This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Enhancing Clinically-relevant Quality Control Through Automated Segmentation Performance Prediction

Daria Laslo[1,2][0009−0005−8771−144X], Victor IJ Strijbis[3][0000−0003−0941−6005], Atlas Haddadi Avval[4][0000−0002−3896−7810], Franziska Vogt[1][0009−0000−6511−8411], Anahita Fathi Kazerooni[5][0000−0001−5830−8890], Zhifan Jiang[6][0000−0001−5713−1675], Abhijeet Parida[6,7][0000−0002−4978−0576], Benjamin Kann[8][0000−0002−4313−2754], Marius George Linguraru[6][0000−0001−6175−8665], Andrea Franson[5][0000−0002−5361−7683], Sabine Müller[4][0000−0002−3452−5150], Andreas M. Rauschecker[4][0000−0003−0633−9876], Catherine R. Jutzeler[1,2][0000−0001−7167−8271], and Sarah Brüningk[3,9][0000−0003−3176−1032]

[1] ETH Zurich, 8092 Zurich, Switzerland - daria.laslo@hest.ethz.ch
[2] SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[3] Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Switzerland
[4] University of California San Francisco, San Francisco, CA 94143, United States
[5] Children's Hospital of Philadelphia, Philadelphia, United States
[6] Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Medical Center, Washington, DC 20010, United States
[7] ETSIT, Universidad Politécnica de Madrid, Madrid, Spain
[8] Brigham and Women's Hospital, Dana-Farber Harvard Cancer Center, Boston, United States
[9] Department of Digital Medicine, University of Bern, Bern, Switzerland

Abstract. Automated segmentation enables large-scale curation of retrospective imaging cohorts, but unrecognized failures can bias downstream analyses, particularly in pediatric neuro-oncology, where multi-institutional datasets are small and longitudinal tumor burden drives response assessments. We study quality control (QC) for automated diffuse midline glioma segmentation, benchmarking methods to identify poor segmentations. We compare supervised QC predictors based on a combination of acquisition metadata, radiomics, MRI volumes with predicted masks, and foundation-model MRI embeddings, Dice estimation, and simple model-derived baselines using inter-model disagreement, softmax entropy, and predicted tumor volume. Crucially, we assess QC by its effect on a clinical endpoint: radiotherapy-response labels derived from longitudinal tumor-volume change. Dice estimation is the strongest failure detector by AUPRC (0.933), but volume-based thresholding recovers more label agreement per re-annotation. At the endpoint level, uncorrected segmentations systematically over-assign favorable response categories relative to manual references, and volume-based selective re-annotation of scans recovers agreement at low review burden. These findings argue for endpoint-aware segmentation QC, optimized for the clinical labels segmentations support rather than for Dice prediction alone.

Keywords: Quality control · Segmentation · Pediatric neuro-oncology

1 Introduction

Pediatric neuro-oncology relies heavily on image-based response assessments, with standardized criteria such as RAPNO (Response Assessment in Pediatric Neuro-Oncology) increasingly incorporating quantitative, volumetric tumor burden alongside conventional diameter-based measures [4,3,14]. In practice, volumetric assessments depend on manual MRI segmentation, which is labor-intensive and subject to inter-rater variability [17], creating a bottleneck for large-scale retrospective and multi-institutional studies.

Automated tumor-volume segmentation is central to overcoming this bottleneck. Deep learning methods now achieve expert-level performance for brain-tumor delineation [10,5,18], including pediatric models [11,5]. However, these models are often trained and evaluated on diagnostic or pre-treatment imaging. Since radiotherapy (RT) and systemic treatment induce necrosis, pseudoprogression, edema, and other phenotypic changes [22], post-treatment images may fall outside model training distributions [1]. In this domain-shift setting, segmentation models usually underperform, leading to segmentation errors that can propagate from automated tumor-volume estimates into response classifications. Quality control (QC) is critical in flagging such cases. Yet current segmentation QC methods are typically judged by scan-level geometric performance, such as Dice, and are rarely tied to the response endpoints. This gap is particularly relevant for longitudinal response assessment, where labels depend on relative tumor-volume change across timepoints.

