

Are your MRI contrast agents cost-effective?

Learn more about generic Gadolinium-Based Contrast Agents.



**AJNR**

This information is current as of May 3, 2024.

## **Systematic Literature Review of Machine Learning Algorithms Using Pretherapy Radiologic Imaging for Glioma Molecular Subtype Prediction**

Jan Lost, Tej Verma, Leon Jekel, Marc von Reppert, Niklas Tillmanns, Sara Merkaj, Gabriel Cassinelli Petersen, Ryan Bahar, Ayyüce Gordem, Muhammad A. Haider, Harry Subramanian, Waverly Brim, Ichiro Ikuta, Antonio Omuro, Gian Marco Conte, Bernadette V. Marquez-Nostra, Arman Avesta, Khaled Bousabarah, Ali Nabavizadeh, Anahita Fathi Kazerooni, Sanjay Aneja, Spyridon Bakas, MingDe Lin, Michael Sabel and Mariam Aboian

*AJNR Am J Neuroradiol* 2023, 44 (10) 1126-1134

doi: <https://doi.org/10.3174/ajnr.A8000>

<http://www.ajnr.org/content/44/10/1126>

# Systematic Literature Review of Machine Learning Algorithms Using Pretherapy Radiologic Imaging for Glioma Molecular Subtype Prediction

 Jan Lost,  Tej Verma, Leon Jekel,  Marc von Reppert,  Niklas Tillmanns,  Sara Merhaj,  Gabriel Cassinelli Petersen,  Ryan Bahar,  Ayyüce Gordem, Muhammad A. Haider,  Harry Subramanian, Waverly Brim,  Ichiro Ikuta,  Antonio Omuro,  Gian Marco Conte, Bernadette V. Marquez-Nostra,  Arman Avesta, Khaled Bousabarah, Ali Nabavizadeh,  Anahita Fathi Kazerooni,  Sanjay Aneja,  Spyridon Bakas,  MingDe Lin, Michael Sabel, and  Mariam Aboian



## ABSTRACT

**BACKGROUND:** The molecular profile of gliomas is a prognostic indicator for survival, driving clinical decision-making for treatment. Pathology-based molecular diagnosis is challenging because of the invasiveness of the procedure, exclusion from neoadjuvant therapy options, and the heterogeneous nature of the tumor.

**PURPOSE:** We performed a systematic review of algorithms that predict molecular subtypes of gliomas from MR Imaging.

**DATA SOURCES:** Data sources were Ovid Embase, Ovid MEDLINE, Cochrane Central Register of Controlled Trials, Web of Science.

**STUDY SELECTION:** Per the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, 12,318 abstracts were screened and 1323 underwent full-text review, with 85 articles meeting the inclusion criteria.

**DATA ANALYSIS:** We compared prediction results from different machine learning approaches for predicting molecular subtypes of gliomas. Bias analysis was conducted for each study, following the Prediction model Risk Of Bias Assessment Tool (PROBAST) guidelines.

**DATA SYNTHESIS:** Isocitrate dehydrogenase mutation status was reported with an area under the curve and accuracy of 0.88 and 85% in internal validation and 0.86 and 87% in limited external validation data sets, respectively. For the prediction of *O6-methylguanine-DNA methyltransferase* promoter methylation, the area under the curve and accuracy in internal validation data sets were 0.79 and 77%, and in limited external validation, 0.89 and 83%, respectively. PROBAST scoring demonstrated high bias in all articles.

**LIMITATIONS:** The low number of external validation and studies with incomplete data resulted in unequal data analysis. Comparing the best prediction pipelines of each study may introduce bias.

**CONCLUSIONS:** While the high area under the curve and accuracy for the prediction of molecular subtypes of gliomas are reported in internal and external validation data sets, limited use of external validation and the increased risk of bias in all articles may present obstacles for clinical translation of these techniques.

**ABBREVIATIONS:** AUC = area under the curve; DL = deep learning; MGMT = *O6-methylguanine-DNA methyltransferase*; ML = machine learning; SVM = support vector machine; WHO = World Health Organization; IDH = *isocitrate dehydrogenase*

**G**liomas account for approximately 33% of brain tumor diagnoses; among adults, >50% of these cases are high-grade

Received February 27, 2023; accepted after revision August 1.

From the Departments of Radiology and Biomedical Imaging (J.L., T.V., L.J., M.v.R., N.T., S.M., G.C.P., R.B., A.G., M.A.H., H.S., W.B., B.V.M.-N., A.A., M.L., M.A.), Neurology and Yale Cancer Center (A.O.), and Therapeutic Radiology (S.A.), Yale School of Medicine, New Haven, Connecticut; Department of Neurosurgery (J.L., M.S.), Heinrich-Heine-University, Duesseldorf, Germany; Department of Radiology (I.I.), Mayo Clinic Arizona, Phoenix, Arizona; Department of Radiology (G.M.C.), Mayo Clinic, Rochester, Minnesota; Visage Imaging Inc (K.B., M.L.), San Diego, California; Department of Radiology (A.N.), Perelman School of Medicine, Hospital of University of Pennsylvania, University of Pennsylvania, Philadelphia, Pennsylvania; Department of Neurosurgery (A.F.K.), Center for Biomedical Image Computing and Analytics (S.B.), and Department of Radiology (S.B.), Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania; Division of Neurosurgery (A.F.K.), and Center for Data-Driven Discovery (A.F.K.), Children's Hospital of Philadelphia, Philadelphia, Pennsylvania; and Richards Medical Research Laboratories (S.B.), Philadelphia, Pennsylvania.

gliomas.<sup>1,2</sup> The 2021 World Health Organization (WHO) classification identified different forms of gliomas based on pathologic characteristics of biopsied or resected tumor and the molecular subtype.<sup>3</sup> This method of diagnosis requires invasive sampling of the tumor, which risks surgery and the potential for tumor misclassification due to tumor heterogeneity with sampling bias.<sup>4,5</sup>

Please address correspondence to Mariam Aboian, MD, Department of Radiology and Biomedical Imaging, Yale School of Medicine, 333 Cedar St, New Haven, CT 06510; e-mail: mariam.aboian@yale.edu

 Indicates open access to non-subscribers at [www.ajnr.org](http://www.ajnr.org)

 Indicates article with online supplemental data.

