Journal for ImmunoTherapy of Cancer

Interpretable multimodal radiopathomics model predicting pathological complete response to neoadjuvant chemoimmunotherapy in esophageal squamous cell carcinoma

Baojia Qi,[1,2] Zhaoyu Jiang,[1,3] Haixia Shen,[1,2] Jiacheng Li,[4] Zhixiang Wang,[5] Min Fang,[1] Changchun Wang,[1] Youhua Jiang,[1] Jingping Yuan,[6] Inigo Bermejo,[7] Andre Dekker,[2] Dirk De Ruysscher,[2] Leonard Wee,[2] Wencheng Zhang,[4] Yongling Ji,[1] Zhen Zhang[1,2,4]

To cite: Qi B, Jiang Z, Shen H, *et al*. Interpretable multimodal radiopathomics model predicting pathological complete response to neoadjuvant chemoimmunotherapy in esophageal squamous cell carcinoma. *Journal for ImmunoTherapy of Cancer* 2025;**13**:e013840. doi:10.1136/jitc-2025-013840

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/jitc-2025-013840).

BQ and ZJ are joint first authors.

Accepted 05 December 2025

Check for updates

For numbered affiliations see end of article.

Correspondence to
Dr Zhen Zhang; zhen.zhang@maastro.nl

Dr Yongling Ji; jiyl@zjcc.org.cn

ABSTRACT

Background Accurate preoperative prediction of pathological complete response (pCR) following neoadjuvant chemoimmunotherapy (nCIT) could help individualize treatment for patients with esophageal squamous cell carcinoma (ESCC). This study aimed to develop and externally validate an interpretable multimodal machine learning framework that integrates CT radiomics and H&E-stained whole-slide images pathomics to predict pCR.

Methods In this multicenter, retrospective study, 335 patients with ESCC who received nCIT followed by esophagectomy were enrolled from three institutions. Patients from one center were divided into a training set (181 patients) and an internal test set (115 patients), while data from the other two centers comprised an external test set (39 patients). We developed unimodal radiomics and pathomics models, and two multimodal fusion models—an intermediate fusion model (MIFM) and a late fusion model (MLFM). Model performance was evaluated using the area under the curve (AUC), accuracy, sensitivity, specificity, and F1 score, with exploratory survival stratification by observed and model-predicted pCR status. Interpretability was treated as a design constraint and operationalized at both the feature and model levels.

Results The MIFM outperformed unimodal models and the MLFM across all cohorts, achieving AUC/accuracy/sensitivity/specificity/F1 score of 0.97/0.93/0.84/0.96/0.86 (training set), 0.78/0.87/0.62/0.93/0.63 (internal test set), and 0.76/0.77/0.54/0.88/0.61 (external test set). Both observed and predicted pCR status showed exploratory prognostic stratification for overall survival. Feature definitions were mathematically or morphologically explicit, and case-level/cohort-level explanations together with decision-pathway views provided insights into model reasoning. We additionally provide a user-friendly Graphical User Interface to facilitate clinical practice.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Neoadjuvant chemoimmunotherapy (nCIT) is a promising treatment for esophageal squamous cell carcinoma (ESCC), but accurately predicting pathologic complete response (pCR) remains challenging. Traditional biomarkers have limited predictive capacity and are hindered by high detection costs and operational complexity. Although the role of multimodal radiopathomics in predicting treatment outcomes has been studied in various cancers, its application in nCIT remains limited. Furthermore, the interpretability of predictive models requires further exploration.

WHAT THIS STUDY ADDS

⇒ This study developed a multimodal radiopathomics model that predicts pCR in patients with ESCC following nCIT by integrating CT-based radiomics and whole-slide images-based pathomics features. The proposed model demonstrated superior performance over unimodal models, achieving high area under the curve, accuracy, sensitivity, specificity, and F1-score across multiple validation cohorts. Interpretability was treated as a design constraint and operationalized at both the feature and model levels. A user-friendly Graphical User Interface is additionally provided to facilitate clinical practice.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study highlights the potential of multimodal radiopathomics model to improve clinical decision-making for ESCC. The model's ability to predict pCR could guide individualized decisions between surveillance and timely surgery. Further refinement and validation through large-scale prospective trials remain essential to establish its utility in clinical practice.

Conclusions We developed and externally validated an interpretable radiopathomics fusion framework that predicts pCR after nCIT in ESCC using standard-of-care data. This model holds promise as an effective tool for guiding individualized decisions between surveillance and timely surgery.

INTRODUCTION

Esophageal squamous cell carcinomas (ESCC) remain one of the most prevalent and aggressive cancers worldwide.[1] Neoadjuvant chemoradiotherapy (nCRT) followed by surgery is the current standard care for locally advanced ESCC.[2 3] Recent clinical trials, however, have highlighted neoadjuvant chemoimmunotherapy (nCIT) followed by surgery as a promising alternative, reporting R0 resection rates ranging from 80.9% to 98.0% and pathologic complete response (pCR) rates between 16.7% and 50.0%.[4 5] A prospective study further suggested that, compared with nCRT, nCIT may yield superior 2-year overall survival (OS) and disease-free survival (DFS) despite similar pCR rates (22.9% vs 25.9%).[6] Achieving pCR correlates with improved long-term survival outcomes and may permit the implementation of watch-and-wait strategies, thereby preserving organ functionality and enhancing quality of life.[4 7 8] Consequently, accurate preoperative prediction of pCR following nCIT is critical for identifying suitable candidates and personalizing therapeutic approaches.

Despite this clinical need, robust biomarkers capable of accurately predicting pCR to nCIT require further exploration. Established tissue biomarkers, including microsatellite instability,[9 10] programmed cell death ligand-1 (PD-L1) expression,[11 12] and tumor mutational burden (TMB),[13–15] have limited predictive capacity and are hindered by high detection costs and operational complexity. Therefore, there is an urgent need to develop accessible, reliable, and cost-effective predictive tools.

Medical imaging provides rich macro-scale and micro-scale information that is well suited to artificial intelligence (AI)-based prediction. Macroscopic radiologic images (eg, contrast-enhanced CT) and microscopic histopathological images (H&E-stained whole-slide images (WSIs)) are complementary, and multimodal fusion may improve predictive accuracy.[16] Radiomics and pathomics enable quantitative characterization of tumor phenotype and microenvironment respectively and have shown promise in outcome prediction across cancers, including ESCC.[17–19] Building on prior work demonstrating the feasibility of radiomics-based pCR prediction following nCIT,[20] and evidence that nuclei-level morphology and texture carry prognostic information,[21 22] integrating radiomics and pathomics features represents a rational strategy to enhance preoperative prediction of pCR in ESCC.

Translating such multimodal predictors into practice requires more than accuracy. Accordingly, we treat interpretability as a design constraint and frame it along two axes—model-level and feature-level. At the model level, we prioritize algorithms with auditable decision functions and stable post-hoc explanations (eg, Shapley-value attribution), enabling visualization of case-level and cohort-level contributions while mitigating the black-box concerns typically associated with deep neural networks.[23] At the feature level, we emphasize mathematically defined radiomics features and explicitly defined pathomics descriptors of nuclear and tissue architecture (eg, nuclear area, eccentricity, perimeter, chromatin texture), selected for their clear clinical semantics and communicability to clinicians.

In this study, we developed an interpretable multimodal machine learning framework to preoperatively predict pCR to nCIT in ESCC using data from three independent patient cohorts. We systematically benchmarked multiple machine learning algorithms and fusion strategies to integrate CT-based radiomics and WSI-based pathomics features, while formalizing transparency at both the model and feature levels. To facilitate clinical communication and workflow fit, we specified case-level and cohort-level explanatory outputs (eg, contribution-based attributions) and implemented a user-facing software prototype to illustrate potential applicability and practicality.

METHODS

Given the retrospective design, the informed consent requirements were waived. The study adhered to the principles of the Declaration of Helsinki and followed established methodological guidance for radiomics research.[24] To promote methodological rigor and transparency, we evaluated protocol adherence using a 12-item methodology-evaluation checklist that we previously proposed.[25] The checklist scoring sheet is provided in the online supplemental table 1. The overarching study flow is presented in figure 1.

Patient enrollment

Consecutive patients with histologically confirmed ESCC who received nCIT followed by curative-intent esophagectomy were retrospectively identified across three academic medical centers—Zhejiang Cancer Hospital, Renmin Hospital of Wuhan University, and Tianjin Medical University Cancer Institute and Hospital—from July 2019 to July 2023 (n=335). At Zhejiang Cancer Hospital (n=296), patients were randomly allocated 6:4 to a training set and an independent internal validation cohort (Test-set-1). The external validation cohort (Test-set-2, n=39) comprised patients treated at Renmin Hospital of Wuhan University from July 2020 to September 2023 (n=22) and at Tianjin Medical University Cancer Institute and Hospital from June 2020 to February 2022 (n=17). For each patient, a contrast-enhanced chest CT was acquired within 14 days prior to nCIT initiation, and H&E-stained WSIs were digitized from pretreatment endoscopic biopsy specimens obtained within 7days of the CT. Detailed inclusion and exclusion criteria and a patient selection flowchart are provided in online supplemental file A1 and figure 1.

