Research Square

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

Brain Tumor Segmentation in the Foundation-Model

Behnam Kiani Kalejahi

kiani-bme@hotmail.com

Central Asian University https://orcid.org/0000-0002-7118-0382

Systematic Review

Keywords: brain tumor segmentation, multimodal MRI, missing-modality robustness, foundation models, uncertainty quantification, external validation

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11205837/v1

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

Additional Declarations: The authors declare no competing interests.

Brain Tumor Segmentation in the Foundation-Model

Architectures, Multimodal MRI, Missing-Modality Robustness, Uncertainty, Domain Shift, and Clinically Trustworthy Evaluation (2015–2026)

A structured critical review, quantitative evidence atlas, and DATA-SEG reporting framework

Behnam Kiani Kalejahi[1, 2]*, Sajid Khan[1], Mohammad Javad Rajabi[2]

[1] Department of Computer Science, School of Engineering, Central Asian University, Tashkent, Uzbekistan

[2] Faculty of Data Science and Information Technology, INTI International University, Putra Nilai, Malaysia

* Corresponding author: Behnam Kiani Kalejahi, b.kiani@centralasian.uz, ORCID 0000-0002-7118-0382

Abstract

Automated brain tumor segmentation has progressed from handcrafted pipelines and U-Net variants to self-configuring 3D CNNs, transformer hybrids, state-space models, diffusion-based segmenters, promptable models and brain-specific foundation models. Yet the field's central scientific problem has changed: another fractional gain in Dice on a familiar BraTS split provides limited evidence that a system will remain reliable when MRI sequences are missing, scanners and institutions change, lesions are small or multifocal, postoperative anatomy is distorted, or uncertainty is high. This structured critical review synthesizes 150 references with emphasis on 2025–2026 and compares architectures, multimodal fusion strategies, preprocessing, missing-modality learning, self-supervision, foundation-model adaptation, federated learning, domain adaptation, uncertainty quantification and clinical translation. We organize evidence around adult glioma, pediatric tumors, brain metastases, meningioma and longitudinal/postoperative applications; distinguish region overlap from lesion detection and volumetric clinical endpoints; and critically examine leakage, site confounding, hidden test reuse, pretraining contamination, annotation uncertainty and challenge-overfitting. Recent evidence is especially informative: brain MRI foundation models such as BrainIAC and BrainFound improve label efficiency and cross-task transfer, while task-specific systems such as nnU-Net and carefully engineered 2.5D/3D networks remain formidable full-data baselines. Missing-modality studies increasingly favor adaptive fusion, distillation and modality dropout over naive zero-filling; uncertainty work is moving toward conformal and evidential approaches that can identify cases requiring expert review; and external studies of metastases and meningiomas demonstrate why lesion-wise, site-stratified and human-in-the-loop evaluation is essential. We propose DATA-SEG, Dataset, Annotation, Train/test integrity, Architecture, Statistics, External robustness and Governance, as a reporting and evaluation framework. The review concludes that the next high-impact phase of brain tumor segmentation should prioritize post-cutoff hidden benchmarks, cross-site robustness, missing-sequence stress tests, calibrated selective segmentation, multi-rater probabilistic ground truth and longitudinal clinical readouts rather than leaderboard optimization alone.

Keywords: brain tumor segmentation; multimodal MRI; missing-modality robustness; foundation models; uncertainty quantification; external validation

1. Introduction: why the field must move beyond leaderboard Dice

Brain tumor segmentation is one of the most mature test beds for medical image analysis. The combination of clinically meaningful subregions, multi-parametric MRI, public challenge infrastructure and expert annotations has made the Brain Tumor Segmentation (BraTS) ecosystem uniquely influential. From the 2015 benchmark onward, methodological progress has been unusually rapid: early multi-scale CNNs were followed by 3D U-Nets and V-Nets, autoencoder-regularized networks, self-configuring pipelines such as nnU-Net, convolution–transformer hybrids, diffusion and state-space models, and now promptable and foundation-model approaches [1–14,19–25].

The very success of BraTS creates a methodological paradox. Standardized datasets make architectures comparable, but repeated optimization on familiar distributions can gradually turn the benchmark into a development environment rather than an independent estimate of clinical performance. A model can achieve an excellent average Dice score while remaining brittle to scanner shift, missing MRI sequences, postoperative cavities, tiny enhancing foci, multifocal metastases or annotation ambiguity. Likewise, a foundation model may have seen related cases or derivative releases during pretraining even when the downstream test set is nominally held out. The resulting distinction between benchmark discrimination and transportable clinical reliability is now more important than the historical distinction between CNN and transformer families.

This review therefore treats segmentation as an end-to-end evidence problem. Architecture matters, but so do the source of images, independence of patients, longitudinal grouping, preprocessing leakage, annotation provenance, external-site testing, lesion-wise sensitivity, calibration, uncertainty, compute requirements and downstream clinical use. We ask not only which method reports the highest overlap, but under what data regime, modality availability, site shift and clinical workflow that result remains meaningful.

A second motivation is the expansion of neuro-oncology beyond a single adult glioma benchmark. Modern challenges and clinical studies include pediatric gliomas, brain metastases, meningiomas, radiotherapy planning and postoperative residual disease. These tasks have different lesion sizes, image contrasts, label ontologies and failure costs. An optimal network which maximizes the overlap of the whole tumor may perform sub-optimally for the clinical question of whether a 3 mm metastasis is present or an enhanced remnant due to operation altered by several milliliters exists. The third reason stems from the transition to reliability-centered approaches in 2025-2026. Missing modality models explicitly account for missing MRI modalities [37-50], diffusion and state space models try to

understand the larger context [51-59], uncertainty quantification methods use conformal prediction and evidential ensembling to detect unreliability [66-74] and brain MRI foundation models focus on transfer learning and label efficiency [19-25]. These directions are scientifically promising, but they also introduce new evaluation requirements. The more reusable the pretrained representation, the more important it becomes to audit pretraining overlap and to distinguish adaptation gains from hidden familiarity with the target benchmark.

Our central thesis is that the next widely cited contribution in brain tumor segmentation is unlikely to be a narrowly modified encoder. Greater scientific value lies in a reproducible benchmark or framework that can reveal when strong models fail, quantify the value of additional sequences, compare foundation models against nnU-Net under matched compute and annotation budgets, and connect segmentation uncertainty to human review. DATA-SEG is proposed as one such organizing framework.

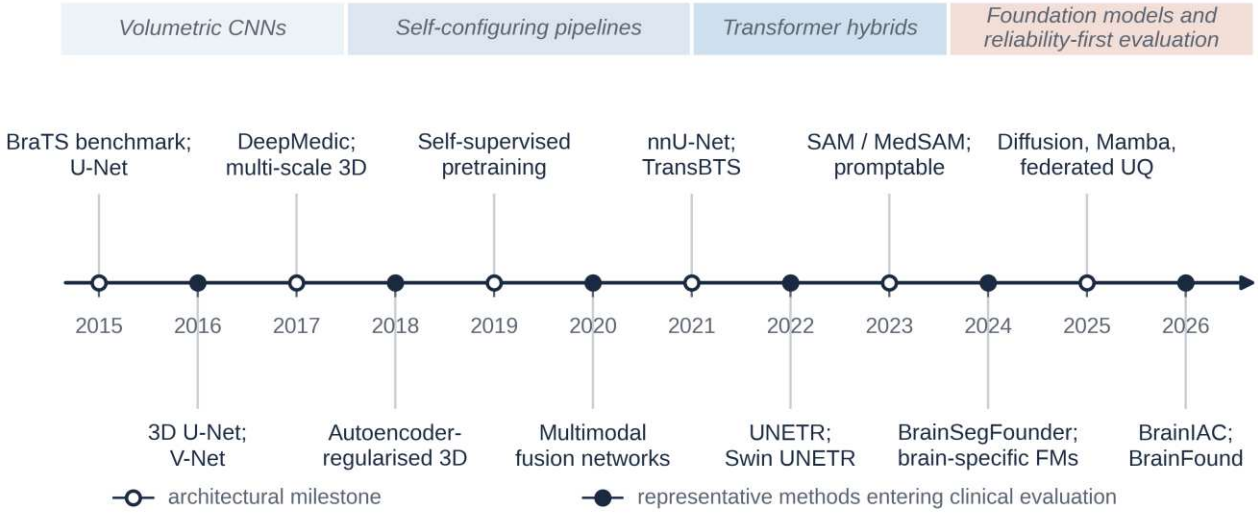


Figure 1. Major methodological transitions in brain tumor segmentation, 2015–2026. The timeline is conceptual and emphasizes changes in evaluation priorities as well as architectures.

2. Review methodology and evidence-verification strategy

The literature search was structured around four evidence streams: (i) brain tumor segmentation methods and BraTS challenge reports; (ii) multimodal MRI, missing-modality, synthesis and domain-shift methods; (iii) foundation, self-supervised and promptable segmentation models; and (iv) clinical translation, uncertainty, annotation variability, metrics and reporting guidance. PubMed and publisher records were prioritized for journal articles; IEEE Xplore/CVF or proceedings records were used for conference papers; official challenge and regulatory sources were used when they were the primary record. Preprints were

retained only where no peer-reviewed version was identified or where the preprint itself is the canonical method description.

Search concepts included combinations of brain tumor, glioma, metastasis, meningioma, MRI, segmentation, BraTS, multimodal, missing modality, incomplete modality, diffusion, Mamba, state-space, transformer, nnU-Net, SAM, MedSAM, foundation model, self-supervised, federated, domain adaptation, uncertainty, conformal prediction, calibration, external validation, postoperative and longitudinal. The evidence cut-off for this manuscript was 19 September 2026. The search was intentionally weighted toward 2025–2026 because the foundation-model and missing-modality literature is changing rapidly, while landmark older methods were retained to preserve methodological continuity.

The review is not labelled a completed PRISMA systematic review. A formal PRISMA claim would require frozen multi-database exports, documented duplicate removal, dual-reviewer title/abstract and full-text screening, predefined exclusion reasons, risk-of-bias assessment and a reproducible study flow. Instead, this manuscript is a structured critical review with a transparent reference-verification ledger. This distinction avoids false precision while still providing a reproducible evidence map.

For quantitative comparisons, results are quoted only when the source reports a sufficiently defined dataset/task. Cross-paper ranking is avoided when data editions, preprocessing, label definitions or evaluation servers differ. We distinguish three evidence levels: internal benchmark validation, multi-dataset or challenge-edition validation, and genuinely external site/scanner/time validation. The latter is treated as the strongest evidence for transportability.

The authors' own peer-reviewed neuroimaging publications are included only where methodologically relevant: 3D GAN and auxiliary-classifier GAN segmentation studies, the 2026 GDNet 2.5D multimodal segmentation framework, and the paired-versus-unpaired MRI synthesis comparison [15–18]. They are not used as privileged evidence; they are situated alongside competing methods and the same limitations are applied.

Table 1. Evidence axes extracted from each included study, and the reason each axis matters for transportability.

Evidence axis	What this review extracts	Why it matters
Dataset	Edition, institution count, modality set, tumor type, longitudinal status	Determines population and domain represented.
Split integrity	Patient/episode grouping, duplicate handling, hidden test status	Prevents leakage and pseudo-replication.
Architecture	2D/2.5D/3D, CNN/Transformer/Mamba/diffusion/FM, parameterization	Separates representation choice from evaluation design.
Training	Loss, sampling, augmentation, pretraining, seeds, checkpoint selection	Reveals sources of performance inflation or instability.
Metrics	WT/TC/ET Dice, HD95, lesion sensitivity, volume error, calibration	Aligns technical output with clinical failure modes.

Evidence axis	What this review extracts	Why it matters
External robustness	Site/scanner/geography/time, missing modality, corruption	Tests transportability.
Clinical readout	Expert correction, RANO/volumetry, time saving, longitudinal consistency	Tests whether the mask changes care or workflow.

3. Clinical and imaging foundations

Brain tumors are biologically and radiologically heterogeneous. Adult diffuse gliomas are now classified using integrated molecular criteria, while pediatric tumors have distinct biology and treatment pathways. Brain metastases are frequently multifocal and small; meningiomas are usually extra-axial and often evaluated on contrast-enhanced T1-weighted MRI; postoperative glioblastoma introduces blood products, resection cavities and treatment-related enhancement. There is no universal benchmark that equally represents the distributions [96-109]. BraTS inputs generally include T1-weighted, T1-weighted post contrast (T1ce), T2-weighted, and FLAIR MR scans. The T1ce scan gives special information about the enhancement of tumors, FLAIR highlights edema and infiltration, and T2/T1 images provide additional structural information. Labels in BraTS are usually categorized into enhancing tumor (ET), tumor core (TC), and whole tumor (WT) classes, which have nested anatomical structure. But there is no one-to-one mapping from historic challenge labels to current clinical entities, and depending on the post operative or therapy response setting, labels need to have different semantics.

Segmentation performance is clinically relevant since volumes and shapes help in treatment planning, radiotherapy contouring, surgical analysis, and follow-up assessment. However, clinical goal does not necessarily coincide with the Dice coefficient. Segmentation can give good overlap results but miss a small enhancing focus; whereas having a slight boundary mismatch in a big lesion could give low surface metrics but will not change the treatment plan. RANO 2.0 and RANO-BM reinforce the importance of serial measurements, lesion burden and standardized response assessment [106,107].

For that reason, a mature evaluation should report both region-level and lesion-level outcomes. Region Dice and HD95 remain useful, but metastasis detection sensitivity, false-positive lesions per case, volume bias, small-lesion strata and longitudinal reproducibility can be more clinically informative. In postoperative disease, the relation between the automated residual volume and expert extent-of-resection assessment may matter more than a global overlap average.

4. BraTS and the evolution of benchmark tasks

Table 2. BraTS benchmark families, their representative scope and the methodological lesson each contributes.

Benchmark family	Representative scope	Core modalities/labels	Methodological lesson
BraTS adult glioma	Preoperative diffuse glioma; multiple annual editions	T1, T1ce, T2, FLAIR; ET/TC/WT	Excellent for standardized comparison; repeated use raises challenge-overfitting concerns.
BraTS-PEDs	Pediatric brain tumors	Multimodal MRI; pediatric tumor labels	Tests age/tumor-spectrum shift; adult-trained models may not transfer.
BraTS-METS	Brain metastases	Contrast MRI with lesion-wise relevance	Requires lesion detection metrics, especially for small lesions.
BraTS-MEN	Meningioma	Primarily contrast-enhanced structural MRI	Extra-axial tumors and volumetry have different failure modes.
BraTS-Africa	Sub-Saharan African glioma population	BraTS-like MRI with geographic/site shift	Directly exposes geographic and acquisition-domain generalization.
BraSyn	Missing-sequence synthesis	Complete/incomplete MRI sequences	Tests whether synthesis preserves pathology, not just image similarity.
BraTS-MEN-RT	Meningioma radiotherapy	Planning-oriented imaging/contours	Connects segmentation to radiotherapy workflow.

BraTS has done more than standardize a dataset; it standardized a way of thinking about multimodal tumor segmentation. Common preprocessing, label definitions and evaluation infrastructure enabled meaningful progress and helped reveal the value of 3D context and multimodal fusion. However, challenge performance should not be interpreted as a universal estimate of clinical accuracy. Annual editions differ in cohort composition and hidden test sets, and some research teams iteratively optimize pipelines across related releases.

