Cureus
Part of SPRINGER NATURE

Open Access Review Article

# Clinical Translation of Automated MRI Segmentation and Deep Learning With Small Datasets in Meningioma

Fengzhang Zhu [1], Hao Ma [1], Heji Ma [1], Shizhou Tang [1], Xianglin Zhang [1], Hui Guo [1]

1. Radiology, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, CHN

**Corresponding author:** Hui Guo, huiguo198509@163.com

**Review began** 10/26/2026
**Review ended** 11/07/2026
**Published** 11/17/2026

**© Copyright** 2026
Zhu et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**DOI:** 10.7759/cureus.119844

## Abstract

Automated segmentation and prediction models have advanced meningioma magnetic resonance imaging (MRI) analysis, but complex procedures and limited case numbers continue to constrain local research and implementation by physicians. Using World Health Organization (WHO) grade prediction as a clinical example, this narrative review examines international and Chinese research on automated segmentation, image pretraining, and tabular foundation models. It compares the roles of convolutional neural networks, vision Transformers, and tabular prior-data fitted network (TabPFN) in small-data settings, focusing on the connections between lesion segmentation, local model development, and physician diagnosis. Current evidence indicates that knowledge reuse can reduce annotation and representation-learning demands, while integrated software can reduce repetitive processing. Clinical value depends on the intended use of contours, outcome labels, validation of the complete workflow, and diagnostic gains when physicians use the outputs. Future research should embed physician review and quality control within automated workflows and assess contour-editing burden, prediction reliability, and the value of the information for clinical care. These priorities can help physicians explore local radiomics models and integrate validated tools into routine practice.

**Categories:** Neurology, Radiology, Oncology
**Keywords:** automated segmentation, clinical translation, deep learning, magnetic resonance imaging, meningioma, small-sample learning, tabular foundation models

## Introduction And Background

Intelligent analysis of meningioma magnetic resonance imaging (MRI) is expanding from lesion detection and grade prediction to automated segmentation, quantitative imaging, and diagnostic support. Extensive algorithm development and performance evaluation provide a basis for considering clinical translation. Can physicians organize these separate technologies into a workflow that supports local research and diagnosis? Horvat et al. identified complex multistep processing, inconvenient interfaces, and inadequate standardization as important barriers to the clinical adoption of radiomics [1]. The research value of routine MRI therefore depends on the additional trustworthy information it can provide, the extra work required to obtain that information, and the clinical judgments it can support.

These issues are particularly important when datasets are small. With limited case numbers, annotation from scratch, model training, and repeated experimentation may demand more resources than a physician team can commit. Clinical prediction methodology distinguishes developing a new model from evaluating or updating an existing one, providing a basis for deciding what genuinely needs to be learned locally [2]. This narrative review follows the progression from reusing existing knowledge through local validation to returning results to diagnostic work. It examines international research, English-language studies from China, and their implications, using World Health Organization (WHO) grading as a recurring example. The focus is on how automated software can connect segmentation with small-sample modeling and which clinical problems remain unresolved along this pathway. Figure *1* summarizes the connection between automated segmentation, local modeling, and physician diagnosis.

**How to cite this article**
Zhu F, Ma H, Ma H, et al. (November 17, 2026) Clinical Translation of Automated MRI Segmentation and Deep Learning With Small Datasets in Meningioma. Cureus 18(11): e119844. DOI 10.7759/cureus.119844