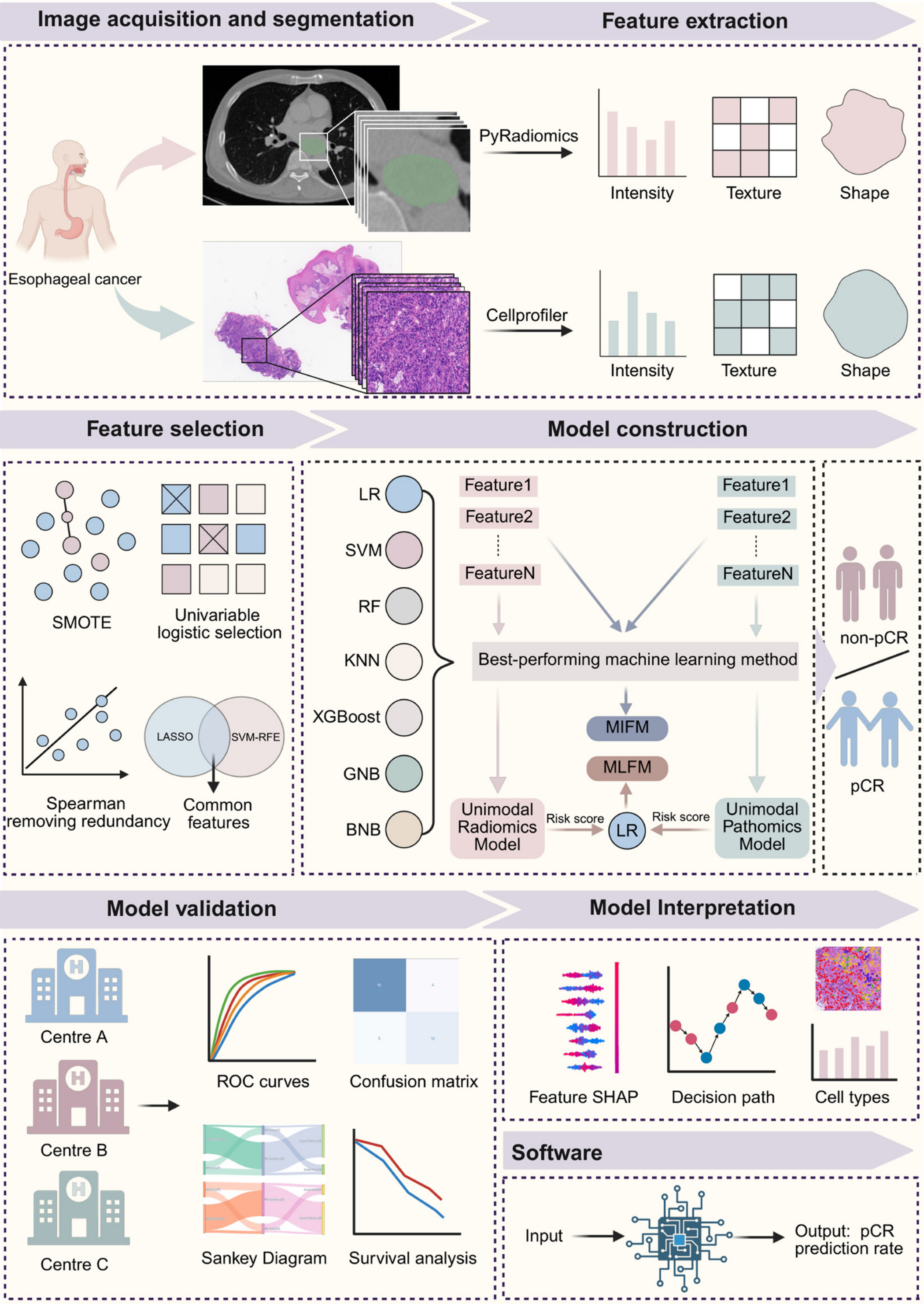

Figure 1 Study pipeline. Preoperative contrast-enhanced CT and H&E-stained WSIs from 335 patients with ESCC across three centers were analyzed. The tumors were manually contoured on CT images and tumor-rich fields were selected on WSIs. Radiomics (PyRadiomics) and pathomics (CellProfiler) features were extracted and screened. Four predictors were built—unimodal radiomics, unimodal pathomics, MIFM, and MLFM—and evaluated in training, internal, and external cohorts using ROC curves, confusion matrix, reclassification Sankey diagrams, and survival analysis. Interpretability analyses included SHAP analysis, case-level decision-pathway views and cell-type quantification. A browser-based Graphical User Interface accepts the CT/ROI and CellProfiler inputs and outputs the patient-level pCR probability. BNB, Bernoulli Naïve Bayes; ESCC, esophageal squamous cell carcinoma; GNB, Gaussian Naïve Bayes; KNN, k-nearest neighbors; LASSO, Least Absolute Shrinkage and Selection Operator; LR, logistic regression; MIFM, multimodal intermediate fusion model; MLFM, multimodal late fusion model; pCR, pathologic complete response; RF, random forest; ROC, receiver operating characteristic; ROI, region of interest; SHAP, SHapley Additive exPlanations; SMOTE, Synthetic Minority Over-sampling Technique; SVM-RFE, Support Vector Machines-Recursive Feature Elimination; XGBoost, eXtreme Gradient Boosting; WSIs, whole-slide images.

Treatment protocol and pathological evaluation

Patients received at least one cycle of neoadjuvant immunotherapy concurrently with chemotherapy. Immunotherapy consisted of standard doses (200 mg every 3 weeks per cycle) of programmed cell death protein 1 or PD-L1 monoclonal antibodies (tislelizumab, sintilimab, durvalumab, envafolimab, pembrolizumab, camrelizumab, or nivolumab). Platinum-based chemotherapy employed two-drug regimens: (1) TC regimen (every 3 weeks): one to four cycles of nab-paclitaxel 260 mg/m^2 (day 1) or paclitaxel 135–175 mg/m^2 (day 1) + carboplatin area under the curve (AUC) 5 mg/mL/min (day 1) every 21 days; (2) TP regimen (every 3 weeks): one to four cycles of nab-paclitaxel 260 mg/m^2 (day 1) or paclitaxel 175 mg/m^2 (day 1) + cisplatin 75 mg/m^2 (day 1).

Radical esophagectomy was undertaken 4–8 weeks after completion of nCIT. Surgical approach (minimally invasive or open) and lymphadenectomy extent (two-field or three-field) were determined by tumor location and surgeon assessment.

Resected specimens were examined by an experienced pathologist and reviewed by a senior esophageal cancer pathologist. Tumor regression grade (TRG) was classified according to the College of American Pathologists Esophageal Carcinoma Protocol[26]: TRG 0 (no histologically identifiable cancer cells); TRG 1 (single cell or rare small groups of cancer cells); TRG 2 (residual cancer with evident tumor regression but more than single cell or rare small groups of cancer cells); TRG 3 (extensive residual cancer with no evident tumor regression). pCR was defined as TRG 0 at the primary site, with TRG 1–3 being classified as non-pCR. This pCR/non-pCR binary outcome served as the prespecified endpoint for model development and evaluation.

Imaging acquisition and segmentation

CT acquisition parameters from the three centers are summarized in online supplemental table 2. Two physicians (HS, XW), each with over 3 years of experience, performed manual segmentation of the primary esophageal tumors on CT images to generate regions of interest (ROIs). Assessors were blinded to pathological outcomes and model outputs. All contours were subsequently reviewed and, when necessary, refined by a senior physician (YJ) with over 25 years of experience. Any discrepancies were resolved by consensus adjudication, and the finalized ROIs served as the ground truth for radiomics feature extraction. Segmentations were performed using 3D Slicer software (V.5.1.0).[27]

Formalin-fixed, paraffin-embedded H&E-stained slides were scanned at 20×magnification and digitized into WSIs. For each WSI, a thoracic pathologist with 3 years of experience (BQ), blinded to clinical outcomes, selected five representative tumor-rich fields of view (FOVs). Each FOV was cropped into a 512×512-pixel patch and saved in PNG format. All patches were visually inspected to guarantee their quality. Visual quality control (QC) was performed to exclude patches with over/under-staining, folds, chatter, inadequate tissue, air bubbles, pen marks or stripping artifacts.

Feature extraction and selection

Radiomics features were computed from the finalized CT ROIs using PyRadiomics[28] (V.3.0.1). A total of 1,094 features were extracted, encompassing shape and size descriptors, first-order intensity statistics, and multiple texture families—gray-level co-occurrence matrix, gray-level size zone matrix, gray-level run length matrix, gray-level dependence matrix and neighboring gray-tone difference matrix—together with wavelet-derived features.

Pathomics features were quantified from H&E-stained WSIs using CellProfiler[29] (V.4.2.8) via an automated pipeline that measures intensity distributions, neighborhood relationships, morphological/shape attributes, texture statistics, and areas-fraction metrics (details in online supplemental file A2 and figure 2). For each case, features were calculated on the five 512×512-pixel patches and averaged to obtain slide-level descriptors, yielding a total of 4,892 quantitative pathomics features covering nuclear, cytoplasmic, and tissue-level characteristics.

Feature selection was conducted independently for radiomics and pathomics, with all procedures confined to the training set. The selection workflow comprised the following steps: first, we used the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. Standardization of extracted radiomics and pathomics features was carried out using Z-Scores (original value−mean value/SD). Then, univariate logistic regression was implemented to identify features with a p value <0.01 (for pathomics features)/0.05 (for radiomics features) for subsequent analysis. Spearman correlation coefficients (ρ) were computed for each pair of features. Redundancy was reduced by computing pairwise Spearman correlations and, for any pair with $|\rho| > 0.85$, retaining the feature showing the stronger association with the outcome. Finally, two selectors—Least Absolute Shrinkage and Selection Operator (LASSO) with 10-fold cross-validation and Support Vector Machines-Recursive Feature Elimination (SVM-RFE)—were applied separately, and their intersection constituted the modality-specific feature set used for final model construction. A schematic of this pipeline is provided in online supplemental figure 3.

