Supplementary Materials

Exploratory whole-tumor and habitat radiomic associations with study-documented *Clonorchis sinensis* status in hepatocellular carcinoma

Yuqi Huang[1,2], Lingfeng Huang[2], Hanjing Zhang[3,4,5,6], Yaoqin Xie[7], Jing Cai[1], Tao Peng[3]

[1]Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong SAR, China.
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, Guangdong, China.
[3]Department of Hepatobiliary Surgery, The First Affiliated Hospital of Guangxi Medical University, Nanning 530021, Guangxi, China.
[4]Guangxi Key Laboratory of Enhanced Recovery after Surgery for Gastrointestinal Cancer (Guangxi Medical University), Nanning 530021, Guangxi, China.
[5]Key Laboratory of Early Prevention and Treatment for Regional High Frequency Tumor (Guangxi Medical University), Ministry of Education, Nanning 530021, Guangxi, China.
[6]Department of Hepatobiliary Surgery, The First Affiliated Hospital, Hengyang Medical School, University of South China, Hengyang 421001, Hunan, China.
[7]Shenzhen University of Advanced Technology, Shenzhen 518107, Guangdong, China.

Correspondence to: Prof. Yaoqin Xie, Shenzhen University of Advanced Technology, Shenzhen 518107, Guangdong, China. E-mail: xieyaoqin@suat-sz.edu.cn; Prof. Jing Cai, Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong SAR, China. E-mail: jing.cai@polyu.edu.hk; Prof. Tao Peng, Department of Hepatobiliary Surgery, The First Affiliated Hospital of Guangxi Medical University, Nanning 530021, Guangxi, China. E-mail: pengtaogmu@163.com

Supplementary Methods

Acquisition records and scope

A separate source-archive audit identified 2,503 NIfTI files representing 212 patients according to the project mapping table, with 985 CT and 1,518 MR image-level labels and 104 CT and 108 MR unique patients. NIfTI headers were readable for 2,501 files; two files were not valid gzip NIfTI files. Among readable headers there were 383 unique dimension triplets, 444 spacing vectors, and 1,126 affine matrices, confirming heterogeneous source geometry rather than a single cohort-wide acquisition geometry. The most frequent readable dimension was (320, 260, 40) in 160 files and the most frequent spacing vector was (1.09375, 1.09375, 4.5) mm in 227 files; representative affine matrices are listed in the companion audit. These source-archive counts are reported for provenance and do not redefine the 161-patient feature-complete modeling cohort.

An audit of one available MR examination package identified 932 images across 28 series, all with modality MR. The headers recorded manufacturer SIEMENS and model Prisma in 932/932 files, 3 T in 926/932 files, and software syngo MR E11 in 932/932 files; the remaining six files used the Secondary Capture SOP class and lacked the field-strength tag. ProtocolName was present in 929/932 files and SeriesDescription in 932/932. Sequence-specific values in the package included localizer 7.0 mm and 7.8/3.69 ms; T2 HASTE 5.0 mm and 1300/121 ms; T1 Dixon 4.5 mm with 3.97/1.29 ms (opposed-phase) or 3.97/2.52 ms (in-phase); T1 VIBE 4.5 mm with TR/TE values spanning approximately 3.31–4.56/1.30–1.43 ms; T2 BLADE-family series 5.0 mm with approximately 5726–6506/81 ms; and DWI 5.0 mm with 3000/43 ms. SpacingBetweenSlices was present in 277/932 files and missing in 655/932. The full series-level table is provided in the companion imaging-provenance audit.

Contrast-agent, contrast-route, contrast-volume, injection-time, and AcquisitionContrast attributes were absent from all 932 files. Dynamic series labels containing A1, A2, V1, or V2 are retained as original DICOM labels and are not reclassified as arterial, portal-venous, or delayed phases. The 6.0-mm MPR entries are derived reconstructions, not original acquisition slice thickness. Because this DICOM package represents one examination, these findings do not establish one uniform acquisition protocol for the 161-patient modeling cohort.

The historical derived\roi_crops cache contains 210 patient-level NPZ files and its source code points to former source locations. Although filenames can be matched to the current image archive, 207/210 cached cases have different original shape metadata and 207/210 have different spacing metadata. These cached crops are therefore not described as regenerated from the current raw

images and current default weight. The more traceable current artifact is the per-image workflow under the current crop output, linked to its mask manifest and QC reports.

Current radiomics extraction

The accompanying radiomics_current.yaml is the available current configuration. It uses the Original image type, bin width 25, label 1, B-spline interpolation, no additional configured voxel resampling, pad distance 5 voxels, three-dimensional extraction and mask correction. It lists first-order, shape, GLCM, GLRLM, GLSZM, GLDM and NGTDM classes. The revision runner disables PyRadiomics shape extraction and calculates shape fields separately. CT-specific acquisition or intensity settings cannot be inferred from this current MR configuration.

The 207 matched revision image/mask pairs represent 106 MR patients. The current feature table contains 107 original radiomic features. The accompanying current_MR_feature_inventory.csv and historical_feature_inventory.csv distinguish this representation from the 173-column historical whole-tumor/habitat schema. A list of historically excluded features and the exact executed historical configuration are unavailable.

The revision geometry calculation uses a padded binary mask and a marching-cubes surface at level 0.5 in physical coordinates. Surface area is computed from mesh triangles, mesh volume from signed tetrahedral volumes, and principal axes from voxel-coordinate covariance. Three-dimensional and projected two-dimensional diameters use convex-hull vertices; if more than 5,000 vertices are present, a deterministic evenly spaced subset is used. The two-dimensional values are whole-mesh projections rather than slice-wise maxima. These are revision-specific descriptors and are not claimed to reproduce historical or IBSI benchmark shape values.

Contour sensitivity and unit of analysis

Each pair has a baseline mask, 1-, 2- and 3-mm erosion/dilation variants, and ten stochastic smooth-displacement variants. Fixed variants threshold a physical signed-distance map. Stochastic displacement uses a coarse Gaussian field with a nominal 12-mm correlation scale, cubic interpolation and smoothing, scaled to a 1.5-mm target within a boundary band and clipped at ±2 mm. The 24 reported model-linked features are the union identified across saved MR model manifests; they are not the feature set selected for the historical pooled candidate.

ICC(2,1) and relative-change summaries compare matching baseline and perturbed image/mask pairs. Forty of 106 patients contribute multiple pairs. Stored pair-level bootstrap intervals do not adjust for this dependence and cannot be interpreted as patient-level intervals. The log-volume/instability Spearman correlation is reported descriptively; the stored unclustered test is not

used for patient-level inference. This analysis provides neither inter-reader agreement nor model-level perturbation predictions.

Revision nested cross-validation

The 90 MR development patients contribute 181 sequence rows, averaged within patient across the 107 numeric original radiomic features. The 16 reserved MR patients contribute 26 rows and are excluded from model development. Sixty-nine development patients also appear in the historical modeling cohort, with consistent stored labels; 14 belong to the historical holdout. These analyses are not independent validation of the historical model.

Three shuffled stratified outer five-fold repeats use seeds 2025, 2026 and 2027. Within each outer training fold, up to four stratified inner folds select a classifier and feature count by mean AUROC. The inner-fold seed is the repeat seed plus 17 times the zero-based outer-fold index. Candidate feature counts are 8, 16, 24, 32 and all available features. Median imputation, standard scaling and ANOVA F-score SelectKBest are fitted within the pipeline. Classifier candidates are balanced L2 logistic regression (C=1, liblinear, maximum 3,000 iterations), balanced RBF SVM (C=1), random forest and ExtraTrees (160 trees, square-root feature sampling, minimum leaf size 2, balanced class weights). The retained candidate is fitted to the outer training fold. Classifier and feature-selection frequencies are supplied as aggregate CSVs.

Each patient receives one out-of-fold score per repeat. The reported summary AUROC uses each patient's mean probability across the three repeats, not the arithmetic mean of repeat-specific AUROCs. The percentile 95% interval uses 2,000 patient resamples with seed 3024. No habitat clustering or habitat-number selection is repeated in this separate original-radiomics analysis.

Learning curves and acquisition controls

Learning curves use five outer folds and five resampling seeds per training fraction, giving 25 fitted models per fraction. Each fraction is applied within an outer training partition, yielding 18, 36, 54 and 72 training patients. The pipeline uses median imputation, standard scaling, up to 24 ANOVA-selected features and balanced L2 logistic regression. All fitted preprocessing remains within its training subset. The figure reports individual replicate AUROCs and the mean with standard deviation.

Acquisition-only controls use sequence signatures, sequence-count signatures, known-phase signatures and sequence/phase presence and count fields, with no reinterpretation of unverified phase names. Categorical variables use training-fitted most-frequent imputation and one-hot encoding; numeric variables use median imputation and standard scaling. Five-fold stratification

uses seed 2025. Logistic regression uses C=1, balanced classes and liblinear; random forest and ExtraTrees use 300 trees, balanced classes and minimum leaf size 2. Common support comprises the four signatures containing both stored label classes (34 patients). The same nested selection procedure is applied to current radiomics in this subset. No cohort-wide ComBat harmonization is claimed.

