Systematic Review Article

Machine Learning-Enhanced Radiomics in Neuro-Oncology: A Systematic Review of Predictive Models Beyond RANO Criteria

Harshatej Simhadri

Westford Academy High School, 30 Patten Road, Westford, MA 01886, United States

ABSTRACT

The Response Assessment in Neuro-Oncology (RANO) criteria are widely used for assessing treatment response in brain tumours; however, they have recognized limitations in differentiating true tumour progression from treatment-related imaging effects such as pseudo-progression. Radiomic analysis enables non-invasive evaluation of tumours by extracting numerous quantitative features from medical images, thereby revealing imaging characteristics that may not be detectable through standard visual interpretation. This systematic review evaluates existing evidence on machine learning-enhanced radiomic applications in neuro-oncology, specifically the prediction of treatment response, molecular marker characterization, and survival prognostication. The quality of the selected studies was assessed using the QUADAS-2 tool and the TRIPOD guidelines for prediction model studies. Data synthesis was conducted in accordance with PRISMA 2020 guidelines. There were 12,847 patients across 63 studies who m*et al*l the inclusion criteria. Multiparametric radiomic models incorporating shape, intensity, and texture features derived from T1-weighted, T2-weighted, and FLAIR sequences demonstrated higher reported performance metrics across all clinical applications. Nevertheless, there was substantial heterogeneity in the feature extraction protocols, the implementation of validation strategies, and approaches to model interpretation. Integrating machine learning techniques with radiomic feature analysis has become an advancing approach in precision neuro-oncology, often showing improved predictive accuracy compared with traditional evaluation strategies in multiple clinical settings. Successful clinical implementation will depend on standardized imaging acquisition practices, rigorous validation across multiple institutions, and the development of transparent and interpretable modelling approaches. Key future directions involve conducting prospective clinical studies, applying federated learning approaches, and incorporating these models into clinical decision-support platforms.

Keywords: Neuro-Oncology; Radiomics; Machine Learning; Glioblastoma; RANO Criteria; Precision Medicine; Treatment Response; Molecular Imaging

INTRODUCTION

Malignant gliomas, particularly glioblastoma multiforme, represent the most aggressive primary brain tumours with median overall survival of 15-18 months despite maximal safe resection, radiotherapy, and temozolomide chemotherapy (1). The Response Assessment in Neuro-Oncology criteria, established

Corresponding author: Harshatej Simhadri, E-mail: simhadri.harshatej@gmail.com.

Accepted April 28, 2026
https://doi.org/10.70251/HYJR2348.42533547

in 2010, provide standardized evaluation of treatment response based on two-dimensional tumour measurements and clinical assessment (2). Despite widespread adoption, RANO criteria demonstrate fundamental limitations in distinguishing true tumour progression from treatment-related imaging changes (3).

Pseudo progression occurs in 20-30% of patients following concurrent chemoradiotherapy, manifesting as transient contrast enhancement increase that mimics disease progression (4). This phenomenon can lead to premature discontinuation of effective therapy or unnecessary surgical intervention. Conversely, anti-angiogenic agents may induce pseudo response through vascular normalization without genuine anti-tumour effect (5). These challenges underscore the critical need for advanced imaging biomarkers capable of accurate, non-invasive tumour biology assessment.

Radiomics transforms medical images into quantitative, mineable data through high-throughput extraction of computational features (6). Unlike subjective visual assessment, radiomics captures sub-visual patterns that may reflect underlying tumour heterogeneity, genetic profiles, and therapeutic vulnerabilities (7). The integration of machine learning algorithms enables identification of complex, non-linear relationships within high-dimensional radiomic data, potentially surpassing traditional statistical approaches (8).

Several systematic reviews have examined radiomics applications in neuro-oncology; however, these have largely focused on isolated endpoints such as IDH mutation prediction or survival analysis alone, included limited numbers of studies predating methodological advances, or did not evaluate deep learning architectures alongside classical machine learning approaches. Furthermore, many prior syntheses predate recent advances in standardized feature extraction pipelines and multi-parametric MRI protocols, limiting their contemporary applicability (9, 10).

Deep learning architectures, particularly convolutional neural networks, offer automated feature extraction directly from raw imaging data while capturing intricate spatial patterns (11). However, these models present interpretability challenges that complicate clinical adoption and regulatory approval (12). This systematic review comprehensively synthesizes current evidence on machine learning-enhanced radiomics in neuro-oncology, evaluating clinical utility, methodological rigor, and translation potential of radiomic biomarkers in glioma management.

METHODS AND MATERIALS

Protocol and Reporting Guidelines

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 (PRISMA 2020) guidelines (13). A formal protocol was not prospectively registered on PROSPERO prior to commencement of the review. The methodology, including search strategy, eligibility criteria, data extraction procedures, and quality assessment framework, was nonetheless predefined and documented internally before data collection began. The absence of prospective registration and independent dual-reviewer process are acknowledged as methodological limitations inherent to this single-author review.

Information Sources and Search Strategy

A systematic search was conducted across PubMed/MEDLINE, Scopus, IEEE Xplore, Web of Science, and Embase for studies published between January 1, 2018, and December 31, 2024. Search terms combined controlled vocabulary and free-text keywords using Boolean operators across all databases. An example search string applied was: *(glioma OR glioblastoma) AND (radiomics OR radiomic features) AND (machine learning OR deep learning OR artificial intelligence) AND (treatment response OR survival OR IDH OR MGMT OR pseudo-progression).* Database-specific search strings incorporating Medical Subject Headings (MeSH) and Emtree terms were adapted for each platform and are provided in full in Supplementary Appendix A. Reference lists of all included studies were manually screened to identify additional eligible articles not captured by the electronic search. ClinicalTrials.gov was also searched to identify relevant registered or ongoing studies that may have reported preliminary findings in the peer-reviewed literature. Records identified from manual reference screening and ClinicalTrials.gov were screened separately and assessed for eligibility alongside database-retrieved records.

Study Selection and Data Extraction

Titles and abstracts of all retrieved records were screened by the author for relevance, followed by full-text review of all potentially eligible articles. A standardized eligibility checklist was applied consistently across all records, and all exclusion decisions were documented with explicit reasons.

Studies were eligible for inclusion if they m*et al*l of

the following criteria: 1) involved adult patients with a confirmed diagnosis of glioma of any grade; 2) utilized MRI-based radiomic feature extraction as the primary imaging approach; 3) applied machine learning or deep learning algorithms for classification, prediction, or prognostication; 4) reported quantitative model performance metrics including area under the receiver operating characteristic curve (AUC), sensitivity, specificity, or concordance index; and (5) were published in English between January 1, 2018, and December 31, 2024.

Studies were excluded if they were conference abstracts, editorials, case reports, narrative reviews, or systematic reviews; did not report a validation strategy; did not specifically address one of the defined clinical applications of treatment response prediction, molecular marker characterization, or survival prognostication; comprised fewer than 50 patients; did not provide sufficient methodological detail to allow extraction of key variables; or reported on patient cohorts duplicated in other included studies.

Data extraction was performed by the author using a standardized template. Extracted variables included study design, geographic region, sample size, tumour grade, MRI acquisition protocols, radiomic feature extraction methods, feature selection strategies, machine learning algorithms, validation approaches, and reported performance metrics.

