Neuro-Oncology Advances

8(1), vdag034, 2026 | https://doi.org/10.1093/noajnl/vdag034 | Advance Access publication February 12, 2026

GlioMODA: Robust glioma segmentation in clinical routine

Julian Canisius[†], **Josef Buchner**[†], **Marcel Rosier**, **Michael Griessmair**, **Jan C. Peeken**, **Jan S. Kirschke**, **Marie Piraud**, **Spyridon Bakas**, **Bjoern Menze**, **Benedikt Wiestler**[‡], **and Florian Kofler**[‡]

All author affiliations are listed at the end of the article
[†]These authors contributed equally as co-first authors.
[‡]These authors acontributed equally as co-senior authors.

Corresponding Author: Julian Canisius, MD, Institute of Neuroradiology, Klinikum rechts der Isar, Technical University of Munich (TUM), Ismaninger Str. 22, 81675 Munich, Germany (julian.canisius@tum.de).

Abstract

Background. Precise glioma segmentation in magnetic resonance imaging (MRI) is essential for accurate diagnosis, optimal treatment planning, and advancing clinical research. However, most deep learning approaches require complete, standardized MRI protocols that are frequently unavailable in routine clinical practice. This study presents and evaluates GlioMODA, a robust deep learning framework designed for automated glioma segmentation that delivers consistent high performance across varied and incomplete MRI protocols.

Methods. GlioMODA was trained and validated on the BraTS 2021 dataset (1251 training, 219 testing cases), systematically assessing performance across 11 clinically relevant MRI protocol combinations. Segmentation accuracy was evaluated using Dice similarity coefficients (DSC) and panoptic quality metrics. Volumetric accuracy was benchmarked against manual ground truth, and statistical significance was established via Wilcoxon signed-rank tests with Benjamini–Yekutieli correction.

Results. GlioMODA demonstrated state-of-the-art segmentation accuracy across tumor subregions, maintaining robust performance with incomplete or heterogeneous MRI protocols. Protocols including both T1-weighted contrast-enhanced and T2-FLAIR sequences yielded volumetric differences vs manual ground truth that were not statistically significant for enhancing tumor (median difference 55 mm^3, $P = .157$) and whole tumor (median difference -7 mm^3, $P = 1.0$), and exhibited median DSC differences close to zero relative to the 4-sequence reference protocol. Omitting either sequence led to substantial and significant volumetric errors.

Conclusions. GlioMODA facilitates reliable, automated glioma segmentation using a streamlined 2-sequence protocol (T1-contrast + T2-FLAIR), supporting clinical workflow optimization and broader implementation of quantitative volumetry compatible with RANO 2.0 criteria. GlioMODA is published as an open-source, easy-to-use Python package at https://github.com/BrainLesion/GlioMODA/.

Key Points

- T1-CE +T2-FLAIR achieves enhancing- and whole-tumor segmentation comparable to 4-sequence MRI.
- Consistent T1-CE +T2-FLAIR volumes enable reliable subsequent analysis.
- Open-source GlioMODA models and code support rapid integration.

Received September 12, 2025; accepted January 30, 2026.

Importance of the Study

Automated glioma segmentation is often limited by incomplete or heterogeneous magnetic resonance imaging (MRI) protocols. GlioMODA addresses this barrier by delivering consistent accuracy across 11 clinically relevant sequence combinations and by identifying a streamlined structural protocol (T1-contrast and T2-FLAIR) whose enhancing- and whole-tumor volumes are not statistically different from expert reference, even when the full 4-sequence protocol is unavailable. This defines a minimal structural sequence set that preserves accuracy and enables robust automated volumetry on preoperative glioma MRI, compatible with the imaging requirements of RANO 2.0. By releasing trained models and code as an open-source package, GlioMODA supports reproducible volumetry on heterogeneous real-world data and facilitates external validation and integration of quantitative volumetry into neuro-oncology workflows.

Gliomas are the most common primary brain tumors in adults, with glioblastoma multiforme (GBM) representing the most aggressive subtype, characterized by imaging heterogeneity, diffuse infiltration, rapid progression, and poor prognosis.[1,2] This study focuses on diffuse gliomas (predominantly high-grade gliomas/glioblastoma), where accurate estimation of enhancing and non-enhancing tumor (ET) burden is central, especially for radiotherapy planning, clinical trial eligibility, and standardized response assessment, even when gross progression is often apparent on visual inspection. Magnetic resonance imaging (MRI) is the diagnostic standard for GBM, enabling visualization of key tumor subregions essential for treatment planning, prognosis assessment, and monitoring therapeutic response.[3,4] These subregions include ET, tumor core (TC), and whole tumor (WT), with accurate segmentation central to clinical decision-making. The TC, including the ET, is essential for defining the gross tumor volume (GTV) in postoperative radiotherapy following the Stupp protocol, while edema (ED) segmentation is vital for assessing the clinical target volume (CTV), which encompasses infiltrative tumor areas, including non-enhancing, FLAIR-hyperintense regions.[5,6] Manual delineation of these subregions remains challenging due to the complex and heterogeneous structure of gliomas, making the process time-consuming, labor-intensive, and subject to substantial inter- and intra-rater variability in both clinical practice and research.[7,8]

In recent years, automated segmentation algorithms based on convolutional neural networks, particularly U-Net variants, have achieved expert-level accuracy and reduced clinician workload.[4] These algorithms substantially reduce the time required for manual segmentation while achieving accuracy comparable to expert reference segmentations, often within inter-rater variability. In addition, automated segmentation has demonstrated prognostic relevance, with volumetric measurements showing significant associations with survival outcomes in glioma patients.[9] Furthermore, automated approaches help standardize response assessment across institutions, reducing inter-observer variability that affects treatment evaluation.[10] However, translation to routine care requires robust performance under incomplete or heterogeneous MRI protocols, compatibility with contemporary response frameworks that accommodate quantitative volumetry (eg Response Assessment in Neuro-Oncology [RANO] 2.0) under standardized imaging conditions.[11]

Yet most state-of-the-art approaches, such as those evaluated in the Brain Tumor Segmentation (BraTS) challenge,[3,12,13] assume complete, standardized 4-sequence MRI—T1-weighted native (T1n), T1-weighted contrast-enhanced (T1c; often referred to as T1-CE), T2-weighted (T2w), and T2-FLAIR (T2f).[14] Hereafter, the abbreviations T1n, T1c, T2w, and T2f are used throughout. A standardized brain tumor imaging protocol (BTIP) for glioblastoma, including these 4 MRI sequences, averages a scan time of approximately 21 min; acquiring all 4 sequences increases scan time, cost, and patient burden; and protocol variability or incomplete data are common in routine care.[15] The prevalence of this issue can be significant; for instance, a multicenter glioma dataset reported missing sequences exceeding 20%.[16]

Recent advances in deep learning suggest that sequence requirements for high-quality glioma segmentation can be reduced without significantly compromising segmentation performance.[17] Concordantly, Buchner et al. previously identified core MRI sequences for reliable automatic brain metastasis segmentation. Here, T1c alone yielded strong lesion segmentation, with T2f necessary to capture surrounding FLAIR-hyperintense ED.[18] More broadly, recent studies indicate that certain MRI modalities may be redundant for segmentation tasks, and that reduced-input models—using only the most informative sequences—can achieve segmentation performance comparable to full-input models, offering practical advantages in terms of efficiency and applicability. Shorter protocols provide clinical advantages, including lower motion risk,[15] greater patient comfort, reduced costs, improved workflow efficiency,[4] and minimized risks associated with contrast agents.[19] They also lower the amount of data that needs to be processed, enabling faster preprocessing times and leaner models, which is advantageous for high-throughput and resource-limited environments.