Model construction and validation

For unimodal modeling, we evaluated seven machine learning algorithms—logistic regression, Gaussian/Bernoulli Naïve Bayes, SVM, random forests, K-nearest neighbors, and eXtreme Gradient Boosting (XGBoost)—for the radiomics and pathomics feature sets. Hyperparameter optimization employed grid search with fivefold cross-validation. A fixed random seed was applied throughout parameter tuning to ensure reproducibility. Key hyperparameters are shown in the online supplemental file A3.

For multimodal learning, we considered two fusion strategies. For the multimodal intermediate fusion model (MIFM), radiomics and pathomics features were concatenated into a joint representation, and the algorithm identified as optimal in the unimodal screen was used to fit the fused model on the development set. For the multimodal late fusion model (MLFM), the best-performing radiomics model and pathomics model from the unimodal stage were first trained on the development set. Their probabilistic risk scores were then used as inputs to a logistic regression model. In total, we established four types of models, including the unimodal radiomics models, unimodal pathomics models, MIFM, and MLFM.

External validation was performed on the original Test-set-1 and the Test-set-2 with no further tuning. Test-set data was processed strictly through the same preprocessing pipeline fitted on the development set, and no data augmentation (eg, SMOTE), no additional normalization or standardization operations were applied to test datasets. Predictive performance was quantified by the area under the receiver operating characteristic (ROC) curve (AUC), accuracy, sensitivity, specificity, and F1 score. To obtain CIs, we used 1,000-iteration bootstrap resampling of each test cohort and reported 95% CIs for all metrics. ROC and precision-recall (PR) curves were plotted for visual comparison. Furthermore, for the best-performing model, decision curve analysis (DCA) was performed to assess its potential clinical utility.

Model interpretation

To interpret the radiomics and pathomics feature contributions to the models, we applied SHapley Additive exPlanations (SHAP) analysis, which quantifies individual feature influence on probability of pCR at both the case level and the cohort level. The Shapley value is defined as follows:

$$\varnothing_j = \sum_{S \subseteq N\{j\}} \frac{|S|!\left(N - |S| - 1\right)!}{N!}\left(v\left(S \cup \{j\}\right) - v\left(S\right)\right)$$

where: $v\left(S \cup \{j\}\right) - v\left(S\right)$ is the specific contribution of j to the coalition S; $\sum\left[S \subseteq N\{j\}\right]$ is the summation over all possible coalitions; and $\left(\left|S\right|!\left(N - \left|S\right| - 1\right)!\right)/N!$ is the weight factor of this particular coalition. Case-level plots display how a patient's specific feature values increase or decrease the predicted risk relative to the model baseline, whereas cohort-level summaries rank features by overall impact and indicate the dominant direction of effect.

To provide a stepwise view of the model's decision logic, we derived patient-specific decision pathways from the trained model. For each patient, the pathway enumerates the sequence of decision rules applied by the model, annotating the feature and threshold at each step and the incremental change in the prediction (on the log-odds/probability scale) contributed by that step, culminating in the final predicted probability. This visualization clarifies which features—and which value ranges—most strongly pushed the prediction toward pCR or non-pCR for that specific case, thereby linking global feature importance with case-level rationale in a single view. The workflow for constructing the contribution plots and decision-path visualizations is presented as a flowchart in online supplemental figure 4.

To characterize the tumor microenvironment (TME) on H&E-stained WSIs, we implemented Hover-Net,[30] an open-source deep learning network model, for nuclear segmentation and coarse cell-type assignment within the analyzed patches. This framework identified the following cell types: tumor cells, lymphocytes, connective cells, necrotic cells, and other cells. Finally, cell-type fractions were compared based on observed pCR status (observed pCR vs observed non-pCR), and separately, based on model-predicted status (predicted pCR vs predicted non-pCR).

Statistical analysis

Patient characteristics were evaluated using SPSS V.27. Pearson's chi-square/Likelihood-ratio tests were applied to compare categorical variables. ANOVA/Kruskal-Wallis H tests analyzed the continuous variables. To assess the predictive value of clinical parameters, univariable logistic regression analyses were conducted. Statistical significance was set at a p value of <0.05 for two-tailed tests. Survival was compared between observed pCR and observed non-pCR, and between predicted pCR and predicted non-pCR (Kaplan-Meier with log-rank tests; threshold fixed a priori). Differences in survival outcomes were analyzed using the log-rank test. HRs and 95% CIs were estimated using the Cox proportional hazards model.

Survival analysis was conducted using R software V.4.4.2 with the "survival" package V.3.8.3, and results were visualized using the "survminer" package V.0.5.0. All machine learning models were constructed using Python V.3.13.1 with the "scikit-learn" V.1.6.1 and "xgboost" V.3.0.0 packages. Details of main packages can be found in the online supplemental table 3.

RESULTS

Patient characteristics

Patient characteristics stratified by observed response status are summarized in table 1. Overall, 77/335 patients (22.99%) achieved pCR and 258/335 (77.01%) did not. No statistically significant differences were observed in most clinicopathological characteristics, except for smoking status, NCIT cycle and s-LN number (p<0.05).

Performance of the unimodal model

After feature selection, 14 radiomics and 11 pathomics features constituted the final unimodal signatures. The definitions of each feature are shown in online supplemental tables 4 and 5. Among the seven unimodal machine learning models, the XGBoost yielded the most consistent discrimination in the Training-set, Test-set-1 and Test-set-2 (figure 2A,B and online supplemental table 6). Therefore, unimodal radiomics and pathomics

Table 1 Patients' clinical characteristics across all data sets

	Overall	Training-set	Test-set-1	Test-set-2	
Characteristics	**(N=335)**	**(n=181)**	**(n=115)**	**(n=39)**	**P value**
Sex					0.189
Female	23 (6.87)	13 (7.18)	5 (4.35)	5 (12.82)	
Male	312 (93.13)	168 (92.82)	110 (95.65)	34 (87.18)	
Age	64 (44–82)	64 (44–82)	65 (46–77)	62 (48–76)	0.825
Smoking status					0.286
Never	115 (34.33)	57 (31.49)	46 (40.00)	12 (30.77)	
Current or former	220 (65.67)	124 (68.51)	69 (60.00)	27 (69.23)	
Drinking status					0.047*
Never	96 (28.66)	44 (24.31)	35 (30.43)	17 (43.59)	
Current or former	239 (71.34)	137 (75.69)	80 (69.57)	22 (56.41)	
ECOG performance status					0.269
0	146 (43.58)	81 (44.75)	44 (38.26)	21 (53.85)	
1	184 (54.93)	96 (53.04)	70 (60.87)	18 (46.15)	
2	5 (1.49)	4 (2.21)	1 (0.87)	0 (0)	
Tumor location					0.251
Upper	46 (13.73)	26 (14.36)	17 (14.78)	3 (7.69)	
Middle	169 (50.45)	99 (54.70)	51 (44.35)	19 (48.72)	
Lower	120 (35.82)	56 (30.94)	47 (40.87)	17 (43.59)	
cT					0.127
1	3 (0.90)	2 (1.10)	1 (0.87)	0 (0)	
2	50 (14.92)	29 (16.02)	19 (16.52)	2 (5.13)	
3	267 (79.70)	145 (80.11)	90 (78.26)	32 (82.05)	
4a	15 (4.48)	5 (2.77)	5 (4.35)	5 (12.82)	
cN					0.707
0	47 (14.03)	24 (13.26)	16 (13.92)	7 (17.95)	
1	170 (50.75)	95 (52.49)	60 (52.17)	15 (38.46)	
2	105 (31.34)	56 (30.94)	35 (30.43)	14 (35.90)	
3	13 (3.88)	6 (3.31)	4 (3.48)	3 (7.69)	
cTNM stage (AJCC Eighth)					0.475
I	3 (0.90)	2 (1.10)	1 (0.87)	0 (0)	
II	74 (22.09)	40 (22.10)	26 (22.61)	8 (20.52)	
III	228 (68.06)	127 (70.16)	78 (67.82)	23 (58.97)	
IVA	30 (8.95)	12 (6.64)	10 (8.70)	8 (20.51)	
Immunotherapy regimen					0.199
PD-1	298 (88.96)	159 (87.85)	101 (87.83)	38 (97.44)	
PD-L1	37 (11.04)	22 (12.15)	14 (12.17)	1 (2.56)	
NCIT cycle					<0.001*
≤2	271 (80.90)	153 (84.53)	96 (83.48)	22 (56.41)	
>2	64 (19.10)	28 (15.47)	19 (16.52)	17 (43.59)	
R0 resection					0.556
No	21 (6.27)	13 (7.18)	7 (6.09)	1 (2.56)	
Yes	314 (93.73)	168 (92.82)	108 (93.91)	38 (97.44)	
Surgical approach					0.103
Minimally	310 (92.54)	171 (94.48)	106 (92.17)	33 (84.62)	
Open	25 (7.46)	10 (5.52)	9 (7.83)	6 (15.38)	

