AJNR

Interrater Agreement of BT-RADS for Evaluation of Follow-Up MRI in Treated Primary Brain Tumor Patients

Michael Essien, Maxwell E. Cooper, Ashwani Gore, Taejin L. Min, Benjamin B. Risk, Gelareh Sadigh, Ranliang Hu, Michael J. Hoch and Brent D. Weinberg

AJNR Am J Neuroradiol published online 29 April 2024
http://www.ajnr.org/content/early/2024/04/29/ajnr.A8322

Published April 29, 2024 as 10.3174/ajnr.A8322

ORIGINAL RESEARCH

Interrater Agreement of BT-RADS for Evaluation of Follow-Up MRI in Treated Primary Brain Tumor Patients

Michael Essien*, Maxwell E. Cooper*, Ashwani Gore, Taejin L. Min, Benjamin B. Risk, Gelareh Sadigh, Ranliang Hu, Michael J. Hoch, Brent D. Weinberg

ABSTRACT

BACKGROUND AND PURPOSE: The Brain Tumor Reporting and Data System (BT-RADS) is a structured radiology reporting algorithm that was introduced to provide uniformity in post-treatment primary brain tumor follow-up and reporting, but its interrater reliability (IRR) assessment has not been widely studied. Our goal is to evaluate the IRR among neuroradiologists and radiology residents in the use of BT-RADS.

MATERIALS AND METHODS: This retrospective study reviewed 103 consecutive MR studies in 98 adult patients previously diagnosed with and treated for primary brain tumor (January 2019 to February 2019). Six readers with varied experience (4 neuroradiologists and 2 radiology residents) independently evaluated each case and assigned a BT-RADS score. Readers were blinded to the original score reports and the reports from other readers. Cases in which at least one neuroradiologist scored differently were subjected to consensus scoring. After the study, a post-hoc reference score was also assigned by 2 readers using future imaging and clinical information previously unavailable to readers. The interrater reliabilities were assessed using Gwet's AC2 index with ordinal weights and percent agreement.

RESULTS: Of the 98 patients evaluated (median age, 53 years; interquartile range, 41-66 years), 53% were males. The most common tumor type was astrocytoma (77%) of which 56% were grade 4 glioblastoma. Gwet's index for interrater reliability among all six readers was 0.83 (95% CI: 0.78, 0.87). The Gwet's index for the neuroradiologists' group (0.84 [95% CI: 0.79, 0.89]) was not statistically different from that for the residents' group (0.79 [95% CI: 0.72, 0.86]) (x^2 = 0.85; p = 0.36). All four neuroradiologists agreed on the same BT-RADS score in 57 of the 103 studies, three neuroradiologists agreed in 21 of the 103 studies, and two neuroradiologists agreed in 21 of the 103 studies. Percent agreement between neuroradiologist blinded scores and post-hoc reference scores ranged from 41%-52%.

CONCLUSIONS: A very good interrater agreement was found when tumor reports were interpreted by independent blinded readers using BT-RADS criteria. Further study is needed to determine if this high overall agreement can translate into greater consistency in clinical care.

ABBREVIATIONS: BI-RADS = Breast Imaging Reporting and Data System; BT-RADS = Brain Tumor Reporting and Data System; IQR = interquartile range; IRR = interrater reliability; NI-RADS = Neck Imaging Reporting and Data System.

Received month day, year; accepted after revision month day, year.
From the Department of Radiology and Imaging Sciences (M.E., M.E.C., A.G., T.L.M., R.H., B.D.W.), and Rollins School of Public Health (B.B.R.), Emory University, Atlanta, GA, USA; Department of Radiological Sciences (G.S.), University of California Irvine, Orange, CA, USA; Department of Radiology (M.J.H.), University of Pennsylvania, Philadelphia, PA, USA

*Co-first authors

The authors declare no conflicts of interest related to the content of this article.

Please address correspondence to Brent D. Weinberg, MD, PhD, Department of Radiology and Imaging Sciences, Division of Neuroradiology, Emory University, 1364 Clifton Rd NE, Suite BG20, Atlanta, GA 30322, USA; Brent.d.weinberg@emory.edu.

SUMMARY SECTION

PREVIOUS LITERATURE: The Brain Tumor Reporting and Data System (BT-RADS) is a structured radiology reporting algorithm developed by a multi-disciplinary team to minimize subjectivity among radiologists interpreting MRI exams of treatment response in patients with brain tumor. However, unlike other reporting and data systems including other neuroradiology standardized reporting system, the interrater agreement of BT-RADS is understudied. To the best of our knowledge, we are only aware of a single study that has reported on the interrater agreement of BT-RADS in a smaller patient population. Our study represents a larger patient population compared to prior study.

KEY FINDINGS: There is a very good interrater agreement when primary brain tumor reports are interpreted by independent blinded readers using BT-RADS criteria. There is no statistical difference between interrater agreement in using BT-RADS among neuroradiologists and radiology trainees.

KNOWLEDGE ADVANCEMENT: BT-RADS can be used to produce consistent and transparent primary brain tumor reports even in less experienced hands, and its implementation can help provide consistent reports across readers.

Copyright 2024 by American Society of Neuroradiology.

INTRODUCTION

The most common adult primary malignant brain tumors are gliomas.[1] They are classified according to the evolving WHO classification of tumors of the central nervous system, currently in its 5th edition (year 2021), and include Astrocytoma, IDH-mutant; Oligodendroglioma, IDH-mutant, and 1p/19q-codeleted; Glioblastoma, IDH-wildtype.[2] Glioblastoma, IDH-wildtype, poses a significant clinical challenge because of its high rate of recurrence despite optimized management strategies.[3] It has a 5-year survival rate as low as 5-7%.[4-6] With high rates of recurrence, long-term monitoring of brain tumor patients is critical for identifying disease progression and guiding clinical management. MRI is an essential tool for diagnosing, monitoring, and managing patients with malignant brain tumors.[7] A previous study surveyed a group of neuro-oncology specialists that reported heavy reliance on MRI reports as part of their management of brain tumor patients.[8] However, there is considerable overlap between MRI findings of tumor progression and treatment-related changes[7, 9], resulting in subjective and inconsistent interpretations. As a result, there is a need for improved standardization of brain tumor reporting that can serve as a guide for clinical management.

