AJNR

Development and Evaluation of Automated Artificial Intelligence–Based Brain Tumor Response Assessment in Patients with Glioblastoma

Jikai Zhang, Dominic LaBella, Dylan Zhang, Jessica L. Houk, Jeffrey D. Rudie, Haotian Zou, Pranav Warman, Maciej A. Mazurowski and Evan Calabrese

AJNR Am J Neuroradiol published online 14 November 2024
http://www.ajnr.org/content/early/2025/04/24/ajnr.A8580

Published April 24, 2025 as 10.3174/ajnr.A8580

Development and Evaluation of Automated Artificial Intelligence–Based Brain Tumor Response Assessment in Patients with Glioblastoma

Jikai Zhang, Dominic LaBella, Dylan Zhang, Jessica L. Houk, Jeffrey D. Rudie, Haotian Zou, Pranav Warman, Maciej A. Mazurowski, and Evan Calabrese

ABSTRACT

SUMMARY: This project aimed to develop and evaluate an automated, AI-based, volumetric brain tumor MRI response assessment algorithm on a large cohort of patients treated at a high-volume brain tumor center. We retrospectively analyzed data from 634 patients treated for glioblastoma at a single brain tumor center over a 5-year period (2017–2021). The mean age was 56 ± 13 years. 372/634 (59%) patients were male, and 262/634 (41%) patients were female. Study data consisted of 3,403 brain MRI exams and corresponding standardized, radiologist-based brain tumor response assessments (BT-RADS). An artificial intelligence (AI)-based brain tumor response assessment (AI-VTRA) algorithm was developed using automated, volumetric tumor segmentation. AI-VTRA results were evaluated for agreement with radiologist-based response assessments and ability to stratify patients by overall survival. Metrics were computed to assess the agreement using BT-RADS as the ground-truth, fixed-time point survival analysis was conducted to evaluate the survival stratification, and associated P-values were calculated. For all BT-RADS categories, AI-VTRA showed moderate agreement with radiologist response assessments (F1 = 0.587–0.755). Kaplan-Meier survival analysis revealed statistically worse overall fixed time point survival for patients assessed as image worsening equivalent to RANO progression by human alone compared to by AI alone (log-rank $P = .007$). Cox proportional hazard model analysis showed a disadvantage to AI-based assessments for overall survival prediction ($P = .012$). In summary, our proposed AI-VTRA, following BT-RADS criteria, yielded moderate agreement for replicating human response assessments and slightly worse stratification by overall survival.

ABBREVIATIONS: 2D = 2-dimensional; AI = artificial intelligence; AI-VTRA = artificial intelligence volumetric tumor response assessment; BT-RADS = Brain Tumor Reporting and Data System; C-index = concordance index; FeTS = Federated Tumor Segmentation; GBM = glioblastoma; *IDH* = *isocitrate dehydrogenase*; NLP = natural language processing; OS = overall survival; RANO = Response Assessment in Neuro-Oncology; RECIST = Response Evaluation Criteria in Solid Tumors; SD = standard deviation; VD_{ET} = volumetric differences for enhancing tumor; VD_{FLAIR} = volumetric differences for FLAIR

Glioblastoma (GBM) is the most common primary brain malignancy in adults and remains difficult to treat even with the benefit of decades of experience.[1] Despite improved understanding of the genetic underpinnings of brain malignancies, treatment options for GBM are limited, and survival remains poor.[2-4] GBM management is further complicated by the complexity and frequency of clinical and radiologic response assessments, which may occur as often as every 4 weeks during active treatment.[5] Brain MRI plays a critical role in GBM treatment response assessments and, along with comprehensive clinical assessment, is central for determining treatment response and/or disease progression.[6,7]

Given the importance of MRI for GBM treatment monitoring, there have been extensive efforts to develop standardized MRI response assessment criteria.[8] Originally proposed in 1990, the McDonald criteria were widely considered the standard for GBM MRI response assessments, particularly for clinical trials.[9] While similar to other solid tumor response assessment criteria, such as the Response Evaluation Criteria in Solid Tumors (RECIST),[10] the McDonald criteria employed 2-dimensional (2D) tumor measurements to better capture the complex shape that is typical of GBM. In the following decades, the Response Assessment in Neuro-Oncology (RANO) criteria and its variations[6,11] superseded the McDonald criteria, with their primary advantage being

Received July 28, 2024; accepted after revision October 19.

From the Departments of Electrical and Computer Engineering (J.Z., M.A.M.) and Computer Science (M.A.M.), Duke University, Durham, North Carolina; Duke Center for Artificial Intelligence in Radiology (J.Z., E.C.) and Departments of Radiation Oncology (D.L.) and Radiology (D.Z., J.L.H., M.A.M., E.C.), Duke University Medical Center, Durham, North Carolina; Department of Biostatistics and Bioinformatics (H.Z., M.A.M.), Duke University School of Medicine (P.W.), Durham, North Carolina; and Department of Radiology (J.D.R.), University of California San Diego, San Diego, California.

This research has been supported in part by an award from the Foundation of the American Society of Neuroradiology to Dr. Evan Calabrese titled "Prospective Evaluation of Automated Pre- and Postoperative Tumor Segmentation for Patients with Glioblastoma."

Please address correspondence to Evan Calabrese, MD, PhD, DUMC Box 3808; Durham, NC 27710; e-mail: evan.calabrese@duke.edu

Indicates article with supplemental data.

http://dx.doi.org/10.3174/ajnr.A8580

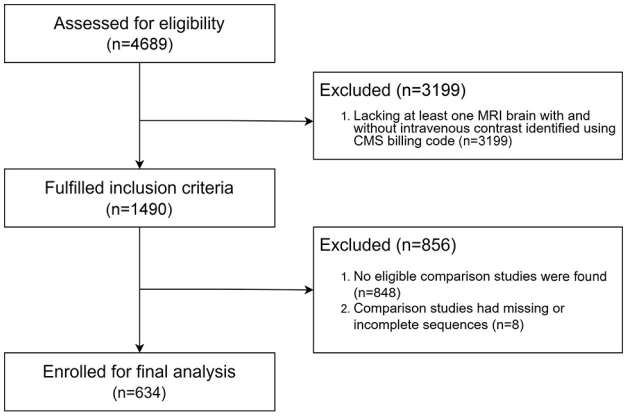


FIG 1. Patient flow diagram for study inclusion.

the consideration of both *enhancing* and *nonenhancing* tumors in addition to relevant treatment modalities. While RANO continues to be widely used in clinical trials, it is not commonly used for routine clinical assessments owing to its complexity.[7] RANO 2.0 updates RANO by providing unified criteria to assess gliomas regardless of their grades and recommends volumetric assessments.[27]