The individual branch reports document these shifts directly: pediatric tumors [90], brain metastases [92], meningioma [91], missing-sequence synthesis [93] and the Sub-Saharan African cohort [89], alongside institutional resources such as UPenn-GBM [94]. Newer challenge branches are therefore scientifically important because they change the population rather than merely enlarging it. BraTS-Africa represents geographic and acquisition shift; BraTS-PEDs changes age and tumor biology; metastasis and meningioma tasks change lesion morphology and clinical endpoints; radiotherapy tasks change the contouring purpose. A robust model should make its intended deployment domain explicit rather than claiming generic 'brain tumor segmentation.'

An additional concern in the foundation-model era is pretraining overlap. Public brain MRI collections can be mirrored, redistributed or incorporated into large pretraining corpora. A downstream test case may be held out from fine-tuning yet not genuinely unseen by the encoder. Future challenges should document post-cutoff acquisition dates and use hidden-server evaluation to reduce this contamination risk.

5. Preprocessing: where reproducibility is often won or lost

Preprocessing is often treated in brain MRI segmentation pipelines as an impartial prelude, but it may affect performance down the line considerably. Actions like registration, resampling, skull stripping, bias correction, intensity normalization, cropping, and quality control change the statistical distribution that the model encounters. Hence, the described architecture is impossible to reproduce without a detailed description of the preprocessing process.

The registration procedure helps the network to learn on a voxel-by-voxel basis and perform multimodal fusion; however, the registration process can cause blurring of the lesion area or distortion of the boundaries. The N4 bias field correction [110] and intensity normalization [114] may help to cope with scanner-based intensity drift, while z-score normalization inside the brain is widely used in BraTS-like workflows. The registration process itself is performed using established tools whose interpolation method needs to be reported [111]. This selection needs to be done without the knowledge of held-out labels. If normalization parameters are derived using the entire dataset, a type of preprocessing leak is possible.

Skull stripping is another hidden variable. BraTS releases have historically provided preprocessed images, whereas real clinical scans may include non-brain anatomy, variable field of view, motion and imperfect registration. Models developed exclusively on challenge-preprocessed inputs can therefore fail before the network itself is reached. Classical and learning-based brain extraction [112,113] help, but external validation should include realistic preprocessing failures rather than assume perfect upstream processing. Shared tooling such as the BraTS Toolkit improves comparability by fixing the preprocessing and fusion steps across submissions [123].

Patch-based sampling and tumor-aware sampling play a significant role in determining the extent of class imbalance. Random 3D patches mainly consist of background areas; oversampling of foreground areas speeds up the learning process but can change the probability calibration and false positives. Tumor-aware slice sampling is utilized in the 2026 GDNet approach in a 2.5D approach [17], showing that sampling of data can address some aspects of the memory overhead required for training in 3D. The augmentation policy requires equal consideration, since the transform family selected and the range of its parameters can affect the training distribution [115].

Table 3. Preprocessing steps, their common implementations, the failure mode introduced when they are under-specified, and the minimum reporting required for reproducibility.

Preprocessing step	Common choice	Failure mode if under-specified	Recommended reporting
Registration	Rigid/affine to template or reference modality	Interpolation artifacts; misregistration across sequences	Transform type, reference, interpolation, QC failure rate.

Preprocessing step	Common choice	Failure mode if under-specified	Recommended reporting
Bias correction	N4 or vendor correction	Residual low-frequency intensity drift	Algorithm/version and mask.
Skull stripping	Challenge-provided, BET, SynthStrip	Boundary errors; missing tumor near skull	Tool/version and manual correction policy.
Intensity normalization	z-score, Nyúl, percentile scaling	Scanner-dependent shift; leakage if global statistics used	Per-case vs training-set statistics and clipping.
Resampling	Isotropic or fixed spacing	Small lesion loss or shape distortion	Target spacing and interpolation by image/mask.
Cropping/patching	Brain box, fixed crop, tumor-aware patches	Context loss; biased prevalence	Crop rule, patch size, positive-sampling probability.
Augmentation	Geometric, intensity, bias/noise	Unrealistic transforms or label inconsistency	Exact transform ranges and whether modality-synchronized.

6. Model architectures across the segmentation literature

6.1 From U-Net to self-configuring 3D systems

The longevity of U-Net comes from its simple inductive bias, whereby local convolutional features are condensed into semantically rich representations and assembled with skip connections of high resolution. 3D U-Net and V-Net extended the same philosophy to 3D volume [5-7]; DeepMedic demonstrated the usefulness of multi-scale 3D pathway in the context of lesions segmentation [8]. These architectures remain popular because new architectures are being compared to versions of older architectures that have been less carefully optimized.

nnU-Net changed the landscape of competition from the introduction of new architecture to the optimization of configuration [10]. Fingerprinting of datasets decides patch size, network topology, normalization, and training schedule. Therefore, nnU-Net becomes not only a robust benchmark, but also a methodological obstacle for the claim of novelty. A new architecture that fails to beat or even match performance of an appropriately configured nnU-Net on the same data split, pre-processing and budgeted compute resources is questionable.

However, the discussion of 2D versus 3D is not binary. 3D architectures have a volumetric context but require higher memory, usually work with smaller batch sizes, and could be somewhat slower. 2D architectures scale well but ignore through-plane information. 2.5D architectures merge adjacent slices or views in order to balance both aspects. For example, GDNet [17] concatenates three slices from four modalities into 12-channel input; this architecture is particularly relevant for cases where budget for computation and latency are important.

Therefore, one can say that selection of architecture should be compatible with data regime. Large 3D architectures may be beneficial when there is a lot of standardized data and GPUs; lightweight 2.5D architectures might be more efficient for edge devices and rapid prototyping. Self-configured baselines are crucial for a fair comparison. Parameters count, inference time, peak memory usage, and energy consumption become standard alongside the Dice score.

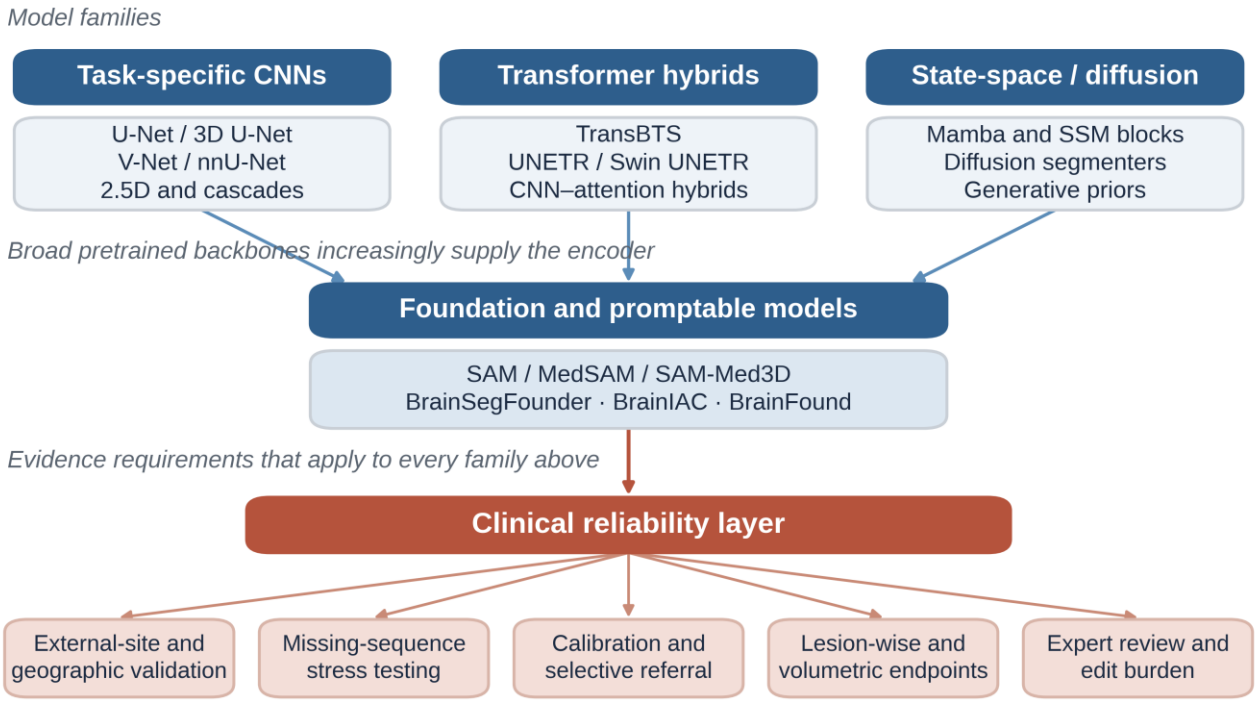


Figure 2. Architecture taxonomy and the reliability layer. The figure emphasizes that clinical evidence requirements sit above all model families.

6.2 Transformers and hybrid CNN–Transformer models

Transformers have found applications in brain tumor segmentation to allow modeling of interactions of a longer range than convolutional kernels can model intrinsically. TransBTS architecture combines a convolutional feature extractor with a transformer bottleneck [11]. UNETR performs processing of 3D volumes using patch tokenization and transformer representations with convolutional decoder [12], while Swin UNETR exploits the idea of shifted-window attention to improve computational locality [13]. nnFormer and its analogs make use of hierarchical attention mechanisms for volumetric segmentation [14].

The main argument in favor of transformers is not a general superiority of attention mechanisms but rather usefulness of global context both for spatially heterogeneous tumors and multimodal feature integration. The disadvantages of this architecture include high computational cost, requirements for data and hyperparameters tuning. When the size of

training cohorts is limited, an appropriately designed CNN can perform similarly or even better than a transformer. This factor becomes particularly important when evaluating a heavily engineered new transformer architecture against a minimalistic U-Net.

Modern hybrid architectures tend to use convolution for local boundary delineation together with attention or state-space blocks for modeling global context. In such manner, these architectures reduce the clear-cut distinction between CNNs and transformers, which makes it necessary to conduct a controlled experiment where preprocessing, augmentation techniques, loss function, optimization algorithm, random seed, training budget and model selection strategy are kept constant. Otherwise, architectural attribution becomes very poor.

6.3 State-space and Mamba models

State-space models with selective architectures, such as Mamba, allow for linear-time sequence modeling and have been instrumental in rapid uptake in volumetric medical imaging. The attraction of state-space models for 3D segmentation applications is straightforward – long-range spatial dependencies could be modeled without the quadratic costs of attention associated with standard transformers. Recent applications in brain tumors combine Mamba blocks with multimodal fusion and, in some cases, with uncertainty constraints [40,56–58]. The Mamba for 2026 incomplete modality application is a relevant example – cross-level fusion transmits context when MRI sequences are not available [40]. This is a more compelling clinical rationale than substituting an attention block to enhance model performance on the leaderboard. MambaDiff and diffusion-state-space hybrids similarly suggest that sequence modeling could be paired with probabilistic generative refinement [56–58].

Still, the empirical base is still quite thin. Many studies on Mamba published between 2025-2026 are limited to single-benchmark analyses with limited generalization. In future comparisons, the parameter count and training budget should be controlled for, and scanner changes should be taken into account. It would be reasonable to assume that at least part of the difference is attributable to training practice.

6.4 Diffusion and generative segmentation

Segmentation techniques have considered denoising as a segmentation algorithm or incorporated representations from diffusion processes in the segmentation task. Multimodal guidance, frequency-based fusion, and light-weight three-dimensional representations are explored for brain-tumour cases by BTSegDiff and DiffBTS [51-54]. Diffusion techniques naturally accommodate multiple possibilities and, perhaps, are compatible with uncertainty-driven segmentation. However, they impose limitations on computational expenses and

inference delay. Synthesis techniques belong to an even more upstream class, as they generate the missing MRI sequence and thereby fill gaps in the data. The unpaired versus paired comparison of MRI synthesis by the author [18] reveals the fundamental dichotomy of the problem: visually plausible generation does not guarantee that pathology is present. Both generative adversarial networks and diffusion models can produce hallucinations, remove enhancing foci, or change contrast in such a way as to create a visually similar image but to distort the tumour segmentation results. Therefore, synthesis-first segmentation techniques should evaluate how well the pathology was retained along with PSNR, SSIM, or FID. The clinical validation of the approach consists in examining if the downstream mask, lesion size and borders remain valid and rare pathology was retained. If a direct approach provides comparable results to synthesis-first pipelines, then one may avoid generation.

6.5 Foundation models and promptable segmentation

The foundation-model era changes both the opportunity and the burden of proof. SAM demonstrated broad promptable segmentation in natural images [23]; MedSAM adapted the paradigm to medical imaging [22]; SAM-Med3D and generalist 3D models extended the concept to volumetric data [24–25]. Brain-specific foundation models such as BrainSegFounder, BrainIAC and BrainFound pursue a different strategy: pretrain representations directly on neuroimaging data to improve transfer across brain tasks [19–21]. Between these poles sits a rapidly growing set of general medical and 3D segmentation foundation models, including 2D and volumetric SAM derivatives, text-driven and vision-language variants, unified 3D segmentation models, anatomy-guided self-supervised encoders and MRI enhancement models [26,30–36]; recent surveys map this landscape and its open problems [142,143].

BrainIAC is particularly important because it was trained and validated across 48,965 brain MRIs and showed its clearest advantages in few-shot and difficult-task settings [20]. That shifts the scientifically interesting question from 'does a foundation model beat nnU-Net on full BraTS?' to 'how rapidly does it adapt when only 1, 5, 10 or 50 labelled cases are available, and does that advantage survive a new institution?' BrainFound similarly targets cross-dataset and label-scarce generalization [21].

The 2026 MRI foundation-model landscape is already diversifying. Decipher-MR combines 3D MRI vision pretraining with report-guided text supervision across more than 200,000 MRI series [148], Triad targets segmentation, classification and registration within a shared 3D MRI representation [150], and a routine-MRI brain tumour foundation model uses large-scale clinical brain MRI pretraining for glioma molecular profiling and outcome prediction [149].

These systems are not all segmentation-specialized, but they matter because segmentation encoders increasingly come from broad pretrained backbones. Their inclusion strengthens the case for post-cutoff tumor benchmarks that can separate true transfer from pretraining familiarity.

Promptable segmentation models deserve equally careful scrutiny. A box or point prompt can make segmentation easier, but the prompt may itself encode substantial expert localization information. Zero-shot performance, automatically generated prompts and expert-supplied prompts should therefore be reported separately. Where parameter-efficient adaptation such as low-rank tuning is used to specialize a promptable model, the comparison should still include a task-specific baseline trained under the same label budget and evaluated on the same external cohort.

Reviews of foundation models in medical imaging caution that generalist pretraining does not by itself displace task-specific specialization, and that a properly tuned task-specific network remains the necessary comparator [141,144]. The practical implication is to evaluate foundation models for data efficiency, transferability, robustness and adaptation cost, not merely parameter scale.