<http://dx.doi.org/10.3174/ajnr.A8000>

In addition, the inability to accurately predict tumor subtype before surgery can result in limited access to neoadjuvant therapies. Of note, the class of tumor affects a patient's predicted survival as well as medical and surgical treatment options.<sup>3,6</sup>

Recent advancements in machine learning (ML) applications in neuro-oncology have shown promise in tumor segmentation,<sup>7,8</sup> differentiating gliomas from other intracranial malignancies such as brain metastases<sup>9</sup> and lymphomas,<sup>10</sup> predicting glioma grade,<sup>11</sup> and predicting the patient's overall survival.<sup>12,13</sup> Most of the published literature on applying ML to neuro-oncology demonstrates a high area under the curve (AUC) and accuracy in the internal testing data sets, but validation of algorithms on external data sets is limited to a very few studies. As a result, there is a limited translation of algorithms among different hospitals and study settings. In addition, previous systematic reviews have demonstrated that most of the literature is focused on small well-curated data sets with an over-representation of commonly used data sets, such as RSNA-ASNR-MICCAI Brain Tumor Segmentation (BraTS; <https://www.kaggle.com/datasets/dschettler8845/brats-2021-task1>) and The Cancer Imaging Archive (TCIA; <https://wiki.cancerimagingarchive.net/display/Public/Collections>).<sup>8-12,14</sup>

This feature further limits translation of algorithms between different institutions. In addition, ML literature in neuro-oncology often scores low in established reporting guidelines such as Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD; <https://www.equator-network.org/reporting-guidelines/tripod-statement/>) and Checklist for Artificial Intelligence in Medial Imaging (CLAIM; <https://pubs.rsna.org/doi/10.1148/ryai.2020200029>) with high bias reported by the Prediction model Risk Of Bias ASsessment Tool (PROBAST).<sup>15-17</sup> These findings suggest that increased use of standardized reporting criteria in publications and the establishment of large databases of annotated images in individual hospitals are critically needed to translate useful algorithms into patient care and to bridge the gap between the diagnosis and targeted therapy. ML can potentially identify patterns that remain invisible to a radiologist's clinical interpretation and can extract information from pretreatment imaging that can influence the implementation of neoadjuvant therapy and the extent of surgical resection.

Our study evaluated the literature that uses ML to predict molecular subtypes of gliomas, such as *isocitrate dehydrogenase (IDH)* mutation status and *O6-methylguanine-DNA methyltransferase (MGMT)* promoter methylation status, which can change treatment options for patients. As an example, *IDH*-mutant astrocytomas have better survival and different treatment options compared with glioblastomas that are uniformly *IDH* wild-type.<sup>3</sup> Glioblastomas have dismal survival and require maximal resection to improve survival.<sup>18</sup> Therefore, predicting *IDH* mutation on preoperative imaging can change the patient's treatment strategy and outcomes. In addition, the prediction of *MGMT* methylation status can predict which patients will respond to temozolomide therapy and, therefore, would influence neoadjuvant options before resection.<sup>19</sup> To guide medical care, the algorithm must have high precision and sensitivity and be translatable to multiple hospitals with different imaging protocols. The goal of our study was to critically evaluate the literature that reports different

algorithms for predicting molecular subtypes of gliomas and to assess the potential obstacles to translating these algorithms into clinical practice.

## MATERIALS AND METHODS

### Study Selection

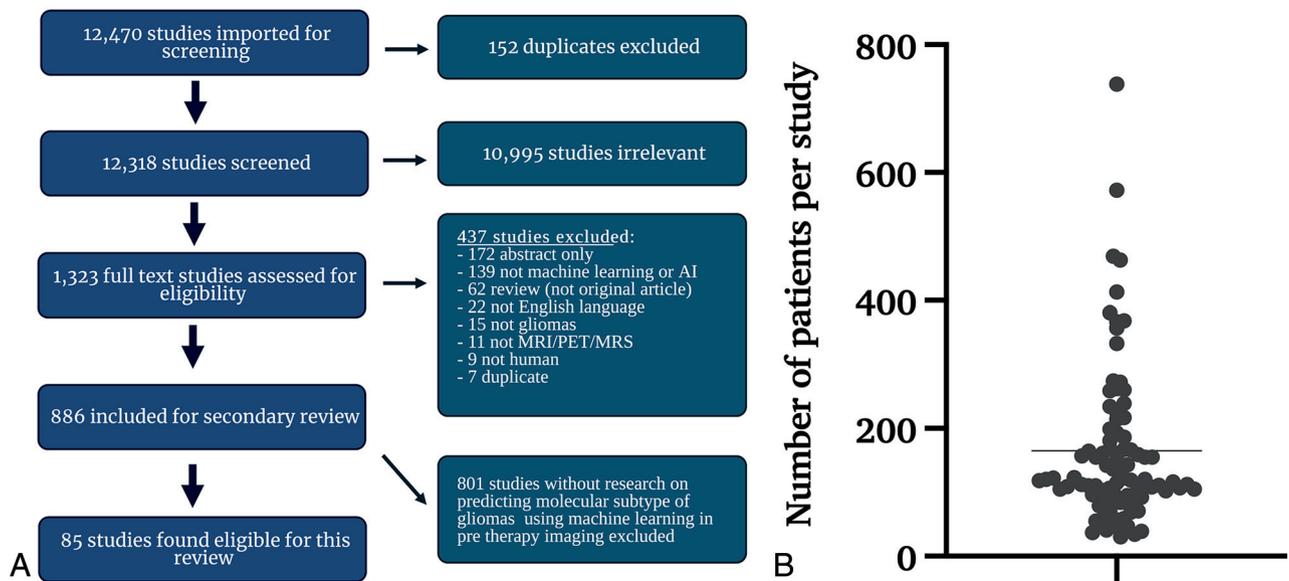
Literature screening for this systematic review conformed to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, using Ovid EMBASE, Ovid MEDLINE, Cochrane Central Register of Controlled Trials, and the Web of Science Core Collection as databases. The study was registered with the Prospective Register of Systematic Reviews (PROSPERO, CRD42020209938). A clinical librarian collected data in September 2020, January 2021, and September 2021, respectively. The literature search strategy involved the use of keywords such as "artificial intelligence," "machine learning," "deep learning," "radiomics," "MR imaging," "glioma," "glioblastoma," and related terms. We identified 12,470 studies (Fig 1A) and added them for further screening to the Covidence Software (Veritas Health Innovation).

After removing 152 duplicates, a neuroradiology attending physician, a neuroradiology resident, and 3 graduate students screened 12,318 studies. Conflicting assessments were resolved by the board-certified neuroradiology attending physician after discussion with screeners. After abstract reviews, 10,995 studies were excluded due to a lack of ML and neuro-oncology applicability, resulting in 1323 full-text reviews. A secondary full-text review was conducted on 886 studies, by either a radiology resident or a graduate student, followed by a second review by a board-certified neuroradiologist attending physician.