To address this gap, we present a systematic comparison of QC strategies for automated tumor segmentation. Using a multi-institutional diffuse midline glioma (DMG) dataset, we benchmark learned predictors and model-derived confidence baselines for identifying poor segmentations, and evaluate their impact on a downstream RT-response endpoint derived from longitudinal tumor-volume change. We show that uncorrected automated segmentations introduce a directional response-assessment bias, over-assigning favorable outcomes relative to manual reference labels. Finally, we simulate a flag–re-annotate–relabel workflow to quantify how selective manual review can recover clinical-label agreement while reducing annotation burden. While our experiments focus on DMG RT response, the framework is applicable to other longitudinal imaging endpoints that depend on automated segmentation.

1.1 Related work

Recent automatic segmentation models span CNN [8,9] and transformer models [7] with self-configuring nnU-Net [8] and their ensembles [13,5,11] being the dominant approach in both adult and pediatric brain MRI segmentation. However, when these models are applied to small pediatric cohorts, reliable QC remains challenging, particularly in the absence of ground-truth annotations.

Most QC methods estimate segmentation quality indirectly via model uncertainty methods such as model ensembles, Monte-Carlo dropout, entropy [12]. Other complementary approaches include the direct estimation of Dice coefficient from calibrated predictive probabilities [16,15]. These have been demonstrated to be effective in scenarios involving moderate performance degradation (min. Dice ~ 0.7), but have not yet been validated for cases of complete segmentation failure.

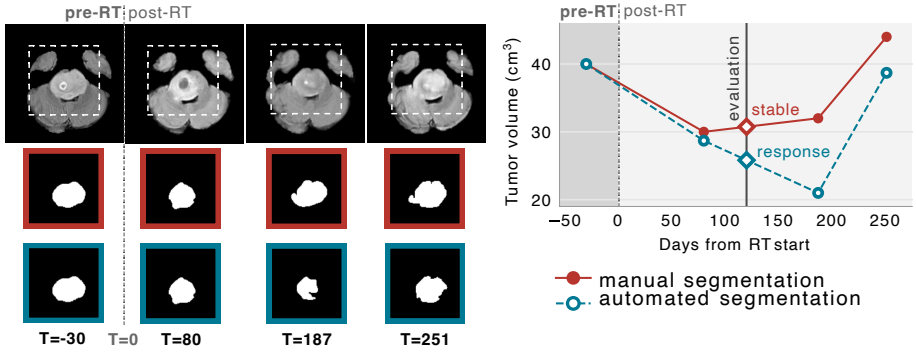

Fig. 1: Tumor-volume trajectories from manual and automated segmentation diverge, flipping the label from stable ($\Delta = -23\%$) to response ($\Delta = -35\%$).

Task-driven proxy approaches, such as reverse classification accuracy [20], instead characterize dataset-specific failures by quantitatively assessing the extent to which a predicted output can be reused as a pseudo-label. Despite their diversity, these methods share a common limitation. Quality is assessed purely in the segmentation space, leaving open whether identified errors actually influence subsequent clinical decision making. To our knowledge, none have been validated for longitudinal DMG assessment or benchmarked against a response label that ultimately guides treatment monitoring.

2 Quality Control for longitudinal RT labeling

We consider a cohort of longitudinal MRI scans, each with an input image x_i, an automated segmentation $m_i = \mathcal{S}(x_i)$, a binary segmentation-quality label $y_i \in \{0, 1\}$, and a downstream RT-response label r_i derived from longitudinal tumor-volume change. We define a QC predictor as a scan-level scoring function

$$q_i = f_\phi(x_i, m_i, z_i) \in [0, 1],$$

where q_i is the probability of poor segmentation quality and z_i denotes any auxiliary information used by the predictor. Different QC strategies correspond to different instantiations of f_ϕ and z_i: model-derived baselines derive z_i from

confidence or disagreement signals produced by $\mathcal{S}$, feature-based predictors use metadata, radiomics, or MRI embeddings, and learned multimodal predictors operate on the image-mask pair. A scan is flagged for review if $q_i \geq \tau$. For flagged scans, the automated mask is replaced by the manual reference mask,$\tilde{m}_i = m_i^{\text{man}}$ if $q_i \geq \tau$, else m_i. The reviewed masks induce revised RT-response labels $\tilde{r}_i$, allowing us to measure both failure-detection performance and the effect of selective re-annotation on clinical label agreement.

3 Implementation Details

3.1 Segmentation-quality predictors

We evaluate four learned predictors of segmentation quality alongside three baselines derived directly from the segmentation model's output and confidence. All models are trained in a 5-fold cross-validation setting with patient-level splits.