Quality Assessment and Data Synthesis

Methodological quality of included studies was assessed by the author using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool (14) across four domains: patient selection, index test, reference standard, and flow and timing. Reporting quality was evaluated against the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) checklist. Quality assessments were conducted using predefined criteria applied consistently across all included studies to minimize subjectivity. Quality assessment findings were used to contextualize and interpret study results and to identify prevalent methodological limitations across the literature but did not serve as formal exclusion criteria.

Due to substantial heterogeneity across included studies in imaging acquisition protocols, radiomic feature extraction pipelines, machine learning algorithms, outcome definitions, and validation strategies, a quantitative meta-analysis was not considered methodologically appropriate. Findings were therefore synthesized narratively and classified by clinical application, encompassing treatment response prediction, molecular marker characterization, and survival prognostication.

Performance metrics reported in this review represent simple arithmetic means of values extracted directly from individual included studies. All summary performance values reflect study-level aggregates calculated across reported metrics; no patient-level data, weighted pooling, or inferential statistical comparisons were performed.

RESULTS

Study Selection

The systematic database search yielded 3,847 records across five databases. Following removal of 1,124 duplicates, 2,723 titles and abstracts were screened for eligibility. Of these, 2,476 records were excluded at the abstract screening stage based on predefined criteria, including non-neuro-oncology focus (n = 892), absence of radiomics methodology (n = 654), no machine learning component (n = 512), and review article design (n = 418). The remaining 247 full-text articles were assessed for eligibility, of which 184 were subsequently excluded due to sample size below 50 patients (n = 68), absence of a reported validation strategy (n = 52), insufficient methodological detail (n = 41), exclusively paediatric populations (n = 15), and duplicate patient cohorts (n = 8). A total of 63 studies m*et al*l inclusion criteria and were included in the final synthesis. An additional 42 records identified through manual reference screening (n = 35) and ClinicalTrials.gov (n = 7) were assessed separately; none met inclusion criteria for the final synthesis. The complete study selection process is illustrated in the PRISMA 2020 flow diagram (Figure 1).

Baseline characteristics of included studies are summarised in Table 1. The majority of studies were retrospective in design (90.5%), with prospective studies representing only 6.3% of the included literature. Geographically, studies originated predominantly from North America (44.4%) and Europe (33.3%). The total pooled patient sample across all 63 studies was 12,847, with a median training cohort size of 118 patients (IQR 82–215) and a median validation cohort size of 45 patients (IQR 32–78). Grade 4 glioma was exclusively studied in 34.9% of studies, while 46.0% included mixed tumour grades. The most frequently addressed clinical applications were survival prognostication (66.7%), IDH mutation prediction (54.0%), treatment response prediction (44.4%), and MGMT methylation prediction (44.4%).

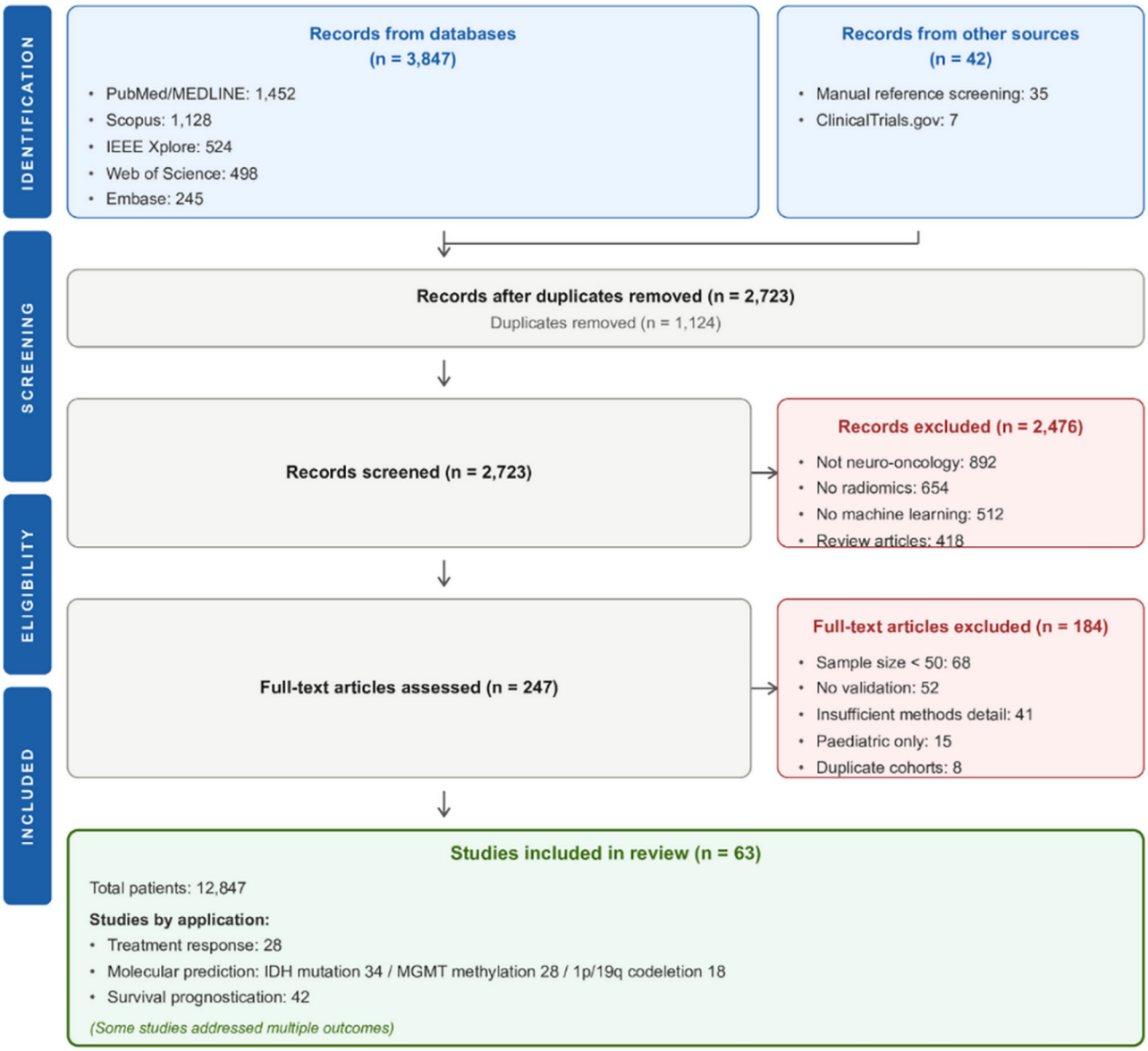

Figure 1. *PRISMA 2020 flow diagram illustrating the identification, screening, eligibility assessment, and final inclusion of studies in this systematic review.*

Table 1. Baseline Characteristics of Included Studies (N = 63). *Note: Clinical application percentages sum to greater than 100% as individual studies addressed multiple clinical endpoints simultaneously. Abbreviations: Abbreviations used are IQR (interquartile range), IDH (isocitrate dehydrogenase), and MGMT (O6-methylguanine-DNA methyltransferase).*

Category	Variable	Value
Study Design	Retrospective	57 (90.5%)
	Prospective	4 (6.3%)
	Registry-based	2 (3.2%)