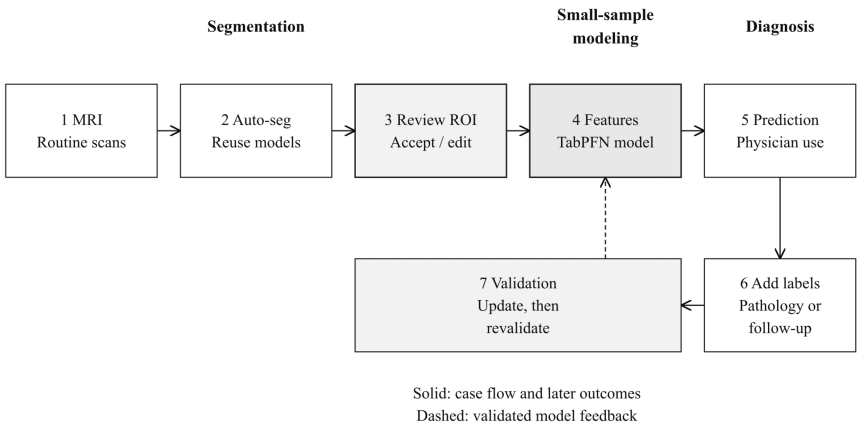



**FIGURE 1: Connecting segmentation, small-sample modeling, and diagnosis in meningioma MRI.**

After physician review and any required ROI correction, image-derived features enter a fixed, validated model for new-case prediction; TabPFN illustrates this pathway. Pathology or follow-up labels are appended to the local database when available, supporting independent evaluation and subsequent local development or updating. Updated versions require revalidation before use (dashed feedback). ROI, region of interest; TabPFN, tabular prior-data fitted network. Author-created conceptual workflow.

# Review

## International research

*Public Data and the Clinical Purpose of Automated Segmentation*

Public meningioma data have progressed from small, single-center collections to multi-institutional segmentation resources. LaBella et al. released 1,141 preoperative MRI examinations, with WHO grade available for 800 examinations; their background review describes approximately 80% of meningiomas as grade 1 [3]. Public data scarcity should therefore be understood in relation to the task: learning contours is different from learning a locally relevant outcome with reliable follow-up. In this dataset, skull stripping excluded extracranial tumor components [3]. For skull-base or invasive lesions, preprocessing therefore helps define the model's scope of applicability. Public resources can reduce the burden of learning anatomy without supplying every label needed for local diagnosis or prognosis.

Segmentation frameworks lower the barrier to repeatedly designing algorithms. nnU-Net organizes preprocessing, network configuration, training, and postprocessing according to dataset characteristics, allowing multiple experience-based development decisions to be reused [4]. Reusing a framework provides a mature training process, whereas reusing weights directly transfers knowledge learned from previous lesions. MedSAM uses large-scale medical image training to provide prompt-driven segmentation and can support contour generation when disease-specific annotations are limited; generating those prompts still requires human effort [5]. MedSegX incorporates cross-task adaptation and out-of-distribution evaluation into general-purpose segmentation research, reflecting a shift from training individual tasks toward reusing segmentation knowledge [6]. These approaches offer different degrees of automation, and their value should be interpreted in terms of the work remaining after prompting, review, and correction.

Segmentation targets are also moving from identifying image contents toward providing the contours required for a clinical purpose. Published in 2026, the BraTS-MEN-RT dataset addresses radiotherapy and uses annotation conventions that differ from those used for preoperative tumor subregions [7]. Enhancing tumor, the entire lesion, and radiotherapy target volumes serve different purposes; their meaning should be checked before numerical performance is compared. Metrics Reloaded recommends selecting evaluation metrics according to the target structure, data characteristics, and intended use [8]. Volume overlap, boundary error, and missed small lesions address different questions. A single Dice coefficient therefore provides an incomplete account of how readily a physician can use a contour.

*Small-Sample Learning Through Pretrained Knowledge*

The feasibility of research with small datasets increasingly depends on pretrained knowledge. RadImageNet showed that the source of pretraining can influence downstream transfer performance [9]. BrainIAC uses self-supervised brain MRI pretraining and evaluates frozen representations, fine-tuning, and adaptation with limited labeled data across several tasks [10]. More specifically, Tabassum et al. transferred glioma

segmentation knowledge to meningioma and metastasis segmentation using a meta-transfer nnU-Net approach [11]. Their selected cohort still included 320 meningioma cases and used a single training-validation-test partition. This is evidence for transferring representations across tumor types, rather than proof that a handful of local cases is sufficient for a dependable clinical system.