The Brain Tumor Reporting and Data System (BT-RADS) is a structured radiology reporting algorithm designed to decrease subjectivity among radiologists interpreting MRI exams of treatment response in brain tumor patients.[8, 10] BT-RADS was developed by a multi-disciplinary team of neuroradiologists, neuro-oncologists, neurosurgeons, and neuro-pathologists. The scoring template consists of a numeric assessment (score 0-4) of MRI findings while considering the patient's previous treatment course, including radiation completion date and use of anti-angiogenic agents (bevacizumab) and steroid therapies. [8] In summary, BT-RADS 0=baseline study; BT-RADS 1a=imaging improvement; BT-RADS 1b=imaging improvement as a result of drug effect; BT-RADS 2=unchanged imaging features; BT-RADS 3a=worsening imaging features 12 weeks post-radiotherapy; BT-RADS 3b=worsening imaging features 12 weeks post-radiotherapy; BT-RADS 3c=unspecified new enhancing lesions within radiotherapy treatment area; BT-RADS 4=specified new lesions outside radiotherapy treatment area.[8]

A critical part of developing a standardized radiology reporting system is assessing interrater reliability. The Breast Imaging Reporting and Data System (BI-RADS) is a structured radiology reporting system that has been adopted as a standard component of breast cancer diagnosis and management. BI-RADS has been demonstrated by several previous studies to have acceptable to good levels of interrater reliability when applied to exams of breast cancer patients.[11-13] Unfortunately, interrater reliability of BT-RADS in post-treatment primary brain tumor is understudied with little information on the transparency and consistency of this scoring scale among its users.

The purpose of this study is to determine the interrater reliability of multiple neuroradiologists using BT-RADS to score the same cohort of post-treatment primary brain tumor MRI exams. In addition, the authors seek to evaluate the interrater reliability among radiology residents utilizing BT-RADS as a part of their neuroradiology training.

MATERIALS AND METHODS

Patient Selection

This was a HIPAA compliant retrospective study approved by the Institutional Review Board of our institution with waiver of patient informed consent. Adult patients (>18 years old) with a primary parenchymal brain tumor undergoing brain MR imaging between January 1, 2019, and February 28, 2019, at a single center were included. The exclusion criteria were patients with secondary or benign brain tumors, meningiomas, and patients younger than 18 years. Each patient's imaging date, diagnosis, tumor grade, BT-RADS score from the radiology report, surgery date, radiation therapy completion date, bevacizumab therapy commencement date (if applicable), and steroid medication use (documented as a yes if used at all) were recorded. Because the study was performed on images obtained before recent revision of the WHO criteria, tumors were classified based on the 2016 4th edition of the WHO classification of tumors of the central nervous system.[14]

Image Acquisition

Patients underwent imaging with a standardized brain tumor imaging protocol without and with IV contrast agent. These studies were performed on a mix of scanners with field strength of 1.5 or 3.0 T. Imaging sequences included the following: axial DWI 5 mm slices, sagittal 3D FLAIR 1 mm slices reformatted into axial and coronal images, axial gradient echo T2WI 5 mm slices, axial fast spin echo T1WI 5 mm slices, sagittal 3D T1WI precontrast MP-RAGE 1 mm slices reformatted into axial and coronal images, axial T2WI fast spin echo fat-suppressed 5 mm slices, and sagittal 3D T1WI postcontrast MP-RAGE 1 mm slices reformatted into axial and coronal images.

Single bolus dynamic susceptibility contrast (DSC) perfusion imaging was performed on each patient and relative cerebral blood volume maps were calculated using Dynasuite (Philips Healthcare, Amsterdam, Netherlands).

Reader Selection

A total of six readers with varied levels of experience were selected for this study - 4 board-certified and CAQ-certified neuroradiologists (Faculty_1, 3 years; Faculty_2, 1 year; Faculty_3, 3 years; Faculty_4, 3 years) and 2 radiology residents (Resident_1, R-3; Resident_2, R-4). BT-RADS had been actively used in our department's standard practice for reading primary brain tumor MRIs for approximately two years prior to the start of this study.

Reader Workflow

Each reader reviewed and scored all 103 MRIs using the BT-RADS system. Readers were blinded to the MRI report for that study, BT-RADS score assigned by the radiologist that initially read the patient's imaging, and scores from other readers. The patient's diagnosis, tumor grade, surgery date, radiation therapy completion date, bevacizumab therapy commencement date (if applicable), and steroid use (if

applicable) were provided to the readers. Readers were given access to previous MRIs and corresponding reports for prior MRI exams used as comparison. However, readers did not view MRIs and corresponding reports that would have been done after the selected MRIs for the study. If an image revealed multiple lesions, readers were instructed to assign the patient's BT-RADS score based on the worst or highest scoring lesion. All readers had access to reference materials with instructions for BT-RADS scoring, including a flow-chart, interactive scoring tool, and reference tables available online (www.btrads.com). Readers also assigned their level of confidence ranging from 1 (least confident) to 5 (most confident), recorded how much DSC perfusion contributed to their score from 1 (very little) to 5 (great deal), and recorded the oldest prior comparison study considered in assigning a score.

Consensus and Post-hoc Reference Scoring

When one or more of the four neuroradiologists assigned a score different from the rest of the neuroradiologists, such MRI exams were subjected to consensus scoring. When three of the four neuroradiologists assigned the same score, the majority vote was assigned the consensus score. When two or fewer neuroradiologists agreed on a score for a particular case, two of the faculty neuroradiologists with the most experience reviewed the imaging in a consensus review session and assigned a consensus score for all such cases, considering both imaging and faculty scores.

