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/2024/11/14/ajnr.A8580

Published November 14, 2024 as 10.3174/ajnr.A8580

ORIGINAL RESEARCH

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

Jikai Zhang [1, 3], Dominic LaBella [4], Dylan Zhang [6], Jessica L. Houk [6], Jeffrey D. Rudie [7], Haotian Zou[5], Pranav Warman[8], Maciej A. Mazurowski [1, 2, 5, 6], Evan Calabrese [3, 6]

ABSTRACT

BACKGROUND AND PURPOSE: 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.

MATERIALS AND METHODS: 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 algorithm was developed using automated, volumetric tumor segmentation. AI-based response assessments 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.

RESULTS: For all BT-RADS categories, AI-based response assessments showed moderate agreement with radiologists' 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=0.007). Cox proportional hazard model analysis showed a disadvantage to AI-based assessments for overall survival prediction (P=0.012).

CONCLUSIONS: AI-based volumetric glioblastoma MRI response assessment following BT-RADS criteria yielded moderate agreement for replicating human response assessments and slightly worse stratification by overall survival.

ABBREVIATIONS: GBM = Glioblastoma; RANO = Response Assessment in Neuro-Oncology; BTRADS = Brain Tumor Reporting and Data System; NLP = Natural Language Processing.

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

[1] Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA
[2] Department of Computer Science, Duke University, Durham, NC, USA
[3] Duke Center for Artificial Intelligence in Radiology, Duke University Medical Center, Durham, NC, USA
[4] Department of Radiation Oncology, Duke University Medical Center, Durham, NC, USA
[5] Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA
[6] Department of Radiology, Duke University Medical Center, Durham, NC, USA
[7] Department of Radiology, University of California San Diego, San Diego, CA, USA
[8] Duke University School of Medicine, Durham, NC, USA

We declare no conflict of interest.

Please address correspondence to Evan Calabrese, MD PhD, Department of Radiation Oncology, Duke University Medical Center, DUMC Box 3808; Durham, NC 27710; USA; evan.calabrese@duke.edu.

SUMMARY SECTION

PREVIOUS LITERATURE: In recent years, there have been extensive efforts to develop standardized MRI response assessment criteria, including RANO and BT-RADS. However, human assessments are prone to error, bias, and inter-rater variability. Artificial intelligence (AI)-based brain tumor MRI segmentation models have potential to improve response assessments through objective volumetric assessment. However, there has been limited prior work focused on developing automated AI-based response assessment tools for glioblastoma incorporating established response assessment criteria.

KEY FINDINGS: AI-based automated volumetric response assessments showed moderate agreement with radiologists' response assessment and comparable survival stratification for patients with glioblastomas. Time-dependent cox proportional hazards models showed similar performance of human and AI assessments for predicting overall survival.

KNOWLEDGE ADVANCEMENT: AI-based volumetric response assessments on MRI may help automate and standardize glioblastoma response assessments. This approach may be particularly useful for certain scenarios where radiologist interpretations are infeasible or as an adjunct to radiologist-based response assessment.

Copyright 2024 by American Society of Neuroradiology.

INTRODUCTION

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 two-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 consideration of both enhancing and non-enhancing tumor 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 a unified criteria to assess gliomas regardless of their grades and recommend volumetric assessments.[39]

More recent efforts towards 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 non-enhancing tumor, 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-grad gliomas, while BT-RADS was designed for high-grade gliomas. Both criteria acknowledge changes in enhancing and non-enhancing tumor, 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 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 three-dimensional (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 due to large collaborative efforts such as the multimodal brain tumor segmentation (BraTS) 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 two key domains: ability to recapitulate human response assessments and ability to stratify patients by overall survival.

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 using Center for Medicaid Services (CMS) Hierarchical Condition Category (HCC) codes (n = 4,689). This included both isocitrate dehydrogenase mutant and wildtype (WT) grade 4 astrocytomas in line with current WHO classifications at the time of diagnosis (referred to as "GBM" hence forth for conciseness). Exclusion criteria were: patients lacking at least one MRI brain exam with and without intravenous contrast (n = 3,199), and patients lacking at least one standardized neuroradiologist response assessment (n = 856). The final study population consisted of 634 patients. A patient flow diagram is provided as FIG 1.

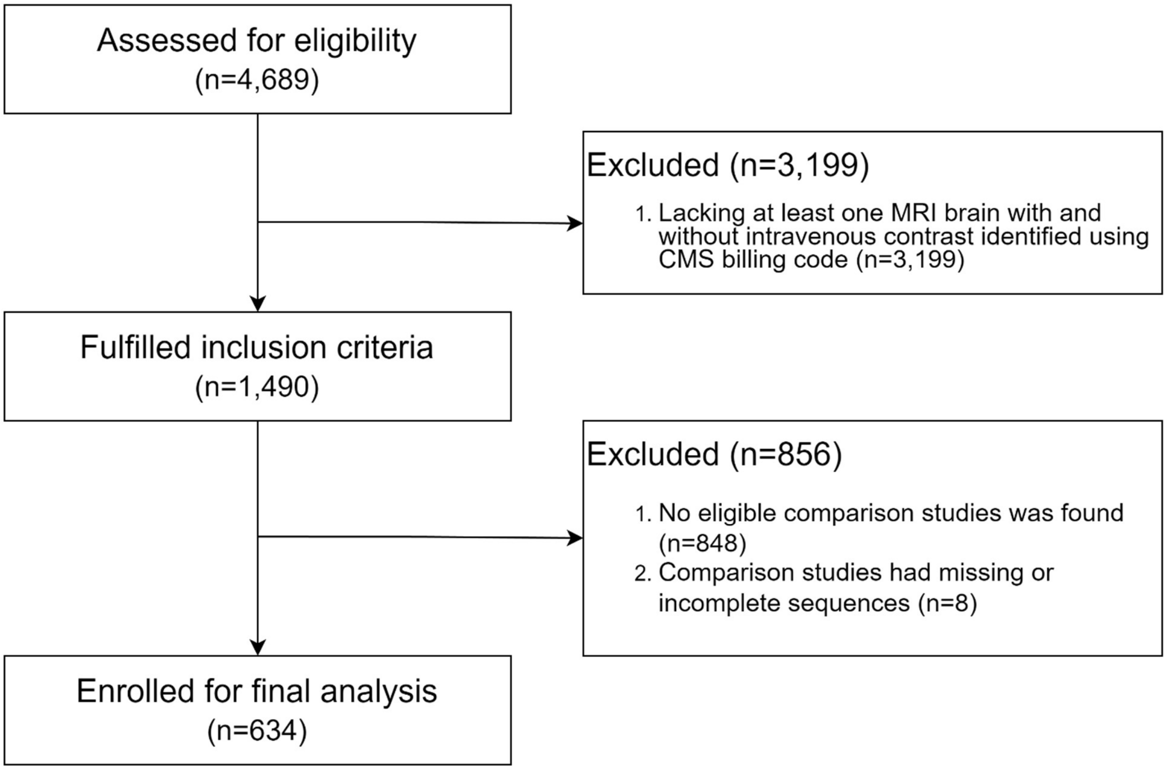
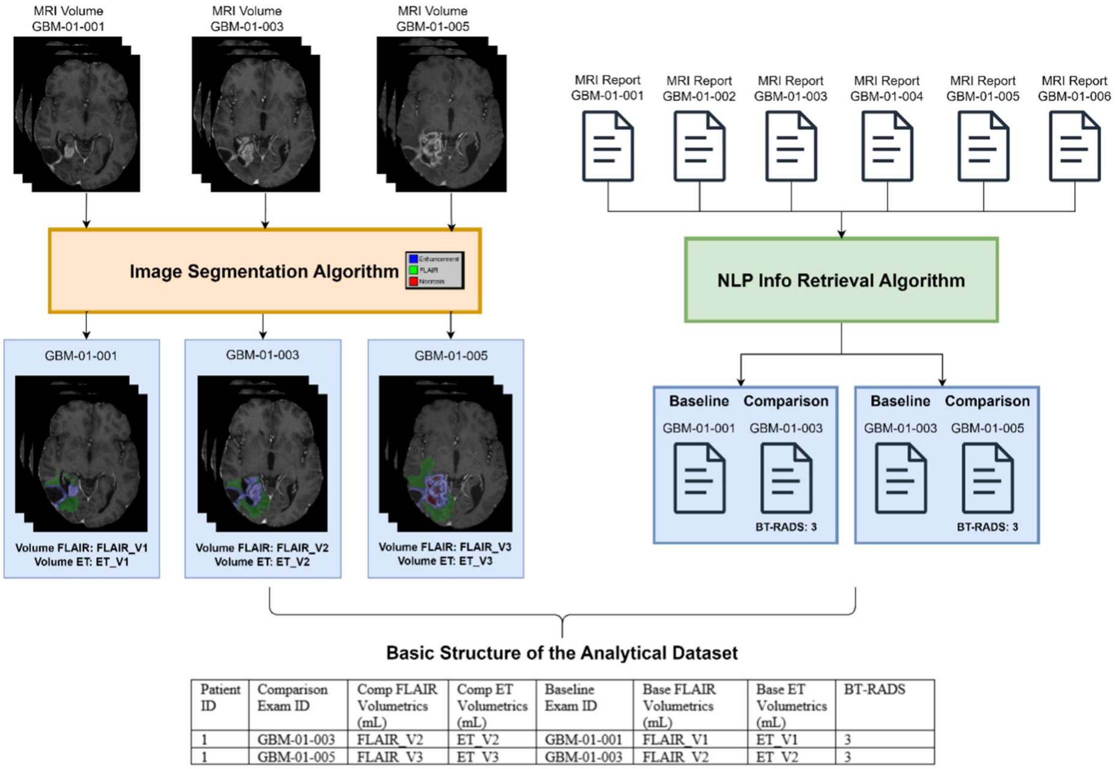

