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.

GAT-26: Release-Path Auditing and Confirmation-Gated Inference Selection for Cross-Tumor Brain Tumor Segmentation

Nathan Chen*

Kang Chiao International School, Xiugang Campus, New Taipei City, Taiwan
naifenchen52@gmail.com

Abstract. We ask whether, for a fixed containerized segmentation system, a commit-frozen multi-gate policy audit can prevent adoption of an inference modification whose aggregate development-subset benefit does not transfer consistently to a same-corpus policy-selection holdout. GAT-26 is a five-fold region-based residual-encoder nnU-Net trained only on subchallenge data from a recorded random initialization. Our contribution is procedural: candidate inference modifications were executed end-to-end through the release machinery on excluded-fold predictions, with candidates, utility and gates fixed in a pushed commit before the specified comparison. The utility is the unweighted mean of six raw component means, ET/TC/WT by DSC and NSD, on common subject support. Eightfold mirroring improved all six DSC/NSD components on the development folds but failed a frozen lesion false-negative veto and stopped. A separately commit-frozen follow-up opened the untouched holdout folds; the gain was descriptively smaller: +0.00073 against +0.00253 at the official-ranking tolerance, +0.00247 against +0.00417 in a lower-tolerance sensitivity analysis. The holdout bootstrap intervals spanned zero across tolerances, one fold changed sign, and the reference-lesion miss rate regressed by +0.0070 against a 0.0025 operational margin, so the baseline was retained. We document a discrepancy easy to ship unnoticed: our cross-validation used eightfold mirroring while the released container does not, so they are not comparable. This is a bounded case study, not evidence the procedure is preferable to nested validation.

Keywords: Brain tumor segmentation · Generalization · nnU-Net · Inference-policy selection · Reproducibility · BraTS.

1 Introduction

The BraTS benchmark standardized evaluation of automated glioma segmentation from multiparametric MRI [14,3,2] and has expanded into subchallenges covering meningioma [12], brain metastases [15], pediatric tumors [7] and a Sub-Saharan Africa population [1], scored by containerized infrastructure [10]. BraTS-GoAT poses the cross-tumor question directly: one model must segment

* Corresponding author.

enhancing tumor (ET), tumor core (TC) and whole tumor (WT) across heterogeneous tumor types and sources, with no per-cohort tuning and no cohort label at inference.

The self-configuring nnU-Net family remains hard to beat on BraTS-style tasks [8,9], so we did not pursue a new architecture. What interested us is that a challenge submission is a *container*, and that its inference policy (checkpoint, test-time augmentation, tiling overlap, threshold, post-processing, ensembling) is routinely chosen on the same data used to report cross-validated performance, with no holdout and no advance commitment. Underspecification makes this fragile: pipelines equivalent on a development distribution can differ elsewhere [6].

Research question. We ask: *for a fixed containerized segmentation system, can a commit-frozen, multi-gate policy audit prevent adoption of an inference modification whose aggregate development-subset benefit does not transfer consistently to a same-corpus policy-selection holdout?* We answer it as a bounded case study on one system, corpus and split, and do not claim it shows the procedure is generally superior to nested cross-validation or an external cohort.

Transferable conclusions. Three claims survive our evidence. *Release-path equivalence must be verified before inference policies are compared*—ours was not, which silently offsets every internal figure. *A small aggregate gain can coexist with fold heterogeneity and worse lesion-miss behaviour*, so an aggregate mean is not sufficient evidence to adopt a change. And *a same-corpus policy holdout reduces direct post-result tuning*, but it is not an external independent cohort and does not replace nested validation.

2 Methods

2.1 Data and Task

We use only data supplied through the 2026 BraTS-GoAT subchallenge, which draws on the BraTS data lineage [14,3,2], its expert-annotated public collections [4,5] and its cohort-specific releases [12,15,7,1]. The training release comprises 1,351 labeled cases, which our audit found complete, geometry-valid and integer-labeled; evaluated regions are the standard nested triple ET = [3], TC = [1, 3], WT = [1, 2, 3]. Validation and test ground truth is withheld. Per-case cohort labels were not reliably derivable and we did not guess them, so cohort-stratified analysis lies outside this work—a real limitation for a task about cohort transfer.

