

Habitat Radiomics-Based Stacking Ensemble Model for Predicting Disease-Free Survival in Early-Stage Cervical Cancer

Hong Yang

Wuzhong People's Hospital Affiliated to Ningxia Medical University

Haiping Tian

General Hospital of Ningxia Medical University

Wei Yang

yangwei_0521@163.com

General Hospital of Ningxia Medical University

Research Article

Keywords: Uterine cervical neoplasms, Radiomics, Magnetic resonance imaging, Disease-free survival, Machine learning

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11324805/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations:

No competing interests reported.

Tables are available in the Supplementary Files section

Habitat Radiomics-Based Stacking Ensemble Model for Predicting Disease-Free Survival in Early-Stage Cervical Cancer

Abstract

Purpose: Despite undergoing curative surgery, about 30% of patients with early-stage cervical cancer (ECC) face the risk of recurrence. However, current prognostic tools are not very effective, with concordance index (C-index) typically ranging from 0.60 to 0.70. This study aimed to develop and validate a stacking survival ensemble model that incorporates intratumoral and peritumoral habitat radiomic features derived from routine preoperative multiparametric MRI to predict postoperative disease-free survival (DFS) in ECC.

Methods: This retrospective-prospective study enrolled 583 ECC patients (training: $n = 293$, internal validation: $n = 165$, external validation: $n = 125$). Habitat radiomic features were extracted from intratumoral and 3-mm peritumoral regions on preoperative T2WI, DWI, and CE-T1WI. A two-tier stacking survival ensemble framework was constructed with five base survival models (Cox regression, Random Survival Forest, Elastic Net-Cox, DeepSurv, and Support Vector Regression) and a Cox regression Meta-learner. Performance was assessed by C-index, time-dependent AUC, calibration curves, and decision curve analysis. The incremental value of the model-derived lymph node metastasis probability (pLNM) from our prior classification model was also evaluated.

Results: The Meta-model achieved C-indices of 0.878, 0.868, and 0.848 in the training, internal validation, and external validation cohorts, outperforming most base models ($P < 0.05$). Externally, it yielded time-dependent AUCs of 0.887, 0.875, and 0.860 at 12, 24, and 36 months. Kaplan-Meier analysis showed that the Meta-model effectively stratified into high- and low-risk groups with distinct DFS outcomes (log-rank $P < 0.001$ for all cohorts). Adding pLNM further improved C-index from 0.878 to 0.913 in the training cohort ($P < 0.001$).

Conclusion: The stacking survival ensemble model enables accurate postoperative risk stratification for ECC patients using routine preoperative MRI without additional imaging, potentially guiding individualized adjuvant therapy and surveillance intensity.

Keywords: Uterine cervical neoplasms, Radiomics, Magnetic resonance imaging, Disease-free survival, Machine learning

1. Introduction

Despite undergoing curative resection and systematic lymphadenectomy, around 30% of patients with early-stage cervical cancer (ECC) experience recurrence within five years post-surgery [1,2]. This notable recurrence rate highlights a significant clinical challenge: existing prognostic tools are inadequate for accurately identifying patients at the highest risk of recurrence. While the Sedlis criteria, FIGO staging, and clinicopathological nomograms are commonly utilized, they exhibit limited discriminative power, with C-indices generally falling between 0.60 and 0.70 [3,4].

As a result, clinicians often grapple with uncertainty when determining which patients would benefit from adjuvant therapy or enhanced surveillance, leading to potential overtreatment in low-risk individuals and undertreatment in those harboring high-risk characteristics that are not readily apparent [5]. This prognostic challenge is particularly pronounced in patients with pathologically node-negative disease, who are traditionally classified as low-risk but still represent a significant proportion of recurrence cases [6].

Conventional pathological evaluation provides only a static assessment of tumor characteristics (e.g., nodal status and stromal invasion depth), failing to capture the dynamic spatial biological heterogeneity that drives tumor evolution and recurrence [7].

MRI radiomics can capture this spatial variation, and several recent reports have shown that such imaging features correlate with outcome in cervical cancer [8,9]. Conventional whole-tumor radiomics collapses the spatial heterogeneity of a lesion into a single summary profile. That strategy works for relatively uniform tumors, but cervical cancers are not uniform [10]. Habitat imaging offers an alternative: it segments the tumor into subregions, each characterized by a distinct imaging phenotype. These subregions carry biological meaning—one area may reflect active proliferation and high perfusion, another hypoxia and necrosis, another dense cellular packing [11,12]. Such spatial variation is not incidental; it shapes how tumors evolve, how they respond to treatment, and whether they ultimately recur [13]. In addition, the peritumoral space adds another layer of information. It captures stromal

remodeling, inflammatory cell infiltration, and lymphovascular invasion—all of which influence whether a tumor extends beyond its original boundary [14].

The features we extracted are high-dimensional and likely interact in non-linear ways. We therefore chose a modeling approach that explicitly accommodates such complexity. Stacking is one such approach: it trains multiple base learners on different feature subsets or modalities, then combines their predictions via a meta-learner [15]. This strategy works well in radiomics because different algorithms capture different aspects of the data—for example, tree-based methods handle interactions, regularized regression manages collinearity, and non-linear kernels model heterogeneity [16,17]. In cervical cancer, stacking has been used almost entirely for classification (e.g., predicting lymph node metastasis). Its use for survival analysis—which must handle censored data and time-to-event outcomes—remains largely unexplored.

The stacking framework itself is not new; it has been applied to survival prediction in hepatocellular carcinoma and head and neck cancer, typically by combining Cox-type models with machine learning survival algorithms [18, 19]. What these studies share is a conventional feature strategy—whole-tumor radiomics or single-region summaries—and a single-task objective: predicting survival from imaging alone. Neither approach explicitly models the spatial heterogeneity within the tumor, nor asks whether a probability score trained for one clinical task (nodal metastasis) might carry prognostic information for another (recurrence). These are the two gaps we set out to address in the present study.

That gap aside, there is another reason to consider this approach. In our prior work, we used the same habitat radiomics framework to build a stacking model that predicts preoperative LNM [20]. That model outputs a per-patient probability—pLNM—that summarizes the imaging evidence of nodal involvement. Intriguingly, we noticed that the radiomic features driving that classification model overlapped substantially with those we found associated with DFS in preliminary analyses. This overlap is not coincidental: both LNM and recurrence reflect the same underlying biological axis of tumor aggressiveness, though with distinct components. This observation led us to hypothesize that pLNM might carry independent prognostic information beyond nodal status alone. We therefore designed the present study to test this hypothesis by extending the stacking framework from classification to survival analysis, and to evaluate whether pLNM provides incremental prognostic value beyond conventional clinicopathological factors.

We extended the stacking framework from classification to survival analysis to: (1) develop and validate a stacking survival ensemble model for postoperative DFS prediction; and (2) evaluate whether pLNM provides incremental prognostic value beyond conventional clinicopathological factors.

2. Materials and Methods

2.1. Study Design and Patient Cohorts

While this study utilized the same cohort and imaging data as our prior LNM prediction study [20], the present work represents a distinct prognostic investigation. We completely re-performed all feature selection and model development using a survival-guided pipeline (DFS endpoint), and propose a novel two-tier stacking survival framework to address the unmet clinical need of postoperative risk stratification. The protocol, including both retrospective and prospective phases, was approved by our institutional review board (Approval No. KYLL-2018-062). From January 2020 to December 2024, 458 eligible patients were retrospectively enrolled and divided into a training cohort ($n = 293$) and an internal validation cohort ($n = 165$). From January 2025 to December 2025, 125 consecutive patients were prospectively enrolled as an external validation cohort with written informed consent. Inclusion criteria were: (a) complete multiparametric MRI within two weeks before surgery; (b) radical hysterectomy with pelvic lymphadenectomy; and (c) histopathologically confirmed cervical cancer with documented nodal status. Exclusion criteria were: (a) prior treatment; (b) poor-quality or incomplete MRI; (c) concurrent malignancies; (d) incomplete follow-up data; and (e) unconfirmed nodal status. The flowchart in Figure 1 displays the details of the patient enrollment.

For the prognostic analysis, we excluded patients without an event who had less than 36 months of follow-up (retrospective cohorts) or less than 24 months (prospective cohort). The primary endpoint was disease-free survival (DFS), defined as time from surgery to local or regional recurrence, distant metastasis, or death from any cause. Patients without these events were censored at their last follow-up.

2.2. MRI Acquisition and Image Preprocessing

The workflow of our research is shown in Figure 2. The MRI acquisition protocols, image preprocessing, evaluation criteria for clinicopathological and MRI features were identical to that described in our previous work [20]. Preprocessing followed the Image Biomarker Standardization Initiative (IBSI) guidelines: all images were resampled to 1 mm³ isotropic voxels with B-spline interpolation, and intensity values were standardized via histogram matching [21,22].

2.3. ROI Segmentation and Habitat Generation

Tumor segmentation and habitat generation followed the same protocol as our prior study, with the same 3D Slicer-based manual slice-by-slice delineation and 3-mm peritumoral expansion [23]. Reproducibility was confirmed via ICC (threshold ≥ 0.75) on 50 randomly selected cases. Intratumoral habitats were generated using K-means clustering ($k=3$, determined by Calinski-Harabasz index [24]) on the training cohort only; cluster centroids were then applied to validation cohorts to prevent data leakage.

2.4. Radiomic Feature Extraction and Selection

Radiomic features were extracted from intratumoral habitats and peritumoral regions using PyRadiomics (v3.0.1) according to IBSI guidelines [22], covering shape, first-order, and texture features.

Although we reused the same raw feature matrix from our prior LNM study, all feature selection steps for this prognostic analysis were re-performed using DFS as

the endpoint. A four-stage pipeline was applied in the training cohort: (1) ICC-based filtering (>0.75); (2) univariate Cox regression with Benjamini-Hochberg FDR correction (adjusted $P < 0.10$); (3) Spearman correlation-based redundancy elimination ($|\rho| > 0.90$); and (4) LASSO-Cox regression with 10-fold cross-validation (1-SE rule). This process yielded prognostic signatures for Rad-score_intra, Rad-score_peri, Rad-score_intra+peri, and ITH_integrated_score. The optimal features and their corresponding coefficients are shown in Supplementary Table S1.

To visualize the relationship between features driving LNM and those driving DFS, we constructed a Venn diagram comparing the final radiomic feature sets selected in our prior LNM prediction model [20] with those selected in the present prognostic Cox model. The overlap between these two independently selected feature sets was examined to identify core imaging phenotypes that underpin both nodal metastasis and recurrence.

2.5. Stacking Survival Ensemble Model Construction

A two-tier stacking survival ensemble framework was developed.

Tier 1: Base Survival Models. Five base survival models were constructed in parallel:

(1) a Cox regression model incorporating independent clinical prognostic factors (Cox-Clinical); (2) a Random Survival Forest (RSF) model trained on intratumoral radiomic features; (3) an Elastic Net-Cox model trained on peritumoral radiomic features; (4) a DeepSurv model trained on combined intratumoral and peritumoral features; and (5) a linear kernel Support Vector Regression (SVR) model trained on

ITH_integrated_score features. Each base model was optimized using 5-fold cross-validation and outputted cumulative risk predictions for each sample.