A post-hoc reference score using subsequent follow-up information was also assigned by the same two faculty neuroradiologists to only the cases that were subjected to consensus scoring. This score used follow-up information obtained after the original scored study, including any available follow-up imaging, clinical worsening, or subsequent pathology results from biopsy or re-resection. This post-hoc reference score was considered the gold standard score for assessment of tumor worsening.

Statistical Analysis

Inter-rater reliability (IRR) was calculated using percent agreement and Gwet's AC2 index, applying linear weights.[15] While Gwet's index corrects for agreement due to chance, percent agreement does not. This leads to overestimation of IRR for percent agreement. IRR was assessed separately for all 6 readers, the neuroradiologists group, and the residents group. Gwet's index was interpreted using the benchmark scale as described by Altman: <0.20 = poor; 0.21-0.40 = fair; 0.41-0.60 = moderate; 0.61-0.80 = good; 0.81-1.00 = very good.[17] Gwet's agreement coefficients between neuroradiologists and residents were compared using the Chi-square test of independence. The value of DSC perfusion and readers' level of confidence in BT-RADS scoring were assessed separately by violin graphs which allows for visualization of data distribution and its probability density. This allows for distribution comparison among multiple groups on the same graphic presentation. Alpha level of 0.05 was set as level of significance. All analyses were performed on R Statistical Software (v4.2.2, R Core Team 2022, Vienna, Austria), using package irrCAC[18].

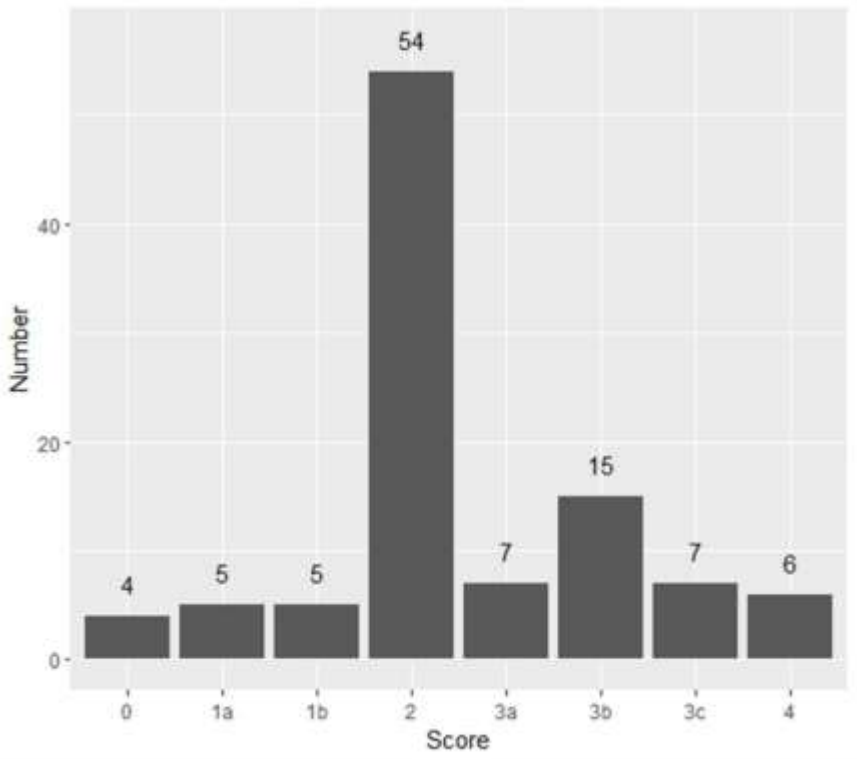

FIG 1. Bar graph showing relative frequency of BT-RADS Scores (from original reports) among selected MRI exams (n=103). BT-RADS = Brain Tumor Reporting and Data System.

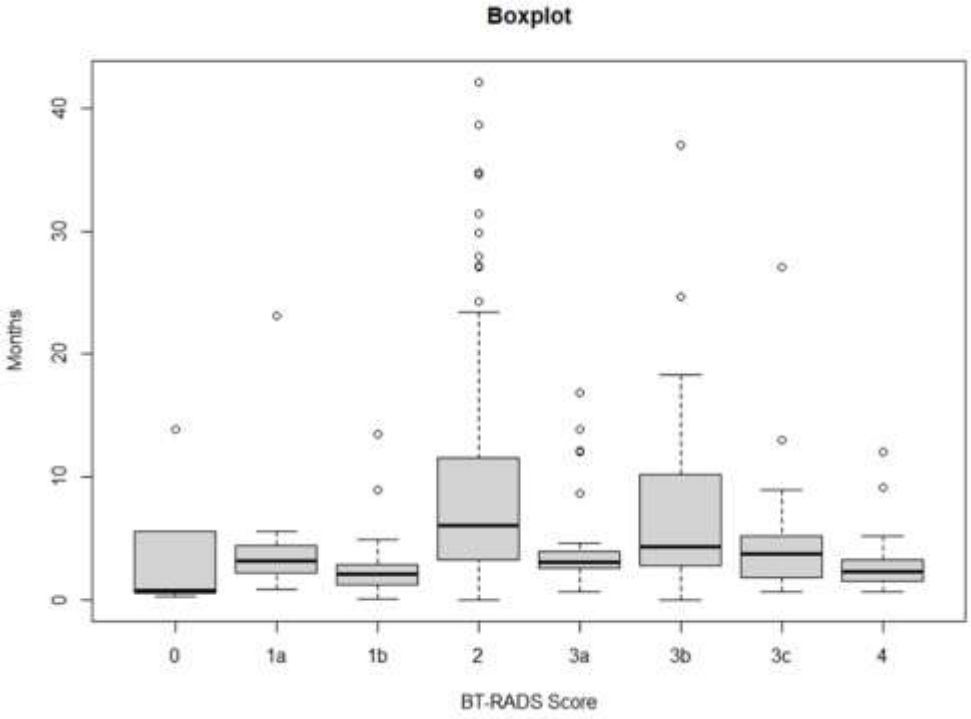

FIG 2. Boxplot graph showing time (in months) from the oldest MRI to the prior MRI used by readers to assign a BT-RADS score. The maximum and minimum values are represented at either end of the whiskers. The box represents the interquartile range (25th percentile to the 75th percentile), with the median value represented by the thick horizontal black line within the box. Outliers are shown away from the whiskers and box by the symbol (○).