More recent efforts toward response assessment standardization have included the Brain Tumor Reporting and Data System (BT-RADS), a standardized MRI reporting system designed to simplify brain MRI reporting for routine clinical follow-up of patients with GBM.[12-14] Similar to RANO, BT-RADS relies on measurements of both enhancing and nonenhancing tumors, and the BT-RADS 4 category was designed to be equivalent to the primary imaging criterion for RANO progression.[6,12] The main advantage of BT-RADS is its ease of use and implementation. In contrast to RANO, BT-RADS has seen more rapid adoption for routine clinical use and has been implemented at several major brain tumor centers since it was first proposed in 2018.[13] RANO 2.0 and BT-RADS differ in scope (RANO 2.0 primarily focused on clinical trials and BT-RADS on routine assessments) and in approach. Specifically, RANO 2.0 proposes a unified set of criteria for high- and lower-grade gliomas, while BT-RADS was designed for high-grade gliomas. Both criteria acknowledge changes in enhancing and nonenhancing tumors, and both share similar criteria for tumor progression (25% increase in enhancing tumor). However, other RANO 2.0 categories do not have straightforward relationships to BT-RADS categories. For example, RANO 2.0 "partial response" requires a 50% 2D/linear decrease in enhancing tumor, while BT-RADS 1 (imaging improvement) does not specify an enhancing tumor decrease threshold. However, BT-RADS, like its predecessors, relies on 2D measurements, which may not accurately capture the complex 3D shape of GBM.[15] In addition, it should be acknowledged that human BT-RADS assessments are an imperfect reference standard as they are somewhat subjective and dependent on manual measurements and interpreting radiologists' adherence to published guidelines. While previous volumetric (3D) response assessment criteria have been proposed, implementation has been hindered by the difficulty in translating volumetric changes into response assessment categories.

Automated artificial intelligence (AI)-based volumetric brain tumor MRI segmentation has recently matured into a clinically viable tool principally because of a large collaborative efforts such as the multimodal brain tumor segmentation challenge[16] and the global Federated Tumor Segmentation (FeTS) initiative.[17] This has led several groups to explore the use of AI-based segmentation tools for automated volumetric GBM MRI response assessment.[18-21] In this work, we evaluate an automated, AI-based, volumetric brain tumor response assessment tool on a large cohort of patients treated at a high-volume brain tumor center. We compare AI-based results to standardized neuroradiologist response assessments in 2 key domains: ability to recapitulate human response assessments and ability to stratify patients by overall survival (OS).

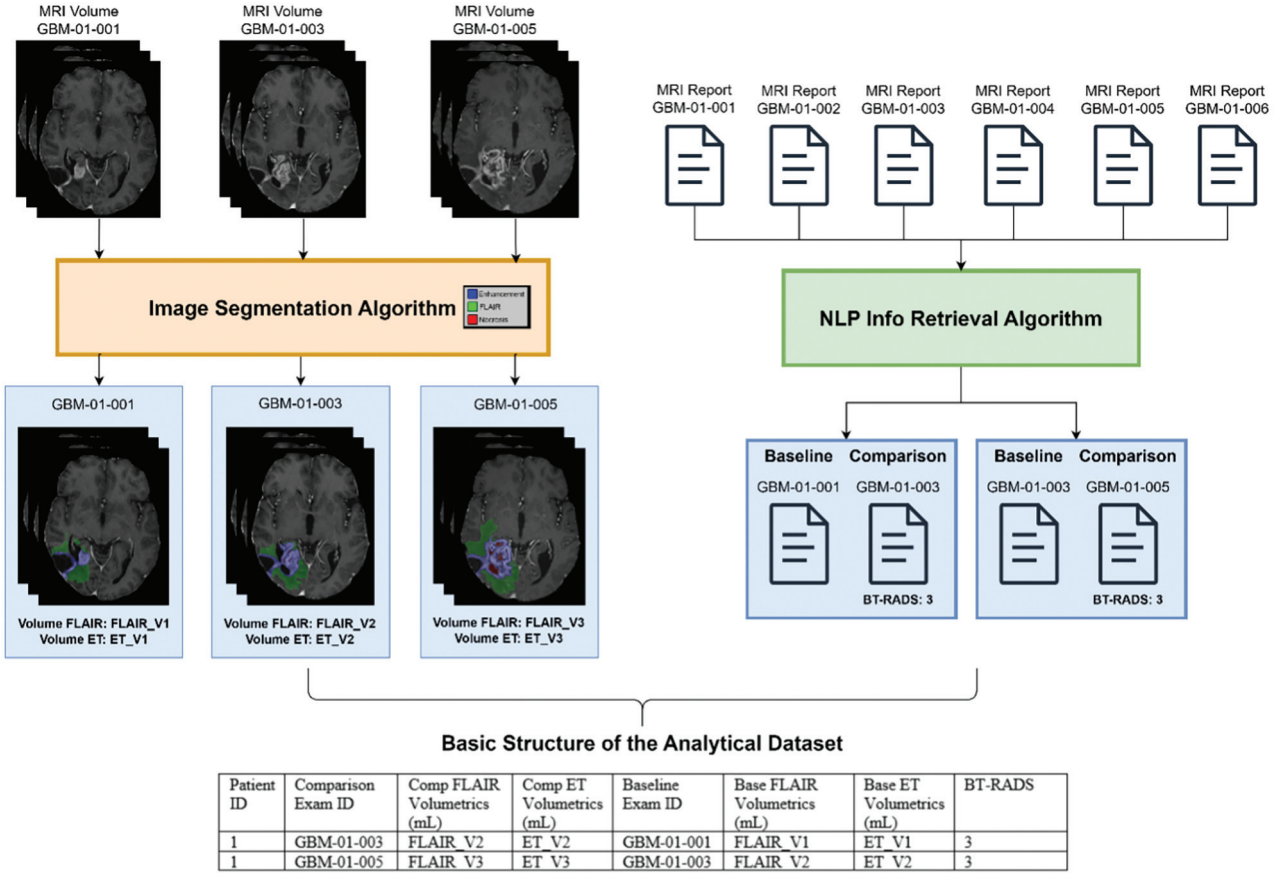
MATERIALS AND METHODS

Study Population

This was a single-center, retrospective, Institutional Review Board–approved study with a waiver for informed consent. Candidate participants were identified by systematic search of electronic health encounter records from 2017–2021 for all adult patients with a diagnosis of "glioblastoma" at a high-volume academic brain tumor center by using Center for Medicaid Services Hierarchical Condition Category codes (n = 4689). This included both *isocitrate dehydrogenase (IDH)* mutant and wild-type grade 4 astrocytomas in line with current WHO classifications at the time of diagnosis (referred to as "GBM" henceforth for conciseness). Exclusion criteria were: patients lacking at least 1 MRI brain examination with and without intravenous contrast (n=3199) and patients lacking at least 1 standardized neuroradiologist response assessment (n=856). The final study population consisted of 634 patients. A patient flow diagram is provided as Fig 1.