We predefined 8 uniform exclusion criteria: 1) abstract-only, 2) no application of ML reported, 3) not an original article, 4) not published in English, 5) no investigation of glioma/glioblastoma, 6) unrelated to MR imaging, MR spectroscopy, or PET imaging, 7) no human research subjects, and 8) duplicates. Overall, 437 studies met  $\geq 1$  of the predefined exclusion criteria and were subsequently excluded (Fig 1A). In the end, for this systematic review, 85 studies were included that specifically analyzed molecular subtype prediction in glioma, using ML techniques based on pretherapy imaging. This resulted in 801 studies being excluded from further review.

### Data Extraction

Data extraction was performed by 3 graduate students and 1 undergraduate in Excel (Microsoft 2022). Studies were reviewed twice, and disagreements were resolved in regular meetings with a supervising neuroradiology attending physician until a consensus was reached. Extracted information included article characteristics (title, authors, publication year), patient data (patient number, cohort sizes, data sources), tumor classification (tumor type, analyzed molecular subtypes), model characteristics (imaging data and best-performing ML algorithms), as well as validation techniques. Generally, accuracy results from the best-performing prediction pipeline were reported. If patient cohorts were split up within a study, different prediction results were reported. We accounted for the use of external validation if patient cohorts were geographically split. Data from the same



**FIG 1.** A, Inclusion/exclusion criteria and resultant study data. Flow chart represents screening workflow and exclusion criteria to visualize the eligibility of studies. The search strategy included keywords “artificial intelligence,” “machine learning,” “deep learning,” “radiomics,” “MR imaging,” “glioma,” “glioblastoma,” and related terms. An independent librarian reviewed the data. We predefined 8 uniform exclusion criteria: 1) abstract-only, 2) no application of ML reported, 3) not an original article, 4) not published in English, 5) no investigation of glioma/glioblastoma, 6) unrelated to MR imaging, MR spectroscopy, or PET imaging, 7) no human research subjects, and 8) duplicates. B, Distribution of all patients per study included in training or validation of a predictive model. We excluded patients whose data were strictly used for models other than the prediction of molecular subtypes. The *line* indicates the mean number of patients. AI indicates artificial intelligence.

source, solely split up by the time of inclusion, were classified as internal validation. To give more insight into the different validation types used, we split internal validation techniques into cross-validation and holdout-validation. For an overview of extracted data, see the Online Supplemental Data.

### Meta-analysis

All studies reporting AUC and 95% confidence intervals were subjected to a meta-analysis using MedCalc for Windows, Version 20.009 (MedCalc software). Heterogeneity was examined using the Higgins  $I^2$  test, publication bias was evaluated using the Egger test, and the results of the quantitative analysis were illustrated with a forest plot (Online Supplemental Data).

### Risk of Bias Assessment

Risk of bias was assessed using PROBAST.<sup>17</sup> Per PROBAST guidelines, studies were classified into development, validation, or development and validation studies. All included studies established predictive models and, therefore, were at least considered development studies; those that additionally tested their developed model on an independent cohort were considered to have both development and validation.

PROBAST uses signaling questions across 4 distinct domains to evaluate potential biases in each study. The first domain (participants) pertains to the data sources and participant enrollment. The second domain (predictors) evaluates the definition and measurement of predictors and their association with the outcome. Domain 3 (outcome) addresses potential biases in defining and terminating the outcome in each study. Last, domain 4 (analysis) evaluates whether inappropriate analysis methods were used or

important statistical considerations were overlooked.<sup>17</sup> Additional information on the application of PROBAST can be found in the work of Moons et al,<sup>17</sup> which is beyond the scope of this article. Signaling questions from each domain were assessed separately for development and validation cohorts. The average item scores from each study and the risk of bias for all 4 domains were evaluated using Excel (Microsoft).

## RESULTS

After a full-text review of 886 studies, 85 studies published between 2016 and 2021 met the eligibility criteria of our analysis.

### Patient Data

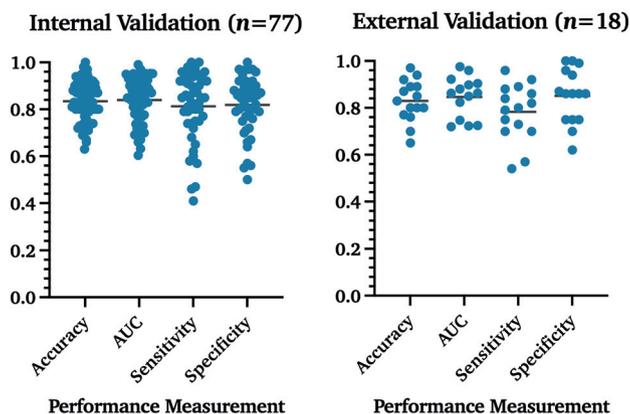
The mean patient number in all 85 studies was 165.32 (Fig 1B). Data were taken from single-center hospital data, public databases, or multicenter hospital data. The overall number of data sources included is higher than the total number of studies because 17 studies included patients from 2 sources, resulting in a total of 102 sources among the 85 studies. While 52% ( $n = 53/102$ ) of patient data was obtained from single-center hospitals, 33% ( $n = 34/102$ ) was from publicly available databases and 15% ( $n = 15/102$ ) was from multiple institutions (Online Supplemental Data).

### Description of Internal and External Validation Data Sets and Techniques

Of 85 studies, 7 reported  $>1$  distinct model pipeline to predict glioma molecular subtypes, resulting in 95 analyzed pipelines.<sup>20-26</sup> The best performing model results for each study are reported for molecular alterations in gliomas, including *IDH*, *MGMT*, *1p/19q* codeletion, histone *H3 K27M*, *ATRX*, *TERT*, and others. The results from external validation were reported. For studies with

**Table 1: Mean performance measurements of all studies**

	Accuracy	AUC	Sensitivity	Specificity
Internal validation	83% ( $n = 66/77$ )	0.84 ( $n = 60/77$ )	81% ( $n = 50/77$ )	82% ( $n = 49/77$ )
External validation	83% ( $n = 15/18$ )	0.85 ( $n = 14/18$ )	78% ( $n = 15/18$ )	85% ( $n = 15/18$ )
<i>P</i> value (Mann-Whitney <i>U</i> test)	.83	.79	.25	.45

**FIG 2.** Algorithm performance measurement in internal and external validation data sets. Percentages are reported as fractions to visualize measurements in 1 graph.

internal validation only, the holdout validation was prioritized over cross-validation. Techniques varied among leave-one-out-cross-validation, *n*-fold cross-validation to holdout, and external validation. In 5% ( $n = 5/95$ ) of the articles, the validation cohorts were separated from the initial cohort by time only and were, therefore, re-classified as internal validation in our analysis. Overall, 81% ( $n = 77/95$ ) used internal validation of any kind, while only 37% ( $n = 35/95$ ) used forms of cross-validation and 39% ( $n = 37/95$ ) used holdout validation. Only 19% ( $n = 18/95$ ) reported accuracy from an external validation cohort.

