Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Contrast-Enhancing Tumor Margin Detection in Gliomas using Non-Contrast MRI: From Human-Only to Human-AI Assisted Assessment

Ivar J.H.G. Wamelink
Amsterdam University Medical Centers
Aynur Azizova
Amsterdam University Medical Centers
Elif Kaya
Ankara Yıldırım Beyazıt University
Emmanouil Koltsakis
Karolinska Institutet
Joao N. P. Ramos
Amsterdam University Medical Centers
Aziz Anil Tan
Ege University
Norman Kornemann
Medizinische Hochschule Hannover
Frederik Barkhof
Amsterdam University Medical Centers
Alle Meije Wink
Amsterdam University Medical Centers
Vera C. Keil
v.c.w.keil@amsterdamumc.nl

Amsterdam University Medical Centers

Research Article

Keywords: Glioblastoma, Segmentation, Artificial Intelligence, Gadolinium-Based Contrast Agents, Glioma

Posted Date: August 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10237490/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Purpose

Contrast-enhanced MRI is the cornerstone of neuro-oncological imaging, but concerns regarding sustainability and cost prompted interest in the development of gadolinium-based contrast agent (GBCA)-free alternatives. We investigated the ability of human raters to guess the contrast-enhancing tumor borders in glioblastoma using GBCA-free MRI and whether a deep learning algorithm can improve this delineation.

Methods

A segmentation model was trained using BraTS (n = 989) and UCSF-PDGM (n = 501) datasets. Ground truth segmentations of the enhancing tumor regions were generated using an established algorithm and corrected by an independent neuroradiologist. T1-weighted (T1w), T2-weighted (T2w), T2-FLAIR, DWI-B0 and B-1000, ADC MRI of 117 glioblastoma, IDH-wildtype, patients were independently re-segmented by three radiologists and one non-clinician. Another radiologist (R-AI) improved segmentations produced by an AI algorithm without post-contrast sequence input. Statistical analysis compared AI segmentations with human performance.

Results

All raters showed a moderate-to-strong correlation with the ground truth ($r = 0.79$; $P < 0.001$). AI assistance increased Dice similarity by 0.09 (0.82 vs. 0.72; $P < 0.001$), greatly reduced Hausdorff distance (9.97 vs. 16.16 mm; $P < 0.001$), and decreased absolute volume error by 10.3 ml (6.75 vs. 17.07 ml; $P < 0.001$). Rater confidence increased with larger tumor volume, despite nonsignificant volumetric error metrics after correction (all $P \geq 0.24$).

Conclusion

Human raters predicted enhancing tumor areas from GBCA-free sequences with moderate accuracy, showing considerable inter-rater variability in glioblastoma. AI-based GBCA-free segmentation assistance significantly improved human segmentations.

Introduction

Contrast-enhanced magnetic resonance imaging (CE-MRI) is central in neuro-oncological practice as it increases lesion visibility and highlights areas of blood-brain barrier disruption providing critical information for lesion characterization.[1] Enhancement pattern, intensity, and extent serve as imaging

biomarkers facilitating histomolecular glioma classification and grading. In glioblastoma, the enhancing tumor borders define the surgical resection margin and guide radiotherapy planning.[2, 3]

Safety, sustainability, and cost issues around GBCAs have prompted growing pharmaceutical surveillance and research into GBCA-free imaging alternatives.[4–8] GBCA-free pulse sequences, such as FLAIR, can guide surgical decision-making and, in some scenarios, superiorly reflect therapeutic outcomes.[9, 10] Deep learning (DL) analyses of GBCA-free imaging are a more recent approach to limit contrast agent exposure.[11–14]

Despite these developments, any transition to GBCA-free imaging must ensure that the enhancing tumor parts can still be reliably identified, because treatment margins fundamentally depend on accurate segmentation of this region. Glioma segmentation typically distinguishes three classes: 1) non-enhancing tumor/edema, 2) necrosis, and 3) enhancing tumor volumes. Manual delineation of these subregions is time-consuming and impractical for routine workflows. Consequently, automated DL-based segmentation approaches using standard MRI sequences have been extensively investigated as alternatives. However, these models rely on CE-MRI as the reference signal, limiting their applicability to GBCA-free imaging.[15, 16]

Recent work explored whether human raters can infer enhancement characteristics from GBCA-free imaging. Their proposed decision tree model for predicting enhancement intensity and morphology demonstrated sufficient diagnostic accuracy.[17] However, the proposed method did not provide spatial information about enhancing areas, which precludes its use for treatment decision-making. Second, knowledge of human benchmark performance and the impact of added AI assistance is crucial for further developing the field of GBCA avoidance using AI approaches in neuroradiology, e.g., explainable AI and supervised learning.

To narrow this hiatus, this study aims to 1) evaluate the ability of human raters to identify the enhancing tumor regions, i.e. the enhancing volume covering the central necrosis, in glioblastoma patients using standard GBCA-free MR sequences. We further assessed how the raters justified their decisions, and if DL-assisted segmentations based on GBCA-free imaging improve human segmentation accuracy.

Materials and Methods

The study was approved by our Institutional Review Board (no. XXX), and informed consent was waived due to the retrospective study design.

Datasets

We included MRI data from 117 untreated glioblastoma patients with IDH-wildtype, WHO grade 4, classified according to the WHO CNS5 (2021) criteria. Cases were randomly drawn from the IMAging in GliOma (IMAGO) database.[18] Clinical information included IDH mutation status, overall survival, and subsequent therapy course.

MRI data included pre- and post-contrast T1-weighted (T1w and CET1), T2-weighted (T2w), T2-FLAIR, diffusion-weighted imaging (DWI B0 and B1000), and apparent diffusion coefficient (ADC) images. Ground truth (GT) segmentations from the IMAGO database included annotations for non-enhancing tumor/edema (1), necrosis (2), and enhancing tumor (3), generated using the PICTURE segmentation algorithm on CET1, T1w, T2w, and T2-FLAIR images, converted into binary masks including the tumor core (enhancing tumor and necrosis).[15] GT segmentations were independently reviewed and refined by a board-certified neuroradiologist with 5 years of experience (XXXX). Delineation of the enhancing tumor margin was performed on axial slices and subsequently validated across coronal and sagittal planes.

The GBCA-free segmentation algorithm was trained on the BraTS 2021 (n = 1,251)[19–21] and UCSF-PDGM (n = 501)[22] datasets. Due to overlap with the UCSF-PDGM dataset, 262 BraTS cases were excluded, leaving 1,490 cases for training. The IMAGO dataset has demographic distributions similar to those of UCSF-PDGM (male: 59% vs. 60%; mean age: 59 ± 10 vs. 57 ± 15 years). Clinical data for BraTS were unavailable. Only 3 of the IMAGO cases were non-enhancing for testing rater and algorithm capacity to identify non-enhancing tumors.

DL algorithm

The GBCA-free segmentation algorithm was developed using the nn-UNet architecture.[23] No additional model fine-tuning was performed on the IMAGO data. For details on model training, see[23]. Three GBCA-free structural MRI sequences (pre-contrast T1w, T2w, and T2-FLAIR) were used as inputs in an 80:20 training-to-validation split, and the output was a three-level mask similar to the one produced in the PICTURE segmentations. Since the human rater segmentations only captured the tumor core (enhancing lesion boundaries plus central necrosis), the AI-generated three-label predictions were also converted into binary masks (enhancing + necrotic regions). Training was performed on one NVIDIA A100-SXM4 (80GB) GPU, using 10 CPU cores and 50GB of RAM. The algorithm is available upon request.