Neuroradiologist Response Assessments

Formal neuroradiologist-based GBM MRI response assessments by using the BT-RADS structured reporting system were available as part of routine clinical care. BT-RADS scores and baseline comparison examination dates were extracted from radiology reports by using a custom semisupervised natural language processing (NLP) algorithm with near-perfect internal validation performance. The full data curation pipeline is shown in Fig 2. For each patient, we searched for all reports containing BT-RADS scores. Then, for each BT-RADS report, the NLP algorithm retrieved the prior examination date and searched for its prior examination with the retrieved examination date (complete methodologic details and performance assessment provided as Supplemental Data). This yielded 2446 pairs of examinations (current and baseline prior) with BT-RADS scores. One baseline prior can be paired with multiple follow-up examinations. BT-RADS scores included the following numerical categories: 1 = imaging improvement, 2 = no appreciable imaging change, 3 = imaging worsening, 4 = imaging worsening with >25% increase in 2D enhancing tumor measurements (equivalent to RANO progression).



Patient ID	Comparison Exam ID	Comp FLAIR Volumetrics (mL)	Comp ET Volumetrics (mL)	Baseline Exam ID	Base FLAIR Volumetrics (mL)	Base ET Volumetrics (mL)	BT-RADS
1	GBM-01-003	FLAIR_V2	ET_V2	GBM-01-001	FLAIR_V1	ET_V1	3
1	GBM-01-005	FLAIR_V3	ET_V3	GBM-01-003	FLAIR_V2	ET_V2	3

FIG 2. Pipeline of data curation process aided by NLP and image segmentation methods.

MRI Data

All routine brain tumor MRI examinations were performed with a Brain Tumor Imaging Protocol,[22] a compliant protocol including 3D, gradient-echo, T1-weighted pre- and postcontrast sequences and 2D, T2-weighted, and T2-FLAIR sequences. MRI data were retrieved for each pair of examinations corresponding to the BT-RADS scores identified in the previous section, which resulted in 3403 unique MRI examinations. Scanner information is included in the Supplemental Data.

Image Processing and Automated Tumor Segmentation

MRI data underwent standard image preprocessing steps including translation-only alignment to the Montreal Neurological Institute brain atlas (MNI352) for FOV standardization[23] and skull stripping by using a publicly available deep learning method.[24] Preprocessed images then underwent automated, volumetric tumor segmentation by using 3D convolutional segmentation neural network. This model was specifically designed for post-treatment examinations including 4 distinct compartments: resection cavity, enhancing tumor, necrotic tumor core, and surrounding nonenhancing T2-FLAIR signal abnormality. The final model was pretrained on an external postoperative brain MRI examination. We utilized nnU-Net[36] to train and validate the model. Internal validation results showed a mean ± standard deviation (SD) of 0.8861 ± 0.2476 for enhancing tumor and 0.9833 ± 0.0372 for surrounding nonenhancing FLAIR signal abnormality (complete methodologic details and performance assessment provided as Supplemental Data).

Artificial Intelligence Volumetric Tumor Response Assessment (AI-VTRA)

An AI scoring system (AI-VTRA) based on volumetric differences for enhancing tumor (VD_{ET}) and surrounding nonenhancing FLAIR hyperintensity (VD_{FLAIR}) was computed for each pair of examinations in the data set and was used to develop AI-based volumetric equivalents to BT-RADS scores. BT-RADS 4 was defined as a ≥40% increase in VD_{ET}, as the extrapolated volumetric threshold derived from 2D measurements, for measurable disease (enhancing tumor volume greater than 1 mL) consistent with multiple previously published studies.[25-27,38] Other relevant volumetric thresholds (notably a ± 10% threshold for no significant change) were determined empirically, as previously published values did not exist. BT-RADS 3 was defined as either 1) VD_{ET} between 10% and 40% increase or 2) VD_{ET} < 10% change and VD_{FLAIR} ≥ 40% increase. BT-RADS 2 was defined as either 1) VD_{ET} < 10% change or 2) VD_{ET} ≥ 10% increase and VD_{FLAIR} ≥ 40% increase. BT-RADS 1 was defined as either 1) VD_{ET} ≥ 10% decrease or 2) VD_{ET} < 10% change and ≥ 40% decrease in VD_{FLAIR}. Complete criteria for AI-VTRA are presented in Table 1. To assess the importance of including VD_{FLAIR}, we also evaluated AI-$VTRA_{ET}$, which was solely based on VD_{ET} (Supplemental Data).

Table 1: Relationship between BT-RADS score and AI-VTRA for each glioblastoma MRI follow-up assessment score[a]

Assessment Category	Assessment System (Rater)	
	BT-RADS Score (Human)	AI-VTRA (Artificial Intelligence)
Imaging improvement	1	$VD_{ET} \leq -10\%$ OR $-10\% > VDET < 10\%$ AND $VD_{FLAIR} \leq -40\%$
No appreciable imaging change	2	$-10\% > VDET < 10\%$ OR $VD_{ET} \leq -10\%$ AND $VD_{FLAIR} \geq 40\%$
Imaging worsening	3	$10\% \leq VDET < 40\%$ OR $-10\% > VDET < 10\%$ AND $VD_{FLAIR} \geq 40\%$
Imaging worsening equivalent to RANO progression	4	$VD_{ET} \geq 40\%$

[a]Detailed rules for determining AI-VTRA are included in Supplemental Data.

Table 2: Basic demographics for the 634 patients included in the study cohort[a]

Age in years Mean +/− SD.	56 +/- 13
Sex N (%)	
Male	372 (59%)
Female	262 (41%)
Primary self-reported race N (%)	
White	566 (89%)
Black or African American	41 (7%)
Asian	9 (1%)
Other	8 (1%)
Not reported/declined	10 (2%)
Self-reported ethnicity N (%)	
Not Hispanic	592 (93%)
Hispanic	11 (5%)
Not reported/declined	31 (2%)
IDH N (%)	
Wild-type	479 (76%)
Mutant	63 (10%)
Inconclusive/missing	92 (14%)
Tumor types N (%)	
Enhancing	2163 (64%)
Nonenhancing edema/FLAIR	3401 (99%)

[a]Patient age was assessed at the time of the first available MRI brain examination date. "Other" self-reported races included American Indian or Alaskan native and other. "Other/missing" IDH types included atypical IDH2 mutation, both positive and negative, not provided, indeterminate, or no records found in the database. The last item, "tumor types," recorded enhancing and nonenhancing tumors (at least 1 mL) across the 3403 examinations in the study cohort.