Moreover, accurate delineation of these tumor subregions provides valuable information for downstream analysis with techniques such as radiomics or neural network-based feature extraction, supporting noninvasive characterization, molecular prediction, and outcome modeling in glioma and brain metastasis patients.[20,21]

Despite these advances, there remains limited determination of which MRI sequence combinations are essential to maintain robust glioma segmentation across real-world scenarios. Addressing this gap is crucial to enable automated analysis when protocols are incomplete and to optimize clinical workflows,[22] aligning with contemporary response frameworks that increasingly accommodate

quantitative volumetry (eg RANO 2.0).[11] Therefore, this study systematically evaluates GlioMODA—a flexible deep learning framework for automated glioma segmentation—across all clinically relevant combinations of standard MRI sequences. Our aim is to identify core MRI sequences that enable accurate, efficient, and broadly applicable glioma segmentation, thereby supporting protocol optimization, improving compatibility with heterogeneous clinical protocols, and expanding access to advanced image analysis with quantitative volumetry in routine practice. By delineating sequence combinations that preserve clinical-grade accuracy even when fewer sequences are available, this work aims to facilitate translation into routine neuro-oncology practice and contemporary quantitative response assessment, ultimately improving patient care.

Methods

Dataset

We used the preoperative BraTS 2021 glioma dataset,[23] comprising 1251 training and 219 test cases of adult patients with diffuse gliomas from multiple international institutions. All MRI examinations represent treatment-naïve scans acquired before any neurosurgical resection or oncological treatment. According to the original BraTS 2021 description, these cases were assembled from several publicly available and institutional diffuse-glioma cohorts (eg TCGA-GBM,[24] TCGA-LGG,[25] IvyGAP,[26] CPTAC-GBM,[27] ACRIN-FMISO-Brain[28]). The cohort predominantly consists of high-grade gliomas/glioblastoma; however, the released dataset does not provide complete molecular annotations (eg IDH status, MGMT, 1p/19q) or WHO-2021 tumor-type labels at the imaging level. Consequently, the cases include both glioblastoma and non-glioblastoma diffuse gliomas, and we did not stratify our analysis by molecular subtype or WHO classification.[2] Following standard practice in machine learning, all models were trained exclusively on the 1251-case training cohort without using a separate validation set, while the 219-case cohort (originally employed as the "validation set" in the BraTS 2021 segmentation challenge) served as our independent test set and remained unseen throughout the training process. All cases included complete preoperative multiparametric MRI with standardized 4-sequence protocols: T1n, T1c, T2w, and T2f images. Ground-truth segmentations were provided by the BraTS 2021 organizers and consist of expert manual delineations of central necrosis (non-enhancing core [NEC]), contrast-enhancing tumor (CET), and surrounding ED, created by board-certified neuroradiologists and other experienced neuro-oncology imaging experts under a standardized, centrally quality-controlled protocol as described in the original BraTS 2021 publication.[23] The multicenter nature of this dataset supports generalizability across diverse clinical environments and imaging protocols. MRI examinations were acquired on 1.5T and 3T scanners from multiple vendors. More detailed annotation protocols and data collection procedures are described in the original BraTS 2021 publication.[23]

MRI Sequence Combinations

To systematically evaluate the impact of protocol completeness on glioma segmentation, we analyzed 11 clinically relevant MRI sequence combinations (Table 1). These were selected based on the BTIP guidelines for glioblastoma[14] and prior evidence of sequence utility in tumor subregion delineation.[6,23]

Neural Network

We used the nnU-Net framework (version 2.5.2) to train and test our neural networks.[29] After analyzing the dataset fingerprint for each sequence combination, we trained a neural network according to the chosen experiment planner for a total of 1000 epochs.

Although Isensee et al.[30] recommend the residual-encoder U-Net presets as the new default, we found no performance differences compared to the previous default experiment planner, despite substantially increasing the training duration on our hardware. Therefore, we chose to use the default planner. We used the 3D-fullres configuration. Furthermore, cross-validation was not performed; instead, we trained on all available training data. Note that the 219 validation patients were never used during training but only for final model testing.

Following the BraTS challenge evaluation protocol, we did not evaluate the individual labels but rather combined them into clinically significant subregions: ET, TC, and WT. In the original BraTS segmentations, 3 labels are provided: necrotic/NEC, CET, and peritumoral ED/non-enhancing FLAIR-hyperintense tumor (ED). Enhancing tumor corresponds to CET only (excluding necrosis), TC comprises NEC + CET, and WT includes NEC + CET + ED, that is, the entire T2/FLAIR hyperintensity including ED. These definitions mirror

Table 1. MRI sequence combinations

Sequence group	Number of sequences (*n*)	MRI sequence combination
Baseline (complete protocol)	4	T1c + T1n + T2f + T2w
Contrast-enhanced (T1c present)	3	T1c + T1n + T2w
	2	T1c + T2f
	2	T1c + T1n
	2	T1c + T2w
	1	T1c
Noncontrast (T1c absent)	3	T2f + T1n + T2w
	2	T2f + T1n
	2	T2f + T2w
	1	T2f
	1	T1n
	1	T2w

Sequence groups, number of sequences, and protocol compositions evaluated in this study.
Abbreviations: T1c, T1-weighted contrast-enhanced; T1n, T1-weighted native; T2f, T2-FLAIR; T2w, T2-weighted.

common clinical usage in neuro-oncology, where ET approximates measurable enhancing disease, TC represents the solid TC, and WT reflects the combined extent of tumor and ED. To reflect this during training, we used the region-based training that is already integrated into the nnU-Net.

All training runs were conducted on Nvidia RTX 6000/8000 GPUs using CUDA version 12.2. Each training run took around 20 h.

Metrics

Segmentation performance was primarily evaluated using the Dice similarity coefficient (DSC),[31] quantifying spatial overlap between predicted and reference segmentations for the ET, TC, and WT subregions. For comprehensive assessment, we also report panoptic quality (PQ), combining segmentation quality (SQ) and recognition quality (RQ) to reflect both boundary accuracy and lesion detection (Table 2).[32] All metrics were calculated with the Python package panoptica.[33]

For Figures 1 and 2, segmentation performance for each sequence combination was compared to the 4-sequence baseline (T1c + T2f + T1n + T2w), with DSC differences (Figure 1) and corresponding Wilcoxon signed-rank tests (Figure 2)[34] to assess statistical significance. The resulting *P*-values were adjusted for multiple comparisons using the Benjamini–Yekutieli procedure[35] and visualized as heatmaps (Figure 2).

For Figures 3 and 4, all protocol combinations—including the baseline—were compared directly to manual ground truth by assessing absolute volumetric errors (Figure 3). Statistical significance of these volumetric differences was evaluated using the same Wilcoxon signed-rank tests with Benjamini–Yekutieli correction, and results were visualized as heatmaps in parallel style and colormap as used for DSC analyses (Figure 4), allowing direct visual and statistical comparison.

All results are reported as median (interquartile range) across the test cohort. Statistical significance was defined as an adjusted *P*-value < .05.

Difference plots, box plots, and statistical significance heatmaps were generated using Matplotlib[36] and seaborn,[37] with protocol-specific color schemes and consistent colormap use for both DSC and volumetric analyses to facilitate clear interpretation and comparability across all figures.