Pretraining contamination is an emerging concern. Large foundation models may be trained on public MRI repositories, challenge data or derivatives. If the target benchmark is included directly or indirectly upstream, the downstream test is no longer independent in the usual sense. Papers should disclose pretraining data provenance and, whenever possible, use post-cutoff hidden cohorts acquired after pretraining was frozen.

Table 4. Model families, their primary advantage, typical adaptation regime and the controlled comparison required to support a claim of superiority.

Model family	Primary advantage	Typical adaptation	Critical comparison
nnU-Net / task-specific 3D CNN	Strong full-data benchmark; self-configuration	Train from scratch or ImageNet/medical pretraining	Matched compute and preprocessing; multiple seeds.
CNN–Transformer	Local detail + global context	Full fine-tuning	Does attention help beyond training recipe?
Mamba/SSM	Long-range context with linear scaling	Full fine-tuning	Parameter/compute-matched transformer and CNN.
SAM/MedSAM	Promptable transfer	Prompting, LoRA, adapters, fine-tuning	Zero-shot vs prompted vs adapted; prompt information burden.
Brain-specific FM	Label efficiency and cross-task reuse	Linear probe, adapters, partial/full fine-tuning	Few-shot curves and external-site transfer.
Diffusion segmentation	Probabilistic/refinement capacity	Task-specific training or conditioning	Latency, calibration, uncertainty usefulness.

7. Multimodal data, missing modalities and label efficiency

7.1 Multimodal MRI fusion: early, intermediate and late strategies

T1, T1ce, T2 and FLAIR are complementary rather than redundant. Early fusion stacks channels before the first encoder, which is simple and memory efficient but assumes all modalities are present and similarly informative. Intermediate fusion uses modality-specific encoders and combines features at one or more semantic levels. Late fusion aggregates predictions or high-level representations, improving modularity but potentially missing fine cross-modal interactions.

Attention-based fusion, cross-modal transformers, gating, prototype alignment and shared-private representations have all been proposed to learn modality interactions. A central methodological question is whether the model truly uses complementary information or learns a dominant-modality shortcut. Sequence-ablation studies should therefore be routine. Performance should be measured for each single modality, clinically plausible pairs, all three-modality subsets and the full four-sequence input.

The value of an additional modality should be evaluated against acquisition cost and robustness. A model that gains 0.003 Dice from a fourth sequence but collapses when that sequence is absent may be less useful than a slightly weaker model designed for incomplete inputs. This is especially relevant in retrospective datasets and low-resource settings where complete standardized imaging is uncommon.

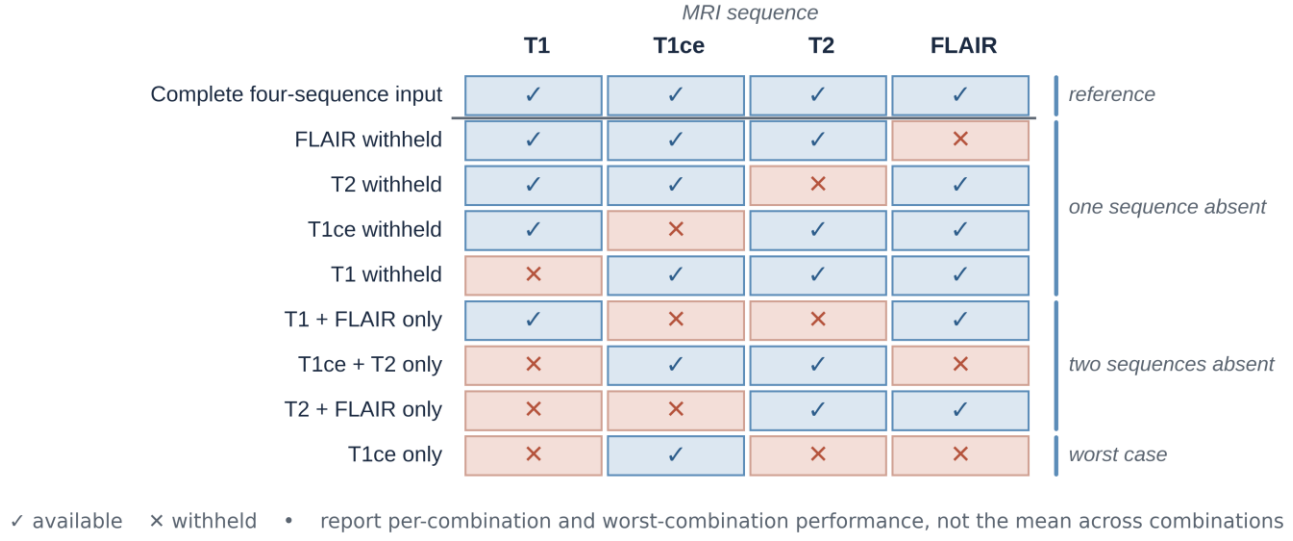


Figure 3. A minimum missing-modality stress matrix for four-sequence brain MRI. Studies should add naturally incomplete and corrupted-input tracks where available.

7.2 Missing-modality segmentation: from imputation to adaptive robustness

Missing MRI sequences are one of the most clinically realistic failure modes in multimodal brain tumor segmentation. Causes include protocol differences, patient intolerance, scanner availability, motion artifacts, contrast agent contraindication, and retrospective data incompleteness. Thus, the conventional assumption of full T1/T1ce/T2/FLAIR overstates readiness for clinical application.

The methodologies span four major groups. Methods of synthesis reconstruct the missing image before segmentation. Shared representation learns modality-invariant features. Knowledge distillation transfers information from the teacher network trained on the complete-modality image to the student trained on the incomplete one. Adaptive fusion conditions the network on the presence of certain sequences. New architectures developed during 2025-2026 incorporate meta-learning, contrastive prototypes, semantic gating, text prompts, or state-space fusion [37-50].

The 2026 study on children is important since it assessed meaningful missing-sequence cases on 715 CBTN/BraTS-PEDs patients and an independent longitudinal PNOC cohort; modality dropout enabled segmentation without more than ~0.04 Dice loss compared with the complete case [96]. This design reflects the real-life task better than purely synthetic modality dropout on a challenge cohort.

Knowledge distillation systems of recent years focus on adaptation to non-major modality combinations [37]. Contrastive prototype learning and meta-guided fusion try to provide generalizable representations regardless of the missing pattern [38,41]. State-space Mamba fusion takes into account long-range state-space interactions [40]. All these ideas need to be tested both on the average missing-modality case and on the most problematic missing-sequence combination. It is important to know which input sets the system supports and which does not, rather than making silent extrapolation.

The review recommends the following evaluation pipelines: 1) complete MRI, 2) naturally incomplete MRI, and 3) withheld-modality stress test. If synthesis is used, then there should be a 4th pipeline testing whether the synthetic images keep tumor morphology and volume. Reporting mean Dice across all missing modalities may mask the system failure for the most informative missing-sequence.

7.3 Self-supervised, semi-supervised, weakly supervised and label-efficient learning

Professional annotation of tumors is an expensive affair, especially if there are several subregions or postoperative changes involved. The fact is what has led researchers to use various techniques such as self-supervised pretraining, semi-supervised consistency learning,

weak supervision, and foundation model adaptation. The fundamental issue here goes beyond that of improvement in the Dice score; rather, it has to do with scalability with respect to labeled data.

Self-supervised methods learn anatomical or intensity representations from unlabeled volumes. Models Genesis, masked image modeling and self-supervised volumetric transformer pretraining established important precedents [27–29]. BrainIAC and BrainFound can be viewed as a large-scale continuation of this direction, emphasizing transfer across tasks and datasets [20–21]. The appropriate evaluation is a label-efficiency curve: performance at 1%, 5%, 10%, 25%, 50% and 100% labels with identical test data.

Semi-supervised methods can exploit unlabeled target-domain data, but leakage is possible if the supposedly unlabeled pool includes cases related to the test set. Weakly supervised methods may use image-level labels, coarse boxes or pseudo-masks; the supervision burden must be quantified because a method using expert boxes is not directly comparable to a fully automatic system. Foundation-model prompts should be treated similarly: a prompt is a form of supervision.

Label-efficient methods are most consequential for small lesions, because small targets are precisely where global overlap metrics are least informative. Future label-efficient work should therefore combine learning curves with lesion-size-stratified sensitivity and calibration rather than reporting a single aggregate score.

8. Tumor-type-specific segmentation

8.1 Adult glioma segmentation

Adult glioma remains the dominant segmentation benchmark because BraTS provides large, standardized multimodal cohorts and established ET/TC/WT metrics. The task has driven much of the architectural progression from DeepMedic to 3D U-Nets, nnU-Net, transformers and ensembles. Current performance on standardized data can be high, but enhancing tumor remains difficult when lesions are small, nonenhancing or postoperative.

An important trend is the shift from a single mean Dice to subregion-aware evaluation. Whole tumor is generally easier because it is larger; ET and TC are more sensitive to boundary ambiguity and tumor biology. Reporting a macro-average without individual subregions can obscure a clinically important failure. Similarly, post-processing that removes small connected components may improve average ET Dice yet erase real tiny enhancing foci.

The most useful new adult-glioma studies therefore combine strong segmentation with robustness, label efficiency or calibration. GDNet [17] contributes a compute-aware 2.5D perspective. Recent cross-modal fusion and context-aggregation models explore representation improvements [59,145]. CONSeg directly addresses uncertainty [67]. The field should increasingly prioritize such orthogonal reliability dimensions rather than incremental architectural complexity alone.

8.2 Pediatric brain tumors

Pediatric brain tumors should not be treated as a smaller version of adult glioma. Histologic and molecular distributions differ, tumor locations differ, acquisition protocols vary by age and institution, and sedation or motion can alter image quality. BraTS-PEDs and CBTN resources therefore provide an essential stress test of model transfer.

The 2026 missing-MRI pediatric study [96] is notable because it couples segmentation with prognosis and includes an independent longitudinal trial cohort. Such designs are closer to real clinical use than single-edition challenge validation. Pediatric research should also report age-stratified performance and explicitly test whether models pretrained on predominantly adult populations exhibit systematic failure in younger patients.

Future pediatric benchmarks should include incomplete sequences, low-field or heterogeneous scanners, rare tumor classes and longitudinal treatment effects. The major opportunity is a shared pediatric benchmark in which the hidden test cohort is genuinely external and temporally later than foundation-model pretraining.

8.3 Brain metastases: lesion-wise detection is indispensable

Brain metastases present a different segmentation problem from diffuse glioma. Patients may have dozens of lesions; many are small; and the clinical cost of missing one lesion can exceed the cost implied by a modest change in global Dice. Lesion-wise sensitivity, false positives per scan and size-stratified performance are therefore mandatory.

Recent studies illustrate the move toward clinically aligned evaluation. A 2025 routine implementation study used automated metastasis segmentation to support RANO-BM follow-up [97]. A data-centric external-validation study reported high lesion sensitivity across internal and external cohorts [98]. A 2026 multi-rater evaluation of a hybrid nnU-Net plus MedSAM framework compared AI contours with experts under bias-controlled conditions [99]. These designs are more informative than a single aggregate overlap metric.

Size stratification is particularly important. Recent longitudinal work linked segmentation to local control/failure and reported different Dice by lesion diameter [100]. Such analyses

should become standard: <3 mm, 3–5 mm, 5–10 mm, 1–2 cm and >2 cm strata may reveal very different clinical reliability.

8.4 Meningioma: volumetry, uncertainty and radiotherapy

Meningiomas are usually extra-axial and often conspicuous on post-contrast T1 MRI, but skull-base location, dural tails, en-plaque growth and postoperative change can make boundaries ambiguous. Volumetric analysis is promising for surveillance since one-dimensional measures fall short when describing tumor volume in an irregular fashion. The recent literature until 2026 presents segmentation of meningioma in addition to overlap measurements. In particular, volumetric validation from the external dataset and multiple parametric detection methods have been published [101,102]. Interestingly, evidential ensemble methods produced a median Dice similarity coefficient of about 0.93 for the internal dataset and 0.92 for the external dataset along with the uncertainty maps and volumetric confidence intervals [66]. This is an example of segmentation work anticipating the following query of the doctor: in what situations is the estimated volume reliable?

BraTS-MEN and BraTS-MEN-RT extend the benchmarking framework for radiotherapy planning [91,95]. In this case, some criteria like contour acceptability, time to edit contours, and impact on the dose are important. A method that slightly outperforms in Dice but requires more time for manual edits in sensitive structures cannot be regarded as successful.

8.5 Postoperative, recurrent and treatment-response segmentation

Postoperative segmentation is harder than preoperative segmentation because blood products, ischemia, enhancement, surgical cavities and treatment effects create appearances that are uncommon in standard preoperative BraTS data. A preoperative model applied unchanged after surgery can therefore fail systematically even if its benchmark performance is excellent.

Recent external-validation work on postoperative glioblastoma and extent of resection demonstrates that dedicated models and evaluation are needed [104]. Longitudinal applications add another requirement: consistency. A small apparent volume change can be caused by segmentation noise rather than tumor biology. Repeatability and serial calibration should therefore accompany Dice.

Clinical translation should connect masks to downstream decisions such as extent of resection, progression, radiotherapy target volume or RANO classification. The most valuable studies will prospectively measure whether AI reduces contouring time without increasing clinically meaningful discrepancies.

9. Evaluation, robustness and trustworthiness

9.1 Evaluation metrics: Dice is necessary but not sufficient

Dice similarity coefficient remains the dominant segmentation metric because it is intuitive and scale-free, but it is sensitive to lesion size and does not distinguish missed lesions from boundary shifts in a straightforward clinical way. While the HD95 and distance to surface errors capture the geometry aspect of the error, they may show instability for highly localized structures. This is a problem for the training process where the loss functions such as generalized Dice, Tversky, and focal losses designed specifically for imbalanced classes maximize an overlap-based approximation and thus make a compromise between lesion localization and boundary agreement which cannot be uncovered through the main metric [117-119]. No metric is enough to cover all types of failure [120-122]. In case of gliomas, WT/TC/ET Dice and HD95 must be reported independently. In case of metastasis, important metrics include lesion detection sensitivity, precision/F1, number of false-positive lesions, and size groups. For uncertainty-aware systems, expected calibration error alone is insufficient; one should report uncertainty–error correlation, risk–coverage curves and performance of accepted versus referred cases.

Confidence intervals should respect the sampling unit. Multiple slices, patches or lesions from the same patient are not independent patients. Bootstrap resampling should occur at patient or episode level, and longitudinal scans may require patient-clustered analysis. Studies with multiple institutions should report worst-site as well as pooled performance.

The field would also benefit from reporting compute-normalized performance. Parameter count, FLOPs, peak GPU memory, training time and inference time contextualize whether a 0.003 Dice gain is worth a tenfold increase in resources. This is especially relevant as foundation models and diffusion networks grow.

Table 5. Task-specific minimum metric sets, and the failure mode each clinical question most commonly conceals.

Clinical question	Minimum metrics	Commonly hidden failure
Adult glioma subregions	WT/TC/ET Dice + HD95 + volume error	Averaging hides weak ET/TC.
Brain metastases	Lesion sensitivity/F1 + FP/scan + Dice + size strata	Tiny lesions disappear in global overlap.
Meningioma surveillance	Dice + volume bias + repeatability + UQ	Small serial volume changes may be model noise.
Postoperative residual	Residual-volume error + lesion detection + expert acceptability	Blood products/cavity mimic tumor.
Missing modality	Per-combination metrics + worst-combination performance	Mean over combinations hides collapse.
Selective segmentation	Risk–coverage + calibration + referred-case error	Uncertainty map may look plausible but not predict failure.
External validation	Patient/site CIs + worst-site performance	Pooled performance masks site failure.