Assessed for eligibility
(n=4,689)
Excluded (n=3,199)
1. Lacking at least one MRI brain with and without intravenous contrast identified using CMS billing code (n=3,199)
Fulfilled inclusion criteria
(n=1,490)
Excluded (n=856)
1. No eligible comparison studies was found (n=848)
2. Comparison studies had missing or incomplete sequences (n=8)
Enrolled for final analysis
(n=634)

FIG 1. Patient flow diagram for study inclusion

Neuroradiologists response assessments

Formal neuroradiologist-based GBM MRI response assessments using the BT-RADS structured reporting system were available as part of routine clinical care. BT-RADS scores and baseline comparison exam dates were extracted from radiology reports using a custom semi-supervised natural language processing (NLP) algorithm with near-perfect internal validation performance. The full data curation pipeline was demonstrated 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 exam date and searched for its prior exam with the retrieved exam date (complete methodologic details and performance assessment provided as Supplementary Data). This yielded 2,446 pairs of exams (current and baseline prior) with BT-RADS scores. One baseline prior can be paired with multiple follow-up exam. BT-RADS scores included the following numerical categories: 1 = imaging improvement, 2 = no significant imaging change, 3 = imaging worsening, 4 = imaging worsening with >25% increase in 2-dimensional 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 exams were performed with a Brain Tumor Imaging Protocol (BTIP)[22] compliant protocol including 3D, gradient echo, T1-weighted pre- and post-contrast sequences and 2D, T2-weighted and T2-Fluid Attenuated Inversion Recovery (FLAIR) sequences. MRI data were retrieved for each pair of exams corresponding to the BT-RADS scores identified in the previous section, which resulted in 3,403 unique MRI exams. Scanner information was included in the Supplementary 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 filed-of-view (FOV) standardization,[23] and skull stripping using a publicly available deep learning method.[24] Preprocessed images then underwent automated, volumetric tumor segmentation using 3D convolutional segmentation neural network. This model was specifically designed for post-treatment exams including four distinct compartments: resection cavity, enhancing tumor, necrotic tumor core, and surrounding non-enhancing T2-FLAIR signal abnormality. The final model was pre-trained on an external post-operative brain MRI exam. We utilized nnU-Net[36] to train and validate the model. Internal validation results showed a mean +/- standard deviation of 0.8861 +- 0.2476 for enhancing tumor and 0.9833 +- 0.0372 for surrounding non-enhancing FLAIR signal abnormality (complete methodologic details and performance assessment provided as Supplementary 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 non-enhancing FLAIR hyperintensity (VD_{FLAIR}) were computed for each pair of exams in the dataset and were used to develop AI-based volumetric equivalents to BT-RADS scores. BT-RADS 4 was defined as ≥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} (Supplementary Data).

AI performance for survival stratification

Performance of automated volumetric criteria for replicating human BT-RADS scores was evaluated across the entire dataset. 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 recapitulating human BT-RADS scores

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 exam. Time from initial diagnosis was not available for all patients and was not included in the analysis. 323/634 (51%) patients had at least 1 BT-RADS assessment in the first 6 months of follow up and were included in this sub-analysis. This cohort was sub-stratified 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 sub-strata 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 exams was considered as the time-varying covariate in the cox model. We removed observations due to unknown IDH status before fitting the cox models. Concordance index (C-index) was calculated for each Cox model. To compare the difference in C-index between two cox models, we applied statistical tests that account for the paired data (see Supplementary Data for details).

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

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

Statistical analyses

Statistical analyses were performed in Python version 3.8 and R version 4.2. Kaplan-Meier estimates were computed using the "lifelines" package in Python. Cox modeling was performed in R 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 0.05 were considered significant.

RESULTS

Patient characteristics

Basic study participant demographic data are reported in Table 2. The mean age was 56 +/- 13 years. 372/634 (59%) patients were male, and 262/634 (41%) patients were female. 566/634 (89%) patients listed their primary self-reported race as Caucasian/white, 41/634 (7%) as black or African American, and 9/634 (1%) as Asian. 8/634 (1%) of patients reported a secondary race, and 10/634 (2%) patients did not report race. 479/634 (76%) patients had an IDH wildtype tumor, 63/634 (10%) patients had an IDH mutant tumor, and 92/634 (14%) patients had missing or inconclusive IDH testing.

MRI image data and segmentation

The 634 included patients had 3,403 qualifying MRI brain exams (average 3.85 exams per patient). The average time between baseline and follow-up studies was 160 days with a standard deviation of 236 days. Automated volumetric tumor segmentation was successfully completed for all exams without errors. The average segmentation time was 11.5 seconds per exam. Representative segmented MRI images from four different patient's exam pairs with each of the different assessment categories are presented in FIG 3.