Human rater segmentation

Five raters (E with 3 months of neuroradiology training and R-AI, B, C, D with 5, 5, 7, and 7 years of neuroradiology experience, respectively) performed GBCA-free manual segmentations using 3D Slicer software. Raters received training on 3D Slicer usage (by XXXX) and GBCA-free prediction of the resectable tumor margin, defined as the outer enhancing margin (by XXXX). The imaging features identifying the enhancing tumor margin were: necrosis, diffusion restriction, and T2 heterogeneity based on the findings of our previous study.[17]

Segmentations were performed in five rounds, each with 20 cases. Four raters separately segmented lesions in the same order. After each set, raters were automatically directed to review post-contrast images without modifying their segmentations. This iterative process allowed them to self-train by identifying errors and improving their segmentation accuracy. During each round of 20 cases, raters

completed a per-case questionnaire assessing (1) which sequence guided the segmentation (T2w, T1w + T2w, DWI, T2w + DWI), (2) how confident they were with their segmentation (5-point scale), (3) how difficult the case was to segment (5-point scale). Segmentation time was automatically tracked by the segmentation tool and was used for a time-quality tradeoff analysis (4). During the post-contrast review of each round, each rater was asked (per case) whether they identified the errors in their segmentation and whether they identified additional imaging clues beyond necrosis, diffusion restriction, and T2 heterogeneity for subsequent decision-making. One Rater (E) was excluded from answering all questions and segmented all 117 cases successively.

Additionally, an independent rater, R-AI, reviewed and refined the GBCA-free AI-based segmentation in 3D Slicer. This rater followed the same structured five-round approach as described earlier. A flowchart of the segmentation workflow is shown in Fig. 1.

Analysis

All analyses were performed in Python (Version 3.11) using the SciPy package (Version 1.14.1). A P-value < .05 adjusted with Bonferroni correction for multiple comparisons was considered statistically significant.

Geometric Accuracy

To evaluate the ability of human raters to define the tumor core on GBCA-free MRI, we performed quantitative analyses compared to the GT. Segmentation similarity was assessed using the: (1) Dice Similarity Coefficient (DC) which measures voxel-level spatial overlap between the segmentation and the GT, (2) Intraclass correlation coefficient (ICC) that quantifies volumetric agreement across raters and with the GT (single random raters; poor < 0.5, moderate 0.5–0.75, good 0.75–0.9, excellent > 0.9)[24], (3) Hausdorff distance (HD) that measures boundary agreement between segmentations, and (4) generalized conformity index (GCI) which quantifies the spatial agreement among multiple segmentations. It generalizes overlap measures such as the dice coefficient to more than two raters.

Questionnaire Analyses

We performed exploratory analyses: (1) Which MRI sequences most strongly guided human rater segmentations, and whether reliance on specific sequences was associated with errors or time spent per case, (2) analysis of segmentation confidence, comparing self-reported rater confidence against actual segmentation performance and time, (3) analysis of case difficulty, examining whether cases rated as more “difficult” correlated with lower segmentation accuracy or longer segmentation times, and (4) analysis of time-quality tradeoff to examine relationships between segmentation time and accuracy (DC, volume error). For the segmentation confidence and case difficulty analyses, we formed case-level groups using the majority rating among raters. This yielded eight confidence groups and eight difficulty

groups. For each construct (confidence and difficulty), five groups corresponded to the ordinal majority ratings from 1 to 5 (e.g., from “easy” (majority scored 1) to “difficult” (majority scored 5), and from “not confident” (majority scored 1) to “highly confident” (majority scored 5)). The remaining three groups captured different agreement patterns: (i) total agreement (all raters gave the same rating), (ii) total disagreement (no clear majority), and (iii) a “two rater pairs agree” pattern (two raters sharing one rating and the other two sharing another rating). Unless stated otherwise, segmentation performance was assessed for both human raters and R-AI. Inter-rater comparisons were evaluated with the Kruskal–Wallis test and post-hoc pairwise Mann-Whitney tests. Cliff’s δ was employed to quantify the effect size of differences in Dice Coefficient (DC) across confidence and difficulty groups. Bowker’s test was applied to evaluate whether raters exhibited systematic asymmetry in rating cases as easy/confident or difficult/not confident. We additionally reviewed the non-enhancing cases.

Results

Quantitative Analysis

Segmentation performance of the five raters is summarized in Table 1. The four manual raters achieved moderate DC scores (mean 0.71–0.76; SD 0.19–0.25), whereas the AI model and R-AI had higher DC scores (0.83 ± 0.18 and 0.82 ± 0.19, respectively; Fig. 2).

R-AI achieved significantly higher DC than each manual rater (mean difference 0.08–0.12; all $P < 0.001$ after Bonferroni correction) and a higher DC than the average of all manual raters (mean difference 0.09; $P < 0.001$).

A similar pattern was observed for Hausdorff Distance, where R-AI showed smaller Hausdorff Distances with the GT (mean reduction of 5–7 mm; $P < 0.001$) than every manual rater and then the average among manual raters. Volume analysis using absolute error relative to GT demonstrated that the R-AI also produced more accurate volumetric estimates, with a reduction in mean absolute error of approximately 10 mL compared to the mean absolute error of the manual raters ($P < 0.001$). All P-values for pairwise and mean comparisons were derived from one-sided paired t-tests and Wilcoxon signed-rank tests, with Bonferroni correction applied for multiple pairwise comparisons. These findings were confirmed in mixed-effects models with patients as a random effect, in which AI assistance was associated with higher DC and lower Hausdorff distance and volume error (all P-values < 0.001). All human raters showed a tendency to oversegment, with one exception (Rater D, Table 1 and Fig. 3).

Table 1
Tumor core segmentation performance across different raters

Rater	Average volume (mL)	Average over/under estimation (mL)	Dice mean (± SD)	HD mean (± SD)	ICC
AI	46	2.65	0.83 ± 0.18	10.08 ± 9.84	0.98
R-AI	47.4	4.05	0.82 ± 0.19	9.49 ± 7.44	0.91
B	48.23	4.88	0.76 ± 0.20	14.05 ± 10.54	0.63
C	66.39	23.04	0.71 ± 0.21	16.54 ± 10.59	0.55
D	42.76	-0.59	0.73 ± 0.22	14.86 ± 14.07	0.52
E	61.02	17.67	0.71 ± 0.25	17.51 ± 13.64	0.46

Four human raters (B-E) and one human rater refining AI GBCA-free segmentations (R-AI) depicted. The GBCA-free AI algorithm was trained on pre-contrast T1w, T2w, and T2-FLAIR images. Mean GT tumor volume was 43.35 mL. *SD: standard deviation, HD: Hausdorff distance, ICC: Intraclass Correlation Coefficient.*

Overall accuracy analysis showed a mean absolute error (MAE) of 13.31 mL, a root mean square error (RMSE) of 24.05 mL, and a correlation coefficient of $r = 0.789$ ($P < 0.001$) with the GT, despite substantial variability between raters.

Inter-Rater Analysis

Inter-rater agreement analysis showed moderate agreement depending on the group of raters considered (Table 2). Excluding the least experienced rater (E) improved the agreement from 0.72 to 0.77. Figure 4 shows pairwise correlations between raters' DC scores and predicted volumes, highlighting patterns of variable but generally moderate-to-strong agreement.

