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.

Supplementary Material Reliability, Not Accuracy, Is the Bottleneck in AI-assisted Post-Treatment Glioma Segmentation: Region-Aware Recalibration and Risk-Controlled Triage

Mostafa Abdalla Mostafa Badr[1⋆], Abdelhadi Nabih Ahmed Ibrahim[1], and Reda Abdelwahab Ahmed El-Khoribi[2]

[1] Faculty of Graduate Studies for Statistical Research, Cairo University, Giza, Egypt
⋆badrmustafa.mb@gmail.com (corresponding author)
[2] Faculty of Computers and Artificial Intelligence, Cairo University, Giza, Egypt

This supplement gives (S1) the ensemble decomposition and its size-matched control, (S2) robustness checks for the cavity-triage guarantee: sensitivity to the calibration fraction, a boundary-localized signal comparison, and the effect of the finite-sample margin, and (S3) the full statistical analysis protocol. All evaluation is on the temporally separated, label-complete 271-case held-out partition. Pooled over all voxels ECE is < 0.001 everywhere (background dominates), so we evaluate region-conditionally; adaptive equal-mass binning changes the ensemble's pooled ECE by ≤ 0.003 at every region, so the result is not a bin-placement artifact.

S1 Ensemble decomposition and size-matched control

Cavity ECE falls $0.243 \rightarrow 0.182 \rightarrow 0.130$ across a single fold, the five-fold within-architecture ensemble, and the 30-checkpoint cross-architecture ensemble: the single-to-five-fold step is $\approx 54\%$ of the reduction, and the remaining $\approx 46\%$ is an irreducible architectural-diversity residual. Holding checkpoint count at the five-fold size, 50 random mixed-architecture draws (mean 2.74 architectures) reach 0.152 ± 0.007 versus 0.182, isolating a diversity effect; the further drop to 0.130 reflects the added checkpoints. Higher-capacity recalibration baselines are left to future work; the two-parameter model already outperforms a global scalar, and a learned nonlinear head reaches only 0.167 ECE, so the signal is low-dimensional.

Table S1. Resection-cavity calibration error (ECE) by ensemble breadth, with the size-matched control.

Configuration	Cavity ECE
Single fold (nnU-Net f0)	0.243
5-fold within-architecture (nnU-Net)	0.182
5-checkpoint mixed-architecture (size-matched, 50 draws)	0.152 ± 0.007
Cross-architecture (30-fold)	0.130

S2 Cavity-triage guarantee: robustness checks

Raising the calibration fraction (50 → 80%) raises feasibility (MedNeXt 29 → 58% of splits) but not the transfer rate: nnU-Net $\geq$ 91%, Swin $\approx$ 88%, MedNeXt 79–85%, never reaching 90%, a shortfall reflecting its high baseline cavity segmentation error. A boundary-localized rim signal is uniformly worse (feasible on 0–5% of splits; AURC 0.24–0.35 vs. 0.16–0.23), so we keep the region signal. Removing the finite-sample Hoeffding–Bentkus margin collapses the transfer rate to $\approx$ 0.52–0.55 for all three; the margin is what makes the guarantee transfer. Restricting entropy to the cavity candidate region costs no measurable inference-time overhead relative to the whole-case signal (135.5 vs. 138.5 ms/case; boundary-restricted, 118.9 ms/case; verbatim source functions on held-out cases, 500 repeats): entropy is computed over the full probability volume regardless of the mask, with region restriction applied only at the final averaging step, so the choice between signals is one of quality, not runtime. Alternative whole-case aggregation strategies (e.g., max- or percentile-based entropy summaries) are likewise left to future work; we would not expect these to close the gap, since the whole-case signal's failure stems from measuring uncertainty over the wrong spatial region, not from how it is aggregated within that region. A full sensitivity sweep (38 combinations: $\alpha \in [0.15, 0.35] \times \text{cov_floor} \in [0.30, 0.70]$ at $\delta = 0.10$, plus $\delta \in [0.05, 0.20]$ at the Table 3 operating point ($\alpha = 0.25$, cov_floor= 0.50), using the certification code on the complete 185-case clean cohort (single-fold nnU-Net), confirms the region-vs-whole-case gap holds at scale: at the operating point in Table 3, REGION certifies on 63% of splits (median 65% retained, test RC-error 0.154, guarantee held on 98.4% of feasible splits), while GLOBAL certifies on 0% at every one of the 38 combinations tested. REGION's feasibility rises monotonically as either bound relaxes (39% → 88% across the δ sweep alone). The ≤ 0.25 risk target reflects published post-operative inter-rater agreement among experts [14, main text], reported as low as 0.32 overlap via a related, not identical, conformity metric.

S3 Statistical analysis protocol

Bootstrap resampling (Methods) used 1000–2000 resamples per metric; a paired difference whose CI excludes zero is treated as a real effect, and the triage guarantee itself uses the distribution-free Hoeffding–Bentkus bound, not a bootstrap.