Tier 2: Meta Survival Model. The risk predictions from the five base models, generated through 5-fold cross-validation, were used as input features to train a Cox regression Meta-learner. The Meta-model outputted the final Ensemble Risk Score (ERS): $ERS = \sum \gamma_m \cdot \hat{h}_m(x)$, where $\hat{h}_m(x)$ is the risk prediction from the m-th base model and γ_m is the coefficient estimated by the Cox Meta-model.

2.6. Evaluation of pLNM Incremental Value to the Meta-model

The predicted pLNM from our previously published classification model was evaluated as an additional prognostic factor. Two stacking survival ensemble models were compared: Model A (without pLNM) and Model B (with pLNM as an additional predictor in Meta model). pLNM was added as an additional predictor at the Meta-learner level, rather than being fed into each base model separately. The incremental value was quantified by changes in C-index, Net Reclassification Improvement (NRI), and Integrated Discrimination Improvement (IDI).

2.7. Model Evaluation and Statistical Analysis

Model discrimination was assessed using Harrell's C-index (primary) and time-dependent AUC at 12, 24, and 36 months. Calibration was evaluated using calibration curves with Brier scores. Clinical utility was assessed via decision curve analysis (DCA). Risk stratification was performed by dividing patients into high-risk

and low-risk groups based on the optimal ERS cutoff determined by the Youden index from the training cohort ROC analysis. Kaplan-Meier curves and log-rank tests were used to compare DFS between groups.

Model comparisons were performed using the DeLong test. NRI and IDI were calculated to quantify incremental predictive value. SHAP analysis was used for model interpretability. The optimal cutoff value for the pLNM probability in the pN0 subgroup was determined by maximizing the Youden index in the training cohort (cutoff = 0.29). All tests were two-sided, with $P < 0.05$ considered significant. Statistical analyses were performed using R (version 4.2.2) and Python (version 3.9.7).

3. Results

3.1. Patient Characteristics

Of the 713 patients initially screened, 583 met the criteria for prognostic analysis: 293 in the training cohort, 165 in internal validation, and 125 in external validation. Median follow-up for the entire cohort was 41.0 months (IQR: 25.5–50.0). Across the three cohorts, 184 patients (31.6%) had recurrence or metastasis (training: 32.4%; internal: 32.7%; external: 28.0%), with no significant between-cohort difference ($P = 0.462$). Baseline clinicopathological characteristics—including age, tumor size, FIGO stage, histological type, grade, LVSI, and adjuvant therapy—were balanced across cohorts (Table 1). pLNM values differed significantly between the training and

validation cohorts (training vs. internal: $P = 0.004$; training vs. external: $P < 0.001$), but not between the two validation cohorts ($P = 0.246$) (Table 1).

3.2. Comparison of Clinicopathological and Radiomic Features Between Recurrence/Metastasis and Disease-Free Groups

In the training cohort, patients who recurred differed from those who remained disease-free across multiple clinical and radiomic dimensions (Supplementary Table S2). The recurrence group had shorter follow-up (median: 18.0 vs. 50.0 months), were younger (50.0 vs. 56.5 years), and had higher SCC-Ag levels (6.90 vs. 1.90 ng/mL). They also had higher rates of lymph node metastasis (42.1% vs. 22.2%), deep stromal invasion (54.7% vs. 40.4%), parametrial infiltration (49.5% vs. 32.8%), and LVI (66.3% vs. 49.5%) (all $P < 0.05$). Fewer patients in the recurrence group received adjuvant therapy (none: 61.1% vs. 26.8%, $P < 0.001$), reflecting standard practice where higher-risk patients are more aggressively treated.

All four radiomic scores were higher in the recurrence group. The largest difference was in ITH_integrated_score (3.45 vs. 1.20, $P < 0.001$), followed by Rad-score_intra+peri (1.88 vs. 0.95), Rad-score_peri (0.82 vs. 0.45), and Rad-score_intra (1.05 vs. 0.52) (all $P < 0.001$). pLNM followed a similar pattern (0.45 vs. 0.44, $P < 0.001$) (Supplementary Fig. S1 and S2).

3.3. Independent Prognostic Factors for DFS

Univariable Cox regression identified 13 variables associated with DFS. Among them, pLNM showed the strongest unadjusted association (HR = 4.53, 95% CI: 2.44–8.39, P

< 0.001), followed by ITH_integrated_score (HR = 3.86, 95% CI: 2.52–5.91, $P < 0.001$). Traditional clinicopathological factors—lymph node metastasis, LVSI, deep stromal invasion, parametrial infiltration, and SCC-Ag—were associated with worse DFS (all $P < 0.05$). Radiotherapy and chemotherapy, in contrast, were associated with better outcomes (HR < 1, $P < 0.001$).

After multivariable adjustment, 11 factors retained significance (Table 2; Supplementary Fig. S3). The strongest independent predictor was again ITH_integrated_score (HR = 3.25, 95% CI: 2.01–5.26, $P < 0.001$), followed by Rad-score_intra+peri (HR = 2.56, 95% CI: 1.66–3.94, $P < 0.001$). pLNM remained independently associated with DFS (HR = 2.86, 95% CI: 1.49–5.48, $P = 0.002$), as did pathological lymph node status (HR = 1.73, 95% CI: 1.13–2.67, $P = 0.012$), LVSI, deep stromal invasion, parametrial infiltration, age, and adjuvant therapy (chemotherapy: HR = 0.29; radiotherapy: HR = 0.39). SCC-Ag lost significance after adjustment ($P = 0.153$), suggesting its effect was mediated by other factors in the model.

We also compared the radiomic features selected in this prognostic model with those from our prior LNM prediction model [20]. The two models shared 8 of 10 and 9 core features, respectively; the overlap and unique features are shown in Supplementary Fig. S4. Specifically, the 8 shared features—including DCE_wavelet-HLL_glrmlm_RunEntropy, DCE_gradient_glrmlm_ShortRunLowGrayLevelEmphasis, DCE_log-sigma-3-0-mm-3D_glrmlm_RunEntropy,

DCE_log-sigma-4-0-mm-3D_glrIm_RunEntropy,
DCE_log-sigma-5-0-mm-3D_glszm_LowGrayLevelZoneEmphasis,
DWI_log-sigma-5-0-mm-3D_glcM_Correlation,
T2WI_wavelet-LLH_glrIm_ShortRunLowGrayLevelEmphasis, and
T2WI_wavelet-HHL_gldm_SmallDependenceLowGrayLevelEmphasis—represent
the core aggressive radiomic phenotype that drives both nodal dissemination and
distant recurrence. The two features unique to the LNM model
(DCE_original_shape_Elongation and DCE_original_shape_Flatness) reflect
morphological changes associated with local lymphatic spread, while the single
feature unique to the prognostic Cox model (DCE_logarithm_firstorder_Entropy)
captures intratumoral heterogeneity linked to time-dependent recurrence risk.

3.4. Performance of the Stacking Survival Ensemble Model

The Meta-model consistently outperformed all base models across the three cohorts (Table 3). C-indices for the Meta-model were 0.878 (95% CI: 0.830–0.920) in the training set, 0.868 (95% CI: 0.829–0.914) in internal validation, and 0.848 (95% CI: 0.804–0.888) in external validation. The corresponding 3-year time-dependent AUCs were 0.889, 0.877, and 0.860. Among base models, SVR—the model based on ITH_integrated_score—ranked first across all cohorts (C-indices: 0.851, 0.834, 0.823; 3-year AUCs: 0.857, 0.842, 0.830), followed by DeepSurv. The clinical Cox model performed worst (C-indices: 0.720, 0.700, 0.690), confirming that radiomic features add predictive value beyond clinical variables alone.

ROC curves (Fig. 3a-c) and radar plots (Supplementary Fig. S5) confirmed the Meta-model's superior and balanced performance. Calibration was good, with Brier scores of 0.095, 0.104, and 0.108 for the Meta-model—lower than all base models (Fig. 3d-f). DCA showed that the Meta-model provided the highest net benefit across a broad range of thresholds (Fig. 3g-i).

DeLong tests (Table 4; Supplementary Fig. S6) showed that the Meta-model significantly outperformed all base models in the training and internal validation cohorts (all $P < 0.05$). In the external validation cohort, the Meta-model outperformed Cox, Elastic Net-Cox, RSF, and DeepSurv (all $P < 0.01$), but the difference with SVR was not statistically significant ($P = 0.102$), although the C-index remained numerically higher (0.848 vs. 0.823).

3.5. Incremental Value of the Meta-Model Over Base Models

NRI and IDI analyses showed that the Meta-model improved risk classification over each base model (Supplementary Table S3; Supplementary Fig. S7). In the training cohort, NRI values ranged from 0.088 (over SVR) to 0.308 (over Cox) (all $P < 0.05$); corresponding IDI values ranged from 0.058 to 0.241. These patterns were similar in the internal and external validation cohorts, with the largest gains consistently observed over the clinical Cox model.

3.6. Time-Dependent AUC Analysis

The Meta-model achieved the highest time-dependent AUCs at 12, 24, and 36 months across all three cohorts (Fig. 4; Supplementary Table S4). In the training cohort, the

corresponding values were 0.915 at 12 months, 0.902 at 24 months, and 0.889 at 36 months. The decline from 12 to 36 months was modest—2.8% in the training cohort, and similarly 2.7% and 3.0% in the internal and external validation cohorts. The model's 12-month AUC exceeded 0.88 in both validation cohorts.

3.7. pLNM Incremental Value

Adding pLNM to the Meta-model improved C-index in all three cohorts (Table 5). In the training cohort, the C-index increased from 0.878 (95% CI: 0.830-0.920) to 0.913 (95% CI: 0.871-0.954); NRI was 0.124 (95% CI: 0.051–0.193, $P < 0.001$) and IDI was 0.058 (95% CI: 0.022–0.095, $P < 0.001$). Similar patterns were observed in the internal validation cohort (C-index: 0.868 $\rightarrow$ 0.900, $P = 0.003$; NRI: 0.105, $P = 0.002$; IDI: 0.048, $P = 0.004$) and external validation cohort (C-index: 0.848 $\rightarrow$ 0.878, $P = 0.012$; NRI: 0.092, $P = 0.008$; IDI: 0.041, $P = 0.006$). We further evaluated the clinical utility of the pLNM risk score in the pathologically lymph node-negative (pN0) subgroup ($n=425$). As shown in Supplementary Figure S8, patients with a pLNM-high risk (≥ 0.29) exhibited significantly worse disease-free survival compared to those with a pLNM-low risk (< 0.29) (Log-rank $P < 0.001$). This finding demonstrated that our preoperative pLNM model could effectively identify a subset of patients at high risk of recurrence, even in patients without pathological evidence of lymph node metastasis.