Continued

Table 1 Continued

Characteristics	Overall (N=335)	Training-set (n=181)	Test-set-1 (n=115)	Test-set-2 (n=39)	P value
Lymphadenectomy extent					0.243
Two-field	36 (10.75)	16 (8.84)	13 (11.30)	32 (82.05)	
Three-field	299 (89.25)	165 (91.16)	102 (88.70)	7 (17.95)	
Tumor pCR					0.144
No	258 (77.01)	138 (76.24)	94 (81.74)	26 (66.67)	
Yes	77 (22.99)	43 (23.76)	21 (18.26)	13 (33.33)	
ypT stage					0.052
0	77 (22.99)	43 (23.76)	21 (18.26)	13 (33.33)	
1	70 (20.90)	33 (18.23)	24 (20.87)	13 (33.33)	
2	62 (18.51)	32 (17.68)	23 (20.00)	7 (17.96)	
3	126 (37.60)	73 (40.33)	47 (40.87)	6 (15.38)	
ypN stage					0.465
0	189 (56.42)	104 (57.46)	67 (58.26)	18 (46.15)	
1	91 (27.16)	44 (24.31)	32 (27.83)	15 (38.46)	
2	41 (12.24)	26 (14.36)	10 (8.70)	5 (12.83)	
3	14 (4.18)	7 (3.87)	6 (5.21)	1 (2.56)	
ypTNM stage (AJCC Eighth)					0.550
I	151 (45.07)	83 (45.86)	51 (44.35)	16 (41.03)	
II	50 (14.93)	27 (14.92)	20 (17.39)	3 (7.69)	
III	134 (40.00)	71 (39.22)	44 (38.26)	20 (51.28)	
s-LN number (median)	23 (5–78)	24.0 (6-63)	22 (5–78)	23 (8–61)	0.042*
Survival time (median)	692 (96–1772)	727 (100–1661)	716 (96–1172)	381 (136–1240)	0.137

Data are n (%), unless otherwise stated.
P value was calculated comparing the Training-set, Test-set-1 and Test-set-2.
s-LN number, defined as the number of lymph nodes removed from surgery.
*P value below 0.05 was considered statistically significant.
AJCC, American Joint Committee on Cancer; cN, clinical node stage; cT, clinical tumor stage; cTNM, Clinical Tumor-Node-Metastasis; ECOG, Eastern Cooperative Oncology Group; NCIT, neoadjuvant chemoimmunotherapy; pCR, pathologic complete response; PD-1, programmed cell death protein 1; PD-L1, programmed cell death ligand 1; s-LN number, surgical lymph node number; ypN, neoadjuvant pathologic node stage; ypT, neoadjuvant pathologic tumor stage; ypTNM, neoadjuvant pathologic Tumor-Node-Metastasis.

models were implemented using XGBoost with their respective optimal feature sets.

The pathomics model achieved AUCs of 0.88 (95% CI 0.82 to 0.94) in the Training-set, 0.68 (95% CI 0.55 to 0.81) in the Test-set-1, and 0.67 (95% CI 0.48 to 0.86) in the Test-set-2. The radiomics model achieved AUCs of 0.90 (95% CI 0.84 to 0.95) in the Training-set, 0.74 (95% CI 0.62 to 0.85) in the Test-set-1, and 0.68 (95% CI 0.51 to 0.85) in the Test-set-2 (figure 2D–F). In the Training-set, the radiomics model showed higher accuracy, sensitivity and specificity (0.84, 0.77 and 0.86, respectively) than the pathomics model (0.81, 0.72 and 0.83, table 2 and online supplemental figure 5).

Given class imbalance, we additionally evaluated PR performance. The pathomics model yielded area under the precision–recall curves (AUPRCs) of 0.73 in the Training-set, 0.37 in the Test-set-1 and 0.50 in the Test-set-2, while the radiomics model achieved AUPRCs of 0.81, 0.45 and 0.55 in the same cohorts (figure 2G–I). The confusion matrix for both unimodal models across test sets is provided in online supplemental figure 6.

Performance of multimodal model

We developed intermediate-fusion and late-fusion multimodal models using the fusion strategies prespecified in Methods. Across cohorts, both fusion approaches performed better than the unimodal radiomics and pathomics models across evaluation metrics (table 2 and figure 2D–I). Between the two multimodal approaches, MIFM demonstrated higher sensitivity, specificity, accuracy, and F1 score compared with MLFM (table 2 and online supplemental figure 5). For MIFM, the confusion matrix (figure 2J–N) indicated strong exclusion of non-pCR cases: true negatives numbered 87 in the Test-set-1 and 23 in the Test-set-2, and specificity was the highest among all models (table 2). The Sankey diagram depicted reclassification from unimodal predictions to MIFM predictions in reference to the ground truth,

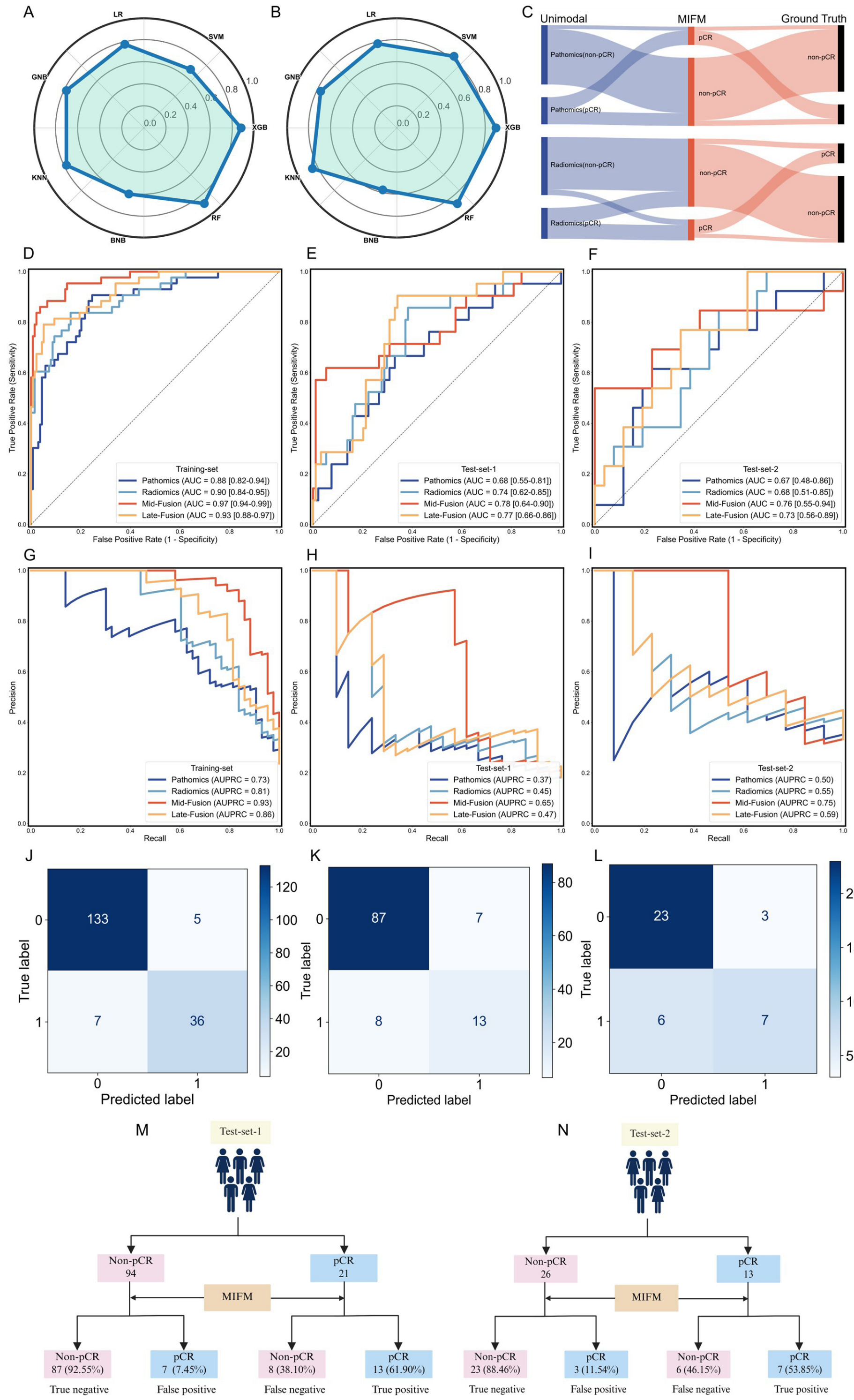

Figure 2 Performance of different models for predicting the pCR. Radar chart comparing the AUC values of seven machine learning algorithms in unimodal pathomics model (A) and unimodal radiomics model (B) in the Training-set. The Sankey diagram depicted reclassification from unimodal predictions to MIFM predictions in reference to the ground truth (C). The ROC curves for the unimodal pathomics model, unimodal radiomics model, MIFM, and MLFM are presented for both the Training-set (D), the Test-set-1 (E) and the Test-set-2 (F). The PR curves for the unimodal pathomics model, unimodal radiomics model, MIFM and MLFM for both the Training-set (G), the Test-set-1 (H) and the Test-set-2 (I). Confusion matrix of the MIFM in the Training-set (J), the Test-set-1 (K) and the Test-set-2 (L). Flow diagrams summarizing MIFM-assigned class versus observed outcome in Test-set-1 (M) and Test-set-2 (N), indicating counts of true positives/negatives and false positives/negatives. AUC, area under the curve; AUPRC, area under the precision-recall curve; BNB, Bernoulli Naïve Bayes; GNB, Gaussian Naïve Bayes; KNN, k-nearest neighbors; LR, logistic regression; MIFM, multimodal intermediate fusion model; MLFM, multimodal late fusion model; pCR, pathological complete response; PR, precision-recall; RF, random forest; ROC, receiver operating characteristic; SVM, Support Vector Machines; XGB, eXtreme Gradient Boosting.