AI Performance for Recapitulating Human BT-RADS Scores

Performance of automated volumetric criteria for replicating human BT-RADS scores was evaluated across the entire data set. Composite performance for all BT-RADS categories was assessed with the macro-F1 score. Performance for individual BT-RADS categories was assessed with sensitivity, specificity, precision, micro-F1 score (calculated globally across all categories), and macro-F1 score (calculated for each category and then averaged).

AI Performance for Survival Stratification

Performance for survival stratification was assessed based on the highest response assessment category assigned within the first 6 months of MRI follow-up, which typically (though not necessarily) corresponded to the second postoperative MRI examination. Time from initial diagnosis was not available for all patients and was not included in the analysis. Three hundred twenty-three of 634 (51%) patients had at least 1 BT-RADS assessment in the first 6 months of follow-up and were included in this subanalysis. This cohort was substratified by response score and whether they were assigned this score by human alone, by AI alone, or by both human and AI simultaneously. We plotted Kaplan-Meier survival curves of each substrata to visualize survival probability. Patients who were still alive at the last available follow-up were censored. Log-rank tests were used to determine the pair-wise differences between survival curves.

Multivariate Survival Modeling

Multivariate Cox proportional hazard models were applied for human (Eq 1) and AI assessments (Eq 2) separately to assess the relative predictive value for survival prediction. Besides the scores, we included normalized age, sex, race, and ethnicity in the model. Time between baseline and follow-up examinations was considered as the time-varying covariate in the Cox model. We removed observations due to unknown IDH status before fitting the Cox models. The Concordance index (C-index) was calculated for each Cox model. To compare the difference in C-index between 2 Cox models, we applied statistical tests that account for the paired data (see Supplemental Data for details).

$$h_{human} = h_{0_{human}}(t)\exp(\alpha_1 * BTRADS + \alpha_2 * Norm(Age) + \alpha_3 * Sex + \alpha_4 * Race + \alpha_5 * Ethnicity + \alpha_6 * IDH) \quad (1)$$

$$h_{AI} = h_{0_{AI}}(t)\exp(\beta_1 * \text{AIVTRA} + \beta_2 * Norm(Age) + \beta_3 * Sex + \beta_4 * Race + \beta_5 * Ethnicity + \beta_6 * IDH) \quad (2)$$

Statistical Analyses

Statistical analyses were performed in Python Version 3.8 and R Version 4.2. Kaplan-Meier estimates were computed by using the "lifelines" package in Python. Cox modeling was performed in R by using the "survival" package. The *scale* method in R was used to normalize *Age*. We set the confidence level as 95%, and *P* values less than .05 were considered significant.

RESULTS

Patient Characteristics

Basic study participant demographic data are reported in Table 2. The mean age was 56 ± 13 years. Three hundred seventy-two of 634 (59%) patients were men, and 262/634 (41%) patients were women. Five hundred sixty-six of 634 (89%) patients listed their primary self-reported race as white, 41/634 (7%) as black or African American, and 9/634 (1%) as Asian. Eight of 634 (1%) patients reported a secondary race, and 10/634 (2%) patients did not report race. Four hundred seventy-nine of 634 (76%) patients

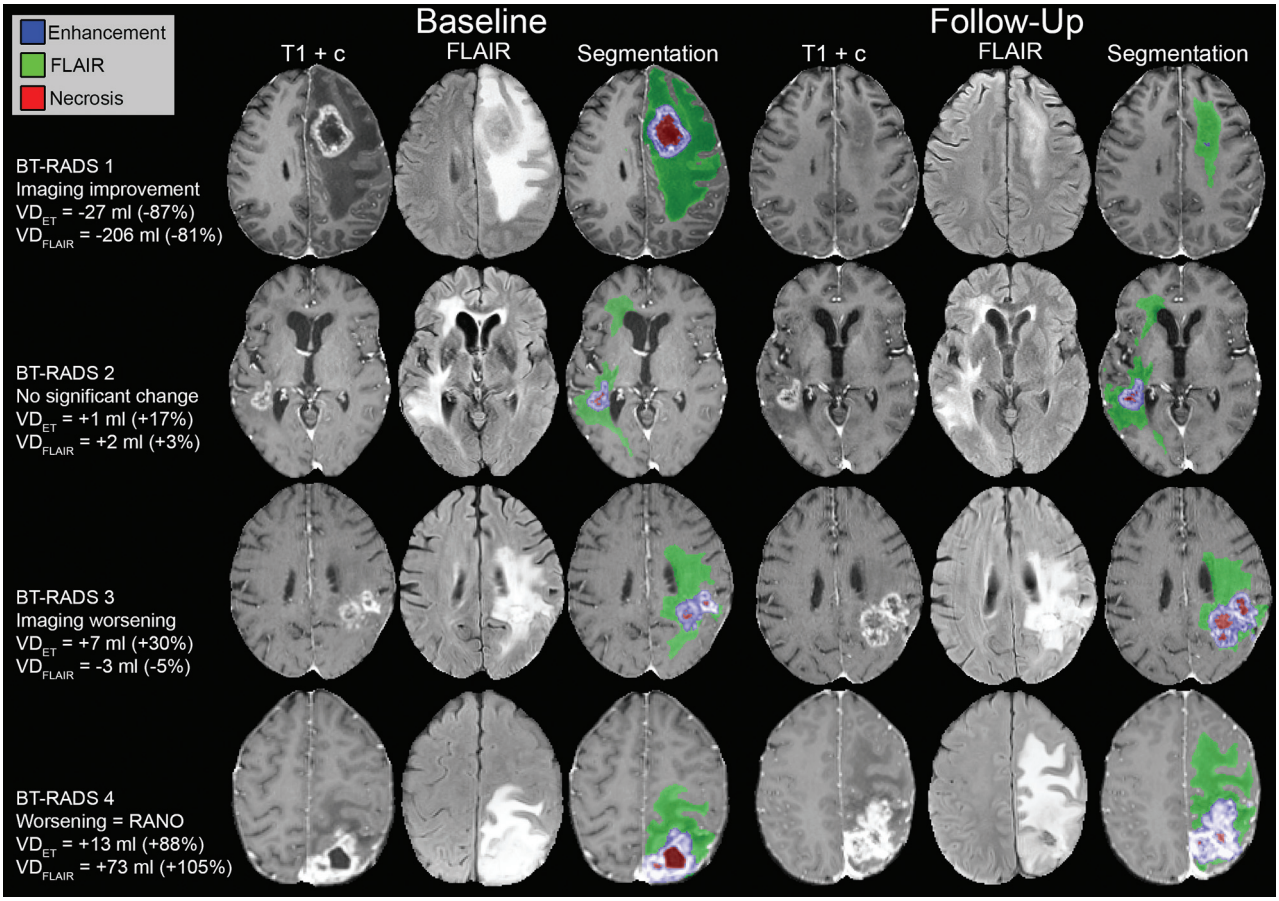


FIG 3. Example MR images, radiologist response assessment categories, and volumetric changes for 4 patients at 2 different time points.