Table 2. Panoptic assessment across MRI sequence combinations

Sequence	RQ (mean ± SD)	SQ (mean ± SD)	PQ (mean ± SD)
Enhancing tumor (ET)			
T1c + T2f + T1n + T2w (baseline)	**0.64 ± 0.33**	**0.81 ± 0.12**	**0.53 ± 0.30**
T1c + T1n + T2w	0.62 ± 0.33	0.81 ± 0.11	0.52 ± 0.30
T2f + T1n + T2w	0.66 ± 0.32	0.65 ± 0.08	0.43 ± 0.22
T1c + T2f	0.62 ± 0.33	0.81 ± 0.11	0.52 ± 0.30
T1c + T1n	0.63 ± 0.33	0.81 ± 0.10	0.52 ± 0.30
T1c + T2w	0.61 ± 0.33	0.81 ± 0.12	0.51 ± 0.29
T2f + T1n	0.66 ± 0.31	0.64 ± 0.08	0.43 ± 0.22
T2f + T2w	0.67 ± 0.32	0.65 ± 0.09	0.43 ± 0.22
T1c	0.61 ± 0.32	0.80 ± 0.11	0.50 ± 0.29
T1n	0.67 ± 0.32	0.61 ± 0.07	0.42 ± 0.21
T2w	0.66 ± 0.32	0.63 ± 0.07	0.42 ± 0.21
T2f	0.66 ± 0.32	0.62 ± 0.08	0.42 ± 0.22
Tumor core (TC)			
T1c + T2f + T1n + T2w (baseline)	**0.71 ± 0.32**	**0.86 ± 0.12**	**0.62 ± 0.30**
T1c + T1n + T2w	0.71 ± 0.31	0.86 ± 0.11	0.63 ± 0.30
T2f + T1n + T2w	0.71 ± 0.32	0.78 ± 0.10	0.56 ± 0.27
T1c + T2f	0.71 ± 0.31	0.86 ± 0.11	0.62 ± 0.30
T1c + T1n	0.71 ± 0.31	0.86 ± 0.11	0.62 ± 0.30
T1c + T2w	0.70 ± 0.32	0.86 ± 0.11	0.61 ± 0.30
T2f + T1n	0.71 ± 0.32	0.77 ± 0.10	0.55 ± 0.27
T2f + T2w	0.71 ± 0.32	0.78 ± 0.10	0.56 ± 0.27
T1c	0.70 ± 0.31	0.86 ± 0.10	0.61 ± 0.30
T1n	0.71 ± 0.32	0.74 ± 0.10	0.53 ± 0.26
T2w	0.71 ± 0.32	0.75 ± 0.11	0.54 ± 0.26
T2f	0.70 ± 0.32	0.76 ± 0.10	0.53 ± 0.26
Whole tumor (WT)			
T1c + T2f + T1n + T2w (baseline)	**0.59 ± 0.31**	**0.87 ± 0.09**	**0.52 ± 0.28**
T1c + T1n + T2w	0.62 ± 0.32	0.84 ± 0.08	0.53 ± 0.28
T2f + T1n + T2w	0.60 ± 0.32	0.86 ± 0.09	0.52 ± 0.29
T1c + T2f	0.59 ± 0.31	0.87 ± 0.09	0.52 ± 0.29
T1c + T1n	0.62 ± 0.31	0.80 ± 0.09	0.51 ± 0.27
T1c + T2w	0.62 ± 0.31	0.84 ± 0.09	0.53 ± 0.28
T2f + T1n	0.59 ± 0.31	0.86 ± 0.09	0.51 ± 0.28
T2f + T2w	0.59 ± 0.31	0.86 ± 0.10	0.52 ± 0.29
T1c	0.63 ± 0.31	0.79 ± 0.10	0.50 ± 0.27
T1n	0.63 ± 0.32	0.78 ± 0.09	0.49 ± 0.26
T2w	0.62 ± 0.32	0.83 ± 0.09	0.52 ± 0.28
T2f	0.59 ± 0.31	0.86 ± 0.10	0.51 ± 0.29

Recognition quality (RQ), segmentation quality (SQ), and panoptic quality (PQ) are reported for ET, tumor core (TC), and WT across all protocol combinations; values are presented as mean ± SD. RQ reflects instance detection (F1-equivalent with an overlap threshold), SQ reflects boundary accuracy for true positives, and PQ combines both measures.
Abbreviations: T1c, T1-weighted contrast-enhanced; T1n, T1-weighted native; T2f, T2-FLAIR; T2w, T2-weighted.

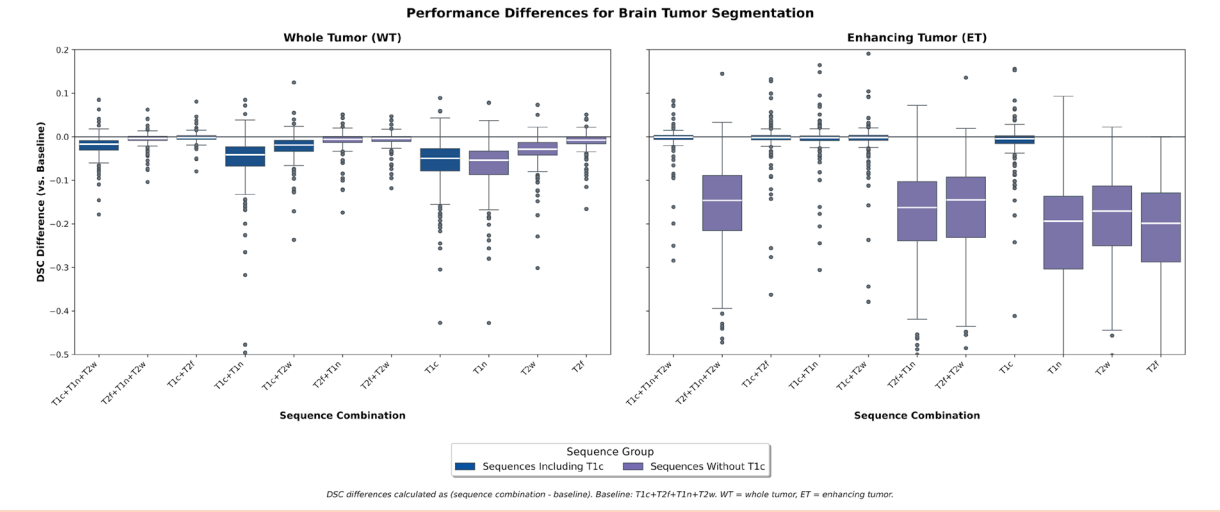

Figure 1. Dice similarity coefficient differences relative to the 4-sequence baseline. Box plots illustrate Dice similarity coefficient (DSC) differences from the 4-sequence baseline protocol (T1c + T2f + T1n + T2w) for whole tumor (WT) and enhancing tumor (ET) segmentation. Boxes show the median (white line) and interquartile range (IQR); whiskers extend to 1.5× IQR, and points indicate outliers. Abbreviations: T1c, T1-weighted contrast-enhanced; T1n, T1-weighted native; T2f, T2-FLAIR; T2w, T2-weighted.

Reporting follows the CLEAR checklist for radiomics and AI imaging studies[38]; the completed checklist is provided as supplementary material.

Results

Sequence Performance and Quality Assessment

To systematically evaluate MRI sequence combinations for glioma segmentation, we used 2 complementary reference levels. First, panoptic SQ (RQ, SQ, and PQ) for all sequence combinations was computed directly against the manual ground-truth annotations (Table 2). Second, for Dice-based SQ (Figure 1), we expressed the performance of each reduced sequence combination relative to the 4-sequence baseline protocol (T1c + T2f + T1n + T2w), which itself is evaluated against the same ground truth. Additionally, absolute DSC values vs the expert manual ground truth for all sequence combinations are provided in Figure S2.

For ET segmentation, T1c-containing combinations demonstrated superior performance with DSC differences clustered tightly around zero. T1c alone achieved near-baseline performance with minimal deviation (median: −0.004, IQR: −0.015 to +0.002), while T1c combinations showed even tighter distributions: T1c + T2f (median: −0.002, IQR: −0.007 to +0.003), T1c + T1n (median: −0.002, IQR: −0.009 to +0.002), and T1c + T2w (median: −0.002, IQR: −0.008 to +0.003). Panoptic segmentation analysis (Table 2) confirmed these findings, with T1c alone achieving RQ, SQ, and PQ scores (0.61 ± 0.32, 0.80 ± 0.11, 0.50 ± 0.29) comparable to baseline (0.64 ± 0.33, 0.81 ± 0.12, 0.53 ± 0.30). All T1c combinations maintained consistently high SQ scores of 0.80-0.81, matching or exceeding baseline performance. In contrast, combinations lacking T1c showed substantial performance degradation with DSC differences of −0.171 to −0.199 and markedly wider variability ranges. Segmentation quality scores dropped to 0.61-0.65 for non-T1c combinations, representing a clinically meaningful 0.16-0.20 decrease compared to baseline and confirming T1c's fundamental importance for accurate ET boundary delineation.