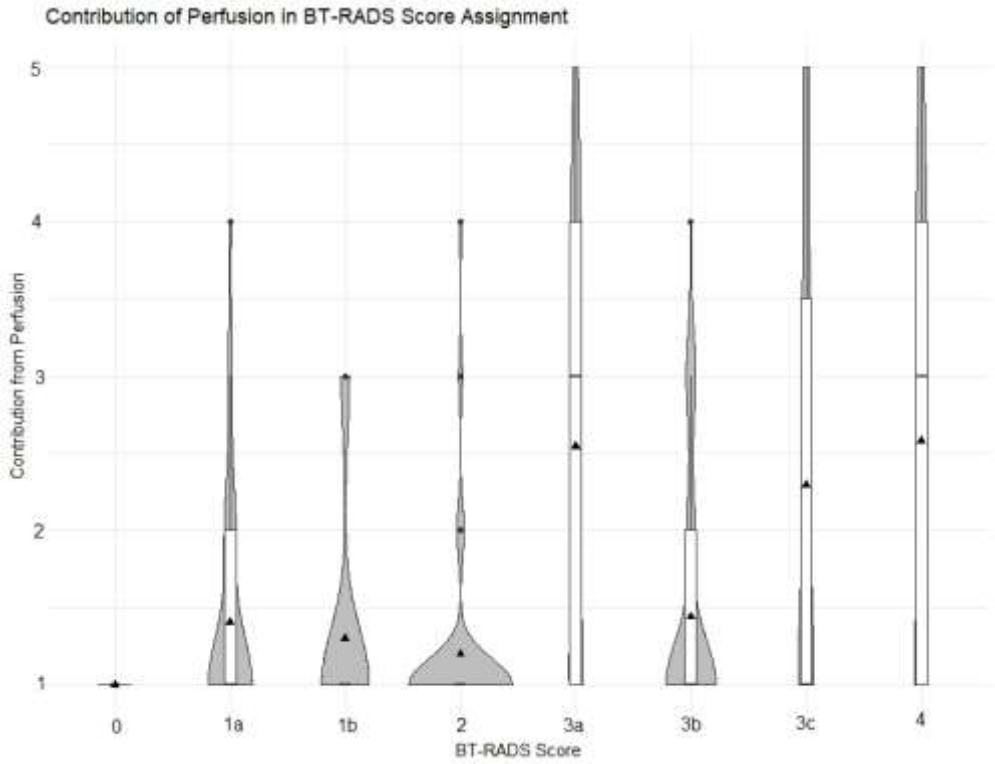

FIG 3. Violin graph showing box plot inside a density plot. The width of the density plot at any region corresponds to the frequency of the data points at that region. For the box plot, the maximum and minimum values are represented at either end of the whiskers. The box represents the interquartile range (25th percentile to the 75th percentile), with the median value represented by a thick horizontal black line either within the box or at either end of the box. The mean value is represented by the black triangle (Δ). Outliers are shown away from the whiskers and box by small black dots (●). BT-RADS = brain tumor reporting and data system.

On a scale of 1 – 5 (with 1 = very little, and 5 = great deal), all 6 readers rated how much perfusion contributed to their interpretation of scan and assignment of BT-RADS score.

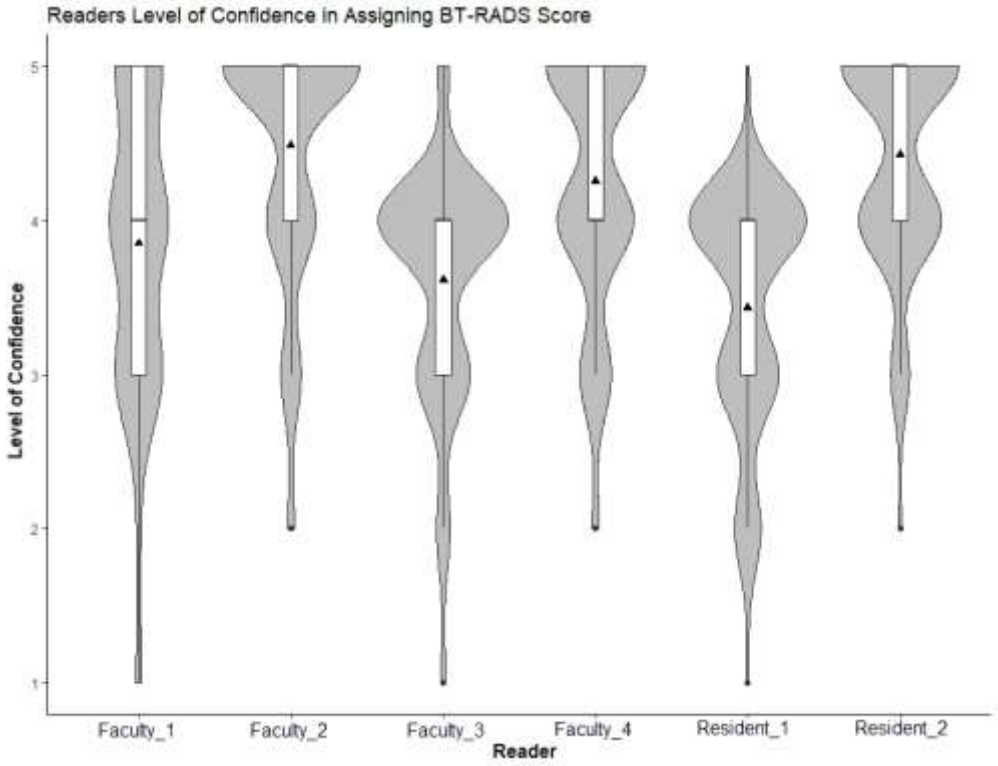

FIG 4. Violin graph showing box plot inside a density plot. The width of the density plot at any region corresponds to the frequency of the data points at that region. For the box plot, the maximum and minimum values are represented at either end of the whiskers. The box represents the interquartile range (25th percentile to the 75th percentile), with the median value represented by a thick horizontal black line either within the box or at either end of the box. The mean value is represented by the black triangle (Δ). Outliers are shown away from the whiskers and box by small black dots (●).