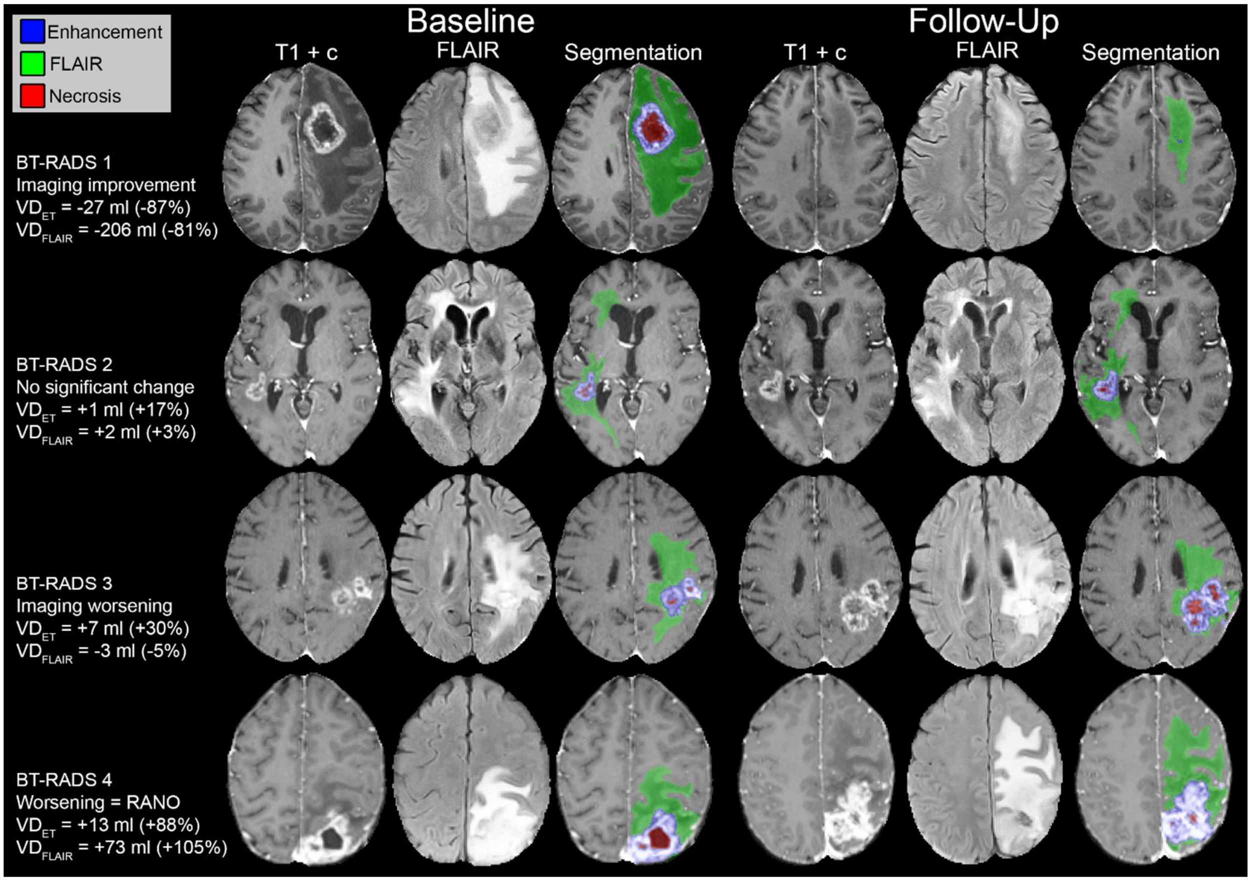


FIG 3. Example MR images, radiologist response assessment categories, and volumetric changes for 4 patients at 2 different timepoints.

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 to 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-VTRAET alone demonstrated improved performance compared to 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 percentage for each score and an analysis of major discrepancies between human and AI assessments are provided in Supplementary Data.

Fixed timepoint survival analysis

465/634 (73%) patients died during the follow-up period. Median overall survival (OS) for the cohort was 443 days from the first available MRI exam, and median survival after the 6-month timepoint selected for the fixed timepoint 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 to by AI alone (log-rank $p = 0.007$). For other assessment categories, S_{6mo} was not significantly different when assessed by AI alone versus human alone.

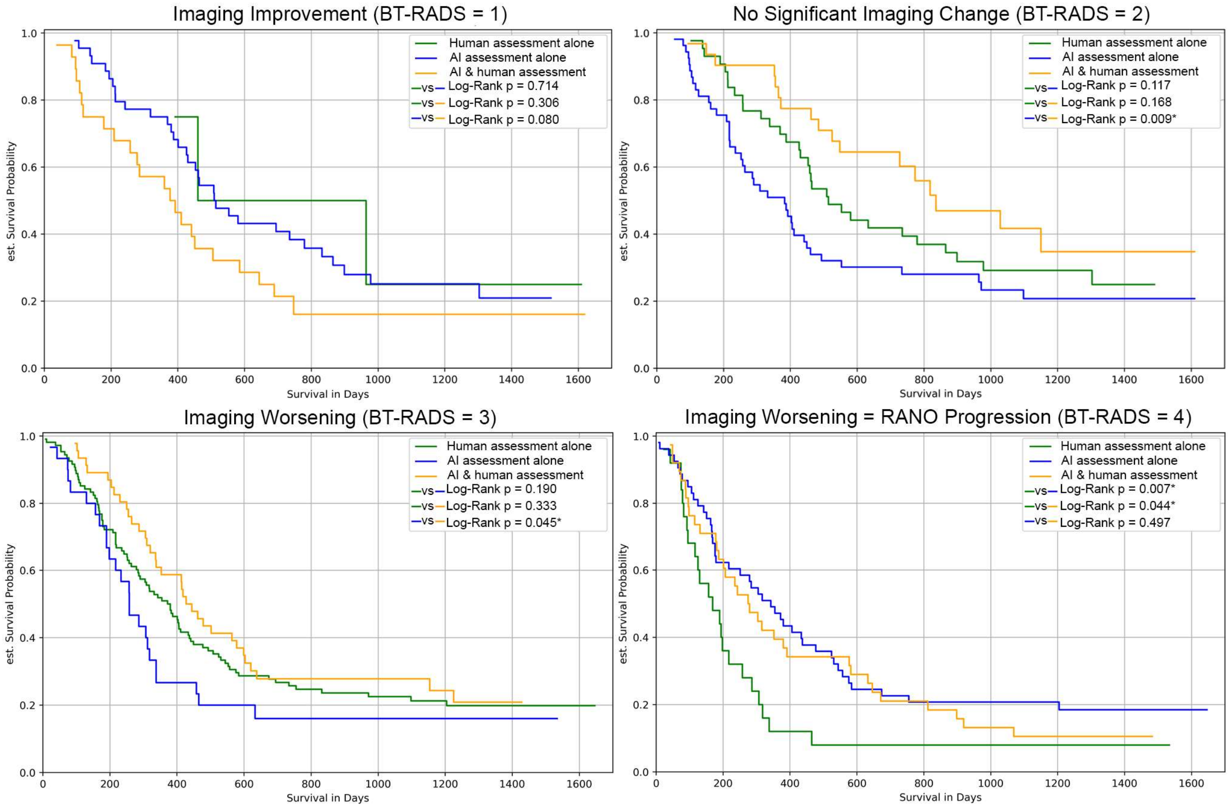


FIG 4. Fixed timepoint Kaplan-Meier survival curves for each response assessment category stratified by AI- and radiologist-based assessment methods. * Indicates a statistically significant difference.

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] vs 0.594 [0.555, 0.633], p-value =0.012), indicating significant improvement in predictive ability for human BT-RADS assessment. We showed hazard ratios and 95% confidence intervals 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 ≈ 0.7) for recapitulating most human response assessment categories (BT-RADS 1, 2, and 4). Performance was lowest (Macro F1 ≈ 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 post-treatment changes, whereas the volumetric segmentation model does not explicitly distinguish non-enhancing tumor versus treatment effect. These results suggest that 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 wildtype grade 4 tumors, they are currently treated the same by BT-RADS. The subgroup analysis of IDH WT tumor dataset (included in the section IDH Subgroup Analysis in Supplementary Material) showed that using the same threshold, that the composite metric AI-VTRA outperforms AI-VTRAET in both IDH WT dataset and the original dataset.

A separate but related domain for evaluating AI-based volumetric response assessment is its ability to stratify patients by overall survival. Compared to a similar survival analysis study on BT-RADS stratification conducted by Kim et al.[39], our study reported the same non-significant 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 overall survival whether assigned by AI alone or human alone. However, overall survival was statistically lower for patients assessed as BT-RADS 4 by human alone compared to by 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, although 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 progression of non-enhancing tumor 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 pseudo-response 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 inter-rater 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 three important observations, 1) simple rule-based AI volumetric response assessments yield only moderate performance for predicting human response assessments, 2) 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. Although 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 of 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 overall survival[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 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 overall survival[21]. 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 post-treatment 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 pseudo-progression (and pseudo-response) is not directly addressed and is a major limitation of this approach. Future work will be required to effectively automate prediction of true versus pseudo-progression 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 generalization of its results. One generalization issue was the class imbalance favoring unchanged/improving conditions in our dataset, which may lead to 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, although our NLP algorithm reached 99% accuracy in our internal validation, we would expect minor NLP- and human-induced errors of information retrieval from the reports, which may cause inexplicable discrepancies between AI vs human evaluations. Fifth, we did not include new lesion, 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 to the AI-VTRA rules. Finally, AI-based response assessments did not benefit from any information on treatment (such as radiation or anti-angiogenic therapy), which fundamentally limits their ability to replicate radiologist-based response assessments.