We performed a Mann-Whitney *U* test on the studies that reported accuracy, AUC, sensitivity, or specificity to analyze whether statistically significant internal and external validation differences were present. No significant differences were found between both groups for all 4 categories (Table 1 and Fig 2).

### Performance of Molecular Subtype Prediction

If studies had multiple predictions with the same model, the highest AUC and accuracy are reported. In internal validation studies, *IDH* mutation status was the most frequently evaluated molecular subtype, with the overall highest mean AUC and accuracy values of 0.88 and 85%, respectively. *MGMT* promoter methylation was reported in 12 studies as the best predicted subtype, with a mean AUC and accuracy of 0.82 and 80%. The *1p/19q* codeletion was predicted in 9, with a mean AUC and accuracy of 0.84 and 85%. Prediction models for histone *H3 K27M* were identified in 6 studies with a mean AUC and accuracy of 0.80 and 81%. *ATRX* status was predicted in 1 study with an AUC and accuracy of 0.93 and 92%. *TERT* promoter mutation was predicted in 2 studies, resulting in a mean AUC and accuracy of 0.77 and 77%. Other subtypes included common subgroups like *EGFR*,<sup>27,28</sup> *p53*,<sup>22</sup> *RBI*<sup>29</sup> and *VEGF*,<sup>30</sup> as well as a pooled accumulation of genetic

information that was predicted as groups (Online Supplemental Data).<sup>31</sup> Buda et al<sup>31</sup> analyzed the prediction of pathologic biomarkers Ki-67, S-100, and glial fibrillary acidic protein,<sup>23</sup> which were also included in this category. Overall, these studies achieved a mean AUC and accuracy of 0.77 and 82% (Fig 3A).

Among studies that used external validation, 7 studies predicted *IDH* status with a mean AUC = 0.89 and accuracy = 86%; 3 studies predicted *MGMT* promoter methylation with a mean AUC = 0.89 and accuracy = 83%; 2 studies predicted *1p/19q* codeletion status with a mean AUC = 0.82 and accuracy = 75%; 2 studies predicted *ATRX* status with a mean AUC = 0.72 and accuracy = 77%; and 1 study predicted *EGFR* status with a mean AUC = 0.82 and accuracy = 85%. Other molecular subtypes, such as *PTEN* (accuracy = 82.5%)<sup>32</sup> and *BRAF* (AUC = 0.85)<sup>33</sup> mutation status, were each evaluated in 1 study. No study with external validation predicted histone *H3 K27M* status (Table 2 and Fig 3A).

### Algorithm-Based Prediction Models

Overall, of internally validated models, 35% ( $n = 27/77$ ) included tree-based; 27% ( $n = 21/77$ ), support vector machine (SVM); 32% ( $n = 25/77$ ), neural networks; and 5% ( $n = 4/77$ ), other classifiers. In externally validated models, 44% ( $n = 8/18$ ) were tree-based; 39% ( $n = 7/18$ ), SVM; 6% ( $n = 1/18$ ), neural networks; and 11% ( $n = 2/18$ ), other classifiers (Table 3 and Fig 4B).

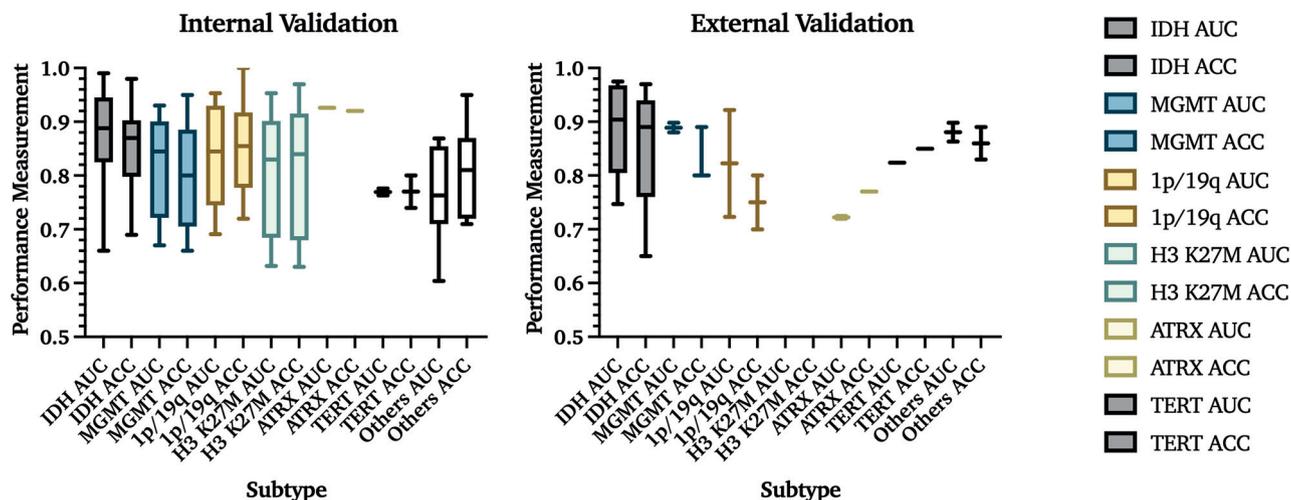
A neural network with a mean AUC and accuracy of 0.88 and 85% achieved the best overall prediction results. No statistically significant difference among all ML classifiers and neural networks was found with the Mann-Whitney *U* test. Tree-based algorithms and SVM performed slightly worse than neural networks with an AUC and accuracy of 0.82 and 82%, and 0.83 and 83%, respectively (Table 3).

In internal validation models, the overall mean AUC and accuracy for deep learning (DL) algorithms (0.82 and 82%) were higher than ML algorithms (0.88 and 86%) and were statistically significant by the Mann-Whitney *U* test ( $P = .02$ ) (Fig 4B).

These results were not observed in external validation pipelines (Table 3).