2.2 Split, Preprocessing and Training

We build a deterministic group-level five-fold split (seed 21072026; fold sizes 271, 270, 270, 270, 270), frozen before training and never changed. A duplicate

analysis found no confirmed same-subject pairs, so cases are singleton groups; we claim only that the audit found no confirmed duplicate group, not that leakage is impossible. Every fold contains ET-absent and TC-absent cases, exercising the empty-region paths the evaluator must handle.

Preprocessing is the standard nnU-Net v2.8.1 pipeline over the four modalities (T1n, T1c, T2w, T2f) in fixed channel order. The backbone is the ResEnc-M residual-encoder U-Net in `3d_fullres` with region-based ET/TC/WT targets, patch size $[128, 160, 112]$, batch size 2, SGD with Nesterov momentum, weight decay 3×10^{-5}, initial learning rate 10^{-2} with polynomial decay, 0.33 foreground oversampling and deep supervision, trained 1000 epochs per fold with the final checkpoint retained; the full plan is in the public repository. All parameters begin from a recorded random initialization; no pretrained or self-supervised weights are used. Folds train independently, and all five are complete.

2.3 Architecture Selection and Its Metric

Selection between the ResEnc-M and ResEnc-L presets followed a policy fixed before any fold result was seen, on fold 0 only, using DSC and HD95. Its statistic differs from the audit utility. For each of the six screened components the presets are ranked against each other, better value first (higher DSC, lower HD95), exact ties taking the average of the tied positions; R is the mean of those six fractional ranks, so *lower R is better*. We report $\Delta R = R(\mathrm{L}) - R(\mathrm{M})$, negative favouring ResEnc-L, with a paired subject-level bootstrap (seed 21072026, 10,000 resamples) summarized by a 95% percentile interval. ResEnc-L could advance from fold 0 to fold-1 confirmation only if the executable rank-gain threshold on $R(\mathrm{M}) - R(\mathrm{L})$, nominally 1/6, was met, the interval for ΔR lay entirely below zero, and every frozen auxiliary gate was supplied and passed; expansion required the same on fold 1, and otherwise ResEnc-M was retained. R is a rank statistic, a different object from U_τ, a mean of raw metric values; the two are never combined, and neither is the challenge's ranking procedure, which uses the Dice similarity coefficient (DSC) and normalized surface distance (NSD) and excludes HD95, so screen and ranking metric do not agree; Section 3.2 rescores the same predictions and Section 4 records the mismatch as a limitation.

2.4 Release Inference Path

Final inference is a sequential five-checkpoint mean-probability ensemble: one fold model is resident at a time and freed before the next, so simultaneous model residency does not scale with the five folds. That bounds model residency only; the running probability accumulator and the current per-region probabilities are also held, so process peak is not that of a bare single-model run. Tiles are extracted with a sliding window at step 0.5 of the patch size and combined with Gaussian importance weighting. Region probabilities are thresholded at a fixed 0.5—no learned threshold or ensemble weight, no test-time augmentation, no component filtering, no presence gate—and reconstructed by a hierarchy-safe projection enforcing WT $\supseteq$ TC $\supseteq$ ET at the exact source geometry. The

runner is near-deterministic but not bit-exact, because deterministic-algorithm enforcement runs in warn-only mode; the supplement quantifies the residual variation.

2.5 Decision Procedure for Policy Audits

Audits were staged (Fig. 1): each compared one frozen candidate with the release baseline under prespecified rules, and development failure stopped advancement.

Which tolerance is authoritative. Organizer-confirmed final ranking uses DSC and NSD at $\tau = 1$, **the authoritative analysis**, against which we report every decision; Panoptica's $\tau = 0.5$ default remains a prespecified sensitivity analysis only.