3.8. Risk Stratification and Survival Analysis

The Meta-model stratified patients into high-risk and low-risk groups based on the ERS cutoff from the Youden index. The two curves diverged early and remained separated throughout follow-up in all three cohorts (log-rank $P < 0.001$ for each; Fig. 5).

3.9. Interpretability Analysis

SHAP analysis (Fig. 6) identified the dominant predictors in the Meta-model. The most influential feature was ITH_integrated_score, followed by Rad-score_intra+peri, pLNM, treatment, and Rad-score_intra (Fig. 6A–B). Radiomic features collectively outweighed clinical variables in mean absolute SHAP impact, with ITH contributing the largest share.

Two representative cases illustrate how the model reaches individual predictions (Fig. 6C–F). For the high-risk patient, elevated ITH_integrated_score and Rad-score_intra+peri pushed the predicted risk upward. For the low-risk patient, lower values across these same features moved the prediction in the opposite direction.

4. Discussion

This study was designed to build a prognostic model for ECC, but the model output was not the only compelling result. The most striking finding was the prognostic dominance of ITH_integrated_score. Its HR (3.25) was nearly twice that of pathological LN status (1.73) and remained significant after adjusting for all clinical

and radiomic covariates. Biologically, this adds to the established evidence that tumors with greater subclonal diversity adapt more readily to therapeutic pressure and immune selection [25–27]. What sets our measure apart is that ITH_integrated_score captures the dispersion of imaging phenotypes across habitat subregions, not their average intensity. In other words, it is the spatial variation—not any single summary metric—that appears to carry the prognostic signal. The SHAP contribution of SVR, the algorithm specifically designed to model heterogeneity, supports this interpretation: linear approaches alone would likely underestimate these relationships. This interpretation is further supported by a separate observation: the LNM prediction model and the prognostic Cox model shared 8 of their 10 and 9 core radiomic features, respectively. In other words, the imaging phenotypes that drive nodal metastasis largely overlap with those that predict recurrence. The separation—two features unique to LNM, one unique to prognosis—suggests that while the two endpoints share a common biological axis, they also have distinct components. The two LNM-exclusive features (DCE_original_shape_Elongation and Flatness) likely reflect mechanical deformation of the tumor as it extends into the lymphatic drainage pathways, whereas the prognosis-exclusive feature (DCE_logarithm_firstorder_Entropy) may capture the degree of internal disorganization that predisposes to late relapse. This compartmentalized biology is clinically relevant: it implies that the same baseline imaging scan can simultaneously inform two distinct clinical decisions—whether to perform nodal dissection (LNM

model) and whether to intensify adjuvant surveillance (prognostic model)—without requiring additional imaging or invasive procedures.

Two additional observations extend this picture. First, combining intra- and peritumoral radiomic features improved prediction over either region alone, echoing a ‘dual-compartment’ model of risk [14,28,29]. Second, pLNM—a probability score we originally developed for preoperative LNM classification—remained independently prognostic after accounting for all radiomic features. The SHAP ranking places pLNM third among all predictors, behind only the two radiomic scores, and ahead of pathological LN status. This cross-domain persistence suggests that pLNM captures overall tumor aggressiveness, not just nodal status. The practical implication is that the same preoperative imaging score used to inform surgical decisions may also help guide adjuvant planning.

In clinical practice, the routine staging MRI serves an additional purpose. Beyond predicting nodal status, the pLNM score effectively identifies a high-risk subset among patients with pathologically negative nodes (pN0). Since pathological assessment offers limited guidance for this subgroup, incorporating pLNM can help clinicians tailor decisions on adjuvant treatment and follow-up intensity, avoiding overtreatment in truly low-risk patients while ensuring adequate therapy for those with unfavorable pLNM results.

From a methodological standpoint, the stacking framework allowed us to avoid a choice that many radiomic studies face: which single algorithm to use. Cox is

interpretable but linear; RSF captures interactions but can struggle with sparse predictors; Elastic Net-Cox handles collinearity but remains linear; DeepSurv models non-linearity but requires larger samples; SVR captures heterogeneity but is sensitive to kernel choices [30,31]. Rather than selecting one, we let the Meta-learner weight each model according to its predictive contribution. The resulting SHAP hierarchy—SVR (heterogeneity) > DeepSurv (intra+peri) > RSF (intra) > Elastic Net-Cox (peri) > Cox (clinical)—is instructive but also confirms what the base model performance already suggested: features derived from heterogeneity and combined intra+peri radiomics carry most of the prognostic signal. The near-negligible contribution of the clinical Cox model, despite the fact that it included lymph node status and other known prognostic factors, aligns with the limited performance of existing clinicopathological nomograms [1-3,32]. In the external validation cohort, the Meta-model outperformed all base models except SVR, where the difference did not reach statistical significance ($P = 0.102$), although the C-index remained numerically higher (0.848 vs 0.823). SVR was the strongest base model across all cohorts, likely because it was specifically trained on ITH_integrated_score—the most prognostic single feature in our study. With SVR already capturing much of the heterogeneity signal, the additional gain from stacking was modest in the external set. The Meta-model still achieved a C-index of 0.848 and a 3-year AUC of 0.860 in that cohort, higher than any single base model except SVR, where the difference was not statistically significant. Whether a simplified model using only ITH and intra+peri features would perform as well as the full stacking framework remains an open

question; our data suggest it might in some settings, but the stacking approach preserved a consistent numerical advantage across all cohorts.

The model achieved its highest AUC at 12 months across all cohorts (0.915 in training, 0.901 and 0.887 in the two validation sets), with only a modest decline thereafter ($\leq 3.0\%$ from 12 to 36 months). This pattern is worth noting because it suggests that the model is picking up something different from what dynamic, post-treatment biomarkers would capture. Those markers require therapy to be delivered before they become informative; here, the relevant signal is already present on preoperative imaging [33]. We interpret this as evidence that the model's strongest contribution is not in tracking how the patient responds to treatment, but in identifying, at baseline, who is likely to have an early recurrence [34]. This timing is clinically relevant: the first postoperative year is when adjuvant decisions are most consequential [35]. If the imaging phenotype can identify those patients upfront, there is an opportunity to intensify therapy before recurrence occurs, rather than detect it after the fact.

The ERS cutoff separated patients into high- and low-risk groups, with log-rank P values below 0.001 in all three cohorts (Fig. 5). The separation was consistent across independent samples—training, internal validation, and external validation—suggesting that it reflects genuine biology rather than overfitting. High-risk patients identified by this stratification could be candidates for intensified surveillance, while low-risk patients might safely avoid unnecessary treatment escalation. This separation is reproducible across independent cohorts—a prerequisite for any clinical application worth pursuing. In practice, a high ERS would prompt

discussion of more frequent postoperative imaging (e.g., 3-month intervals in the first two years rather than the standard 6-month schedule) and whether the patient's risk profile justifies adjuvant chemotherapy. A low ERS would support a less intensive follow-up plan. The score is derived from the preoperative MRI already acquired for staging; no additional imaging sequences or post-processing software are required beyond what is already available in most radiology departments. Implementation would require integrating the model into the reporting workflow, but the computational burden is negligible and the output—a single risk score—is interpretable to clinicians familiar with nomogram-based tools.

Several aspects of this study should be considered when interpreting the results. The validation cohort, although prospective, was from a single institution; multi-center testing would strengthen generalizability. Manual segmentation, while reproducible in our hands ($ICC > 0.75$), is not scalable for routine use—automated methods would be a logical next step. We did not incorporate molecular biomarkers, and whether adding them would further improve performance is unknown. Finally, the median follow-up of 41 months is sufficient for DFS but not for overall survival; longer observation is ongoing.

5. Conclusion

We developed and validated a stacking survival ensemble model integrating habitat radiomics for postoperative DFS prediction in ECC. The model performed consistently across cohorts (C-indices: 0.878–0.848). ITH_integrated_score was the

dominant predictor (HR = 3.25), and pLNM independently improved prediction (C-index: 0.878→0.913). This framework offers a practical, imaging-derived tool from routine preoperative MRI to support individualized adjuvant therapy and surveillance planning.

References

1. Hou W, Ma Y, Sun S, et al. Predictive factors for postoperative recurrence in early cervical cancer patients: a meta-analysis. *Front Surg.* 2025; 12:1588558. doi: 10.3389/fsurg.2025.1588558.
2. Bae BK, Cho WK, Kim BG, et al. Patterns and risk factors of recurrence in low-risk early-stage cervical adenocarcinoma treated with surgery alone: implications on risk group stratification. *Int J Gynecol Cancer.* 2022; 32(12):1524-1530. doi: 10.1136/ijgc-2022-003971.
3. Wright JD, Matsuo K, Huang Y, et al. Prognostic performance of the 2018 International Federation of Gynecology and Obstetrics cervical cancer staging guidelines. *Obstet Gynecol.* 2019; 134(1):49–57. doi: 10.1097/AOG.0000000000003311.
4. Levinson K, Beavis AL, Purdy C, et al. Beyond Sedlis-A novel histology-specific nomogram for predicting cervical cancer recurrence risk: An NRG/GOG ancillary analysis. *Gynecol Oncol.* 2021; 162(3):532-538. doi: 10.1016/j.ygyno.2021.06.017.
5. Gómez-Hidalgo NR, Acosta Ú, Rodríguez TG, et al. Adjuvant therapy in

early-stage cervical cancer after radical hysterectomy: are we overtreating our patients?

A meta-analysis. Clin Transl Oncol. 2022; 24(8):1605-1614. doi: 10.1007/s12094-022-02808-9.

6. Cibula D, Abu-Rustum NR, Fischerova D, et al. Surgical treatment of "intermediate risk" lymph node negative cervical cancer patients without adjuvant radiotherapy-A retrospective cohort study and review of the literature. Gynecol Oncol. 2018; 151(3):438-443. doi: 10.1016/j.ygyno.2018.10.018.

7. Andor N, Graham TA, Jansen M, et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nat Med. 2016; 22(1):105-113. doi: 10.1038/nm.3984.

8. Xin W, Rixin S, Linrui L, et al. Machine learning-based radiomics for predicting outcomes in cervical cancer patients undergoing concurrent chemoradiotherapy. Comput Biol Med. 2024; 177:108593. doi: 10.1016/j.compbimed.2024.108593.

9. Zhang H, Teng C, Yao Y, et al. MRI-based radiomics models for noninvasive evaluation of lymphovascular space invasion in cervical cancer: a systematic review and meta-analysis. Clin Radiol. 2024; 79(11):e1372-e1382. doi: 10.1016/j.crad.2024.07.018.

10. Chandra S, Goswami A, Mandal P. Molecular Heterogeneity of Cervical Cancer Among Different Ethnic/Racial Populations. J Racial Ethn Health Disparities. 2022; 9(6):2441-2450. doi: 10.1007/s40615-021-01180-8.

11. Chen S, Zhang Y, Su Y, et al. Habitat Radiomics Based on Dynamic Contrast-Enhanced Magnetic Resonance Imaging for Assessing Axillary Lymph Node

Burden in Clinical T1-T2 Stage Breast Cancer: A Multicenter and Interpretable Study.