Model-derived baselines. Three approaches assess the scan directly from the segmentation model's output. *Ensemble disagreement* scores each scan by the volumetric disagreement (1 − Dice) between two state-of-the-art pediatric segmentation models [11,2]. *Softmax entropy* averages per-voxel binary entropy over the predicted foreground from calibrated nnU-Netv2 [9] probabilities, with higher entropy indicating poorer quality. *Volume thresholding* flags scans with fewer than 524 predicted tumor voxels, the volumetric extension of the minimal measurable lesion size in RAPNO (1cm diameter sphere) [4].

Metadata and radiomics. As an interpretable baseline, we train gradient boosting classifiers on acquisition metadata (scanner, acquisition parameters, view plane, image dimensions) and radiomic features [6] extracted from the predicted mask, evaluating metadata alone or combined with radiomics and predicted volume. Features are reduced to $k = 30$ via ANOVA F-test selection on the validation set.

Learned image and image-mask representations. For MRI-only prediction, we fine-tune BrainIAC [19], a 3D ViT-B foundation model pretrained with SimCLR contrastive learning. Per-contrast 768-dimensional class-token embeddings are concatenated, passed through dropout ($p = 0.2$) and a linear head, and fine-tuned end-to-end with Adam under cosine-annealing warm restarts. A complementary ResNet-style predictor instead learns image-mask compatibility directly, taking a 5-channel scan-plus-segmentation volume as input. The backbone is a 3D ResNeXt-style residual encoder (ExtResNetBlock) with global average pooling and an MLP head trained with binary cross-entropy on foreground-biased 128^3 patches (Adam, lr 10^{-5}, weight decay 10^{-4}, batch size 4, up to 1000 epochs, early stopping).

Calibrated Dice estimation. Finally, we estimate Dice directly from calibrated nnU-Netv2 [9] softmax probabilities. Pseudo-logits z are rescaled by a temperature parameter T, $p_{\text{cal}} = \sigma(z/T)$, and Dice is estimated with the soft-Dice estimator of Li et al. [16],

$$\widehat{\text{Dice}} = \frac{2\sum_{p>0.5} p}{|\{p > 0.5\}| + \sum_p p}. \tag{1}$$

T is fit per fold on a calibration split ($\sim 20\%$ of patients) by minimizing the absolute difference between mean estimated and mean true Dice, then evaluated on the held-out majority ($\sim 80\%$). We threshold $\widehat{\text{Dice}}$ at 0.8 to flag poor quality.

3.2 Dataset

We use a multi-institutional DMG dataset comprising 110 patients with 468 multiparametric MRI scans (T1w, T1w-CE, T2w, T2w-FLAIR) from three contributing centers (anonymized A, B, C). Prior to segmentation, all images were registered, skull-stripped, and resampled to 1mm isotropic voxel resolution. All scans were manually corrected by a medical trainee and reviewed by a certified neuroradiologist, starting from an automated segmentation. All patients received comparable irradiation regimens, with 83 of them having both pre- and post-RT scans enabling the assessment of response. Information on possible concurrent systemic therapy was not available.

Automatic Segmentation and quality label generation. Automated segmentations were generated using two models drawn from successive generations: an nnU-Net/Swin UNETR ensemble [11] and a recent large-scale pretrained pipeline refined through adaptive post-processing [2]. Per-scan Dice is strongly correlated across these models (Spearman ρ=0.82). We build all quality predictors on the best-performing model [2], which achieves a median Dice of 0.92 yet still exhibits a distinct sub-population of near-zero failures. We label a scan as *poorly segmented* when the Dice falls below 0.8, set at the natural break (rounded antimode of kernel density estimate), yielding 29% poor scans.

3.3 Clinical evaluation.

We assess segmentation QC through its clinical relevance in two steps: (1) how automated segmentation affects a radiotherapy-response endpoint relative to reference labels derived from purely manual segmentations, and (2) simulate a QC-driven flag–re-annotate–relabel pipeline to measure how much of that discrepancy manual review of selected scans recovers.

Radiotherapy response endpoint. This analysis is restricted to the first RT course of the eligible 83 patients ($\sim$ 350 scans). In line with RAPNO [4], we use a paired scan comparison between the most recent scan before RT start and the one closest to 120 days after as the evaluation. The relative tumor-volume change was calculated. If more than one post-RT scan is available, we linearly interpolate the volume at 120 days. Based on the volumetric change, we classify patients as *progressive disease* (PD), *stable disease* (SD), or *partial response* (PR) using volume-based thresholds (+25% for PD, −25% for PR and SD otherwise). We compute the RT-response label from both the automated and manual segmentations, and report agreement between the two label sets, and the transition matrix between them to quantify label changes.