For TC, panoptic segmentation performance showed a similar overall pattern to ET. Recognition quality was largely stable across all sequence combinations, whereas T1c-containing protocols achieved clearly higher SQ and PQ than combinations without T1c. Across all combinations, absolute SQ and PQ values for TC were higher than for ET, consistent with the more compact, solid morphology of the TC compared with the irregular enhancing margins. Complete TC panoptic metrics for all sequence combinations are reported together with ET and WT in Table 2.

For WT segmentation, T2f-containing combinations provided optimal performance stability across all evaluated metrics. The T1c + T2f combination achieved near-baseline performance (median DSC difference: −0.002, IQR: −0.006 to +0.002; Figure 1) with SQ scores identical to baseline (SQ: 0.87 ± 0.09; Table 2). T2f alone demonstrated robust performance (median DSC difference: −0.008, IQR: −0.016 to −0.001; SQ: 0.86 ± 0.10), while other T2f combinations maintained excellent stability: T2f + T2w (median: −0.003, IQR: −0.011 to +0.001), T2f + T1n + T2w (median: −0.004, IQR: −0.009 to +0.001), and T2f + T1n (median: −0.006, IQR: −0.013 to +0.001). These combinations achieved SQ scores close to baseline

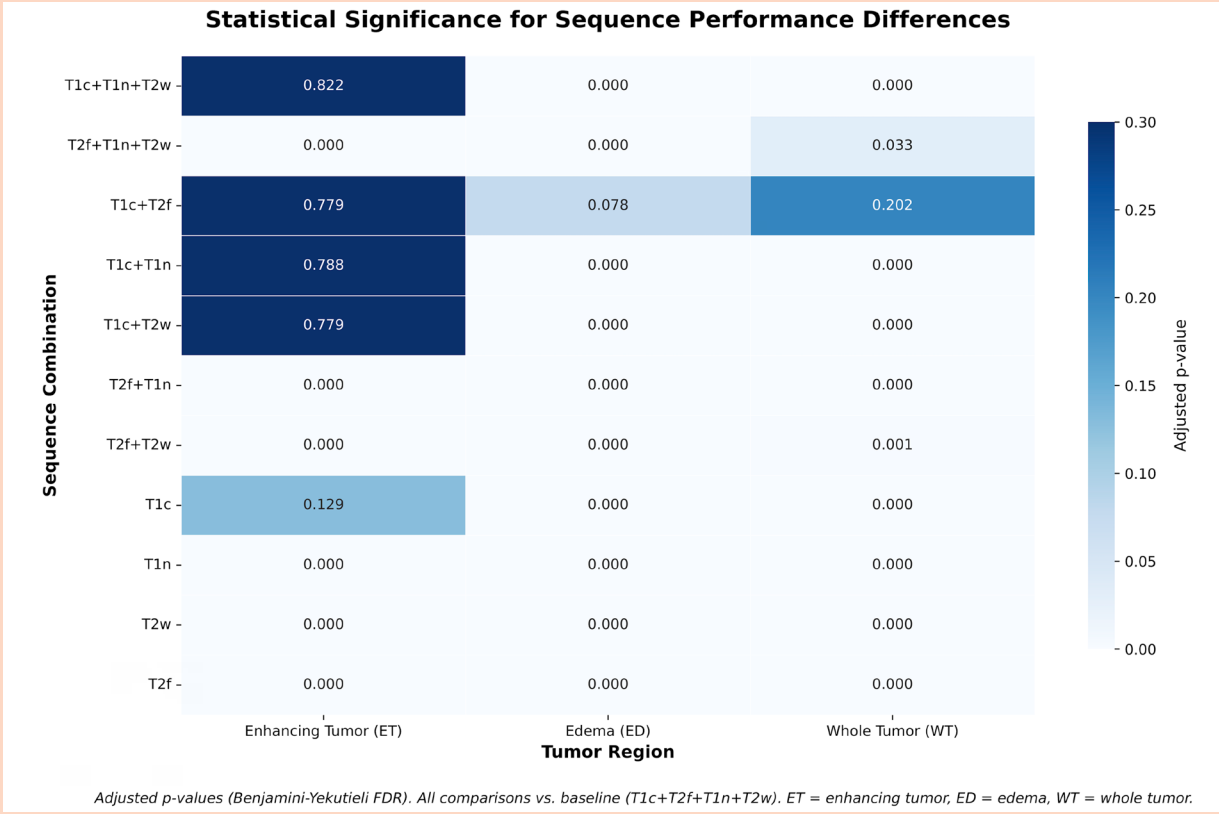

Figure 2. Statistical significance for sequence performance differences. The heatmap displays adjusted *P*-values from Wilcoxon signed-rank tests comparing each sequence combination with the 4-sequence baseline across enhancing tumor (ET), edema (ED), and whole tumor (WT). *P*-values were adjusted using the Benjamini–Yekutieli procedure for multiple comparisons correction. Color intensity represents statistical significance levels, with darker shades indicating higher *P*-values (non-significant differences) and lighter shades showing lower *P*-values (significant differences). Abbreviations: T1c, T1-weighted contrast-enhanced; T1n, T1-weighted native; T2f, T2-FLAIR; T2w, T2-weighted.

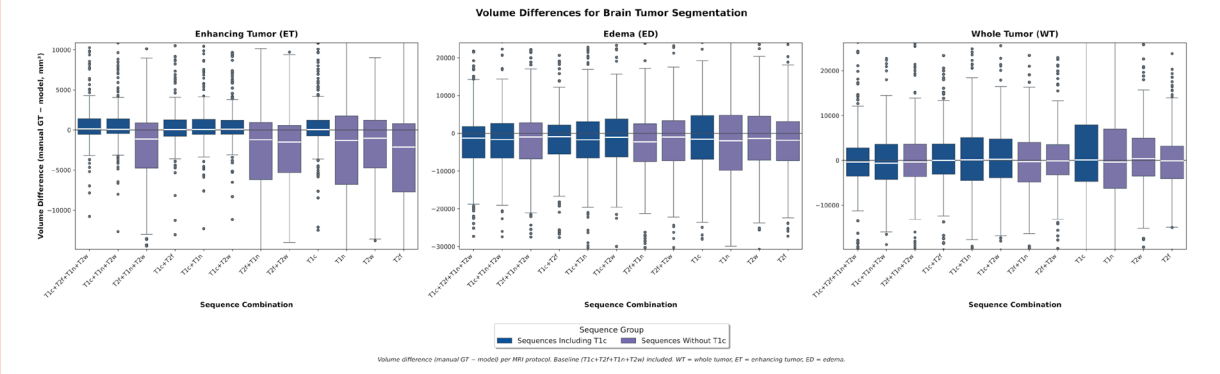

Figure 3. Volumetric accuracy across MRI sequence combinations. Box plots show signed volume differences between model segmentation and manual ground truth (manual GT—model, mm³) for ET, ED, and WT across MRI sequence combinations. Negative values indicate over-segmentation and positive values indicate under-segmentation relative to ground truth. Boxes show the median (white line) and IQR; whiskers extend to 1.5× IQR. The baseline protocol is T1c + T2f + T1n + T2w. Abbreviations: T1c, T1-weighted contrast-enhanced; T1n, T1-weighted native; T2f, T2-FLAIR; T2w, T2-weighted.

