Supplementary File: Risk-of-Bias Assessment Rubric

This supplementary file provides the operational criteria used to classify methodological risk of bias in studies included in the systematic review. The rubric was used qualitatively and was not intended to exclude studies automatically; rather, it supported transparent interpretation of the strength and limitations of the evidence.

Domain	Low risk	Moderate risk	High risk	Rationale
Dataset representativeness	Multi-centre, externally validated, or diverse cohort with clear source reporting.	Single-centre or public benchmark plus limited robustness testing.	Exclusive benchmark dependence, unclear cohort source, or highly selected dataset.	Determines generalisability beyond benchmark conditions.
Train-test separation/data leakage	Explicit patient-level separation and leakage controls.	Split described but leakage-control detail incomplete.	Unclear split, slice-level split, or possible overlap.	Leakage can inflate segmentation performance.
Sample size/class imbalance	Adequate sample size and explicit imbalance-handling strategy.	Moderate sample or incomplete class-imbalance reporting.	Small sample, no imbalance strategy, or rare subregion neglect.	Small enhancing tumour regions are vulnerable to overfitting and poor sensitivity.
External validation	Independent external test set or multi-institutional holdout.	Internal cross-validation/holdout with limited perturbation testing.	No external validation or robustness testing.	External testing is central to clinical translation.
Reporting transparency	Full architecture, preprocessing, hyperparameters, metrics, and limitations reported.	Some reproducibility details missing.	Insufficient methods, selective reporting, or unclear metric definitions.	Transparency determines reproducibility and interpretability.
Uncertainty/calibration	Uncertainty and calibration quantified and clinically interpreted.	Uncertainty or calibration partially reported.	No uncertainty, calibration, or reliability assessment.	Overconfident segmentation can be unsafe in clinical workflows.

Overall judgement rule

A study was considered low overall risk when most domains were low risk and no critical domain, such as train-test separation or external validation, was high risk. Moderate risk indicated mixed reporting or limited validation without clear evidence of leakage. High risk indicated major uncertainty in data separation, exclusive benchmark optimisation, no external validation, or insufficient reporting for reproducibility.