J Magn Reson Imaging. 2025; 62(3):751-764. doi: 10.1002/jmri.29796.

12. Wu Y, Zhang W, Liang X, et al. Habitat radiomics analysis for progression free survival and immune-related adverse reaction prediction in non-small cell lung cancer treated by immunotherapy. J Transl Med. 2025; 23(1):393. doi: 10.1186/s12967-024-06057-y.

13. Chen X, Jiang H, Pan M, et al. Habitat radiomics predicts occult lymph node metastasis and uncovers immune microenvironment of head and neck cancer. J Transl Med. 2025; 23(1):498. doi: 10.1186/s12967-025-06474-7.

14. Shang Y, Zeng Y, Luo S, et al. Habitat Imaging With Tumoral and Peritumoral Radiomics for Prediction of Lung Adenocarcinoma Invasiveness on Preoperative Chest CT: A Multicenter Study. AJR Am J Roentgenol. 2024; 223(4):e2431675. doi: 10.2214/AJR.24.31675.

15. Wolpert DH. Stacked generalization. Neural Networks. 1992; 5(2):241-259. doi:10.1016/S0893-6080(05)80023-1.

16. Wang P, Li Y, Reddy CK. Machine learning for survival analysis: A survey. ACM Comput Surv. 2019; 51(6):110. doi:10.1145/3214306.

17. Sui CX, Chen K, Su K, et al. Predictive value of 18F-FDG PET/CT habitat radiomics combining stacking ensemble learning for prognosis in patients with hepatocellular carcinoma. Chin J Nucl Med Mol Imaging. 2025; 45(3):145-150. doi:10.3760/cma.j.cn321828-20240613-00205.

18. Liang Y, Wang Z, Peng Y, et al. Development of ensemble learning models for

prognosis of hepatocellular carcinoma patients underwent postoperative adjuvant transarterial chemoembolization. *Front Oncol.* 2023; 13:1169102. doi: 10.3389/fonc.2023.1169102.

19. Wang B, Liu J, Zhang X, et al. A stacking ensemble framework integrating radiomics and deep learning for prognostic prediction in head and neck cancer. *Radiat Oncol.* 2025; 20(1):127. doi: 10.1186/s13014-025-02695-8.

20. Yang H, Yang W, Tian H, et al. Multi-modal Stacking Machine Learning Model for Predicting Lymph Node Metastasis in Cervical Cancer. *Acad Radiol.* 2026. doi: 10.1016/j.acra.2026.07.022.

21. Nyul LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. *IEEE Trans Med Imaging.* 2000; 19(2):143-150. doi: 10.1109/42.836373.

22. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology.* 2020; 295(2):328-338. doi: 10.1148/radiol.2020191145.

23. Yang T, Zhang T, Yan W, et al. DWI-derived intratumoral, peritumoral, and habitat features for preoperative prediction of lymph node metastasis in early-stage cervical cancer using machine learning method. *Abdom Radiol (NY).* 2026. doi: 10.1007/s00261-026-05463-8.

24. Caliński T, Harabasz J. A dendrite method for cluster analysis. *Communications in Statistics.* 1974;3(1):1-27. doi:10.1080/03610927408827101.

25. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer*. 2012;12(5):323-334. doi:10.1038/nrc3261.
26. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*. 2012; 366(10):883-892. doi: 10.1056/NEJMoa1113205.
27. McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*. 2017; 168(4):613-628. doi: 10.1016/j.cell.2017.01.018.
28. Cui L, Yu T, Kan Y, et al. Multi-parametric MRI-based peritumoral radiomics on prediction of lymph-vascular space invasion in early-stage cervical cancer. *Diagn Interv Radiol*. 2022; 28(4):312–321. doi: 10.5152/dir.2022.20657.
29. Xiao ML, Fu L, Wei Y, et al. Intratumoral and peritumoral MRI radiomics nomogram for predicting parametrial invasion in patients with early-stage cervical adenocarcinoma and adenosquamous carcinoma. *Eur Radiol*. 2024; 34(2):852-862. doi: 10.1007/s00330-023-10042-2.
30. Zeng L, Ma P, Li Z, et al. Multimodal Machine Learning-Based Marker Enables Early Detection and Prognosis Prediction for Hyperuricemia. *Adv Sci (Weinh)*. 2024; 11(34):e2404047. doi: 10.1002/advs.202404047.
31. Fu Y, Wang X, Yi X, et al. Ensemble machine learning model incorporating radiomics and body composition for predicting intraoperative HDI in PPGL. *J Clin Endocrinol Metab*. 2024; 109(2):351-360. doi: 10.1210/clinem/dgad543.
32. Cibula D, Slama J, Dostálek L, et al. Tumour-free distance: a novel prognostic marker in patients with early-stage cervical cancer treated by primary surgery. *Br J*

Cancer. 2021; 124(6):1121-1129. doi: 10.1038/s41416-020-01204-w.

33. Lin W, Lan X, Huang W, et al. Preoperative prediction of cervical cancer recurrence using explainable MRI radiomics: a SHAP-Guided machine learning study. *Abdom Radiol (NY)*. 2026; 51(5):2654-2667. doi: 10.1007/s00261-025-05261-8.

34. Li J, Liu G, Luo J, et al. Cervical cancer prognosis and related risk factors for patients with cervical cancer: a long-term retrospective cohort study. *Sci Rep*. 2022; 12(1):13994. doi: 10.1038/s41598-022-17733-8.

35. Hanprasertpong J, Jiamset I, Geater A, et al. Impact of time interval between radical hysterectomy with pelvic node dissection and initial adjuvant therapy on oncological outcomes of early stage cervical cancer. *J Gynecol Oncol*. 2017; 28(4):e42. doi: 10.3802/jgo.2017.28.e42.

Figure Legends

Fig. 1. Patient enrollment flowchart for the prognostic analysis. The study comprised a retrospective development cohort (n = 458), which was further divided into a training cohort (n = 293) and an internal validation cohort (n = 165), and a prospectively collected external validation cohort (n = 125). The primary endpoint was disease-free survival (DFS), with patient outcomes stratified into DFS and Recurrence/Metastasis groups.

Fig. 2. Overview of the two-tier stacking survival ensemble framework. Tier 1 comprises five base survival models: Cox (clinical), RSF (intratumoral), Elastic Net-Cox (peritumoral), DeepSurv (combined intra/peri), and SVR (ITH). Tier 2

integrates these via a Cox regression Meta-learner. The workflow includes survival-guided feature selection (ICC, Cox+FDR, Spearman, LASSO) and the external input of pLNM from the prior LNM model.

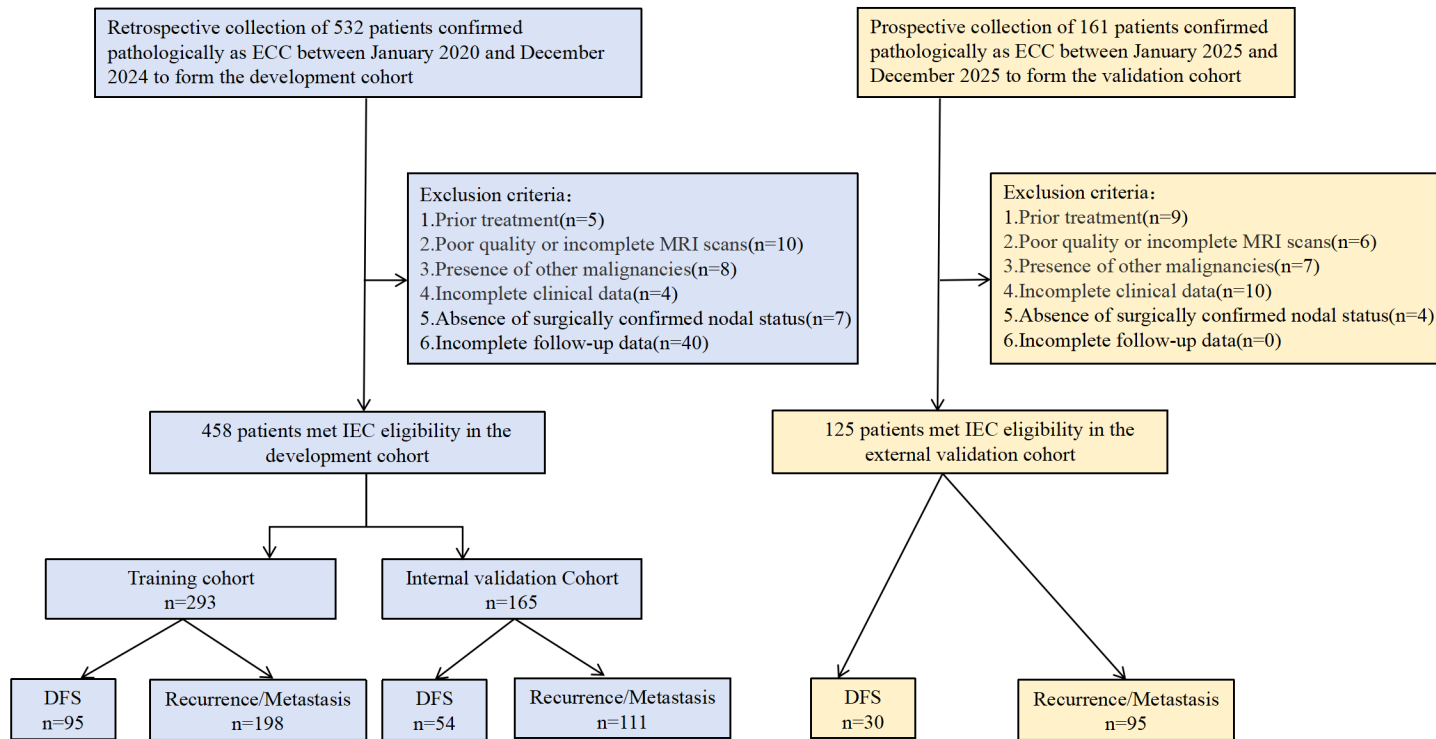
Fig. 3. Performance evaluation of the Meta-model and base models in the three cohorts. (a-c) ROC curves for the training, internal validation, and external validation cohorts. (d-f) Calibration curves for the three cohorts. (g-i) Decision curve analysis for the three cohorts.

Fig. 4. Time-dependent AUC curves for the six models at 12, 24, and 36 months across the three cohorts. The Meta-model consistently achieves the highest AUC values at all time points, with only modest declines with extended follow-up.

Fig. 5. Kaplan-Meier survival curves for high-risk and low-risk groups stratified by the ERS cutoff determined by the Youden index in the training, internal validation, and external validation cohorts. The Meta-model effectively separates patients into distinct risk groups with significantly different DFS outcomes (log-rank $P < 0.001$ for all cohorts).

Fig. 6. SHAP analysis of the Meta-model. (a) Global SHAP summary plot showing the distribution of SHAP values for each predictor across all patients. Red indicates higher feature values, blue indicates lower values. (b) Ranking of predictors by mean absolute SHAP values. ITH_integrated_score showed the highest contribution, followed by Rad-score_intra+peri, pLNM, and treatment. (c-d) Waterfall plots for

representative high-risk (c) and low-risk (d) patients, demonstrating how individual features shift the prediction from the base value to the final prediction.