Table 2
Inter-rater reliability for tumor core segmentation

Group	ICC	GCI (mean ± SD)
All raters + AI	0.71*	0.61 ± 0.2
All raters	0.72*	0.6 ± 0.19
Without rater E	0.77*	0.63 ± 0.19
Without AI and without rater E	0.77*	0.6 ± 0.19
AI and R-AI	0.97*	0.92 ± 0.15

Intraclass correlation coefficients (ICC) and Generalized Conformity Index were calculated across different groups of raters. *P-value was smaller than 0.0001.

Segmentation Justification Questionnaire

Segmentation accuracy differed significantly across sequence-decision subsets, reflecting the varying contribution of MRI sequences to delineating the tumor core margin (Kruskal–Wallis H = 18.0, P < 0.001). The highest performance (mean DC 0.84, 95% CI 0.78–0.89) was achieved when T1w and T2w images were used for segmentation. Lower accuracy was observed when T2w alone (mean DC 0.78, 95% CI 0.75–0.81) and T2w combined with DWI (mean DC 0.76, 95% CI 0.74–0.78) were the main guidance and even lower when DWI was the primary driver (mean DC 0.58, 95% CI 0.44–0.72). Pairwise comparisons confirmed significant differences between T2w vs T2w + DWI (P < 0.01, effect size = 0.20) and T2w vs DWI (P = 0.02, effect size = 0.34), whereas other contrasts were not significant after correction (Table 3).

Table 3
Pairwise comparisons of sequence subsets

Comparison	P-value	P-value corrected	Effect size
T2w vs T2w + DWI	*0.001*	**0.00**	0.20
T2w vs DWI	*0.001*	**0.02**	0.34
T2w + DWI vs DWI	*0.04*	0.27	0.23
T1w + T2w vs DWI	0.30	1	0.38
T2w + DWI vs T1w + T2w	0.48	1	0.24
T2w vs T1w + T2w	0.80	1	0.09

Pairwise comparisons of sequence subsets deemed most helpful for segmentation. Only T2w vs T2w + DWI and T2w vs DWI remained significantly different after correction.

Segmentation Confidence Questionnaire

Rater agreement on confidence scores was significantly associated with segmentation performance (DC), but not with time or volumetric error. Global testing revealed differences in DC across confidence groups (Kruskal–Wallis H = 15.99, P = 0.01). DC was highest when raters had perfect agreement (median 0.88) or majority high confidence (0.83–0.84) and lowest when confidence was split or very low (median was 0.23 when not confident, though n = 2). Time (H = 4.74, P = 0.58) showed no effect with segmentation confidence. Larger tumors tended to fall into the high-agreement groups, but volumetric error metrics (all P ≥ 0.24) did not remain significant after correction.

Case Difficulty Questionnaire

Global testing across case-difficulty groups revealed no significant differences in segmentation time or volumetric error (all P > 0.05). DC was different between difficulty groups (Kruskal–Wallis H = 26.34, P < 0.001), but the only pairwise contrast that remained significant after correction was between moderately-easy and difficult, with a strong separation (Cliff's δ = 0.979). Other contrasts showed trends toward differences, but none were significant after adjustment for multiple comparisons, likely due to very small subgroup sizes (n ≤ 3 for easy and difficult).

Rater B consistently rated cases as more difficult than other raters (Bowker P ≤ 10^{-6}), while rater C tended to rate cases as easier. R-AI and rater D showed the most similar rating patterns.

Time–Quality Trade-off

Segmentation time analysis revealed significant differences between raters (ANOVA F = 88.384, P < 0.001; Kruskal–Wallis H = 265.714, P < 0.001). Rater D was the fastest (4.29 ± 2.79 min per case), followed by R-AI (12.08 ± 7.19 min), B (16.68 ± 14.81 min), and C (33.37 ± 22.83 min) (Fig. 5A, C). A significant decrease in time per case was observed (r=–0.304, P = 0.001), consistent with a learning effect (Fig. 5B). Importantly, segmentation time did not correlate with DC scores (r=–0.001, P = 0.977), indicating that dedicating more time did not yield higher segmentation accuracy.

Non-enhancing GBM identification

Identification of non-enhancing cases was difficult, as well as tumors with an enhancing volume of less than 5mL (Fig. 6).

Supplementary Table S1 contains an overview of missed or wrongly segmented cases. Figure 7 shows a left-sided enhancing lesion that was consistently and accurately identified by all raters. In contrast, the right side shows a non-enhancing lesion that most raters misclassified.

Discussion

In this retrospective study, we evaluated whether contrast enhancement marking the tumor core of glioblastoma IDH-wt patients can be delineated on GBCA-free MR images by human raters alone and if an AI algorithm trained exclusively on GBCA-free sequences (R-AI) can improve this delineation. R-AI

achieved higher segmentation accuracy than human raters when only provided with GBCA-free sequences, as evidenced by quantitative analysis showing statistically significantly higher DC scores, lower Hausdorff distances, and closer volumetric agreement with the ground truth. Human performance was varied and rater-dependent, with systematic volume bias (over- or undersegmentation) and moderate inter-rater agreement. Segmentation performance varied primarily with the decision-guiding images: T1w+T2w yielded the best performance, whereas DWI-guided segmentation decisions were most error-prone. Overall, GBCA-free delineation of the tumor core is feasible, and consistency and reliability improve with AI support and careful sequence selection. Segmentation time per case decreased while segmentation performance did not decrease, this suggests a learning effect. Segmentation performance was highest when raters had high agreement in their confidence ratings and was also associated with perceived case difficulty, while segmentation time and volumetric error were unaffected.

Most prior brain tumor segmentation studies relied on post-contrast T1-weighted imaging to define enhancing tumor margins. In those settings, strong models achieve enhancing tumor margins DC values between 0.70 and 0.90 for strong models, with whole-tumor DC often higher due to non-enhancing/edema contributions.[16, 25, 26] Our GBCA-free model, trained only on pre-contrast T1w, T2w, and T2-FLAIR, achieved a similar DC score to that of many contrast-dependent segmentation algorithms. This indicates that a substantial fraction of enhancing-margin information is encoded in structural sequences via secondary markers (e.g., necrosis, diffusion restriction, and T2 heterogeneity).[17] High average DCs for AI and R-AI (0.83 and 0.82, respectively) suggest that, in many cases, contrast may be unnecessary for defining the resectable tumor core margin in glioblastoma IDH-wt patients. This would be especially beneficial for patients with renal insufficiency, prior adverse reactions, vulnerable populations, including pregnant and breastfeeding women and children, and circumstances where shorter protocols and lower cost are prioritized.[27–29] However, our qualitative analysis of non-enhancing cases underscores that identifying these cases remains difficult for both humans and AI, even in our small sample.