On a scale of 1 – 5 (with 1 = not sure at all, and 5 = absolutely sure), all 6 readers rated their level of confidence for the BT-RADS score they assigned to each case. Neuroradiologists (Faculty_1, Faculty _2, Faculty _3 and Faculty _4); Radiology residents (Resident_1 and Resident_2).

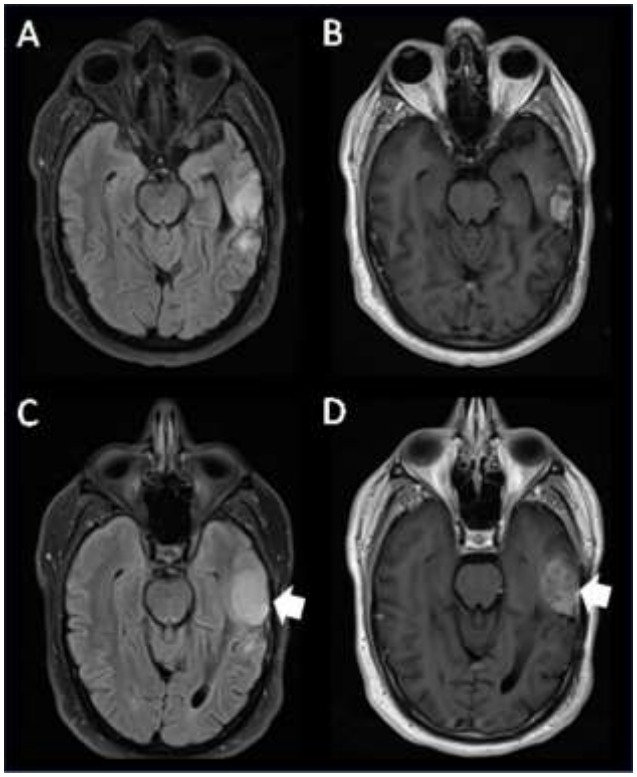

FIG 5. Imaging from a patient with agreement among all readers. 48-year-old man with IDH wild type glioblastoma. FLAIR (A) and T1 post-contrast (B) imaging 18 months after surgery showing abnormal FLAIR and enhancing treated tumor in the left temporal lobe. FLAIR (C) and T1 post-contrast (D) two months later showing marked increase in the size of abnormal mass-like FLAIR and enhancement (white arrows) with > 25% increase in cross-sectional area. All readers gave the study a score of BT-RADS 4. BT-RADS = brain tumor reporting and data system.

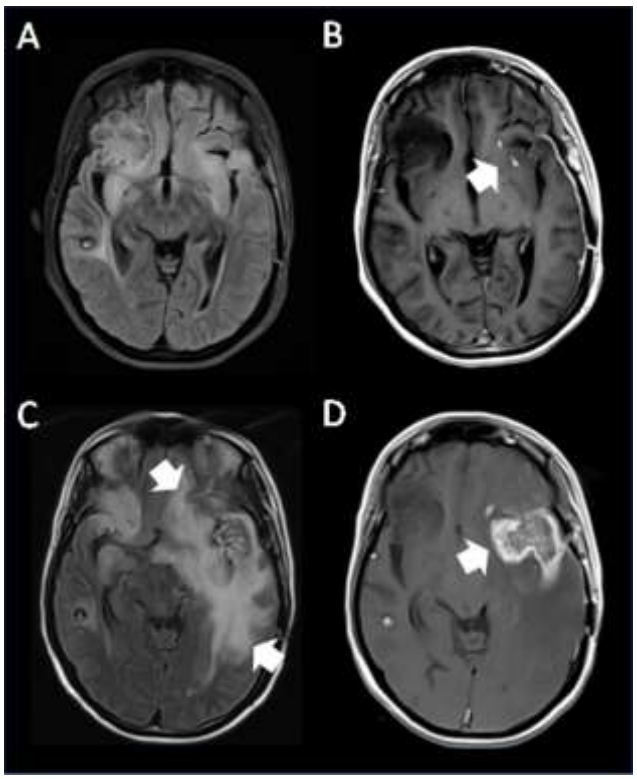

FIG 6. Imaging from a patient with disagreement among readers. 45-year-old woman with IDH wild type glioblastoma. FLAIR (A) and T1 post-contrast (B) imaging 3 months after surgery and 1 month after completing radiation showing multifocal abnormal FLAIR in the bilateral frontal lobes with minimal enhancement in the left frontal lobe (white arrow). FLAIR (C) and T1 post-contrast (D) two months later (3 months after completing radiation) showing marked increase in left frontal edema (white arrows) and

enhancement (white arrow). Half the readers (2 neuroradiologists and 1 resident) gave the study a score of BT-RADS 3a (pseudoprogression) and half the readers gave the study a score of BT-RADS 4 (progression). The post-hoc reference score was BT-RADS 4. BT-RADS = brain tumor reporting and data system.

RESULTS

A total of 103 consecutive MR imaging studies of primary brain tumors from 98 patients were evaluated in this study. Five of the patients had two MRI exams each included in the study because they were scanned twice during the 2-month imaging search period for this study. The median patient age was 53 years (interquartile range [IQR], 41-66 years). Fifty three percent of the patient population (52 of 98) were males, with 18% and 20% of the sample population on steroids and bevacizumab respectively. Seventy seven percent of the MR studies (79 of 103) had a diagnosis of an astrocytoma type, of which 56% of the astrocytoma type (44 of 79) were classified as Grade 4 glioblastoma. Of the grade 4 glioblastoma, 20% (9 of 44) were of IDH-mutant type. Forty five percent of all astrocytoma tumor type (36 of 79) were of IDH-mutant type (Table 1). Analysis of the BT-RADS scores from the original radiology reports assigned by the initial reading radiologist showed that 52% of the sampled MR images (54 of 103) had a score of 2 (Figure 1).