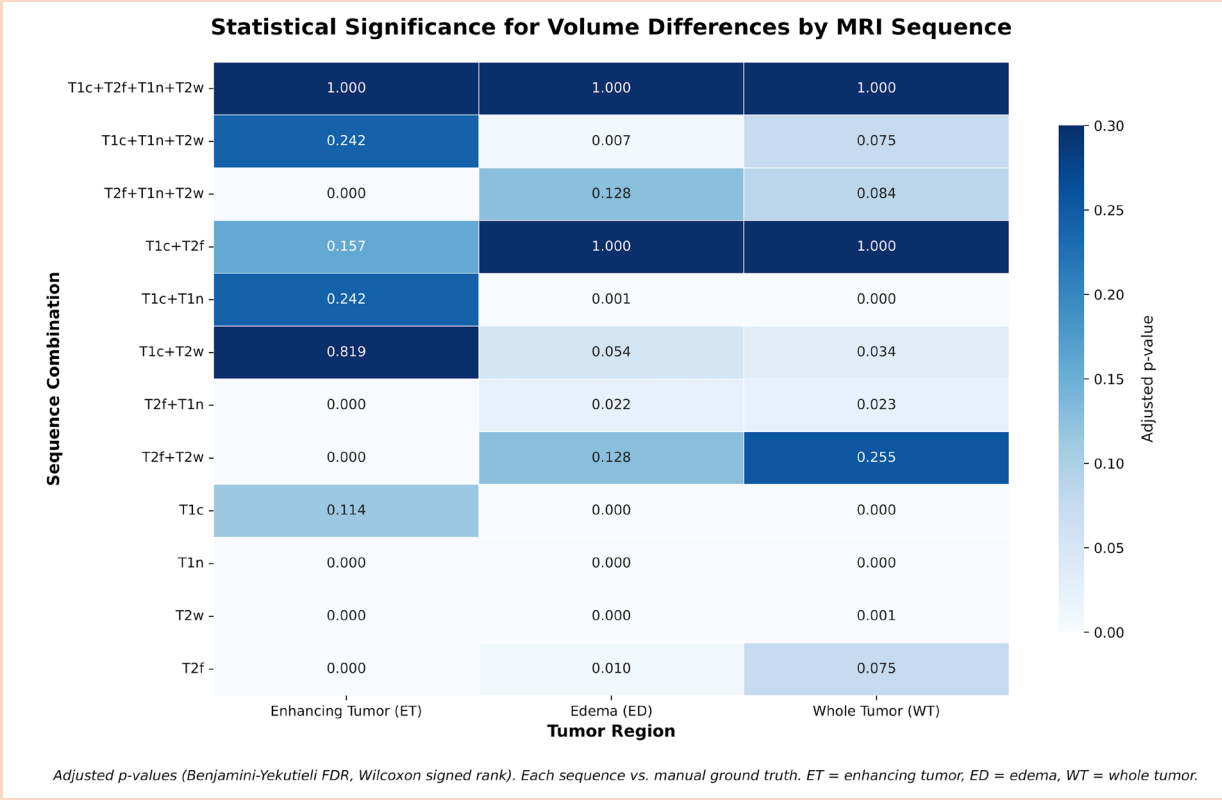

Figure 4. Statistical significance of volumetric differences across MRI sequence combinations. The heatmap shows Benjamini–Yekutieli-adjusted *P*-values (Wilcoxon signed-rank) for volumetric differences between model and manual ground truth across sequence combinations and tumor regions (ET, ED, WT). Color intensity corresponds to adjusted *P*-value, with lighter shades reflecting stronger statistical significance ($P < 0.05$). Abbreviations: T1c, T1-weighted contrast-enhanced; T1n, T1-weighted native; T2f, T2-FLAIR; T2w, T2-weighted.

(0.86 ± 0.09-0.10), corresponding to only a ~0.01 absolute decrease in SQ compared to the baseline protocol. Single-sequence models without T2f showed greater performance variability and larger systematic deviations: T1c alone (median: −0.049, IQR: −0.079 to −0.027), T1n alone (median: −0.054, IQR: −0.087 to −0.033), and T2w alone (median: −0.029, IQR: −0.042 to −0.013). This pattern emphasizes T2f's critical role in capturing complete tumor extent, including edematous components.

Statistical Validation

Pairwise Wilcoxon signed-rank tests with Benjamini–Yekutieli false discovery rate correction revealed distinct statistical patterns that reinforce the performance findings across tumor regions (Figure 2).

For ET segmentation, all T1c-containing combinations showed non-significant differences from baseline, with T1c alone achieving $P = .129$ and T1c combinations demonstrating even higher *P*-values: T1c + T1n + T2w ($P = .822$), T1c + T2f ($P = .779$), T1c + T1n ($P = .788$), and T1c + T2w ($P = .779$). These statistical results align directly with the minimal median DSC differences observed for these combinations (all ≤.004). Conversely, all non-T1c combinations demonstrated highly significant performance degradation ($P < .001$), corresponding to their substantial median differences of −0.171 to −0.199.

For ED segmentation, the statistical analysis revealed that only T1c + T2f showed non-significant differences from baseline ($P = .078$), reflecting its superior median difference of −0.002 and highlighting the synergistic importance of combining both sequences. All other combinations were significantly inferior to the baseline ($P < .001$), demonstrating that neither sequence alone is sufficient for optimal ED delineation.

For WT segmentation, T1c + T2f achieved non-significant differences from baseline ($P = .202$), corresponding to its minimal median difference of −0.002, with T2f + T1n + T2w also approaching non-significance ($P = .033$). All other sequence combinations, including single-sequence models, were significantly inferior to the baseline (adjusted $P \leq .001$).

Volumetric Accuracy Assessment

Volumetric evaluation showed distinct patterns across MRI sequence protocols, with important implications for reliable tumor measurement in the ET, ED (corresponding to the peritumoral FLAIR-hyperintense label, that is WT minus TC), and WT compartments (Figures 3 and 4).

Protocols containing both T1c and T2f produced the most accurate segmentations for ET and WT. The T1c + T2f

combination achieved nearly perfect volumetric agreement with ground truth, with median differences close to zero in both ET (55 mm³, IQR: −788 to 1279 mm³) and WT (−7 mm³, IQR: −3069 to 3664 mm³), and no significant difference from manual reference (adjusted $P = .157$ and $P = 1.0$). All T1c-containing protocols had nonsignificant ET errors (all adjusted $P \geq .114$). In contrast, protocols lacking T1c—such as T2f, T1n, and T2w—showed pronounced over-segmentation in ET (eg T2f: −2139 mm³, IQR: −7723 to 790 mm³) and were all highly significant vs ground truth (all adjusted $P < .001$).

For WT, the highest accuracy was achieved by the baseline and T1c + T2f (adjusted $P = 1.0$ for both), as well as T2f + T2w ($P = .255$), T2f + T1n + T2w ($P = .084$), T1c + T1n + T2w ($P = .075$), and T2f ($P = .075$), which all yielded nonsignificant differences from ground truth (adjusted $P \geq .05$). All other combinations returned statistically significant volumetric errors (adjusted $P < .05$).

Edema segmentation showed mixed volumetric accuracy across protocols. The baseline and T1c + T2f protocols yielded small median volume differences and non-significant deviations from the manual ground truth (adjusted $P = 1.0$ for both), and T2f-based combinations such as T2f + T2w and T2f + T1n + T2w (adjusted $P = .128$ for both) also did not reach statistical significance, whereas many other protocols, except T1c + T2w (adjusted $P = .054$), exhibited significant volumetric errors (adjusted $P < .05$).

In summary, T1c + T2f is the only simplified protocol consistently providing nonsignificant volumetric differences compared to ground truth for both ET and WT, providing robust segmentation with reduced input requirements when full multisequence MRI is unavailable, while not replacing clinically required diagnostic imaging.

Discussion

This study demonstrates that automated glioma segmentation can be delivered with clinically applicable accuracy using a streamlined 2-sequence protocol (T1-contrast and T2f), achieving nonsignificant volumetric differences relative to expert reference for enhancing and whole-tumor compartments and maintaining overlap-based segmentation performance comparable to a 4-sequence reference. Critically, this identifies a minimal protocol that preserves volumetric and segmentation accuracy for targets most relevant to clinical decision-making, supporting implications for response assessment and clinical implementation under contemporary practice standards, including RANO 2.0.[11]

However, the BraTS 2021 dataset comprises only treatment-naïve preoperative examinations and does not include longitudinal posttreatment timepoints; therefore, our findings should be interpreted as demonstrating technical feasibility and workflow compatibility with volumetric RANO 2.0 criteria rather than as a validation of treatment response assessment in recurrent disease.