The suggested segmentation protocol provides a robust and practical foundation for GBCA-free brain tumor segmentations and could serve as a standard for future studies. However, human raters tended to oversegment, and accuracy analysis revealed rater-specific differences in error patterns. Some raters consistently showed positive signed errors with larger absolute errors, while one rater tended to underestimate. This pattern aligns with prior reports of inter-rater variability and systematic bias in glioma delineation.[30] Dice, Hausdorff distance, and volumetric error improved significantly when a human was assisted by AI. Improvements to the protocol, such as advanced imaging biomarkers that are derived from APT-CEST and ASL might reduce these consistent errors.[31, 32]

The differences in volume definition compared to the standard clinical GT may have clinical consequences, given that the mean tumor volume was larger than tumor volume derived from CET1 sequences. Similarly, the mean resected volume is larger when approaching surgical removal of lesions based on T2-FLAIR (FLAIRectomy), which may extend far beyond contrast-enhancing regions.[9, 10]

While resection or radiation of the enhancing tumor core remains the gold standard, the approach taken in this research is likely to have intermediate results between the gold standard (CE-based resection) and FLAIRectomy. Whereas with FLAIRectomy the additional morbidity is estimated from larger areas using FLAIR as a basis, the resection based on the approach suggested in this paper may not exceed the disease burden caused by FLAIR and will likely result in resection of all CE tumor, leading to clinically comparable results.

When raters used T1w+T2w sequences to predict enhancing tumor areas, they obtained the highest DC; when T2w alone was used, the DC was intermediate; and DWI-led decisions were the least accurate. This is consistent with the biological and imaging expectations: structural heterogeneity, necrotic architecture, and mass effect are visible on T1w/T2w/FLAIR, whereas DWI signal can be nonspecific in GBM and may reflect cytotoxic edema, cellularity, or treatment effects heterogeneously.

All raters showed a learning effect, with time per case decreasing, but there was no correlation between time and accuracy. This suggests that segmentation efficiency gains were not at the expense of segmentation quality.

Identifying non-enhancing GBM cases was difficult. The AI correctly identified one of three; R-AI missed two and introduced one false non-enhancing label. Among human raters, identification of non-enhancing GBM cases ranged from failure to detect non-enhancing disease to marked overestimation of very small enhancing lesions. A retrospective review of misclassified cases by rater R-AI suggested a lack of canonical GBCA surrogates in non-enhancing patients: no T2 heterogeneity, no diffusion restriction, and no necrosis-like signal. While informative, this single-rater perspective does not explain why others confidently labeled the same lesions as enhancing and, in some instances, inflated volumes. These error types are clinically relevant: small-volume overcalls may drive overtreatment or misleading response assessment, while misses increase the risk of undertreatment. Segmentation of small lesions and non-enhancing cases is notoriously difficult.

Several limitations should be acknowledged. The study was limited to glioblastoma, IDH-wt, patients, a selection that may bias raters to find enhancement and could explain the poor identification of non-enhancing GBMs. While we observed good performance on this specific subset, additional research across other glioma subtypes is needed to assess performance in the general glioma population. Such broader analysis would strengthen the non-enhancing tumor analyses, particularly since lower-grade gliomas are generally non-enhancing.[33] In the questionnaire analyses, some subgroup analyses had small sample sizes after stratification, reducing the power of pairwise contrasts even when global tests were significant.

Conclusion

A GBCA-free sequence-based delineation of the enhancing tumor core in GBM is feasible and, when augmented by a GBCA-free AI algorithm, achieves accuracy that approaches typical contrast-dependent benchmarks and improves correlation. Non-enhancing and small GBM lesions remain difficult to identify

and segment. AI assistance offers improved tumor core identification and may assist in the safe reduction of contrast use in routine practice.

Declarations

Author Contribution

Overview of each co-author's contribution:I.J.H.G. W.: Concept, Design, Programming, Statistics, WritingA. A.: Rating, Segmenting, Writing, DesignE. Ka.: Rating, Segmenting, WritingE. Ko.: Rating, Segmenting, WritingJ.N.P. R.: Rating, Segmenting, WritingA.A. T.: Rating, Segmenting, WritingN. K.: Rating, Segmenting, WritingF.B.: Concept, Writing, EditingA.M. W.: Programming, Segmenting, EditingV.C. K.: Project Lead, Supervision, Writing, Editing

Data Availability

The MRI data can be found via XNAT Health-RI or EUCAIM: https://www.nature.com/articles/s41597-026-07424-8.The code for the deep learning model can be found here: https://github.com/IWamelink/NEURAL-MRI.git. (Request to Ivar Wamelink).

References

1. Weller M, van den Bent M, Preusser M et al (2020) EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. Nat Rev Clin Oncol 18:170–186
2. Wamelink IJHG, Azizova A, Booth TC et al (2024) Brain tumor imaging without gadolinium-based contrast agents: Feasible or fantasy? Radiology 310:e230793
3. Morana G, Cugini C, Scatto G et al (2013) Use of contrast agents in oncological imaging: magnetic resonance imaging. Cancer Imaging 13:350–359
4. FDA (ed) (2018) https://www.fda.gov/drugs/drug-safety-and-availability/fda-drug-safety-communication-fda-warns-gadolinium-based-contrast-agents-gbcas-are-retained-body
5. Runge VM (2017) Critical Questions Regarding Gadolinium Deposition in the Brain and Body After Injections of the Gadolinium-Based Contrast Agents, Safety, and Clinical Recommendations in Consideration of the EMA’s Pharmacovigilance and Risk Assessment Committee Recommendation for Suspension of the Marketing Authorizations for 4 Linear Agents. Invest Radiol 52:317–323
6. Scarciglia A, Papi C, Romiti C et al (2025) Gadolinium-based contrast agents (GBCAs) for MRI: A benefit-risk balance analysis from a chemical, biomedical, and environmental Point of View. Global Chall 9:2400269
7. Krohn LM, Klimpel F, Béziat P, Bau M (2024) Impacts of COVID-19 and climate change on wastewater-derived substances in urban drinking water: Evidence from gadolinium-based contrast agents in tap water from Berlin, Germany. Water Res 259:121847

8. Zhang Z, Jiang W, Gu T et al (2024) Anthropogenic gadolinium contaminations in the marine environment and its ecological implications. Environ Pollut 359:124740
9. Certo F, Altieri R, Maione M et al (2021) FLAIRectomy in supramarginal resection of glioblastoma correlates with clinical outcome and survival analysis: A prospective, single institution, case series. Oper Neurosurg (Hagerstown) 20:151–163
10. Haddad AF, Young JS, Morshed RA, Berger MS (2022) FLAIRectomy: Resecting beyond the contrast margin for glioblastoma. Brain Sci 12:544
11. Ruffle JK, Mohinta S, Gray R et al (2023) Brain tumour segmentation with incomplete imaging data. Brain Commun 5:fcad118
12. Liu J, Jakary A, Villanueva-Meyer JE et al (2024) Automatic brain tissue and lesion segmentation and multi-parametric mapping of contrast-enhancing gliomas without the injection of contrast agents: A preliminary study. Cancers (Basel) 16:1524
13. Iliadis G, Kotoula V, Chatzisotiriou A et al (2012) Volumetric and MGMT parameters in glioblastoma patients: survival analysis. BMC Cancer 12:3
14. Ellingson BM, Kim GHJ, Brown M et al (2022) Volumetric measurements are preferred in the evaluation of mutant IDH inhibition in non-enhancing diffuse gliomas: Evidence from a phase I trial of ivosidenib. Neuro Oncol 24:770–778
15. Pemberton HG, Wu J, Kommers I et al (2023) Multi-class glioma segmentation on real-world data with missing MRI sequences: comparison of three deep learning algorithms. Sci Rep 13:18911
16. Wen L, Sun H, Liang G, Yu Y (2025) A deep ensemble learning framework for glioma segmentation and grading prediction. Sci Rep 15:4448
17. Azizova A, Wamelink IJHG, Prysiazhniuk Y et al (2024) Human performance in predicting enhancement quality of gliomas using gadolinium-free MRI sequences. J Neuroimaging 34:673–693
18. Wamelink IJHG, Wink AM, Verburg N et al (2026) Amsterdam IMAging and clinical GliOma dataset; IMAGO. Sci Data 1–8
19. Baid U, Ghodasara S, Mohan S et al (2021) The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv [cs.CV]
20. Menze BH, Jakab A, Bauer S et al (2015) The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans Med Imaging 34:1993–2024
21. Bakas S, Akbari H, Sotiras A et al (2017) Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data 4:170117
22. Calabrese E, Villanueva-Meyer JE, Rudie JD et al (2022) The university of California San Francisco preoperative diffuse glioma MRI dataset. Radiol Artif Intell 4:e220058
23. Isensee F, Jaeger PF, Kohl SAA et al (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18:203–211

24. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15:155–163
25. Hatamizadeh A, Nath V, Tang Y et al (2022) Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. Lecture Notes in Computer Science. Springer International Publishing, Cham, pp 272–284
26. Zeng X, Zeng P, Tang C et al (2023) DBTrans: A Dual-Branch Vision Transformer for Multi-Modal Brain Tumor Segmentation. Lecture Notes in Computer Science. Springer Nature Switzerland, Cham, pp 502–512
27. Anazodo UC, Ng JJ, Ehiogu B et al (2023) A framework for advancing sustainable magnetic resonance imaging access in Africa. NMR Biomed 36:e4846
28. Proença F, Guerreiro C, Sá G, Reimão S (2021) Neuroimaging safety during pregnancy and lactation: a review. Neuroradiology 63:837–845
29. Blumfield E, Moore MM, Drake MK et al (2017) Survey of gadolinium-based contrast agent utilization among the members of the Society for Pediatric Radiology: a Quality and Safety Committee report. Pediatr Radiol 47:665–673
30. Visser M, Müller DMJ, van Duijn RJM et al (2019) Inter-rater agreement in glioma segmentations on longitudinal MRI. NeuroImage Clin 22:101727
31. Henriksen OM, Del Mar Álvarez-Torres M, Figueiredo P et al (2022) High-Grade Glioma Treatment Response Monitoring Biomarkers: A Position Statement on the Evidence Supporting the Use of Advanced MRI Techniques in the Clinic, and the Latest Bench-to-Bedside Developments. Part 1: Perfusion and Diffusion Techniques. Front Oncol 12:810263
32. Booth TC, Wiegers EC, Warnert EAH et al (2021) High-Grade Glioma Treatment Response Monitoring Biomarkers: A Position Statement on the Evidence Supporting the Use of Advanced MRI Techniques in the Clinic, and the Latest Bench-to-Bedside Developments. Part 2: Spectroscopy, Chemical Exchange Saturation, Multiparametric Imaging, and Radiomics. Front Oncol 11:811425
33. Guillevin R, Herpe G, Verdier M, Guillevin C (2014) Low-grade gliomas: the challenges of imaging. Diagn Interv Imaging 95:957–963

Figures

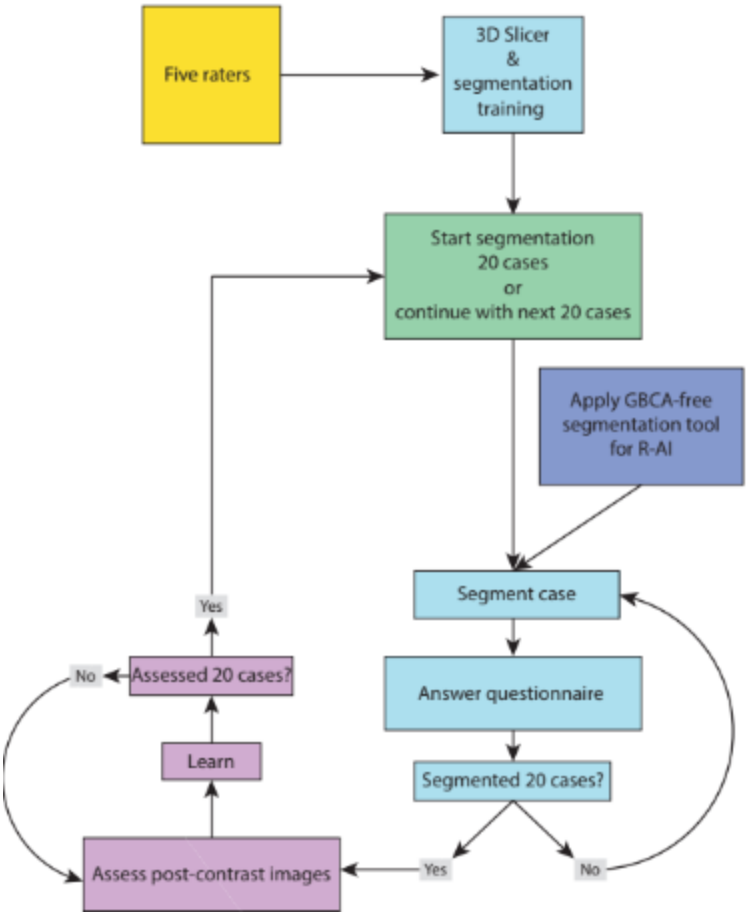

Figure 1

Segmentation analysis workflow.