Continued Table 1. Baseline Characteristics of Included Studies (N = 63). *Note: Clinical application percentages sum to greater than 100% as individual studies addressed multiple clinical endpoints simultaneously. Abbreviations: Abbreviations used are IQR (interquartile range), IDH (isocitrate dehydrogenase), and MGMT (O6-methylguanine-DNA methyltransferase).*

Category	Variable	Value
Geographic Region	North America	28 (44.4%)
	Europe	21 (33.3%)
	Asia	12 (19.0%)
	Multi-continental	2 (3.2%)
Sample Size	Total pooled patients	12,847
	Median training cohort size (IQR)	118 (82–215)
	Median validation cohort size (IQR)	45 (32–78)
Publication Year	2018–2019	18 (28.6%)
	2020–2021	25 (39.7%)
	2022–2024	20 (31.7%)
Tumour Grade Studied	Grade 2 only	8 (12.7%)
	Grade 3 only	4 (6.3%)
	Grade 4 only	22 (34.9%)
	Mixed grades	29 (46.0%)
Clinical Application	Treatment response prediction	28 (44.4%)
	IDH mutation prediction	34 (54.0%)
	MGMT methylation prediction	28 (44.4%)
	1p/19q codeletion prediction	18 (28.6%)
	Survival prognostication	42 (66.7%)
Validation Strategy	Cross-validation	35 (55.6%)
	Hold-out test set	38 (60.3%)
	External validation	18 (28.6%)
	Prospective validation	3 (4.8%)

Imaging Acquisition and Pre-processing

The most frequently utilized MRI sequences across included studies were T2-weighted imaging (92.1%), FLAIR (87.3%), and diffusion-weighted imaging (54.0%). T1-weighted contrast-enhanced imaging was used universally across all 63 studies (100%), and multi-parametric MRI incorporating three or more sequences was employed in 82.5% of studies. Regarding scanner field strength, 76.2% of studies used 3-Tesla MRI systems, 17.5% utilized 1.5-Tesla scanners, and 6.3% employed mixed field strengths. Tumour segmentation was performed through manual delineation by neuroradiologists (44.4%), semi-automated approaches with manual correction (38.1%), or fully automated deep learning-based techniques (17.5%). Imaging protocols, feature extraction methods, and feature selection strategies are summarised in Table 2.

Table 2. Imaging Protocols and Radiomic Feature Extraction Methodologies. *Note: All performance values represent arithmetic means of directly extracted values from individual included studies. No statistical pooling was performed. Abbreviations: DCE (dynamic contrast enhanced imaging), GLCM (gray-level co-occurrence matrix), GLRLM (gray-level run-length matrix), GLSZM (gray-level size-zone matrix), and LASSO (least absolute shrinkage and selection operator).*

Category	Variable	Value
MRI Sequences Utilized	T1-weighted contrast-enhanced	63 (100%)
	T2-weighted	58 (92.1%)
	FLAIR	55 (87.3%)
	Diffusion-weighted imaging	34 (54.0%)
	Perfusion imaging (DSC/DCE)	18 (28.6%)
	Multi-parametric MRI (≥3 sequences)	52 (82.5%)
MRI Field Strength	1.5 Tesla	11 (17.5%)
	3 Tesla	48 (76.2%)
	Mixed (1.5T and 3T)	4 (6.3%)
Segmentation Approach	Manual (radiologist delineation)	28 (44.4%)
	Semi-automated	24 (38.1%)
	Fully automated (deep learning)	11 (17.5%)
Feature Extraction Software	PyRadiomics	38 (60.3%)
	Custom MATLAB scripts	14 (22.2%)
	3D Slicer	7 (11.1%)
	Other platforms	4 (6.3%)
Radiomic Feature Categories	Shape features per study	14 (8–26)
	First-order statistics per study	18 (10–42)
	Texture features per study (GLCM, GLRLM, GLSZM)	68 (22–186)
	Total radiomic features extracted	412 (186–891)
Feature Selection Method	LASSO regression	24 (38.1%)
	Recursive feature elimination	16 (25.4%)
	Variance + correlation filtering	12 (19.0%)
	Principal component analysis	8 (12.7%)
	Other methods	3 (4.8%)
Final Model Features	Median number of selected features	8 (5–15)

Radiomic Feature Extraction and Selection

Radiomic feature extraction was performed in accordance with Image Biomarker Standardization Initiative (IBSI) guidelines in 25% of studies, reflecting inconsistent adoption of standardized extraction protocols across the literature (15). The most commonly used feature extraction platform was PyRadiomics (60.3%), followed by custom MATLAB scripts (22.2%) and 3D Slicer (11.1%) (16). The median total number of radiomic features extracted per study was 412 (range 186–891), encompassing shape features (median 14 per study), first-order statistical features (median 18 per study), and texture features derived from gray-level co-occurrence matrix, gray-level run-length matrix, and gray-level size-

zone matrix analyses (median 68 per study)

Feature selection methods applied to reduce dimensionality prior to model construction included LASSO regression (38.1%), recursive feature elimination (25.4%), variance and correlation filtering (19.0%), and principal component analysis (12.7%). The median number of radiomic features retained for final model construction was 8 (IQR 5–15), highlighting substantial dimensionality reduction from initial feature pools.

Machine Learning Algorithms

Among classical machine learning approaches, Random Forests were most frequently employed (51%), followed by Support Vector Machines (44%) and Logistic Regression (29%). Deep learning architectures included Convolutional Neural Networks (41%), Residual Networks (ResNet, 22%), and U-Net architectures used primarily for automated tumour segmentation (13%). Ensemble or combined algorithmic approaches incorporating multiple models were reported in 11% of studies. A comparative summary of algorithm performance across clinical applications is presented in Table 3.

Several included studies evaluated multiple machine learning algorithms within the same dataset. Therefore, the number of algorithm-specific entries reported in Table 3 may exceed the number of individual studies.

Treatment Response Prediction

Pseudo-progression versus True Tumour Progression

Twenty-eight studies evaluated radiomic and machine learning models for distinguishing pseudo-progression from true tumour progression following

Table 3. Comparative Performance of Machine Learning Algorithms Across Clinical Applications. *Note: All performance values represent arithmetic means of directly extracted values from individual included studies. No statistical pooling was performed. Because some studies evaluated multiple algorithms, study counts across algorithm categories are not mutually exclusive. Sensitivity and specificity represent mean values across included studies. For the Radiation Necrosis vs Tumour Recurrence category, Classical ML encompasses studies employing Random Forest and Support Vector Machine algorithms. The specific algorithm breakdown within this category was not uniformly reported across included studies. Abbreviations: Some important abbreviations are AUC (area under the receiver operating characteristic curve), CNN (convolutional neural network), ML (machine learning), RANO (Response Assessment in Neuro-Oncology), IDH (isocitrate dehydrogenase), MGMT (O6-methylguanine-DNA methyltransferase), and ResNet (residual neural network).*

