Faanes, M.G.; Helland, R.H.; Solheim, O.; Muller, S.; Reinertsen, I. Automatic Brain Tumor Segmentation in 2D Intra-Operative Ultrasound Images Using Magnetic Resonance Imaging Tumor Annotations. *J. Imaging* **2025**, *11*, 365. <https://doi.org/10.3390/jimaging11100365>

Contents

1 One-Sentence Verdict

Training nnU-Net with MRI tumor annotations transferred to ultrasound space via rigid registration yields segmentation performance statistically indistinguishable from models trained with direct iUS annotations. **Deep read — highly relevant to my project.**

2 Research Question & Background Gap

Intraoperative ultrasound (iUS) provides real-time guidance during brain tumor resection, but its noisy, limited-field-of-view images make developing automatic segmentation models

difficult — the only publicly annotated iUS dataset is RESECT-SEG with just 23 cases. MRI tumor annotations are comparatively easy to obtain (public datasets + automatic segmentation software like Raidionics). Can MRI annotations, mapped to iUS space via registration, substitute for manual iUS annotations in training segmentation models?

P.S. The practical significance: if MRI annotations work, clinical teams no longer need to spend extensive time manually annotating ultrasound images, enabling rapid training set expansion.

3 Methods & Data

1. 3.1 Data Composition

- **iUS annotated data** (29 cases): 23 from RESECT + 6 CuRIOUS-SEG test cases, all pre-resection 3D ultrasound
- **MRI annotated data** (180 cases): 103 from ReMIND (55 glioma/metastasis patients) + 77 from St. Olavs in-house (43 patients); missing annotations filled with Raidionics v1.2 auto-segmentation

P.S. **Patient count discrepancy** (added after engineering practice): Our first-surgery filtering from the ReMIND Clinical Data Excel yielded 62 patients — 7 more than Faanes' 55. The difference likely comes from: (1) ReMIND-101 missing MRI tumor segmentation annotation (−1), (2) ~5–6 cases excluded by Faanes due to iUS quality issues (ImFusion visual QC). Faanes' exact exclusion criteria are not detailed in the paper.

Data Type	Source	Count	Annotation Method
iUS annotated	RESECT + CuRIOUS-SEG	29 3D volumes	Manual
MRI annotated	ReMIND	103 volumes	Manual / Raidionics
MRI annotated	St. Olavs in-house	77 volumes	Manual / Raidionics

2. 3.2 Registration

MRI→iUS **rigid registration** using ImFusion Suite (commercial software), algorithm from Wein et al. [22], all registration results visually inspected. Post-registration MRI annotations are transferred to iUS space, forming "pseudo labels."

3. 3.3 Model Training

3D ultrasound volumes sliced along three orthogonal directions into 2D NIfTI, retaining only tumor-containing slices. nnU-Net v2.2 2D configuration with `nnUNetTrainerDA5`, five-fold cross-validation (same patient kept within the same fold), early stopping patience=30. The `nnUNetTrainerDA5` data augmentation option is notably non-default — I need to verify how it differs from the standard trainer.

4. 3.4 Experimental Design

Experiment 1: 9 models with different tumor area thresholds (0–300 mm²) filtering training slices, evaluated on all 29 iUS annotated cases (14,107 test slices). Goal: find the optimal threshold.

Experiment 2: At the optimal threshold (200 mm²), compare three annotation-source models:

- `MRI_200`: MRI pseudo labels only
- `US_200`: iUS annotations only (23 RESECT for training, 6 for testing)
- `MRI+US_200`: both combined
  Also compared against an expert neurosurgeon's manual annotations (Annotator) as inter-observer reference.