DATA & INPUT

- Patients (n = 583)
- Preoperative MRI (T2WI, DWI, CE-T1WI)
- Habitat radiomic features extracted as previously described

FEATURE SELECTION

Survival-Guided Feature Selection (DFS endpoint)

- Step 1: ICC > 0.75 (reproducibility)
- Step 2: Univariate Cox + FDR ($P < 0.10$)
- Step 3: Spearman correlation ($|\rho| > 0.90$)
- Step 4: LASSO-Cox regression (10-fold CV, 1-SE rule)

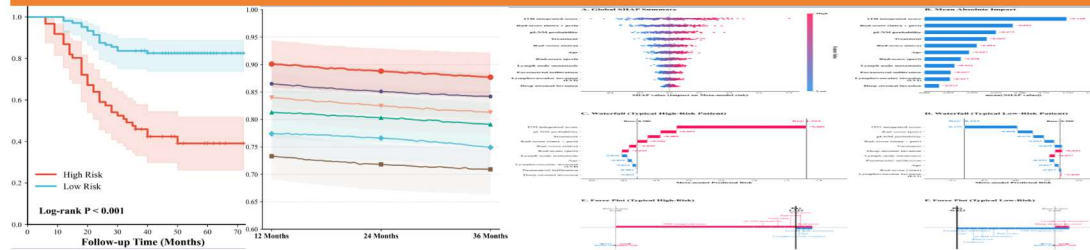
MODEL DEVELOPMENT

Tier 1: Base Models → COX RSF Elastic Net-Cox DeepSurv SVR

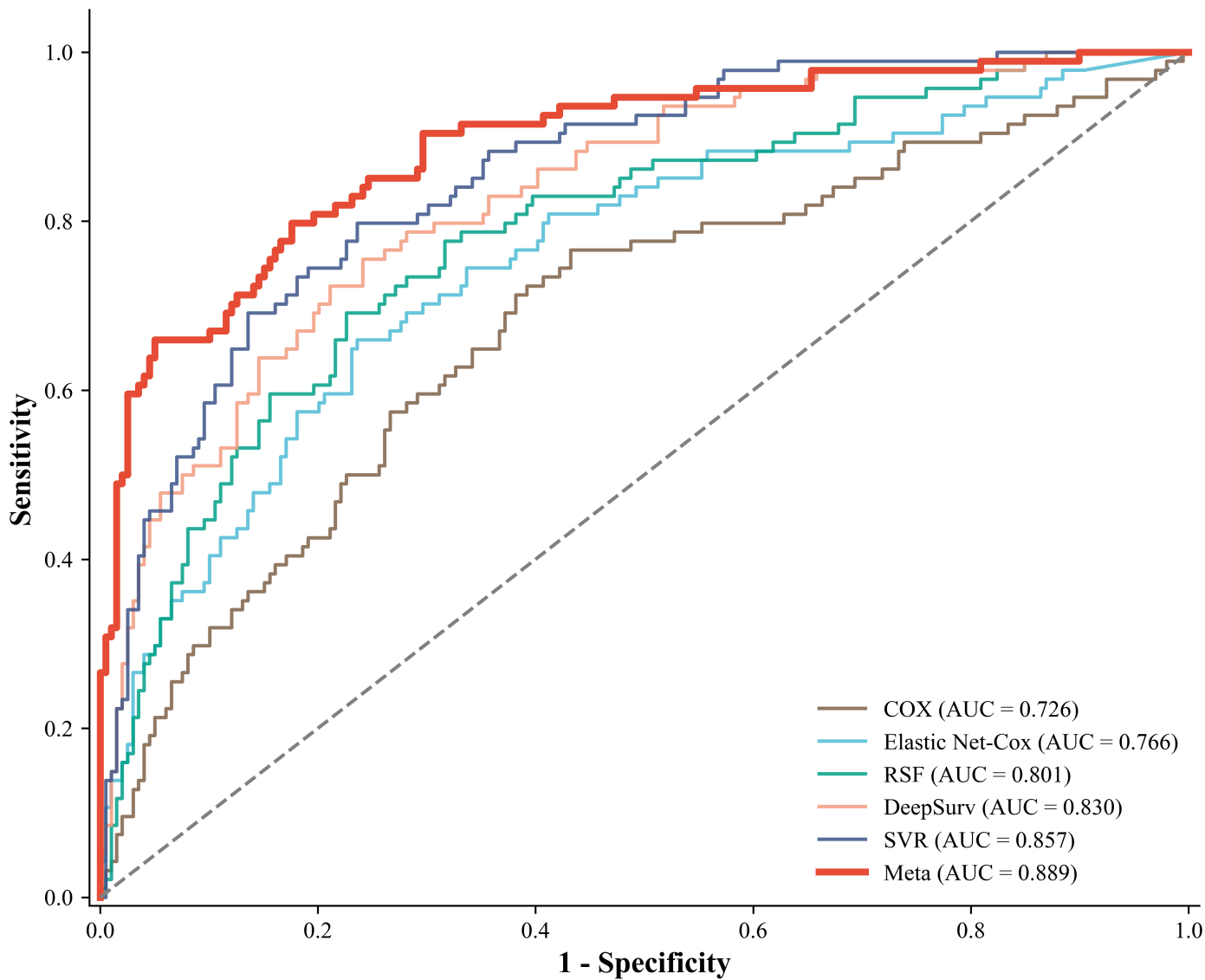
Tier 2: Meta Model → COX Regression Meta-model

External Input: pLNM (from prior LNM prediction model)