The overall Gwet's AC2 value of interrater agreement among the 6 readers (4 neuroradiologists and 2 resident readers) was calculated to be 0.83 (95% CI: 0.78, 0.87) with a percent agreement of 91%. Gwet's AC2 value of interrater agreement among the 4 neuroradiologists was calculated to be 0.84 (95% CI: 0.79, 0.89) with a percent agreement of 91%. Gwet's AC2 value of interrater agreement between the 2 residents was calculated to be 0.79 (95% CI: 0.72, 0.86) with a percent agreement of 90%. Although there was a slight difference between neuroradiologists and radiology residents interrater agreement, this difference was not statistically significant (χ^2 = 0.85; *p* value = 0.36). For cases showing improvement or no change (BT-RADS 0-2), the Gwet's index among the neuroradiologists was 0.94 (95% CI: 0.88, 0.96) with a percent agreement of 94%; that for the residents was 0.83 (95% CI: 0.75, 0.91) with a percent agreement of 90%. For cases with worsening imaging (BT-RADS 3a-4), the Gwet's index among the neuroradiologists was 0.66 (95% CI: 0.57, 0.76) with a percent agreement of 86%; that for the residents was 0.58 (95% CI: 0.34, 0.76) with a percent agreement of 81%.

All four neuroradiologists agreed on the same BT-RADS score in 57 of the 103 studies (55%) – the remaining 46 studies were subjected to consensus scoring and were given post-hoc reference scores as well. The percent agreement between consensus and post-hoc reference scoring was 74%. The percent agreement between neuroradiologist blinded scores and consensus scores ranged from 46%-65%; that between blinded scores and post-hoc reference scores ranged from 41%-52%. Faculty members with three years of clinical experience had higher agreement rates (Table 2). For cases showing disagreement between blinded scores and post-hoc reference scores, neuroradiologists generally underestimated BT-RADS scores (Table 3).

Across all studies, the median oldest comparison used was 4 months (IQR, 3-8 months). Raters used older MRI studies when assigning a BT-RADS score of 2 (median, 6 months; IQR, 3-12 months) compared to all other score categories combined (median, 3 months; IQR, 2-5 months). Based on the Mann-Whitney U test, the two compared median values are statistically different (*p* value, <.001). The month range for the oldest MR comparison used for BT-RADS score 3b appeared modestly increased compared to baseline (Figure 2).

For cases showing improvement or no change (BT-RADS scores 0-2), perfusion provided little contribution in the scoring process (mean, 1.2; range, 1.0-4.0). For cases with imaging worsening (BT-RADS scores 3a-4), there was a combined relative increase in the contribution from perfusion (mean, 2.1; range, 1.0-5.0). Contribution from perfusion was highest for score 3a (mean, 2.5; range, 1.0-5.0) and score 4 (mean, 2.6; range, 1.0-5.0) as shown in Figure 3. The mean perfusion contribution for the BT-RADS score 3a-4 group is larger than that for the BT-RADS score 0-2 group (two-sample t-statistic, 8.10; *p* value, <.001).

Overall, the readers' level of confidence in assigning score was moderate, regardless of level of training or years of clinical practice (Figure 4). The mean confidence for faculty scoring (mean, 4.0; range 1-5) and that for resident scoring (mean, 3.9; range 1-5) was not statistically different (paired t-statistic, 0.90; *p* value, 0.32).

Representative examples of scored studies are shown in Figures 5 and 6. Figure 5 shows a case of IDH-wild type glioblastoma in which all readers agreed. Enhancing treated tumor was noted 18 months post-surgery, and 2 months later, the abnormal enhancement had increased more than 25%. This study was scored with a BT-RADS score of 4 and met Response Assessment in Neuro-Oncology (RANO) criteria for progression. Figure 6 shows a case of IDH-wild type glioblastoma in which there was disagreement among the readers. Abnormal FLAIR was noted 3 months post-surgery and 1-month post-radiation therapy, with increasing enhancement 3 months post-radiation. This study was an early post-treatment study in which there was disagreement about whether imaging worsening constituted pseudoprogression or progressive tumor. The post-hoc reference score was BT-RADS 4.

DISCUSSION

Changes associated with brain tumor treatment may mimic tumor progression, making longitudinal assessment of brain tumor MR imaging difficult. BT-RADS was developed to improve reporting consistency, similar to other standardized reporting systems such as BI-RADS, which has shown acceptable to good interrater reliability (IRR). IRR of BT-RADS has not been widely studied, and the aim of our study was to assess variation between six readers with differing levels of clinical training in a department where BT-RADS has been implemented and readers had 1-2 years' experience prior to this study. A previous smaller study evaluated the IRR of BT-RADS using MR images from 23 patients,[19] but to the best of our knowledge this is the largest study on IRR of BT-RADS to date.

We found very good overall interrater agreement (Gwet's AC2 index, 0.83) among six readers, suggesting high BT-RADS consistency in routine clinical practice. Interrater agreement (Gwet's AC2 index, 0.83) was higher compared to that estimated by Parillo et al (Fleiss' kappa, 0.70)[19]. Compared to Fleiss' kappa, Gwet's index is more robust and resistant to the kappa paradox and less impacted by marginal probability and prevalence,[20, 21] which may result in Gwet's index being a better measure for this purpose.[16] Additionally, we used a larger patient sample and also assessed the added value of perfusion MRI. Our study produced similar interrater agreement results compared to other neuroradiology standardized reporting systems. For the Neck Imaging Reporting and Data System (NI-RADS), Elshotz et al reported Kendall's coefficient of concordance of 0.74 and 0.80 for the primary site and nodes respectively[22], and Hsu et al calculated

percent agreement of 84% and 93% for primary sites and nodes respectively[23]. The percent agreement rates for the Thyroid Imaging Reporting and Data System (TI-RADS) ranged from 70% to 87%[24]. All of these prior reports are comparable to this study (Gwet's AC2 index, 0.83, percent agreement 91%). In general, percent agreement values tend to be higher than Gwet's index or other kappa values because percent agreement does not account for agreement due to chance.