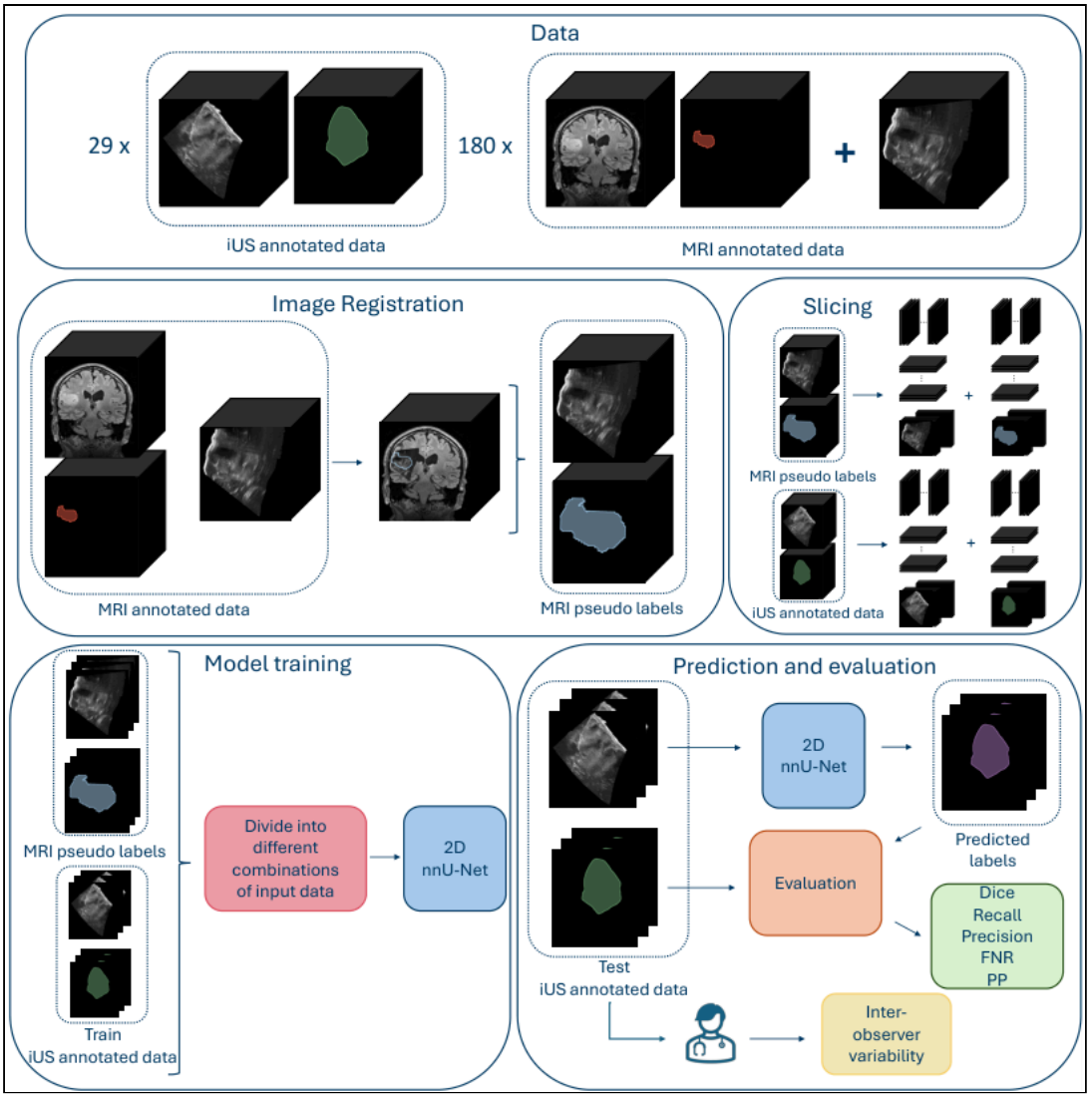

Figure 1: Method overview – data preparation, registration, slicing, training, evaluation

4 Key Evidence

5. 4.1 Experiment 1: Tumor Area Threshold

Table 1 and Figure 2 show that **as the threshold increases from 0 to 200 mm², Dice and Recall improve significantly, FNR drops significantly**, while Precision shows no significant change. 200 mm² is the performance inflection point — MRI_200 achieves overall Dice 0.60±0.28 with the largest effect size (Cohen's d = 0.25). MRI_300 is slightly better on large tumors >200 mm² (0.76), but overall Dice is identical (0.60) and training samples drop from 51k to 35k.