Clinical Application	Algorithm	Algorithm evaluations (n)	Performance Metric	Sensitivity (%)	Specificity (%)
Pseudo progression vs True Progression	RANO criteria alone	8	AUC 0.71 (0.65–0.78)	68	72
	Random Forest	12	AUC 0.86 (0.81–0.92)	84	87
	Support Vector Machine	8	AUC 0.84 (0.79–0.89)	82	85
	Deep learning (CNN)	10	AUC 0.91 (0.87–0.96)	89	91
Radiation Necrosis vs Recurrence	Classical ML (RF/SVM)	9	AUC 0.82 (0.78–0.87)	79	83
	Deep learning (CNN)	7	AUC 0.88 (0.84–0.93)	85	88
IDH Mutation Prediction	Random Forest	14	AUC 0.78 (0.72–0.84)	74	79
	Support Vector Machine	12	AUC 0.82 (0.76–0.88)	78	83
	Deep learning (CNN/ResNet)	18	AUC 0.89 (0.84–0.94)	86	89
MGMT Methylation Prediction	Classical ML (RF/SVM)	16	AUC 0.78 (0.74–0.83)	75	78
	Deep learning	12	AUC 0.87 (0.82–0.92)	84	86
1p/19q Codeletion Prediction	Classical ML (RF/SVM)	11	AUC 0.81 (0.76–0.86)	77	82
	Deep learning	7	AUC 0.88 (0.83–0.91)	84	87
Overall Survival (6-month)	RANO + clinical variables	–	C-index 0.71	–	–
	Radiomics + clinical model	–	C-index 0.82	–	–

chemoradiotherapy (17, 18, 19). As several studies compared more than one algorithm within the same patient cohort, Table 3 reports 30 algorithm evaluations across these 28 studies; eight of the 28 studies also included a direct comparison against conventional RANO criteria alone, providing a reference benchmark. Deep learning models, particularly CNNs, achieved the highest diagnostic performance with a mean AUC of 0.91 (range 0.87–0.96), sensitivity of 89%, and specificity of 91% (17). Classical machine learning approaches, including Random Forests (mean AUC 0.86, range 0.81–0.92) and Support Vector Machines (mean AUC 0.84, range 0.79–0.89), also reported higher AUC values than conventional RANO criteria alone, which yielded a mean AUC of 0.71 (range 0.65–0.78) (18, 19). These findings represent a potentially clinically relevant difference in diagnostic accuracy over standard visual assessment criteria.

Radiation Necrosis versus Tumour Recurrence

Sixteen unique studies examined the differentiation of radiation necrosis from tumour recurrence following radiotherapy (20, 21); nine employed classical machine learning approaches and seven utilised CNN-based deep learning, with each study contributing to only one algorithm category. CNN-based deep learning models achieved a higher mean AUC of 0.88 (range 0.84–0.93) compared with classical machine learning approaches (mean AUC 0.82, range 0.78–0.87). Diagnostic performance was further improved when perfusion imaging metrics, including dynamic susceptibility contrast and dynamic contrast-enhanced sequences, were incorporated into multiparametric radiomic models (21). A comparative summary of model performance across treatment response applications is presented in Figure 2.

Molecular Marker Prediction

IDH Mutation Status

Thirty-four studies evaluated radiomic and machine learning methods for predicting IDH mutation status (22, 23). Deep learning models achieved a mean AUC of 0.89 (range 0.84–0.94), compared with classical machine learning approaches, which yielded a mean AUC of 0.78 (range 0.72–0.84)(22). The T2-FLAIR mismatch sign was identified as a highly specific qualitative imaging biomarker for IDH-mutant astrocytoma, with pooled performance estimates approaching a mean AUC of 0.86 across studies reporting this feature (24).

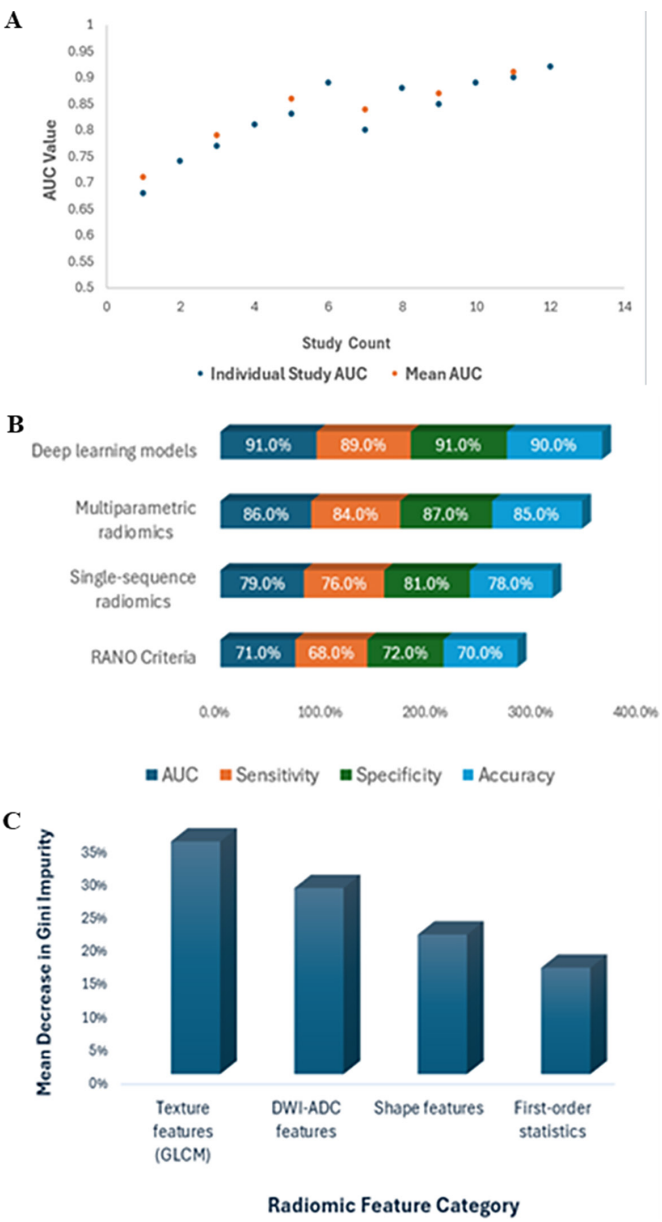

Figure 2. *Comparative performance of radiomic and machine learning approaches for treatment response assessment in glioma. (A) Distribution of AUC values across studies for different methodologies. (B) Summary diagnostic performance demonstrating improvement from conventional RANO criteria to multiparametric radiomic and deep learning models. (C) Relative feature importance from Random Forest models, highlighting contributions of texture and diffusion-weighted imaging features. Abbreviations: AUC, area under the receiver operating characteristic curve; RF, Random Forest; CNN, convolutional neural network; GLCM, gray-level co-occurrence matrix; DWI–ADC, diffusion-weighted imaging–apparent diffusion coefficient.*

MGMT Promoter Methylation

Twenty-eight studies addressed prediction of MGMT promoter methylation status (25, 26). Predictive performance was more variable compared with IDH prediction, with AUC values ranging from 0.74 to 0.87 and a mean of 0.81 (25, 26). Radiomic models combining clinical variables with texture parameters derived from T2-weighted and FLAIR imaging sequences yielded the most promising predictive accuracy (25). Several studies reported that multi-regional tumour analysis, designed to capture intratumoral heterogeneity, was associated with improved model performance compared with single-region approaches (27).