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

Assessment Category	Assessment System (rater)	
	BT-RADS (human)	AI-VTRA (AI)
Imaging Improvement	1	$VD_{ET} \leq -10\%$ OR $-10\% > VD_{ET} < 10\%$ AND $VD_{FLAIR} \leq -40\%$
No significant imaging change	2	$-10\% > VD_{ET} < 10\%$

Imaging worsening	3	OR $VD_{ET} \leq -10\%$ AND $VD_{FLAIR} \geq 40\%$ $10\% \leq VD_{ET} < 40\%$ OR $-10\% > VD_{ET} < 10\%$ AND $VD_{FLAIR} \geq 40\%$
Imaging worsening equivalent to RANO progression	4	$VD_{ET} \geq 40\%$

* Detailed rules of determining AI-VTRA are included in **Supplementary Data**.

Table 2: Basic demographics for the 634 patients included in the study cohort. Patient age was assessed at the time of the first available MRI brain exam date. SD is standard deviation. "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 non-enhancing tumors (at least 1 milliliters) across the 3403 exams in the study cohort.

Demographic Characteristics	
Age (yr)	
Mean	56
Standard Deviation	14
Sex N (%)	
Male	372 (59%)
Female	262 (41%)
Primary Self-Reported Race N (%)	
Caucasian/White	566 (89%)
Black or African American	41 (7%)
Asian	9 (1%)
Other	8 (1%)
Not Responded/Declined	10 (2%)
Self-Reported Ethnicity N (%)	
Not Hispanic	592 (93%)
Hispanic	11 (5%)
Not Reported/Declined	31 (2%)
IDH N (%)	
Wildtype	479 (76%)
Mutant	63 (10%)
Inconclusive/Missing	92 (14%)
Tumor Types N (%)	
Enhancing	2,163 (64%)
Non-enhancing Edema/FLAIR	3,401 (99%)

Table 3: Performance metrics (Macro-F1, Micro-F1, sensitivity, specificity, and precision) for AI-VTRA/AI-VTRAET 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*

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	REF
	1 - Improve	1.75 [1.05, 2.92]	0.03
	3 - Worsening	1.48 [1.03, 2.13]	0.03
	4 – RANO Progression	2.62 [1.71, 4.00]	< 0.001
Age		1.12 [0.95, 1.30]	0.16
Sex	Female	REF	REF
	Male	1.30 [0.99, 1.71]	0.06
Primary Self-reported Race	Caucasian	REF	REF
	Black or African American	0.93 [0.56, 1.56]	0.78
	Asian	0.48 [0.11, 2.04]	0.32
	American Indian or Alaskan Native	0.93 [0.11, 8.00]	0.95
	Not Reported/Declined	2.12 [0.54, 8.40]	0.28
	Other	1.54 [0.31, 7.57]	0.60
Self-Reported Ethnicity	Hispanic/Latino	REF	REF
	Not Hispanic/Latino	1.24 [0.57, 2.67]	0.58
	Not Reported/Declined	N/A	N/A
IDH	Negative (WT)	REF	REF
	Positive (Mutant)	0.69 [0.31, 1.51]	0.35

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	REF
	1 - Improve	1.26 [0.86, 1.87]	0.24
	3 - Worsening	1.20 [0.81, 1.77]	0.36
	4 – RANO Progression	1.54 [1.07, 2.21]	0.02
Age		1.11 [0.95, 1.30]	0.18
Sex	Female	REF	REF
	Male	1.25 [0.95, 1.65]	0.11
Primary Self-reported Race	Caucasian	REF	REF
	Black or African American	1.02 [0.61, 1.71]	0.95
	Asian	0.45 [0.11, 1.94]	0.28
	American Indian or Alaskan Native	1.29 [0.15, 11.12]	0.82
	Not Reported/Declined	2.04 [0.51, 8.20]	0.32
	Other	1.34 [0.27, 6.56]	0.72
Self-Reported Ethnicity	Hispanic/Latino	REF	REF
	Not Hispanic/Latino	1.27 [0.59, 2.72]	0.54
	Not Reported/Declined	N/A	N/A
IDH	Negative (WT)	REF	REF
	Positive (Mutant)	0.68 [0.31, 1.51]	0.34

CONCLUSIONS

In conclusion, 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 overall survival stratification. While this approach is unlikely to be useful for standalone response assessment, it may be useful for certain scenarios where radiologist interpretations are infeasible or as an adjunct to radiologist-based response assessment.

ACKNOWLEDGMENTS

Acknowledgments have to be in a separate section but it does not have to be a freestanding page. Obtaining permission to include a name in this section from the acknowledged individual is advised.

REFERENCES

1. Tamimi AF, Juweid M. Epidemiology and Outcome of Glioblastoma. Exon Publications https://doi.org/10.15586/codon.glioblastoma.2017.ch8.
2. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro-Oncology 2021;23:1231–51.
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.
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.
5. Reardon DA, Ballman KV, Buckner JC, et al. Impact of imaging measurements on response assessment in glioblastoma clinical trials. Neuro-Oncology 2014;16:vii24–35.
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.
7. Chukwueke UN, Wen PY. Use of the Response Assessment in Neuro-Oncology (RANO) criteria in clinical trials and clinical practice. CNS Oncology 2019;8:CNS28.
8. Ramakrishnan D, von Reppert M, Krycia M, et al. Evolution and implementation of radiographic response criteria in neuro-oncology. Neuro-Oncology Advances 2023;5:vdad118.
9. Macdonald DR, Cascino TL, Schold SC, et al. Response criteria for phase II studies of supratentorial malignant glioma. JCO 1990;8:1277–80.
10. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). European Journal of Cancer 2009;45:228–47.
11. Ellingson BM, Wen PY, Cloughesy TF. Modified Criteria for Radiographic Response Assessment in Glioblastoma Clinical Trials. Neurotherapeutics 2017;14:307–20.
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. Journal of the American College of Radiology 2018;15:767–71.
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 Radiol 2019;26:974–80.
14. Zhang JY, Weinberg BD, Hu R, et al. Quantitative Improvement in Brain Tumor MRI Through Structured Reporting (BT-RADS). Academic Radiology 2020;27:780–4.
15. Chappell R, Miranpuri SS, Mehta MP. Dimension in defining tumor response. J Clin Oncol 1998;16:1234.
16. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging 2015;34:1993–2024.
17. Pati S, Baid U, Edwards B, et al. Federated learning enables big data for rare cancer boundary detection. Nat Commun 2022;13:7346.
18. Rudie JD, Calabrese E, Saluja R, et al. Longitudinal Assessment of Posttreatment Diffuse Glioma Tissue Volumes with Three-dimensional Convolutional Neural Networks. Radiol Artif Intell 2022;4:e210243.
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-Oncology 2019;21:1412–22.
20. Suter Y, Notter M, Meier R, et al. Evaluating automated longitudinal tumor measurements for glioblastoma response assessment. Front Radiol 2023;3:1211859.
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.
22. Ellingson BM, Bendszus M, Boxerman J, et al. Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials. Neuro Oncol 2015;17:1188–98.
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.
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.
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-Oncology 2017;19:853–61.
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 Radiol 2012;81:2370–4.
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.
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-Oncology 2023;25:533–43.
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.
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 methods1. Neuro-Oncology 2006;8:156–65.
31. Dempsey MF, Condon BR, Hadley DM. Measurement of Tumor "Size" in Recurrent Malignant Glioma: 1D, 2D, or 3D? American Journal of Neuroradiology 2005;26:770–6.
32. Yang D. Standardized MRI assessment of high-grade glioma response: a review of the essential elements and pitfalls of the RANO criteria. Neuro-

Oncology Practice 2016;3:59–67.