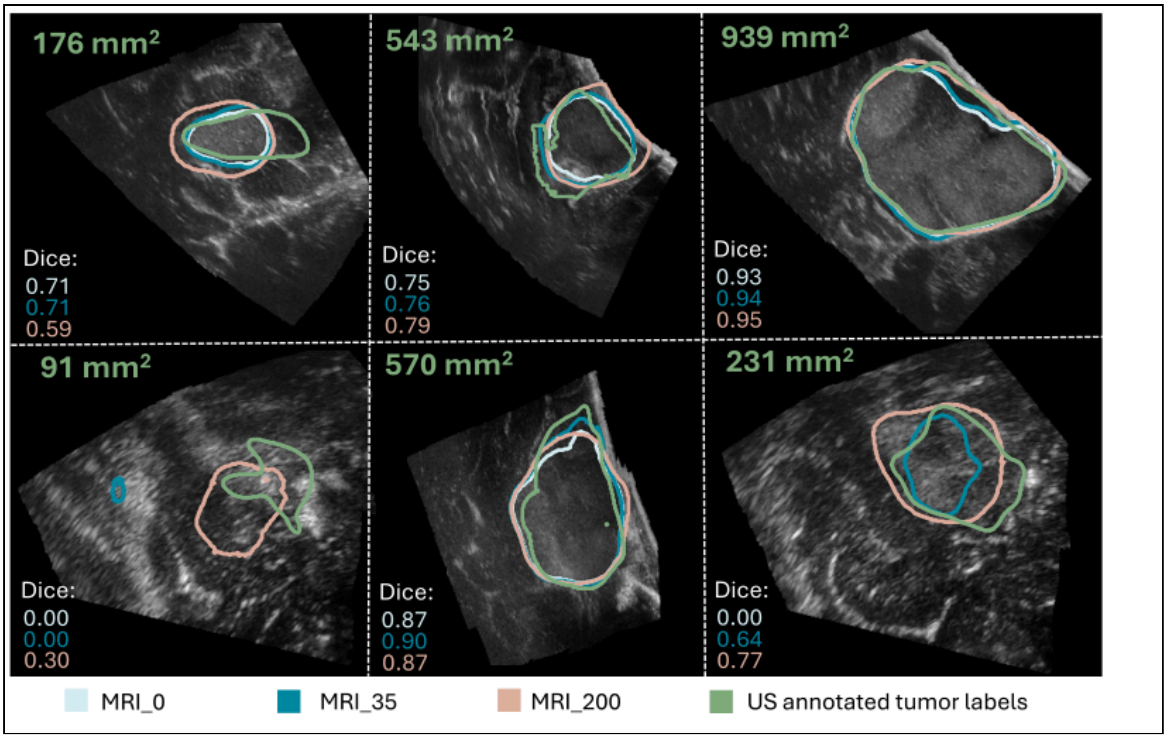

Figure 3: Segmentation comparison across thresholds – large tumors show similar performance across models, small tumors frequently missed by MRI_0 and MRI_35

Figure 3 visually demonstrates: for large tumors (>200 mm²) all three models' contours nearly overlap, but for small tumors (91 mm²) MRI_0 completely misses (Dice=0.00) while MRI_200 still achieves rough localization at 0.30. The authors do not over-interpret these figures, but only 6 slices are shown — representativeness is limited.

6. 4.2 Experiment 2: Annotation Source Comparison

The three annotation-source models show **no statistically significant difference** (p > 0.0085, Bonferroni), with negligible effect sizes (d ~ 0.01):

Model	Overall Dice	>200 mm² Dice	PP
MRI+US_200	0.62±0.31	**0.81±0.12**	99.0%
MRI_200	0.58±0.32	0.79±0.15	96.6%
US_200	0.59±0.29	0.77±0.12	100%
Annotator	**0.67±0.25**	0.77±0.14	95.2%

The expert annotation performs comparably to models on large tumors but shows clear advantage on small tumors (0-35 mm²): 0.20 vs 0.06–0.08, confirming that small tumors are difficult for all approaches. Figure 5's example selection aligns with Table 1 data, with no apparent over-interpretation.