Historical uncertainty estimates

Historical AUROC intervals use 2,000 patient bootstrap resamples of the saved predictions, omitting one-class samples. Seeds are 43 for pooled image-model AUROC, 48 for CT, 49 for MR, 50 for modality-only and 51 for MR acquisition-only estimates. Paired image-minus-control differences use the same resampled patients in both models, with seeds 51 and 52 for the modality and MR acquisition comparisons. The two-sided P value is twice the smaller zero-tail proportion with a one-count correction, capped at 1. The archived scikit-learn AUROC arithmetic and exact zero comparison are retained. Equivalent rank arithmetic can classify floating-point near-zero differences differently in the MR comparison; the difference, interval and interpretation remain unchanged. These intervals are conditional on saved scores and do not account for historical model search.

Recorded software

The revision environment records Python 3.10.19, NumPy 2.2.6, pandas 2.3.3, scikit-learn 1.7.2, SciPy 1.15.3, SimpleITK 2.5.3, PyRadiomics 3.0.1 and PyTorch 2.7.1+cu118. This is the revision environment, not an exact historical runtime record.

Original model-development records and configuration limits

The primary model used image-derived features only. Modality, MR protocol bucket, sequence name, scanner identifier, and is_vibe were excluded from the primary training features. The archived candidate pipelines included training-side preprocessing, correlation pruning, feature selection, classifier fitting, and probability prediction. The recorded feature-selection families included K-best, L1, and elastic-net variants; the classifier candidates included logistic regression, SVM, random forest, and ExtraTrees.

The implementation records a variance-penalized training selection score of mean cross-validated AUROC minus 0.5 times the cross-validated AUROC standard deviation. The archived inventory contains 1,908 model-configuration rows, 26 train-side cluster-construction rows, and 16 baseline/control rows (1,950 inventory rows in total). Of the model-configuration rows, 1,882 are historical non-selected or derived execution rows and 26 have selected-model/holdout evidence.

These counts include shared, copied, and derived layers and do not represent independent replications. The route used for the main quantitative narrative was the patient-balanced pooled CT/MR candidate with K=2, ExtraTrees, and enet10. It is called a historical candidate throughout. The rationale for K=2 is separated into three parts. It was one of the values evaluated in the patient-balanced K=2–6 search; the saved cluster-validity indices did not uniformly favor K=2, and no biological meaning is assigned to its habitats. The fused label-invariant summary schema contains 173 raw columns for each evaluated K, so K=2 was not selected because it produced a smaller downstream feature count. K=2 is retained because it is the K of the archived pooled candidate selected by the historical training-side cross-validation/variance-penalized route record. This is a predictive-selection and archive-identity rationale, not evidence of a biologically optimal or resampling-stable K.

The repository does not preserve a complete executed software-version lock or the external configuration needed to reconstruct every upstream parameter. Implementation defaults are therefore not presented as proof of the exact historical execution settings. The selected pooled candidate has recorded training-side 5-fold, 3-repeat cross-validation mean AUROC 0.6915 (SD 0.0887) and a historical holdout AUROC of 0.833; these are reported as an archive trace, not as a multiplicity-corrected performance estimate. Additional MR sensitivity analyses of a frozen current specification are described in main-text Section 2.7; they did not alter the historical candidate, labels, split, or holdout predictions and do not retroactively validate the historical search.

Available segmentation and image-processing records

The archived feature and prediction artifacts were derived from tumor masks and CT/MR images in the project pipeline. The accessible default segmentation weight loaded into the code-defined MONAI DynUNet with zero missing and zero unexpected state-dict keys. The corresponding inference output contains 2,502 predicted mask files, and the current mask manifest has 2,502 parsed rows with all referenced image and mask paths present. The current per-image workflow produced 2,337 NPZ crops and has an accompanying per-image and per-patient QC record. These are evidence of an available automated segmentation/crop pipeline and its recorded QC outputs; they are not accuracy or human-reader reproducibility measurements.

The historical manuscript lineage describes an nnU-Net-based segmentation workflow followed by physician review, and nnU-Net is cited as the relevant segmentation method [28]. The current archive still lacks reader identities or expertise, manual-correction criteria and frequency, endpoint blinding, multifocal-lesion handling, independent reader masks, Dice or intraclass correlation

coefficient measurements, and human-reader reproducibility records. Contour-perturbation sensitivity analysis supplies a separate synthetic feature-level contour-sensitivity analysis. These records limit interpretation to the archived image-derived association and do not support human inter-reader reproducibility, segmentation-accuracy, or model-level mask-perturbation claims; they do not invalidate the current feature-level association.

Cohort-wide acquisition metadata were incomplete. One available MR DICOM examination contained 932 images across 28 series, recorded a Siemens Prisma scanner and 3-T field strength in 926 images, and did not preserve contrast-agent or injection-time fields. The remaining six secondary-capture images lacked the field-strength tag. These observations cannot be generalized to the full cohort. Sequence-level details, NIfTI geometry summaries and the historical crop-cache mismatch are reported in Supplementary Methods and the accompanying acquisition table.

The accessible implementation and provenance records do not by themselves establish a complete cohort-wide acquisition or ROI parameter set. In particular, the external PyRadiomics configuration, historical runtime environment, and definitive mapping between the manuscript's historical ROI wording and extraction logs remain unavailable. The companion file historical_feature_inventory.csv in Supplementary Methods lists the 173 columns present in the saved K=2 pooled fused feature table; an exhaustive excluded-feature list is not reconstructable without the executed configuration.

The repository contains the split manifest, route manifests, model-selection records, frozen prediction files, result tables, and figure/table source artifacts used for this study. PyRadiomics and the Image Biomarker Standardization Initiative are cited for the general radiomics framework [29,30], and the radiomics reporting limitations literature is cited for methodological context [31].

The accessible code records a 96×96×96 inference patch, 1-mm target resampling, intensity scaling/clipping from [-175, 250] to [0, 1], Gaussian sliding-window blending with batch size 1, and a 10-mm isotropic crop margin with nearest-neighbour mask resampling when image/mask geometry differs. These are code-level settings, not a verified historical environment lock. The source archive contains 2,503 NIfTI files representing 212 patients, while the available DICOM audit covers one 932-image MR examination across 28 series; neither should be interpreted as one uniform acquisition protocol for the modeling cohort. The complete historical runtime, command, external PyRadiomics configuration, and raw execution logs remain unavailable. The current clinical-covariate audit records that the requested variables were not collected in the project data; it is an availability/data-collection audit, not a clinical baseline comparison.

The revision analyses used Python 3.10.19, NumPy 2.2.6, pandas 2.3.3, scikit-learn 1.7.2, SciPy 1.15.3, SimpleITK 2.5.3 and PyRadiomics 3.0.1. The complete available current extraction configuration is supplied as Supplementary Methods (radiomics_current.yaml); it is a revision configuration and does not recover the missing historical configuration. AI-assisted figure layout and deterministic plotting are disclosed in the declarations and the accompanying figure-design record.

Display-case selection and image presentation

Source images use tumor-ROI 1st–99th percentile intensity rescaling. ROI scale bars use verified in-plane spacing (CT, 0.66796875 mm; MR, 1.125 mm). Display cases were selected from complete image/mask/map triplets with one connected component on the largest tumor-mask slice and both stored habitat labels. Within each modality, the case closest to the median total tumor-mask volume was chosen, with ties resolved by larger occupied map area. This deterministic rule does not establish a typical lesion. Stored maps are displayed in their own processing grids; only empty margins are cropped, retaining every nonzero category without an assumed raw-to-map registration transform.

Detailed MR sensitivity procedures and results

The revision analyses evaluated the separately available current MR radiomics data. Analysis specifications were documented in an internal project log; no public prospective registration is claimed. They did not refit the historical CT/MR whole-tumor-plus-habitat candidate or modify its saved predictions. The eligible revision data contained 207 image/mask pairs from 106 MR patients. The feature representation comprised 107 original radiomic features, rather than the historical fused 173-column representation. Features from multiple sequences were averaged within each patient without using the label or acquisition metadata as predictors.

For current-setting radiomic re-extraction, first-order and texture features were re-extracted under the supplied current PyRadiomics configuration. Shape features used a revision-specific geometry implementation described in Supplementary Methods. These settings do not reconstruct the missing historical configuration.