Table 3: Performance metrics (macro-F1, micro-F1, sensitivity, specificity, and precision) for AI-VTRA/AI-VTRA$_{ET}$ predictions of radiologist-based response assessment. Within each category, we binarized the BT-RADS and AI predictions based on the target score and computed the metrics

	Imaging Improvement (BT-RADS 1)		No Significant Imaging Change (BT-RADS 2)		Imaging Worsening (BT-RADS 3)		Imaging Worsening Equivalent to RANO Progression (BT-RADS 4)	
	AI-VTRA$_{ET}$	AI-VTRA	AI-VTRA$_{ET}$	AI-VTRA	AI-VTRA$_{ET}$	AI-VTRA	AI-VTRA$_{ET}$	AI-VTRA
Macro-F1	0.747	0.755	0.760	0.750	0.561	0.587	0.705	0.705
Micro-F1	0.857	0.870	0.765	0.757	0.695	0.689	0.831	0.831
Sensitivity	0.747	0.700	0.793	0.746	0.222	0.298	0.596	0.596
Specificity	0.873	0.895	0.746	0.765	0.920	0.875	0.872	0.872
Precision	0.474	0.526	0.672	0.675	0.568	0.530	0.450	0.450

had an IDH wild-type tumor, 63/634 (10%) patients had an IDH mutant tumor, and 92/634 (14%) patients had missing or inconclusive IDH testing.

MRI Data and Segmentation

The 634 included patients had 3403 qualifying MRI brain examinations (average of 3.85 examinations per patient). The average time between baseline and follow-up studies was 160 days, with an SD of 236 days. Automated volumetric tumor segmentation was successfully completed for all examinations without errors. The average segmentation time was 11.5 seconds per examination. Representative segmented MRI from 4 different patients' examination pairs with each of the different assessment categories are presented in Fig 3.

AI Performance for Recapitulating Human BT-RADS Scores

For recapitulating human BT-RADS scores, AI-VTRA had a higher macro-F1 score (AI-VTRA macro-F1 = 0.548) compared with AI-VTRA$_{ET}$ (AI-VTRA$_{ET}$ macro-F1 = 0.535). Performance metrics for predicting each of the individual BT-RADS scores are provided in Table 3. AI-VTRA$_{ET}$ alone demonstrated improved performance compared with AI-VTRA for a single score, BT-RADS 2 (no significant change). Overall, automated volumetrics yielded moderate performance (F1 > 0.7) for predicting neuroradiologist BT-RADS scores of 1, 2, and 4, and yielded moderate performance (F1 > 0.55) for predicting BT-RADS 3. Total counts and percentages for each score and an analysis of major discrepancies between human and AI assessments are provided in the Supplemental Data.

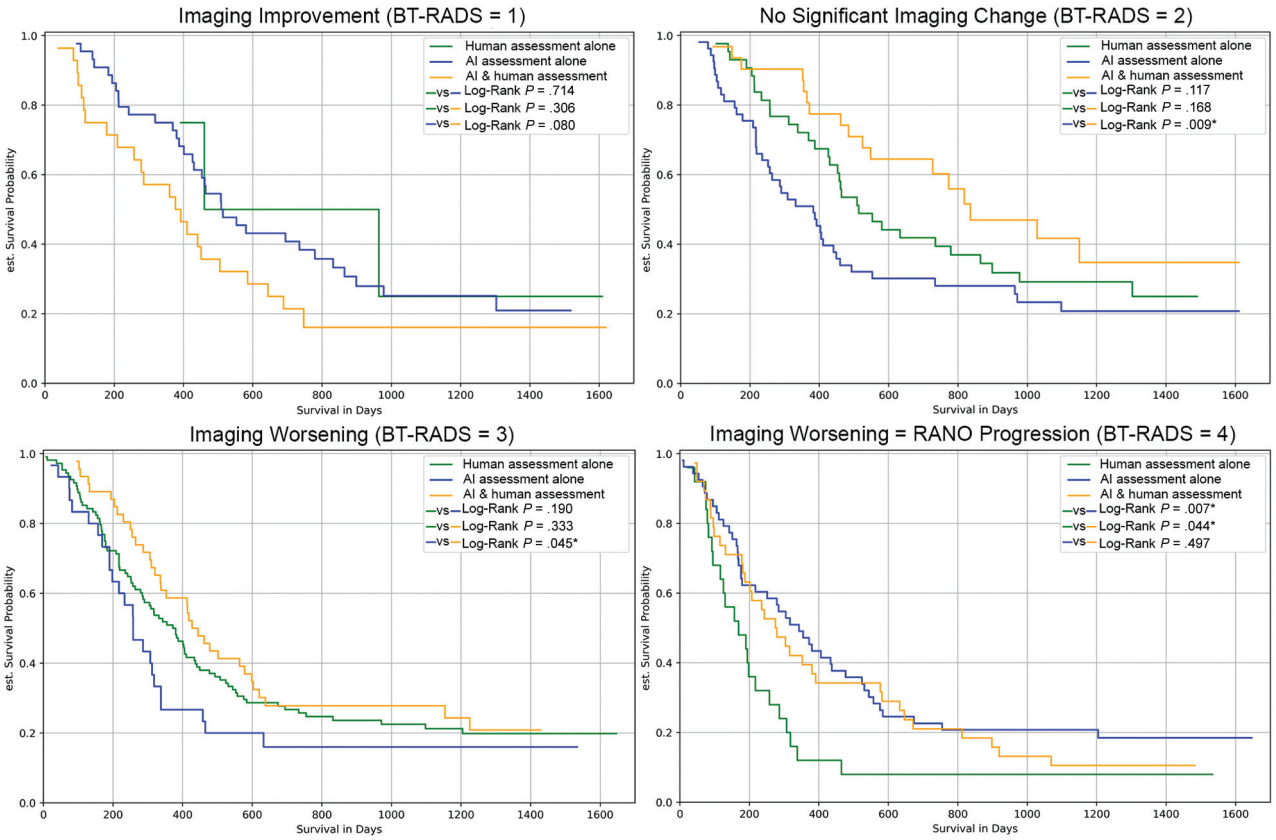


FIG 4. Fixed time point Kaplan-Meier survival curves for each response assessment category stratified by AI- and radiologist-based assessment methods. * Indicates a statistically significant difference.