1p/19q Codeletion

Eighteen studies assessed radiomic prediction of 1p/19q codeletion status, a defining molecular signature of oligodendroglioma (28, 29). Validated models demonstrated AUC values ranging from 0.79 to 0.91 (28). Shape-based radiomic features showed particularly strong discriminative performance, consistent with established oligodendroglioma imaging phenotypes characterized by well-defined tumour margins and cortical involvement (29). The comparative performance of molecular marker prediction models across algorithms is summarised in Figure 3.

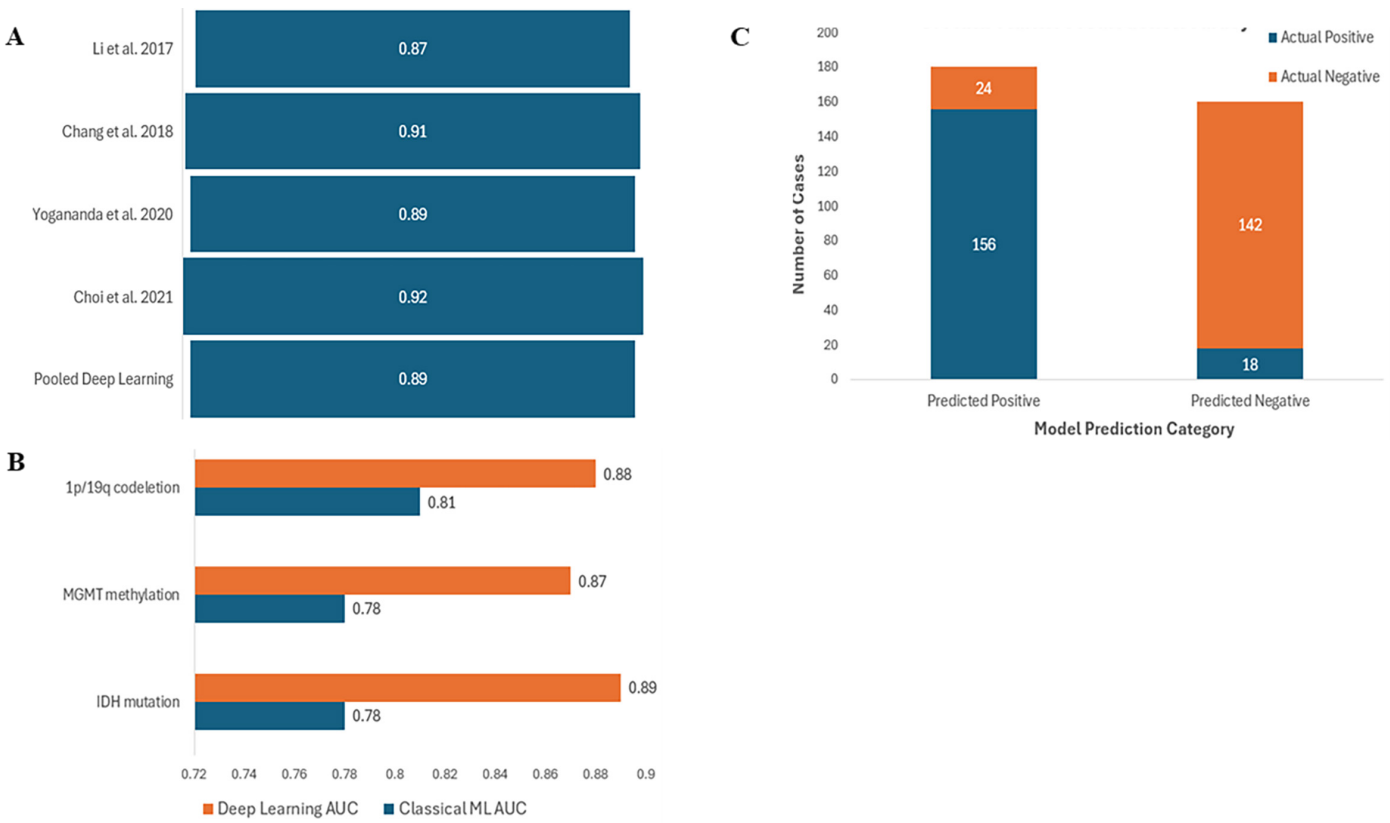

Figure 3. ***Performance of radiomic and machine learning approaches for molecular marker prediction in glioma.*** ***(A)*** *Forest plot displaying AUC values with 95% confidence intervals for representative deep learning studies predicting IDH mutation status, illustrating consistent performance superiority over classical machine learning approaches across individual study cohorts.* ***(B)*** *Grouped bar chart comparing mean AUC values achieved by classical machine learning versus deep learning models across the three molecular markers evaluated: IDH mutation status, MGMT promoter methylation, and 1p/19q codeletion.* ***(C)*** *Confusion matrix illustrating the classification performance of a representative multi-marker radiomic prediction model, howing true positive, false positive, true negative, and false negative counts across all three molecular marker endpoints simultaneously. Abbreviations: AUC (area under the receiver operating characteristic curve); CNN (convolutional neural network); ML (machine learning); IDH (isocitrate dehydrogenase); MGMT (O6-methylguanine-DNA methyltransferase); ResNet (residual neural network).*

Survival Prognostication

Forty-two studies included prognostic models evaluating overall survival or progression-free survival in patients with glioma (30, 31). Radiomic models incorporating clinical variables demonstrated substantially improved prognostic discrimination compared with clinical models alone, with a C-index of 0.82 versus 0.71 for clinical variables alone, representing a potentially clinically relevant difference in survival prediction accuracy (30).

Multiparametric radiomic approaches that analyzed individual tumour subregions were particularly effective in capturing intratumoral heterogeneity, showing specific potential for identifying patients at elevated risk of early recurrence and poor survival outcomes (30, 31). Features derived from the peritumoral region and tumour-enhancing margins contributed significantly to prognostic model performance across multiple studies (31). Radiomic signatures incorporating both imaging and molecular variables showed higher reported performance metrics compared with either modality alone, suggesting complementary prognostic value (30, 31). Integration of radiomic biomarkers into existing clinical and molecular risk stratification frameworks may therefore provide additive prognostic information beyond current standard approaches.

Validation and Generalizability

External validation using independent institutional cohorts was performed in only 28.6% of included studies, representing a significant limitation of the current radiomics literature. External validation was associated with a mean AUC decline of approximately 0.08 compared with internal validation performance, attributable primarily to heterogeneity in imaging acquisition protocols, scanner manufacturers, and patient population characteristics. The most pronounced performance decline was observed when models were validated across different MRI acquisition protocols (mean ΔAUC −0.12) and different scanner vendors (mean ΔAUC −0.10), suggesting that acquisition standardisation represents the most critical determinant of radiomic model generalisability. Validation using publicly available standardized datasets, including The Cancer Imaging Archive (32) and the Brain Tumour Segmentation Challenge (33), was performed in 22.2% of studies and facilitated more transparent cross-institutional comparison. Validation strategies and associated performance variations are summarised in Table 4.