For contour-perturbation sensitivity analysis, each image/mask pair was evaluated with the baseline mask and 16 perturbations: 1–3-mm erosions, 1–3-mm dilations and ten stochastic smooth displacements. ICC(2,1), bootstrap intervals and median relative absolute changes were calculated for 24 features identified across saved MR model manifests. This subset is not the selected feature set of the historical pooled model. Forty of the 106 patients contributed multiple pairs. The stored

pair-level intervals and descriptive correlation therefore do not establish patient-level uncertainty; we do not use the unclustered correlation test to claim a patient-level association. No model-level perturbation predictions were generated.

For repeated nested cross-validation, the revision split provided 90 MR development patients (70 label-1, 20 label-0) and a reserved 16-patient MR holdout. The reserved holdout was excluded. The development cohort included 69 patients from the historical modeling cohort, of whom 14 belonged to its 33-patient holdout; labels agreed for all 69 shared patients. The two evaluations therefore are not independent. Three repeats of five-fold outer patient-level cross-validation each used inner cross-validation to choose among four classifiers and five feature-count settings. Median imputation, standardization and feature selection were fitted inside the resampling procedure. Details of the candidate models and seeds are supplied in Supplementary Methods. We report repeat-specific AUROCs and AUROC from each patient's mean out-of-fold probability across repeats. Habitat construction and selection of K were not repeated; this is an evaluation of a separate current radiomics specification, not a correction for the historical habitat-model search.

For learning-curve analysis, learning curves evaluated four fractions of each outer training partition, with 25 fitted models per fraction. These corresponded to training sizes of 18, 36, 54 and 72 patients and used training-fitted imputation, standardization, selection of up to 24 features and balanced L2 logistic regression. This analysis measures sample-size sensitivity and is not a power calculation.

For acquisition-support sensitivity analysis, filename-derived sequence/phase signatures were evaluated as predictors within the same 90-patient MR development cohort. Acquisition-only controls used logistic regression, random forest and ExtraTrees. We also reevaluated radiomics in the subset of signatures containing both labels. These signatures were not treated as verified contrast phases. No batch harmonization was performed because a defensible cohort-wide scanner or institution variable was unavailable.

The revision-only audit added evidence about the current feature specification but did not replace the historical holdout analysis. radiomic re-extraction matched all 207 eligible source rows to current-setting re-extractions, with zero extraction failures. This establishes an executable current configuration for the matched image/mask pairs, but the historical PyRadiomics contract remained not exactly reproducible.

The ROI perturbation analysis (Contour-perturbation sensitivity analysis) produced 3,519 rows from 207 tokens and 17 mask variants per token. All 1,712 feature-variant summaries were valid. Across the 24 model-linked features and 16 non-baseline variants, median ICCs ranged from 0.8613

to 0.9713 and median relative changes ranged from 0.0141 to 0.1126. Across 207 image/mask pairs from 106 MR patients, log volume and the instability summary had a descriptive Spearman correlation of −0.3109. Repeated sequences from 40 patients preclude interpreting an unclustered pair-level test as patient-level inference. The analysis quantified synthetic contour sensitivity at the feature level; it did not quantify reader agreement, reference-mask overlap, or model prediction stability, because the historical model-level branch was not run.

The separate MR nested reevaluation (Repeated nested cross-validation) used only the 90-patient development cohort (70 label-1 and 20 label-0 patients), 3 repeats, and 5 outer folds per repeat. AUROC from the probabilities averaged across repeats was 0.8536 (95% CI 0.7650–0.9257), with repeat-specific values from 0.7550 to 0.8771. The median pairwise feature-selection Jaccard was 0.2991 (IQR 0.2991–1.0000), indicating variation in selected feature sets under the frozen specification. The reserved holdout of the revision split was excluded, but 14 development patients overlapped the historical holdout. This analysis used current MR radiomics and did not repeat habitat construction.

The learning curve (Learning-curve analysis) showed mean AUROCs of 0.7050±0.1402, 0.7786±0.0869, 0.7493±0.1138, and 0.7321±0.1112 at 25%, 50%, 75%, and 100% of each outer training partition, respectively (18, 36, 54 and 72 training patients; 25 replicates per fraction). These estimates describe sensitivity to available development sample size and do not establish adequate power.

The protocol-signature analysis (Acquisition-support sensitivity analysis) identified 10 available sequence/phase signatures in the development cohort. Protocol-only AUROCs were 0.8000, 0.8004, and 0.7829 for ExtraTrees, logistic regression, and random forest, respectively. Only 34 patients had common support across signatures (18 label-1 and 16 label-0); nested reevaluation of the frozen radiomics specification in this subset yielded AUROC 0.6840 (95% CI 0.4861–0.8523). These findings make protocol-label coupling visible, but they do not provide cohort-wide harmonization or protocol independence; ComBat was not run because no defensible batch variable was available.

The revision analyses characterize sensitivity under the current specification. Figure 6 summarizes MR resampling and acquisition sensitivity; contour distributions appear in Supplementary Figure 5; Supplementary Figure 5, Supplementary Figure S6, and Supplementary Figure S7 show the underlying ROI-sensitivity, selection/learning-curve, and protocol-support distributions, respectively. The evidence-provenance map is provided in Supplementary Figure S8. In particular, the feature-level ROI analysis does not support a human segmentation reproducibility claim, the

frozen-specification AUROC is not a corrected estimate for the historical model search, and the protocol controls do not establish scanner- or sequence-invariant performance. Feature-provenance details are provided in Supplementary Table S6, the clinical-covariate audit in Supplementary Table S7, and the revision-only details in Supplementary Table S8.

Original versus patient-balanced clustering

Supplementary Table S1 and Supplementary Figure S1 now distinguish the original K=2–9 construction from the patient-balanced K=2–6 construction. The former sampled 708,681 CT and 268,894 MR training voxels. The latter sampled 145,174 CT and 110,876 MR voxels from 61 and 67 training patients. Summing the per-patient sampling logs reproduces these totals, and no patient contributes more than 3,000 voxels. The theoretical maxima are 183,000 CT and 201,000 MR voxels; actual totals are smaller because small tumors contribute all available ROI voxels. Patient-balanced Calinski–Harabasz and Davies–Bouldin values are supplied for K=2–6. Its Silhouette entries are missing and its silhouette sample size is 0, so no patient-balanced Silhouette values are claimed. The number of maps saved includes training and holdout application and is not the number of patients used for centroid fitting.

Whole-tumor versus per-habitat column counts

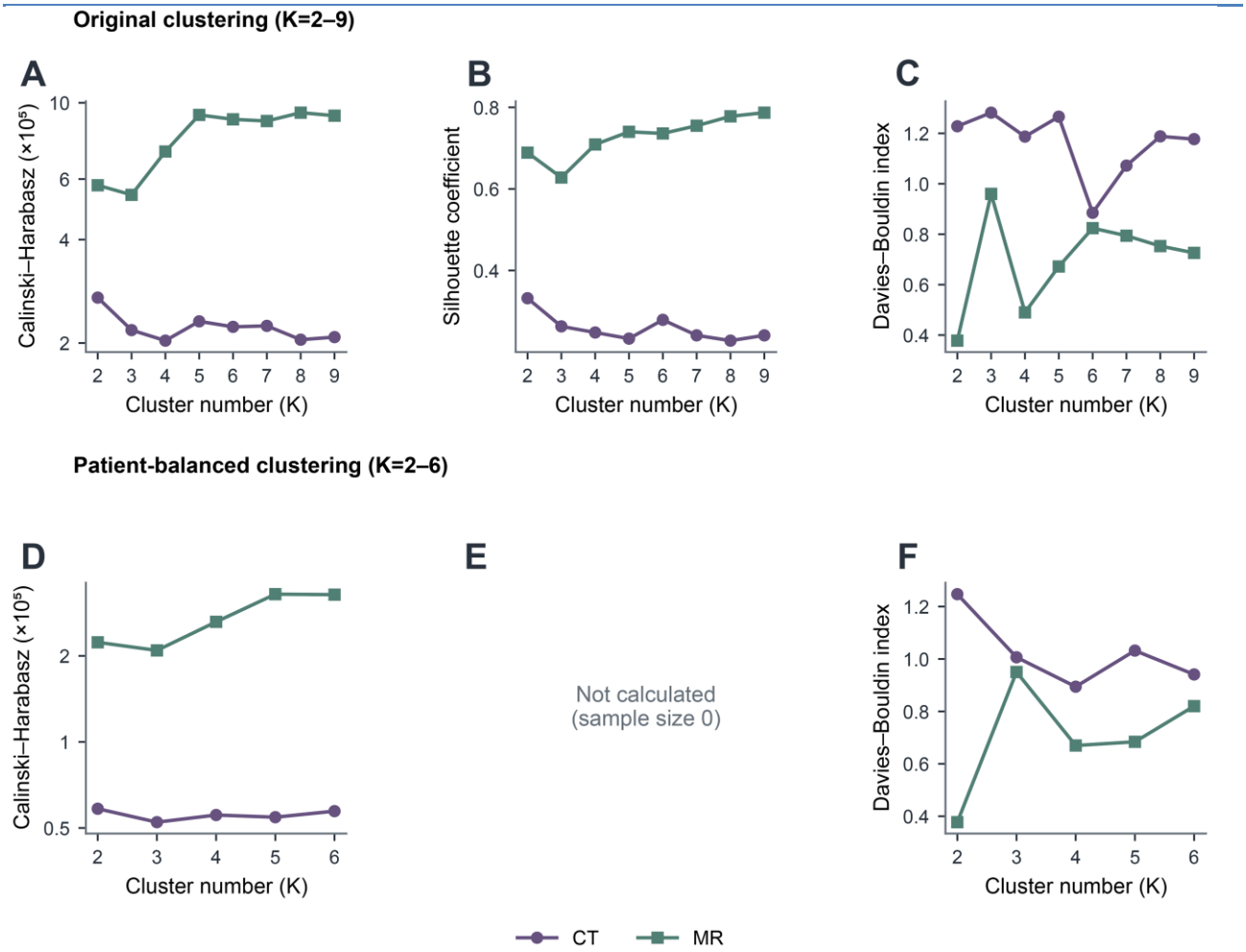
The saved whole-tumor table contains 19 named radiomic descriptors and one voxel-count column (20 columns). Each unsummarized habitat contains those same 19 named descriptors, its voxel count and its tumor fraction (21 columns). Thus, the extra column is habitat fraction; it is not an extra radiomic biomarker. Direct concatenation gives 42, 63, 84, 105 and 126 columns for K=2–6. The label-invariant habitat summaries are a different representation, with 153 columns, yielding 173 when combined with the 20 whole-tumor columns. The companion whole_tumor_per_habitat_column_comparison.csv lists all 20 and 21 columns and their roles. The exact executed extraction configuration is unavailable, so the reason this historical run retained these 19 named radiomic descriptors, rather than a larger PyRadiomics feature set, remains unresolved. The current 107-feature MR configuration does not explain or reconstruct that historical choice.