Utility. Let S_τ be the subjects for which all six DSC/NSD components are finite under *both* policies, and $\bar{m}_{c,\tau}(P; S_\tau)$ the mean of component c over S_τ. The audit utility is the unweighted arithmetic mean of those six *raw* component means,

$$U_\tau(P; S_\tau) = \frac{1}{6} \sum_c \bar{m}_{c,\tau}(P; S_\tau), \qquad \Delta U_\tau = U_\tau(P_{\text{cand}}; S_\tau) - U_\tau(P_{\text{base}}; S_\tau), \quad (1)$$

with c over ET/TC/WT $\times$ DSC/NSD$_\tau$. Since both policies share S_τ, ΔU_τ is the equally weighted mean of six component deltas; unlike R (Section 2.3) it contains no ranking, tie-breaking or HD95.

Common support and fail-closed evaluation. A subject contributes only if all six components are finite under both policies, so a change in scorable cases cannot itself create a gain; we report $|S_\tau|$ beside every ΔU_τ. Empty reference regions follow the evaluator's convention and non-finite entries are excluded, not imputed. Common support is a comparability, not an error-tolerance, device: a candidate whose expected evaluation is missing, errored or membership-mismatched is ineligible outright and the audit fails closed. Three of the twenty-three holdout checks enforce this, and all three passed.

Uncertainty and safety. Uncertainty is a paired *subject-level* bootstrap—subjects resampled, never region-cases—with seed 20260730, 10,000 resamples, a percentile interval and the fraction of resamples with $\Delta U_\tau > 0$; that fraction is the share in which the candidate is ahead, not a posterior probability. Ties resolve to the baseline. Safety uses 26-connected lesion components in the spirit of instance-wise evaluation [11], metric choice materially affecting such conclusions [13]: a reference component counts as missed if no predicted component of the same region overlaps it. The frozen margins were 0.0025 on the point change in miss rate, 0.0050 on its one-sided 95% upper bound and per region, and a 5% cap on the relative increase in predicted false-positive components. These are *operational* margins fixed in the protocol, which records no clinical derivation; they are not externally validated clinical thresholds.

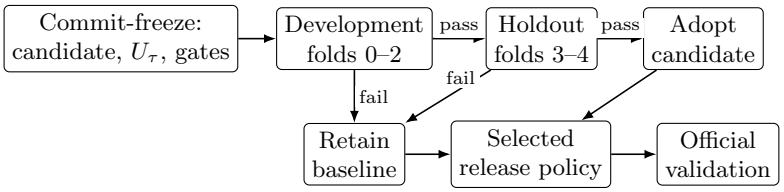

Fig. 1. The staged decision procedure in general form: a candidate failing the development subset stops there, one that passes reaches the policy-selection holdout and is adopted or rejected, and either branch yields one *selected release policy*, which alone is run on official validation. This is the procedure, not the outcome. Here no candidate advanced under its own gate—all three audits stopped on the development subset, Audit C on a frozen lesion false-negative veto—and the folds-3–4 evidence comes from a separate, development-informed follow-up of that same frozen M8 candidate, commit-frozen after Audit C stopped and before those folds were opened, not advancement under Audit C. C0 remained the selected policy.

2.6 Container Contract

The submission is a `linux/amd64` zero-network container for the official Task-3 budget (NVIDIA A10G, 24 GB VRAM, 12 h), with all dependencies and five fold checkpoints baked in. **Each output file name is the complete validated basename of its input case folder**, with no truncation to a fixed-width identifier and no hardcoded cohort prefix; an earlier revision assuming a fixed-width suffix was corrected, and the preflight tests multiple prefixes and variable-length basenames. Labels are restricted to $\{0, 1, 2, 3\}$, geometry is preserved exactly, and modality discovery fails closed before any output is written. We make no claim that the runner refuses to overwrite. The supplement carries the remaining operational detail. The source is public under the Apache License 2.0 at `https://github.com/Neethan-hub/gat26-brats-2026/tree/brats-goat-2026-camera-ready-21-r11`; it contains no images, labels, checkpoints or predictions.

3 Results

3.1 Descriptive Out-of-Fold Performance

Under the DSC/HD95 screen ResEnc-M won four of six fold-0 components, the paired bootstrap did not favour ResEnc-L, and ResEnc-L failed both declared tail gates, so the policy selected ResEnc-M; the supplement gives all six components.