33. Ellingson BM. On the promise of artificial intelligence for standardizing radiographic response assessment in gliomas. Neuro-Oncology 2019;21:1346–7.
34. Sotoudeh H, Shafaat O, Bernstock JD, et al. Artificial Intelligence in the Management of Glioma: Era of Personalized Medicine. Frontiers in Oncology 2019;9.
35. Rudie JD, Rauschecker AM, Bryan RN, et al. Emerging Applications of Artificial Intelligence in Neuro-Oncology. Radiology 2019;290:607–18.
36. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods. 2021 Feb;18(2):203-11.
37. Thust SC, van den Bent MJ, Smits M. Pseudoprogression of brain tumors. J Magn Reson Imaging. Published online May 7, 2018. doi:10.1002/jmri.26171
38. Linhares P, Carvalho B, Figueiredo R, Reis RM, Vaz R. Early pseudoprogression following chemoradiotherapy in glioblastoma patients: the value of RANO evaluation. Journal of Oncology. 2013;2013(1):690585.
39. Kim S, Hoch MJ, Peng L, Somasundaram A, Chen Z, Weinberg BD. A brain tumor reporting and data system to optimize imaging surveillance and prognostication in high-grade gliomas. J Neuroimaging. 2022 Nov;32(6):1185-1192. doi: 10.1111/jon.13044. Epub 2022 Aug 31. PMID: 36045502.
40. Bianconi A, Rossi LF, Bonada M, Zeppa P, Nico E, De Marco R, Lacroce P, Cofano F, Bruno F, Morana G, Melcarne A. Deep learning-based algorithm for postoperative glioblastoma MRI segmentation: a promising new tool for tumor burden assessment. Brain Informatics. 2023 Dec;10(1):26.
41. Bangalore Yogananda CG, Wagner B, Nalawade SS, Murugesan GK, Pinho MC, Fei B, Madhuranthakam AJ, Maldjian JA. Fully automated brain tumor segmentation and survival prediction of gliomas using deep learning and MRI. InInternational MICCAI Brainlesion Workshop 2020 (pp. 99-112). Springer, Cham.

SUPPLEMENTAL FILES

Additional Details of the Initial Cohort

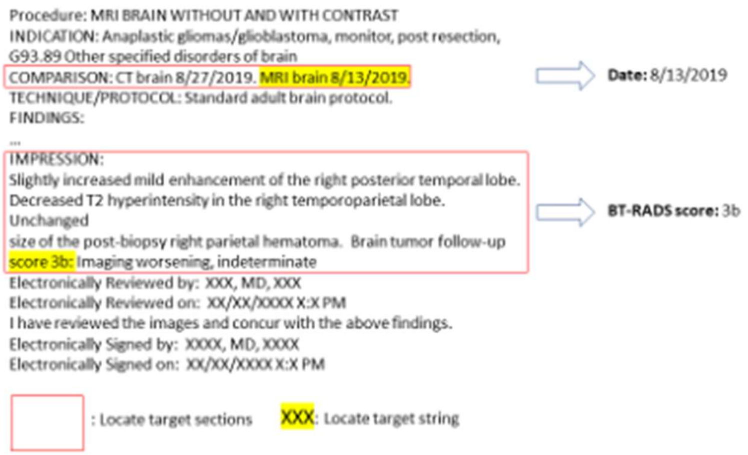
Our institution has implemented routine BT-RADS assessment for all high-grade adult gliomas since 2018. After the initial adoption period, all neuro-radiologists at our institution started using BT-RADS routinely. In this cohort specifically, we counted 8310 radiology reports in total, 3570 (43%) reports had BT-RADS assessments (identified by either human annotations or the NLP algorithms), and 22 individual radiologists generated these BT-RADS assessments. From the original cohort, the average follow-up days to the prior were 74 days.

Natural Language Processing Algorithm Methods and Evaluation

We developed a custom rule-based Natural Language Processing (NLP) algorithm to retrieve BT-RADS score and comparison date from unstructured MRI reports. The NLP algorithm takes the unstructured brain MRI report as the input and provides a structured output, including pairs of baseline comparison and follow-up studies and assigned BT-RADS scores. Our institutional report has a specific section to document the baseline prior, and that baseline is derived from the treating neuro-oncologists' clinical notes in the electronic health record. The specified baseline may be the immediate prior or the exam closest to the last change in treatment. BT-RADS scores are always in reference to the baseline prior regardless of whether there are other interval exams.

To develop the NLP algorithm, we randomly sampled 450 patients with 2444 reports for the algorithm development, 150 patients with 886 reports for validation, and 89 patients with 452 reports for testing.

The NLP algorithm consisted of three modules to simulate the process of manually retrieving data from the unstructured report. Supplementary Figure 1 shows a brief illustration of the first two steps of the NLP algorithm. First, the algorithm retrieved BT-RADS score by searching the impression section of the report. The string next to the text "score" was extracted and processed to match with the BT-RADS scoring system. If the algorithm detected more than one score in the impression, no result was provided. Second, in baseline and comparison sections, the algorithm retrieved the date of the comparison exam using regular expression-based fuzzy logic date identification. In cases where a baseline comparison was not specified, we selected the most recent MRI exam date in listed in the comparison section. Comparison dates corresponding to CT or reference only exams were excluded. Finally, the existence of the baseline comparison study was confirmed by identifying prior MRI exams performed on the same patient on the extracted baseline date. From the test set, we compared the retrieved BT-RADS score, comparison exam date, and existence of the comparison study with the manually labeled results and reported the precision and recall for each labeling component. We found 99.4% precision and 99.4% recall for extracting BT-RADS scores (165 exams labeled), 97.5% precision and 96.3% recall for extracting comparison date (162 exams labeled), and 98.0% precision and 95.1% recall for confirming existence of the baseline exam (103 exams labeled). The final NLP model was applied to the entire dataset consisting of 634 patients with 3403 available MRI scans.



Procedure: MRI BRAIN WITHOUT AND WITH CONTRAST
INDICATION: Anaplastic gliomas/glioblastoma, monitor, post resection, G93.89 Other specified disorders of brain
COMPARISON: CT brain 8/27/2019. MRI brain 8/13/2019.
TECHNIQUE/PROTOCOL: Standard adult brain protocol.
FINDINGS:
...
IMPRESSION:
Slightly increased mild enhancement of the right posterior temporal lobe.
Decreased T2 hyperintensity in the right temporoparietal lobe.
Unchanged
size of the post-biopsy right parietal hematoma. Brain tumor follow-up
score 3b: Imaging worsening, indeterminate
Electronically Reviewed by: XXX, MD, XXX
Electronically Reviewed on: XX/XX/XXXX X:X PM
I have reviewed the images and concur with the above findings.
Electronically Signed by: XXXX, MD, XXXX
Electronically Signed on: XX/XX/XXXX X:X PM

Date: 8/13/2019

BT-RADS score: 3b

Supplementary Figure 1. Illustration of the first two steps of the NLP algorithm. Sensitive information was manually masked in this example. The NLP algorithm took a radiology report as input and retrieved relevant information from the free-text. BT-RADS score was retrieved in the impression section in the main body. Date information was retrieved from the comparison section at top.