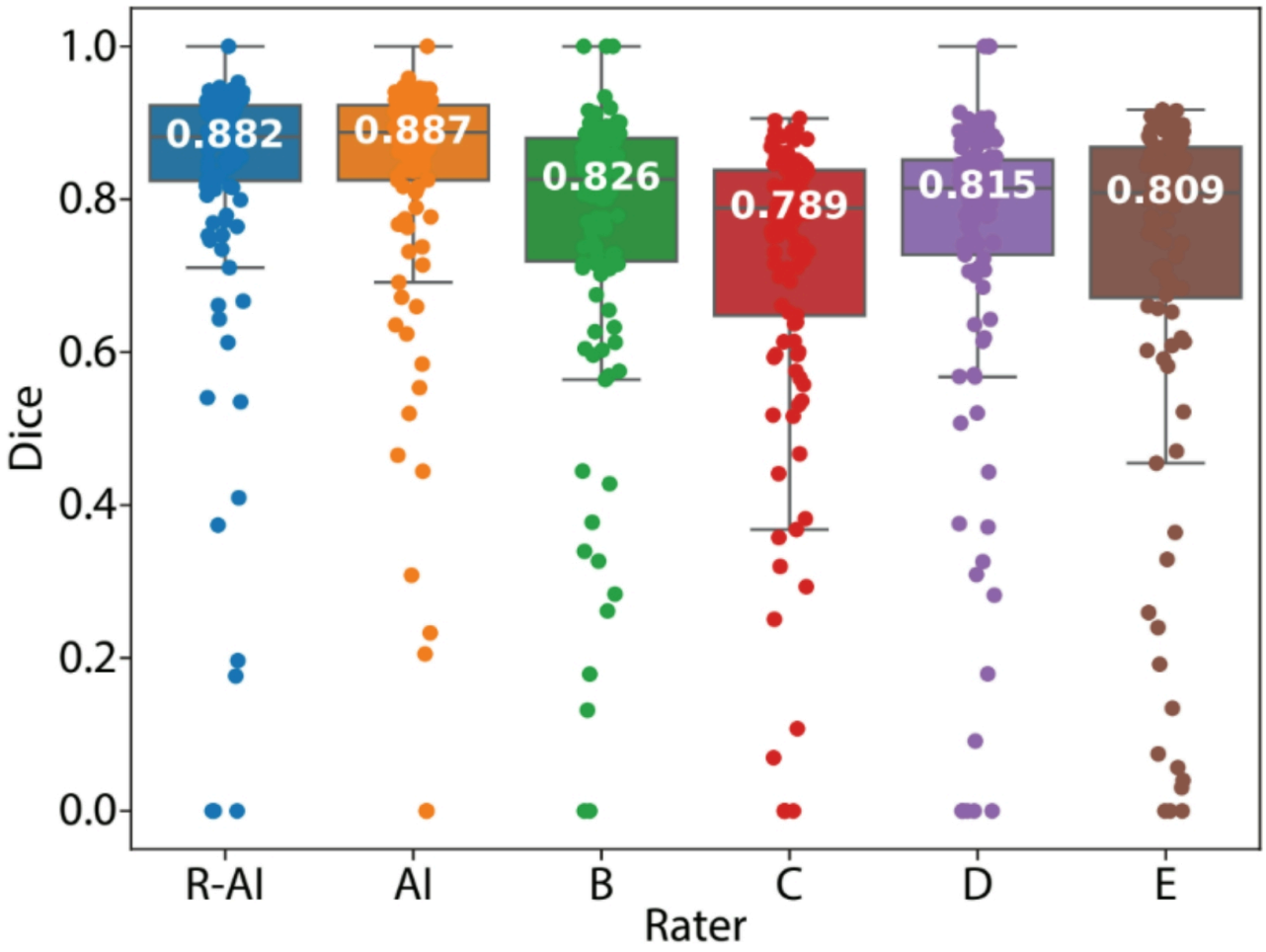

Figure 2

Dice similarity coefficients for AI and human raters with the enhancing tumor segmentations as ground truth. Boxplots show the distribution of Dice scores for the AI model, the human rater refining the AI output (rater+AI), and individual manual raters. Median values are indicated on each box.

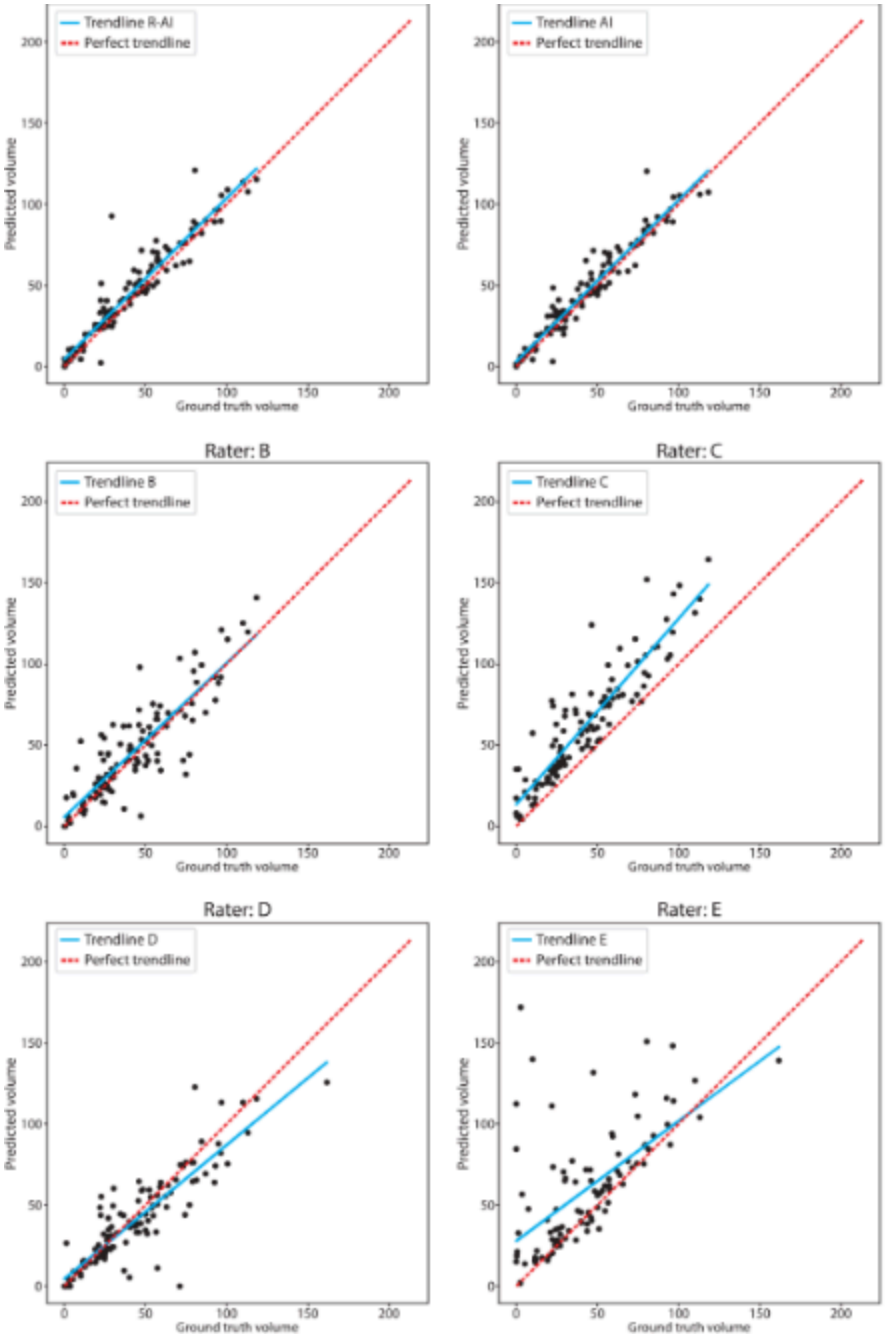

Figure 3

Predicted tumor core volumes versus ground truth for AI and human raters. Scatter plots show the relationship between predicted and ground truth volumes for each rater and the AI model. The yellow line indicates perfect agreement (identity line), while the colored lines represent regression fits. The predicted tumor core volumes generated by the AI and rater+AI correlated excellently with the ground truth. All human raters tended to oversegment compared with the ground truth.