In a strictly out-of-fold set—each case predicted once by the fold model that did not train on it—the union covers 1,351 cases, each exactly once, with no missing case, duplicate or fold overlap. Scored in one pass by the official evaluator with zero errors and all six denominators equal to 1,351, it gives DSC 0.865 (ET), 0.915 (TC), 0.927 (WT) and HD95 13.39, 6.75, 6.34 mm. This is *descriptive*

post-selection out-of-fold performance, conditional on the chosen model family and split—neither an unbiased generalization estimate nor in-sample accuracy: fold 0 chose the architecture family and then appears inside this summary, and the corpus was reused across audits. It is not the released ensemble's score, since every training case trained four of the five models and it came from a different inference path.

3.2 Rescoring the Architecture Comparison Under the Ranking Metric

Because the architecture screen used DSC/HD95, we rescored the same fold-0 predictions under DSC/NSD at $\tau = 0.5$ as a diagnostic. **The ordering reverses at the point estimate**: ResEnc-L is better on four of six official components, ResEnc-M on TC and WT DSC. On the same statistic $\Delta R = R(\mathrm{L}) - R(\mathrm{M}) = -0.333$, nominally favouring ResEnc-L, but its paired-bootstrap 95% interval $[-1.000, +0.667]$ includes zero, so no rank advantage is established. Fold-0 ResEnc-L weights already existed—they are what the diagnostic scores—so no five-fold expansion, architecture change or retraining campaign followed. The diagnostic used existing fold-0 end-of-training predictions at $\tau = 0.5$, not release-path predictions; the supplement gives all six components. The ordering therefore reversed without a robust advantage, and deployed ResEnc-M had been selected under a metric pair the challenge does not rank on.

3.3 Two Inference Paths

Every internal figure above came from the official evaluator applied to nnU-Net's own end-of-training validation predictions, whose recorded settings include mirroring axes $(0, 1, 2)$: eightfold mirroring test-time augmentation. The released container disables mirroring entirely. We found this by trying to reproduce an internal development baseline through the release path and failing. On the same 811 development subjects, with the same evaluator and denominators, the augmented lineage reproduces the recorded internal values to a worst absolute error of 4.3×10^{-7}, whereas the release path is lower by roughly 0.002 on each DSC component and 0.002–0.007 on each NSD component. The gap is single-signed across all twelve component/tolerance combinations, reproduces fold by fold, and is orders of magnitude larger than the run-to-run variation we measured; it is consistent with the mirroring-path difference and not with nondeterminism. No internal number here therefore estimates what the released container produces.

3.4 Checkpoint and Weight-Averaging Audits

Audit A tested the best-validation checkpoint, mirroring over the recorded axes, both combined, and ET component cleanup at three declared volume thresholds; no candidate improved robustly. **Audit B** tested two declared per-fold weight averages of the final and best-validation checkpoints [16] and is the sharpest

negative: all six DSC/NSD components moved the wrong way for both soups, though not monotonically—the 0.50 soup beat the 0.25 one. Both stopped on the development subset; per-component values are in the supplement.

3.5 Mirroring: A Development Gain That Did Not Transfer Consistently

Audit C tested one candidate, M8, differing from baseline C0 only in mirroring over axes exactly $(0, 1, 2)$, executed end-to-end by the release machinery on raw NIfTI inputs. One scoping point is essential: each subject is predicted by *its one excluded-fold checkpoint*, whereas the deployed submission ensembles *all five*, so the audit tests a closely related inference-policy effect and does not directly measure mirroring on the deployed five-model ensemble.