Fixed Time Point Survival Analysis

Four hundred sixty-five of 634 (73%) patients died during the follow-up period. Median OS for the cohort was 443 days from the first available MRI examination, and median survival after the 6-month time point selected for the fixed time point survival analysis (S_{6mo}) was 401 days. Median S_{6mo} stratified by the highest human (BT-RADS) response category assessed during the first 6 months of follow-up was 401 days for BT-RADS 1, 625 days for BT-RADS 2, 394 days for BT-RADS 3, and 207 days for BT-RADS 4. Median S_{6mo} stratified by the highest AI (AI-VTRA) category assessed during the first 6 months of follow-up was 450 days for imaging improvement, 501 days for no significant change, 346 days for imaging worsening, and 305 days for image worsening equivalent to RANO progression. Survival curves for each BT-RADS and AI-VTRA category are presented in Fig 4. There was statistically worse overall S_{6mo} for patients assessed as image worsening equivalent to RANO progression by human alone compared with by AI alone (log-rank $P = .007$). For other assessment categories, S_{6mo} was not significantly different when assessed by AI alone versus human alone.

Multivariate Survival Modeling

A multivariate Cox proportional hazard model for S_{6mo} yielded a C-index for human assessments versus AI assessments (0.637 [0.600, 0.674] versus 0.594 [0.555, 0.633], $P = .012$), indicating significant improvement in predictive ability for human BT-RADS assessment. We showed hazard ratios and 95% CI of fitted fixed effects in Table 4 and Table 5 for BT-RADS and AI-VTRAS, respectively. Both models suggested that *Imaging RANO Progression* (score of 4) had significantly worse survival than *No change* (score of 2). The model that included BT-RADS suggested significantly worse survival in *Improving* (score of 1) and *Worsening* (score of 3) than *No change*.

DISCUSSION

The goal of this study was to compare AI-based volumetric GBM MRI response assessment with standardized radiologist response assessments. First, we addressed the ability of AI to recapitulate radiologist response assessments. Our results show that AI-based volumetric response assessment yielded overall moderate performance (Macro F1 $\approx$ 0.7) for recapitulating most human response assessment categories (BT-RADS 1, 2, and 4). Performance was lowest (Macro F1 $\approx$ 0.6) for predicting BT-RADS 3. This is likely related to the high variability of this assessment category, which ranges from minimal changes to relatively large tumor volume increases that do not meet the threshold for RANO progression. Prediction of this category is further complicated by the need to specify a volumetric threshold for "no significant change," which is incongruous with human response assessments where this threshold may differ depending on the clinical scenario. For example, radiologists may intuitively ignore T2/FLAIR signal attributed to posttreatment changes, whereas the volumetric segmentation model does not explicitly distinguish nonenhancing tumor versus treatment effect. These results suggest that

Table 4: Hazard ratios, confidence intervals, and *P* values of BT-RADS, age, sex, primary self-reported race, self-reported ethnicity, and IDH

Variable	Level	HR [95% CI]	*P* Value
BT-RADS	2 – No change	REF[a]	REF[a]
	1 – Improve	1.75 [1.05, 2.92]	.03
	3 – Worsening	1.48 [1.03, 2.13]	.03
	4 – RANO progression	2.62 [1.71, 4.00]	< .001
Age		1.12 [0.95, 1.30]	.16
Sex	Female	REF[a]	REF[a]
	Male	1.30 [0.99, 1.71]	.06
Primary self-reported race	White	REF[a]	REF[a]
	Black or African American	0.93 [0.56, 1.56]	.78
	Asian	0.48 [0.11, 2.04]	.32
	American Indian or Alaskan Native	0.93 [0.11, 8.00]	.95
	Not reported/declined	2.12 [0.54, 8.40]	.28
	Other	1.54 [0.31, 7.57]	.60
Self-reported ethnicity	Hispanic/Latino	REF[a]	REF[a]
	Not Hispanic/Latino	1.24 [0.57, 2.67]	.58
	Not reported/declined	N/A	N/A
IDH	Negative (wild-type)	REF[a]	REF[a]
	Positive (mutant)	0.69 [0.31, 1.51]	.35

[a] All subsequent comparisons are relative to this reference level.

Table 5: Hazard ratios, confidence intervals, and *P* values of AI-VTRA, age, sex, primary self-reported race, self-reported ethnicity, and IDH

Variable	Level	HR [95% CI]	*P* Value
AI-VTRA	2 – No change	REF[a]	REF[a]
	1 – Improve	1.26 [0.86, 1.87]	.24
	3 – Worsening	1.20 [0.81, 1.77]	.36
	4 – RANO progression	1.54 [1.07, 2.21]	.02
Age		1.11 [0.95, 1.30]	.18
Sex	Female	REF[a]	REF[a]
	Male	1.25 [0.95, 1.65]	.11
Primary self-reported race	White	REF[a]	REF[a]
	Black or African American	1.02 [0.61, 1.71]	.95
	Asian	0.45 [0.11, 1.94]	.28
	American Indian or Alaskan Native	1.29 [0.15, 11.12]	.82
	Not reported/declined	2.04 [0.51, 8.20]	.32
	Other	1.34 [0.27, 6.56]	.72
Self-reported ethnicity	Hispanic/Latino	REF[a]	REF[a]
	Not Hispanic/Latino	1.27 [0.59, 2.72]	.54
	Not reported/declined	N/A	N/A
IDH	Negative (wild-type)	REF[a]	REF[a]
	Positive (mutant)	0.68 [0.31, 1.51]	.34

[a] All subsequent comparisons are relative to this reference level.

AI-based volumetric response assessments may be better suited as a clinical decision support adjunct rather than a replacement for radiologists' assessments.[28] While different thresholds may ultimately be relevant for IDH mutant versus wild-type grade 4 tumors, they are currently treated the same by BT-RADS. The subgroup analysis of the IDH wild-type tumor data set (included in the section IDH Subgroup Analysis in Supplemental Data) showed that by using the same threshold, the composite metric AI-VTRA outperforms the AI-VTRA$_{ET}$ in both the IDH wild-type data set and the original data set.