Table 2 Performance of the models for predicting pathologic complete response

	AUC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	F1 score (95% CI)
Training-set					
Unimodal pathomics model	0.88 (0.82 to 0.94)	0.81 (0.75 to 0.86)	0.72 (0.58 to 0.85)	0.83 (0.77 to 0.89)	0.64 (0.52 to 0.74)
Unimodal radiomics model	0.90 (0.84 to 0.95)	0.84 (0.79 to 0.90)	0.77 (0.64 to 0.89)	0.86 (0.80 to 0.92)	0.69 (0.58 to 0.80)
MIFM	0.97 (0.94 to 0.99)	0.93 (0.90 to 0.97)	0.84 (0.71 to 0.95)	0.96 (0.93 to 0.99)	0.86 (0.77 to 0.93)
MLFM	0.93 (0.88 to 0.97)	0.89 (0.85 to 0.93)	0.79 (0.67 to 0.91)	0.92 (0.87 to 0.96)	0.77 (0.68 to 0.86)
Test-set-1					
Unimodal pathomics model	0.68 (0.55 to 0.81)	0.68 (0.59 to 0.77)	0.52 (0.32 to 0.75)	0.71 (0.61 to 0.80)	0.37 (0.21 to 0.52)
Unimodal radiomics model	0.74 (0.62 to 0.85)	0.69 (0.61 to 0.77)	0.62 (0.39 to 0.83)	0.70 (0.61 to 0.80)	0.42 (0.26 to 0.57)
MIFM	0.78 (0.64 to 0.90)	0.87 (0.81 to 0.92)	0.62 (0.41 to 0.83)	0.93 (0.87 to 0.98)	0.63 (0.44 to 0.79)
MLFM	0.77 (0.66 to 0.86)	0.70 (0.61 to 0.77)	0.57 (0.35 to 0.78)	0.72 (0.63 to 0.81)	0.41 (0.24 to 0.56)
Test-set-2					
Unimodal pathomics model	0.67 (0.48 to 0.86)	0.69 (0.56 to 0.82)	0.54 (0.27 to 0.82)	0.77 (0.60 to 0.92)	0.54 (0.27 to 0.74)
Unimodal radiomics model	0.68 (0.51 to 0.85)	0.59 (0.44 to 0.74)	0.54 (0.29 to 0.80)	0.62 (0.42 to 0.81)	0.47 (0.22 to 0.67)
MIFM	0.76 (0.55 to 0.94)	0.77 (0.64 to 0.90)	0.54 (0.27 to 0.83)	0.88 (0.74 to 1.00)	0.61 (0.33 to 0.82)
MLFM	0.73 (0.56 to 0.89)	0.64 (0.49 to 0.79)	0.54 (0.25 to 0.82)	0.69 (0.50 to 0.86)	0.50 (0.24 to 0.71)

AUC, area under curve; MIFM, multimodal intermediate fusion model; MLFM, multimodal late fusion model.

highlighting the net movement toward correct labels (figure 2C). The DCA was illustrated in online supplemental figure 7.

Exploratory prognostic stratification by observed and model-predicted pCR status

We examined whether observed pCR status and model-predicted pCR status (from the MIFM at its fixed operating threshold) stratified OS. In the Training-set, patients with observed pCR showed longer OS with visible separation of Kaplan-Meier curves (figure 3A), whereas this separation did not reach statistical significance in the Test-set-1 or Test-set-2 (figure 3C,E). Stratifying patients by the model's predicted status yielded a qualitatively similar pattern but did not achieve statistical significance in any cohort (figure 3B,D and F). Univariate Cox regression analysis confirmed that observed pCR, and model-predicted pCR were significantly associated with OS (p value<0.005; online supplemental figure 8).

Interpretability analyses and software prototype

SHAP summaries illustrated the contribution of individual radiomics and pathomics features to the predictions of the MIFM (figure 4A). Across all features retained in the final model, cross-modality correlations were mild to low (figure 4B and online supplemental figure 9), with values between −0.57 and 0.69, suggesting that the two modalities may capture complementary aspects of tumor biology and treatment response.

We compared the cell-type fractions by observed pCR status and, separately, by model-predicted pCR status (figure 4C). Relative to their respective non-pCR groups, both the observed pCR group and the predicted pCR group tended to show higher tumor and lymphocyte fractions and lower necrotic fractions.

Figure 4D presents two representative samples to make the decision process transparent. In one case, the unimodal radiomics and pathomics models offered discordant opinions, and the fused model reconciled these by weighting modality-specific evidence; in the other, both modalities were concordant. For each sample, feature maps displayed the spatial distribution of salient imaging and tissue descriptors (online supplemental figure 10), per-feature contribution plots show how specific values pushed the prediction toward pCR or non-pCR, and a decision-pathway view summarizes the stepwise reasoning leading to the final probability. Together, these views link feature-level signals with the model's case-level rationale.

Finally, we provide a browser-based, Graphical User Interface tool that requires no coding. Users upload the pretreatment CT, the corresponding tumor ROI mask, and the CellProfiler-derived comma-separated values (CSV) file; the tool returns the patient-level predicted probability of pCR along with basic input checks. A concise user guide and screenshots are available in the online supplemental figure 11. (This prototype is intended for research use).

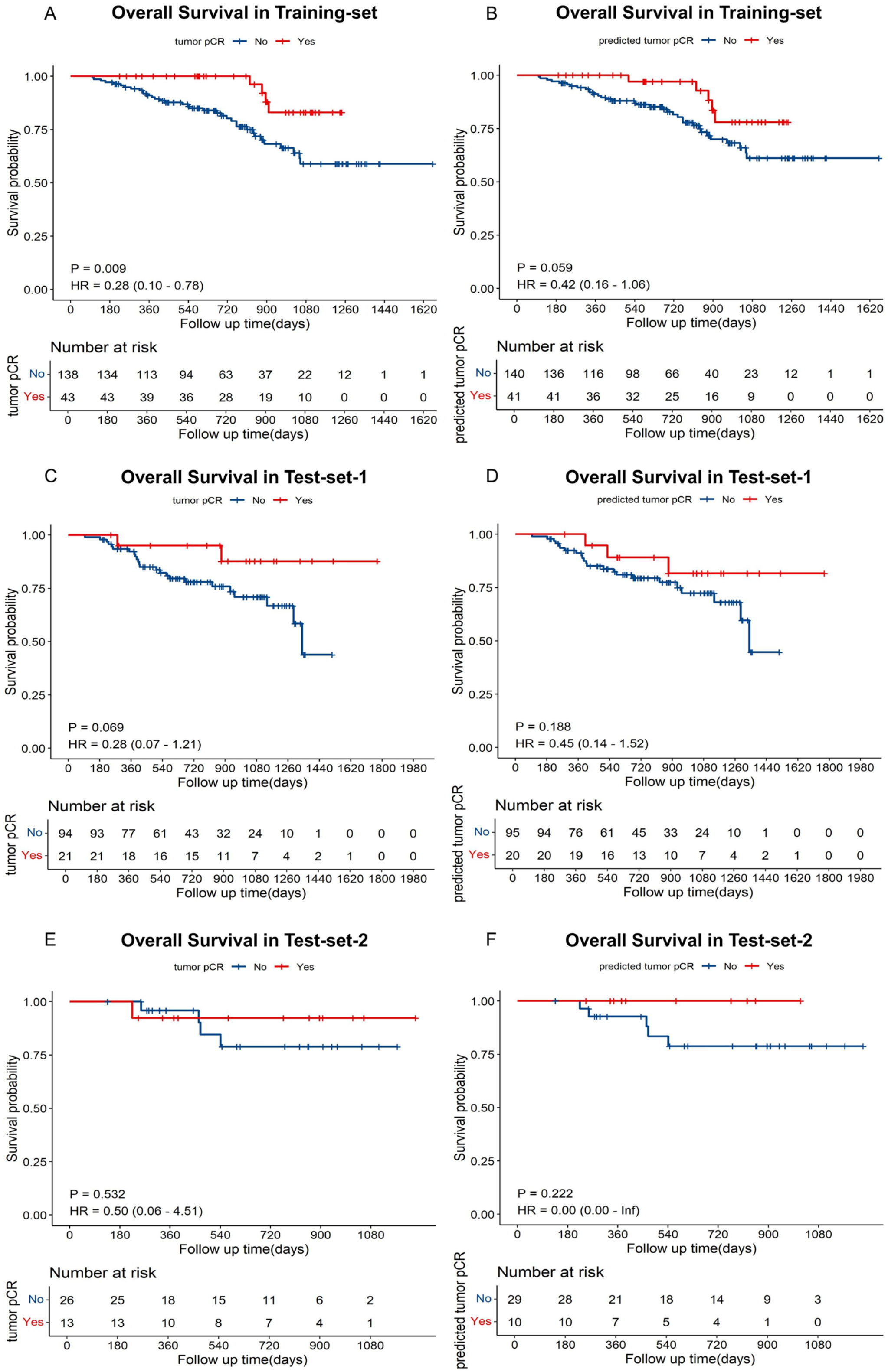

Figure 3 Prognostic stratification performance. KM curves for OS stratified by observed pCR versus non-pCR in the Training-set (A), Test-set-1 (C) and Test-set-2 (E) and stratified by MIFM predicted pCR versus non-pCR in the Training-set (B), Test-set-1 (D) and Test-set-2 (F). KM, Kaplan-Meier; MIFM, multimodal intermediate fusion model; OS, overall survival; pCR, pathologic complete response.