Timing of analytical choices

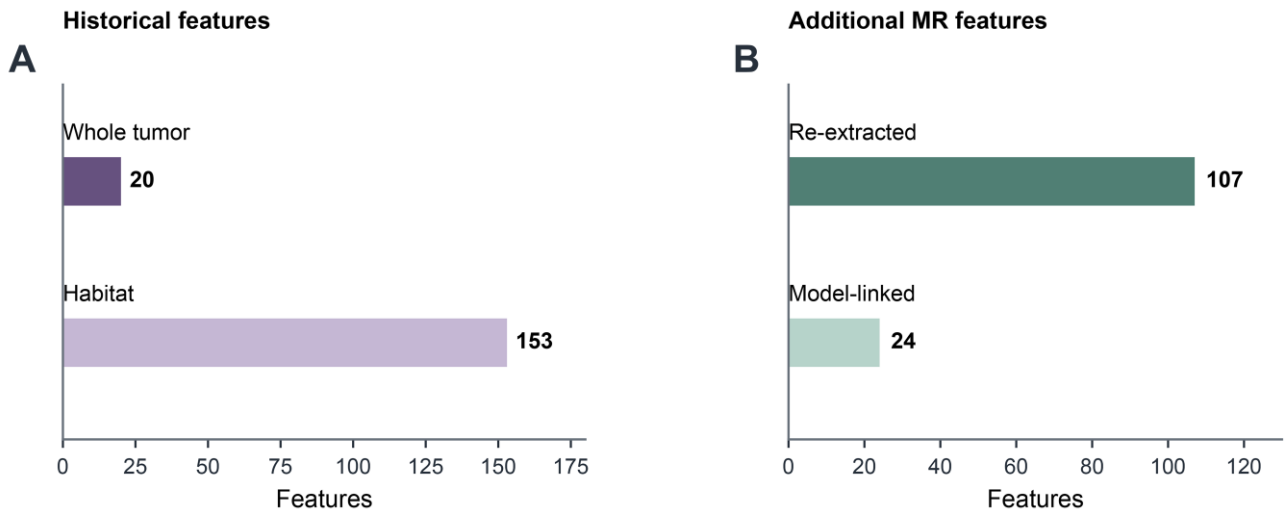
The available records establish the fixed patient split, training-only centroid fitting, original and patient-balanced clustering ranges, feature-summary definitions, classifier candidates and training cross-validation selection score. The responsible author confirms that holdout results were not viewed before model selection. A complete time-stamped record of when patient balancing, K

ranges, label-invariant summaries and candidate classifiers were specified, and whether later manuscript decisions were informed by holdout results, is unavailable. Directory/run names alone are not used as proof of decision timing. Multiple models share the same 33-patient evaluation set, so the manuscript does not describe a one-time locked test or independent validation. Nested MR resampling characterizes a separate specification and does not correct the complete original model search.