The tabular prior-data fitted network (TabPFN) extends this approach to tabular data. Hollmann et al. pretrained a Transformer on synthetic tabular tasks, enabling predictions through in-context learning once labeled examples are supplied [12]. In radiomics, the inputs can comprise processed imaging features and clinical variables, creating an opportunity to reduce repeated training and extensive parameter tuning. A key distinction from a convolutional neural network (CNN) is the input: CNNs learn spatial representations from images, whereas TabPFN processes an already constructed feature table. Feature quality, selection procedures, and label credibility consequently remain determinants of TabPFN output. Pretrained priors cannot supply local outcome information that has not been collected.

Direct comparisons in meningioma are particularly informative. Huryol et al. studied 455 cases, including an external cohort of 109, and used TabPFN for radiomics and clinical feature tables. External AUCs were 0.628 for radiomics, 0.645 for clinical variables, and 0.732 for the nnU-Net pathway; the best combined model reached 0.742 and did not significantly outperform nnU-Net ($p = 0.737$) [13]. Thus, a convenient tabular pathway can remain limited by the information and transportability of its input features. These results do not establish a universal ranking of architectures. Manual segmentation was still used to define cropping and peritumoral regions, so the study also illustrates work that remained outside the learned predictors. Knowledge reuse should be assessed by the work it replaces and by independently tested predictions.

*Clinical Evaluation of Model Performance in Use*

International research has clarified the evidence required for translation. A systematic evaluation of the Radiomics Quality Score found that the literature remains dominated by model development, with limited high-level evidence supporting clinical translation [14]. The 2024 update of the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) emphasizes transparent reporting of imaging artificial intelligence (AI) data, models, and evaluation [15]. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis guidelines for regression and AI methods (TRIPOD+AI) require sufficient information to understand and reuse prediction models, including their implementation and conditions of use [16]. These standards improve the scrutiny of research and make the gap between high performance and independent validation or use easier to identify.

Further translation depends on human-AI collaboration. DECIDE-AI places early clinical evaluation in the context of actual use, examining interaction between physicians and systems and its effects on clinical care [17]. FUTURE-AI includes usability, robustness, traceability, and postdeployment monitoring in its lifecycle recommendations [18]. Representative data, physician oversight, and workflow integration are therefore shared requirements for the reuse of foundation models. The object of automation research is consequently expanding from individual models to the diagnostic processes in which those models participate.

## Research in China

*Clinical Meaning in Segmentation and Annotation*

Chinese-affiliated groups have contributed both general segmentation tools and disease-specific applications. MedSegX explicitly addresses negative transfer and evaluates out-of-distribution segmentation, recognizing that a model can lose useful knowledge when tasks or data sources differ [6]. For meningioma, Yang et al. used SegResNet on routine contrast-enhanced T1-weighted MRI and linked automated contours to radiomics grading [19]. This direction brings contour generation closer to the analysis physicians want to perform. A model's generality and its suitability for a particular lesion boundary remain separate questions, especially around the skull base, adjacent dura, and tumor-brain interface.

Clinical annotation meaning must be retained when public models are reused. CLAIM 2024 asks investigators to describe acquisition protocols, preprocessing, reference-standard definitions, annotator qualifications, and test-set annotation procedures [15]. It is a reporting guideline for imaging AI, not a radiomics-specific scoring tool. For local workflow design, these requirements identify information that should accompany each image and contour. Scanner and sequence differences, lesion extent, and the treatment context should be documented together. They cannot be reconstructed reliably from a mask file alone. A local database is therefore a linked record of images, reviewed contours, clinical questions, and subsequently confirmed outcomes.