On the development subset M8 improved all six DSC/NSD components at both tolerances and seventeen of eighteen frozen checks passed. The one failure was a strict veto on any increase in lesion false-negative *count*, which rose by 13 (+0.3%); under that frozen gate Audit C *stopped* and the baseline was retained. Because it stopped, the candidate-side phases of that protocol—runtime qualification of M8, and any official submission of M8—were conditional on advancement and were never reached. They are unreached branches of a frozen protocol, not omitted steps, and they are unrelated to the separately documented official-validation submission of the retained baseline C0. Inspecting the veto exposed a limitation rather than an error: the number of reference lesion components is *policy-invariant*, but an absolute count neither normalizes by that fixed number of opportunities nor carries an interval. We therefore fixed, in a pushed commit, a separate reference-lesion miss-rate noninferiority analysis together with a follow-up examination of the same frozen M8 candidate on the untouched folds 3–4. That follow-up is *development-informed*: it was designed after seeing Audit C's development outcome and commit-frozen before those folds were opened, so it is not globally result-blind and we say so plainly. It is a separate subsequent analysis, not continuation or advancement under Audit C's failed gate, and it does not annul the earlier audit, which correctly applied its own frozen rule. Its candidate-side phases were likewise conditional and were not reached.

Holdout outcome. These holdout values come from that separate follow-up, and any comparison against the development values is a descriptive cross-analysis rather than evidence that one globally result-blind protocol carried through. At $\tau = 1$ the follow-up gain was $+0.00073$ against the development $+0.00253$, below the required $+0.0010$; the positive-resample fraction was 0.734 against 1.000 and the 95% interval $[-0.0025, +0.0026]$ straddles zero. The $\tau = 0.5$ sensitivity analysis behaves the same way: $+0.00247$ against $+0.00417$, fraction 0.961, interval $[-0.0003, +0.0042]$, also including zero. Descriptively the holdout retained 28.7% of the development gain at $\tau = 1$ and 59.3% at $\tau = 0.5$; neither is an effect size. The two holdout folds gave $+0.0023$ and -0.0009—a sign change—zero-DSC region cases rose from 24 to 26, and the reference-lesion miss rate regressed by $+0.0070$ against the 0.0025 margin, with a one-sided 95% upper bound of

+0.0124 against the 0.0050 margin and all three regions over their own limits. Eight of 23 checks failed, including the primary utility and safety gates. No denominator and no non-finite pattern changed, so neither the residual gain nor the failure is a denominator artifact, and the released no-augmentation policy was retained.

What failed, qualitatively. We publish no patient-level examples. In aggregate, mirroring consistently *reduced* predicted false-positive lesion components (development 1243 → 1087; holdout 864 → 781) while consistently *increasing* missed reference components (4103 → 4116; 2726 → 2764). That is consistent with a *possible* precision–sensitivity trade-off, but it remains only a hypothesis: we did not stratify by lesion size and did not test the mechanism. What is established is the direction of the counts, which the overlap means—all improved—do not reveal.

Bounded inventory. Our search was bounded: a connected-component proposal, a 40-epoch augmentation and tumor-type-mixture screen whose bound cannot speak to convergence, and an untested dense-overlap variant (D25) were considered and none adopted; larger ensembles, cascades, cohort routing and alternative thresholds were not tested. The supplement lists each with its outcome.

3.6 Official Validation Result

The frozen released ensemble was run over the complete official Task-3 validation cohort and submitted once; Table 1 gives the returned values, rounded. *No scientific element of the inference policy*—model, checkpoint, threshold, post-processing or ensemble rule—changed at any point, before or after the score was seen. One operational change followed, and it was not a response to the score: the output-name parser was corrected after the organizers' execution of an earlier image failed on a folder-name assumption (Section 2.6).

Table 1. Organizer-reported official validation values for the frozen released ensemble, from a single submission, rounded to four decimals. DSC and NSD are the ranked metrics, HD95 is diagnostic, n is the finite per-case denominator. The organizers confirmed that the *final ranking* uses NSD at $\tau = 1$, but did not disclose which tolerance produced these particular participant-visible values, so no tolerance is attached to them here.