9.2 Uncertainty, calibration and selective segmentation

Segmentation uncertainty is clinically useful only if it changes action. MC dropout, deep ensembles, evidential learning, probabilistic U-Nets and conformal prediction can all produce uncertainty estimates [66–74]. The critical test is whether uncertainty is higher where the model is wrong and whether that relationship holds under domain shift.

CONSeg demonstrated that conformal uncertainty could identify glioma cases with worse Dice and HD95, with coverage close to the nominal target on internal and external data [67]. The 2026 meningioma study used evidential ensembles and external validation to produce calibrated volumetric uncertainty [66]. These are stronger demonstrations than qualitative heatmaps because uncertainty is linked to measurable error and potential human referral.

Selectively segmenting reinterprets the issue of deployment as one of coverage. Instead of forcing automatic segmentation to be used in all scans, the method can accept easy cases and send hard cases for manual segmentation. The useful analysis will need to provide a measurement of the segmentation error at predetermined levels of automatic coverage (for instance, 50%, 70%, 80%, 90%, and 100%). The desirable effect should be monotonic in nature: as the algorithm becomes more selective, the error among accepted cases should decrease, and the referred cases should include more true negatives. It must be separated into calibration and overlapping. The algorithm that has a good Dice score may still be overly confident. The use of approaches such as temperature scaling, ensembles, or conformal sets at a voxel level can help to calibrate without hurting segmentation scores based on rankings.

9.3 Domain shift, external validation and geographic robustness

The cross-site generalization problem is a major deficiency separating challenge success from clinical applicability. Differences in scanner manufacturer, field strength, sequence execution, slice thickness, reconstruction, contrast timing, and hospital procedures may affect the MRI appearance. The pre-processing steps of skull stripping and image registration may work differently in uncurated data. Scanner- and site-specific effects are quantifiable confounders, and medical imaging algorithms have been found to consistently employ site-correlated shortcuts, thus overfitting internal performance at the cost of external validity; the implications vary for different patient populations. BraTS-Africa is significant exactly due to the introduction of population and acquisition diversity not covered by random resampling of the original BraTS population. There are several domain adaptation and generalization strategies, including UDA-GS; however, the strongest evidence is still external validation on institutions which were completely left out of the training and fine-tuning of the method. External validation should be done in hierarchical manner: (1) check the same kind of dataset

but on another set; (2) evaluate on independent institutions and scanners; (3) do temporal validation on future acquisitions; (4) validate prospectively. A model cannot be considered "generalizable" simply because it works on another edition of BraTS, based on the same challenge dataset structure. The result of the worst performing institution must be reported separately. The pooled Dice score can be deceptive in case of dominating easy institution in the data. Site-stratified confidence intervals, scanner-vendor stratification analysis, and failure galleries are more informative metrics. Geography must be treated as a domain factor.

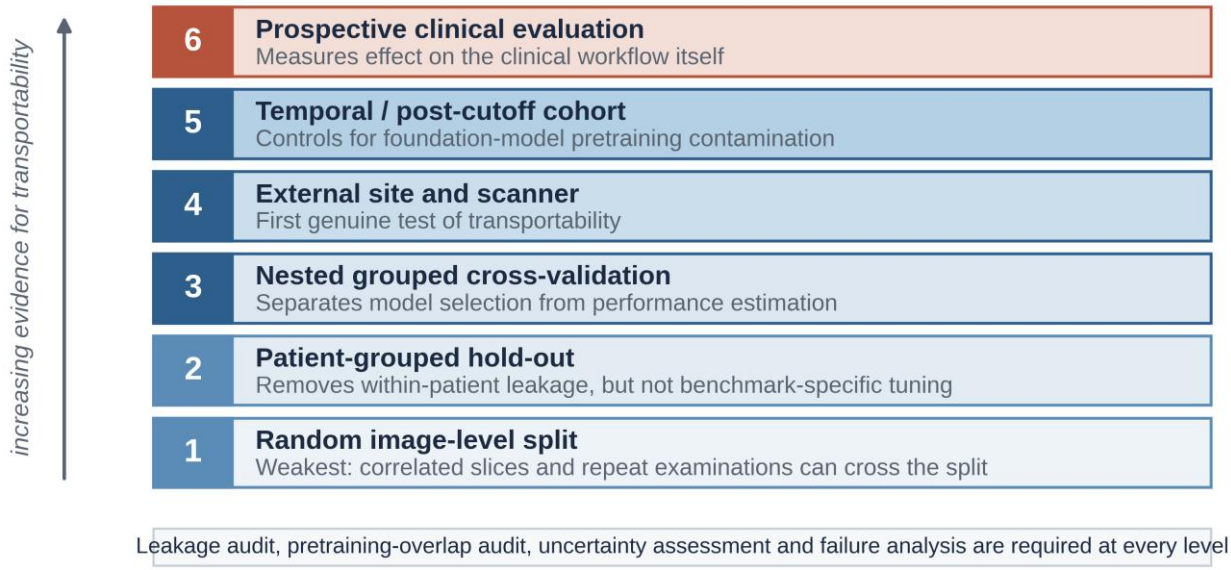


Figure 4. Evidence ladder for segmentation transportability. Each level should retain leakage audits, failure analysis and uncertainty assessment.

9.4 Data leakage, challenge overfitting and benchmark contamination

Medical imaging segmentation in particular is subject to risks of non-independence caused by the number of observations each patient provides, including different slices, patches, timepoints or derivations. Therefore, random splitting of image-level dataset might put highly correlated data points both in train and test sets. While respecting patient identifiers is a start, longitudinal scans of the same patient and their derivatives might span across the split. Duplicate checks are necessary. Simple hashing will detect bit-by-bit identical files, perceptual hashing or embedding comparison could help to identify resized/re-encoded duplicates, and metadata or filename patterns could point to the same series of scans. Thus, while building confidence intervals, the proper unit should be an independent patient or episode, and not the individual slice. Overfitting issues are harder to spot multiple submissions to the leaderboard or use of public validation annotations might bias architecture decisions or even hyperparameters, leading to validation leakage. Hidden test servers help

but cannot prevent adaptation to the distribution of the benchmark. Pretraining poses a contamination risk: a publicly released BraTS scan might be prohibited from fine-tuning but still present in pretraining or its derivative repository. It is essential for the authors to specify sources, dates and any overlap of the pretraining material. The best protection against the issue is to collect a hidden post-cutoff dataset that is acquired after the foundation model is trained. Therefore, the paper suggests inclusion of stable patient identifiers, site, scanner, acquisition date, longitudinal episode, source repository and duplicates in the manifest of future benchmark. Such information does not reveal patients' identity but acts as a provenance control.

9.5 Annotation uncertainty and multi-rater ground truth

Manual tumor segmentation is not an absolute truth. Experts disagree at infiltrative margins, around necrosis, after surgery and where enhancement is subtle. Consensus masks can hide this uncertainty by collapsing multiple plausible boundaries into one binary target.

Probabilistic segmentation methods such as the probabilistic U-Net and PHiSeg explicitly model multiple plausible annotations [61,116]. Brain tumor benchmarks could go further by releasing several expert masks per difficult case and reporting inter-rater distributions. Models could then be judged against the range of expert disagreement rather than a single mask.

A particularly valuable future dataset would include five to ten specialist annotations for a subset of cases, region-level ambiguity tags and adjudicated clinical notes. Such data would allow separation of epistemic model uncertainty from irreducible annotation ambiguity and would provide a rigorous target for uncertainty-aware foundation models.

10. Federated learning and explainability

10.1 Federated learning, privacy and multi-institutional training

Brain tumor segmentation is well suited to federated learning because data are distributed across institutions and central pooling can be difficult. Early feasibility work showed that multi-institutional tumor segmentation could be trained without directly sharing patient images [76,77], and related neuroimaging work combined federated training with domain adaptation across sites [78]. More recent algorithms target stronger generalization and incomplete modalities [47,75].

The 2025 generalized federated algorithm reported a 2.2% Dice improvement over a federated baseline with only a modest gap from centralized training [75]. Such results are encouraging,

but federated studies should report client heterogeneity, communication cost, privacy threat model and worst-client performance. A federated average can hide a hospital that benefits little from collaboration.

Privacy-preserving learning is not automatically private. Updates to models could lead to leaking information, and the embeddings generated using the foundational models might carry patient identities. In cases where claims move outside the local datasets, differential privacy, secure aggregation, and attack evaluations need to be thought of. Benchmarking platforms like MedPerf are the infrastructural way forward for hidden multi-site evaluations [79].

10.2 Explainability and human interaction

Segmentation explanations should be customized based on the task in question. As a mask represents spatial output, saliency heatmaps usually provide little value. Useful explanations could include uncertainty visualizations, ablation studies for different modalities, counterfactual sequence testing, and illustration of contour changes due to sequence removal. Explainability via human-in-the-loop evaluation provides better results than mere visualizations. For instance, the 2026 study aimed at reducing bias in the metastasis detection task has humans reviewing or modifying AI-derived contours [99]. Future trials should record correction time, voxels changed, clinically critical boundary changes and disagreement by lesion size/location.

Modality reliance is another interpretable target. A system that appears multimodal may depend overwhelmingly on T1ce. Systematic single-sequence and missing-sequence tests reveal this behavior more directly than attention maps. For foundation models, prompt sensitivity should also be audited because small changes in points or boxes can alter the mask.

11. Efficiency, reproducibility and the case for strong baselines

Computational efficiency is becoming a first-class research question. Diffusion models and large foundation models can consume substantially more memory and inference time than a compact nnU-Net or 2.5D network. The correct comparison therefore depends on the intended deployment: large offline radiotherapy planning may tolerate slower inference; near-real-time clinical tools or low-resource sites may not.

Recent lightweight systems such as LCBTS-Net, TransLiteUNet, FFLUNet and LATUP-Net show that small parameter counts can remain competitive [60,62–65,147]. GDNet adds an

author contribution in the 2.5D efficiency space [17]. These studies motivate Pareto-front reporting: Dice or lesion sensitivity versus parameters, memory, latency and training energy.

Reproducibility requires more than code release. The exact dataset edition, split manifest, preprocessing, seed, software versions, pretrained weights, checkpoint-selection criterion, threshold/post-processing and evaluation code should be frozen. Prediction files should retain patient IDs and site identifiers so bootstrap and subgroup analyses can be regenerated.

A strong paper should include nnU-Net or an equivalently optimized baseline unless there is a clear reason not to. Architectural novelty should be interpreted only after controlling training recipe and compute. Multiple seeds are particularly important when reported gains are small.

12. DATA-SEG: a reliability-first reporting framework

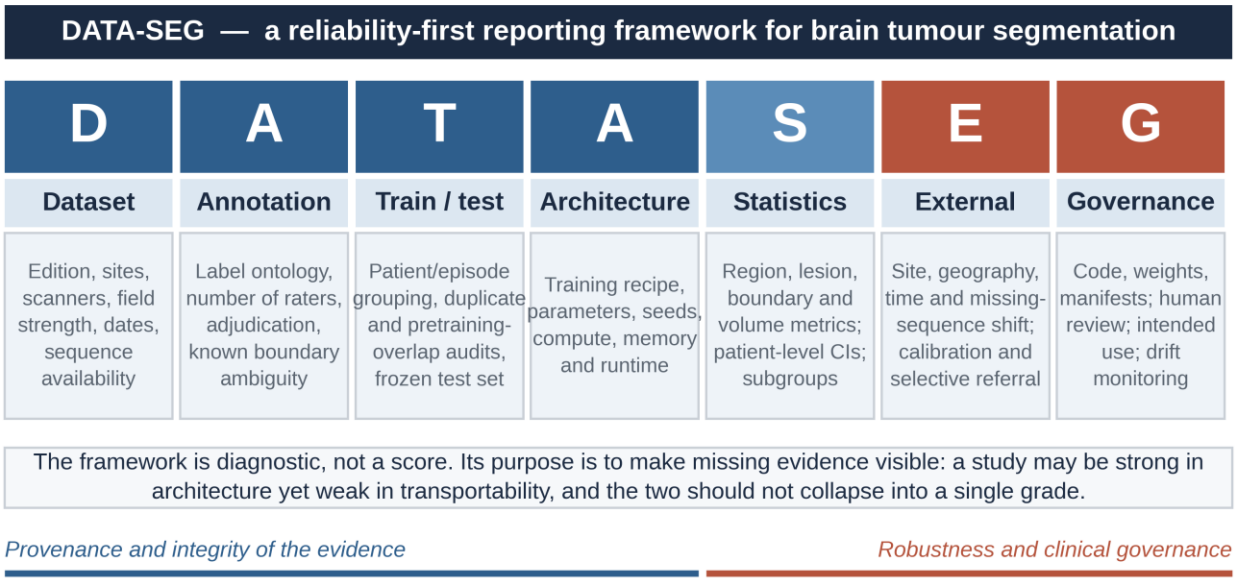


Figure 5. DATA-SEG framework proposed for brain tumor segmentation studies.

DATA-SEG is proposed as a compact checklist for designing and reviewing brain tumor segmentation research. It intentionally combines data provenance with statistical and clinical evidence because failures frequently originate outside the network architecture.

D, Dataset: report dataset/version, acquisition dates, institutions, scanners, field strengths, sequence availability, tumor spectrum and inclusion/exclusion. A, Annotation: report label ontology, annotator expertise, number of raters, adjudication and known ambiguity. T, Train/test integrity: group by patient and longitudinal episode, audit duplicates and pretraining overlap, and lock the final test set before model selection.

A, Architecture: specify dimensionality, parameter count, pretraining, input channels, loss, sampling, augmentation, optimization, checkpoint selection, post-processing and compute. S, Statistics: report task-appropriate region, lesion, boundary and volume metrics with patient-clustered uncertainty, multiple seeds and subgroup/worst-site analyses.

E, External robustness: include site/scanner/geography/time shift, missing-modality and corruption stress tests, calibration, uncertainty and selective referral. G, Governance/clinical readout: release code/weights/manifests where possible; report runtime, human correction, downstream clinical agreement, privacy and monitoring plans.

The framework is not intended to create another score that can be gamed. Its purpose is to make missing evidence visible. A study can be strong in architecture but weak in transportability; DATA-SEG should reveal that distinction rather than collapse it into a single grade.

Table 6. DATA-SEG domains: minimum manuscript evidence and high-value extensions.

DATA-SEG domain	Minimum manuscript evidence	High-value extension
Dataset	Edition, sites, scanners, modality availability, dates	Post-cutoff external cohort and geography.
Annotation	Ontology, raters, adjudication	Multi-rater probabilistic subset.
Train/test integrity	Patient/episode grouping, duplicate audit, frozen test	Foundation-model pretraining-overlap audit.
Architecture	Full training recipe, compute, seeds	Compute-matched ablations and Pareto analysis.
Statistics	WT/TC/ET + lesion/boundary/volume metrics, CIs	Worst-site and size-stratified analysis.
External robustness	Independent sites, missing modalities, corruptions	Temporal/prospective validation.
Governance	Code/weights/manifest, human review, monitoring	Prospective reader study and drift monitoring.