Post-RT segmentation post-processing. We further investigated a systematic inflation of post-RT volumes by computing a correction factor, defined as the

Table 1: AUROC and AUPRC across segmentation quality prediction approaches, mean and standard deviation across folds is reported.

Metric	Model-derived				Learned classifiers					
	Ensemble disagr.	Volume thresh.	Softmax entropy	Dice estim.	Meta-data	Metadata & pred. vol	Radio-mics	Radiomics & pred. vol	ResNet pred	Brain-IAC
AUROC	**0.908** ±0.01	0.800 ±0.02	0.781 ±0.01	0.906 ±0.003	0.549 ±0.06	0.668 ±0.04	0.767 ±0.06	0.855 ±0.05	0.636 ±0.05	0.727 ±0.10
AUPRC	0.783 ±0.02	0.701 ±0.04	0.623 ±0.03	**0.933** ±0.004	0.300 ±0.06	0.377 ±0.06	0.447 ±0.07	0.750 ±0.11	0.355 ±0.08	0.505 ±0.02

median ratio of reference to predicted post-RT volume across a held-out 20% calibration split, and applied it to the remaining scans.

Flag–re-annotate–relabel pipeline. Let a segmentation quality predictor select scans for manual review. We simulate re-annotation by replacing the automated segmentation on those scans with the manual-segmentation, recompute RT-response labels, and measure agreement of the resulting *reviewed* labels relative to the *true* (fully manual) labels. In the case of the ensemble disagreement or entropy we flag the top 29% of scans for review. For the learned predictors, we flag all scans predicted as poor quality. We present these as a function of the percentage of label agreement with the true labels, and the number of scans flagged for review representing the effort. Additionally, we compute prioritization curves, where scans are reviewed worse-first as predicted by the different approaches, recomputing agreement after each review. We summarize curves by the area under the agreement curve (AUAC).

Statistical analysis. Post-treatment undersegmentation was assessed via the relative volume discrepancy $(V_{\text{man}} - V_{\text{auto}})/V_{\text{man}}$, comparing pre- vs post-RT with a paired Wilcoxon test. Automated-vs-manual label shift was tested with the McNemar–Bowker test of symmetry. Confidence intervals (CI) were computed by patient-level bootstrap (1000 resamples).

4 Experimental Results

4.1 Performance of segmentation quality predictors

Table 1 presents a comparison across all tested strategies, both model-derived and learned classifiers, for the binary prediction of poor/adequate segmentations. Overall, model confidence-based methods outperform most of the supervised models. Calibration-based Dice estimation achieves the highest AUPRC (0.933 ± 0.004) and a near-best AUROC (0.906 ± 0.003), both with negligible variance across folds and comparable to ensemble disagreement. Most notably among the tested baselines, simple volume thresholding reaches an AUROC of 0.800 ± 0.02, outperforming all learned classifiers except the classifier building on radiomic features and predicted volume which yields a comparable AUROC of 0.855. The other notable performer is the foundation-model embedding, reaching an AUROC of 0.727.

4.2 Clinical-endpoint impact

We systematically evaluated automated segmentation impact on fixed time-point volumetric threshold-based RT-response labels. Fig. 2B summarizes these results as a Sankey diagram as an aggregation over the 83 patients with RT information. While auto-segmentation and manual segmentation derived labels agree for 80% of patients (n=67), in the remaining 20% of disjoint cases (n= 16) we observe a systematic shift of autosegmentation-based response labels towards improved response categories: automated segmentation over-assigns response (PR: 47 vs. 38 manual) while under-assigning progression (PD: 21 vs. 24) and stable disease (SD: 15 vs. 21). A McNemar–Bowker test of symmetry rejects the null of incidental, non-directional disagreement ($\chi^2(3) = 10.36$, $p = 0.016$). However, the magnitude is estimated from 16 discordant patients and should be treated as indicative rather than a calibrated effect size. Notably, the shift persisted even when the volume threshold was relaxed to 35%. The net effect is an inflation of favorable outcomes and traces to a post-RT under-segmentation as shown in Fig. 1. We assess volume difference across the manually and auto-segmented contours by statistical comparison. We observed a relative volume discrepancy (manual − automated) rising from a median of 1.0% before RT to 17.5% after (paired Wilcoxon $p < 10^{-8}$).