DISCUSSION

In this multicenter retrospective study across three academic hospitals, we developed and externally validated an interpretable multimodal framework that integrates routinely available contrast-enhanced CT radiomics with H&E-stained WSI pathomics to preoperatively predict

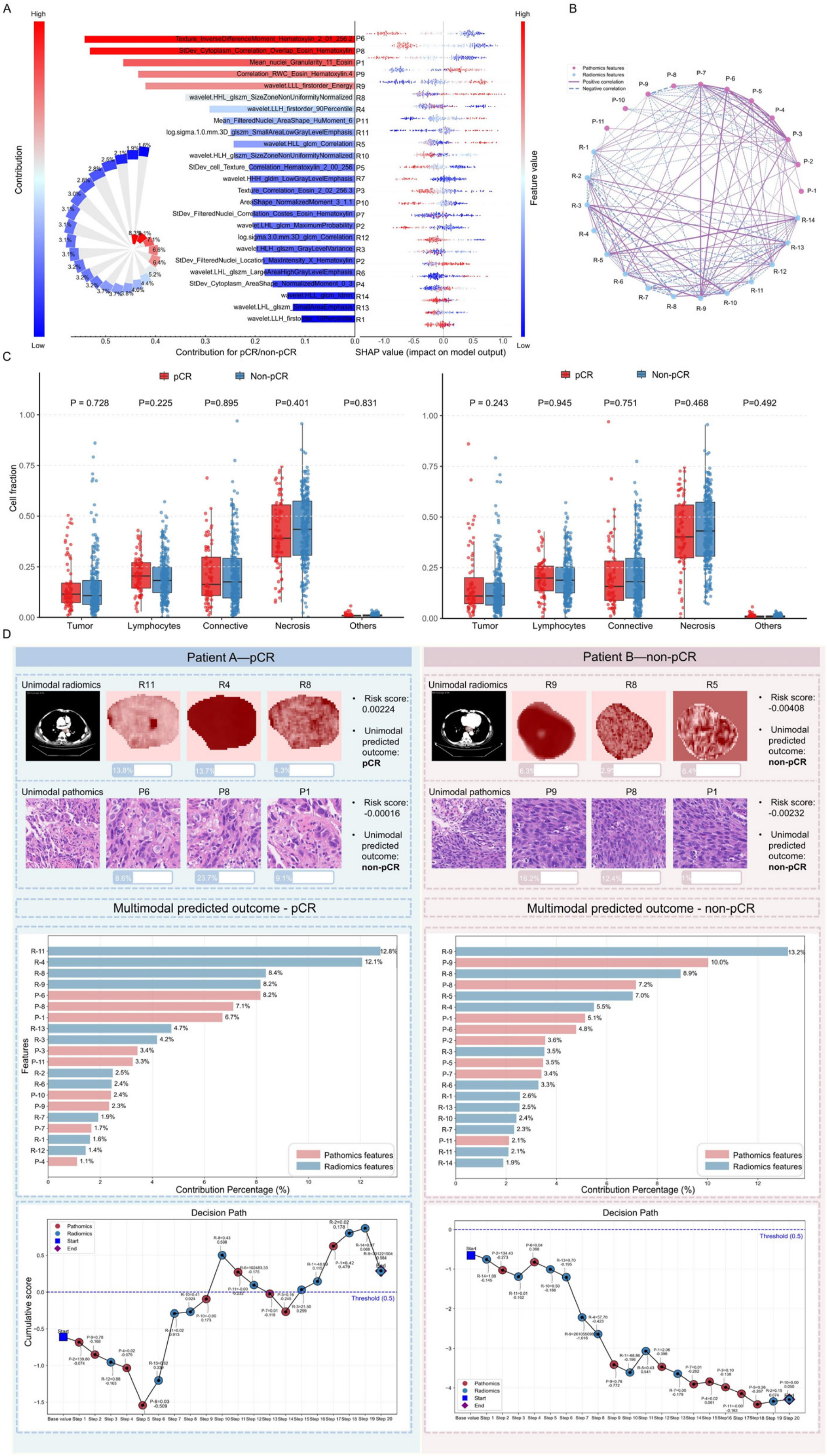

Figure 4 Interpretability analyses. (A) The SHAP summary of the MIFM: right, beeswarm plot showing per-feature effects on model output; left, bar chart ranking global importance by mean |SHAP|; inset shows proportional contributions. (B) Cross-modality Spearman correlation network for features retained in the final model (edge width reflects correlations). (C) Boxplots of WSI-derived cell-type fractions (tumor, lymphocyte, stromal, necrotic, other): left, comparison by observed status (observed pCR vs observed non-pCR); right, comparison by model-predicted status (predicted pCR vs predicted non-pCR). (D) Case vignettes illustrating discordant unimodal predictions reconciled by MIFM (Patient A, observed pCR) and concordant unimodal predictions (Patient B, observed non-pCR); for each, feature maps, top-20 contribution ranking, and the decision-pathway depict how modality-specific features accumulate to the final prediction. MIFM, multimodal intermediate fusion model; pCR, pathologic complete response; SHAP, SHapley Additive exPlanations; WSI, whole-slide image.

pCR after nCIT in ESCC. Compared with unimodal radiomics or pathomics, the intermediate-fusion model showed more robust discrimination across the development set and two validation cohorts, underscoring the value of combining complementary imaging and tissue information. The work's key strengths are its clinical practicality—it relies solely on standard-of-care data without additional testing or cost—and its interpretability-first design, where feature definitions are mathematically or morphologically explicit and case-level/cohort-level explanations (eg, SHAP summaries, feature maps) together with decision-pathway views render model reasoning transparent. Finally, we provide a browser-based Graphical User Interface tool that requires no coding and returns a patient-level pCR probability, facilitating exploratory use in multidisciplinary settings. Together, these elements highlight a feasible and transparent pathway toward translating multimodal AI into decision support for nCIT in ESCC.

The chemoradiotherapy for esophageal cancer followed by surgery study (CROSS) trial established the superiority of nCRT over surgery alone for locally advanced ESCC.[31] Nonetheless, distant metastasis remains the dominant mode of failure after nCRT—far exceeding local recurrence (22.0% vs 5.9%)—underscoring the need for enhanced systemic therapies to improve outcomes. Recently, it was shown that intensive chemotherapy improved the OS and local control over nCRT.[32 33] These improved regimens could lead to the omission of esophagectomy in patients achieving a pCR after induction therapy.[34] However non-invasive ways to determine pCR are not available. In the era of immunotherapy, multiple studies have reported that the combination of immunotherapy and chemotherapy has achieved favorable outcomes as the first-line treatment for advanced esophageal cancer,[35 36] suggesting translational potential in the neoadjuvant setting. Head-to-head comparisons of nCIT versus nCRT in locally advanced ESCC are still accruing. In a prospective multicenter study across eight high-volume centers, Guo *et al* reported superior 2-year OS (81.3% vs 71.3%) and DFS (73.9% vs 63.4%) with nCIT compared with nCRT, while pCR rates were similar (22.9% vs 25.9%) and major pathologic response favored nCRT (61.5% vs 71.8%).[6] Taken together, although the optimal neoadjuvant strategy is not yet settled, current evidence indicates substantial promise for chemoimmunotherapy in this population.

This evolving landscape motivates the present work. In routine care, pCR can only be histologically confirmed postoperatively. For patients likely to achieve pCR after neoadjuvant therapy, a watch-and-wait strategy may avert unnecessary esophagectomy, preserve organ function, and improve quality of life.[37–39] Conversely, for patients unlikely to achieve pCR, proceeding to timely esophagectomy to eradicate residual disease remains the standard curative pathway. Consequently, there is a clear clinical need for an accurate, preoperative, non-invasive predictor of pCR to guide individualized decision-making between surveillance and prompt surgery. Our multimodal, interpretability-constrained framework directly addresses this gap, aiming to inform neoadjuvant pathways in ESCC.

Integration of multimodal data has emerged as a promising approach for predicting treatment response across various cancers. For example, Mao *et al* combined pretreatment MRI, WSIs, and clinical risk factors to predict pCR following neoadjuvant chemotherapy in breast cancer,[40] although with deep learning-derived features whose semantics were less explicit. For ESCC, Qi *et al* reported that incorporating CT images and WSIs could predict pCR after nCIT, supporting the utility of multimodal fusion[41]; however, pathomics features were extracted by deep learning models, and the paired CT-WSI cohort was relatively limited (n=89). Against this backdrop, our study provides, to our knowledge, one of the largest multimodal evaluations of pCR prediction after nCIT in ESCC, with paired CT and biopsy WSIs across three centers and external validation. By employing mathematically defined radiomics features and explicitly defined morphologic/texture descriptors from WSIs, the framework enhances feature-level interpretability while maintaining competitive discrimination. compared with unimodal models, the MIFM achieved higher specificity across all cohorts (table 2), indicating strong exclusion of non-pCR cases. Clinically, such operating characteristics primarily reduce the risk of misclassifying non-pCR as pCR, thereby avoiding inappropriate surveillance and supporting timely surgery for those unlikely to achieve pCR. Conversely, patients predicted as pCR could be considered for cautious watch-and-wait, where confirmatory assessment and close monitoring are in place. Prospective studies are warranted to determine thresholds and workflows that safely translate these findings into practice.

To reduce risk inherent to high-dimensional features and modest cohort sizes,[42 43] we implemented a rigorous, training-only feature screening pipeline. From 1,094 radiomics features and 4,892 pathomics features, we applied two complementary selectors—LASSO and SVM-RFE—and used the intersection of their selected features for the final model construction. For multimodal models, we implemented intermediate-fusion and late-fusion techniques to integrate the 14 selected radiomics features and 11 selected pathomics features. While some prior reports favor late-fusion models for its robustness,[44] our results indicate that intermediate-fusion—which directly models complementary information across modalities—can yield superior specificity in this setting (table 2 and figure 2D–I). We hypothesize that retaining original, modality-specific feature information and explicitly leveraging cross-modality complementarity facilitates building predictors that are both discriminative and robust.