Deep Learning Segmentation Model Methods and Evaluation

We developed a 3D deep convolutional neural network to segment tumor sub-compartments on multi-sequence MRI. The model was based on the 3D U-Net architecture and implemented in nnU-Net version 1. nnU-Net is an automated pipeline which configures the best U-Net-based network based on the provided training data. We trained the nnU-Net model in a 5-fold cross-validation scheme using the University of California San Francisco Adult Longitudinal Posttreatment Diffuse Glioma dataset (n=596 manually annotated brain MRI exams). Data is available at the following link: https://imagingdatasets.ucsf.edu/dataset/2. Training comprised 1000 epochs per fold iterating through the entire dataset and lasted approximately 20 hours per fold using an NVIDIA A6000 GPU.

The trained model was formally evaluated using a previously unseen testing dataset from our institution. The testing cohort consisted of a subset of 497 MRI scans chosen to have equal proportions of exams scored as BT-RADS 1, 2, 3, and 4, respectively, as identified by the NLP algorithm described in the previous section. The trained segmentation model was applied to each exam to generate automated segmentations. Tumor segmentations for each exam in the testing cohort were manually corrected and approved by a fellowship trained neuroradiologist. Automated tumor segmentations were compared to manually corrected segmentations using the Dice coefficient as the primary metric. Results showed a mean +/- standard deviation of 0.8861 +- 0.2476 for enhancing tumor and 0.9833 +- 0.0372 for surrounding non-enhancing FLAIR signal abnormality.

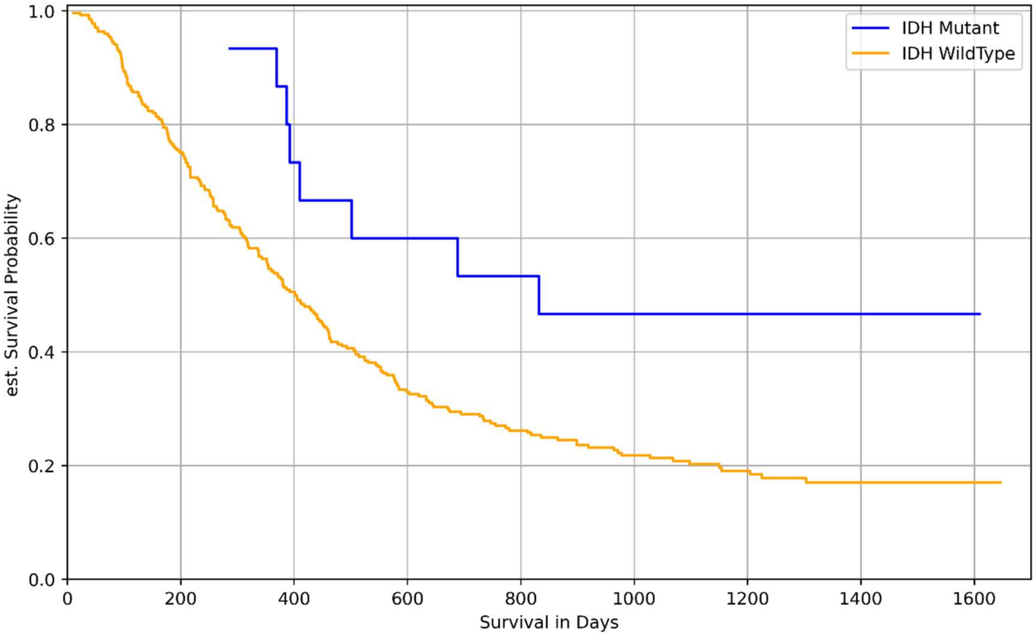
Concordance Index Difference Test

We computed the difference in C-index between the two survival models. The variance of the difference was computed based on the covariance matrix of the C-index from the two survival models, which accounts for the correlation between the two C-indices. We then computed the test statistic and p-value based on an approximation from the standard Normal distribution under large sample size approximation.

IDH Subgroup Analysis

Supplementary Figure 2 shows Kaplan-Meier curves between IDH-wildtype and IDH-mutant. The log-rank test shows a significantly worse overall survival in IDH-wildtype comparted to IDH-Mutant (P-value = 0.018). This finding aligns with the poor prognosis and survival outcomes in IDH-wildtype GBMs reported in literature.

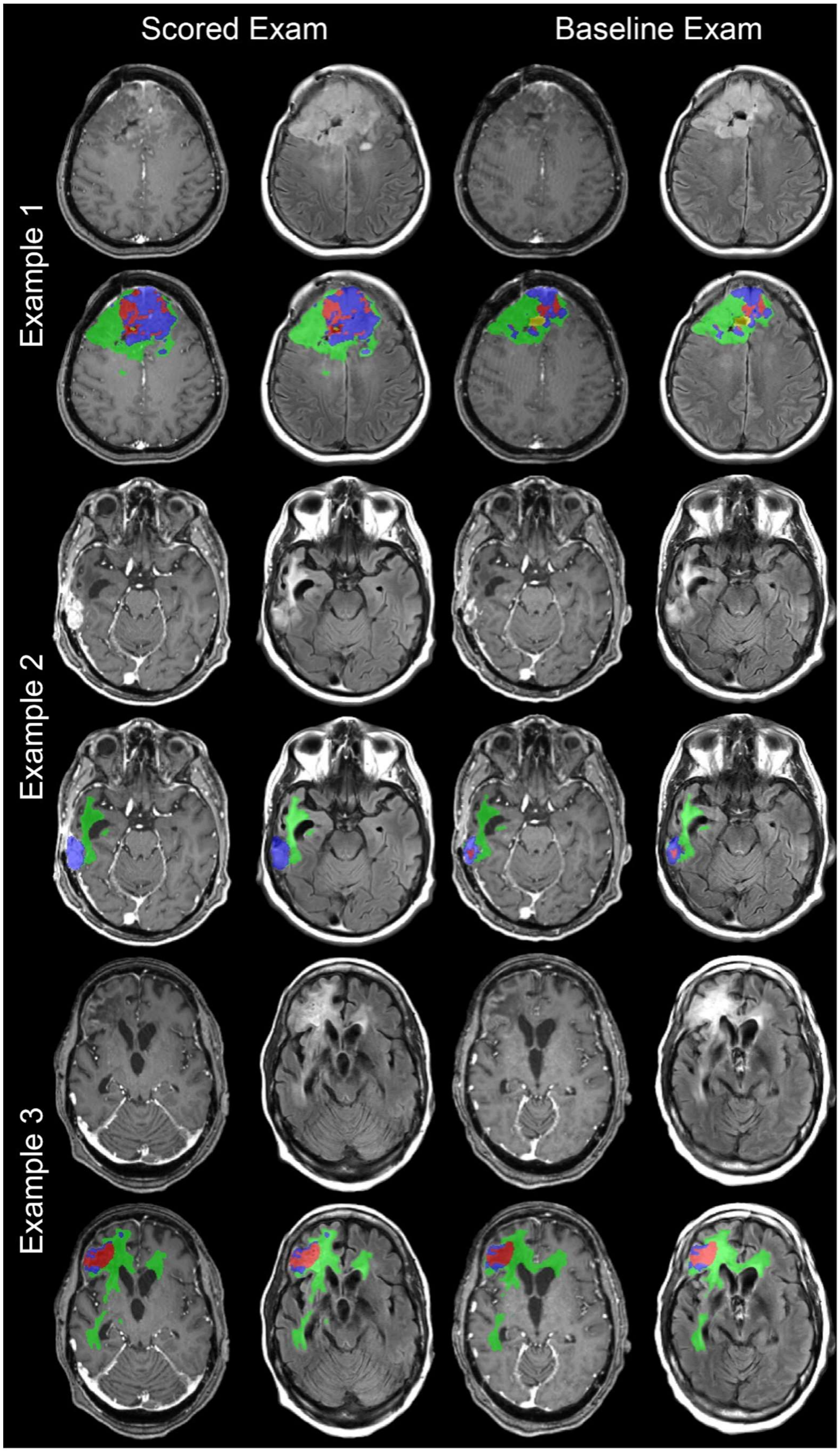
Supplementary Table 1 provides full performance metrics (similar as Table 3 in the manuscript) in IDH-wildtype patients. AI-VTRA had a higher macro-F1 score (AI-VTRA macro-F1 = 0.532) compared to AI-VTRAET (AI-VTRAET macro-F1 = 0.541) alone. AI-VTRAET alone demonstrated improved performance compared to AI-VTRA for 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. These findings are aligned with the findings in the manuscript when assessed both IDH wildtype and IDH mutant.