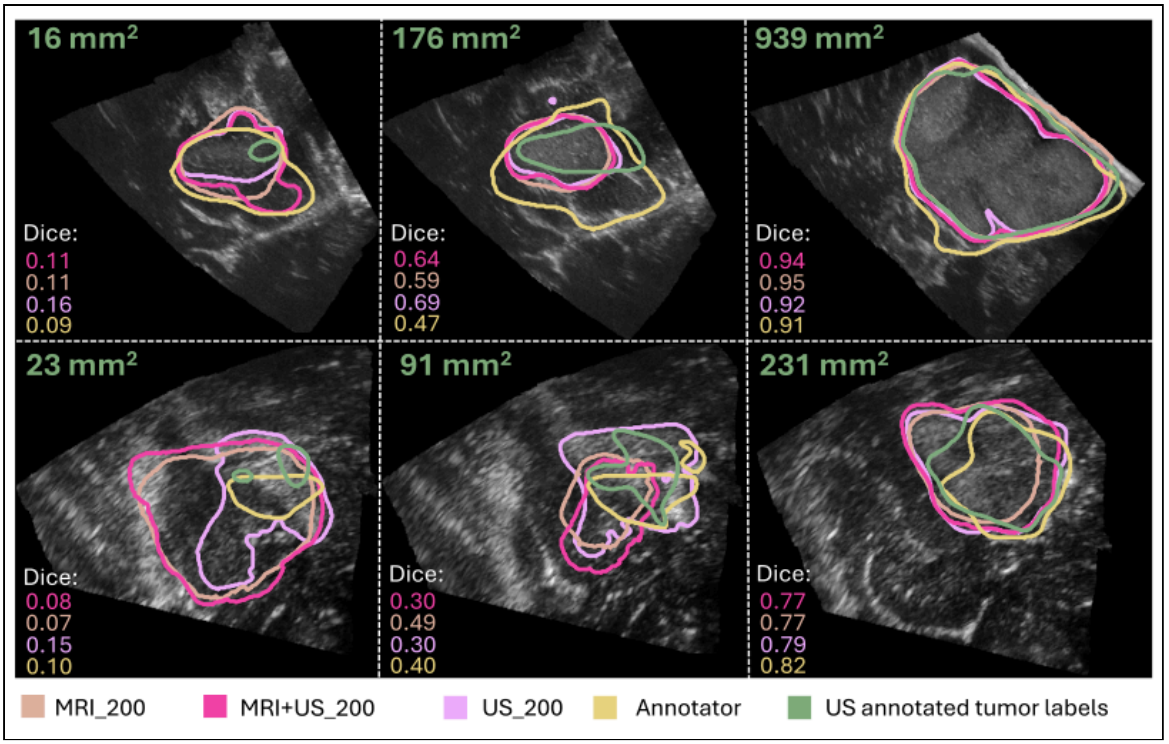

Figure 5: All models vs expert annotation – small tumors (16 mm², 23 mm²) are poor across all methods

5 Author Claims & My Critical Assessment

7. 5.1 What the Paper Explicitly States

- MRI annotations via rigid registration can substitute for iUS annotations in training segmentation models
- Removing small tumor slices (<200 mm²) improves training outcomes
- The best model (MRI+US_200) achieves Dice 0.62±0.31, not significantly different from the expert (0.67±0.25)
- Small tumor segmentation is challenging for both models and human experts

8. 5.2 What Can Be Reasonably Inferred

Registration quality is the bottleneck. The authors speculate that poor small-tumor performance stems from registration errors having outsized impact on small regions (Section 4, para 2). This reasoning is sound — edge slices have small tumor areas at infiltration boundaries, where a few-pixel registration shift can completely misalign the pseudo label.

More data does not necessarily help. MRI+US_200's training set (57k slices) is 7× larger than US_200's (8k slices), yet Dice improves by only 0.03. MRI pseudo-label noise cancels out the volume advantage.

Dice is overly sensitive to small structures. The authors themselves acknowledge this (Section 4, para 5) — a few-pixel shift can drop Dice from 0.5 to 0.0.

9. 5.3 What Remains Uncertain

The **test set is too small** (only 6 patients), giving insufficient statistical power. "No significant difference" does not equate to "truly equivalent" — a larger test set is needed for robust conclusions.

Only rigid registration was used, with no affine or deformable alternatives attempted. Would better registration improve pseudo-label quality? Could it mitigate the small-tumor problem?