We further used SHAP to interpret our machine learning model, quantifying the global influence of each feature. For instance, radiomics feature R8 (wavelet.HHL_glszm_SizeZoneNonUniformityNormalized) captures textural heterogeneity—higher values indicate a more uneven distribution of same-intensity zones after high-frequency filtering, which could be related to mixed viable/necrotic components or perfusion variability. Radiomics feature R4 (wavelet.LLH_firstorder_90Percentile) summarizes the brightest voxel intensities in the low–low–high (LLH) sub-band. Higher values may indicate areas of rich vascular supply and active tumor proliferation, while lower values could correspond to necrotic or low-density regions. Among pathomics, P6 (Texture_InverseDifferenceMoment_Hematoxylin) quantifies the uniformity of the staining intensity of hematoxylin within the nuclear region, with higher values indicating high nuclear heterogeneity, while lower values indicate low nuclear heterogeneity. Pathomics feature P11 (Mean_FilteredNuclei_AreaShape_HuMoment) quantifies nuclear asymmetry and morphological heterogeneity, with higher values indicating greater nuclear irregularity and lower values reflecting more symmetrical nuclear morphology. Nuclear morphology reflects cellular proliferation status and abnormal development, where generally enlarged or irregular nuclei are associated with malignant potential.[45] These key features provide, to some extent, biologically plausible links between image-derived measurements and tumor phenotype, helping bridge model outputs with clinical reasoning.

We evaluated case-level interpretability using the decision-pathway views in figure 4D, making the stepwise reasoning of the tree-based learner explicit. Patient A (observed pCR) represents a discordant unimodal scenario: unimodal radiomics and pathomics models predicted pCR (risk score: 0.00224) and non-pCR (−0.00016), respectively, while MIFM correctly predicted pCR. Feature map panels display the three most influential radiomics and pathomics features, and a top-20 contribution ranking shows radiomics features predominating for this case. The decision pathway traces how successive radiomics thresholds progressively increased the cumulative score above the fixed operating threshold, while several pathomics features exerted negative contributions toward non-pCR. Patient B (observed non-pCR) illustrates a concordant scenario: both unimodal models correctly predicted non-pCR (radiomics: −0.00408; pathomics: −0.00232), and MIFM also yielded accurate predictions. Here, the contribution histogram shows a more balanced mix of radiomics and pathomics influences, and the decision pathway depicts cumulative decrements that keep the prediction below the threshold. These two examples not only enhance the transparency of the decision-making process but also suggest a complementary and synergistic relationship between macroradiological information and micropathological features.

The TME constitutes a complex network comprising diverse components including cancer cells, stromal cells, blood vessels, nerve fibers, and extracellular matrix.[46 47] This system plays a crucial role in tumor progression, prognosis, and response to immunotherapy.[48] In our cohort, we quantified cell-type fractions on H&E-stained WSIs (tumor/epithelial, lymphocyte, stromal/spindle, necrotic, other) and compared distributions by observed pCR status and, separately, by model-predicted status (figure 4C). In both comparisons, the pCR groups tended to exhibit higher tumor and lymphocyte fractions and lower necrotic fractions relative to their respective non-pCR groups—directionally consistent with prior reports linking viable tumor architecture and lymphocytic infiltration to treatment sensitivity.[49 50] Although these trends did not reach statistical significance, plausibly reflecting limited sample size, biopsy sampling variability, and potential variability in Hover-Net segmentation, they are biologically plausible and hypothesis-generating, warranting confirmation in larger, prospective datasets.

Although the results are encouraging, our study has several limitations. First, the retrospective design and limited external validation cohorts may introduce potential bias, necessitating prospective validation in larger populations. Second, established predictive biomarkers for immunotherapy efficacy—including TMB, PD-L1 expression levels, and combined positive score—were excluded from our analysis because these tests were not uniformly available and would add cost; integrating such markers could further improve performance. Third, despite predefined procedures, manual segmentation and visual QC for CT and WSIs inevitably introduce subjectivity. Developing and validating automated QC and segmentation pipelines should be prioritized. Fourth, no stain normalization or color augmentation was used in this study, and these strategies should be explored in future work. Fifth, statistically significant Kaplan-Meier separation was observed only for observed pCR in the training cohort, and the lack of statistical significance in the other cohorts may be attributable to the limited number of deaths. In addition, pCR is not the only determining prognostic factor for long-term survival. Therefore, these findings should be considered exploratory and hypothesis-generating. Finally, although our discussion of transparency in this study focuses on both the machine learning model level and the handcrafted feature level, interpretable deep learning could potentially enhance both model performance and transparency in future work. Furthermore, the biological hypotheses generated from this study remain preliminary, and genomic-level evidence will be essential to validate and substantiate these interpretations.

In summary, we developed and externally validated an interpretable multimodal machine learning framework that integrates contrast-enhanced CT radiomics with H&E-stained WSI pathomics to preoperatively predict pCR after nCIT in ESCC. Our findings demonstrate the clinical potential of this multimodal approach for guiding individualized decisions between surveillance and timely surgery. Further refinement and validation through large-scale prospective trials remain essential to establish its utility in clinical practice.

Author affiliations
[1]Zhejiang Cancer Hospital, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou, Zhejiang 310022, China

[2]Department of Radiation Oncology (Maastro), GROW Research Institute of Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands
[3]School of Public Health, Nanjing Medical University, Nanjing, Jiangsu, China
[4]Department of Radiation Oncology, Key Laboratory of Cancer Prevention and Therapy, Tianjin Medical University Cancer Institute & Hospital, National Clinical Research Center for Cancer, Tianjin's Clinical Research Center for Cancer, Tianjin, 300060, China
[5]Department of Ultrasound, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China
[6]Department of Pathology, Renmin Hospital of Wuhan University, Wuhan, Hubei, China
[7]Data Science Institute (DSI), Hasselt University, Hasselt, Belgium

Acknowledgements Figure 1 was created in BioRender. Zhang, Z (2025) https://BioRender.com/ijbkjms.

Contributors BQ: Investigation, Methodology, Visualization, Writing—original draft. ZJ: Visualization, Methodology, Writing—review and editing. HS: Investigation, Writing—review and editing. JL: Investigation, Writing—review and editing. ZW: Software, Writing—review and editing. MF: Writing—review and editing. CW: Writing—review and editing. YJia: Writing—review and editing. JY: Writing—review and editing. IB: Supervision, Writing—review and editing. AD: Supervision, Writing—review and editing. DDR: Supervision, Writing—review and editing. LW: Supervision, Writing—review and editing. WZ: Supervision, Writing—review and editing. YJi: Supervision, Writing—review and editing. ZZ: Supervision, Funding acquisition, Conceptualization, Writing—review and editing. ZZ is the guarantor.

Funding This study was funded by the National Natural Science Foundation of China (No. 82303672, 82573437) and Zhejiang Medical and Health Science and Technology Project (No. 2022KY637).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This retrospective study was approved by the Institutional Review Boards (IRB) of Zhejiang Cancer Hospital (IRB-2023-88).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Data are private institutional collections, which may be made available to other researchers upon reasonable request and subject to data sharing agreements—please contact the corresponding author.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

Wencheng Zhang https://orcid.org/0000-0003-3730-5361
Zhen Zhang https://orcid.org/0000-0001-6335-9529