13. Evidence synthesis across method families

13.1 Evidence atlas: representative method families and current interpretation

Table 7. Evidence atlas of representative method families, their principal contribution and the critical interpretation each requires.

Representative approach	Family	Primary contribution	Critical interpretation
nnU-Net (2021)	Self-configuring 2D/3D CNN	Strong baseline across biomedical segmentation	Still difficult to beat fairly; mandatory comparator for most new methods.
TransBTS (2021)	CNN + Transformer bottleneck	Global context for multimodal BraTS	Established hybrid reference; gains must be separated from training recipe.
UNETR/Swin UNETR (2022)	Transformer encoder / hierarchical windows	3D global context	Powerful but compute/data intensive; strong pretraining often matters.
BrainSegFounder (2024)	Brain segmentation foundation model	Transfer across neuroimaging tasks	Tests whether pretrained brain representations reduce task-specific labels.
MedSAM/SAM-Med3D (2024–25)	Promptable medical segmentation	Broad transfer and interactive use	Prompt burden and 2D/3D consistency must be reported.
BTSegDiff/DiffBTS (2025)	Diffusion segmentation	Probabilistic refinement /	Promising uncertainty link;

Representative approach	Family	Primary contribution	Critical interpretation
		multimodal modeling	inference cost and external validation remain gaps.
LCBTS/FFLUNet (2025–26)	Lightweight CNN/hybrid	Compute-efficient segmentation	Relevant for deployment; compare on a Pareto frontier, not Dice alone.
Missing-modality distillation (2025–26)	Teacher–student / adaptive fusion	Robustness to absent sequences	Worst-combination performance is more important than mean.
BrainIAC (2026)	Brain MRI foundation model	Few-shot/general transfer	High-impact direction: label efficiency + external transfer.
BrainFound (2026)	Self-supervised brain MRI FM	Cross-dataset label-scarce transfer	Supports scalable slice-based pretraining; tumor-specific segmentation needs matched tests.
CONSeg (2025)	Conformal uncertainty	Error-aware glioma segmentation	Moves UQ from maps to coverage and referral.
Evidential meningioma ensembles (2026)	Evidential UQ ensembles	Calibrated volumes + external validation	Strong template for clinically trustworthy segmentation.

13.2 Quantitative evidence atlas: representative 2021–2026 studies

Table 8 consolidates representative studies across task-specific networks, lightweight systems, incomplete-modality methods, uncertainty approaches and foundation models. The table is deliberately descriptive rather than a league table. Scores reported by different BraTS editions or external cohorts are not directly exchangeable; the purpose is to expose the data setting, evaluation design and the type of evidence each method contributes.

Table 8. Quantitative evidence atlas of representative 2021–2026 studies. Values are reproduced as reported by the original authors; datasets, label definitions and evaluation servers differ, so entries are not directly comparable and are not ranked.

Year	Study / family	Data setting	Method	Reported contribution	Interpretation
2021	nnU-Net [10]	Multiple biomedical / BraTS-like	Self-configuring 2D/3D CNN	Strong task-specific baseline	General baseline, not a single tumor-specific score
2021	TransBTS [11]	BraTS multimodal MRI	CNN + transformer bottleneck	Global multimodal context	Established hybrid reference
2022	UNETR [12]	3D medical segmentation	Transformer encoder + CNN decoder	Volumetric transformer baseline	High compute/data demand
2022	Swin UNETR [13]	BraTS/3D MRI	Hierarchical shifted-window transformer	Strong 3D context modeling	Often benefits from pretraining
2023	Multimodal Transformer for incomplete MRI [42]	BraTS 2018–2020	Modality encoders + multimodal transformer	Missing-modality compensation	Tests many incomplete combinations
2023	3D GAN , Kiani et al. [15]	BraTS	3D GAN segmentation	Generative regularization	Author contribution; early GAN pathway
2023	ACGAN , Kiani et al. [16]	BraTS	Auxiliary-classifier GAN	Joint adversarial/classification learning	Author contribution
2024	BrainSegFounder [19]	Multi-dataset neuroimaging	Brain-specific foundation model	Transfer learning for neurosegmentation	Evidence for pretraining across brain tasks
2024	MedSAM [22]	Large medical segmentation corpus	SAM adapted to medical data	Promptable transfer	Prompt dependence must be disclosed
2024	Postoperative GBM model [104]	External postoperative cohorts	Dedicated postoperative DL	Residual / EOR evaluation	Shows pre-op and post-op are distinct domains
2025	BTSegDiff [51]	Multimodal brain MRI	Dynamically guided diffusion	Diffusion-based probabilistic segmentation	Latency and external validation remain key
2025	DiffBTS [52]	3D multimodal MRI	Lightweight diffusion	Efficiency-oriented diffusion	Useful compute/accuracy trade-off
2025	UKAN-EP [146]	BraTS-GLI 2024	3D U-KAN + efficient attention + pyramid fusion	WT Dice 0.9001 ± 0.0127; 11.30M params	Strong benchmark result; needs true external-site test
2025	Federated generalized BTS [75]	Multi-client brain tumor data	Federated learning	+2.2% DSC vs FL baseline; ~3% below centralized	Tests collaboration without data pooling
2025	RANO-BM routine implementation [97]	132 patients / 27,456 images	UNETR transfer learning	Mean Dice ~0.77; RANO-BM workflow	Clinical integration rather than leaderboard optimization
2025	CONSeg [67]	Internal + external	Conformal prediction on	Uncertain-case	UQ linked to actual

Year	Study / family	Data setting	Method	Reported contribution	Interpretation
		glioma cohorts	U-Net	DSC/HD95 significantly worse than certain cases ($p<0.001$)	error/referral
2025	Incomplete-modality optimal transport FL [47]	Incomplete multimodal MRI	Federated distillation + optimal transport	Privacy + missing-modality learning	Important combined stress scenario
2025	FFLUNet [62]	BraTS-style 3D MRI	Lightweight frequency-learning U-Net	~1.45M parameters; faster than nnU-Net in report	Efficiency-focused alternative
2026	GDNet [17]	BraTS 2024	2.5D multimodal U-Net + EMA	3 adjacent slices × 4 modalities; tumor-aware sampling	Compute-aware author contribution
2026	BrainIAC [20]	48,965 brain MRIs	Self-supervised brain MRI foundation model	Strong few-shot / difficult-task transfer	High-value label-efficiency evidence
2026	BrainFound [21]	Diverse brain MRI datasets	Slice-based self-supervised foundation model	Cross-dataset and label-scarce gains	Promising generalization; tumor segmentation needs matched benchmark
2026	No-modality-left-behind distillation [37]	Incomplete multimodal MRI	Knowledge distillation + graph refinement	Robustness to absent sequences	Focus on difficult modality combinations
2026	Contrastive prototype missing-modality [38]	Incomplete multimodal MRI	Contrastive prototypes	Resilience to missing sequences	Modern representation alignment
2026	LFC-SATP-SGG [39]	Incomplete multimodal MRI	Latent feature completion + text prompting + gating	State-aware missingness modeling	Interesting multimodal prompting direction
2026	Mamba missing-modality network [40]	BraTS 2018/2020	Mamba fusion + uncertainty constraint	~1 DSC point over strongest baseline reported	Needs external site validation
2026	MGML [41]	Incomplete multimodal MRI	Meta-guided plug-and-play fusion	Adaptation across modality subsets	Targets general incomplete inputs
2026	CMHF-3DNet [145]	BraTS 2023–2025	Cross-modal transformer + hierarchy-aware loss	Mean DSC 0.8527; HD95 6.47 mm on BraTS 2025	Multi-edition robustness, not equivalent to external hospital
2026	TransLiteUNet [147]	Two public 3D datasets	CNN–Transformer hybrid	0.43M params; 14.98 GFLOPs	Highlights extreme parameter efficiency
2026	Meningioma evidential ensemble [66]	Internal 788 pts + external 353 pts	Evidential deep ensembles	Median Dice 0.93 internal / 0.92 external	Strong template for UQ + external validation
2026	TUM-SAM metastasis [99]	301 train pts + 105 external pts	nnU-Net detection + tumor-adapted MedSAM	External DSC ~0.84; expert comparison	Human-rater evaluation strengthens translation

13.3 Quantitative synthesis by method family

In terms of task-specific architectures, however, the key point that emerges from all such approaches is that full-data performance levels have become saturated such that the design of evaluation itself often matters more than the backbone label. nnU-Net still represents a stringent benchmark by virtue of automated tuning and solid heuristics that minimize the contributions of a novel architecture that could erroneously be taken as the explanation of superior performance. While both Transformer and KAN studies may boast of impressive scores, their true importance lies in the contributions of multimodal fusion, hierarchical constraints, efficiency, and robustness beyond a modest Dice increase.

New quantitative results from 2025-2026 highlight several particular niches in which such approaches can compete effectively. UKAN-EP scored 0.9001 ± 0.0127 in the whole-tumor Dice and used 11.30 million parameters in BraTS-GLI 2024, proving the viability of efficient KAN designs. In turn, CMHF-3DNet showed a DSC of 0.8527 and HD95 of 6.47 mm in BraTS 2025 while implementing anatomical constraints of ET $\subset$ TC $\subset$ WT. TransLiteUNet is another approach advancing the efficiency boundaries, with only 0.43 million parameters used. None

of these numbers can be compared straightforwardly due to differences in dataset, labels, and metrics used, but they define separate points on an accuracy-compute-robustness surface.

Finally, foundation-modeling evidence shows yet another benefit of the technique. While BrainIAC was trained and validated on 48,965 MRIs and excels in low-data and hard tasks setting, BrainFound emphasizes the ability to work well in cross-dataset and label-scarce transfer setting. Therefore, fair comparisons should focus not just on the full-data scores but learning curves, with the former being plotted for each foundation model. Foundation model that outperforms nnU-Net on 5 labeled samples have its own scientific importance regardless of relative failure in case of 100+ labeled cases.

Uncertainty is one of the emerging lines of work in recent times. The CONSeg framework includes glioma segmentation using external segmentation together with conformal coverage, where it shows that increased uncertainty is associated with poor segmentation performance. The 2026 evidential-ensemble paper for the meningioma case has a very high median Dice score in an independent external test set using calibrated volumetric uncertainty. Overall, these studies raise an answer to a clinically relevant question that is not usually addressed in regular architecture papers: does the model know when to apply automated processing?

The metastasis papers also highlight the need for lesion-based metrics and human comparison. In such a way, even an automated approach for RANO-BM may prove useful despite its average Dice score being approximately 0.77. Conversely, a model with excellent global overlap may still be unacceptable if it misses subcentimeter lesions. The unit of clinical failure must therefore determine the unit of evaluation.

Table 9. Quantitative synthesis by method family: where each family is strongest, its main unresolved risk and the most informative next experiment.

Family	Where it is strongest	Main unresolved risk	Most informative next experiment
nnU-Net / 3D CNN	Full-data task-specific accuracy and reproducibility	May overfit benchmark preprocessing/domain	Post-cutoff external site with missing sequences.
2.5D/lightweight	Compute-constrained deployment and rapid inference	Reduced through-plane context	Pareto comparison vs 3D models across lesion sizes.
Transformer hybrid	Global context and multimodal interaction	Compute/data sensitivity	Compute-matched external-site ablation.
Mamba / SSM	Long-range context with linear scaling	Young evidence base	Matched transformer/CNN test on external sites.
Diffusion	Probabilistic refinement and generative modeling	Slow inference / uncertain calibration	Risk–coverage + latency + external cohort.
SAM/MedSAM	Interactive prompting and transfer	Prompt carries expert information	Automatic vs expert prompt head-to-head.
Brain-specific FM	Few-shot and cross-task adaptation	Pretraining overlap/contamination	Post-cutoff label-efficiency curves.
Missing-modality adaptive fusion	Incomplete clinical MRI	Worst modality subset can still fail	Natural missingness + deliberate stress matrix.
Federated learning	Cross-institution learning without pooling	Client heterogeneity and privacy leakage	Worst-client external benchmark with attack audit.
UQ/selective segmentation	Failure detection and human referral	Calibration drift under site shift	Prospective selective automation reader study.

13.4 Failure-mode atlas and stress tests

Table 10. Failure-mode atlas: stress conditions, the typical failure each induces, the mitigation to be tested and the endpoint required to demonstrate it.

Stress condition	Typical failure	Mitigation to test	Required endpoint
T1ce absent	ET signal weakens or disappears	Train modality dropout; adaptive fusion	ET Dice, lesion sensitivity, volume error
FLAIR absent	Edema/infiltration under-segmented	Missing-modality distillation	WT Dice and worst-combination performance
Motion	Ghosting and blurred boundaries	Motion augmentation + QC gate	HD95, uncertainty, rejection rate
Low resolution / thick slices	Tiny lesions vanish	Resolution augmentation / super-resolution with caution	Size-stratified lesion sensitivity
Scanner/vendor shift	Intensity and texture changes	Domain generalization / normalization	Worst-site Dice and calibration
Misregistration between modalities	False cross-modal correspondences	Registration QC; robust fusion	Boundary error and failure-detection AUROC
Postoperative cavity	Blood/cavity mimics tumor	Dedicated postoperative training	Residual-volume bias and expert correction
Treatment effect	Enhancement not equal progression	Longitudinal/clinical context	Downstream progression agreement
Small metastasis	Global Dice ignores miss	Lesion-aware loss / detection branch	Sensitivity by diameter; FP/scan
Multifocal disease	Component removal deletes real lesions	Audit post-processing	Lesion recall before/after cleanup
Annotation ambiguity	Single mask overstates truth	Multi-rater probabilistic labels	Coverage of expert distribution
Pretraining contamination	Inflated FM transfer results	Post-cutoff hidden benchmark	Performance drop vs contaminated/public benchmark

A useful stress test should be prospective in spirit even when performed retrospectively: define the perturbation before looking at final test performance, freeze the model, and evaluate all systems under exactly the same corruption or missingness. Stress-test severity should be clinically plausible, not chosen only to produce dramatic model failure.

The most informative analysis is often a failure surface rather than a single number. For example, plot ET Dice against slice thickness, or lesion sensitivity against metastasis diameter, or risk–coverage under progressively stronger motion. This reveals whether a method degrades gracefully or fails abruptly.

Foundation models should be included in the same stress tests. Scale does not guarantee robustness. A brain-specific pretrained model may improve low-label performance yet still inherit scanner or source shortcuts from pretraining. Robustness and data efficiency are separate axes.

13.5 Statistical analysis blueprint for future segmentation studies

The independent statistical unit should match the clinical unit. For one scan per patient, patient-level bootstrap is usually appropriate. For longitudinal imaging, resampling should occur by patient so all time points remain together. For multi-site studies, a hierarchical bootstrap or site-stratified analysis can reflect both patient and institutional variation.

Point estimates should be accompanied by confidence intervals, especially for small external cohorts. Multiple seeds are useful for training variance but do not replace patient-level

sampling uncertainty. When comparing two models on the same patients, paired bootstrap differences or paired nonparametric tests are preferable to independent tests.