A separate but related domain for evaluating AI-based volumetric response assessment is its ability to stratify patients by OS. Compared with a similar survival analysis study on BT-RADS stratification conducted by Kim et al,[39] our study reported the same nonsignificant hazard ratios for IDH status and significantly high hazard ratios for score 4. For all assessment categories other than BT-RADS 4, there was no statistically significant difference in OS, whether assigned by AI alone or human alone. However, OS was statistically lower for patients assessed as BT-RADS 4 by human alone compared with AI alone, with a median S_{6mo} of 207 versus 305 days, respectively. Interestingly, when BT-RADS 4 was assigned by both AI and human assessments, survival was more similar to those assigned by AI alone. In addition, though there were no statistically significant survival differences in other assessment categories whether assigned by AI alone versus human alone, human assessment resulted in larger differences in survival between assessment categories. These findings are indicative of the fact that human assessments often draw on additional findings or clinical history not captured by the proposed AI method, such as the progression of nonenhancing tumors in the setting of anti-angiogenic therapy. Incorporating this additional data will likely be important for improving AI-based GBM assessment methods in the future. We observed poorer S_{6mo} for BT-RADS 1 (401 days) than BT-RADS 2 (625 days), almost equivalent to BT-RADS 3 (394 days). We suspect that this might be due to bevacizumab pseudoresponse in patients with late-stage recurrent disease, which is consistent with the prior published report[40] on survival following BT-RADS assessment.

GBM MRI response assessments are highly complex owing to the highly variable appearance of recurrent tumor and treatment changes. There are several well-known issues with current response assessments that could be addressed with AI, including the inherent inaccuracies and high interrater variability of 2D measurements.[11,29-32] The results of this study add to a growing body of literature focused on AI-based GBM MRI response assessments,[28,33,34] which, like many applications of AI in neuro-oncology, have yet to deliver promised benefits in a meaningful way.[35] However, our results highlight 3 important observations: 1) simple rule-based AI volumetric response assessments yield only moderate performance for predicting human response assessments, 2) by using this approach, human assessments yielded a small but significant improvement in survival stratification performance, and 3) major discrepancies between human and AI assessments were rare, and both human and AI error were identified as causes. Overall, these results highlight the need for better AI models that can incorporate additional clinical and imaging variables into the response assessment. Though potentially incomplete segmentation of the lesion from the AI model

may contribute to the survival discrepancy between AI and human assessment, we do not believe that incomplete segmentation of tumors was a major factor in our study. Based on the defined rules for AI-VTRA and our evaluations, we believe that the primary factors are 1) a 25% 2D increase does not precisely correspond to a 40% volumetric increase and 2) the fact that human determination of "no significant change" does not necessarily correspond to any specific volumetric threshold.

Several prior studies have investigated automated brain tumor MRI segmentation as a means of assessing longitudinal tumor burden and even predicting time to progression and OS.[18,19,21,40,41] However, these studies have largely focused on automated volumetrics as an alternative to standard response assessment criteria rather than as a comprehensive method for automation of these criteria. For example, Kickingereder et al[21] compared brain tumor growth dynamics derived from automated segmentation with central RANO assessment for a longitudinal multi-institution cohort of 532 patients and found that the volumetric assessment was superior for predicting OS. However, to our knowledge, no prior work has evaluated automated volumetrics for predicting human BT-RADS scores or RANO progressive disease assessments. Our study differs from prior work in that it includes a larger number of patients and focuses on recapitulating human response assessments. This approach focuses on a paradigm of automating existing assessments rather than proposing new ones.

Pseudoprogression of GBM is a posttreatment phenomenon, with variable incidences from at least 9%, that can confuse the interpretations of tumor growth due to the pathology.[37] This study focuses exclusively on objective imaging change rather than subjective interpretation of the reason for this change. As such, the problem of pseudoprogression (and pseudoresponse) is not directly addressed and is a major limitation of this approach. Future work will be required to automate prediction of true versus pseudoprogression effectively and will likely require additional inputs such as treatment history and advanced imaging modalities like perfusion-weighted MRI.[37,38]

This study has several important limitations. First, this was a single-center retrospective study, which limits the generalization of its results. One generalization issue was the class imbalance favoring unchanged/improving conditions in our data set, which may lead to the under-representation of progression cases in this study. As an attempt to account for this issue, we reported multiple classification metrics. Second, this study used a relatively simplistic logic-based approach for assigning tumor volumetric differences to response categories. Third, this study relied on BT-RADS scores for radiologist-based response assessments. The BT-RADS system has been previously validated in several studies[13]; however, it is not yet as widely utilized as other response assessment criteria such as RANO. Fourth, though our NLP algorithm reached 99% accuracy in our internal validation, we would expect minor NLP- and human-induced errors in information retrieval from the reports, which may cause inexplicable discrepancies between AI versus human evaluations. Fifth, we did not include new lesions, which is part of the RANO progression criteria.[7] In future studies, we propose to apply a connected component algorithm to evaluate the growth of each separate lesion region and incorporate this analysis into the AI-VTRA rules. Finally, AI-based response assessments did not benefit from any information on treatment (such as radiation or antiangiogenic therapy), which fundamentally limits their ability to replicate radiologist-based response assessments.

CONCLUSIONS

AI-based volumetric GBM MRI response assessment following BT-RADS criteria can provide moderate performance for replicating human response assessments and show comparable performance for OS stratification. While this approach is unlikely to be useful for stand-alone response assessment, it may be useful for certain scenarios where radiologist interpretations are infeasible or as an adjunct to radiologist-based response assessment.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.

REFERENCES

1. Tamimi AF, Juweid M. **Epidemiology and outcome of glioblastoma**. In: De Vleeschouwer S, ed. *Glioblastoma*. Exon Publications; 2017 CrossRef
2. 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
3. Delgado-López PD, Corrales-García EM. **Survival in glioblastoma: a review on the impact of treatment modalities**. *Clin Transl Oncol* 2016;18:1062–71 CrossRef Medline
4. Marenco-Hillembrand L, Wijesekera O, Suarez-Meade P, et al. **Trends in glioblastoma: outcomes over time and type of intervention: a systematic evidence based analysis.** *J Neurooncol* 2020;147:297–307 CrossRef Medline
5. Reardon DA, Ballman KV, Buckner JC, et al. **Impact of imaging measurements on response assessment in glioblastoma clinical trials.** *Neuro Oncol* 2014;16 Suppl 7:vii24–35 CrossRef Medline
6. 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;28:1963–72 CrossRef Medline
7. Chukwueke UN, Wen PY. **Use of the Response Assessment in Neuro-Oncology (RANO) criteria in clinical trials and clinical practice.** *CNS Oncol* 2019;8:CNS28 CrossRef Medline
8. Ramakrishnan D, von Reppert M, Krycia M, et al. **Evolution and implementation of radiographic response criteria in neuro-oncology.** *Neurooncol Adv* 2023;5:vdad118 CrossRef Medline
9. Macdonald DR, Cascino TL, Schold SC, et al. **Response criteria for phase II studies of supratentorial malignant glioma.** *J Clin Oncol* 1990;8:1277–80 CrossRef Medline
10. Eisenhauer EA, Therasse P, Bogaerts J, et al. **New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1).** *Eur J Cancer* 2009;45:228–47 CrossRef Medline
11. Ellingson BM, Wen PY, Cloughesy TF. **Modified criteria for radiographic response assessment in glioblastoma clinical trials.** *Neurotherapeutics* 2017;14:307–20 CrossRef Medline
12. Weinberg BD, Gore A, Shu H-KG, et al. **Management-based structured reporting of posttreatment glioma response with the Brain Tumor Reporting and Data System.** *J Am Coll Radiology* 2018; 15:767–71 CrossRef Medline
13. Gore A, Hoch MJ, Shu H-KG, et al. **Institutional implementation of a structured reporting system: our experience with the brain tumor reporting and data system.** *Acad Radiology* 2019;26:974–80 CrossRef Medline
14. Zhang JY, Weinberg BD, Hu R, et al. **Quantitative improvement in brain tumor MRI through structured reporting (BT-RADS).** *Acad Radiology* 2020;27:780–84 CrossRef Medline