REFERENCES

1 Bray F, Laversanne M, Sung H, *et al*. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J Clinicians* 2024;74:229–63.
2 Yang H, Liu H, Chen Y, *et al*. Neoadjuvant Chemoradiotherapy Followed by Surgery Versus Surgery Alone for Locally Advanced Squamous Cell Carcinoma of the Esophagus (NEOCRTEC5010): A Phase III Multicenter, Randomized, Open-Label Clinical Trial. *J Clin Oncol* 2018;36:2796–803.
3 Ajani JA, D'Amico TA, Bentrem DJ, *et al*. Esophageal and Esophagogastric Junction Cancers, Version 2.2023, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* 2023;21:393–422.
4 Liu J, Yang Y, Liu Z, *et al*. Multicenter, single-arm, phase II trial of camrelizumab and chemotherapy as neoadjuvant treatment for locally advanced esophageal squamous cell carcinoma. *J Immunother Cancer* 2022;10:e004291.
5 Yan X, Duan H, Ni Y, *et al*. Tislelizumab combined with chemotherapy as neoadjuvant therapy for surgically resectable esophageal cancer: A prospective, single-arm, phase II study (TD-NICE). *Int J Surg* 2022;103:106680.
6 Guo X, Chen C, Zhao J, *et al*. Neoadjuvant Chemoradiotherapy vs Chemoimmunotherapy for Esophageal Squamous Cell Carcinoma. *JAMA Surg* 2025;160:565–74.
7 Wang H, Jiang Z, Wang Q, *et al*. Pathological response and prognostic factors of neoadjuvant PD-1 blockade combined with chemotherapy in resectable oesophageal squamous cell carcinoma. *Eur J Cancer* 2023;186:196–210.
8 Wang H, Tang H, Fang Y, *et al*. Morbidity and Mortality of Patients Who Underwent Minimally Invasive Esophagectomy After Neoadjuvant Chemoradiotherapy vs Neoadjuvant Chemotherapy for Locally Advanced Esophageal Squamous Cell Carcinoma. *JAMA Surg* 2021;156:444.
9 Luchini C, Bibeau F, Ligtenberg MJL, *et al*. ESMO recommendations on microsatellite instability testing for immunotherapy in cancer, and its relationship with PD-1/PD-L1 expression and tumour mutational burden: a systematic review-based approach. *Ann Oncol* 2019;30:1232–43.
10 Raimondi A, Lonardi S, Murgioni S, *et al*. Tremelimumab and durvalumab as neoadjuvant or non-operative management strategy of patients with microsatellite instability-high resectable gastric or gastroesophageal junction adenocarcinoma: the INFINITY study by GONO. *Ann Oncol* 2025;36:285–96.
11 Wang C, Ju C, Du D, *et al*. CircNF1 modulates the progression and immune evasion of esophageal squamous cell carcinoma through dual regulation of PD-L1. *Cell Mol Biol Lett* 2025;30:37.
12 Yang F, Dan M, Shi J, *et al*. Efficacy and safety of PD-1 inhibitors as second-line treatment for advanced squamous esophageal cancer: a systematic review and network meta-analysis with a focus on PD-L1 expression levels. *Front Immunol* 2025;15.
13 Anagnostou V, Bardelli A, Chan TA, *et al*. The status of tumor mutational burden and immunotherapy. *Nat Cancer* 2022;3:652–6.
14 Zhou KI, Peterson B, Serritella A, *et al*. Spatial and Temporal Heterogeneity of PD-L1 Expression and Tumor Mutational Burden in Gastroesophageal Adenocarcinoma at Baseline Diagnosis and after Chemotherapy. *Clin Cancer Res* 2020;26:6453–63.
15 Niknafs N, Najjar M, Dennehy C, *et al*. Of Context, Quality, and Complexity: Fine-Combing Tumor Mutational Burden in Immunotherapy-Treated Cancers. *Clin Cancer Res* 2025;31:2850–63.
16 Boehm KM, Khosravi P, Vanguri R, *et al*. Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer* 2022;22:114–26.
17 Lin H, Hua J, Gong Z, *et al*. Multimodal radiopathological integration for prognosis and prediction of adjuvant chemotherapy benefit in resectable lung adenocarcinoma: A multicentre study. *Cancer Lett* 2025;616:217557.
18 Qi Y-J, Su G-H, You C, *et al*. Radiomics in breast cancer: Current advances and future directions. *Cell Rep Med* 2024;5:101719.
19 Li B, Qin W, Yang L, *et al*. From pixels to patient care: deep learning-enabled pathomics signature offers precise outcome predictions for immunotherapy in esophageal squamous cell cancer. *J Transl Med* 2024;22:195.
20 Zhang Z, Luo T, Yan M, *et al*. Voxel-level radiomics and deep learning for predicting pathologic complete response in esophageal squamous cell carcinoma after neoadjuvant immunotherapy and chemotherapy. *J Immunother Cancer* 2025;13:e011149.
21 Kumar N, Verma R, Chen C, *et al*. Computer-extracted features of nuclear morphology in hematoxylin and eosin images distinguish stage II and IV colon tumors. *J Pathol* 2022;257:17–28.
22 Wang Y, Pan X, Lin H, *et al*. Multi-scale pathology image texture signature is a prognostic factor for resectable lung adenocarcinoma: a multi-center, retrospective study. *J Transl Med* 2022;20:595.
23 van der Velden BHM, Kuijf HJ, Gilhuijs KGA, *et al*. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med Image Anal* 2022;79:102470.
24 Lambin P, Leijenaar RTH, Deist TM, *et al*. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017;14:749–62.
25 Shi Z, Zhang Z, Liu Z, *et al*. Methodological quality of machine learning-based quantitative imaging analysis studies in esophageal

cancer: a systematic review of clinical outcome prediction after concurrent chemoradiotherapy. *Eur J Nucl Med Mol Imaging* 2022;49:2462–81.

26 Shi C, Jordan B. Protocol for the examination of specimens from patients with carcinoma of the esophagus. College of American Pathologists Cancer Protocols, 2017:1–17.

27 Fedorov A, Beichel R, Kalpathy-Cramer J, *et al*. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging* 2012;30:1323–41.

28 van Griethuysen JJM, Fedorov A, Parmar C, *et al*. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res* 2017;77:e104–7.

29 Stirling DR, Carpenter AE, Cimini BA. CellProfiler Analyst 3.0: accessible data exploration and machine learning for image analysis. *Bioinformatics* 2021;37:3992–4.

30 Graham S, Vu QD, Raza SEA, *et al*. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med Image Anal* 2019;58:101563.

31 Eyck BM, van Lanschot JJB, Hulshof MCCM, *et al*. Ten-Year Outcome of Neoadjuvant Chemoradiotherapy Plus Surgery for Esophageal Cancer: The Randomized Controlled CROSS Trial. *JCO* 2021;39:1995–2004.

32 Hoeppner J, Brunner T, Schmoor C, *et al*. Perioperative Chemotherapy or Preoperative Chemoradiotherapy in Esophageal Cancer. *N Engl J Med* 2025;392:323–35.

33 Hoeppner J, Schmoor C, Brunner T, *et al*. Recurrence Patterns of Esophageal Adenocarcinoma in the Phase III ESOPEC Trial Comparing Perioperative Chemotherapy With Preoperative Chemoradiotherapy. *J Clin Oncol* 2025;43:3451–6.

34 van der Wilk BJ, Eyck BM, Wijnhoven BPL, *et al*. Neoadjuvant chemoradiotherapy followed by active surveillance versus standard surgery for oesophageal cancer (SANO trial): a multicentre, stepped-wedge, cluster-randomised, non-inferiority, phase 3 trial. *Lancet Oncol* 2025;26:425–36.

35 Sun J-M, Shen L, Shah MA, *et al*. Pembrolizumab plus chemotherapy versus chemotherapy alone for first-line treatment of advanced oesophageal cancer (KEYNOTE-590): a randomised, placebo-controlled, phase 3 study. *Lancet* 2021;398:759–71.

36 Janjigian YY, Shitara K, Moehler M, *et al*. First-line nivolumab plus chemotherapy versus chemotherapy alone for advanced gastric, gastro-oesophageal junction, and oesophageal adenocarcinoma (CheckMate 649): a randomised, open-label, phase 3 trial. *The Lancet* 2021;398:27–40.

37 Kubo Y, Nozaki R, Igaue S, *et al*. Neoadjuvant Chemotherapy Improves Feasibility of Larynx Preservation and Prognosis in Resectable Locally Advanced Cervical Esophageal Cancer. *Ann Surg Oncol* 2024;31:5083–91.

38 Song X-Y, Lin L, Yang Y, *et al*. Radiotherapy as an organ-preserving alternative to surgery in patients with locally advanced esophageal squamous cell carcinoma achieving major pathologic response after induction immunochemotherapy. *Int J Cancer* 2025;157:1680–93.

39 Matsuda S, Kawakubo H, Irino T, *et al*. Role sharing between minimally invasive oesophagectomy and organ preservation approach for surgically resectable advanced oesophageal cancer. *Jpn J Clin Oncol* 2022;52:108–13.

40 Mao N, Dai Y, Zhou H, *et al*. A multimodal and fully automated system for prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer. *Sci Adv* 2025;11:eadr1576.

41 Qi Y, Hu Y, Lin C, *et al*. A preoperative predictive model based on multi-modal features to predict pathological complete response after neoadjuvant chemoimmunotherapy in esophageal cancer patients. *Front Immunol* 2025;16:1530279.

42 Mao Y, Hou X, Fu S, *et al*. Transcriptomic and machine learning analyses identify hub genes of metabolism and host immune response that are associated with the progression of breast capsular contracture. *Genes Dis* 2024;11:101087.

43 Li Y, Yu J, Li R, *et al*. New insights into the role of mitochondrial metabolic dysregulation and immune infiltration in septic cardiomyopathy by integrated bioinformatics analysis and experimental validation. *Cell Mol Biol Lett* 2024;29:21.

44 Captier N, Lerousseau M, Orlhac F, *et al*. Integration of clinical, pathological, radiological, and transcriptomic data improves prediction for first-line immunotherapy outcome in metastatic non-small cell lung cancer. *Nat Commun* 2025;16:614.

45 Conner S, Guarin JR, Le TT, *et al*. Cell morphology best predicts tumorigenicity and metastasis *in vivo* across multiple TNBC cell lines of different metastatic potential. *bioRxiv* 2023:2023.06.14.544969.

46 Zheng S, Wang W, Shen L, *et al*. Tumor battlefield within inflamed, excluded or desert immune phenotypes: the mechanisms and strategies. *Exp Hematol Oncol* 2024;13:80.

47 Khosravi G-R, Mostafavi S, Bastan S, *et al*. Immunologic tumor microenvironment modulators for turning cold tumors hot. *Cancer Commun (Lond)* 2024;44:521–53.

48 Jin MZ, Jin WL. The updated landscape of tumor microenvironment and drug repurposing. *Sig Transduct Target Ther* 2020;5:166.

49 Hwang HW, Jung H, Hyeon J, *et al*. A nomogram to predict pathologic complete response (pCR) and the value of tumor-infiltrating lymphocytes (TILs) for prediction of response to neoadjuvant chemotherapy (NAC) in breast cancer patients. *Breast Cancer Res Treat* 2019;173:255–66.

50 Sang S, Sun Z, Zheng W, *et al*. TME-guided deep learning predicts chemotherapy and immunotherapy response in gastric cancer with attention-enhanced residual Swin Transformer. *Cell Rep Med* 2025;6:102242.