### Performance of Studies on the Prediction of *IDH* Mutation and *MGMT* Methylation Status

Mean AUC and accuracy for the prediction of *IDH* mutation status were 0.88 and 85% in internal validation studies and 0.85 and 86% in external validation studies. For the prediction of *MGMT* promoter methylation, AUC and accuracy in internal validation studies were 0.79 and 80% and 0.89 and 83% in external validation. While a *P* value of .02 indicated a statistically significant difference between AUC values for *IDH* and *MGMT* for internal validation cohorts, Mann-Whitney *U* tests did not show such a difference between accuracies, with a *P* value = .16 (Fig 5).



**FIG 3.** AUC and accuracy (ACC) results from internal and external validation studies. Results from 76 internal and 18 external validation studies are demonstrated on the basis of the molecular subtype that is being predicted. The *central line* in each result indicates the median value of the labeled subtype. Percentages are reported as fractions to provide visualization.

**Table 2: Types of algorithms used to predict the molecular subtypes of gliomas and the number of studies that used them**

	IDH	MGMT	1p/19q	H3 K27M	ATRX	TERT	Others
Internal validation							
Tree-based	11	4	4	4	0	0	4
SVM	9	1	2	2	1	2	4
Neural networks	14	6	3	0	0	0	2
Others	3	1	0	0	0	0	0
External validation							
Tree-based	4	1	0	0	1	1	1
SVM	2	1	2	0	1	0	1
Neural networks	1	0	0	0	0	0	0
Others	0	1	0	0	0	1	0

**Table 3: Performance of algorithms**

	Accuracy	AUC
Internal validation		
Tree-based	82% ( $n = 23/27$ )	0.82 ( $n = 23/27$ )
SVM	83% ( $n = 12/21$ )	0.83 ( $n = 17/21$ )
Neural network	85% ( $n = 22/25$ )	0.88 ( $n = 18/25$ )
Others	84% ( $n = 3/4$ )	0.85 ( $n = 2/4$ )
External validation		
Tree-based	85% ( $n = 6/8$ )	0.84 ( $n = 7/8$ )
SVM	82% ( $n = 7/7$ )	0.84 ( $n = 4/7$ )
Neural network	86% ( $n = 1/1$ )	0.86 ( $n = 1/1$ )
Others	89% ( $n = 1/2$ )	0.88 ( $n = 2/2$ )

### Risk of Bias Assessment

A risk of bias assessment was performed per PROBAST guidelines (Fig 6 and Online Supplemental Data).<sup>17</sup>

The risk of bias was low in the participant, predictor, and outcome sections of development and validation studies. The risk of bias was high in all of the analysis sections of the studies, regardless of whether these were development or validation studies. The main reasons for high bias in the analysis sections were due to items 4.1 and item 4.4.

Item 4.1 assesses the number of patients relative to imaging features extracted, suggesting an overfitting issue. It scored as “no” in 77.64% ( $n = 66/85$ ) of the studies. Additionally, item 4.4

suggesting a lack of reporting or incorrect handling of missing data, such as simple exclusion, was found in 98% ( $n = 83/85$ ) of studies.

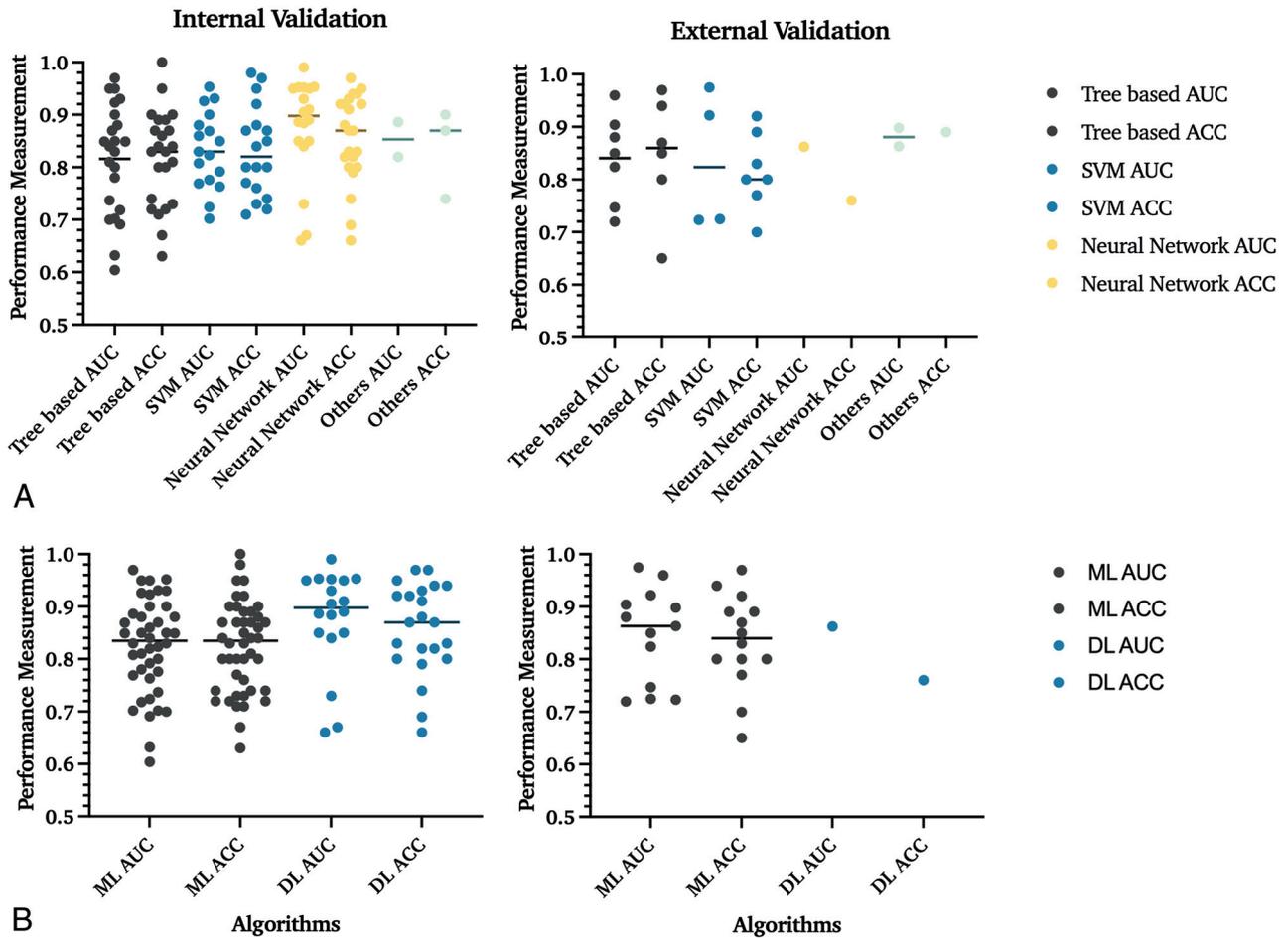
### DISCUSSION

Diagnosis and treatment of gliomas are based on pathologic and molecular classifications outlined in the 2021 WHO Classification of CNS tumors. Noninvasive methods for image-based prediction of glioma molecular subtype on preoperative images are the next

frontier in neuro-oncology because they will provide information before surgical intervention and have the potential to change treatment options for patients with brain tumors. Since 2017, published literature has significantly increased, showing high prediction results for this classification task. However, a thorough assessment of this literature to identify algorithms that can be used for clinical translation have yet to be performed.