Multiple tumor regions and multiple models generate multiplicity. A primary endpoint should be prespecified, for example mean ET/TC/WT Dice or lesion-wise sensitivity, while secondary endpoints provide failure characterization. If many pairwise tests are performed, false-discovery control should be considered.

Calibration and selective segmentation need their own statistics. Report calibration error or coverage with confidence intervals, then evaluate whether uncertainty predicts case-level segmentation error. Risk–coverage curves should be summarized at clinically meaningful coverage targets rather than optimized retrospectively on the test set.

For external validation, pooled analysis alone is insufficient. Report site-specific estimates and a worst-site endpoint. If demographics or scanner types are imbalanced, descriptive subgroup analysis should be prioritized over unstable significance testing in tiny subgroups.

Table 11. Statistical analysis blueprint: recommended analyses and the corresponding analyses to avoid.

Design question	Recommended analysis	Avoid
Two models, same test patients	Paired patient bootstrap of metric differences	Independent t-test on slice-level observations
Longitudinal scans	Patient-cluster bootstrap or mixed-effects model	Treating visits as independent patients
Multi-site external validation	Site-stratified estimates + worst-site + pooled CI	Only pooled Dice
Multiple random seeds	Mean/SD across seeds plus fixed-patient test CI	Using seed SD as the only uncertainty
Lesion detection	Patient-clustered lesion sensitivity + FP/scan	Global voxel Dice alone
UQ/selective automation	Risk–coverage, error–uncertainty correlation, calibration	Qualitative uncertainty maps only
Many model comparisons	Prespecified primary endpoint; paired tests; multiplicity control	Selecting best model from test-set ranking

14. Designing a clinically credible reader study

The purpose of a reader study is not to assess the ability of experts to recognize the algorithm but to determine if segmentation assistance improves work performance. To test this scientifically, a well-designed experiment should use a randomized crossover, where neuroradiologists or radiation oncologists would do contouring with and without assistance from AI on specific cases after a washout period. There should be randomization of cases and assistance and blindness of model identity to the readers.

Outcomes could include time to acceptable contours and rate of clinically significant errors. There might also be secondary outcomes of number of edits, volumetric difference, variability between readers, subjective level of trust and downstream conformity with RANO criteria or radiotherapy decisions. In case uncertainty information is used in the study, there should be evaluation if it helps correction rather than just increase perceived sophistication.

For metastases, it is important to have over-sampled small and multifocal cases; for postoperative glioblastoma - to have over-sampled cases with presence of blood products and radiation changes; for meningioma - to consider skull base and en-plaque variants. A benchmark consisting of only easily contoured and large lesions will provide overly optimistic estimates of benefits for workflow. Failure narratives should be collected in the study. Short labels like "missed lesion," "false positive vessel," "cavity confusion," "infiltration margin under-segmentation" and "registration error" may be useful to create a reusable taxonomy.

15. Decision matrix: which model family should be tested for which research question?

Table 12. Decision matrix linking research questions to a priority baseline, candidate advanced model and the critical experimental control.

Research question	Priority baseline	Candidate advanced model	Critical control
Best full-data BraTS performance	nnU-Net	Swin UNETR / CMHF-3DNet / Mamba	Matched training budget and multiple seeds
Low-label adaptation	nnU-Net from scratch	BrainIAC / BrainFound / BrainSegFounder	1/5/10/50-case learning curves
Interactive contouring	nnU-Net automatic	MedSAM / SAM-Med3D / TUM-SAM	Automatic vs expert prompts
Missing modalities	Modality dropout nnU-Net	AdaMM / MGML / Mamba fusion	All modality subsets + natural missingness
Low compute	Compact U-Net / GDNet	TransLiteUNet / FFLUNet	Latency, memory and lesion-size analysis
Uncertainty/referral	Deep ensemble	CONSeg / evidential ensemble	External risk–coverage
Multi-institution privacy	Centralized upper bound	Federated segmentation	Worst-client + privacy/communication audit
Longitudinal response	Dedicated longitudinal CNN	FM + temporal adaptation	Volume repeatability + RANO agreement
Postoperative residual	Post-op nnU-Net	Specialized FM/transformer	Pre-op model as negative control
Contamination-resistant FM test	Scratch nnU-Net	Brain FM / SAM family	Post-cutoff hidden cohort

16. What should a 2026–2030 benchmark actually test?

A future benchmark should be designed around failure modes rather than another static leaderboard. At minimum it should include a locked development cohort, a hidden external-site cohort, a post-cutoff temporal cohort, incomplete-sequence variants and a subset with repeated expert annotations. Foundation models should be evaluated without knowing the hidden cohort identity.

The benchmark should enforce patient/episode independence and provide a reproducible duplicate audit. If longitudinal imaging is included, all time points from the same patient must remain in one partition. Site and scanner metadata should be retained for stratified analysis even if not exposed during model development.

Evaluation should include region Dice/HD95, lesion-wise metrics, volume error, small-lesion strata, worst-site performance, calibration and risk–coverage. Compute should be measured. The primary score should not be a single opaque weighted composite; a Pareto view is more transparent.

For foundation models, the central experiment should compare adaptation regimes under matched label budgets: zero-shot where feasible, linear probing, adapters/LoRA, partial fine-tuning and full fine-tuning versus scratch nnU-Net. Learning curves should be produced at 1, 5, 10, 50 and all labelled patients.

Robustness to missing modalities needs to be studied with respect to all combinations of clinical relevance, including incomplete ones. The method of synthetic reconstruction must be evaluated in terms of the preservation of the downstream pathology and not just the image appearance. Evaluation of uncertainty needs to be done with respect to pristine and shifted scenarios, as well.

17. High-impact research agenda for 2027–2030

Table 13. Proposed high-impact research programs for 2027–2030, with the core experiment defining each.

Research program	Core experiment	Why it could be highly citable
Global prospective hidden-server benchmark	Recruit ≥50 institutions across geographic and economic regions; freeze post-cutoff cases; evaluate average and worst-site performance.	Would create a durable community resource and directly test transportability.
Missing-modality and acquisition-shift challenge	Complete/incomplete T1/T1ce/T2/FLAIR plus motion, low resolution, bias field and scanner shifts; compare direct robust segmentation vs synthesis-first.	Targets the most common real-world deviation from BraTS-like assumptions.
Probabilistic multi-rater glioma dataset	5–10 expert masks on difficult cases; ambiguity tags and adjudication; evaluate probabilistic segmentation and UQ.	Provides ground truth for aleatoric/annotation uncertainty.
Post-cutoff foundation-model challenge	Acquire hidden cases after model pretraining cutoff; compare nnU-Net, brain FMs, SAM-family adaptation and 3D FMs at fixed label budgets.	Directly addresses contamination and tests the claimed value of foundation pretraining.
Longitudinal RANO 2.0 challenge	Serial pre/post-treatment MRI with clinical response, progression and treatment-effect adjudication.	Connects segmentation to clinically meaningful change.
Selective-segmentation benchmark	Compare ensemble, evidential, conformal and Bayesian methods under clean/external/missing-modality conditions.	Turns UQ into an actionable referral policy.
Federated global robustness benchmark	Train without centralizing data; measure privacy, communication, mean/worst-site segmentation and adaptation to underrepresented sites.	Combines privacy with equity and domain robustness.
Human-AI contouring trial	Randomized crossover study of manual vs AI-assisted contouring for radiotherapy/response assessment; capture time, edits and critical errors.	Measures clinical benefit rather than technical overlap alone.

18. Proposed controlled head-to-head experiment

One immediately executable study emerging from this review is a controlled comparison of architecture families under one frozen, leakage-audited manifest. Candidate systems would include nnU-Net, a 2.5D compute-efficient model, Swin UNETR, a Mamba/SSM model, a

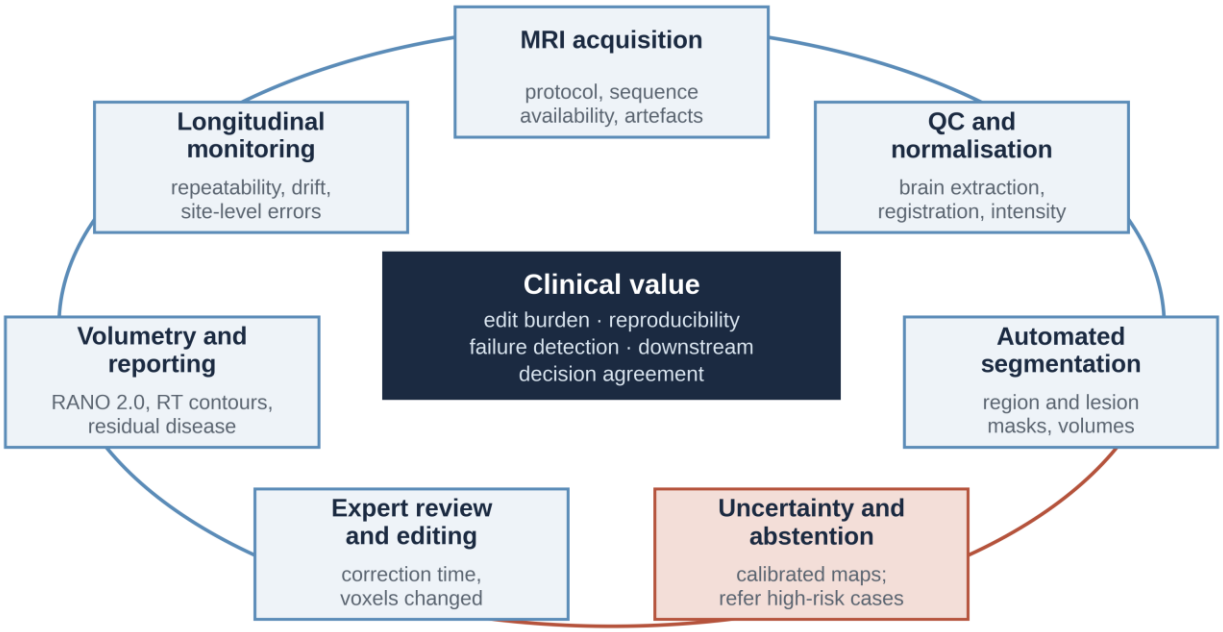
diffusion segmenter, BrainSegFounder or another brain-specific foundation model, and SAM/MedSAM adaptation. Every model would use the same train/validation/test patients and the same preprocessing outputs.

Model selection would be performed on grouped validation only. At least three seeds would be used for stochastic methods. The hidden test would be evaluated once after architecture and thresholds were frozen. An external-site cohort would remain untouched until the internal protocol was complete. Foundation models would disclose pretraining datasets and dates; cases with known overlap would be excluded or analyzed separately.

The primary analysis would report WT/TC/ET Dice and HD95, patient-clustered confidence intervals, parameter count, inference time and peak memory. Secondary analyses would include missing-modality combinations, lesion-size strata, calibration, risk–coverage, site/scanner subgroups and volume error. Rather than nominate a single winner, the study would identify Pareto-optimal models across accuracy, robustness and compute.

The design is more informative compared to the other design since it allows for isolating the exact contributions made by the foundation models, transformers, Mamba, and diffusion in value creation. It also makes negative results publishable such as where nnU-Net is better under full-data while a foundation model is dominant only with five labeled samples.

19. Clinical translation and workflow integration



Dice alone measures none of the endpoints in the centre of this loop

Figure 6. Clinical translation loop for brain tumor segmentation. Automated masks should be integrated with QC, uncertainty, expert editing and downstream longitudinal decisions.

Clinical deployment involves a socio-technical workflow rather than just a model endpoint. The segmentation algorithm needs to be able to integrate into a viewer environment, match to the appropriate imaging sequence, survive pre-processing failure, display uncertainty, facilitate quick fixups, and export the volumes or contours to the appropriate clinical system. Latency and user interface design are the dominant factors in adoption in cases where accuracy is good. The best reader studies will involve realistic tasks and a blinded comparison. Performance metrics could include time until acceptable contour, number of edits, critical misses, inter-reader variation, and downstream impact on response classification. AI assistance outputs need to be compared not only to the ground truth but also to the performance of experts without assistance. Post-deployment monitoring is just as important. Imaging protocols change, software versions change, and patient populations change. A segmentation service needs to monitor its inputs, availability of modalities, distribution of uncertainty, volume distribution, edit burden, and site-specific errors. Significant drift would require re-validation of the model rather than its continuous application. Regulation and governance frameworks put more emphasis on data provenance, intended use, and post-market surveillance. Claims should therefore be limited: a model validated for pre-operative adult gliomas using standardized four sequence MRI is not validated for pediatrics, post-op residual disease, or partial acquisitions.

20. Limitations of the current literature and of this review

However, there is still heterogeneity in the literature. Discrepancies between BraTS version numbers, preprocessing pipelines, access to hidden test sets, label definition, and even implementations of metrics themselves hinder comparability and reduce confidence in rankings across papers. A few papers from 2025–2026 have been either published ahead of print or are available for pre-publication.

External validation is still uncommon relative to internal benchmark reporting. The most recent systematic reviews also identify limited external testing and robustness evaluation as persistent gaps [124–133]. Reported state-of-the-art Dice values should therefore not be equated with expected performance in routine neuro-oncology.

This manuscript is a structured critical review rather than a completed PRISMA systematic review. Although 150 references were curated and checked against publisher, PubMed, proceedings or official records, no screening counts, study-flow diagram or formal risk-of-bias assessment are reported, because the corresponding dual-reviewer procedures were not

performed. A journal requiring PRISMA should add frozen database exports and dual-reviewer screening before submission.

Finally, this review is a synthesis of the published record rather than a re-analysis of primary data. Reported metrics are reproduced as stated by the original authors and were not independently recomputed; where a study reports results on a challenge edition whose hidden test labels are unavailable, the reported value cannot be verified by any third party. Readers should therefore treat the quantitative evidence atlas as a map of what has been claimed and under what evaluation design, not as a validated ranking.

21. Novelty map: twenty research questions with high scientific leverage

The following questions are intentionally framed as falsifiable studies rather than broad future-work statements. Each addresses a limitation that recurs across the 2025–2026 literature and could support a reusable benchmark, dataset, statistical framework or clinical trial rather than a one-off architectural variant.

Table 14. Novelty map: twenty falsifiable research questions and the definitive experiment for each.