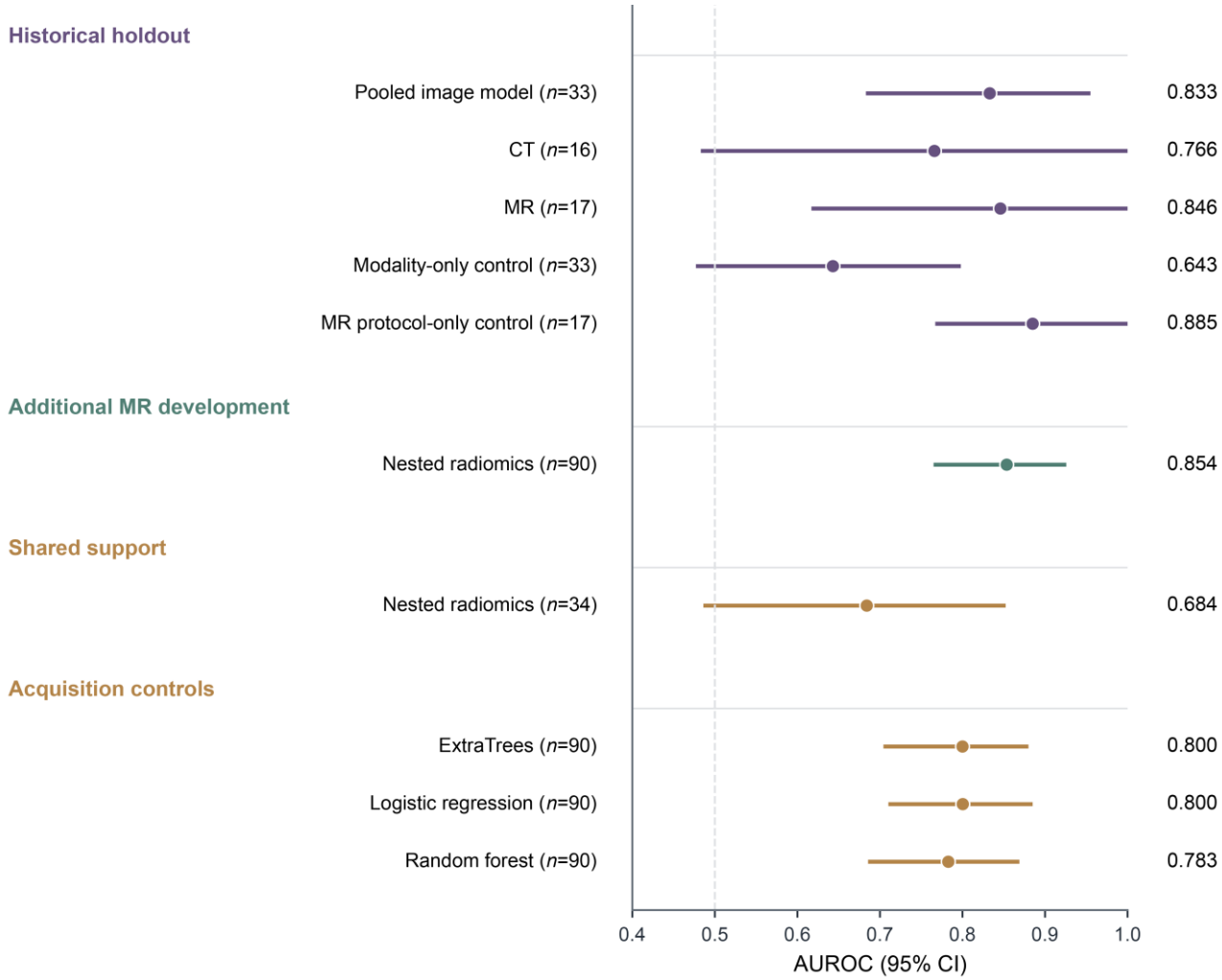
Supplementary Figure



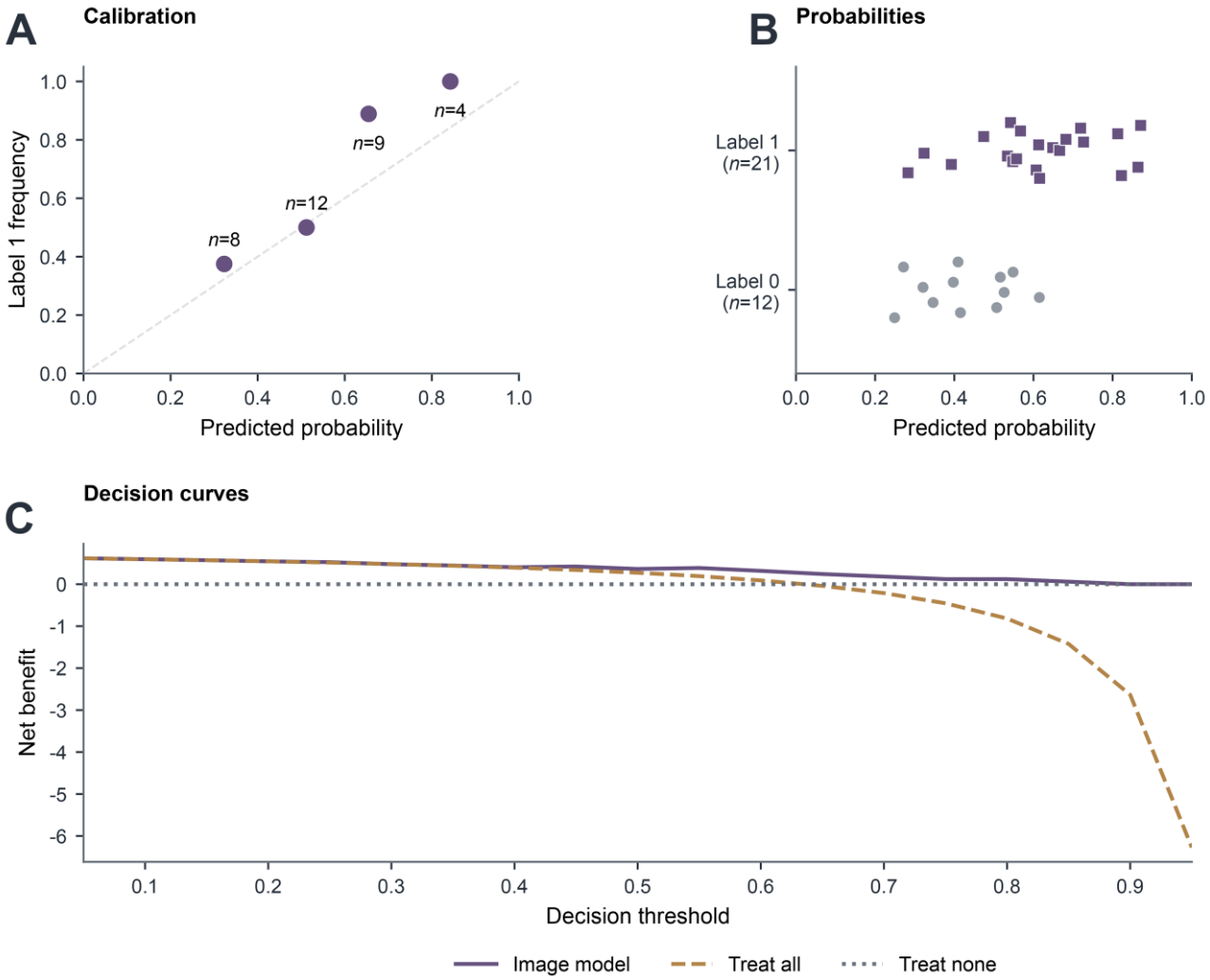
Supplementary Figure 1. Cluster-number assessment. (A-C) Original clustering at K = 2-9 and (D-F) patient-balanced clustering at K = 2-6. Columns show the Calinski-Harabasz index, Silhouette coefficient and Davies-Bouldin index. Higher Calinski-Harabasz and Silhouette values and lower Davies-Bouldin values are preferable. CT (computed tomography) and MR (magnetic resonance) are indicated by purple circles and green squares. Original clustering used 708,681 CT and 268,894 MR training voxels. Patient-balanced clustering used 145,174 CT and 110,876 MR training voxels from 61 and 67 patients, respectively, with at most 3,000 per patient. Patient-balanced Silhouette values were not calculated in the saved run (silhouette sample size 0), and panel (E) explicitly shows this absence; original Silhouette values are not substituted. Calinski-Harabasz axes use logarithmic scaling. These training-side indices do not identify a uniformly optimal or biologically defined K. All values are provided in Supplementary Table 1.



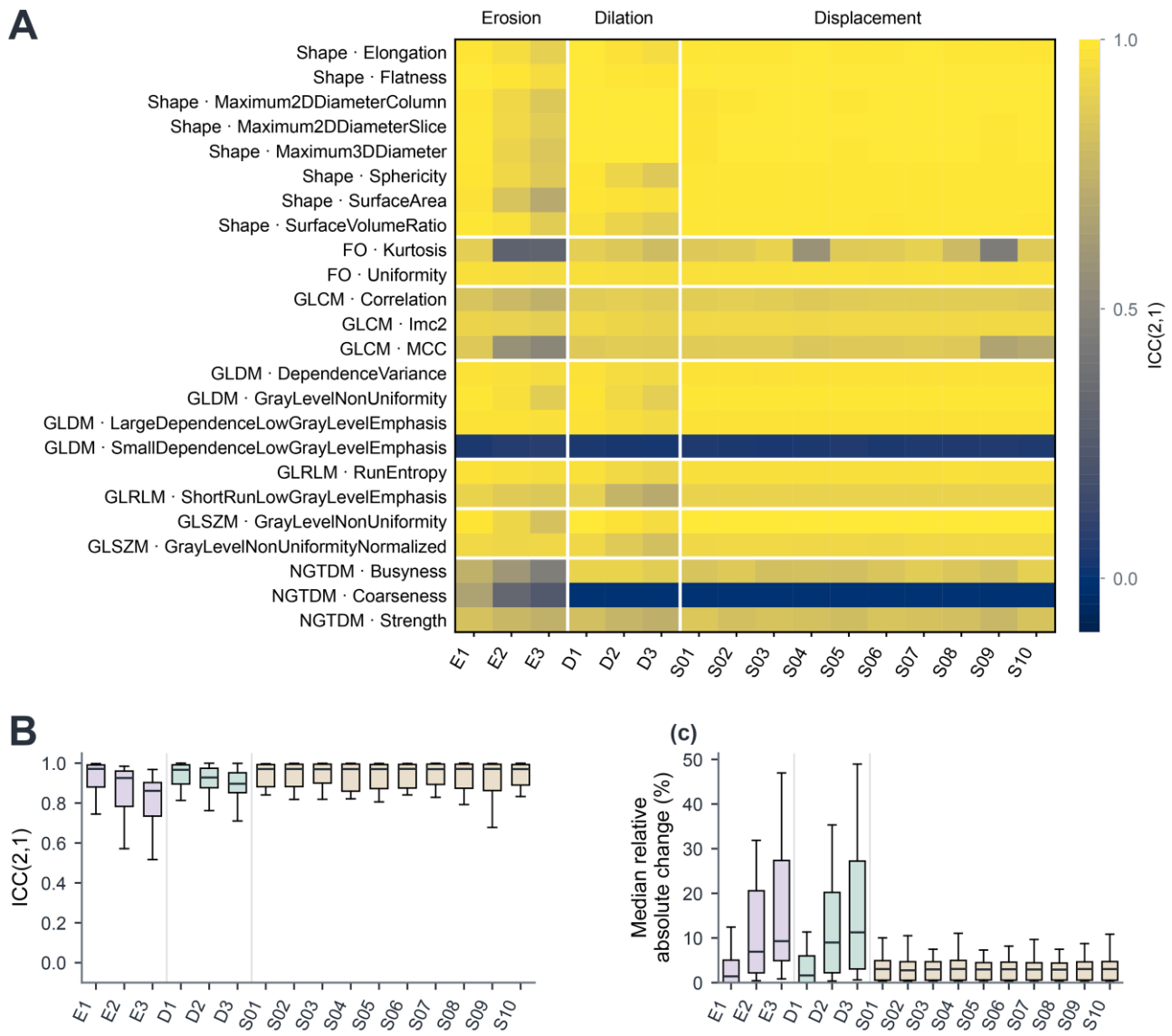
Supplementary Figure 2. Radiomic feature sets. (A) The archived fused K = 2 schema contains 20 whole-tumor and 153 habitat-summary columns, totaling 173 extracted columns after metadata columns are excluded; (B) The current MR re-extraction contains 107 features. The 24 model-linked features were identified across saved MR model manifests and used for the displayed contour-sensitivity summaries; they are not the selected features of the historical pooled classifier. The two schemas are not interchangeable, and the display does not imply historical feature equivalence, feature ranking, or biological mechanism.