4.3 Re-annotation efficiency

Finally, we assess each strategy as a flagging-for-re-annotation tool, quantifying the agreement with the fully manual labels recovered per unit of review effort (Fig. 2A). For each approach, the identified poor quality scans are re-annotated and the resulting label agreement is benchmarked against a random-selection of the same number of scans. All strategies improve over the uncorrected automated baseline (agreement 0.807, 95% CI [0.72,0.89] patient-level bootstrap, n=83). Given wide CIs due to the small cohort, we focus on the relative efficiency of the methods: confidence-based and most learned classifiers lie on or above the null band, while metadata-only offers no advantage over blind review. Thresholding the volume, a simple, yet effective approach achieves 87% agreement with the manual labels and, when combined with a systematic inflation of post-RT volumes, agreement reaches $\sim$0.90, for only $\sim$20 re-annotations.

Beyond this fixed operating point, we also consider prioritized review independent of any classification threshold: scans are ranked by predicted quality and re-annotated worst-first (Fig. 2C). All methods increase agreement from the uncorrected baseline (0.80) and are tightly clustered (AUAC 0.911–0.926), with radiomics-and-predicted-volume being the most efficient. The curves show that prioritized review recovers most of the achievable agreement early, but reaching 90% by re-annotation alone still requires inspecting $\sim$36% of the cohort.

4.4 Discussion

Evaluating segmentation QC by its downstream impact rather than segmentation-space performance alone exposes a divergence: methods that predict low Dice well

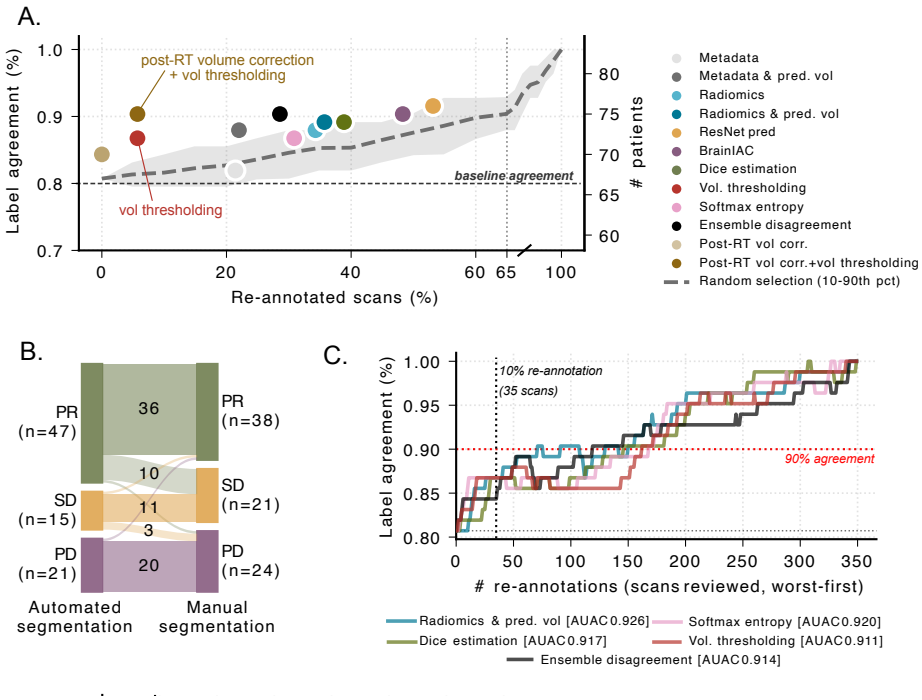

Fig. 2: A. Agreement with true labels as a function of percentage of re-annotations for different approaches. B. Transitions between automatic segmentation-based labeling and manual-annotation-based labeling. C. Prioritization curves (re-annotations in the order of poor quality likelihood) for top 5 performing approaches. Area under agreement curve (AUAC) is also shown.

do not necessarily flag the samples that alter the radiotherapy-response label. In line with [21], we observe systematic post-RT undersegmentation, which induces not random error but a directional shift toward favorable outcomes: over-assigning response while under-assigning stable and progressive disease. Calibration-based Dice estimation was the most reliable detector. In the endpoint setting, however, the simplest approach was most effective: volume thresholding with an explicit correction for the post-treatment inflation recovered comparable label agreement (~ 0.90) for only ~ 20 re-annotations, whereas prioritized review by any flagger required inspecting $\sim$36% of the cohort to reach 90%. Modeling the dominant failure mode thus outperforms data-specific learned predictors.