Region	DSC	NSD	HD95 (mm)	n
ET	0.7722	0.5398	44.82	440
TC	0.8235	0.4999	23.64	449
WT	0.8794	0.4834	17.56	450

Four qualifications travel with this table. *(i)* No rank was exposed, so we record none. *(ii)* The submitted archive contained all 451 predictions and its

member set equalled the frozen expected set, whereas the participant-visible per-case file contains 450 scored rows; the cause was not disclosed, so we attribute the discrepancy to neither party. *(iii)* The finite denominators are 440/449/450, and the organizers' mean summary row equals the posted values exactly. *(iv)* We do not claim to reproduce the official scorer: its version, the units of τ and parts of the non-finite handling are not exposed.

These values fall substantially below our internal cross-validation figures. One component of the difference is identified: the inference-path difference of Section 3.3, worth roughly 0.002 DSC. The remainder cannot be assigned to a specific mechanism with the evidence available. Per-case cohort labels were not derivable, and there are no subgroup results and no controlled comparison, so domain shift, case composition, ET heterogeneity, adaptive reuse and scorer differences all remain consistent with the observation and cannot be separated. The gap is reported as a measurement, without attribution to any single cause.

3.7 Runtime Evidence

Our full-cohort evidence is the run of the exact release policy that produced Table 1: 451 of 451 cases, exit code 0, zero inference errors, peak framework-reserved VRAM 2.48 GiB, wall time 3 h 35 m (28.6 s per case) on a single NVIDIA A40. Both figures describe that A40 run only; the 2.48 GiB is framework-reserved memory, not total device memory, and is neither an A10G measurement nor an A10G headroom claim. An earlier, superseded revision was exercised on one A10G using synthetic volumes. That test does not qualify the corrected image: its fixtures used fixed-width folder names and so never exercised the variable-length condition the correction addresses. We hold *no* A10G measurement of the corrected image, no organizer execution log for that image, and no hidden-test performance result or rank for it. Eightfold mirroring multiplies forward passes per tile by eight; we measured neither its full-ensemble runtime nor memory profile on the target platform and make no feasibility claim in either direction.

4 Discussion and Limitations

Our quantities are not interchangeable: out-of-fold cross-validation is descriptive and post-selection, audit results are policy-selection comparisons rather than performance estimates, the single official measurement carries no rank, and hidden-test performance, which alone determines the ranking, does not exist at the time of writing, so we report and estimate none. The procedure did what it was built to do—a candidate improving all six DSC/NSD components on the development subset, every $\tau = 1$ resample positive, was stopped before release by gates fixed in advance—but this is one demonstration on one system. We have not shown that the gates generalize, that the margins suit another model or corpus, or that the procedure outperforms nested cross-validation, which we did not run.

Limitations. The policy holdout is same-corpus: folds 3–4 carry prior development history, so they bound direct post-result tuning rather than adaptive reuse of the corpus, and multiplicity across audits is uncontrolled; we hold no external independent cohort. The architecture–metric mismatch and its diagnostic reversal are as reported in Section 3.2. The mirroring audit is a single-model out-of-fold comparison (Section 3.5) and does not determine how mirroring would affect the deployed five-checkpoint ensemble, which could mitigate, preserve or amplify the lesion-miss behaviour; we neither adopted it nor claim it would have helped. The absent per-case row, the non-derivability of per-case cohort labels, and the non-attribution of the official gap are as stated in Section 3.6; the absence of any hidden-test result or rank, organizer execution log or A10G measurement for the corrected image is as stated in Section 3.7. Group-level splitting cannot prove the absence of same-subject leakage, a small minority of cases dominates the boundary-error mean, and the development-subset bootstrap exists only at $\tau = 1$. GAT-26 trains only on GoAT-provided data from a recorded random initialization, with no external, prior-BraTS, private or pretrained data or weights; validation and test labels were never used, and we performed no pseudo-labeling, transductive adaptation or leaderboard-driven tuning.

