysis and Artificial Intelligence, Danube Private University, Krems an der Donau, Austria (J.F.); Department of Neurosurgery, St. Olavs Hospital, Trondheim, Norway (O.S.); Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway (O.S.); Department of Neurosurgery, GHU Paris Psychiatrie & Neurosciences, Paris, France (J.P.); Department of Neurosurgery, University of California, San Francisco, San Francisco, California, USA (M.B.); Department of Neurosurgery, Medical University of Vienna, Vienna, Austria (G.W.); Department of Neurosurgery, Karolinska University Hospital, Stockholm, Sweden (J.B.); Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden (J.B.); Department of Neurosurgery, Rigshospitalet, Copenhagen, Denmark (J.B.); Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden (I.H., I.Y.H.G.); Department of Medical Radiation Sciences, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden (I.H.); Department of Neurosurgery, Sahlgrenska University Hospital, Gothenburg, Sweden (A.S.J.)

References

1. 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(8):1231–1251.
2. Lasocki A, Anjari M, Örs Kokurcan S, Thust SC. Conventional MRI features of adult diffuse glioma molecular subtypes: a systematic review. *Neuroradiology*. 2021;63(3):353–362.
3. Wen PY, van den Bent M, Vogelbaum MA, Chang SM. RANO 2.0: The revised Response Assessment in Neuro-Oncology (RANO) criteria for high- and low-grade glial tumors in adults designed for the future. *Neuro Oncol*. 2024;26(1):2–4.
4. Miller JJ, Gonzalez Castro LN, McBrayer S, et al. Isocitrate dehydrogenase (IDH) mutant gliomas: a Society for Neuro-Oncology (SNO) consensus review on diagnosis, management, and future directions. *Neuro Oncol*. 2023;25(1):4–25.
5. Weller M, van den Bent M, Preusser M, et al. EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. *Nat Rev Clin Oncol*. 2021;18(3):170–186.