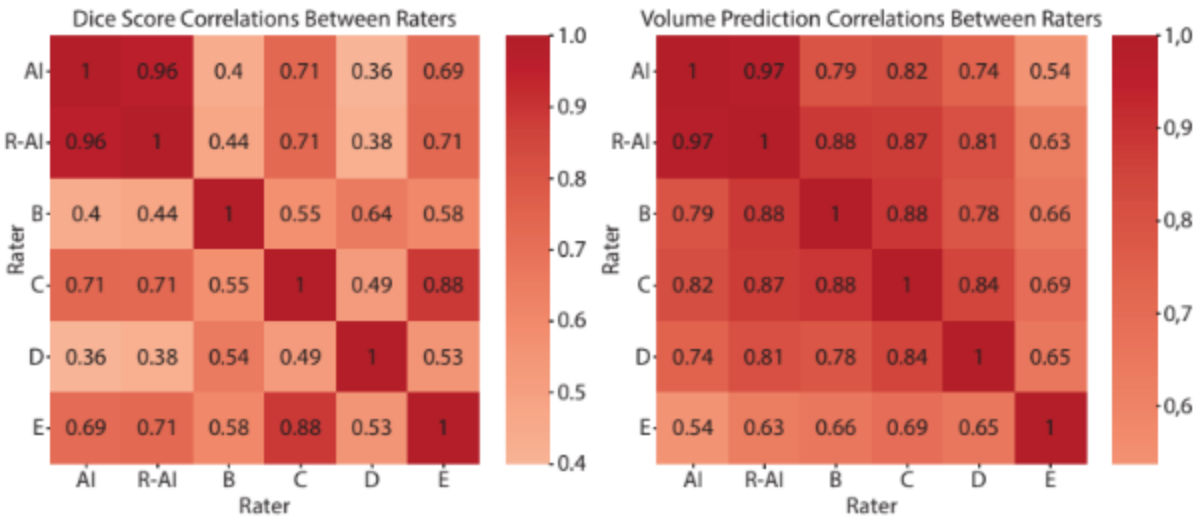

Figure 4

Pairwise correlations between human raters. Left: Heatmap of Dice score correlations between raters shows fair-to-substantial agreement. Right: Heatmap of predicted tumor core volume correlations demonstrates generally substantial agreement for volume compared with Dice.

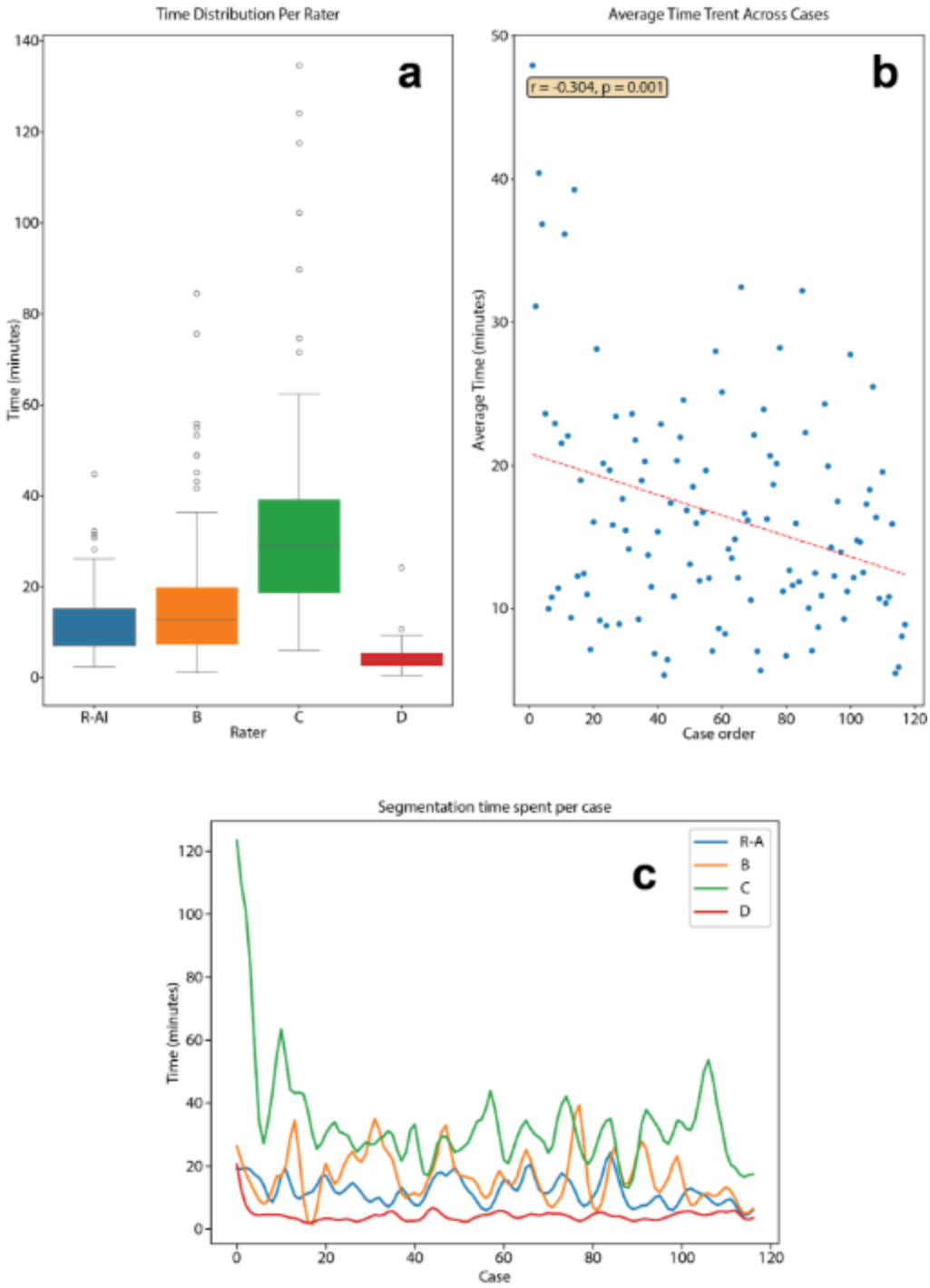

Figure 5

Segmentation time analysis across raters. (a) Distribution of segmentation times per case for individual raters, showing rater D as the fastest, followed by R-AI, B, and C, who was the slowest. (b) Average segmentation time per case across all raters shows a significant decreasing trend consistent with a learning effect (r=-0.304, P=0.001). (c) Segmentation times are shown for each rater across all cases. Rater E was consistently the fastest, followed by R-AI, then B, while C required substantially longer times. The differences were very large, with near-complete separation between raters D and C (rank-biserial ≈ +1) and a strong effect between raters R-AI and C (rank-biserial ≈ -0.92).