Supplementary Figure 3. Performance uncertainty. Point estimates and 95% confidence intervals (CI) for four explicitly separated evidence layers: Historical internal holdout, Additional MR development, Common-support analysis, and Protocol-only controls. AUROC denotes area under the receiver operating characteristic curve; the vertical reference line denotes Chance. The revision development cohort overlaps the historical holdout in 14 patients. Historical holdout and additional MR estimates are not independent, are not pooled or interpreted as one validation cohort, and rows without an available interval are shown as point estimates only. CT denotes computed tomography and MR magnetic resonance.

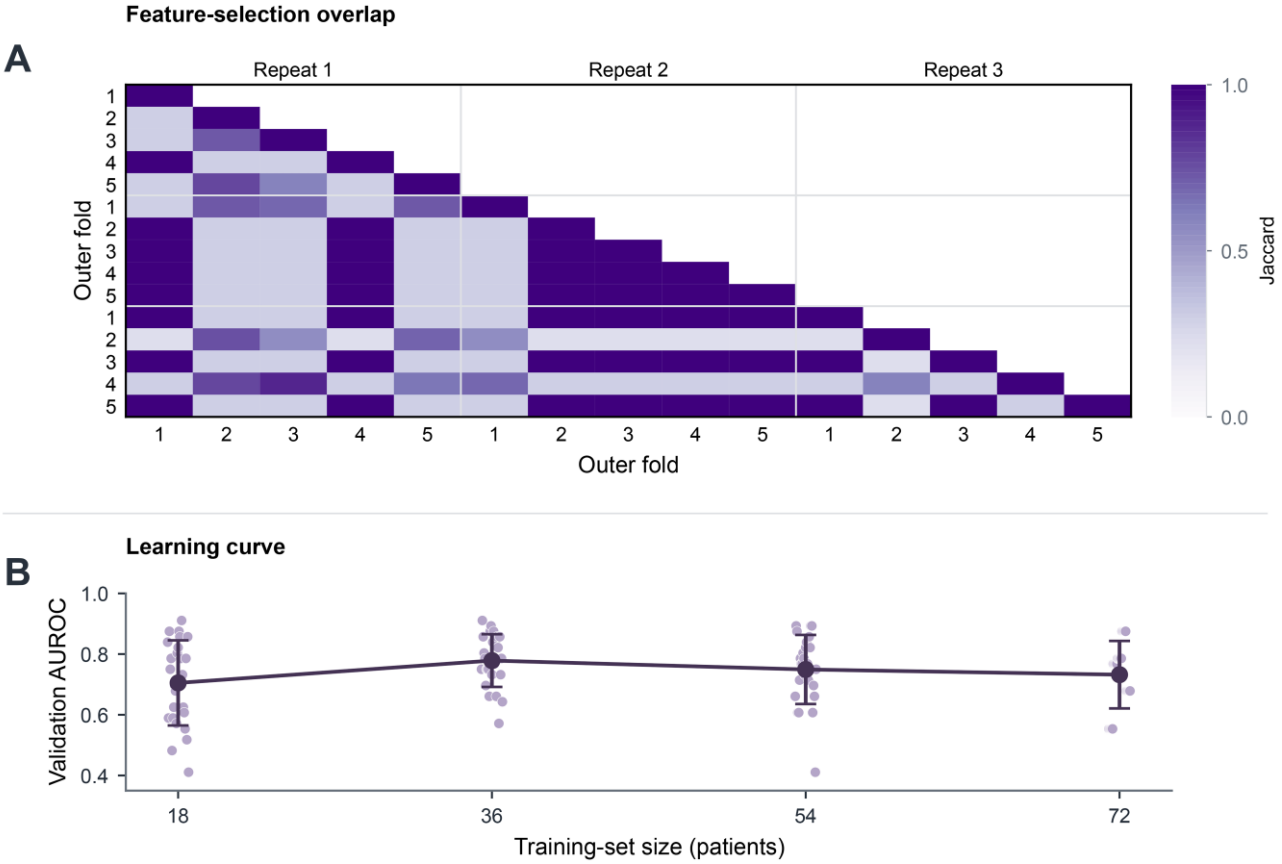


Supplementary Figure 4. Calibration and decision analysis. (A) Aggregate calibration-bin points with bin denominators (no fitted calibration curve); (B) individual frozen probabilities by stored label with deterministic vertical jitter for visibility; and (C) decision-curve analysis (DCA) for the n = 33 historical holdout over the complete evaluated threshold grid (0.05-0.95), including all negative treat-all net benefits. The stored summaries were Brier score 0.182 [95% CI (confidence interval) 0.143-0.225], ECE (expected calibration error) 0.100, and Hosmer-Lemeshow (HL) statistic 6.228 with stored P = 0.101; these values are retained in the caption/table rather than a fourth plot panel. Calibration is descriptive because the holdout is small and no prespecified clinical threshold exists; the display does not establish threshold validity, clinical utility, or deployment benefit. The dashed diagonal in (A) denotes ideal calibration.

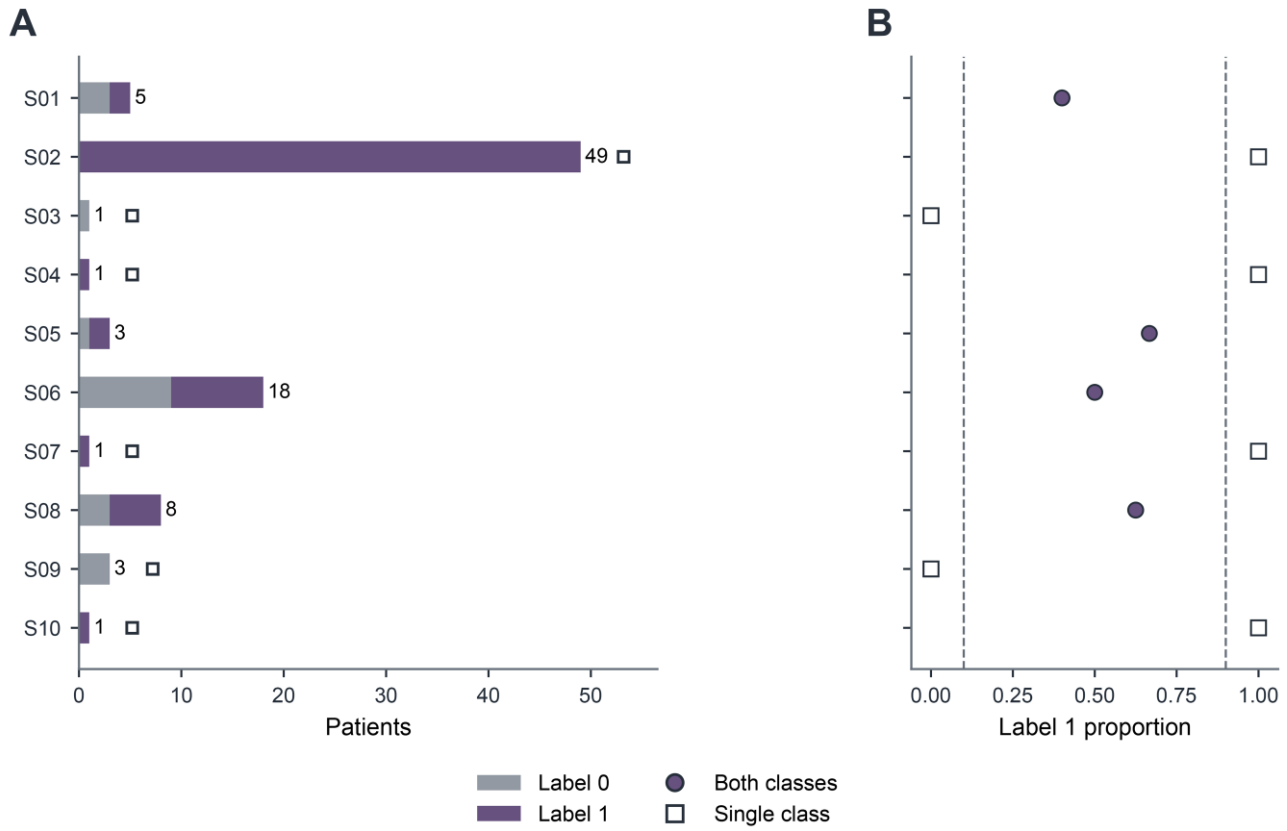


Supplementary Figure 5. ROI sensitivity. (A) Heatmap of the reported point ICC(2,1) values for 24 model-linked features across 16 non-baseline contour-perturbation conditions; (B) ICC(2,1) distributions and (C) median relative absolute changes, expressed as percentages, across the same conditions. The denominator is 207 MR image/mask pairs from 106 patients, evaluated with 17 mask variants per pair. Forty patients contribute more than one pair. The 24 model-linked features come from saved MR model manifests, not the historical pooled classifier. E1-E3 are erosion conditions, D1-D3 are dilation conditions, and S01-S10 are specified stochastic smooth-displacement conditions. ICC(2,1) denotes the reported single-measure intraclass correlation coefficient; FO, GLCM, GLDM, GLRLM, GLSZM, and NGTDM denote first-order, gray-level co-occurrence, gray-level dependence, gray-level run-length, gray-level size-zone, and neighborhood gray-tone-difference feature families, respectively. Boxes show the interquartile range with the median marked; outliers are suppressed for display. This is a computational contour-sensitivity

analysis, not inter-reader agreement, Dice against a reference mask, or model prediction robustness. ROI denotes region of interest. Whiskers extend to the most extreme observations within 1.5 interquartile ranges.