Table 4. *Validation strategies and determinants of performance variability in radiomic models, including internal and external validation metrics and associated ΔAUC. ΔAUC represents the mean decline in model performance between internal and external validation. Values are descriptive and not derived from pooled statistical analysis. Abbreviations: Some important abbreviations are AUC (area under the receiver operating characteristic curve), CNN (convolutional neural network), ML (machine learning), RANO (Response Assessment in Neuro-Oncology), IDH (isocitrate dehydrogenase), MGMT (O6-methylguanine-DNA methyltransferase), and ResNet (residual neural network). TCIA, The Cancer Imaging Archive; BraTS, Brain Tumour Segmentation Challenge; MRI, magnetic resonance imaging.*

Category	Variable	Studies n (%)	Mean Internal AUC	Mean External AUC	ΔAUC (Decline)
Validation Approach	Cross-validation only	35 (55.6)	0.86 ± 0.07	—	—
	Hold-out test set	38 (60.3)	0.84 ± 0.08	—	—
	External validation – same vendor	8 (12.7)	0.87 ± 0.06	0.84 ± 0.07	−0.03
	External validation – different vendor	10 (15.9)	0.86 ± 0.07	0.76 ± 0.09	−0.10
	Public datasets (TCIA/BraTS)	14 (22.2)	0.85 ± 0.08	0.79 ± 0.10	−0.06
	Prospective validation	3 (4.8)	0.83 ± 0.09	0.77 ± 0.11	−0.06
Factors for Decline	Different MRI vendor	18	—	—	−0.10 ± 0.04
	Different field strength	12	—	—	−0.09 ± 0.05
	Different acquisition protocol	24	—	—	−0.12 ± 0.06
	Different patient population	16	—	—	−0.07 ± 0.04
	Time period difference	8	—	—	−0.05 ± 0.03

Quality Assessment

Quality assessment using the QUADAS-2 tool revealed that approximately 38% of included studies demonstrated a high risk of bias in at least one methodological domain. The most prevalent methodological limitations identified were inadequate clinical utility evaluation (86%), poor justification of sample size (81%), absence of external validation (71%), and insufficient examination of model interpretability (68%).

Adherence to TRIPOD reporting guidelines across included studies was inconsistent. Complete reporting of model performance measures was present in 65% of studies, while pre-specified statistical analysis plans were reported in only 29% of studies. These findings highlight widespread methodological and reporting deficiencies that limit the reproducibility and clinical translatability of current radiomic research in neuro-oncology. The quality assessment findings are illustrated in Figure 4.

DISCUSSION

Principal Findings

This systematic review synthesized evidence from 63 studies encompassing 12,847 patients, demonstrating that machine learning-enhanced radiomics consistently demonstrated higher AUC values than conventional RANO criteria across multiple neuro-oncological applications (15, 17, 18). Multiparametric MRI strategies combining complementary imaging sequences demonstrated performance improvements of 15–20% over single-sequence or visual evaluation approaches (17, 19). Deep learning architectures offered additional improvements of 5–12% over classical machine learning methods, at the cost of reduced interpretability and greater computational demands (22, 25).

The ability to non-invasively predict molecular markers with AUCs approaching 0.90 represents a

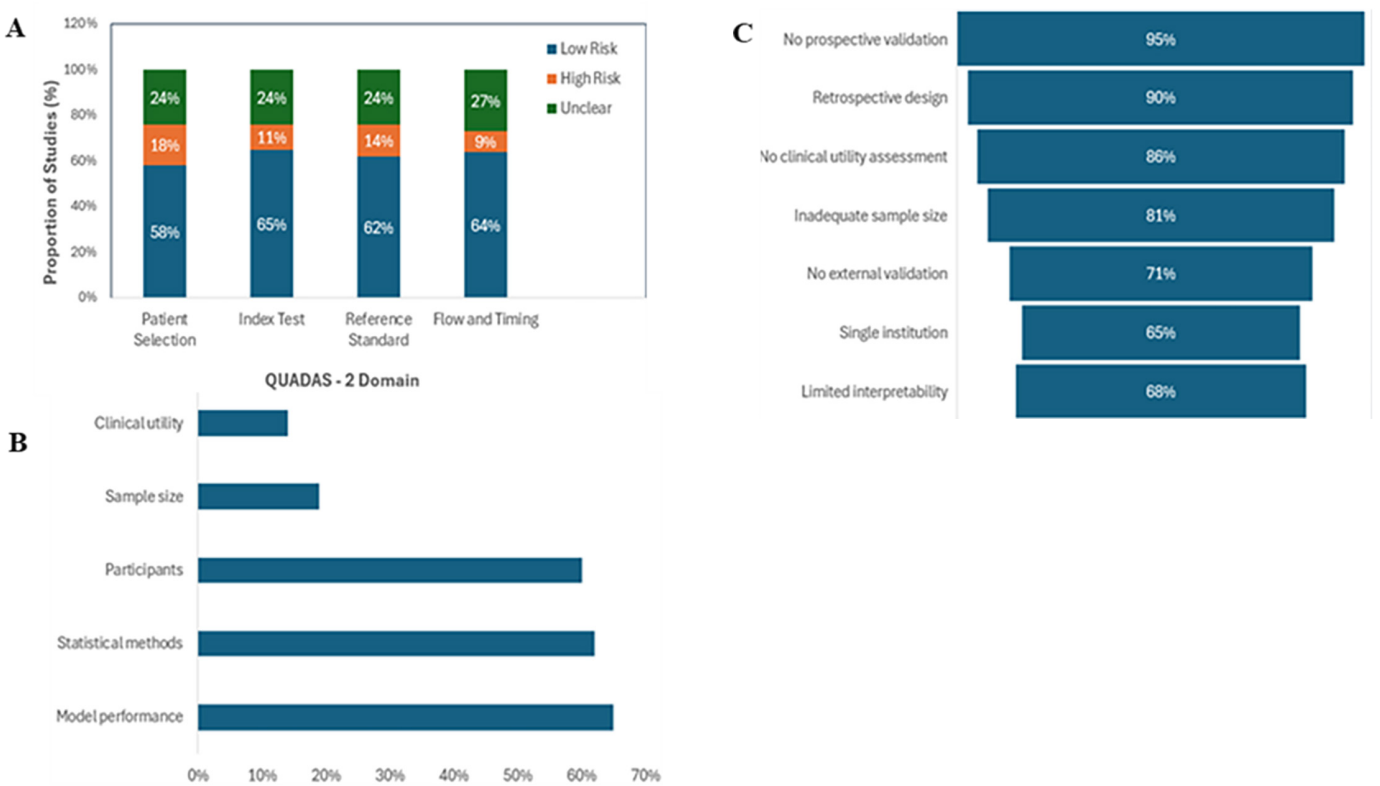

Figure 4. *Quality assessment of included studies.(A) The risk-of-bias evaluation based on the QUADAS-2 framework took into account four methodological domains and classified the studies based on low, high, or unclear risk. (B) Adherence to TRIPOD reporting guideline items across included studies. (C) Prevalence of common methodological limitations identified in the literature. Abbreviations: QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2), and TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis).*

clinically meaningful advance, potentially enabling treatment stratification and longitudinal molecular monitoring without repeated tissue sampling (22, 28). However, the consistent performance decline observed during external validation — averaging approximately 0.08 AUC units — underscores the critical importance of generalizability testing before clinical deployment (34). These findings collectively indicate that while the technical promise of radiomic biomarkers is well established, the pathway to routine clinical integration requires substantial additional methodological work.

Clinical Implications and Translation Potential

Accurate discrimination of pseudo-progression from true tumour progression has direct and immediate clinical relevance, as misclassification can lead to premature discontinuation of effective therapy or unnecessary surgical intervention in patients already burdened by a serious diagnosis (17, 18). The reported 15–20% improvements in sensitivity and specificity over RANO criteria are clinically significant, particularly given that pseudo-progression occurs in 20–30% of patients receiving concurrent chemoradiotherapy (4, 19).