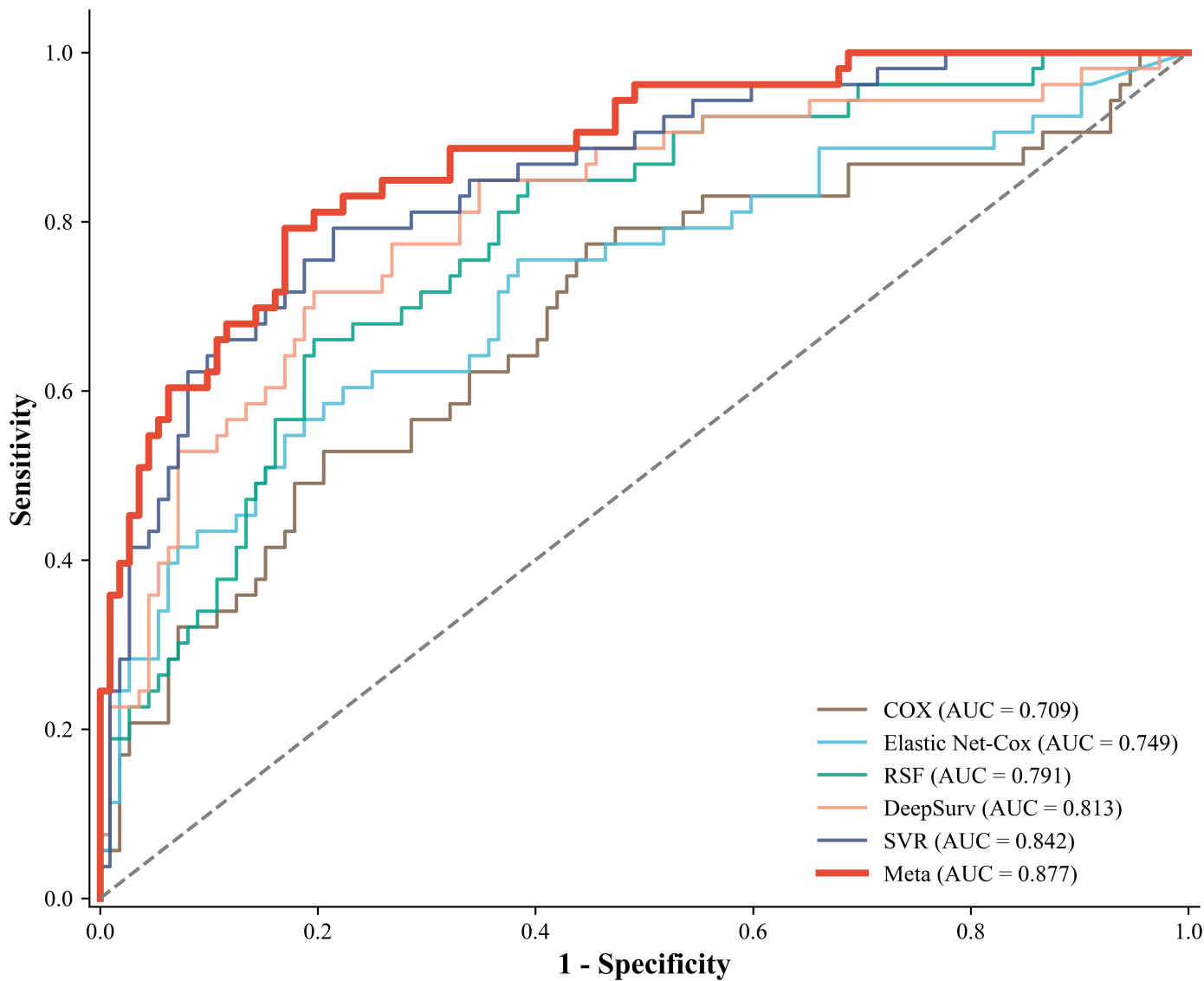
OUTPUT & EVALUATION



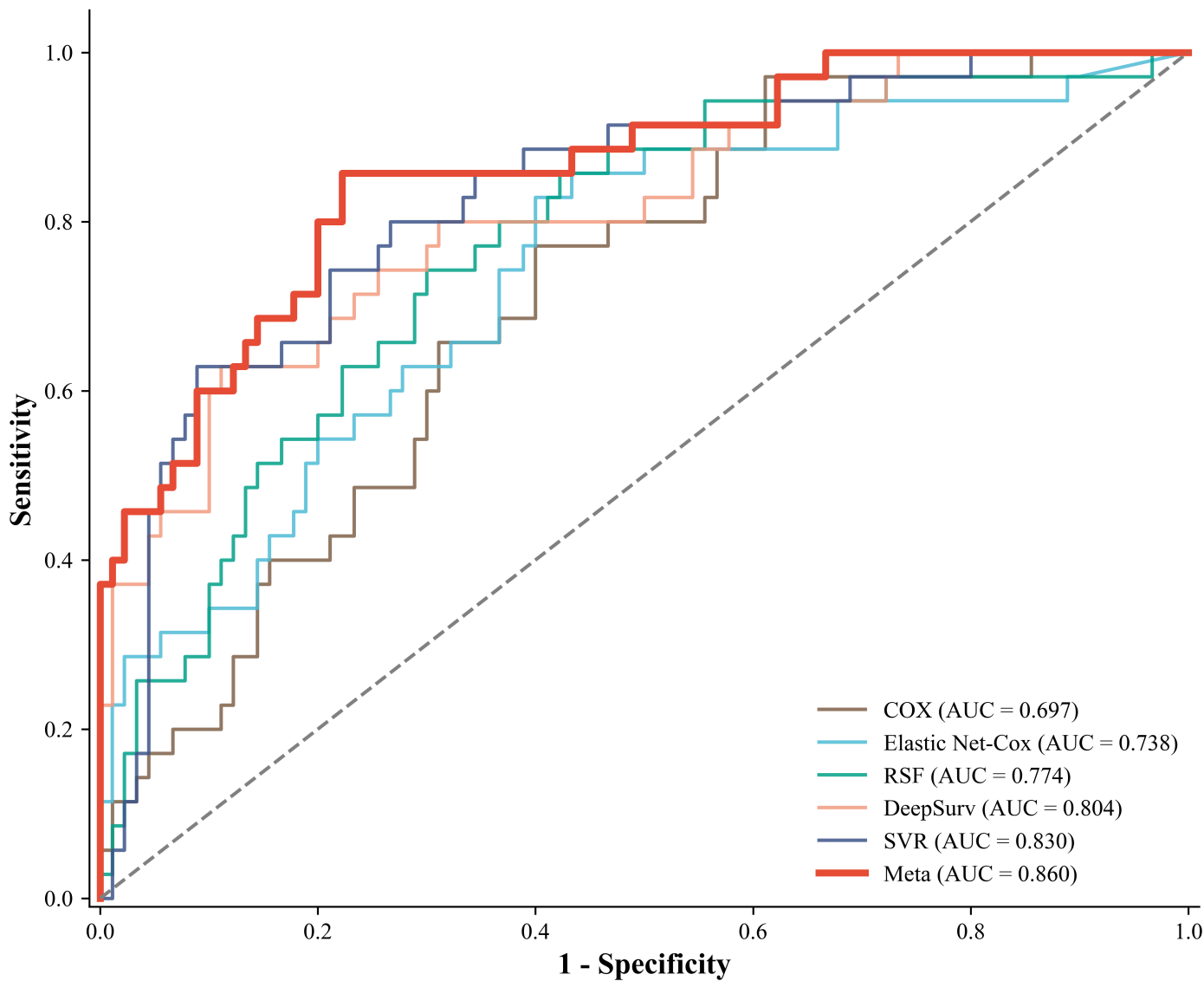
ROC Curves (Training Cohort, n=293)



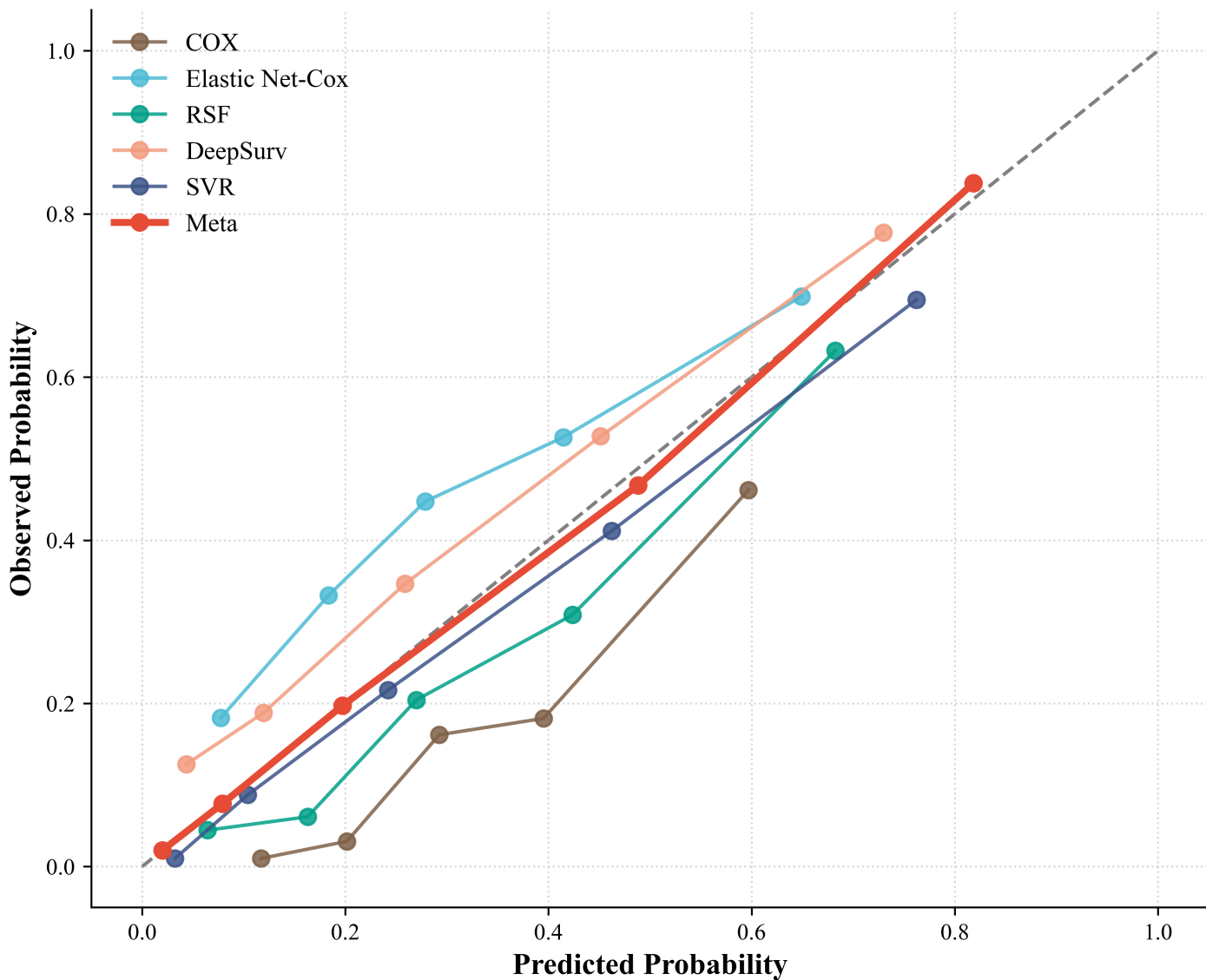
ROC Curves (Internal Validation Cohort, n=165)



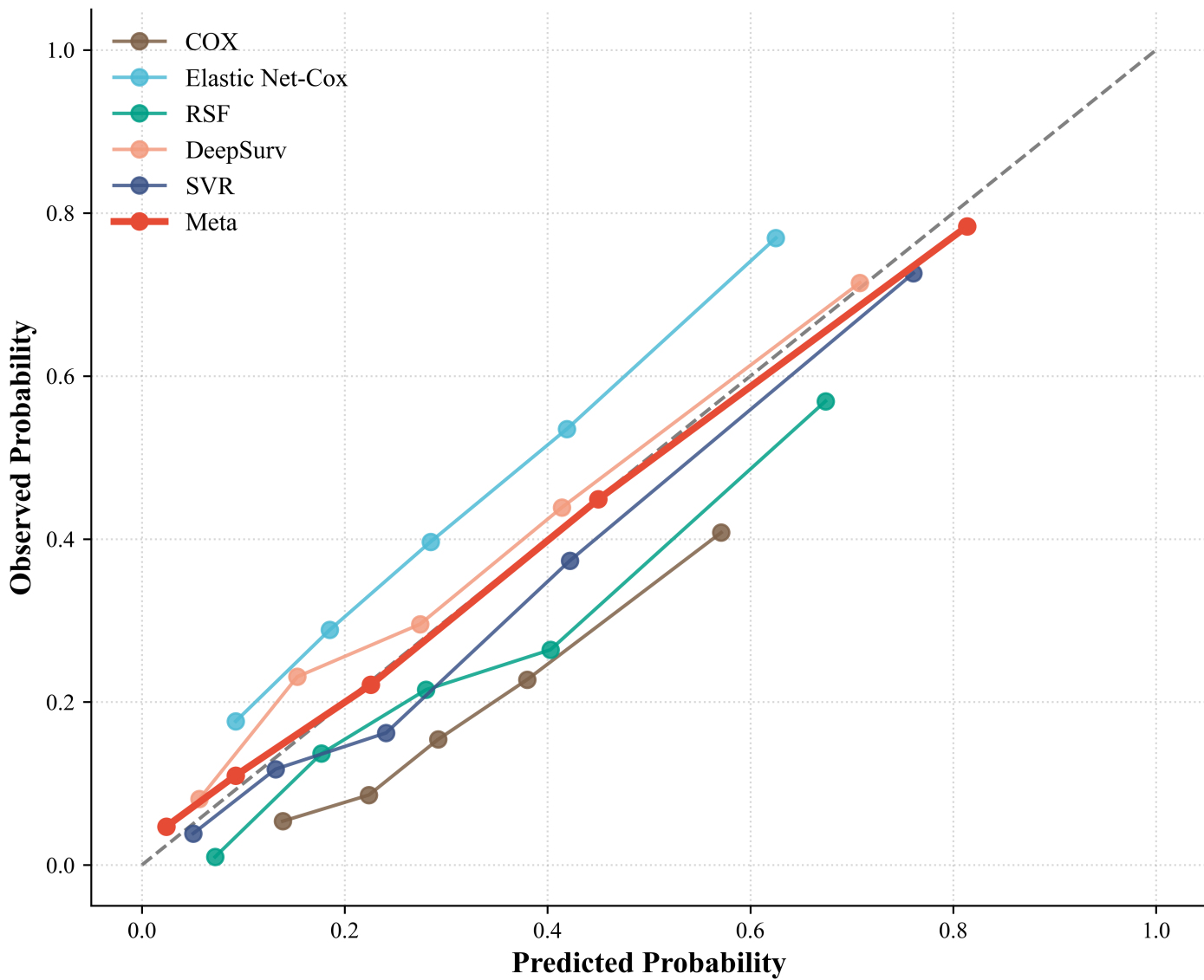
ROC Curves (External Validation Cohort, n=125)