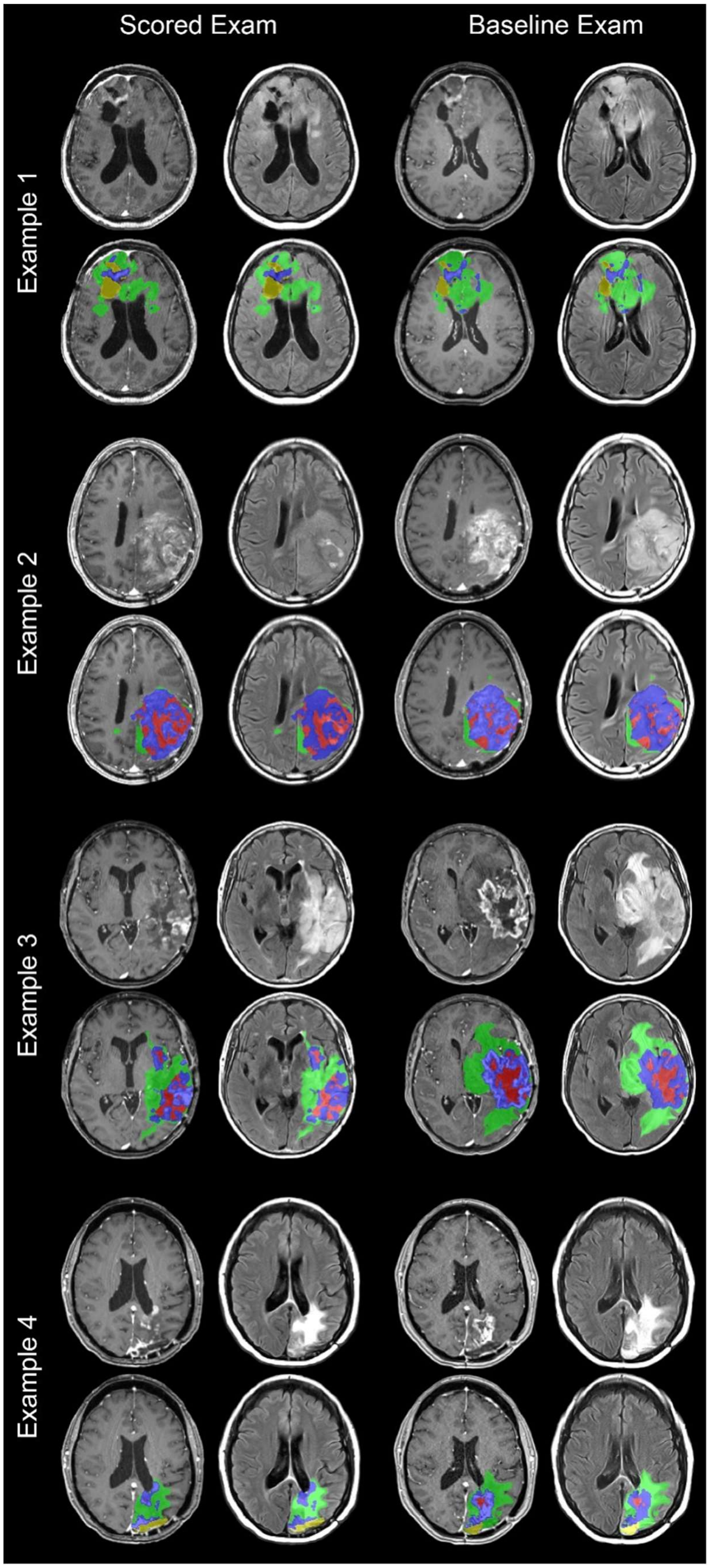
Supplementary Figure 2. Kaplan-Meier curves for IDH-wildtype (orange) and IDH-mutant (blue).

AI-human major discrepancy analysis

All major discrepancies between AI-VTRA and human assessments (defined as BT-RADS score of 4 assigned by one modality and BT-RADS score of 1 assigned by the other) were identified for further analysis. For each major discrepancy, MR images and corresponding reports were manually reviewed by a board certified neuroradiologist (author [REDACTED FOR REVIEW]) to determine the cause of the discrepancy. 23/2,446 (0.9%) of analyzed exam pairs yielded major discrepancies between AI and human assessments. Of these, 3 were scored as BT-RADS 1 by human and BT-RADS 4 by AI (1-4 discrepancy) and the remaining 20 were scored as BT-RADS 4 by human and BT-RADS 1 by AI (1-4 discrepancy). Considering the 3 cases of 1-4 discrepancy, 2/3 (67%) were due human error from comparison to a prior other than the specified baseline, and 1/3 (33%) was due to poor AI segmentation of enhancement in the setting of bevacizumab therapy (Supplementary Figure 3). Considering the 20 cases of 4-1 discrepancy, 11/20 (55%) were due to mixed change (i.e. one enhancing lesion smaller and another larger) that was assigned as BT-RADS 4 by human. 8/20 (40%) of 4-1 discrepancies were found to have decreased volume of enhancement but increased overall lesion size (including FLAIR abnormality) in the setting of bevacizumab therapy. A single (5%) 4-1 discrepancy was due to human error from comparison to a prior other than the specified baseline (Supplementary Figure 4).



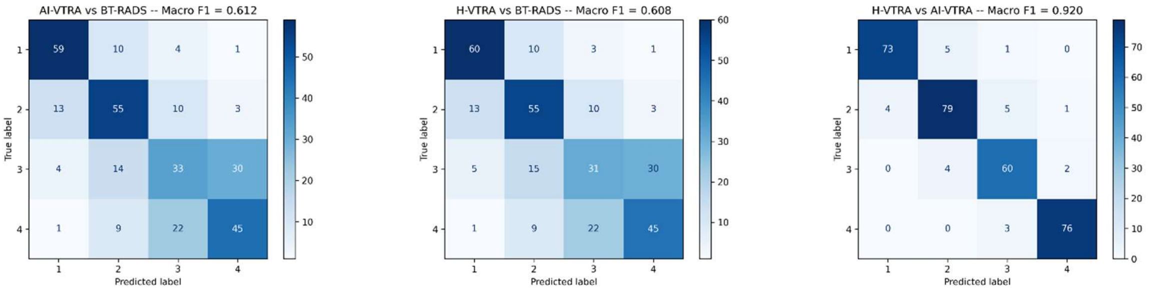
Supplementary Figure 3. All three examples of exams that were scored as BT-RADS 1 by a radiologist and BT-RADS 4 by the AI algorithm. Example 1: This exam was inappropriately scored as a BT-RADS 1 due to “decreased enhancement and FLAIR signal abnormality”. Review of imaging reveals grossly accurate segmentation and a significant increase in tumor enhancement consistent with a score of BT-RADS 4. It is possible that the interpreting radiologist mistakenly compared this exam to a prior other than the specified baseline exam. Example 2: This exam was inappropriately scored as a BT-RADS 1 due to “decrease in enhancing tumor” compared to an interval prior even though a significant increase in enhancement compared to the baseline prior was also described. Review of imaging reveals grossly accurate segmentation and a significant increase in tumor enhancement consistent with a score of BT-RADS 4. Example 3: This exam was scored as a BT-RADS 1 due to “decreased enhancement”. Review of imaging and clinical history revealed a largely non-enhancing lesion in the setting of bevacizumab therapy. Automated segmentation was grossly inaccurate, particularly in areas of coagulative necrosis, and therefore the AI score of BT-RADS 4 was inaccurate. Review of imaging revealed no significant change in tumor related signal abnormality consistent with a score of BT-RADS 2.