Non-invasive molecular marker prediction offers multiple clinical use cases, including rapid treatment stratification while awaiting definitive pathology, monitoring of molecular evolution during treatment, identification of therapeutic targets for precision medicine, and patient selection for clinical trial enrolment (22, 24, 28). Radiomic biomarkers may also complement existing clinical and pathological variables to enhance risk stratification, in a manner analogous to established scoring systems in other oncological settings (29, 30). Improved survival prognostication can support more informed patient-physician discussions regarding treatment intensity, clinical trial participation, and quality-of-life considerations (30, 31). However, clinical utility must extend beyond predictive accuracy to encompass decision impact, measurable outcome improvement, and cost-effectiveness — dimensions that remain poorly examined in the existing literature (34, 35).

Methodological Challenges and Reporting Standards

The substantial performance decline observed during external validation highlights imaging acquisition heterogeneity as a primary barrier to clinical translation (34). Systematic variation in radiomic features attributable to scanner manufacturers, field strengths, sequence parameters, and reconstruction algorithms can be partially addressed through harmonization approaches such as ComBat correction and histogram normalization, though these methods themselves require further prospective validation before routine implementation (34, 36). The Image Biomarker Standardization Initiative provides essential guidance for reproducible feature extraction, yet full adoption remains limited — only 25% of included studies reported IBSI compliance (15, 37).

The absence of external validation in 71% of included studies and prospective assessment in fewer than 5% represents a critical evidence gap (38, 39). Multi-institutional external validation with transparent performance reporting across sites should be considered a minimum standard for studies seeking clinical translation (39). Adherence to TRIPOD reporting recommendations was incomplete across included studies, particularly regarding sample size justification and clinical utility evaluation — deficiencies identified as systemic across the neuro-oncology radiomics literature (40).

Model Interpretability and Clinical Trust

Interpretability analyses were included in only 32% of reviewed studies, despite growing recognition that clinical adoption requires physicians to understand and trust the reasoning underlying model decisions (35, 37). Explainable artificial intelligence methods — including gradient-weighted class activation mapping (Grad-CAM), SHapley Additive exPlanations (SHAP), and attention-based mechanisms — can provide valuable insight into model decision-making processes and may facilitate regulatory approval and physician acceptance. (35, 37). Importantly, interpretability analyses in the reviewed studies frequently revealed that models prioritised imaging characteristics such as tumour margin morphology, peritumoral oedema patterns, and enhancement features — regions biologically known to carry prognostic and diagnostic information — suggesting meaningful alignment between algorithmic reasoning and established clinical knowledge (37). Nevertheless, interpretability must extend beyond simply describing model predictions; it should also identify the circumstances under which models may be unreliable, to ensure appropriate clinical integration and safe deployment (35, 37).

Limitations

Several limitations of this review warrant acknowledgement. Publication bias may have resulted in under-representation of negative studies, potentially leading to overestimation of model performance across

the literature (40). Restriction to English-language publications may have excluded relevant international studies. The rapid evolution of the field means that some recent methodological advances may not be fully captured within the defined search period. Substantial methodological heterogeneity across included studies precluded formal quantitative meta-analysis, limiting the precision of pooled effect estimates. Furthermore, as a single-author review, the absence of independent dual-reviewer screening and data extraction introduces the possibility of selection bias and extraction error, which should be considered when interpreting findings. To mitigate this risk, a standardised eligibility checklist was applied throughout screening and all extracted data were independently re-verified against source articles in a second pass conducted at least one week after initial extraction. Quality assessment, while conducted using validated tools, retains an inherent degree of subjectivity. The absence of prospective protocol registration on PROSPERO represents an additional methodological limitation, as it precludes independent verification that methods were not modified following data collection.

CONCLUSION

This systematic review synthesized evidence from 63 studies encompassing 12,847 patients to evaluate machine learning-enhanced radiomics across three neuro-oncological applications: treatment response prediction, molecular marker characterization, and survival prognostication. Across all three domains, radiomic models incorporating multiparametric MRI features reported higher performance metrics than conventional RANO criteria, with deep learning architectures reporting the highest AUC values, particularly for distinguishing pseudo-progression from true tumour progression and for non-invasive prediction of IDH mutation status.

Despite these promising findings, external validation remains a critical limitation of the current literature, with independent institutional validation performed in fewer than 29% of included studies and associated with a mean AUC decline of approximately 0.08. Substantial heterogeneity in imaging acquisition protocols, feature extraction pipelines, and validation strategies limits direct comparison across studies and constrains generalizability of reported performance metrics. These findings should therefore be interpreted within the context of a narrative synthesis rather than formal statistical pooling.

Addressing persistent barriers to clinical translation will require standardized imaging acquisition protocols aligned with IBSI guidelines, multi-institutional external validation frameworks, adoption of explainable artificial intelligence methods to improve model transparency, and full adherence to TRIPOD and PRISMA reporting standards. Federated learning approaches represent a particularly promising mechanism for achieving large-scale multi-institutional validation while preserving patient data confidentiality.

Future research must prioritise prospective randomised validation studies that evaluate not only predictive accuracy but also clinical decision impact, patient outcome improvement, and cost-effectiveness. With continued methodological refinement and collaborative validation, radiomics and machine learning hold meaningful potential to advance precision neuro-oncology and support more individualised management of glioma patients.

ACKNOWLEDGEMENTS

I am grateful to the library staff at Westford Academy, who assisted me with literature searches and obtaining the manuscripts.

CONFLICT OF INTEREST

The author does not assert any conflict of interest with this work.

REFERENCES

1. Stupp R, Mason WP, van den Bent MJ, *et al*. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med*. 2005; 352 (10): 987-996. https://doi.org/10.1056/NEJMoa043330
2. Wen PY, Macdonald DR, Reardon DA, *et al*. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J Clin Oncol*. 2010; 28 (11): 1963-1972. https://doi.org/10.1200/JCO.2009.26.3541
3. Ellingson BM, Wen PY, Cloughesy TF. Modified criteria for radiographic response assessment in glioblastoma clinical trials. *Neurotherapeutics*. 2017; 14 (2): 307-320. https://doi.org/10.1007/s13311-016-0507-6
4. Brandes AA, Franceschi E, Tosoni A, *et al*. MGMT promoter methylation status can predict the incidence and outcome of pseudoprogression after concomitant