5 Conclusion

GAT-26 is a five-fold region-based residual-encoder nnU-Net ensemble whose release policy was selected and released following commit-frozen rules. All three audits stopped on the development subset; for the mirroring candidate a separate, development-informed follow-up then examined the untouched same-corpus policy-selection holdout, where the gain was descriptively smaller at both tolerances, both intervals spanned zero, one fold changed sign and a lesion miss-rate gate failed. The simpler baseline was retained. Two observations travel further than the architecture: internal cross-validation numbers and the shipped container came from different inference paths, and a development-subset gain that looked unambiguous was only partially reproduced on the holdout—28.7% at the official-ranking-aligned $\tau = 1$—while lesion-miss behaviour worsened. Official validation fell well below our internal figures, a gap we report without attributing. This is a bounded case study.

Acknowledgments. The author thanks Professor Pin-Yuan Chen for his clinical guidance, methodological feedback, review of the manuscript, and mentorship. Data used in this publication were obtained as part of the Challenge project through Synapse ID (syn74274097).

Data use and ethics. This work analyses only de-identified imaging data distributed by the BraTS-GoAT 2026 subchallenge, under that challenge's data-use agreement. No participants were recruited, no new data were collected, and no re-identification was attempted. No patient images, labels, per-case scores or case identifiers appear in this paper, the supplement or the repository.

Disclosure of Interests. The author declares no competing interests.

References

1. Adewole, M., Rudie, J.D., Gbadamosi, A., Toyobo, O., Raymond, C., Zhang, D., Omidiji, O., Akinola, R., et al.: The brain tumor segmentation (BraTS) challenge 2023: Glioma segmentation in Sub-Saharan Africa patient population (BraTS-Africa). arXiv preprint arXiv:2305.19369 (2023). https://doi.org/10.48550/arXiv.2305.19369, arXiv:2305.19369
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021). https://doi.org/10.48550/arXiv.2107.02314, arXiv:2107.02314
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., et al.: Advancing the Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**, 170117 (2017). https://doi.org/10.1038/sdata.2017.117
4. Bakas, S., Akbari, H., Sotiras, A., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
5. Bakas, S., Akbari, H., Sotiras, A., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF
6. D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., et al.: Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research **23**(226), 1–61 (2022), arXiv:2011.03395
7. Fathi Kazerooni, A., Khalili, N., Liu, X., Haldar, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., et al.: The brain tumor segmentation (BraTS) challenge 2023: Focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). arXiv preprint arXiv:2305.17033 (2023). https://doi.org/10.48550/arXiv.2305.17033, arXiv:2305.17033
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 (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. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15009, pp. 488–498. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72114-4_47
10. Karargyris, A., Umeton, R., Sheller, M.J., Aristizabal, A., George, J., Wuest, A., Pati, S., et al.: Federated benchmarking of medical artificial intelligence with MedPerf. Nature Machine Intelligence **5**, 799–810 (2023). https://doi.org/10.1038/s42256-023-00652-2
11. Kofler, F., Möller, H., Buchner, J.A., de la Rosa, E., Ezhov, I., Rosier, M., et al.: Panoptica – instance-wise evaluation of 3d semantic and instance segmentation maps. arXiv preprint arXiv:2312.02608 (2023). https://doi.org/10.48550/arXiv.2312.02608, arXiv:2312.02608
12. LaBella, D., Adewole, M., Alonso-Basanta, M., Altes, T., Anwar, S.M., Baid, U., Bergquist, T., Bhalerao, R., et al.: The ASNR-MICCAI brain tumor segmentation (BraTS) challenge 2023: Intracranial meningioma. arXiv preprint arXiv:2305.07642 (2023). https://doi.org/10.48550/arXiv.2305.07642, arXiv:2305.07642

13. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., et al.: Metrics reloaded: Recommendations for image analysis validation. Nature Methods **21**, 195–212 (2024). https://doi.org/10.1038/s41592-023-02151-z
14. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
15. Moawad, A.W., Janas, A., Baid, U., Ramakrishnan, D., Saluja, R., Ashraf, N., Maleki, N., Jekel, L., et al.: The brain tumor segmentation (BraTS-METS) challenge 2023: Brain metastasis segmentation on pre-treatment MRI. arXiv preprint arXiv:2306.00838 (2023). https://doi.org/10.48550/arXiv.2306.00838, arXiv:2306.00838
16. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., et al.: Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 23965–23998. PMLR (2022)