Our systematic review shows that literature in this field has several limitations, which include low patient numbers (mean = 165.31), limited use of geographically distinct validation data sets (18.95%), and limited use of multicenter hospital data (14.71%). The most predicted molecular biomarkers were *IDH* and *MGMT*, which are critical for classifying glioblastoma from lower-grade gliomas and for predicting response to temozolomide therapy. Testing for these is standard of care in clinical practice. Therefore, these results are more available than other molecular biomarkers.

We show that DL algorithms result in significantly higher AUC values in internal validation studies than ML ( $P$  value = .02), leading us to recommend developing DL algorithms for future applications. Our review highlights the feasibility of accurately predicting glioma molecular subtypes; however, only some studies addressed the need for transparency and interpretability of these prediction models and demonstrated a high risk of bias.



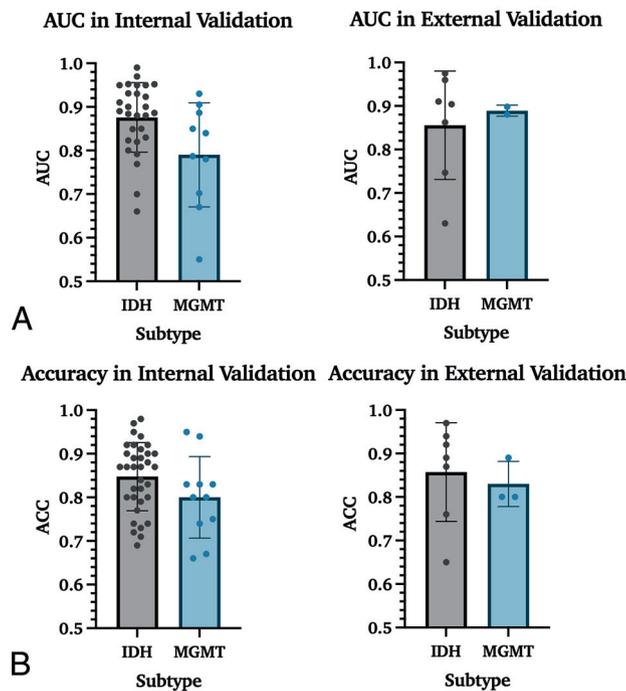
**FIG 4.** Comparison of performance of different ML algorithms in internal and external validation data sets. *A*, In internal validation studies, 35% ( $n = 27/77$ ) used tree-based; 27% ( $n = 21/77$ ), SVM; 32% ( $n = 25/77$ ), neural network; and 5% ( $n = 4/77$ ), other classifiers. The section named “Others” includes machine and deep learning algorithms, which cannot be classified into these 3 groups, and mixed classifiers with characteristics of multiple techniques. *Lines* indicate the mean value. *B*, Comparison of ML and DL algorithms. This figure refers to SVM and tree-based algorithms as overall ML algorithms. At the same time, all neural network–based classifiers are DL classifiers. In internal validation studies, 68% ( $n = 52/77$ ) used ML algorithms, and 32% ( $n = 25/77$ ) used DL classifiers. The DL algorithms demonstrated higher AUCs and statistically significant internal validation data sets. In the 95 patient cohorts analyzed, 68 studies used classic ML classifiers for their predictive models, while 26 used DL networks. The comparison of algorithms in external validation data sets was limited due to the small number of studies that validated DL algorithms. ACC indicates accuracy.

A unique aspect of our study is the broad inclusion criteria, which required screening of >12,000 abstracts before full-text review. Our study also contains the largest number of evaluated articles ( $n = 85$ ) with the most characteristic features extracted ( $n = 18$ ) (Online Supplementary Data).<sup>34</sup> In our analysis, we extracted not only feature characteristics of patient data sets but also imaging sequences, including advanced imaging modalities, algorithms, and types of ML algorithms with outcome assessment on internal and external validation data sets.

In addition to a thorough feature-extraction process from individual articles, we performed bias analysis with the most relevant assessment tool currently available, PROBAST.<sup>17</sup> Prior studies demonstrated that ML approaches for the evaluation of gliomas have significant deficiencies in reporting quality as assessed by TRIPOD,<sup>9-11</sup> aligning with our results of high bias. Unlike the previous systematic review, which had more restrictive inclusion criteria, our systematic review included 35 studies that did not report AUC ( $n = 21/95$ ) or accuracy ( $n = 14/95$ ) in evaluating their

respective models. Our less restrictive inclusion criteria, inclusion of studies that did not report both AUC and accuracy allowed us to better evaluate the distribution of different ML approaches in this field. Furthermore, our review did not exclude studies with non-MR imaging modalities or those with missing data, resulting in a total of 44 studies being included in our analysis.<sup>34</sup> While our findings generally agree with the results of prior work, we provide additional information on the differences in internal and external validation studies (conclusions), prediction performances of analyzed glioma molecular subtypes (section 4.4), and details of different image modalities with their corresponding performance metrics (Online Supplementary Data).

We recommend future studies evaluating ML algorithms in the imaging of gliomas to develop prediction algorithms based on larger data sets with geographically distinct data for model validation to provide generalizable results.<sup>35</sup> Until now, most studies relied on single-center hospital data and publicly available data sets. We recommend building databases of annotated images in



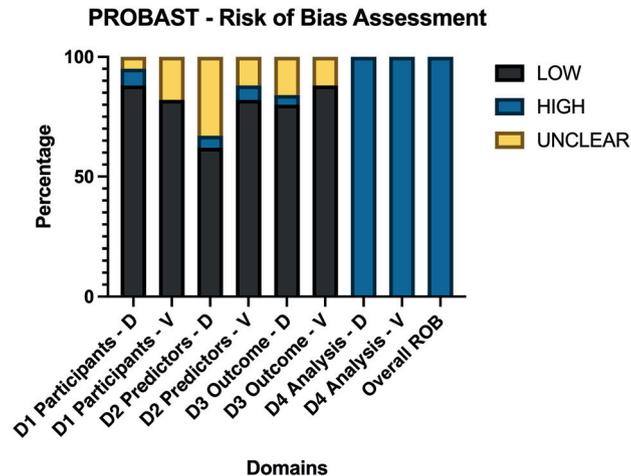
**FIG 5.** Performance measurements for the prediction of *IDH* and *MGMT* in internal and external data sets. **A.** AUC results for *IDH* mutation status and *MGMT* promoter methylation prediction. These data include all studies with *IDH* and/or *MGMT* prediction results, increasing the number of studies from 59 to 63. **B.** Accuracy results for *IDH* mutation and *MGMT* promoter methylation status.