Supplementary Figure 4. Four examples of exams that were scored as a BT-RADS 4 by a radiologist and as a BT-RADS 1 by the AI algorithm. Example 1: This exam was scored as a BT-RADS 4 by a human due to “new subependymal enhancement”. The automated tumor segmentation is grossly accurate and shows an overall decrease in volume of FLAIR abnormality and enhancement consistent with a score of BT-RADS 1. Example 2: This exam was scored as a BT-RADS 4 by a human due to “increase in enhancement”. Review of imaging and clinical history reveals enlargement of tumor related signal abnormality but decreased overall enhancement, which was presumably related to interval initiation of bevacizumab therapy. According to guidelines, this exam should have been scored as a BT-RADS 1B. Example 3: This exam was scored as BT-RADS 4 by a human; however, imaging clearly shows decreased FLAIR abnormality and enhancement. In this case the human reader mistakenly scored the exam compared to a different prior other than the specified baseline. Example 4: This exam was scored as a BT-RADS 4 by a human due to a new subependymal enhancing lesion. Review of the imaging reveals mixed change with overall decrease in FLAIR abnormality and enhancing tumor volume but emergence of a new small subependymal enhancing lesion.

Intermediary Comparison

Inter-rater variability of BT-RADS, considered as source of heterogeneity, has been previously reported by the Emory University neuroradiology group (Essien et al. AJNR Am J Neuroradiol. 2024;45(9):1308-1315. doi:10.3174/ajnr.A8322) and is quoted at Gwet index = 0.83 for a group of 6 neuroradiologist and trainee readers. Although we did not formally evaluate the inter-rater variability, we added the intermediary comparison as suggested to examine if the major source of heterogeneity came from the automated segmentation models. We utilized 497 manually reviewed and annotated MRIs for this heterogeneity analysis. Filtering these MRIs from the analytics cohort (2446 paired data from 634 patients) resulted in a subset of 313 paired data from 139 patients, each of which contains both the baseline and the comparison MRI annotated by human. We applied the same thresholds and methods from the main analysis to keep the results consistent. We denoted results from human-segmented data as H-VTRA as compared to AI-VTRA. We compared among H-VTRA, AI-VTRA, and BT-RADS, and reported the main metric Macro-F1 scores, and confusion matrices (Supplementary Figure 5). We found that the classification performance was almost the same when compared to BT-RADS (H-VTRA vs BT-RADS: 0.608, AI-VTRA vs BT-RADS: 0.602), and both have similar assessments (AI-VTRA vs H-VTRA: 0.920). These findings suggest that the segmentation model was not the major source of heterogeneity between AI and human assessments.



Supplementary Figure 5. Confusion matrices for pairwise comparison among H-VTRA, AI-VTRA, and BT-RADS. Macro-F1 scores were included in the subplot titles.

Supplementary Table 1. Performance metrics (Macro-F1, Micro-F1, sensitivity, specificity, and precision) for AI-VTRA/AI-VTRAET predictions of radiologist-based response assessment in IDH-wildtype patients (N=479).

	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	***0.749***	***0.756***	***0.756***	***0.741***	***0.559***	***0.580***	***0.696***	***0.696***
Micro-F1	***0.856***	***0.868***	***0.768***	***0.758***	***0.673***	***0.664***	***0.812***	***0.812***
Sensitivity	***0.768***	***0.710***	***0.768***	***0.707***	***0.235***	***0.311***	***0.597***	***0.597***
Specificity	***0.869***	***0.893***	***0.768***	***0.786***	***0.908***	***0.853***	***0.854***	***0.854***
Precision	***0.474***	***0.506***	***0.645***	***0.644***	***0.578***	***0.533***	***0.442***	***0.442***

Supplementary Table 2. Class distribution of BT-RADS and AI-VTRA scores with respect to each comparison pair (N=2,446) in the dataset.

Score	BT-RADS N (%)	AI-VTRA N (%)
1	324 (13%)	449 (18%)
2	968 (40%)	1070 (44%)
3	788 (32%)	443 (18%)
4	366 (15%)	484 (20%)

Supplementary Table 3. Detailed Relationship between BT-RADS score and AI-VTRAFLAIR, AI-VTRAET, and AI-VTRA for each glioblastoma MRI follow up assessment category. AI-VTRA is a composite metric derived from AI-VTRAFLAIR, AI-VTRAET. % VD is the percentage change of VD. Abs VD is the absolute change of VD.

Assessment Category	BT-RADS (human)	AI-VTRA (AI)		
		$AI\text{-}VTRA_{FLAIR}$	$AI\text{-}VTRA_{ET}$	**AI-VTRA**
Imaging improvement	1	$\% VD_{FLAIR} \leq -40\%$ AND $Abs\ VD_{FLAIR} \geq 1mL$	$\% VD_{ET} \leq -10\%$ AND $Abs\ VD_{ET} \geq 1mL$	$AI\text{-}VTRA_{ET} = 1$ OR $(AI\text{-}VTRA_{ET} = 2$ AND $AI\text{-}VTRA_{FLAIR} =1)$
No significant imaging change	2	$-40\% < \% VD_{FLAIR} < 40\%$ OR $Abs\ VD_{FLAIR} < 1mL$	$-10\% < \% VD_{ET} < 10\%$ OR $Abs\ VD_{ET} < 1mL$	$AI\text{-}VTRA_{ET} = 2$ OR $(AI\text{-}VTRA_{ET} = 1$ AND $AI\text{-}VTRA_{FLAIR} = 3)$
Imaging worsening	3	$\% VD_{FLAIR} \geq 40\%$ AND $Abs\ VD_{FLAIR} \geq 1mL$	$10\% <= \% VD_{ET} < 40\%$ AND $Abs\ VD_{ET} \geq 1mL$	$AI\text{-}VTRA_{ET} = 3$ OR $(AI\text{-}VTRA_{ET} = 2$ AND $AI\text{-}VTRA_{FLAIR} = 3)$
Imaging worsening equivalent to RANO progression	4	N/A	$\% VD_{ET} \geq 40\%$ AND $Abs\ VD_{ET} \geq 1mL$	$AI\text{-}VTRA_{ET} = 4$

Supplementary Table 4. Full metrics table (Macro sensitivity, specificity, precision, and F1) of AI-VTRAET, and AI-VTRA when compared with the reference standard BT-RADS score at the empirical thresholds.

Metrics	$AI\text{-}VTRA_{ET}$	AI-VTRA
Macro-F1	0.535	0.548
Macro-Sensitivity	0.590	0.585
Macro-Specificity	0.853	0.852
Macro-Precision	0.541	0.540

Supplementary Table 5. Imaging parameters, including field of view (FOV), resolution, bandwidth (BW), slice thickness, TR, and TE, for T2 flair, T1-weighted pre-contrast, T2, T1-weighted post-contrast used in the study.

Scanner	Plane	FOV (mm)	Resolution (mm/pixel)	BW	Slice Thickness (mm)	TR	TE
Siemens 1.5T	T2 flair	256	1	592	1	5000	355
	T1-weighted pre-contrast	256	1	360	1	2200	3.11
	T2	240	0.5	191	4	3000	103
	T1-weighted post-contrast	256	1	360	1	2200	3.95
Siemens 3T	T2 flair	256	1		1	5000	390
	T1-weighted pre-contrast	256	1		1	2110	3.95
	T2	240	0.5		4	3000	103
	T1-weighted post-contrast	256	1		1	2110	3.95
GE 1.5T	T2 flair	240	1.03	91	1	6200	90
	T1-weighted pre-contrast	256	1	31	1	7.7	min*
	T2	240	0.47	50	4	3137	102
	T1-weighted post-contrast	256	1	31	1	7.7	min*

*Minimum allowed by the scanner