Supplementary Figure 6. Selection stability and learning curves. (A) Lower-triangular pairwise feature-selection Jaccard matrix across 15 outer models, grouped as Repeat 1-3 with five outer folds per repeat, from three repeats of nested MR development cross-validation on the separate current 107-feature representation; the axis terminology is Outer fold; (B) Raw AUROC (area under the receiver operating characteristic curve) points for 25 learning-curve replicates at each of four training-set sizes (18, 36, 54, and 72 patients), with the overlaid line connecting the four means and error bars denoting mean±standard deviation. The reserved 16-patient MR holdout of the revision split was excluded; 14 revision development patients overlapped the historical holdout. The display reports variability under a frozen current specification and does not provide independent validation, correct the historical route search, or establish prospective power. MR denotes magnetic resonance.



Supplementary Figure 7. Acquisition and label support. (A) Stored-label counts within the ten available development signatures; signatures with one class only are marked separately from signatures with both classes; (B) Proportion of stored label 1 within each signature, with 10% and 90% reference lines. Filled circles indicate both-class support, whereas open squares indicate single-class support; shape and fill provide redundant noncolor encoding. The plot uses neutral IDs S01-S10; the recorded sequence/phase signatures remain in the source table and are not reclassified, and no absent contrast-injection or timing metadata are inferred. The development denominator is 90 MR patients; the totals at the bar ends sum to this denominator.

Source	Evidence	Not established
Imaging archive	Image geometry	Uniform acquisition
DICOM examination	Sequence metadata	Cohort-wide metadata
Tumor masks	Automated segmentation	Reader agreement
Revision analyses	Computational sensitivity	Model-level ROI robustness
Study endpoint	Label association	Endpoint ascertainment

Supplementary Figure 8. Evidence scope. Source, evidence, and the claims not established are aligned in three columns. The imaging archive contains 2,503 NIfTI (Neuroimaging Informatics Technology Initiative) files from 212 patients and supports image geometry summaries, without establishing uniform acquisition. The available DICOM (Digital Imaging and Communications in Medicine) package contains 932 MR (magnetic resonance) images from one examination and 28 series; sequence metadata apply only to this package. Contrast-agent identity and injection timing were absent, so dynamic labels were not reclassified as arterial, portal-venous, or delayed phases. The segmentation archive contains 2,502 masks and 2,337 current image crops; current automated segmentation does not establish reader agreement or accuracy against a reference mask. Revision analyses include 207 image/mask tokens and computational contour sensitivity; the model-level ROI (region of interest) prediction-robustness branch was not run. Clinical records support association with the study-documented endpoint, without case-level endpoint ascertainment or collected clinical covariates.

Supplementary Tables

Supplementary Table 1. Cluster-validity results by construction

Clustering	Modality	K	Train cases	Train voxels	Calinski-Harabasz	Silhouette	Davies-Bouldin	Maps saved
Original	CT	2	61	708681	271130.548	0.332	1.228	77
Original	CT	3	61	708681	218127.45	0.263	1.282	77
Original	CT	4	61	708681	203234.572	0.248	1.187	77
Original	CT	5	61	708681	231288.524	0.233	1.266	77
Original	CT	6	61	708681	222926.947	0.279	0.885	77
Original	CT	7	61	708681	224376.693	0.241	1.072	77
Original	CT	8	61	708681	204565.751	0.228	1.188	77
Original	CT	9	61	708681	208113.689	0.241	1.177	77
Original	MR	2	67	268894	575633.494	0.689	0.378	84
Original	MR	3	67	268894	539876.903	0.628	0.96	84
Original	MR	4	67	268894	721680.074	0.709	0.49	84
Original	MR	5	67	268894	922195.274	0.74	0.672	84
Original	MR	6	67	268894	895414.616	0.736	0.824	84
Original	MR	7	67	268894	885342.42	0.755	0.794	84
Original	MR	8	67	268894	936276.254	0.778	0.753	84
Original	MR	9	67	268894	917414.527	0.787	0.726	84
Patient-balanced	CT	2	61	145174	58379.483	—	1.247371	77
Patient-balanced	CT	3	61	145174	52405.1081	—	1.006543	77
Patient-balanced	CT	4	61	145174	55507.8691	—	0.894294	77
Patient-balanced	CT	5	61	145174	54538.9288	—	1.032053	77
Patient-balanced	CT	6	61	145174	57263.6235	—	0.941041	77
Patient-	MR	2	67	110876	223081.0858	—	0.377343	84

balanced								
Patient-balanced	MR	3	67	110876	209003.124	—	0.950356	84
Patient-balanced	MR	4	67	110876	263082.2757	—	0.669828	84
Patient-balanced	MR	5	67	110876	328780.1354	—	0.683956	84
Patient-balanced	MR	6	67	110876	327528.0608	—	0.820126	84

Original clustering (K=2–9; 708,681 CT and 268,894 MR voxels) and patient-balanced clustering (K=2–6; 145,174 CT and 110,876 MR voxels) are labeled separately. Both use 61 CT and 67 MR training patients. Patient-balanced sampling is capped at 3,000 voxels per patient, giving theoretical maxima of 183,000 and 201,000; actual samples are smaller because some tumors contain fewer voxels. Patient-balanced Silhouette values are unavailable because the saved run used a silhouette sample size of 0. Calinski–Harabasz and Davies–Bouldin values are provided without substituting original-route values. Maps saved include application to training and holdout cases; only training voxels fit centroids. CT, computed tomography; MR, magnetic resonance. AUROC, area under the receiver operating characteristic curve; CI, confidence interval; CT, computed tomography; MR, magnetic resonance; CV, cross-validation; SD, standard deviation; ROI, region of interest. Missing results are shown as unavailable.

Supplementary Table 2. Pooled-model CT/MR subgroup and macro-AUROC metrics

Subgroup	Label	*n*	Label 1	Label 0	Auc	Ci low	Ci high	Balanced accuracy	Brier	Model identity
overall	Overall	33	—	—	0.833333	0.683	0.955	0.738095	0.182158	Patient-balanced pooled K=2 ExtraTrees enet10
CT	CT	16	8.0	8.0	0.765625	0.483	1.0	0.75	0.211186	Patient-balanced pooled K=2 ExtraTrees enet10
MR	MR	17	13.0	4.0	0.846154	0.617	1.0	0.586538	0.154837	Patient-balanced pooled K=2 ExtraTrees enet10
macro_CT_MR	Macro CT/MR	33	—	—	0.805889	—	—	0.668269	0.183011	Patient-balanced pooled K=2 ExtraTrees enet10

Overall, CT, MR, and macro CT/MR metrics from the same patient-balanced pooled candidate, with denominators and available secondary metrics. The macro AUROC is the unweighted mean of the CT and MR subgroup AUROCs. AUROC, area under the receiver operating characteristic curve; CI, confidence interval; CT, computed tomography; MR, magnetic resonance; CV, cross-validation; SD, standard deviation; ROI, region of interest. Missing results are shown as unavailable.

Supplementary Table 3. Historical route and search inventory

Row type	Status	Rows	Interpretation
baseline_control	VERIFIED	16	Repeated baseline/control evidence rows
cluster_configuration	VERIFIED	26	Train-side habitat construction/validity rows
model_configuration	EXECUTED	1882	Shared or derived historical evidence layer
model_configuration	VERIFIED	26	Shared or derived historical evidence layer

Aggregate route/search evidence layers and the full editable inventory source. The 1,908 model-configuration rows, 26 cluster-construction rows, and 16 repeated baseline/control rows contain shared, derived, copied, or repeated evidence and are not independent replications. AUROC, area under the receiver operating characteristic curve; CI, confidence interval; CT, computed tomography; MR, magnetic resonance; CV, cross-validation; SD, standard deviation; ROI, region of interest. Missing results are shown as unavailable.