*Automated Processing and Multicenter Evaluation of Meningioma Prediction*

Chinese meningioma research illustrates the progression toward larger, externally tested workflows. Cheng et al. included 952 patients, with 314 in an external test cohort, and connected automated segmentation to attention-enhanced 2D and 2.5D ResNet50 and Swin Transformer classifiers [20]. Histopathology was

reviewed under the 2021 WHO classification. This matters because the model learns the meaning of the assigned label, including the classification version and reference process. Historical grade records cannot simply be assumed equivalent to labels reviewed under a newer framework. Larger cohorts improve the opportunity for evaluation, while the relevant local question remains whether the same inputs and label definitions are available.

Yang et al. provide a direct comparison of the contour-to-prediction link. In 326 patients, the automated-contour radiomics model had a test AUC of 0.842 (95% confidence interval: 0.807-0.895), compared with 0.813 (0.799-0.862) for the manual-contour model; the test set comprised 66 patients, including nine high-grade tumors [19]. This supports technical feasibility without establishing equivalence between contouring methods or clinical benefit after deployment. In a complementary Korean study, Jun et al. trained a 3D U-Net/ResNet workflow on 257 patients and tested it externally in 61; the combined T1-contrast and T2 model achieved a segmentation Dice score of 0.910 and a grading AUC of 0.770 (95% confidence interval: 0.644-0.895) [21]. These studies show why contour overlap and downstream prediction need to be evaluated together.

Conventional CNNs remain useful sources of image representations. Gui et al. compared VGG-, ResNet-, and DenseNet-derived features with radiomics for sinus invasion in a Chinese multicenter cohort of 601 meningiomas. The radiomics-ResNet fusion model had an external AUC of 0.769, compared with 0.747 for the ResNet-feature model [22]. The task was sinus invasion, rather than WHO grading, and the classifiers were random forests operating on selected features. This is an example of combining learned representations with a tabular predictor, rather than requiring every physician team to train an end-to-end CNN. Comparing a model with independent physician readings also provides different evidence from measuring the benefit of assistance during reading; the clinical claim must match the actual evaluation design.

*Operational Value of Software Integration and Physician Participation*

Reusable software helps connect these methods. PyRadiomics provides an open-source feature-extraction platform that can be used independently or through 3D Slicer [23]. Gui et al. used PyRadiomics alongside CNN feature extraction and random-forest modeling, demonstrating the combination of reusable components in disease-specific research [22]. Such components make physician-led local analysis technically possible, but their availability alone does not demonstrate ease of use or reduced learning time. Integration should expose the clinical meaning of inputs and outputs, minimize repeated file handling, and explain choices such as feature filtering. The intended result is a reviewable research process in which physicians can focus on the disease question.

A study from Taiwan provides direct evidence about assisted contouring workload. Lu et al. evaluated nine physicians contouring 10 mixed brain-tumor cases for radiosurgery, including three meningioma cases. Among experts, median contouring time decreased from 11.2 to 7.3 minutes per case, while inter-reader Dice agreement increased from 0.862 to 0.900 [24]. These are results for the mixed case series, rather than meningioma-only performance estimates. They show that review and correction of generated contours can produce measurable benefits in an actual task. Whether grade-prediction assistance provides a comparable diagnostic benefit requires corresponding reader studies.

## Critical appraisal and translational priorities

*Growing Reusable Capabilities and Limited Evidence Matched to Clinical Tasks*

The central tension is the growth of readily available segmentation and learning capabilities alongside comparatively limited evidence that they fit particular clinical workflows. Small datasets impose at least three constraints: insufficient segmentation annotations, limited data for learning image representations, and too few independent patients with credible outcomes. Automated segmentation, image pretraining, and TabPFN address different parts of these constraints. Reusing segmentation and representation-learning capabilities can reduce repeated work when local case numbers are small; the availability of independent outcome-labeled patients determines how thoroughly a prediction model can be evaluated clinically. Prediction model methodology emphasizes assessing data adequacy against the objectives of development and validation [2]. Additional slices, image augmentation, or synthetic features do not increase the number of independent patients or their outcome information.