15. Chappell R, Miranpuri SS, Mehta MP. **Dimension in defining tumor response.** *J Clin Oncol* 1998;16:1234 CrossRef Medline
16. 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
17. Pati S, Baid U, Edwards B, et al. **Federated learning enables big data for rare cancer boundary detection.** *Nat Commun* 2022;13:7346 CrossRef Medline
18. Rudie JD, Calabrese E, Saluja R, et al. **Longitudinal assessment of posttreatment diffuse glioma tissue volumes with three-dimensional convolutional neural networks.** *Radiology Artif Intell* 2022;4: e210243 CrossRef Medline
19. Chang K, Beers AL, Bai HX, et al. **Automatic assessment of glioma burden: a deep learning algorithm for fully automated volumetric and bidimensional measurement.** *Neuro Oncol* 2019;21:1412–22 CrossRef Medline
20. Suter Y, Notter M, Meier R, et al. **Evaluating automated longitudinal tumor measurements for glioblastoma response assessment.** *Front Radiology* 2023;3:1211859 CrossRef Medline
21. 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;20:728–40 CrossRef Medline
22. Ellingson BM, Bendszus M, Boxerman J; Jumpstarting Brain Tumor Drug Development Coalition Imaging Standardization Steering Committee, et al. **Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials.** *Neuro Oncol* 2015; 17:1188–98 CrossRef Medline
23. Mazziotta JC, Toga AW, Evans A, et al. **A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM).** *Neuroimage* 1995;2:89–101 CrossRef Medline
24. Pati S, Baid U, Edwards B, et al. **The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research.** *Phys Med Biol* 2022;67:204002 CrossRef
25. Gahrmann R, van den Bent M, van der Holt B, et al. **Comparison of 2D (RANO) and volumetric methods for assessment of recurrent glioblastoma treated with bevacizumab—a report from the BELOB trial**. *Neuro Oncol* 2017;19:853–61 CrossRef Medline
26. Wang M-Y, Cheng J-L, Han Y-H, et al. **Measurement of tumor size in adult glioblastoma: classical cross-sectional criteria on 2D MRI or volumetric criteria on high resolution 3D MRI?** *Eur J Radiology* 2012;81:2370–74 CrossRef Medline
27. Wen PY, van den Bent M, Youssef G, et al. **RANO 2.0: update to the Response Assessment in Neuro-Oncology criteria for high- and low-grade gliomas in adults.** *J Clin Oncol* 2023;41:5187–99 CrossRef Medline
28. Vollmuth P, Foltyn M, Huang RY, et al. **Artificial intelligence (AI)-based decision support improves reproducibility of tumor response assessment in neuro-oncology: an international multi-reader study.** *Neuro Oncol* 2023;25:533–43 CrossRef Medline
29. Vos MJ, Uitdehaag BMJ, Barkhof F, et al. **Interobserver variability in the radiological assessment of response to chemotherapy in glioma.** *Neurology* 2003;60:826–30 CrossRef Medline
30. Galanis E, Buckner JC, Maurer MJ, et al. **Validation of neuroradiologic response assessment in gliomas: measurement by RECIST, two-dimensional, computer-assisted tumor area, and computer-assisted tumor volume methods.** *Neuro Oncol* 2006;8:156–65 CrossRef Medline
31. Dempsey MF, Condon BR, Hadley DM. **Measurement of tumor "size" in recurrent malignant glioma: 1D, 2D, or 3D?** *AJNR Am J Neuroradiol* 2005;26:770–76 Medline
32. Yang D. **Standardized MRI assessment of high-grade glioma response: a review of the essential elements and pitfalls of the RANO criteria.** *Neurooncol Pract* 2016;3:59–67 CrossRef Medline
33. Ellingson BM. **On the promise of artificial intelligence for standardizing radiographic response assessment in gliomas.** *Neuro Oncol* 2019;21:1346–47 CrossRef Medline
34. Sotoudeh H, Shafaat O, Bernstock JD, et al. **Artificial intelligence in the management of glioma: era of personalized medicine.** *Front Oncol* 2019;9:768 CrossRef Medline
35. Rudie JD, Rauschecker AM, Bryan RN, et al. **Emerging applications of artificial intelligence in neuro-oncology.** *Radiology* 2019;290: 607–18 CrossRef Medline
36. Isensee F, Jaeger PF, Kohl SA, et al. **nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.** *Nat Methods* 2021;18:203–11 Medline
37. Thust SC, van den Bent MJ, Smits M. **Pseudoprogression of brain tumors.** *J Magn Reson Imaging* 2018;48:571–89
38. Linhares P, Carvalho B, Figueiredo R, et al. **Early pseudoprogression following chemoradiotherapy in glioblastoma patients: the value of RANO evaluation.** *J Oncol* 2013;2013:690585 CrossRef Medline
39. Kim S, Hoch MJ, Peng L, et al. **A brain tumor reporting and data system to optimize imaging surveillance and prognostication in high-grade gliomas.** *J Neuroimaging* 2022;32:1185–92 CrossRef Medline
40. Bianconi A, Rossi LF, Bonada M, et al. **Deep learning-based algorithm for postoperative glioblastoma MRI segmentation: a promising new tool for tumor burden assessment.** *Brain Inform* 2023;10: 26 CrossRef Medline
41. Bangalore Yogananda CG, Wagner B, Nalawade SS, et al. **Fully automated brain tumor segmentation and survival prediction of gliomas using deep learning and MRI**. In: Crimi A, Bakas S, eds. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. BrainLes 2019. Lecture Notes in Computer Science. Springer; 2020;11993:99–112