individual hospitals and considering federated learning.<sup>36,37</sup> These changes will result in less reliance on publicly available data sets that are highly curated and will provide a diversity of hospital training data that will overcome the low data set patient number and allow translation of algorithms between different hospitals.<sup>38</sup> Additionally, we recommend that journal editors require validation of data set results, despite the expected drop in AUC and accuracy. This is well-represented in the recent RSNA-MICCAIS AI challenge, which demonstrated that the highest prediction for *MGMT* methylation based on an unseen external validation set was an AUC of 0.62.<sup>39-42</sup> Our literature review demonstrates 17 articles predicting *MGMT* methylation with AUC results ranging from 0.55 to 0.93 (mean, 0.81) in internal and external validation sets.

Some of the reasons for the inability to replicate these results could be low patient numbers in the reported studies, lack of description of model validation, high bias within the articles with the potential for overtraining, differences in segmentations among data sets, and differences in methods of *MGMT* methylation assessment between different hospitals. Future studies should consider that other pathologic testing protocols for molecular characterization at different hospitals and tumor heterogeneity with sampling bias can be a source of error because it is currently used as the criterion standard in ML algorithm development and, therefore, introduces reference bias in the respective findings.

### Limitations

A limitation of this review includes analysis of studies with incomplete data, which resulted in unequal data analysis for criteria such as the accuracy of results. Additionally, the number of



**FIG 6.** Average PROBAST scores of all 85 included articles. Domains were scored as high ROB if  $\geq 1$  item for each domain were scored as “No” or “Probably No.” PROBAST questions are assessed so that answering with no indicates ROB for this respective item. If 1 domain was considered a high ROB, the overall ROB of the study was considered high. As a result, each study was overall rated as having a high ROB. ROB indicates risk of bias; D, development studies; V, validation studies; D1, domain 1 participants; D2, domain 2 predictors; D3, domain 3 outcome; D4, domain 4 analysis.

studies with external validation was low, limiting the generalizability of the findings and raising an important point for editors and authors to consider the need to include external validation data sets in their publications. Furthermore, publication bias was not assessed because it was considered beyond the scope of this review. Finally, we compared the best prediction pipeline each study had to offer, which might introduce bias in our findings despite our efforts to minimize bias in our analyses.

### CONCLUSIONS

The results of prediction algorithms for molecular subtypes of gliomas in published studies demonstrate high AUC and accuracy. Still, there is an increased risk of bias based on the PROBAST assessment, which can result in poor data reproducibility. Improvement in reporting the quality of articles, development of large hospital-annotated data sets, and performance of external validation studies in future literature are critical for identifying algorithms that can be translated into clinical practice. This issue raises the need to develop novel tools for efficient data curation and annotation within the clinical workflow.

### ACKNOWLEDGMENTS

The American Society of Neuroradiology Fellow Award 2018 was presented to M.A. This publication was made possible by KL2 TR001862 (M.A.) from the National Center for Advancing Translational Science, components of the National Institutes of Health, and the National Institutes of Health Roadmap for Medical Research. A.A. is a PhD student in the Investigative Medicine Program at Yale, which is supported by Clinical and Translational Science Awards grant No. UL1 TR001863 from the National Center for Advancing Translational Science, a component of the National Institutes of Health. A.A. holds securities in

Hyperfine. M.L. is an employee and stockholder of Visage Imaging Inc and, unrelated to this work, receives funding from National Institutes of Health/National Cancer Institute R01 CA206180 and is a board member of Tau Beta Pi engineering honor society. Michael Sabel is a consultant for Novocure and Codman.

The content of this article is solely the authors' responsibility and does not necessarily represent the official view of National Institutes of Health.

Disclosure forms provided by the authors are available with the full text and PDF of this article at [www.ajnr.org](http://www.ajnr.org).