Local modeling should start with an unmet diagnostic need. If an existing model fits the local task, assessing its applicability before considering recalibration or adaptation can reduce redevelopment. Clinical prediction methodology distinguishes internal evaluation, external evaluation, and model updating, and explains why a simple split of a small dataset can weaken both development and evaluation [2]. Local development is therefore an option, not an assurance of superior accuracy on local patients. Its rationale may be a distinct target, available sequences, or a different case mix; any advantage must be demonstrated in patients who did not determine the model. Physician research autonomy includes choosing which knowledge to reuse and which differences require local evidence.

*Connecting ROI Review With Prediction Reliability*

In clinical use, physician review of automated contours, followed by selective correction before analysis, is compatible with routine work. Review should establish whether the lesion is covered, whether adjacent air, skull, or other non-target tissues have been included, and whether the contour serves the intended downstream analysis. Lifecycle guidance connects postprocessing, quality assessment, and clinical use within a continuous quality-management process [18]. Expert inspection and correction of existing contours provide evidence of clinical acceptability and editing burden. Quantifying agreement with an independent reference contour requires a separate reference annotation; these are distinct forms of evidence.

A frequently overlooked issue is propagation of contour error through the workflow. If a prediction model is developed using detailed manual ROIs but deployed using automated ROIs, its input has changed. The Image Biomarker Standardization Initiative (IBSI) standardizes feature definitions and software calculations, providing a common basis for investigating these differences [25]. Subsequent IBSI work extended standardization to convolutional filter configurations, demonstrating that identically named features also require consistent processing conventions [26]. Evaluation of the complete workflow should therefore establish whether inputs derived from automated or physician-corrected contours preserve the reliability of downstream predictions and risk interpretation. Acceptable contours and accurate classification need to be assessed together to support a usable system.

*Connecting Models by Comparing What Must Be Learned Locally*

Suitability for small datasets should be compared in terms of inputs, existing priors, and what remains to be learned locally (Table *1*). A CNN trained from scratch, a fine-tuned CNN, and a lightweight predictor using a frozen encoder can have very different local data demands. TabPFN reuses tabular priors and can offer physicians an accessible way to explore a pretrained deep learning model without first designing a new image architecture. Its incremental value should be compared with penalized logistic regression or other baselines using the same input information and patient-level evaluation. Cross-paper AUC rankings confound architecture with case mix, labels, and evaluation design.

| Method and input | Reusable capability | Disease-specific evidence | What still needs local assessment |
|---|---|---|---|
| U-Net or nnU-Net family; MRI volumes | Segmentation workflow or task-specific weights | SegResNet: test Dice 0.881 in a 326-patient cohort [19] | Coverage, boundary errors, correction time and downstream feature changes |
| CNN; 2D, 2.5D or 3D images | Pretrained spatial representations | 2.5D ResNet50: external grading AUC 0.90 in 314 test patients [20] | Frozen encoding versus fine-tuning; representative outcome labels |
| Vision Transformer; images | Pretrained contextual representations | BrainIAC tested frozen representations and limited-label adaptation across brain MRI tasks [10] | Target relevance; added sequences and adaptation burden |
| TabPFN; feature tables | Priors learned on synthetic tabular tasks | External meningioma AUC: radiomics 0.628; clinical 0.645 [13] | Feature transportability, label quality, calibration and matched baselines |
| Penalized logistic regression; features | Constrained statistical modeling | Regularized logistic regression aggregated nnU-Net outputs in Huryol et al. [13] | Discrimination and calibration with the same patient-level splits |

**TABLE 1: Roles and conditions of commonly used models in meningioma MRI workflows with small datasets.**

CNN, convolutional neural network; ViT, vision Transformer; ROI, region of interest; AUC, area under the receiver operating characteristic curve. Numerical examples are task-specific and do not constitute cross-study rankings. A segmentation Dice score measures contour overlap; a grading AUC measures discrimination, with 0.5 indicating chance-level ranking. Neither measures calibration or benefit to a physician using the system.