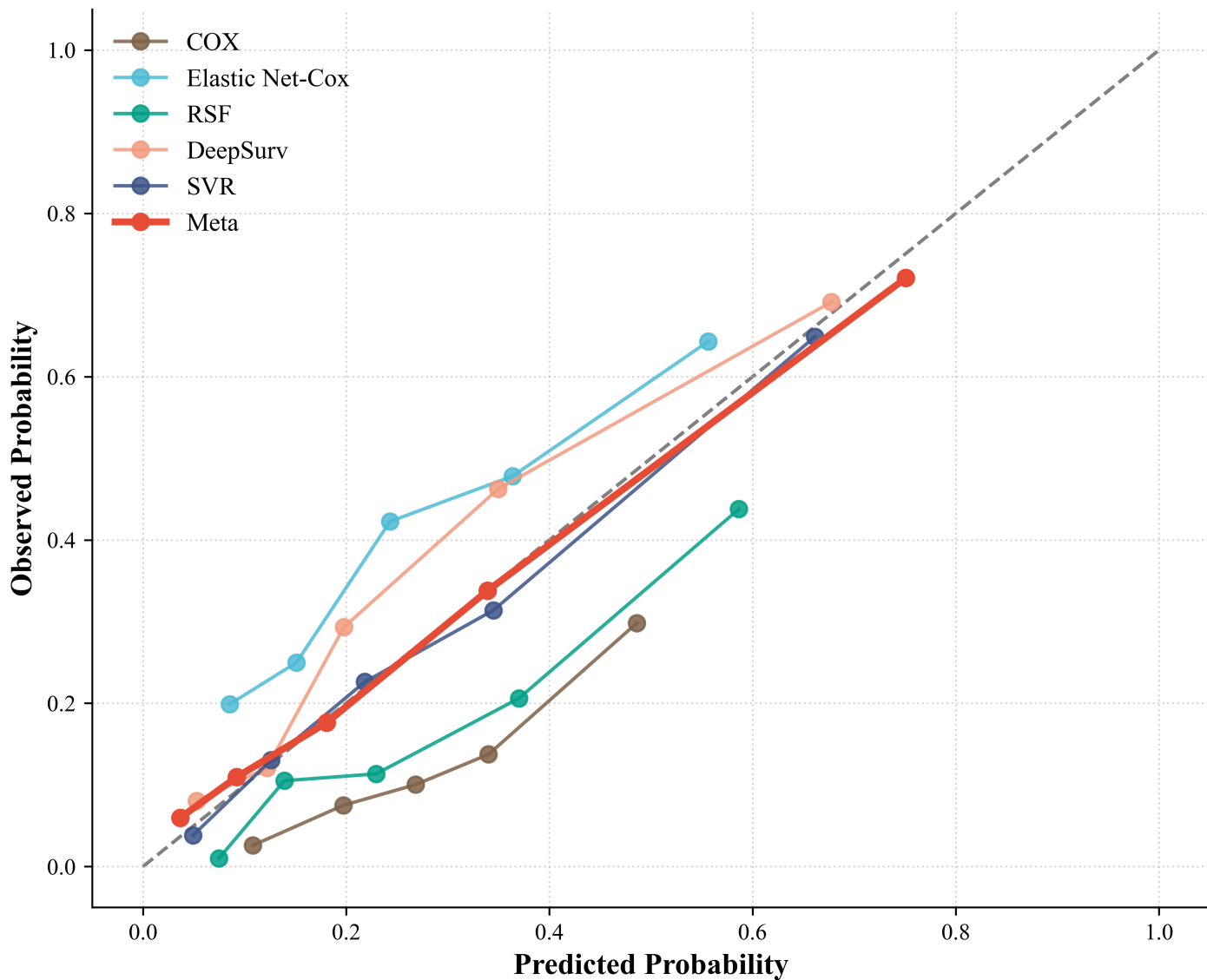
Calibration Curves (Training Cohort, n=293)



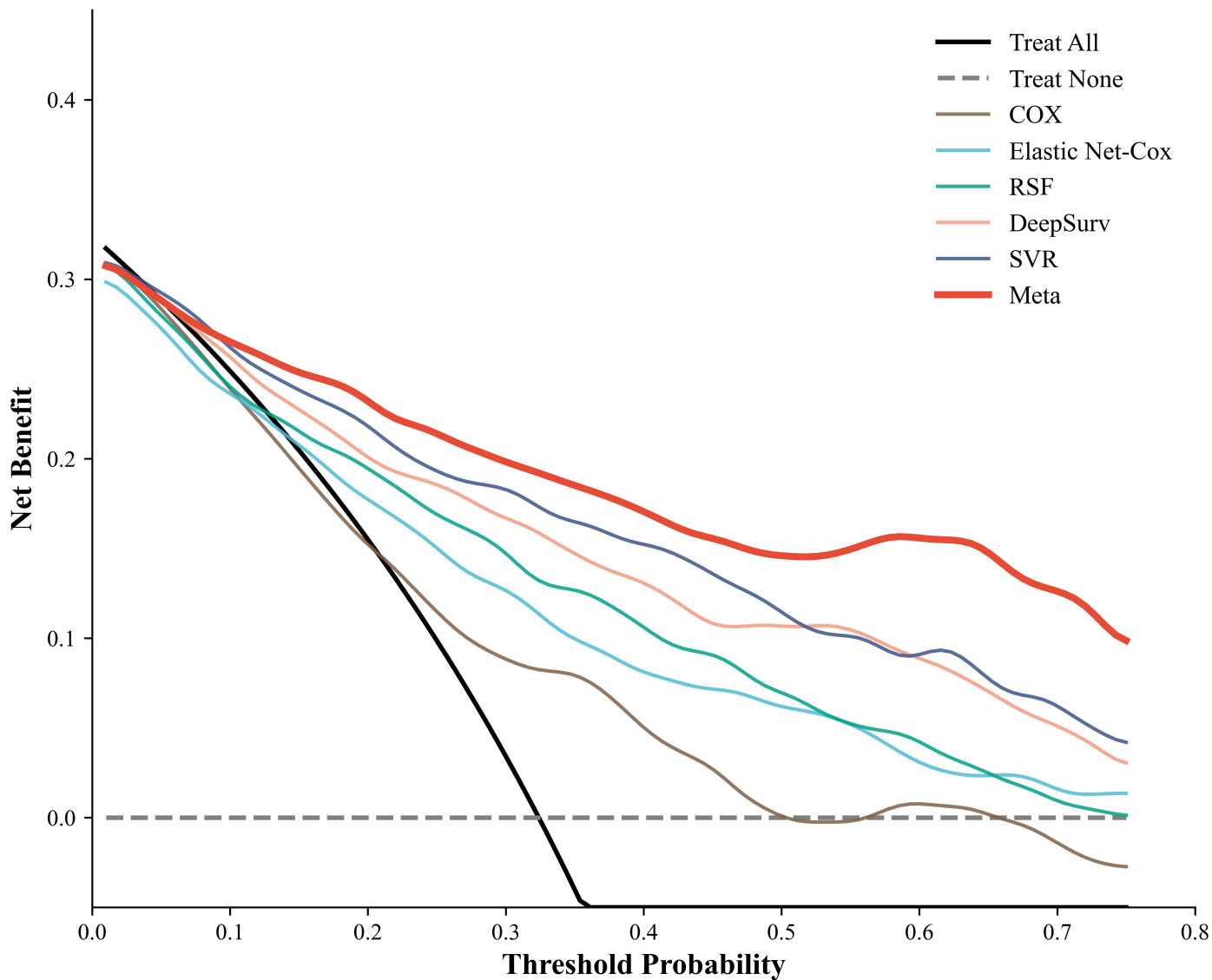
Calibration Curves (Internal Validation Cohort, n=165)



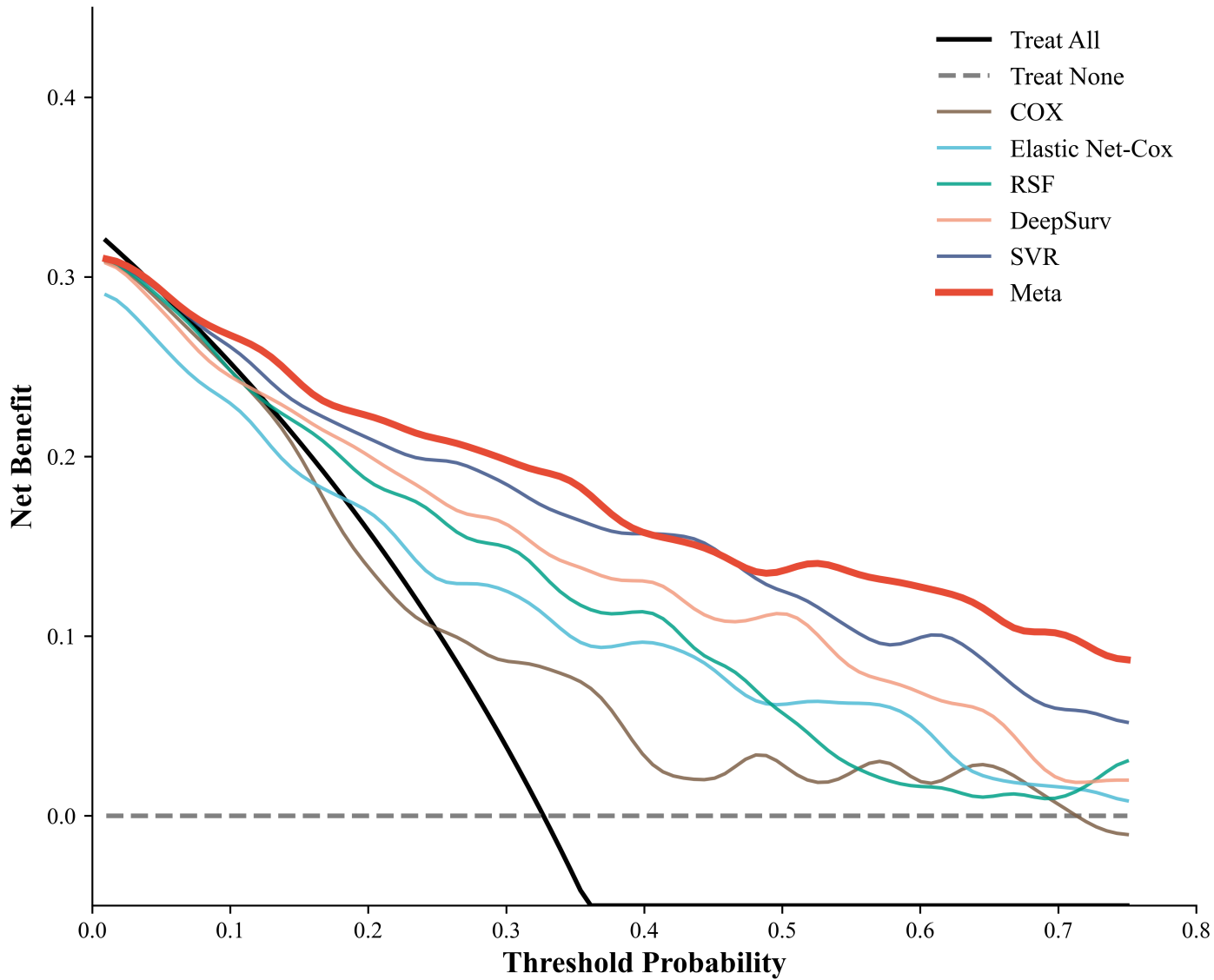
Calibration Curves (External Validation Cohort, n=125)