Gliomas, particularly GBM, remain among the most difficult challenges in neuro-oncology due to their aggressive growth, infiltrative behavior, and marked heterogeneity on imaging, which complicate consistent measurement and longitudinal assessment in practice. Although the BraTS 2021 dataset includes both glioblastoma and non-glioblastoma diffuse gliomas, our primary clinical focus is on high-grade gliomas/glioblastoma, which constitute the majority of cases. Comprehensive overviews, including the MICCAI 2022 BraTS volume,[39] underscore the central role of automated segmentation in both research and clinical neuro-oncology and the need for models that are robust, generalizable, and clinically interpretable. Multicenter evaluations have demonstrated clinically acceptable glioblastoma segmentation across diverse datasets, and reviews emphasize that rigorous external validation and explicit management of heterogeneous clinical data—including missing sequences—are prerequisites for translation and reproducible response assessment.[40,41] Nevertheless, many models still require complete, standardized MRI protocols, whereas real-world examinations frequently lack one or more sequences due to protocol variability,[42] scan time constraints, motion, or technical limitations, limiting dependable deployment for quantitative response workflows.[16]

This gap motivates the development of tools like GlioMODA, explicitly designed to maintain performance under missing or heterogeneous input data. Such segmentation tools are key enablers for reliable quantitative response assessment in longitudinal cohorts under RANO 2.0.[11] To this end, this study systematically evaluates segmentation and volumetric performance across 11 clinically relevant sequence combinations to define where accuracy is preserved and where it degrades when key inputs are omitted—information directly actionable for protocol design and site-level implementation. Addressing missing modalities during training, prior work has proposed sequence-dropout strategies to improve robustness when specific inputs are unavailable, without compromising accuracy when all modalities are present[43]; these observations align with the current findings and support flexible frameworks for real-world adoption.

Precisely, our results demonstrate that GlioMODA achieves state-of-the-art segmentation performance across all tumor subregions, even when using fewer input sequences, and provide new insights into the sequence dependencies of automated glioma segmentation. T1c was indispensable for accurate enhancing-tumor delineation, and models using T1c alone or in combination with other sequences performed comparably to the full-sequence reference, whereas omission of T1c produced systematic degradation. Conversely, T2f was critical for capturing non-ET extent, essential for accurate radiotherapy planning and CTV delineation,[6] reinforcing its central role in whole-tumor assessment and longitudinal comparability. Consistently, Wilcoxon signed-rank test analysis revealed that T1c-containing combinations were not statistically inferior to the full protocol for ET segmentation, while T2f-containing combinations also maintained acceptable ED and WT performance. Reduced accuracy was most frequently observed for very small or faint enhancing foci, ambiguous FLAIR margins, and cases with motion artifacts, which can blur compartment boundaries and contribute to under- or over-segmentation.

By evaluating PQ, RQ, and SQ, this analysis provides a comprehensive performance profile that couples lesion detectability with boundary precision. High RQ scores (>0.6) across all sequence combinations confirm robust tumor detection, while SQ more sensitively captures the effect of

sequence completeness on boundary accuracy. This pairing—reliable recognition with context-dependent precision—supports longitudinal clinical workflows, maintaining dependable detection when boundary detail varies.

The relevance of this result is underscored by the updated RANO 2.0 criteria, which permit volumetric response assessment alongside 2-dimensional measures and emphasize standardized baseline and follow-up imaging.[11] While RANO 2.0 retains 2-dimensional assessments as the primary standard, it provides explicit volumetric thresholds—approximately 65% reduction for partial response and 40% increase for progression—that closely mirror traditional area-based cutoffs, thereby operationalizing quantitative categorization in clinical trials and care. In this context, our results suggest that using T1c + T2f as the segmentation input can preserve preoperative volumetric accuracy and support workflow compatibility for volumetric measurements in RANO 2.0-oriented settings. By retaining the 2 most informative sequences, this strategy reduces avoidable measurement bias introduced by missing key modalities—an important prerequisite for subsequent longitudinal applications of volumetry. Conversely, omission of either T1c or T2f was associated with systematic volume errors that could bias volumetric assessments and risk misclassification if applied in longitudinal response settings, including those based on RANO 2.0, potentially compromising trial endpoints or treatment decisions. Importantly, the ability to maintain high volumetric accuracy with a simplified protocol addresses a central challenge highlighted in the RANO 2.0 update: ensuring standardized, longitudinally consistent imaging across diverse clinical settings with practical acquisition constraints. Overall, these findings support T1c + T2f as a pragmatic, segmentation-focused minimum input for accurate preoperative volumetry and for RANO 2.0-compatible baseline measurements in multicenter settings. For clinical MRI interpretation, however, precontrast T1-weighted imaging remains important (eg to identify hemorrhage or fat), and thus this proposed minimum refers to segmentation input requirements rather than a complete diagnostic protocol. Longitudinal posttreatment validation remains required before applying this approach to response assessment.

While progression of glioblastoma is frequently obvious on visual inspection of MRI, automated volumetric assessment can still add value in several clinically relevant settings. First, enhancing- and whole-tumor volumes contribute to the definition and adaptation of radiotherapy target volumes (eg GTV and CTV) and to stratification and response assessment in clinical trials, where reproducible, observer-independent measurements are essential. Second, quantitative volumetry enables more objective evaluation in borderline or equivocal situations, such as slow volumetric growth or discordant clinical and imaging findings, and can support contemporary frameworks like RANO 2.0, which explicitly accommodate volumetric thresholds for response categorization. Finally, robust automated measurements facilitate standardized reporting and multicenter comparisons, which are increasingly important in neuro-oncology research and trial design. The isolated volume of central necrosis itself usually has a limited direct impact on day-to-day treatment decisions. In GlioMODA, necrosis is segmented primarily to define the TC consistently and to separate enhancing from nonenhancing tissue, which is important for radiotherapy planning, longitudinal assessment, and radiomics or prognostic modeling. Accordingly, the main clinically relevant outputs of our framework are enhancing-tumor and whole-tumor volumes, while necrosis volumetry is included mainly for structural completeness and future research applications.

A recent study by De Sutter et al.[17] systematically investigated the influence of different MRI modality combinations on automated glioblastoma segmentation, primarily using nnU-Net and SwinUNETR architectures. Consistent with our findings, they reported that T1c suffices for accurate enhancing-tumor and tumor-core delineation, and that T1c + T2f achieves whole-tumor performance comparable to a 4-sequence input; adding further modalities did not statistically improve accuracy and could introduce redundancy, although it reduced epistemic uncertainty. However, our study extends these findings in several important ways. First, we conducted a more comprehensive systematic evaluation by testing all clinically relevant sequence combinations, reflecting the spectrum of real-world protocol variability rather than a subset, thereby directly informing protocol design under routine constraints. Second, we complemented overlap measures with RQ, SQ, and PQ to separate lesion detectability from boundary accuracy, providing a more nuanced operational profile for clinical deployment. Third, we used Wilcoxon signed-rank tests and heatmap visualizations to rigorously determine which modality combinations are not significantly different from the baseline, offering clearer guidance for protocol optimization. Finally, we release trained models and code as open source, supporting reproducibility, site-level validation, and PACS/DICOM integration—key prerequisites for clinical and multicenter trial implementation. Together, these advances emphasize technical rigor and practical implementability and deliver decision-ready guidance for implementation.