Our study demonstrated high agreement for the neuroradiologist's group (Gwet's AC2 index, 0.84) and the resident's group (Gwet's AC2 index, 0.79). This finding suggests that BT-RADS is applicable to all levels of experience, including trainees. Similarity in agreement between the two groups is likely related to the simplicity in the structure of BT-RADS in its reference scoring flow cart and detailed category guide.[8, 25] Even though the difference in agreement between the two groups was not statistically significant, there may be small differences below the level of detection for this study. Generally, agreement within the neuroradiologist and resident's groups was lower among cases with worsening imaging (BT-RADS 3a-4) compared to cases showing improvement or no change (BT-RADS 0-2). However, the values were much lower in the resident's group compared to the neuroradiologist's group. This is explained by the complexity of such cases even for experienced neuroradiologists. This variation likely exists with free text report as structured reporting does not create this variation, but likely allows us to quantitate that variation.

The overall percent agreement between consensus score and post-hoc reference score was good, compared to a moderate agreement between blinded neuroradiologists scores and post-hoc reference scores. During consensus scoring, the two scoring neuroradiologists utilized available information that individual readers were blinded to during the blinded scores, including imaging scores and scores from all 4 neuroradiologists. This provided more information that aided in the consensus scoring and resulted in higher percent agreement when compared with the post-hoc reference score. When compared to post hoc scores, BT-RADS scores did slightly underestimate progression. This is partially due to system design, which favors undercalling progression in the early posttreatment period to leave treatment decisions in the hands of the oncology treatment team, which may prefer to wait for subsequent follow-up before changing treatment or removing a patient from a clinical trial. DSC perfusion was found to be useful in scoring more complex cases (BT-RADS scores 3a-4) but provided little information for lower scoring cases (BT-RADS scores 0–2). This is consistent with previous reports that perfusion is useful adjunct or troubleshooting tool for indeterminate studies[26], but may not contribute to many reports on primary brain tumor cases. Although high confidence in a reporting system does not necessarily equate to assigning the correct score, it is a secondary measure of ease with use of the system. We recorded a moderate level of confidence in using BT-RADS which did not differ between the neuroradiologist and resident groups.

Our study had some limitations. This was a retrospective study conducted in a single facility where BT-RADS is used as part of standard practice in reading primary brain tumor. Differences in practice patterns, such as follow-up at shorter or longer clinical intervals, and range of experience of neuroradiologists interpreting brain tumor imaging may affect outcomes. In addition, we classified the tumors according to the 2016 WHO classification and they were not reclassified. For this study, we recruited only two radiology residents compared to four neuroradiologists. Lastly, we did not perform intra-rater agreement analysis to assess internal consistency among readers as it was beyond the scope of this study, but it is certainly a topic of future interest. A multi-center prospective study involving evaluation of MR images with varied distribution of more complex cases, inclusion of readers without prior institutional use of BT-RADS, inclusion of more radiology trainees, and conduction of intra-rater reliability as secondary analysis is warranted to further support the use of BT-RADS for standardization of brain tumor MRI reporting.

Table 1: Patient and Tumor Characteristics.

Characteristic	
Number of patients	98
Number of men	52 (53%)
Number of MRI exams evaluated	103
Median age and [interquartile range] (yr.)	53 [41 – 66]
Number of patients with previous surgery	97 (99%)
Number of patients with completed radiotherapy	89 (91%)
Number of patients on steroids	18 (18%)
Number of patients on bevacizumab	20 (20%)
Tumor Classification	
Astrocytoma	79 (77%)
Grade 2	14 (14%)
Grade 3	20 (19%)
Grade 4	44 (43%)
Anaplastic pilocytic astrocytoma	1 (1%)
Oligodendroglioma	24 (23%)
Grade 2	16 (15%)
Grade 3	8 (8%)
Tumor Type by IDH-mutational status	
Astrocytoma	
IDH-wild type	40 (51%)
IDH-mutant	36 (45%)
Unknown	3 (4%)
Oligodendroglioma	
IDH-mutant	19 (79%)
Unknown	5 (21%)

Table 2: Percent Agreement Between Neuroradiologist Blinded Score and Consensus Score; Blinded Score and Post-hoc Reference Score (n=46).

Reader	Agreement (%) with Consensus Score	Agreement (%) with Post-hoc Reference Score
Faculty_1	65	52
Faculty_2	46	46
Faculty_3	52	41
Faculty_4	63	48

Note:- Faculty_1, Faculty_3 and Faculty_4 had 3 years of practice experience. Faculty_2 had 1 year of practice experience

Table 3: Under- and Over- estimation of Disagreement of Neuroradiologist Blinded Score compared to Post-hoc Reference Score

Reader	Underestimation / Overestimation
Faculty_1	14 / 8
Faculty_2	12 / 13
Faculty_3	20 / 7
Faculty_4	17 / 7

Note:- Faculty_1, Faculty_3 and Faculty_4 had 3 years of practice experience. Faculty_2 had 1 year of practice experience

CONCLUSIONS

Our study shows good to very good agreement in using BT-RADS for post-treatment follow-up among experienced neuroradiologists and radiology residents. BT-RADS can thus be used to produce consistent and transparent reports even in less experienced hands. Our study also demonstrated similar interrater agreement results as compared to other RADS systems used in neuroradiology. BT-RADS has been adopted at multiple institutions within the US as well as internationally, and this report suggests implementing BT-RADS can help provide consistent reports across readers. With a moderate agreement between blinded scores and post-hoc scores, future studies to better understand the predictiveness of BT-RADS scores on future outcomes will be a great addition to the literature.