#	Research question	Definitive experiment
1	Do brain-specific foundation models retain a few-shot advantage after pretraining-overlap is excluded?	Post-cutoff hidden external cohort; 1/5/10/50-case fine-tuning curves.
2	Which MRI sequence contributes unique information for each tumor subregion?	Conditional modality ablation with Shapley-style contribution analysis and missing-sequence stress tests.
3	Can uncertainty identify segmentation failure before expert review?	Prospective risk–coverage study with prespecified referral thresholds.
4	Does synthetic missing-modality reconstruction preserve pathology?	Compare direct incomplete-input segmentation vs synthesis-first pipelines using lesion volume and boundary endpoints.
5	What is the true value of multimodal fusion under scanner shift?	Train on one institutional mixture, test on vendors/sites excluded from development.
6	Are foundation-model gains caused by stronger pretraining or simply larger compute?	Compute- and parameter-matched adaptation study.
7	Can a lightweight 2.5D model outperform 3D systems on a clinical efficiency Pareto front?	Measure Dice/HD95, memory, latency, energy and editing time.
8	How much benchmark performance is lost under naturally incomplete MRI?	Curate a real incomplete-sequence cohort rather than synthetic channel deletion only.
9	What is the smallest metastasis size at which each method remains reliable?	Lesion-size sensitivity curve with false positives per scan.
10	Does post-processing improve Dice by deleting clinically important small lesions?	Report lesion recall before and after connected-component filtering.
11	Can multi-rater probabilistic labels improve calibration?	Train deterministic vs probabilistic segmentation on 5–10 expert masks per difficult case.
12	How stable are automated volumes across serial scans without biological change?	Test-retest/short-interval repeatability study.
13	Can models distinguish progression from treatment-related change when segmentation is used longitudinally?	Link serial masks to adjudicated RANO 2.0 outcomes.
14	Does federated training improve the worst hospital or only the average client?	Report client-wise and worst-client performance plus communication/privacy cost.
15	Can domain generalization outperform target-site adaptation when no target labels exist?	Prospective unseen-site comparison of normalization, DG and UDA.
16	How robust are SAM/MedSAM results to prompt location and prompt author?	Randomized prompt perturbation and expert-vs-automatic prompt experiment.
17	Do Mamba/state-space gains persist under identical training recipes?	Matched CNN, transformer and SSM benchmark with fixed optimizer/augmentation/compute.

#	Research question	Definitive experiment
18	Do diffusion segmenters provide better calibrated uncertainty than deterministic networks?	Compare deep ensembles, conformal, diffusion and evidential models on external sites.
19	Can a hidden evaluation server prevent adaptive overfitting to BraTS?	Limit submissions; maintain rolling post-cutoff test cohorts.
20	Does AI assistance reduce clinically relevant contouring error and time?	Randomized crossover reader study with edit burden and downstream decision agreement.

One common feature of these studies is the consideration of architecture as just one variable from several others. Such reconsideration changes the motivational aspect, because even if something goes wrong, the result still holds some value in science for pointing out weak spots of the foundation model, missing modality approach, or uncertainty quantification. High-citation synthesis cannot merely predict the universality of the foundation models. Instead, a more reasonable hypothesis is that the foundation models will be more beneficial in situations where there are few labels, the tasks evolve, or reusing representations is required, while self-configured task-specific systems are difficult to replace in situations where labeled data is abundant and the way to obtain them is clear.

The field also needs explicit cost curves. Adding a modality, increasing model size or invoking a diffusion sampler has a price in acquisition time, memory, latency or annotation effort. Future papers should quantify the marginal clinical value of that cost rather than assume that more modalities and larger models are automatically superior.

Finally, research agendas should be designed to survive rapid architectural change. A post-cutoff benchmark, multi-rater dataset, selective-segmentation protocol or reader-study design will remain useful even if today's best encoder is replaced next year. That durability is precisely what can make a review framework highly citable.

22. DATA-SEG author and reviewer checklist

Table 15. DATA-SEG author and reviewer checklist.

ID	Checklist item	Status for a new study
D1	Dataset edition/version and acquisition dates reported	□
D2	Institutions, scanners/field strength and geography described	□
D3	MRI sequence availability and missingness reported	□
A1	Label ontology and annotator expertise reported	□
A2	Consensus/adjudication and ambiguity described	□
T1	Patient-level grouping confirmed	□
T2	Longitudinal scans grouped by patient/episode	□
T3	Duplicate/near-duplicate audit performed	□
T4	Foundation-model pretraining overlap	□

ID	Checklist item	Status for a new study
	considered	
A3	Input dimensionality, parameters and pretraining reported	□
A4	Loss, sampling, augmentation, optimizer and checkpoint rule reported	□
A5	Compute/memory/runtime reported	□
S1	WT/TC/ET or task-appropriate region metrics reported	□
S2	Lesion-wise metrics included when clinically relevant	□
S3	Boundary/volume metrics and patient-level CIs reported	□
S4	Multiple seeds or variance analysis reported	□
E1	Independent external site/scanner evaluated	□
E2	Missing-modality stress test performed	□
E3	Calibration/UQ/risk–coverage evaluated	□
E4	Worst-site/subgroup performance reported	□
G1	Code/weights/split manifest or reproducible equivalent provided	□
G2	Human review/edit burden assessed for clinical claims	□
G3	Intended use and monitoring/drift plan defined	□

23. Conclusions

Brain tumor segmentation has entered a phase in which architectural novelty is no longer the only, or necessarily the main, scientific bottleneck. U-Net descendants, nnU-Net, transformers, Mamba/state-space models, diffusion methods and foundation models can all produce strong benchmark masks. The critical question is whether those masks remain accurate, calibrated and clinically useful when the patient, site, scanner, sequence set or tumor phenotype changes.

Recent 2025–2026 work points to a more mature research agenda. Missing-modality methods explicitly model absent sequences; brain-specific foundation models prioritize data efficiency and transfer; uncertainty methods support referral; metastasis and meningioma studies increasingly use external cohorts and clinical endpoints. These developments should be consolidated through better benchmark design rather than fragmented into isolated architectural competitions.

DATA-SEG summarizes the proposed standard: document dataset provenance and annotation, prove train/test independence, report architecture and compute transparently, use patient-aware statistics and clinically appropriate metrics, stress-test external robustness and uncertainty, and connect outputs to governance and workflow. A segmentation model should

be considered robust only after plausible shortcuts and acquisition failures have been deliberately tested.

The highest-impact next step is a post-cutoff, multi-institutional, hidden-server benchmark that compares strong task-specific baselines with foundation-model adaptation under identical label and compute budgets, includes incomplete MRI and multi-rater uncertainty, and measures lesion-wise and longitudinal clinical outcomes. Such a benchmark would move the field beyond BraTS leaderboard optimization toward reliable neuro-oncology AI.

Declarations

Author contributions. B.K.K. conceived the review, designed the evidence-extraction framework and the DATA-SEG reporting framework, and drafted the manuscript. S.K. and M.J.R. contributed to reference screening and verification and to the quantitative evidence atlas. All authors reviewed, edited and approved the final manuscript and agree to be accountable for its content.

Funding. [State funding sources and grant numbers, or: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.]

Conflicts of interest. [State any competing financial or non-financial interests, or: The authors declare no competing interests.]

Ethics approval and consent. Not applicable. This review analyses previously published studies and publicly available benchmark descriptions; no new human or animal data were collected, and no identifiable patient data were accessed.

Data availability. No new data were generated. All studies analysed are cited in the reference list.

Use of generative AI. [Declare here, in the form required by the target journal, any use of generative artificial intelligence or AI-assisted technologies in the preparation of this manuscript, together with the tool, version and purpose. The authors take full responsibility for the content, for the accuracy of every cited reference, and for the integrity of the manuscript.]

Self-citation disclosure. Four cited studies are the authors' own work [15–18]. They are discussed alongside competing methods, are not used as privileged evidence, and are subject to the same limitations applied throughout this review.

References

1. Menze BH, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans Med Imaging. 2015;34(10):1993-2024. https://doi.org/10.1109/TMI.2014.2377694
2. Bakas S, 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
3. Bakas S, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv. 2018;1811.02629.
4. Baid U, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv. 2021;2107.02314.
5. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. MICCAI. 2015:234-241. https://doi.org/10.1007/978-3-319-24574-4_28
6. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. MICCAI. 2016:424-432. https://doi.org/10.1007/978-3-319-46723-8_49
7. Milletari F, Navab N, Ahmadi SA. V-Net: fully convolutional neural networks for volumetric medical image segmentation. 3DV. 2016:565-571. https://doi.org/10.1109/3DV.2016.79
8. Kamnitsas K, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal. 2017;36:61-78. https://doi.org/10.1016/j.media.2016.10.004
9. Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. BrainLes/MICCAI. 2018:311-320. https://doi.org/10.1007/978-3-030-11726-9_28
10. 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
11. Wang W, Chen C, Ding M, Li J, Yu H, Zha S. TransBTS: multimodal brain tumor segmentation using Transformer. MICCAI. 2021:109-119. https://doi.org/10.1007/978-3-030-87193-2_11
12. Hatamizadeh A, Tang Y, Nath V, et al. UNETR: Transformers for 3D medical image segmentation. WACV. 2022:574-584.
13. Hatamizadeh A, Nath V, Tang Y, et al. Swin UNETR: Swin Transformers for semantic segmentation of brain tumors in MRI images. BrainLes/MICCAI. 2022.
14. Zhou H-Y, Guo J, Zhang Y, et al. nnFormer: volumetric medical image segmentation via a 3D Transformer. IEEE Trans Image Process. 2023;32:4036-4045.
15. Kiani Kalejahi B, Meshgini S, Danishvar S. Segmentation of Brain Tumor Using a 3D Generative Adversarial Network. Diagnostics. 2023;13(21):3344. https://doi.org/10.3390/diagnostics13213344
16. Kiani Kalejahi B, Meshgini S, Danishvar S. Brain tumor segmentation by auxiliary classifier generative adversarial network. Signal Image Video Process. 2023;17(7):3339-3345. https://doi.org/10.1007/s11760-023-02555-6
17. Kiani Kalejahi B, Khan S, Rajabi MJ. GDNet: A Robust 2.5D Multimodal MRI Brain Tumor Segmentation Framework with EMA Stabilization and Tumor-Aware Sampling. J Imaging. 2026;12(7):288. https://doi.org/10.3390/jimaging12070288
18. Kiani Kalejahi B, Danishvar S, Rajabi MJ. GAN-Based Cross-Modality Brain MRI Synthesis: Paired Versus Unpaired Training and Comparison with Diffusion and Transformer Models. Biomimetics. 2026;11(3):175. https://doi.org/10.3390/biomimetics11030175
19. Cox B, et al. BrainSegFounder: towards 3D foundation models for neuroimage segmentation. Med Image Anal. 2024;97:103301. https://doi.org/10.1016/j.media.2024.103301
20. Tak D, Garomsa BA, Zapaishchykova A, et al. A generalizable foundation model for analysis of human brain MRI. Nat Neurosci. 2026;29(4):945-956. https://doi.org/10.1038/s41593-026-02202-6
21. Mazher M, Parker GJM, Alexander DC. Towards generalisable foundation models for brain MRI. npj Imaging. 2026. https://doi.org/10.1038/s44303-026-00176-5
22. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun. 2024;15:654. https://doi.org/10.1038/s41467-024-44824-z
23. Kirillov A, Mintun E, Ravi N, et al. Segment Anything. ICCV. 2023:4015-4026.
24. Wang H, et al. SAM-Med3D: a vision foundation model for general-purpose segmentation on volumetric medical images. IEEE Trans Neural Netw Learn Syst. 2025;36:17599-17612.
25. Zhang S, et al. A generalist foundation model and database for open-world medical image segmentation. Nat Biomed Eng. 2026;10:1026-1041.

26. Sun Y, Wang L, Li G, Lin W, Wang L. A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. Nat Biomed Eng. 2025;9:521-538. https://doi.org/10.1038/s41551-024-01283-7
27. Chen Z, Agarwal D, Aggarwal K, Safta W, Balan MM, Brown K. Masked image modeling advances 3D medical image analysis. WACV. 2023:1970-1980. https://doi.org/10.1109/WACV56688.2023.00201
28. Zhou Z, Sodha V, Rahman Siddiquee MM, et al. Models Genesis: generic autodidactic models for 3D medical image analysis. Med Image Anal. 2021;67:101840. https://doi.org/10.1016/j.media.2020.101840
29. Tang Y, Yang D, Li W, et al. Self-supervised pre-training of Swin Transformers for 3D medical image analysis. CVPR. 2022:20730-20740.
30. Zhao T, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. Nat Methods. 2025;22:166-176. https://doi.org/10.1038/s41592-024-02499-w
31. Peiris H, et al. Segment any tumour: an uncertainty-aware vision foundation model for whole-body analysis. Nat Commun. 2026;17:9640. https://doi.org/10.1038/s41467-026-76531-2
32. Perez-Garcia F, et al. Exploring scalable medical image encoders beyond text supervision. Nat Mach Intell. 2025;7:119-130.
33. He Y, et al. VISTA3D: a unified segmentation foundation model for 3D medical imaging. CVPR. 2025:20863-20873.
34. Cheng J, et al. SAM-Med2D. arXiv. 2023;2308.16184.
35. Koleilat T, Asgariandehkordi H, Rivaz H, Xiao Y. MedCLIP-SAMv2: towards universal text-driven medical image segmentation. Med Image Anal. 2025;106:103749.
36. Taher MRH, Gotway MB, Liang J. Towards foundation models learned from anatomy in medical imaging via self-supervision. DART/MICCAI Workshop. 2024:94-104.
37. Zhu S, Chen Y, Chen W, et al. No modality left behind: Adapting to missing modalities via knowledge distillation for brain tumor segmentation. Med Image Anal. 2026;112:104108. https://doi.org/10.1016/j.media.2026.104108
38. Xi H, Ye Y, Zhu J, Li J. Multimodal Contrastive Prototype Learning for Resilient Brain Tumor Segmentation With Missing Modalities. IEEE J Biomed Health Inform. 2026;30(8):6841-6853. https://doi.org/10.1109/JBHI.2025.3649819
39. Ren X, Shi Y, Yuan S, Cheng J, Zhu Z, Liu Y. LFC-SATP-SGG: Latent Feature Completion and State-Aware Text Prompting with Semantic-Guided Gating for Incomplete Multi-Modality Brain Tumor Segmentation. IEEE J Biomed Health Inform. 2026. https://doi.org/10.1109/JBHI.2026.3720675
40. Liu C, Li XL, Xu D, Wang H, Jiang J. Mamba-based brain tumor segmentation of incomplete multi-modal MR images. Quant Imaging Med Surg. 2026;16(2):142. https://doi.org/10.21037/qims-2025-1913
41. Zou Y, Liu B, Zheng CJ, et al. MGML: A plug-and-play meta-guided multi-modal learning framework for incomplete multimodal brain tumor segmentation. Neural Netw. 2026;197:108525. https://doi.org/10.1016/j.neunet.2025.108525
42. Ting H, Liu M. Multimodal Transformer of Incomplete MRI Data for Brain Tumor Segmentation. IEEE J Biomed Health Inform. 2023. https://doi.org/10.1109/JBHI.2023.3286689
43. Li Z, et al. Joint learning-based feature reconstruction and enhanced network for incomplete multi-modal brain tumor segmentation. Comput Biol Med. 2023;162:107234. https://doi.org/10.1016/j.compbiomed.2023.107234
44. Zhang Y, et al. A literature survey of MR-based brain tumor segmentation with missing modalities. Comput Biol Med. 2023;152:106339.
45. Zhou W, Shi Y, Fang M, Yue G. Decouple, Collaborate, Match: Prototype-Driven Cognitive Mutual Learning for Brain Tumor Segmentation. IEEE J Biomed Health Inform. 2026. https://doi.org/10.1109/JBHI.2026.3701696
46. Does Adding a Modality Really Make Positive Impacts in Incomplete Multi-Modal Brain Tumor Segmentation? IEEE Trans Med Imaging. 2025;44(5):2194-2205. https://doi.org/10.1109/TMI.2025.3526818
47. Progressive Distillation With Optimal Transport for Federated Incomplete Multi-Modal Learning of Brain Tumor Segmentation. IEEE J Biomed Health Inform. 2025;29(5):3159-3170. https://doi.org/10.1109/JBHI.2025.3525854
48. MedMAP: Promoting Incomplete Multi-Modal Brain Tumor Segmentation With Alignment. 2025. PMID:40833905.
49. Unsupervised generative AI for enhancing brain tumor segmentation in multi-center, incomplete real-world data scenarios. 2025. PMID:41310177.
50. SGAFNet: Robust brain tumor segmentation via learnable sequence-guided adaptive fusion in available MRI acquisitions. Comput Med Imaging Graph. 2026. https://doi.org/10.1016/j.compmedimag.2026.102703
51. BTSegDiff: Brain tumor segmentation based on multimodal MRI dynamically guided diffusion probability model. Comput Biol Med. 2025;186:109694. https://doi.org/10.1016/j.compbiomed.2025.109694
52. DiffBTS: A Lightweight Diffusion Model for 3D Multimodal Brain Tumor Segmentation. Sensors. 2025;25(10):2985. https://doi.org/10.3390/s25102985