radiochemotherapy in newly diagnosed glioblastoma patients. *J Clin Oncol*. 2008; 26 (13): 2192-2197. https://doi.org/10.1200/JCO.2007.14.8163
5. Hygino da Cruz LC Jr, Rodriguez I, Domingues RC, *et al*. Pseudoprogression and pseudoresponse: imaging challenges in the assessment of posttreatment glioma. *AJNR Am J Neuroradiol*. 2011; 32 (11): 1978-1985. https://doi.org/10.3174/ajnr.A2397
6. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016; 278 (2): 563-577. https://doi.org/10.1148/radiol.2015151169
7. Lambin P, Rios-Velazquez E, Leijenaar R, *et al*. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012; 48 (4): 441-446. https://doi.org/10.1016/j.ejca.2011.11.036
8. Bi WL, Hosny A, Schabath MB, *et al*. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA Cancer J Clin*. 2019; 69 (2): 127-157. https://doi.org/10.3322/caac.21552
9. Tabassum M, Suman AA, Suero Molina E, Pan E, Di Ieva A, Liu S. Radiomics and machine learning in brain tumors and their habitat: a systematic review. *Cancers*. 2023; 15 (15): 3845. https://doi.org/10.3390/cancers15153845
10. Tabatabaei M, Razaei A, Sarrami AH, Saadatpour Z, Singhal A, Sotoudeh H. Current status and quality of machine learning-based radiomics studies for glioma grading: a systematic review. *Oncology*. 2021; 99 (7): 433-443. https://doi.org/10.1159/000515597
11. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015; 521 (7553): 436-444. https://doi.org/10.1038/nature14539
12. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019; 1 (5): 206-215. https://doi.org/10.1038/s42256-019-0048-x
13. Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021; 372: n71. https://doi.org/10.1136/bmj.n71
14. Whiting PF, Rutjes AW, Westwood ME, *et al*. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011; 155 (8): 529-536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
15. Zwanenburg A, Vallières M, Abdalah MA, *et al*. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020; 295 (2): 328-338. https://doi.org/10.1148/radiol.2020191145
16. van Griethuysen JJM, Fedorov A, Parmar C, *et al*. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*. 2017; 77 (21): e104-e107. https://doi.org/10.1158/0008-5472.CAN-17-0339
17. Patel M, Zhan J, Bhatt D, *et al*. Machine learning-based radiomic evaluation of treatment response prediction in glioblastoma. *Clin Radiol*. 2021; 76 (8): 628.e17-628.e27. https://doi.org/10.1016/j.crad.2021.03.019
18. Sun YZ, Yan LF, Han Y, *et al*. Differentiation of pseudoprogression from true progression in glioblastoma patients after standard treatment: a machine learning strategy combined with radiomics features from T1-weighted contrast-enhanced imaging. *BMC Med Imaging*. 2021; 21 (1): 17. https://doi.org/10.1186/s12880-020-00545-5
19. Jang BS, Jeon SH, Kim IH, Kim IA. Machine learning model to predict pseudoprogression versus progression in glioblastoma using MRI: a multi-institutional study (KROG 18-07). *Cancers*. 2020; 12 (9): 2706. https://doi.org/10.3390/cancers12092706
20. Kickingereder P, Isensee F, Tursunova I, *et al*. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol*. 2019; 20 (5): 728-740. https://doi.org/10.1016/S1470-2045(19)30098-1
21. Lohmann P, Kocher M, Ceccon G, *et al*. Combined FET PET/MRI radiomics differentiates radiation injury from recurrent brain metastasis. *NeuroImage Clin*. 2018; 20: 537-542. https://doi.org/10.1016/j.nicl.2018.08.024
22. 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. https://doi.org/10.1093/neuonc/noaa177
23. Zhang S, Sun H, Su X, *et al*. Automated machine learning to predict the co-occurrence of isocitrate dehydrogenase mutations and O6-methylguanine-DNA methyltransferase promoter methylation in patients with gliomas. *J Magn Reson Imaging*. 2021; 54 (1): 197-205. https://doi.org/10.1002/jmri.27498, https://doi.org/10.1002/jmri.26995
24. Sun C, Fan L, Wang W, *et al*. Radiomics and qualitative features from multiparametric MRI predict molecular subtypes in patients with lower-grade glioma. *Front Oncol*. 2022; 11: 756828. https://doi.org/10.3389/fonc.2021.756828
25. Chen S, Xu Y, Ye M, *et al*. Predicting MGMT promoter methylation in diffuse gliomas using deep learning with radiomics. *J Clin Med*. 2022; 11 (12): 3445. https://doi.org/10.3390/jcm11123445

26. Han W, Qin L, Bay C, *et al*. Diagnostic performance of radiomics using machine learning algorithms to predict MGMT promoter methylation status in glioma patients: a meta-analysis. *Front Oncol*. 2021; 11: 734303.
27. Zhang Y, Yan LF, Zhang J, *et al*. Preoperative prediction of MGMT promoter methylation in glioblastoma based on multiregional and multi-sequence MRI radiomics analysis. *Sci Rep*. 2024; 14: 15817. https://doi.org/10.1038/s41598-024-66653-2
28. Bhandari AP, Liong R, Koppen J, Murthy SV, Lasocki A. Noninvasive determination of IDH and 1p/19q status of lower-grade gliomas using MRI radiomics: a systematic review. *AJNR Am J Neuroradiol*. 2021; 42 (1): 94-101. https://doi.org/10.3174/ajnr.A6875
29. Kha QH, Le VH, Hung TNK, Le NQK. Development and validation of an efficient MRI radiomics signature for improving the predictive performance of 1p/19q co-deletion in lower-grade gliomas. *Cancers.* 2021; 13 (21): 5398. https://doi.org/10.3390/cancers13215398
30. Shaheen A, Bukhari ST, Nadeem M, *et al*. Overall survival prediction of glioma patients with multiregional radiomics. *Front Neurosci.* 2022; 16: 911065. https://doi.org/10.3389/fnins.2022.911065
31. Kiesel B, Freudiger C, Gatterbauer B, *et al*. Deep learning-assisted radiomics facilitates multimodal prognostication for personalized treatment strategies in low-grade glioma. *Sci Rep.* 2023; 13: 9767. https://doi.org/10.1038/s41598-023-36298-8
32. 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 (6): 1045-1057. https://doi.org/10.1007/s10278-013-9622-7
33. 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. https://doi.org/10.1038/sdata.2017.117
34. Orlhac F, Lecler A, Savatovski J, *et al*. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *Eur Radiol*. 2021; 31 (4): 2272-2280. https://doi.org/10.1007/s00330-020-07284-9
35. Severn C, Suresh K, Görg C, *et al*. A pipeline for the implementation and visualization of explainable machine learning for medical imaging using radiomics features. *Sensors.* 2022; 22 (14): 5205. https://doi.org/10.3390/s22145205
36. Orlhac F, Eertink JJ, Cottereau AS, *et al*. A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *J Nucl Med*. 2022; 63 (2): 172-179. https://doi.org/10.2967/jnumed.121.262464
37. Lohmann P, Franceschi E, Vollmuth P, *et al*. Radiomics in neuro-oncological clinical trials. *Lancet Digit Health*. 2022; 4 (12): e841-e849. https://doi.org/10.1016/S2589-7500(22)00144-3
38. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015; 162 (1): 55-63. https://doi.org/10.7326/0003-4819-14-1-55, https://doi.org/10.7326/M14-0697
39. Vils A, Bogowicz M, Tanadini-Lang S, *et al*. Radiomic analysis to predict outcome in recurrent glioblastoma based on multi-center MR imaging from the prospective DIRECTOR trial. *Front Oncol.* 2021; 11: 636672. https://doi.org/10.3389/fonc.2021.636672
40. Park JE, Kim HS, Kim N, *et al*. A systematic review reporting quality of radiomics research in neuro-oncology: toward clinical utility and quality improvement using high-dimensional imaging features. *BMC Cancer.* 2020; 20 (1): 29. https://doi.org/10.1186/s12885-019-6504-5