## REFERENCES

1. Norden AD, Wen PY. **Glioma therapy in adults.** *Neurologist* 2006;12:279–92 [CrossRef Medline](#)
2. Johns Hopkins Medicine. **Gliomas.** June 27, 2022. <https://www.hopkinsmedicine.org/health/conditions-and-diseases/gliomas>. Accessed June 27, 2022
3. Louis DN, Perry A, Wesseling P, et al. **The 2021 WHO Classification of Tumors of the Central Nervous System: a summary.** *Neuro Oncol* 2021;23:1231–51 [CrossRef Medline](#)
4. Parker NR, Khong P, Parkinson JF, et al. **Molecular heterogeneity in glioblastoma: potential clinical implications.** *Front Oncol* 2015;5:55 [CrossRef Medline](#)
5. Friedmann-Morvinski D. **Glioblastoma heterogeneity and cancer cell plasticity.** *Crit Rev Oncog* 2014;19:327–36 [CrossRef Medline](#)
6. Olar A, Aldape KD. **Using the molecular classification of glioblastoma to inform personalized treatment.** *J Pathol* 2014;232:165–77 [CrossRef Medline](#)
7. Tillmanns N, Lum AE, Cassinelli S, et al. **Identifying clinically applicable machine learning algorithms for glioma segmentation: recent advances and discoveries.** *Neurooncol Adv* 2022;4:vdac093 [CrossRef Medline](#)
8. 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.** 2019. [arxiv.org/abs/1811.02629](https://arxiv.org/abs/1811.02629). Accessed June 27, 2022
9. Jekel L, Brim WR, von Reppert M, et al. **Machine learning applications for differentiation of glioma from brain metastasis: a systematic review.** *Cancers(Basel)* 2022;14:1369 [CrossRef Medline](#)
10. Cassinelli Petersen GI, Shatalov J, Verma T, et al. **Machine learning in differentiating gliomas from primary CNS lymphomas: a systematic review, reporting quality, and risk of bias assessment.** *AJNR Am J Neuroradiol* 2022;43:526–33 [CrossRef Medline](#)
11. Bahar RC, Merkaş 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 [CrossRef Medline](#)
12. Sanghani P, Ang BT, King NK, et al. **Overall survival prediction in glioblastoma multiforme patients from volumetric, shape and texture features using machine learning.** *Surg Oncol* 2018;27:709–14 [CrossRef Medline](#)
13. Bakas S, Shukla G, Akbari H, et al. **Overall survival prediction in glioblastoma patients using structural magnetic resonance imaging (MRI): advanced radiomic features may compensate for lack of advanced MRI modalities.** *J Med Imaging (Bellingham)* 2020;7:031505 [CrossRef Medline](#)
14. Clark K, Vendt B, Smith K, et al. **The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository.** *J Digit Imaging* 2013;26:1045–57 [CrossRef Medline](#)
15. Heus P, Damen J, Pajouheshnia R, et al. **Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies.** *BMJ Open* 2019;9:e025611 [CrossRef Medline](#)
16. 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:e200029 [CrossRef Medline](#)
17. Moons KG, Wolff RF, Riley RD, et al. **PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration.** *Ann Intern Med* 2019;170:W1–33 [CrossRef Medline](#)
18. Li YM, Suki D, Hess K, et al. **The influence of maximum safe resection of glioblastoma on survival in 1229 patients: can we do better than gross-total resection?** *J Neurosurg* 2016;124:977–88 [CrossRef Medline](#)
19. Hegi ME, Diserens AC, Gorlia T, et al. **MGMT gene silencing and benefit from temozolomide in glioblastoma.** *N Engl J Med* 2005;352:997–1003 [CrossRef Medline](#)
20. 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;9:20311 [CrossRef Medline](#)
21. Ren Y, Zhang X, Rui W, et al. **Noninvasive prediction of IDH1 mutation and ATRX expression loss in low-grade gliomas using multiparametric MR radiomic features.** *J Magn Reson Imaging* 2019;49:808–17 [CrossRef Medline](#)
22. Zhang X, Tian Q, Wang L, et al. **Radiomics strategy for molecular subtype stratification of lower-grade glioma: detecting IDH and TP53 mutations based on multimodal MRI.** *J Magn Reson Imaging* 2018;48:916–26 [CrossRef Medline](#)
23. Gao M, Huang S, Pan X, et al. **Machine learning-based radiomics predicting tumor grades and expression of multiple pathologic biomarkers in gliomas.** *Front Oncol* 2020;10:1676 [CrossRef Medline](#)
24. Lu CF, Hsu FT, Hsieh KL, et al. **Machine learning-based radiomics for molecular subtyping of gliomas.** *Clin Cancer Res* 2018;24:4429–36 [CrossRef Medline](#)
25. Chen X, Zeng M, Tong Y, et al. **Automatic prediction of MGMT status in glioblastoma via deep learning-based MR image analysis.** *Biomed Res Int* 2020;2020:9258649 [CrossRef Medline](#)
26. Li Y, Wei D, Liu X, et al. **Molecular subtyping of diffuse gliomas using magnetic resonance imaging: comparison and correlation between radiomics and deep learning.** *Eur Radiol* 2022;32:747–58 [CrossRef Medline](#)
27. Hedyehzadeh M, Maghooli K, MomenGharibvand M, et al. **A comparison of the efficiency of using a deep CNN approach with other common regression methods for the prediction of EGFR expression in glioblastoma patients.** *J Digit Imaging* 2020;33:391–98 [CrossRef Medline](#)
28. Akbari H, Bakas S, Pisapia JM, et al. **In vivo evaluation of EGFRvIII mutation in primary glioblastoma patients via complex multiparametric MRI signature.** *Neuro Oncol* 2018;20:1068–79 [CrossRef Medline](#)
29. Park JE, Kim HS, Park SY, et al. **Prediction of core signaling pathway by using diffusion- and perfusion-based MRI radiomics and next-generation sequencing in isocitrate dehydrogenase wild-type glioblastoma.** *Radiology* 2020;294:388–97 [CrossRef Medline](#)
30. Sun Z, Li Y, Wang Y, et al. **Radiogenomic analysis of vascular endothelial growth factor in patients with diffuse gliomas.** *Cancer Imaging* 2019;19:68 [CrossRef Medline](#)
31. Buda M, AlBadawy EA, Saha A, et al. **Deep radiogenomics of lower-grade gliomas: convolutional neural networks predict tumor genomic subtypes using MR images.** *Radiol Artif Intell* 2020;2:e180050 [CrossRef Medline](#)
32. Li Y, Liang Y, Sun Z, et al. **Radiogenomic analysis of PTEN mutation in glioblastoma using preoperative multi-parametric magnetic resonance imaging.** *Neuroradiology* 2019;61:1229–37 [CrossRef Medline](#)
33. Wagner MW, Hainc N, Khalvati F, et al. **Radiomics of pediatric low-grade gliomas: toward a pretherapeutic differentiation of BRAF-mutated and BRAF-fused tumors.** *AJNR Am J Neuroradiol* 2021;42:759–65 [CrossRef Medline](#)
34. Jian A, Jang K, Manuguerra M, et al. **Machine learning for the prediction of molecular markers in glioma on magnetic resonance**

- imaging: a systematic review and meta-analysis. *Neurosurgery* 2021;89:31–44 [CrossRef Medline](#)
35. Cabitza F, Campagner A, Soares F, et al. **The importance of being external. Methodological insights for the external validation of machine learning models in medicine.** *Comput Methods Programs Biomed* 2021;208:106288 [CrossRef Medline](#)
36. Rieke N, Hancox J, Li W, et al. **The future of digital health with federated learning.** *NPJ Digit Med* 2020;3:119 [CrossRef Medline](#)
37. Sheller MJ, Edwards B, Reina GA, et al. **Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data.** *Sci Rep* 2020;10:12598 [CrossRef Medline](#)
38. Pfitzner B, Steckhan N, Arnrich B. **Federated learning in a medical context: a systematic literature review.** *ACM Trans Internet Technol* 2021;21:1–31 [CrossRef](#)
39. Baid U, Ghodasara S, Mohan S, et al. **The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification.** *Computer Vision and Pattern Recognition*. 2021. <https://arxiv.org/abs/2107.02314>. Accessed June 27, 2022
40. Menze BH, Jakab A, Bauer S, et al. **The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS).** *IEEE Trans Med Imaging* 2015;34:1993–2024 [CrossRef Medline](#)
41. 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 [CrossRef Medline](#)
42. **RSNA-MICCAI Brain Tumor Radiogenomic Classification.** July 13, 2021 to October 15, 2021. <https://www.kaggle.com/rsna-miccai-brain-tumor-radiogenomic-classification>. Accessed June 27, 2022