REFERENCES

1. Ostrom QT, Bauchet L, Davis FG, Deltour I, Fisher JL, Langer CE, et al. The epidemiology of glioma in adults: a "state of the science" review. Neuro Oncol. 2014;16(7):896-913.
2. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro Oncol. 2021;23(8):1231-51.
3. Willems E, Lombard A, Dedobbeleer M, Goffart N, Rogister B. The unexpected role of Aurora A kinase in glioblastoma recurrences. Targeted Oncology. 2017;12:11.
4. Delgado-Lopez PD, Corrales-Garcia EM. Survival in glioblastoma: a review on the impact of treatment modalities. Clin Transl Oncol. 2016;18(11):1062-71.
5. Wu W, Klockow JL, Zhang M, Lafortune F, Chang E, Jin L, et al. Glioblastoma multiforme (GBM): An overview of current therapies and mechanisms of resistance. Pharmacol Res. 2021;171:105780.
6. Ostrom QT, Patil N, Cioffi G, Waite K, Kruchko C, Barnholtz-Sloan JS. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013-2017. Neuro Oncol. 2020;22(12 Suppl 2):iv1-iv96.
7. Quant EC, Wen PY. Response assessment in neuro-oncology. Curr Oncol Rep. 2011;13(1):50-6.
8. Weinberg BD, Gore A, Shu HG, Olson JJ, Duszak R, Voloschin AD, et al. Management-Based Structured Reporting of Posttreatment Glioma Response With the Brain Tumor Reporting and Data System. J Am Coll Radiol. 2018;15(5):767-71.
9. Malik DG, Rath TJ, Urcuyo Acevedo JC, Canoll PD, Swanson KR, Boxerman JL, et al. Advanced MRI Protocols to Discriminate Glioma From Treatment Effects: State of the Art and Future Directions. Frontiers in Radiology. 2022;2.
10. Gore A, Hoch MJ, Shu HG, Olson JJ, Voloschin AD, Weinberg BD. Institutional Implementation of a Structured Reporting System: Our Experience with the Brain Tumor Reporting and Data System. Acad Radiol. 2019;26(7):974-80.
11. Bignotti B, Calabrese M, Signori A, Tosto S, Valdora F, Tagliafico A, et al. Background parenchymal enhancement assessment: Inter- and intra-rater reliability across breast MRI sequences. Eur J Radiol. 2019;114:57-61.
12. El Khoury M, Lalonde L, David J, Labelle M, Mesurolle B, Trop I. Breast imaging reporting and data system (BI-RADS) lexicon for breast MRI: interobserver variability in the description and assignment of BI-RADS category. Eur J Radiol. 2015;84(1):71-6.
13. Grimm LJ, Anderson AL, Baker JA, Johnson KS, Walsh R, Yoon SC, et al. Interobserver Variability Between Breast Imagers Using the Fifth Edition of the BI-RADS MRI Lexicon. AJR Am J Roentgenol. 2015;204(5):1120-4.
14. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol. 2016;131(6):803-20.
15. Gwet KL. Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters: Advanced Analytics, LLC; 2014.
16. Popplewell M, Reizes J, Zaslawski C. Appropriate Statistics for Determining Chance-Removed Interpractitioner Agreement. The Journal of Alternative and Complementary Medicine. 2018;25(11):1115-20.
17. Altman DG. Practical Statistics for Medical Research: Taylor & Francis; 1990.
18. Gwet KL. irrCAC: Computing Chance-Corrected Agreement Coefficients (CAC). R package version 1.0 2019 [Available from: https://CRAN.R-project.org/package=irrCAC. Accessed September 18, 2023
19. Parillo M, Mallio CA, Pileri M, Dirawe D, Romano A, Bozzao A, et al. Interrater reliability of Brain Tumor Reporting and Data System (BT-RADS) in the follow up of adult primary brain tumors: a single institution experience in Italy. Quantitative Imaging in Medicine and Surgery. 2023;0(0):0-.
20. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol. 2013;13:61.
21. Tong F, Tang S, Irby BJ, Lara-Alecio R, Guerrero C. The determination of appropriate coefficient indices for inter-rater reliability: Using classroom observation instruments as fidelity measures in large-scale randomized research. International Journal of Educational Research. 2020;99.
22. Elsholtz FHJ, Ro SR, Shnayien S, Erxleben C, Bauknecht HC, Lenk J, et al. Inter- and Intrareader Agreement of NI-RADS in the Interpretation of Surveillance Contrast-Enhanced CT after Treatment of Oral Cavity and Oropharyngeal Squamous Cell Carcinoma. AJNR Am J Neuroradiol. 2020;41(5):859-65.
23. Hsu D, Rath TJ, Branstetter BF, Anzai Y, Phillips CD, Juliano AF, et al. Interrater Reliability of NI-RADS on Posttreatment PET/Contrast-enhanced CT Scans in Head and Neck Squamous Cell Carcinoma. Radiol Imaging Cancer. 2021;3(3):e200131.
24. Chung R, Rosenkrantz AB, Bennett GL, Dane B, Jacobs JE, Slywotzky C, et al. Interreader Concordance of the TI-RADS: Impact of Radiologist Experience. American Journal of Roentgenology. 2020;214(5):1152-7.
25. Abidi SA, Hoch MJ, Hu R, Sadigh G, Voloschin A, Olson JJ, et al. Using Brain Tumor MRI Structured Reporting to Quantify the Impact of Imaging on Brain Tumor Boards. Tomography. 2023;9(2):859-70.
26. Yang Y, Yang Y, Wu X, Pan Y, Zhou D, Zhang H, et al. Adding DSC PWI and DWI to BT-RADS can help identify postoperative recurrence in patients with high-grade gliomas. J Neurooncol. 2020;146(2):363-71.