GlioMODA has been validated on the publicly available BraTS 2021 preoperative glioma dataset (1251 training; 219 validation/test cases with expert segmentations), demonstrating robust performance across diverse imaging protocols. Recognizing that clinical integration of automated segmentation tools faces practical barriers, including workflow adaptation, regulatory validation, and interoperability challenges, GlioMODA is provided as an open-source Python package available at https://github.com/BrainLesion/GlioMODA/ within the BrainLesion suite.[44] This open release enables transparent, auditable use and local verification, and its modular design supports PACS/DICOM integration—providing a practical, validation-ready route toward clinical deployment and multicenter trials in settings that adopt RANO 2.0—aligned volumetric workflows.[11] The software is intended for local, on-premise use; imaging data remain under the control of the deploying institution, and all data anonymization, data-protection compliance (eg GDPR or corresponding local regulations), and medical-device regulatory responsibilities rest with the local site.

Future research directions include extending GlioMODA to pediatric gliomas[45] and rare tumor subtypes, where distinct imaging characteristics and age-specific biology may require adapted training strategies and evaluation endpoints, building on previous work investigating multiple pediatric brain tumor types.[46] Prospective, multisite

validation in these populations will be essential to ensure safe clinical use and generalizability. Integration of advanced imaging modalities such as FET-PET represents a promising approach for enhanced molecular characterization. However, recent work indicates both potential gains in volumetric accuracy and practical challenges, including volume underestimation and occasional segmentation errors,[47] underscoring the need for careful calibration and cross-modality harmonization. The development of multitask models that simultaneously perform segmentation, tumor grading, and molecular prediction represents another promising direction, particularly where shared representations improve data efficiency and support end-to-end clinical workflows, as demonstrated by recent deep ensemble learning frameworks.[48] Additionally, integration of GlioMODA into radiomics pipelines could enable personalized treatment strategies by linking imaging-derived features with molecular and clinical outcomes, while providing standardized, reproducible volumetry for response assessment in line with contemporary neuro-oncology criteria.

Several limitations must be acknowledged in our study. First, our analysis is based on the preoperative gliomas from the BraTS 2021 dataset, which predominantly contains high-grade gliomas/glioblastoma but does not provide complete molecular or WHO-2021 subtype information at the imaging level. As a result, we could not distinguish isocitrate dehydrogenase (IDH)-wildtype glioblastoma from IDH-mutant astrocytoma grade 4, and our findings may not fully capture potential differences in the biological and imaging significance of FLAIR hyperintensity between these entities. Moreover, because BraTS 2021 includes only treatment-naïve preoperative examinations, our results are restricted to untreated preoperative tumors and do not cover postoperative imaging or treated/recurrent disease. Dedicated longitudinal studies in treated and recurrent disease, as well as in these additional patient groups, will be required to validate GlioMODA for RANO 2.0 response assessment. Second, the exclusive focus on this preoperative adult glioma cohort limits generalizability to other glioma subtypes and pediatric populations, where different imaging characteristics and segmentation challenges may apply. Third, while our systematic evaluation demonstrates robust performance across diverse sequence combinations, prospective validation in routine clinical workflows is required to confirm real-world applicability. Fourth, our current framework is optimized for structural MRI sequences and does not incorporate advanced imaging modalities such as perfusion MRI, diffusion tensor imaging, or molecular imaging techniques like FET-PET, which may provide additional diagnostic value in specific clinical scenarios. Fifth, the retrospective nature of our study using curated research datasets may underrepresent scanner/vendor variability, artifacts, and protocol deviations encountered in routine practice. Finally, while our volumetric accuracy analysis supports baseline RANO 2.0 implementation, longitudinal, multi-timepoint validation across different treatment response scenarios is needed to establish clinical utility for serial tumor monitoring and trial endpoints.

In conclusion, GlioMODA represents a practical advancement for automated glioma segmentation, demonstrating that T1c +T2f alone is sufficient for clinical-grade preoperative volumetric accuracy and is compatible with the imaging requirements of RANO 2.0. This approach enables workflow simplification, addresses key barriers to volumetric assessment, and provides a robust, open-source AI solution for routine neuro-oncology and future clinical trials.

Supplementary Material

Supplementary material is available online at *Neuro-Oncology Advances* (https://academic.oup.com/noa).

Keywords

glioblastoma | MRI segmentation | deep learning | volumetric assessment | RANO 2.0

Conflict of Interest Statement

None declared.

Funding

National Institutes of Health (U01CA242871); National Cancer Institute; Informatics Technology for Cancer Research (ITCR) funding program. The content of this publication is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

Ethics Approval

This study used exclusively publicly available, de-identified BraTS 2021 data; accordingly, institutional review board approval and informed consent were not required.

Data Availability

The data underlying this article are available in https://github.com/BrainLesion/GlioMODA/

Affiliations

Department of Neuroradiology, TUM School of Medicine, TUM University Hospital rechts der Isar, Technical University of Munich, Munich, Germany (J.C., J.S.K.); Department of Radiation Oncology, TUM School of Medicine, TUM University Hospital rechts der Isar, Technical University of Munich, Munich, Germany (J.B., J.C.P.); Department of Quantitative Biomedicine, University

of Zurich, Zurich, Switzerland (M.R., B.M., F.K.); Helmholtz AI, Helmholtz Munich, Neuherberg, Germany (M.R., M.P., F.K.); Department of Diagnostic, Interventional, and Pediatric Radiology, Inselspital Bern, University of Bern, Bern, Switzerland (M.G.); Deutsches Konsortium für Translationale Krebsforschung (DKTK), Partner Site Munich, Munich, Germany (J.C.P.); Institute of Radiation Medicine (IRM), Helmholtz Center Munich, Munich, Germany (J.C.P.); Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, Indiana, USA (S.B.); Indiana University Melvin and Bren Simon Comprehensive Cancer Center, Indianapolis, Indiana, USA (S.B.); Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University, Indianapolis, Indiana, USA (S.B.); Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, Indiana, USA (S.B.); Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, Indiana, USA (S.B.); Department of Neurological Surgery, Indiana University School of Medicine, Indianapolis, Indiana, USA (S.B.); AI for Image-Guided Diagnosis and Therapy, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany (B.W., F.K.); Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany (F.K.)