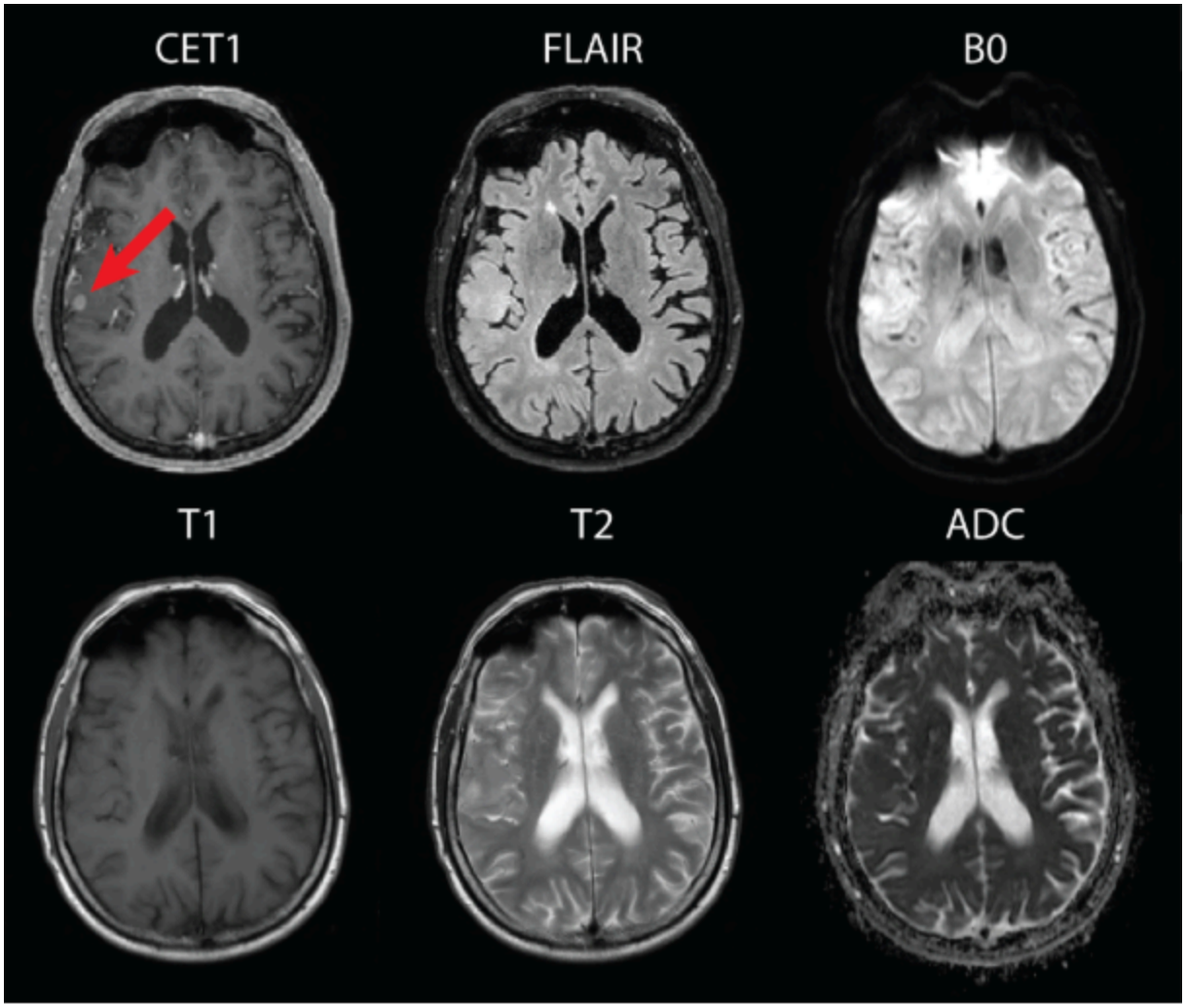

Figure 6

A small enhancing lesion in the right temporal lobe, misclassified by raters. The axial post-contrast T1-weighted image shows a 0.6 mL enhancing lesion (red arrow) in the right temporal lobe. The AI correctly classified this case as enhancing, but the human rater refining the AI output incorrectly reclassified it as non-enhancing. Two raters (B and D) also labeled this lesion as non-enhancing. In comparison, the other two raters segmented it with volumes of 30.7 mL and 10.9 mL, illustrating substantial variability in rater performance.

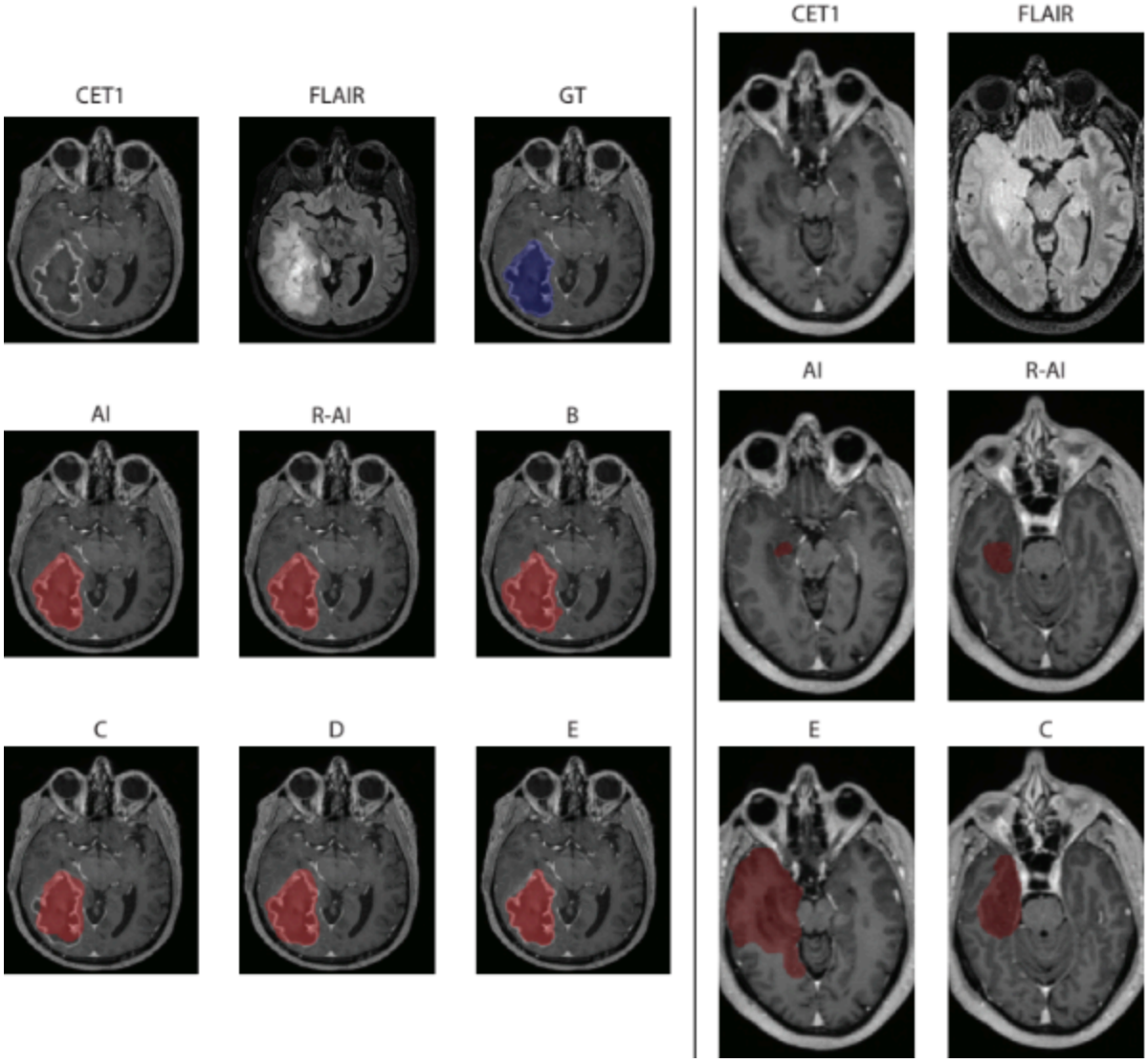

Figure 7

Left: Enhancing lesion consistently classified by all raters as enhancing. The axial post-contrast T1-weighted image shows an enhancing lesion (red arrow) that was correctly identified by all raters, with segmented volumes in close agreement and overlapping with the gold-standard segmentation. Right: Non-enhancing lesion misclassified by most raters as enhancing. The axial CET1 image shows a non-enhancing lesion (red arrow) that was incorrectly classified as enhancing by the majority of raters, with some raters (C and E) substantially overestimating the lesion volume. AI: AI-based segmentation, A-E: raters A-E, CET1: contrast-enhanced T1-weighted, FLAIR= fluid-attenuated inversion recovery, GT: ground truth, R-AI: AI-supported rater.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.docx