Our findings should be interpreted in light of some limitations. The directional shift is statistically robust, but it rests on 16 discordant cases, so its precise magnitude may be specific to this cohort and should be confirmed in larger, independent datasets. Our manual segmentations also reflect a biased protocol in which reviewers corrected the predicted mask, chosen to promote review speed, which limits interpretation of the absolute segmentation performance and

prevents direct comparison with blinded assessment. Method ranking should be less affected, as all predictors share the same reference, though this anchoring likely understates achievable QC performance. These points motivate future work extending the framework beyond DMGs, developing segmentation adapted to the altered post-treatment phenotype, and incorporating explainable-AI approaches that could inform QC decisions.

5 Conclusion

Segmentation QC should be assessed not through single-metric optimization alone, but by quantifying its effect on the downstream task. In our longitudinal setting, this reframing exposed a systematic, directional post-treatment bias that traditional metrics mask. Because these response labels are central to our research question and to longitudinal monitoring, such a bias is clinically relevant regardless of segmentation-space performance. Explicitly correcting discovered biases can minimize manual re-annotation effort while preserving the clinical relevance of the measurements, and offers a practical route to efficient cohort curation for longitudinal response assessment.

Acknowledgments. We acknowledge the ChadTough Defeat DIPG Foundation (#1477715) and the Swiss National Science Foundation (Starting Grant TMSGI3_225913, ADMIRA Grant 10001696) for funding this project. Computational data analysis was performed at Leonhard Med (https://sis.id.ethz.ch/services/sensitiveresearchdata), the secure trusted research environment at ETH Zurich. We thank Ariana M. Familiar for her support with data processing.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Code availability The source code is available at: https://gitlab.ethz.ch/BMD Slab/publications/oncology/clinically-relevant-qc-for-automated-segmentation

References

1. Bibault, J.E., Giraud, P.: Deep learning for automated segmentation in radiotherapy: a narrative review. The British Journal of Radiology **97**(1153), 13–20 (Dec 2023). https://doi.org/10.1093/bjr/tqad018
2. Capellán-Martín, D., Parida, A., Jiang, Z., Kulkarni, N., Iyer, K., Tapp, A., Anwar, S.M., Ledesma-Carbayo, M.J., Linguraru, M.G.: Adaptable Segmentation Pipeline for Diverse Brain Tumors with Radiomic-guided Subtyping and Lesion-Wise Model Ensemble. Segmentation, classification, and synthesis for brain tumors and traumatic brain injuries : MICCAI 2025 Challenges: BraTS-Lighthouse 2025 and AIMS-TBI 2025, held in conjunction with MICCAI 2025, Daejeon, South Korea, September 23, 2025, **16376**, 456–467 (2026). https://doi.org/10.1007/978-3-032-16365-3_41