6. Munkvold BKR, Solheim O, Bartek J Jr, et al. Variations in the management of diffuse low-grade gliomas-A Scandinavian multicenter study. *Neurooncol. Pract.*. 2021;8(6):706–717.
7. Aabedi AA, Young JS, Zhang Y, et al. Association of neurological impairment on the relative benefit of maximal extent of resection in chemoradiation-treated newly diagnosed isocitrate dehydrogenase wild-type glioblastoma. *Neurosurgery.* 2022;90(1):124–130.
8. Patel SH, Bansal AG, Young EB, et al. Extent of surgical resection in lower-grade gliomas: differential impact based on molecular subtype. *AJNR Am J Neuroradiol.* 2019;40(7):1149–1155.
9. Kalaroopan D, Lasocki A. MRI-based deep learning techniques for the prediction of isocitrate dehydrogenase and 1p/19q status in grade 2-4 adult gliomas. *J Med Imaging Radiat Oncol.* 2023;67(5):492–498.
10. Li L, Wang Y, Li Y, Fang S, Jiang T. Role of molecular biomarkers in glioma resection: a systematic review. *Chin Neurosurg J.* 2020;6:18.
11. Sun W, Song C, Tang C, et al. Performance of deep learning algorithms to distinguish high-grade glioma from low-grade glioma: a systematic review and meta-analysis. *iScience.* 2023;26(6):106815.
12. Pallud J, Capelle L, Taillandier L, et al. Prognostic significance of imaging contrast enhancement for WHO grade II gliomas. *Neuro Oncol.* 2009;11(2):176–182.
13. Widhalm G, Kiesel B, Woehrer A, et al. 5-Aminolevulinic acid induced fluorescence is a powerful intraoperative marker for precise histopathological grading of gliomas with non-significant contrast-enhancement. *PLoS One.* 2013;8(10):e76988.
14. Widhalm G, Olson J, Weller J, et al. The value of visible 5-ALA fluorescence and quantitative protoporphyrin IX analysis for improved surgery of suspected low-grade gliomas. *J Neurosurg.* 2019;133(1):79–88.
15. Gomez Vecchio T, Ryden I, Ozanne A, et al. Global health status and fatigue score in isocitrate dehydrogenase-mutant diffuse glioma grades 2 and 3: a longitudinal population-based study from surgery to 12-month follow-up. *Neurooncol. Pract.*. 2024;11(3):347–357.
16. Kunz M, Albert NL, Unterrainer M, et al. Dynamic 18F-FET PET is a powerful imaging biomarker in gadolinium-negative gliomas. *Neuro Oncol.* 2019;21(2):274–284.
17. Kamble AN, Agrawal NK, Koundal S, et al. Imaging-based stratification of adult gliomas prognosticates survival and correlates with the 2021 WHO classification. *Neuroradiology.* 2023;65(1):41–54.
18. Bahar RC, Merkaj S, Cassinelli Petersen GI, et al. Machine learning models for classifying high- and low-grade gliomas: a systematic review and quality of reporting analysis. *Front Oncol.* 2022;12:856231.
19. Bhandari A, Scott L, Weilbach M, Marwah R, Lasocki A. Assessment of artificial intelligence (AI) reporting methodology in glioma MRI studies using the Checklist for AI in Medical Imaging (CLAIM). *Neuroradiology.* 2023;65(5):907–913.
20. Zhang S, Yin L, Ma L, Sun H. Artificial intelligence applications in glioma with 1p/19q co-deletion: a systematic review. *J Magn Reson Imaging.* 2023;58(5):1338–1352.
21. 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.
22. 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(6):803–820.
23. Corell A, Ferreyra Vega S, Hoefling N, et al. The clinical significance of the T2-FLAIR mismatch sign in grade II and III gliomas: a population-based study. *BMC Cancer.* 2020;20(1):450.
24. Gomez Vecchio T, Neimantaite A, Corell A, et al. Lower-grade gliomas: an epidemiological voxel-based analysis of location and proximity to eloquent regions. *Front Oncol.* 2021;11(3687):748229.
25. Masci J, Meier U, Cireşan D, Schmidhuber J. Stacked convolutional auto-encoders for hierarchical feature extraction. Paper presented at: Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 212011.
26. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016; Las Vegas, NV.
27. Huang G, Liu Z, Weinberger KQ. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017:2261–2269; Honolulu, HI.
28. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. Mobilenetv2: Inverted residuals and linear bottlenecks. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018; Salt Lake City, UT.
29. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell.* 2020;2(2):e200029.
30. Gryska E, Bjorkman-Burtscher I, Jakola AS, et al. Deep learning for automatic brain tumour segmentation on MRI: evaluation of recommended reporting criteria via a reproduction and replication study. *BMJ Open.* 2022;12(7):e059000.
31. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ.* 2015;350:g7594.
32. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging.* 2023;14(1):75.
33. Liu L, Chang J, Zhang P, Qiao H, Xiong S. SASG-GCN: self-attention similarity guided graph convolutional network for multi-type lower-grade glioma classification. *IEEE J Biomed Health Inform.* 2023;27(7):3384–3395.
34. van der Voort SR, Incekara F, Wijnenga MMJ, et al. Combined molecular subtyping, grading, and segmentation of glioma using multi-task deep learning. *Neuro Oncol.* 2023;25(2):279–289.
35. Matsui Y, Maruyama T, Nitta M, et al. Prediction of lower-grade glioma molecular subtypes using deep learning. *J Neurooncol.* 2020;146(2):321–327.
36. Ali MB, Gu IY, Berger MS, et al. Domain mapping and deep learning from multiple MRI clinical datasets for prediction of molecular subtypes in low grade gliomas. *Brain Sci.* 2020;10(7):463.
37. Fukuma R, Yanagisawa T, Kinoshita M, et al. Prediction of IDH and TERT promoter mutations in low-grade glioma from magnetic resonance images using a convolutional neural network. *Sci Rep.* 2019;9(1):20311.
38. Li Z, Wang Y, Yu J, Guo Y, Cao W. Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. *Sci Rep.* 2017;7(1):5467.
39. Bangalore Yogananda CG, Wagner BC, Truong NCD, et al. MRI-based deep learning method for classification of IDH mutation status. *Bioengineering (Basel).* 2023;10(9):1045.
40. Chakrabarty S, LaMontagne P, Shimony J, Marcus DS, Sotiras A. MRI-based classification of IDH mutation and 1p/19q codeletion status of gliomas using a 2.5D hybrid multi-task convolutional neural network. *Neurooncol. Adv.*. 2023;5(1):vdad023.
41. Cluceru J, Interian Y, Phillips JJ, et al. Improving the noninvasive classification of glioma genetic subtype with deep learning and diffusion-weighted imaging. *Neuro Oncol.* 2022;24(4):639–652.
42. Choi YS, Bae S, Chang JH, et al. Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics. *Neuro Oncol.* 2021;23(2):304–313.
43. Chen M, Zhang M, Yin L, et al. Medical image foundation models in assisting diagnosis of brain tumors: a pilot study. *Eur Radiol.* 2024;34(10):6667–6679.