53. Conditional diffusion model for high-accuracy brain tumor segmentation in MRI images. Sci Rep. 2025. https://doi.org/10.1038/s41598-025-25137-7

54. SF-Diff: Diffusion network with spatial channel attention infusion and frequency spatial attention for brain tumor segmentation. Med Phys. 2025. https://doi.org/10.1002/mp.17482

55. Brain tumor segmentation using synthetic MR images - A comparison of GANs and diffusion models. Sci Data. 2024. https://doi.org/10.1038/s41597-024-03073-x

56. MambaDiff: A Mamba-based diffusion model for medical image segmentation. IEEE Trans Image Process. 2025;34:5761-5775. https://doi.org/10.1109/TIP.2025.3607615

57. WAS-Mamba: state-space modeling for medical image segmentation. IEEE Trans Image Process. 2026;35:8152-8166. https://doi.org/10.1109/TIP.2026.3691019

58. MamTrans: attention and state-space modeling for brain tumor segmentation. 2025. PMID:40606374.

59. Dilated convolution V-Net with Transformer for brain tumor segmentation. Comput Biol Med. 2025. https://doi.org/10.1016/j.compbiomed.2025.111210

60. LCBTS-Net: lightweight contextual brain tumor segmentation network. Comput Med Imaging Graph. 2026. https://doi.org/10.1016/j.compmedimag.2026.102727

61. Kohl SAA, Romera-Paredes B, Meyer C, et al. A probabilistic U-Net for segmentation of ambiguous images. NeurIPS. 2018;31:6965-6975.

62. FFLUNet: lightweight full-frequency learning U-Net for brain tumor segmentation. 2025. PMID:40517595.

63. LATUP-Net: lightweight attention-based tumor segmentation network. Comput Biol Med. 2025;184:109353. https://doi.org/10.1016/j.compbiomed.2024.109353

64. LIU-NET: lightweight improved U-Net for brain tumor segmentation. PeerJ Comput Sci. 2025;11:e2787. https://doi.org/10.7717/peerj-cs.2787

65. FALS-YOLO for brain tumor segmentation and classification. Sensors. 2025;25(19):5993. https://doi.org/10.3390/s25195993

66. Guennoun Y, Nedelec P, McArthur M, et al. Segmenting with confidence through uncertainty quantification for brain tumor imaging. npj Digit Med. 2026. https://doi.org/10.1038/s41746-026-02902-0

67. CONSeg: Voxelwise Uncertainty Quantification for Glioma Segmentation Using Conformal Prediction. 2025. PMID:40610235.

68. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. ICML. 2016:1050-1059.

69. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. NeurIPS. 2017;30.

70. Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? NeurIPS. 2017;30.

71. Huang L, Ruan S, Xing Y, Feng M. A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods. Med Image Anal. 2024;97:103223. https://doi.org/10.1016/j.media.2024.103223

72. Ziaee SS, Ovens K. Uncertainty Quantification in Medical Image Segmentation: A Comprehensive Survey. J Imaging. 2026;12(8):341. https://doi.org/10.3390/jimaging12080341

73. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. ICML. 2017:1321-1330.

74. Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. 2021. arXiv:2107.07511.

75. Wen J, Li X, Ye X, et al. A highly generalized federated learning algorithm for brain tumor segmentation. Sci Rep. 2025;15:21053. https://doi.org/10.1038/s41598-025-05297-2

76. Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. Brainlesion/MICCAI. 2019.

77. Sheller MJ, Edwards B, Reina GA, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep. 2020;10:12598. https://doi.org/10.1038/s41598-020-69250-1

78. Li X, Gu Y, Dvornek N, Staib LH, Ventola P, Duncan JS. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Med Image Anal. 2020;65:101765.

79. Karargyris A, et al. Federated benchmarking of medical artificial intelligence with MedPerf. Nat Mach Intell. 2023;5:799-810.

80. UDA-GS: unsupervised domain adaptation for cross-center multimodal brain tumor segmentation. Comput Biol Med. 2025;109472. https://doi.org/10.1016/j.compbiomed.2024.109472

81. DA-MLM: domain-adaptive MRI learning for central nervous system tumors. 2026. PMID:41595769.

82. Glocker B, Robinson R, Castro DC, Dou Q, Konukoglu E. Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects. 2019. arXiv:1910.04597.

83. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15:e1002683.

84. Roberts M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3:199-217.

85. Hill BG, Koback FL, Schilling PL. The risk of shortcutting in deep learning algorithms for medical imaging research. Sci Rep. 2024;14:29224. https://doi.org/10.1038/s41598-024-79838-6

86. Ross H, et al. Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. npj Digit Med. 2024;7:124. https://doi.org/10.1038/s41746-024-01118-4

87. Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. Nat Med. 2024;30:2838-2848. https://doi.org/10.1038/s41591-024-03113-4

88. Pulido-Arias D, et al. Domain Generalization Mitigates Scanner-Induced Domain Shift in Medical Imaging. J Imaging Inform Med. 2026. https://doi.org/10.1007/s10278-026-02082-z

89. Adewole M, et al. Brain tumor segmentation in Sub-Saharan Africa patient population: the BraTS-Africa challenge. Neurooncol Adv. 2026;8(1):vdag082. https://doi.org/10.1093/noajnl/vdag082

90. BraTS-PEDs 2023 challenge report: pediatric brain tumor segmentation in multi-parametric MRI. 2023. PMID:37292481.

91. BraTS 2023 meningioma segmentation challenge report. 2023. PMID:37608937.

92. BraTS 2023 brain metastases segmentation challenge report. 2023. PMID:37396600.

93. BraSyn 2023: missing MRI synthesis challenge report. 2023. PMID:37608932.

94. Baid U, et al. The UPenn-GBM dataset: advanced MRI, clinical, genomics, and radiomics. Sci Data. 2022. PMID:35906241.

95. BraTS-MEN-RT: a multi-institutional MRI dataset for radiotherapy planning in meningioma. Sci Data. 2026;13:306. https://doi.org/10.1038/s41597-026-06649-x

96. Chrysochoou D, et al. AI-powered segmentation and prognosis with missing MRI in pediatric brain tumors. npj Precis Oncol. 2026;10(1):63. https://doi.org/10.1038/s41698-025-01269-x

97. Development and routine implementation of deep learning algorithm for automatic brain metastases segmentation on MRI for RANO-BM criteria follow-up. Neuroimage. 2025;306:121002. https://doi.org/10.1016/j.neuroimage.2025.121002

98. A Data-Centric Approach to Deep Learning for Brain Metastasis Analysis at MRI. 2025. PMID:40552999.

99. Blinded, bias-controlled multi-rater evaluation of human-versus-AI brain metastasis segmentation using a hybrid foundation-model framework. 2026. PMID:42329664.

100. Longitudinal deep learning segmentation of brain metastases with outcome-linked evaluation. 2026. PMID:41527513.

101. Deep Learning for Automated Detection and Segmentation of Meningioma on Multiparametric MRI. 2026. PMID:42711640.

102. Automated deep learning-based segmentation and volumetric analysis of meningiomas. 2026. PMID:42622848.

103. CNN-based meningioma segmentation: systematic review and meta-analysis. 2025. PMID:39777602.

104. Deep learning-based postoperative glioblastoma segmentation and extent of resection evaluation: development, external validation, and model comparison. Neurooncol Adv. 2024. https://doi.org/10.1093/noajnl/vdae199

105. Challenges in Building of Deep Learning Models for Glioblastoma Segmentation: Evidence from Clinical Data. 2021. PMID:34042753.

106. Wen PY, 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.

107. Lin NU, et al. Response assessment criteria for brain metastases: proposal from the RANO group. Lancet Oncol. 2015;16:e270-e278.

108. Ostrom QT, et al. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States. Neuro Oncol. 2025.

109. 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.

110. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging. 2010;29:1310-1320.

111. Avants BB, Tustison NJ, Song G, et al. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage. 2011;54:2033-2044.

112. Smith SM. Fast robust automated brain extraction. Hum Brain Mapp. 2002;17:143-155.

113. Hoopes A, Mora JS, Dalca AV, Fischl B, Hoffmann M. SynthStrip: skull-stripping for any brain image. Neuroimage. 2022;260:119474.

114. Nyúl LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. IEEE Trans Med Imaging. 2000;19:143-150.

115. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6:60.

116. Baumgartner CF, Tezcan KC, Chaitanya K, et al. PHiSeg: capturing uncertainty in medical image segmentation. MICCAI. 2019:119-127. https://doi.org/10.1007/978-3-030-32245-8_14

117. Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. DLMIA. 2017:240-248.

118. Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. MLMI. 2017:379-387.

119. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. ICCV. 2017:2980-2988.

120. Maier-Hein L, et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024;21:195-212.

121. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15:29.

122. Reinke A, et al. Common limitations of image processing metrics: a picture story. arXiv. 2021;2104.05642.

123. Kofler F, et al. BraTS Toolkit: translating BraTS brain tumor segmentation algorithms into clinical and scientific practice. Front Neurosci. 2020;14:125.

124. Zhou et al. Deep learning for multimodal brain tumor segmentation: Architectures, fusion, robust learning, and deployment perspectives. Comput Med Imaging Graph. 2026;134:102807. https://doi.org/10.1016/j.compmedimag.2026.102807

125. From CNNs to diffusion models: a decade of advances in brain tumor and stroke lesion segmentation across BraTS, ATLAS, and ISLES benchmarks. Biomed Signal Process Control. 2026;118:109780. https://doi.org/10.1016/j.bspc.2026.109780

126. Recent advances in brain tumor segmentation: A review of self-supervised learning, federated learning, and graph neural networks. Eng Appl Artif Intell. 2026;173:114473. https://doi.org/10.1016/j.engappai.2026.114473

127. Deep learning for brain tumor segmentation in multimodal MRI images: A review of methods and advances. Image Vis Comput. 2025;156:105463. https://doi.org/10.1016/j.imavis.2025.105463

128. Verma R, Yadav A. Brain tumor segmentation with deep learning: Current approaches and future perspectives. J Neurosci Methods. 2025;418:110424. https://doi.org/10.1016/j.jneumeth.2025.110424

129. Jin et al. Transformer-based architectures in MRI brain tumor segmentation: A review. Comput Med Imaging Graph. 2026;129:102729. https://doi.org/10.1016/j.compmedimag.2026.102729

130. Transformer-based neuroimage segmentation: systematic review. J Med Internet Res. 2025;27:e57723. https://doi.org/10.2196/57723

131. Advancing Precision: A Comprehensive Review of MRI Segmentation Datasets from BraTS Challenges (2012-2025). Sensors. 2025;25(6):1838.

132. Almudaimeegh et al. Brain Tumor Segmentation and Grading on MRI Using Deep Learning: A Systematic Literature Review and Benchmark-Driven Comparative Analysis. Diagnostics. 2026;16(17):2806. https://doi.org/10.3390/diagnostics16172806

133. A Systematic Review of Machine Learning Approaches for MRI-Based Brain Tumor Analysis. Front Comput Neurosci. 2026.

134. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell. 2020;2:e200029.

135. Tejani AS, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiol Artif Intell. 2024;6:e240300.

136. Collins GS, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.

137. Lekadir K, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388:e081554.

138. Sounderajah V, et al. Developing the STARD-AI reporting guideline for diagnostic accuracy studies of AI. Nat Med. 2025;31:3283-3289.

139. Liu X, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: CONSORT-AI extension. Nat Med. 2020;26:1364-1374.

140. Cruz Rivera S, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: SPIRIT-AI extension. Nat Med. 2020;26:1351-1363.

141. Zhang S, Metaxas D. On the challenges and perspectives of foundation models for medical image analysis. Med Image Anal. 2024;91:102996.

142. Awais M, et al. Foundation models defining a new era in vision: a survey and outlook. IEEE Trans Pattern Anal Mach Intell. 2025;47:2245-2264. https://doi.org/10.1109/TPAMI.2024.3506283

143. D'Antonoli TA, et al. Foundation models for radiology: fundamentals, applications, opportunities, challenges, risks, and prospects. Diagn Interv Radiol. 2026. https://doi.org/10.4274/dir.2025.253445

144. Bian Y, Li J, Ye C, Jia X, Yang Q. Artificial intelligence in medical imaging: from task-specific models to large-scale foundation models. Chin Med J. 2025;138:651-663.

145. Hussain A, Jungshen W, Turab A, Gilani Y, Ahmed KI, Wajahat A. CMHF-3DNet: A Transformer-Based Framework for Improved Brain Tumor Segmentation Across Modalities. Acad Radiol. 2026;33(9):4042-4059. https://doi.org/10.1016/j.acra.2026.04.045

146. Chen Y, Tang T, Kim T, Shu H. UKAN-EP: enhancing U-KAN with efficient attention and pyramid aggregation for 3D multi-modal MRI brain tumor segmentation. BMC Med Imaging. 2025;25:517. https://doi.org/10.1186/s12880-025-02053-w

147. Zhou L, Yang Y, Yang Y. TransLiteUNet: A Lightweight CNN-Transformer Hybrid for Efficient 3D Brain Tumor Segmentation with Sub-0.5 M Parameters. J Imaging. 2026;12(7):290. https://doi.org/10.3390/jimaging12070290

148. Yang Z, DSouza N, Megyeri I, et al. Decipher-MR: a vision-language foundation model for 3D MRI representations. npj Digit Med. 2026;9:419. https://doi.org/10.1038/s41746-026-02596-4

149. Li J, Liu R, Xing Y, Gao X, Yin Q, Su Q. Foundation Model Based on Routine Magnetic Resonance Imaging for Brain Tumor Molecular Profiling and Progression Prediction. JCO Precis Oncol. 2026;10:e2500930. https://doi.org/10.1200/PO-25-00930

150. Wang S, Safari M, Li Q, et al. Vision foundation model for 3D magnetic resonance imaging segmentation, classification, and registration. Med Image Anal. 2026;110:103992. https://doi.org/10.1016/j.media.2026.103992