Automated feature selection should retain decision points that physicians can understand. Feature selection is a methodological decision within radiomics. Software should explain whether a step removes unstable measurements, compresses redundant information, or selects variables associated with an outcome, rather than returning only a set labeled as optimal. Users can be offered choices of method, threshold, and retained feature count, provided the selection process is recorded and evaluated with the model. The Prediction Model Risk of Bias Assessment Tool for regression and AI methods (PROBAST+AI) includes analytical procedures in its assessment of bias, supporting scrutiny of leakage during preprocessing, selection, and tuning [27]. Research with small datasets particularly requires patient-level data partitioning, restriction of

standardization, selection, and tuning to training data, and resampling-based evaluation of the complete process, including these choices. The distinction between development quality and validation bias helps physicians assess whether a successfully executed program has produced trustworthy results.

### Connecting clinical workflows by making outputs reviewable

The routine modeling sequence is image preparation, reviewed segmentation, feature extraction or encoding, feature processing, model development, and independent evaluation. To make this sequence part of daily work, we propose triggering validated segmentation when the required MRI series becomes available, transferring contours to the physician workstation, and retaining reviewed ROIs during reporting; this architecture should be assessed locally rather than assumed already proven for every scanner. Pathological or follow-up labels can be appended later, creating a database for subsequent analysis and appropriately governed exchange between centers. New cases should use a fixed model, with later updates evaluated separately [2]. The distinction connects routine case accumulation with research while preserving the version used for each prediction.

Software needs to retain input provenance, ROI corrections, feature processing, and model versions. TRIPOD+AI supports transparent reporting of model development, evaluation, and implementation details [16]. A further proposed use of language-model agents is to help physicians find documented tools and translate natural-language requirements into draft scripts that connect existing steps. This may lower a programming barrier, but it has not established clinical benefit in this meningioma workflow. Agent-generated code must be fixed to a version, reviewed by a qualified person, and pass regression checks before entering the operational process. At the workstation, the physician should see lesion coverage, the prediction target, applicability conditions, and any processing failure alongside the risk estimate.

WHO grade makes the clinical target concrete. EANO recommends annual MRI follow-up for grade 1 and six-monthly follow-up for grade 2 during the first five years, followed by longer intervals, and at least three- to six-monthly review for grade 3; management also depends on symptoms, resection status, location, and treatment feasibility [28]. A preoperative risk estimate can therefore inform discussion of aggressive disease and anticipated management, while imaging does not replace histomolecular grading or determine radiotherapy by itself. Tumor consistency is a distinct supplementary target: Shi et al. prospectively linked MR elastography measurements to intraoperative consistency in 88 patients with 89 meningiomas [29]. When specialized elastography is unavailable, routine MRI merits study: Han et al. examined T2-weighted and diffusion histogram parameters for consistency in 103 patients [30]. This supports investigating additional information in available scans, without equating histological subtype with hardness or claiming routine MRI is equivalent to elastography. Feature attribution likewise describes model behavior; biological or treatment meaning requires separate evidence.

### Evaluating workload and reliability in clinical use

Workload evaluation should include physician inspection, correction, and confirmation time, together with algorithmic inference time. It should also report the proportion of cases requiring no correction, failures, and reprocessing burden. Diagnostic evaluation should distinguish independent physician reading from reading with model assistance. Multireader, multicase evaluation can examine human-AI collaboration and differences between physicians [24]. Such evaluation can establish whether time saved through automation becomes effective diagnostic support and can reveal burdens concentrated in difficult cases that are obscured by average time savings.

Prediction evaluation should examine discrimination, calibration, and utility related to clinical decisions. Decision curves estimate potential net benefit under specified thresholds and assumptions; actual changes in physician diagnosis and management require evaluation during use [2]. FUTURE-AI treats usability, robustness, traceability, and postdeployment monitoring as lifecycle requirements [18]. A local workflow should therefore be judged by its conditions of use, failures, maintenance needs, and diagnostic contribution, as well as its predictive performance. If automated contours save work but grading remains unreliable, the contouring benefit can still be reported and used within its validated scope while unsupported grading output is withheld from diagnostic use.