References

1. Price M, Ballard C, Benedetti J, et al. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2017-2021. *Neuro-Oncol*. 2024;26:vi1-vi85. https://doi.org/10.1093/neuonc/noae145
2. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol*. 2021;23:1231-1251. https://doi.org/10.1093/neuonc/noab106
3. Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34:1993-2024. https://doi.org/10.1109/TMI.2014.2377694
4. Kofler F, Berger C, Waldmannstetter D, et al. BraTS Toolkit: translating BraTS brain tumor segmentation algorithms into clinical and scientific practice. *Front Neurosci*. 2020;14:125. https://doi.org/10.3389/fnins.2020.00125
5. Stupp R, Mason WP, Van den Bent MJ, et al.; National Cancer Institute of Canada Clinical Trials Group. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med*. 2005;352:987-996. https://doi.org/10.1056/NEJMoa043330
6. Niyazi M, Andratschke N, Bendszus M, et al. ESTRO-EANO guideline on target delineation and radiotherapy details for glioblastoma. *Radiother Oncol*. 2023;184:109663. https://doi.org/10.1016/j.radonc.2023.109663
7. Holtzman Gazit M, Faran R, Stepovoy K, Peles O, Shamir RR. Post-operative glioblastoma multiforme segmentation with uncertainty estimation. *Front Hum Neurosci*. 2022;16:932441. https://doi.org/10.3389/fnhum.2022.932441
8. Beser-Robles M, Castellá-Malonda J, Martínez-Gironés PM, et al. Deep learning automatic semantic segmentation of glioblastoma multiforme regions on multimodal magnetic resonance images. *Int J Comput Assist Radiol Surg*. 2024;19:1743-1751. https://doi.org/10.1007/s11548-024-03205-z
9. 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;20:728-740. https://doi.org/10.1016/S1470-2045(19)30098-1
10. Vollmuth P, Foltyn M, Huang RY, et al. Artificial intelligence (AI)-based decision support improves reproducibility of tumor response assessment in neuro-oncology: an international multi-reader study. *Neuro Oncol*. 2023;25:533-543. https://doi.org/10.1093/neuonc/noac189
11. 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;41:5187-5199. https://doi.org/10.1200/JCO.23.01059
12. Kofler F, Rosier M, Astaraki M, et al. BraTS orchestrator: democratizing and disseminating state-of-the-art brain tumor image analysis. [published online ahead of print June 13, 2025]. *arXiv*. https://arxiv.org/abs/2506.13807.
13. Bakas S, Akbari H, Sotiras A, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*. 2017;4:170117. https://doi.org/10.1038/sdata.2017.117
14. Ellingson BM, Bendszus M, Boxerman J, et al.; Jumpstarting Brain Tumor Drug Development Coalition Imaging Standardization Steering Committee. Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials. *Neuro-Oncol*. 2015;17:1188-1198. https://doi.org/10.1093/neuonc/nov095
15. Zaitsev M, Maclaren J, Herbst M. Motion artifacts in MRI: a complex problem with many partial solutions. *J Magn Reson Imaging*. 2015;42:887-901. https://doi.org/10.1002/jmri.24850
16. Pemberton HG, Wu J, Kommers I, et al. Multi-class glioma segmentation on real-world data with missing MRI sequences: comparison of three deep learning algorithms. *Sci Rep*. 2023;13:18911. https://doi.org/10.1038/s41598-023-44794-0
17. De Sutter S, Wuts J, Geens W, Vanbinst AM, Duerinck J, Vandemeulebroucke J. Modality redundancy for MRI-based glioblastoma segmentation. *Int J Comput Assist Radiol Surg*. 2024;19:2101-2109. https://doi.org/10.1007/s11548-024-03238-4
18. Buchner JA, Peeken JC, Etzel L, et al. Identifying core MRI sequences for reliable automatic brain metastasis segmentation. *Radiother Oncol*. 2023;188:109901. https://doi.org/10.1016/j.radonc.2023.109901
19. Hao D, Ai T, Goerner F, Hu X, Runge VM, Tweedle M. MRI contrast agents: basic chemistry and safety. *J Magn Reson Imaging*. 2012;36:1060-1071. https://doi.org/10.1002/jmri.23725
20. Hooper GW, Ginat DT. MRI radiomics and potential applications to glioblastoma. *Front Oncol*. 2023;13:1134109. https://doi.org/10.3389/fonc.2023.1134109
21. Buchner JA, Kofler F, Mayinger M, et al. Radiomics-based prediction of local control in patients with brain metastases following postoperative stereotactic radiotherapy. *Neuro Oncol*. 2024;26:1638-1650. https://doi.org/10.1093/neuonc/noae098
22. Peeken JC, Wiestler B, Combs SE. Image-guided radiooncology: the potential of radiomics in clinical application. In: Schober O, Kiessling F, Debus J, eds. Molecular Imaging in Oncology. Vol 216. Recent Results in Cancer Research. Springer International Publishing; 2020:773–794.
23. Baid U, Ghodasara S, Mohan S, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification.[published online ahead of print September 12, 2021]. *arXiv*. https://arxiv.org/abs/2107.02314.
24. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008;455:1061-1068. https://doi.org/10.1038/nature07385

25. The Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med*. 2015;372:2481-2498. https://doi.org/10.1056/NEJMoa1402121
26. Puchalski RB, Shah N, Miller J, et al. An anatomic transcriptional atlas of human glioblastoma. *Science*. 2018;360:660-663. https://doi.org/10.1126/science.aaf2666
27. Wang LB, Karpova A, Gritsenko MA, et al.; Clinical Proteomic Tumor Analysis Consortium. Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell*. 2021;39:509-528.e20. https://doi.org/10.1016/j.ccell.2021.01.006
28. Gerstner ER, Zhang Z, Fink JR, et al.; ACRIN 6684 Trial Group. ACRIN 6684: assessment of tumor hypoxia in newly diagnosed glioblastoma using 18F-FMISO PET and MRI. *Clin Cancer Res*. 2016;22:5079-5086. https://doi.org/10.1158/1078-0432.CCR-15-2529
29. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18:203-211. https://doi.org/10.1038/s41592-020-01008-z
30. Isensee F, Wald T, Ulrich C, et al. nnU-Net revisited: a call for rigorous validation in 3D medical image segmentation. In: Linguraru MG, Dou Q, Feragen A, et al., eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science*. Vol. 15009. Cham: Springer; 2024.
31. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol*. 2004;11:178-189. https://doi.org/10.1016/s1076-6332(03)00671-8
32. Kirillov A, He K, Girshick R, Rother C, Dollár P. Panoptic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2019:9404–9413.
33. Kofler F, Möller H, Buchner JA, et al. Panoptica—instance-wise evaluation of 3D semantic and instance segmentation maps.[published online ahead of print December 5, 2023]. *arXiv*. https://arxiv.org/abs/2312.02608.
34. Wilcoxon F. Individual comparisons by ranking methods. *Biom Bull*. 1945;1:80. https://doi.org/10.2307/3001968
35. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29:1165-1188.
36. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9:90-95. https://doi.org/10.1109/MCSE.2007.55
37. Waskom M. Seaborn: statistical data visualization. *JOSS*. 2021;6:3021. https://doi.org/10.21105/joss.03021
38. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging*. 2023;14:75. https://doi.org/10.1186/s13244-023-01415-8
39. S Bakas, A Crimi, U Baid, et al., eds. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries: 8th International Workshop, BrainLes 2022, Held in Conjunction with MICCAI 2022; September 18; 2022; Singapore. Revised Selected Papers, Part I. Vol. 13769. Switzerland: Springer Nature; 2023.
40. Dorfner FJ, Patel JB, Kalpathy-Cramer J, Gerstner ER, Bridge CP. A review of deep learning for brain tumor analysis in MRI. *NPJ Precis Oncol*. 2025;9:2. https://doi.org/10.1038/s41698-024-00789-2
41. Tian S, Liu Y, Mao X, et al. A multicenter study on deep learning for glioblastoma auto-segmentation with prior knowledge in multimodal imaging. *Cancer Sci*. 2024;115:3415-3425. https://doi.org/10.1111/cas.16304
42. Li HB, Conte GM, Hu Q, et al. The Brain Tumor Segmentation (BraTS) Challenge 2023: brain MR image synthesis for tumor segmentation (BraSyn). [published online ahead of print November 24, 2024]. *arXiv*. https://arxiv.org/abs/2305.09011.
43. Feng X, Ghimire K, Kim DD, et al. Brain tumor segmentation for multi-modal MRI with missing information. *J Digit Imaging*. 2023;36:2075-2087. https://doi.org/10.1007/s10278-023-00860-7
44. Kofler F, Rosier M, Astaraki M, et al. BrainLesion suite: a flexible and user-friendly framework for modular brain lesion image analysis. [published online ahead of print July 11, 2025]. *arXiv*. https://arxiv.org/abs/2507.09036.
45. Kazerooni AF, Khalili N, Liu X, et al. The Brain Tumor Segmentation (BraTS) challenge 2023: focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). [published online ahead of print May 23, 2024]. *arXiv*. https://arxiv.org/abs/2404.15009.
46. Piffer A, Buchner JA, Gennari AG, et al. Enhancing efficiency in paediatric brain tumour segmentation using a pathologically diverse single-center clinical dataset. *Neuro-Oncol. Adv*. 2026;vdag024. https://doi.org/10.1093/noajnl/vdag024
47. Barry N, Kendrick J, Rowshanfarzad P, et al. An external, independent validation of an O-(2-[18 F]fluoroethyl)-l-tyrosine PET automatic segmentation network on a single-center, prospective dataset of patients with glioblastoma. *J Nucl Med*. 2025;66:948-953. https://doi.org/10.2967/jnumed.124.268925
48. Wen L, Sun H, Liang G, Yu Y. A deep ensemble learning framework for glioma segmentation and grading prediction. *Sci Rep*. 2025;15:4448. https://doi.org/10.1038/s41598-025-87127-z