3. Cooney, T.M., Cohen, K.J., Guimaraes, C.V., Dhall, G., Leach, J., et al.: Response assessment in diffuse intrinsic pontine glioma: recommendations from the Response Assessment in Pediatric Neuro-Oncology (RAPNO) working group. Lancet Oncol. **21**(6), e330–e336 (Jun 2020). https://doi.org/10.1016/S1470-2045(20)30166-2
4. Erker, C., Tamrazi, B., Poussaint, T.Y., Mueller, S., Mata-Mbemba, D., et al.: Response assessment in paediatric high-grade glioma: recommendations from the Response Assessment in Pediatric Neuro-Oncology (RAPNO) working group. Lancet Oncol. **21**(6), e317–e329 (Jun 2020). https://doi.org/10.1016/S1470-2045(20)30173-X
5. Fathi Kazerooni, A., Khalili, N., Liu, X., Haldar, D., Jiang, Z., et al.: BraTS-PEDs: Results of the multi-consortium international pediatric Brain Tumor Segmentation challenge 2023. J. Mach. Learn. Biomed. Imaging **3**(June 2025), 72–87 (Jun 2025). https://doi.org/10.59275/j.melba.2025-f6fg
6. van Griethuysen, J.J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., et al.: Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research **77**(21), e104–e107 (Nov 2017). https://doi.org/10.1158/0008-5472.CAN-17-0339
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images (Jan 2022). https://doi.org/10.48550/arXiv.2201.01266, arXiv:2201.01266 [eess.IV]
8. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (Feb 2021). https://doi.org/10.1038/s41592-020-01008-z
9. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation (Jul 2024). https://doi.org/10.48550/arXiv.2404.09556, arXiv:2404.09556 [cs.CV]
10. Jain, I., Willems, S., Latre, S., Schepper, T.D.: On-the-Fly Data Augmentation for Brain Tumor Segmentation (Sep 2025). https://doi.org/10.48550/arXiv.2509.24973, arXiv:2509.24973 [cs.CV]
11. Jiang, Z., Capellán-Martín, D., Parida, A., Tapp, A., Liu, X., Ledesma-Carbayo, M.J., Anwar, S.M., Linguraru, M.G.: Magnetic Resonance Imaging Feature-Based Subtyping and Model Ensemble for Enhanced Brain Tumor Segmentation (Dec 2024). https://doi.org/10.48550/arXiv.2412.04094, arXiv:2412.04094 [eess.IV]
12. Jungo, A., Balsiger, F., Reyes, M.: Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation. Front. Neurosci. **14** (Apr 2020). https://doi.org/10.3389/fnins.2020.00282
13. Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., Glocker, B.: Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation. Lect. Notes Comput. Sci. **10670 LNCS**, 450–462 (Nov 2017). https://doi.org/10.1007/978-3-319-75238-9_38
14. Kann, B.H., Vossough, A., Brüningk, S.C., Familiar, A.M., Aboian, M., et al.: Artificial Intelligence for Response Assessment in Pediatric Neuro-Oncology (AI-RAPNO), part 1: review of the current state of the art. The Lancet Oncology **26**(11), e597–e606 (Nov 2025). https://doi.org/10.1016/S1470-2045(25)00484-X
15. Köhler, P., Fadugba, J., Berens, P., Koch, L.M.: Efficiently correcting patch-based segmentation errors to control image-level performance in retinal images. In: Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning. pp. 841–856. PMLR (Dec 2024)

16. Li, Z., Kamnitsas, K., Islam, M., Chen, C., Glocker, B.: Estimating Model Performance under Domain Shifts with Class-Specific Confidence Scores (Jul 2022). https://doi.org/10.48550/arXiv.2207.09957, arXiv:2207.09957 [cs.CV]
17. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (Oct 2015). https://doi.org/10.1109/TMI.2014.2377694
18. Parida, A., Capellán-Martín, D., Jiang, Z., Kulkarni, N., Iyer, K., Tapp, A., Anwar, S.M., Ledesma-Carbayo, M.J., Linguraru, M.G.: Improving pre-trained adult glioma segmentation models using only post-processing techniques. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 237–247. Springer (2025)
19. Tak, D., Garomsa, B.A., Zapaishchykova, A., Chaunzwa, T.L., Climent Pardo, J.C., Ye, Z., Zielke, J., Ravipati, Y., Pai, S., Vajapeyam, S., Mahootiha, M., Parker, M., Pike, L.R.G., Smith, C., Familiar, A.M., Liu, K.X., Prabhu, S., Arnaout, O., Bandopadhayay, P., Nabavizadeh, A., Mueller, S., Aerts, H.J.W.L., Huang, R.Y., Poussaint, T.Y., Kann, B.H.: A generalizable foundation model for analysis of human brain MRI. Nature Neuroscience **29**(4), 945–956 (2026). https://doi.org/10.1038/s41593-026-02202-6
20. Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B.: Reverse Classification Accuracy: Predicting Segmentation Performance in the Absence of Ground Truth (Feb 2017). https://doi.org/10.48550/arXiv.1702.03407, arXiv:1702.03407 [cs.CV]
21. Verdier, M.C.d., Saluja, R., Gagnon, L., LaBella, D., Baid, U., Tahon, N.H., et al.: The 2024 Brain Tumor Segmentation (BraTS) Challenge: Glioma Segmentation on Post-treatment MRI (May 2024). https://doi.org/10.48550/arXiv.2405.18368, arXiv:2405.18368 [cs.CV]
22. Walker, A.J., Ruzevick, J., Malayeri, A.A., Rigamonti, D., Lim, M., Redmond, K.J., Kleinberg, L.: Postradiation imaging changes in the CNS: how can we differentiate between treatment effect and disease progression? Future oncology (London, England) **10**(7), 1277–1297 (May 2014). https://doi.org/10.2217/fon.13.271