## Conclusions

Connecting automated segmentation with learning from small datasets provides a practical pathway for physicians to explore local meningioma MRI models. Automated segmentation supports contour acquisition, pretraining, and TabPFN reuse of existing learning, and software integration can reduce switching between steps and repeated handling. The aim is to make reviewable modeling more accessible and return validated information to diagnosis. Independent patients and dependable labels remain necessary, and TabPFN's convenience does not guarantee generalization. Future studies should examine contour-editing burden, downstream reliability, and physician diagnostic gains in the same workflow. Routine accumulation of reviewed images and later outcomes can support local evaluation and exchange between centers. Automation can take on repetitive work while physicians retain responsibility for reviewing ROIs, labels, and interpretation.

# Additional Information

## Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

**Concept and design:** Fengzhang Zhu, Hao Ma, Hui Guo, Shizhou Tang

**Acquisition, analysis, or interpretation of data:** Fengzhang Zhu, Hao Ma, Heji Ma, Xianglin Zhang

**Drafting of the manuscript:** Fengzhang Zhu, Hao Ma, Xianglin Zhang

**Critical review of the manuscript for important intellectual content:** Fengzhang Zhu, Hao Ma, Hui Guo, Heji Ma, Shizhou Tang

**Supervision:** Fengzhang Zhu, Hao Ma, Hui Guo, Xianglin Zhang

## Disclosures

**Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

## Acknowledgements

We extend our heartfelt appreciation to all the staff of the Department of Radiology of the First Affiliated Hospital of Jinzhou Medical University for their devoted efforts and contributions. Hao Ma and Fengzhang Zhu contributed equally to this work as co-first authors. No artificial intelligence was used.

# References

1. Horvat N, Papanikolaou N, Koh DM: Radiomics beyond the hype: a critical evaluation toward oncologic clinical use. Radiol Artif Intell. 2024, 6:e230437. 10.1148/ryai.230437
2. Collins GS, Dhiman P, Ma J, et al.: Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024, 384:e074819. 10.1136/bmj-2023-074819
3. LaBella D, Khanna O, McBurney-Lin S, et al.: A multi-institutional meningioma MRI dataset for automated multi-sequence image segmentation. Sci Data. 2024, 11:496. 10.1038/s41597-024-03350-9
4. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021, 18:203-11. 10.1038/s41592-020-01008-z
5. Ma J, He Y, Li F, Han L, You C, Wang B: Segment anything in medical images. Nat Commun. 2024, 15:654. 10.1038/s41467-024-44824-z
6. Zhang S, Zhang Q, Zhang S, et al.: A generalist foundation model and database for open-world medical image segmentation. Nat Biomed Eng. 2026, 10:1026-41. 10.1038/s41551-025-01497-3
7. LaBella D, Schumacher K, Mix M, et al.: The 2024 Brain Tumor Segmentation Challenge Meningioma Radiotherapy (BraTS-MEN-RT) dataset. Sci Data. 2026, 13:306. 10.1038/s41597-026-06649-x
8. Maier-Hein L, Reinke A, Godau P, et al.: Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024, 21:195-212. 10.1038/s41592-023-02151-z
9. Mei X, Liu Z, Robson PM, et al.: RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. Radiol Artif Intell. 2022, 4:e210315. 10.1148/ryai.210315
10. Tak D, Garomsa BA, Zapaishchykova A, et al.: A generalizable foundation model for analysis of human brain MRI. Nat Neurosci. 2026, 29:945-56. 10.1038/s41593-026-02202-6
11. Tabassum M, Di Ieva A, Liu S: Meta transfer learning for brain tumor segmentation using nnUNet in meningioma and metastasis cases. Sci Rep. 2025, 15:37599. 10.1038/s41598-025-20957-z
12. Hollmann N, Müller S, Purucker L, et al.: Accurate predictions on small data with a tabular foundation model. Nature. 2025, 637:319-26. 10.1038/s41586-024-08328-6
13. Huryol C, Sulaimanov U, Yuce M, et al.: MRI-based preoperative WHO grade prediction in meningioma: development and comparison of deep learning, radiomics, clinical, and combined models in multicenter data. J Neurooncol. 2026, 180:18. 10.1007/s11060-026-05827-6
14. Barry N, Kendrick J, Molin K, et al.: Evaluating the impact of the Radiomics Quality Score: a systematic review and meta-analysis. Eur Radiol. 2025, 35:1701-13. 10.1007/s00330-024-11341-y
15. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, Kahn CE Jr: Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiol Artif Intell. 2024, 6:e240300. 10.1148/ryai.240300
16. Collins GS, Moons KG, Dhiman P, et al.: TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024, 385:e078378. 10.1136/bmj-2023-078378
17. Vasey B, Nagendran M, Campbell B, et al.: Reporting guideline for the early-stage clinical evaluation of

decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022, 28:924-33. 10.1038/s41591-022-01772-9
18. Lekadir K, Frangi AF, Porras AR, et al.: FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025, 388:e081554. 10.1136/bmj-2024-081554
19. Yang L, Wang T, Zhang J, Kang S, Xu S, Wang K: Deep learning-based automatic segmentation of meningioma from T1-weighted contrast-enhanced MRI for preoperative meningioma differentiation using radiomic features. BMC Med Imaging. 2024, 24:56. 10.1186/s12880-024-01218-3
20. Cheng X, Li H, Li C, et al.: Attention-based deep learning network for predicting World Health Organization meningioma grade and Ki-67 expression based on magnetic resonance imaging. Eur Radiol. 2026, 36:2518-30. 10.1007/s00330-025-11958-7
21. Jun Y, Park YW, Shin H, et al.: Intelligent noninvasive meningioma grading with a fully automatic segmentation using interpretable multiparametric deep learning. Eur Radiol. 2023, 33:6124-33. 10.1007/s00330-023-09590-4
22. Gui Y, Hu W, Ren J, Tang F, Wang L, Zhang F, Zhang J: Preoperative diagnosis of meningioma sinus invasion based on MRI radiomics and deep learning: a multicenter study. Cancer Imaging. 2025, 25:20. 10.1186/s40644-025-00845-5
23. van Griethuysen JJ, Fedorov A, Parmar C, et al.: Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017, 77:e104-7. 10.1158/0008-5472.CAN-17-0339
24. Lu SL, Xiao FR, Cheng JC, et al.: Randomized multi-reader evaluation of automated detection and segmentation of brain tumors in stereotactic radiosurgery with deep neural networks. Neuro Oncol. 2021, 23:1560-8. 10.1093/neuonc/noab071
25. Zwanenburg A, Vallières M, Abdalah MA, et al.: The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. 2020, 295:328-38. 10.1148/radiol.2020191145
26. Whybra P, Zwanenburg A, Andrearczyk V, et al.: The Image Biomarker Standardization Initiative: standardized convolutional filters for reproducible radiomics and enhanced clinical insights. Radiology. 2024, 310:e231319. 10.1148/radiol.231319
27. Moons KG, Damen JA, Kaul T, et al.: PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025, 388:e082505. 10.1136/bmj-2024-082505
28. Goldbrunner R, Stavrinou P, Jenkinson MD, et al.: EANO guideline on the diagnosis and management of meningiomas. Neuro Oncol. 2021, 23:1821-34. 10.1093/neuonc/noab150
29. Shi Y, Huo Y, Pan C, et al.: Use of magnetic resonance elastography to gauge meningioma intratumoral consistency and histotype. Neuroimage Clin. 2022, 36:103173. 10.1016/j.nicl.2022.103173
30. Han T, Liu X, Sun J, et al.: T2-weighted imaging and apparent diffusion coefficient histogram parameters predict meningioma consistency. Acad Radiol. 2024, 31:2511-20. 10.1016/j.acra.2023.12.014