Supplementary Table 4. MR protocol sensitivity and support boundary

Experiment	Protocol subset	*n* train	*n* test	Holdout auc	Role	Support note
whole_plus_habitat_T1_VIBE_only	T1_VIBE_FS	22	7	0.833333	Historical sensitivity only	—
patient_balanced_T1_VIBE_only	T1_VIBE_FS	22	7	0.916667	Historical sensitivity only	—
MR_OTHER_support	MR_OTHER	45	10	—	Support failure	10 label-1 / 0 label-0; AUROC not estimable

T1_VIBE_FS sensitivity routes (*n*=7; primary-route AUROC 0.833 and patient-balanced-route AUROC 0.917) and the MR_OTHER one-class support boundary (10 stored label-1 / 0 stored label-0; AUROC not estimable). T1_VIBE_FS is exploratory and neither result establishes general MR robustness or protocol-independent performance. AUROC, area under the receiver operating characteristic curve; CI, confidence interval; CT, computed tomography; MR, magnetic resonance; CV, cross-validation; SD, standard deviation; ROI, region of interest. Missing results are shown as unavailable.

Supplementary Table 5. Calibration and decision-curve metadata

Record type	Item	Value	Ci or range	*n*	Note
calibration_summary	Brier score	0.182	0.143–0.225	33	Stored descriptive summary
calibration_summary	ECE (5 bins)	0.1	—	33	Stored descriptive summary
calibration_summary	Hosmer–Lemeshow statistic	6.228	p=0.101	33	Stored descriptive summary
dca_metadata	Threshold grid	0.05–0.95	step 0.05	33	Aggregate display derived from frozen pooled predictions; descriptive only

Stored calibration summaries and aggregate decision-curve metadata for the frozen pooled holdout predictions. The values are descriptive and are not evidence of clinical utility or deployment benefit. AUROC, area under the receiver operating characteristic curve; CI, confidence interval; CT, computed tomography; MR, magnetic resonance; CV, cross-validation; SD, standard deviation; ROI, region of interest. Missing results are shown as unavailable.

Supplementary Table 6. Feature-family and extraction-provenance notes

Feature set	Arm	*K*	Cases	Features	Missing values	Provenance note
Whole tumor	CT	0	77	20	0	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Whole tumor	MR	0	84	20	19	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Whole tumor	pooled	0	161	20	19	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-habitat	CT	2	77	42	69	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-habitat	CT	3	77	63	307	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-habitat	CT	4	77	84	937	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-habitat	CT	5	77	105	1032	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-habitat	CT	6	77	126	1196	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-habitat	MR	2	84	42	113	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-	MR	3	84	63	156	Family-level aggregate; exact ROI/configuration semantics remain

habitat						unresolved
Patient-balanced per-habitat	MR	4	84	84	351	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-habitat	MR	5	84	105	590	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-habitat	MR	6	84	126	633	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-habitat	pooled	2	161	42	182	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-habitat	pooled	3	161	63	463	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-habitat	pooled	4	161	84	1288	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-habitat	pooled	5	161	105	1622	Family-level aggregate; exact ROI/configuration semantics remain unresolved
Patient-balanced per-habitat	pooled	6	161	126	1829	Family-level aggregate; exact ROI/configuration semantics remain unresolved

Feature-family counts, missingness, and provenance notes for the archived representation. The supplementary feature inventory lists the 173 saved K=2 pooled feature columns, but the excluded-feature universe and exact executed configuration remain unavailable. The table does not provide feature-ranking, causal, mechanistic, or segmentation-reproducibility evidence; ROI/configuration semantics remain unresolved. AUROC, area under the receiver operating characteristic curve; CI, confidence interval; CT, computed tomography; MR, magnetic resonance; CV, cross-validation; SD, standard deviation; ROI, region of interest. Missing results are shown as unavailable.

Supplementary Table 7. Clinical covariate data-collection audit

Variable	Linked patient-level data in current archive	Definition or time point	Use in current revision	Interpretation boundary
Age	Not collected in project table	Not collected	No label-group or split comparison	No inference about age balance
Sex	Not collected in project table	Not collected	No label-group or split comparison	No inference about sex balance
HBV status	Not collected in project table	Not collected	No clinical adjustment	No inference about HBV confounding
HCV status	Not collected in project table	Not collected	No clinical adjustment	No inference about HCV confounding
Alcohol-related liver disease	Not collected in project table	Not collected	No clinical adjustment	No inference about alcohol-related disease balance
Metabolic liver disease	Not collected in project table	Not collected	No clinical adjustment	No inference about metabolic disease balance
Cirrhosis	Not collected in project table	Not collected	No clinical adjustment	No inference about cirrhosis balance
Liver function	Not collected in project table	Not collected	No clinical adjustment	No inference about liver-function balance
Alpha-fetoprotein	Not collected in project table	Not collected	No clinical adjustment	No inference about AFP balance
Tumor size	Not collected in project table	Not collected	No clinical adjustment	No inference about tumor-burden balance
Tumor number or multifocality	Not collected in project table	Not collected	No clinical adjustment	No inference about lesion-count balance
Tumor stage	Not collected in project table	Not collected	No clinical adjustment	No inference about stage balance
Histological	Not collected in	Not collected	No clinical	No inference about

differentiation	project table		adjustment	differentiation balance
Prior treatment	Not collected in project table	Not collected	No clinical adjustment	No inference about treatment balance
Imaging institution or center	Not collected in project table	Not collected	No clinical adjustment or center balance analysis	No inference about center composition

The table records that the project data used for this analysis did not collect the clinical variables requested by the reviewers. This means that a Cs-group baseline comparison, train/holdout clinical balance analysis, and adjusted clinical-radiomic modeling are outside the current data boundary; it is not evidence that the groups are clinically balanced. AUROC, area under the receiver operating characteristic curve; CI, confidence interval; CT, computed tomography; MR, magnetic resonance; CV, cross-validation; SD, standard deviation; ROI, region of interest. Missing results are shown as unavailable.

Supplementary Table 8. Additional MR sensitivity and robustness results

Experiment	Denominator design	Verified result	Interpretation boundary
REV-A-RADIOMICS-REPRO	207 matched current image/mask tokens	207 current-setting re-extractions; 0 extraction failures	Current configuration executable for matched pairs; historical PyRadiomics YAML/version/command/full environment unavailable; shape fields revision-only
REV-B-ROI-PERTURB	207 MR pairs from 106 patients (40 with multiple pairs); baseline plus 16 fixed/stochastic perturbations (17 masks/token)	3519 long-form rows; 1712 valid feature-variant summaries; selected-feature median ICC(2,1) 0.8613-0.9713; median relative change 0.0141-0.1126; descriptive pair-level log-volume vs instability Spearman rho -0.3109, n=207; no unclustered patient-level inference	Synthetic feature-level contour sensitivity; pair-level uncertainty does not account for patient clustering; not reader agreement, reference-mask Dice/ICC, or prediction robustness; model-level prediction was not performed.
REV-C-NESTED-STABILITY	MR development n=90 (70 label-1, 20 label-0); 3 repeats x 5 outer folds; 2000 bootstrap	AUROC of probabilities averaged across repeats 0.8536; 95% CI 0.7650-0.9257; repeat range 0.7550-	Nested evaluation of patient-averaged 107 current MR original features; not the historical 173-column fused model; reserved revision holdout excluded; 14 development patients overlap the

	replicates	0.8771; median pairwise feature-selection Jaccard 0.2991 (IQR 0.2991-1.0000)	historical holdout; upstream K=2 frozen; not a historical search correction
REV-D-LEARNING-CURVE	MR development n=90; 25 replicates at each fraction; training n=18,36,54,72	Mean AUC +/- SD: 25% 0.7050+/-0.1402; 50% 0.7786+/-0.0869; 75% 0.7493+/-0.1138; 100% 0.7321+/-0.1112	Sample-size sensitivity only; not prospective/post-hoc power analysis or external validation
REV-E-PROTOCOL-SENSITIVITY	MR development n=90; 10 sequence/phase signatures; 2000 bootstrap replicates	Protocol-only AUC: ExtraTrees 0.8000 (CI 0.7042-0.8799); logistic regression 0.8004 (CI 0.7102-0.8850); random forest 0.7829 (CI 0.6856-0.8689); common support n=34 (18 label-1, 16 label-0); nested radiomics AUC 0.6840 (95% CI 0.4861-0.8523)	Protocol-label coupling visible; no defensible cohort-wide scanner/site/batch variable; no protocol-independent or external claim

Outputs for current-setting radiomics re-extraction (REV-A), fixed ROI perturbation feature sensitivity (REV-B), development-only nested reevaluation (REV-C), learning-curve sensitivity (REV-D), and protocol-signature/common-support analysis (REV-E). The table records the analysis denominator, point estimates, bootstrap settings where applicable, and the specific unresolved boundary for each result. REV-B model-level perturbation prediction was not performed; no result in S8 is a human-reader agreement, external-validation, or clinical-adjustment result. AUROC, area

under the receiver operating characteristic curve; CI, confidence interval; CT, computed tomography; MR, magnetic resonance; CV, cross-validation; SD, standard deviation; ROI, region of interest. Missing results are shown as unavailable.