ImFusion Suite is commercial software with limited algorithmic transparency, constraining reproducibility. Its core is the LC2 metric (Wein et al.), which incorporates an ultrasound scattering physics model, making it fundamentally more suitable for MRI-iUS cross-modal registration than purely statistical MI metrics. The CuRIOUS 2018 benchmark shows LC2 achieving mTRE=1.57mm, significantly outperforming MI-based methods.

Raidionics auto-annotation quality: some MRI labels come from automatic segmentation rather than manual annotation, but the paper does not separate their impact on final results.

Only pre-resection ultrasound was tested — no validation on mid/post-resection scans where brain shift is more severe.

Warning **Registration exclusion rate comparison** (added after engineering practice): Our reproduction using SimpleITK Mattes MI had a 22.8% exclusion rate (30/114), with ~13% of patients producing pseudo labels that passed all automatic metrics but were visually incorrect. Faanes using ImFusion LC2 had only ~4–8% exclusion. The core difference is the metric function — MI lacks ultrasound physics priors and more easily converges to local optima unrelated to anatomical alignment. This was the biggest unexpected difficulty in reproducing the pipeline.

10. 5.4 Confirmation Bias Self-Check

The paper's conclusion — "MRI annotations can substitute for iUS annotations" — aligns with my expectations (since my pipeline is designed exactly this way), making me potentially less sensitive to its limitations. In reality, the test set has only 6 patients, and "no significant difference" is very likely due to insufficient statistical power rather than true equivalence. This must be kept in mind when designing my own experiments.

6 Relevance to My Project

Warning This paper's pipeline is nearly identical to my project: MRI annotations → registration → pseudo labels → nnU-Net 2D segmentation.

The experimental design framework (threshold experiment + annotation source comparison) can be directly referenced, with the 200 mm² area threshold serving as a starting point for training data filtering. The evaluation metric suite (Dice + Precision + Recall + FNR + PP) is comprehensive. The ReMIND dataset usage (55 patients, 103 pre-resection ultrasound volumes) provides a clear data processing template for my work.

For adaptation, I use SimpleITK for registration (open-source, controllable) while the paper uses ImFusion Suite (commercial), so registration algorithm and quality may differ. The `nnUNetTrainerDA5` data augmentation strategy needs verification. The St. Olavs in-house data (77 cases) is inaccessible, but the public ReMIND (103 cases) and RESECT (23 cases) suffice for initial experiments.

Potential improvements worth exploring include better registration methods (affine or

deformable, or even manual fine-tuning), boundary-based metrics (normalized surface distance, boundary IoU) to supplement Dice for small tumor evaluation, alternative architectures (Vision Transformer, USFM foundation models), and incorporating mid/post-resection MRI annotations to expand the small-tumor training set and address data imbalance.

7 My Questions & Ideas

How large is the rigid registration error, exactly? If the offset between pseudo labels and true iUS tumor boundaries could be quantified, low-quality pseudo labels could be automatically filtered based on displacement rather than the blunt area-threshold approach. The 200 mm² threshold was tuned on this specific dataset and may need re-tuning on my data, but the direction is right — quality over quantity.

The paper did not explore semi-supervised learning. If the 180 unlabeled iUS volumes were also leveraged (e.g., pseudo-label self-training), could small-tumor detection improve? Dorent et al. [11]'s patient-specific approach (synthetic ultrasound + fine-tuning) achieves median Dice 0.84–0.87, far above generic models. Could both strategies be combined — first train a generic model with MRI pseudo labels, then fine-tune with patient-specific synthetic ultrasound from MRI?

8 Key References

- **Dorent et al. [11]**: Patient-specific synthetic ultrasound segmentation, median Dice 0.84–0.87 — entirely different but complementary approach
- **Qayyum et al. [8]**: CuRIOUS-SEG challenge winner, self-supervised 3DResUNet, Dice 0.57 — current baseline
- **Juvekar et al. [12]**: ReMIND database, the core data source shared by this paper and my project
- **Wein et al. [22,23,24]**: LC2 metric MRI-US registration series
- **Reinke et al. [29]**: Systematic analysis of Dice and other metrics' sensitivity to small structures

#segmentation #ultrasound #MRI #nnUNet #cross-modal-annotation-transfer #rigid-registration #high-priority