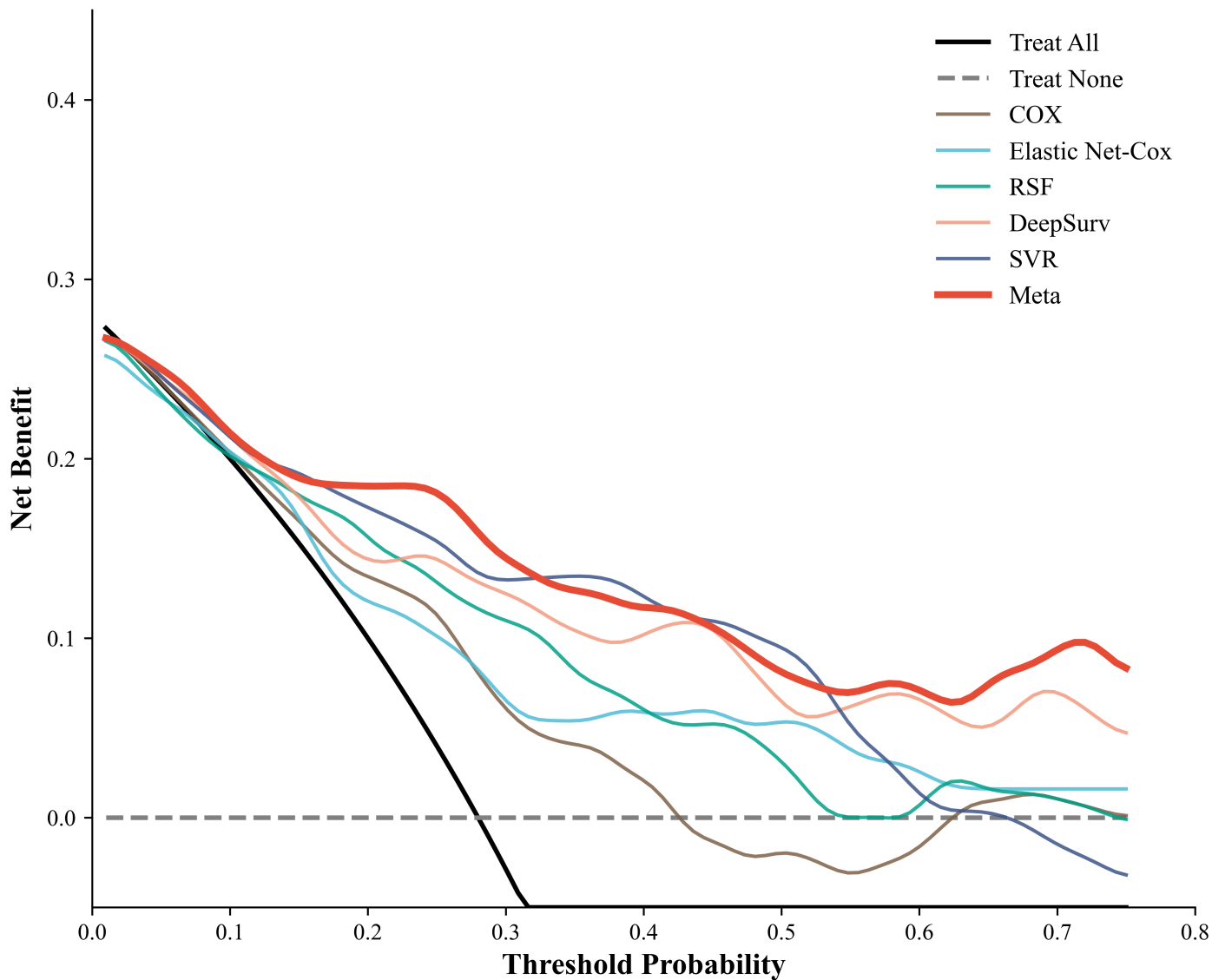
Decision Curve Analysis (Training Cohort, n=293)



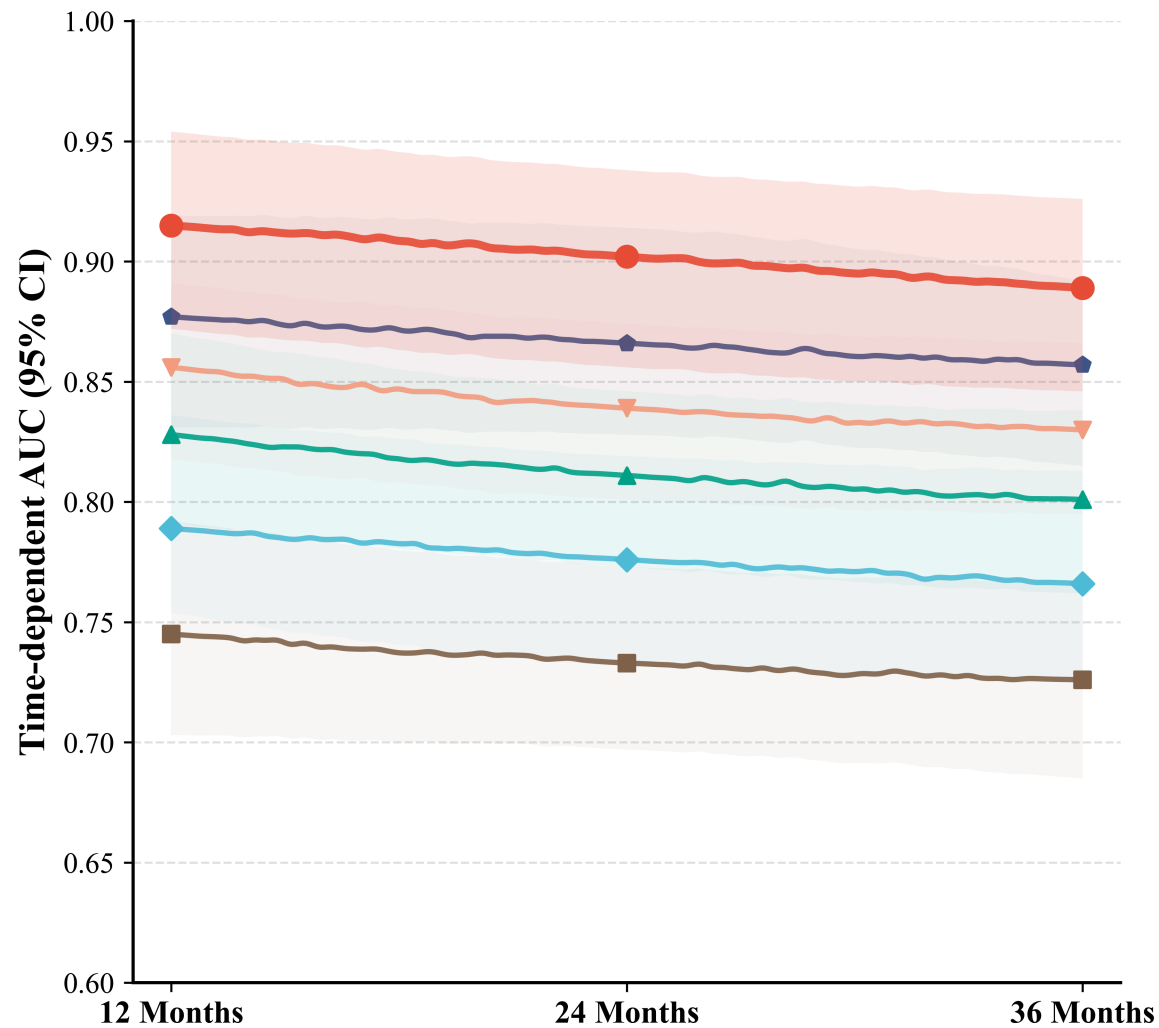
Decision Curve Analysis (Internal Validation Cohort, n=165)



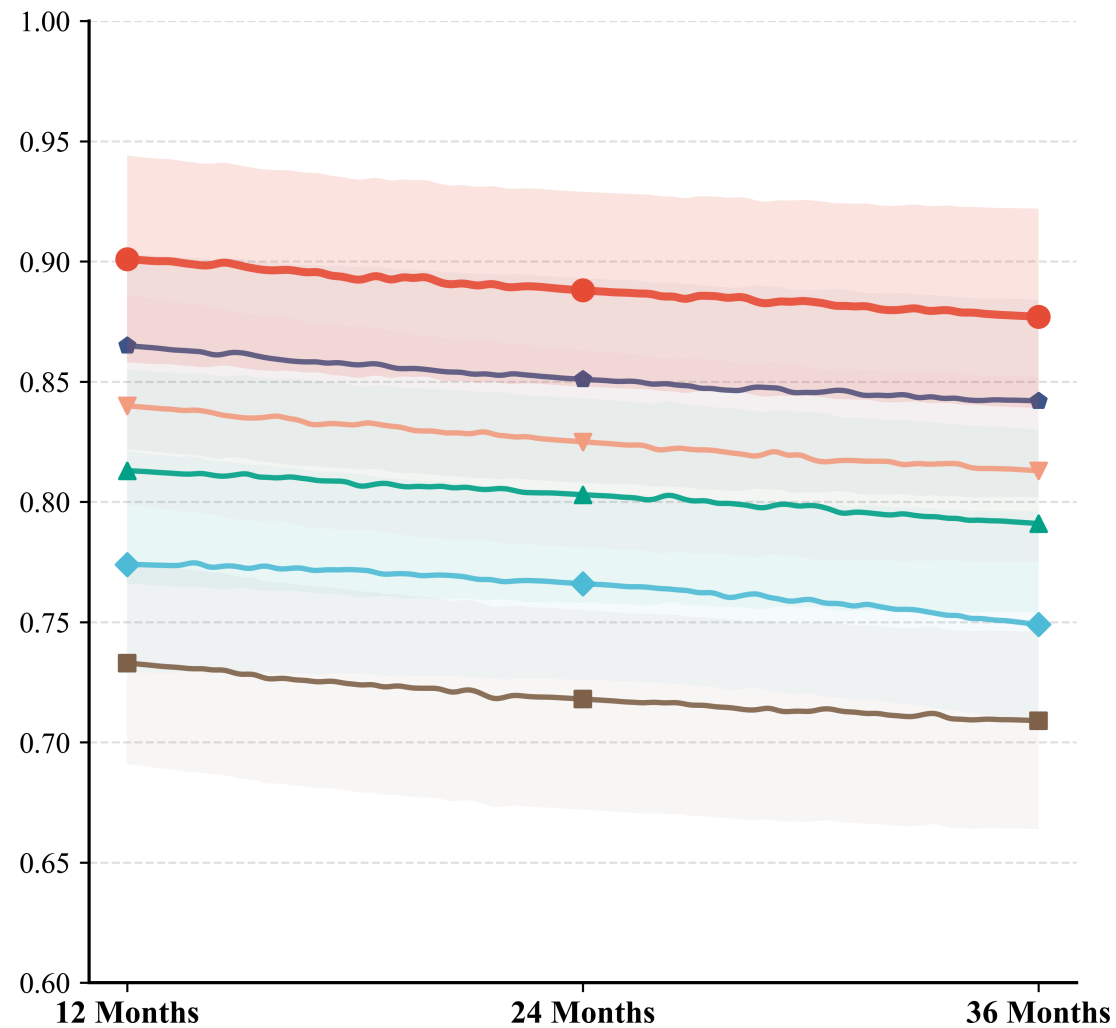
Decision Curve Analysis (External Validation Cohort, n=125)



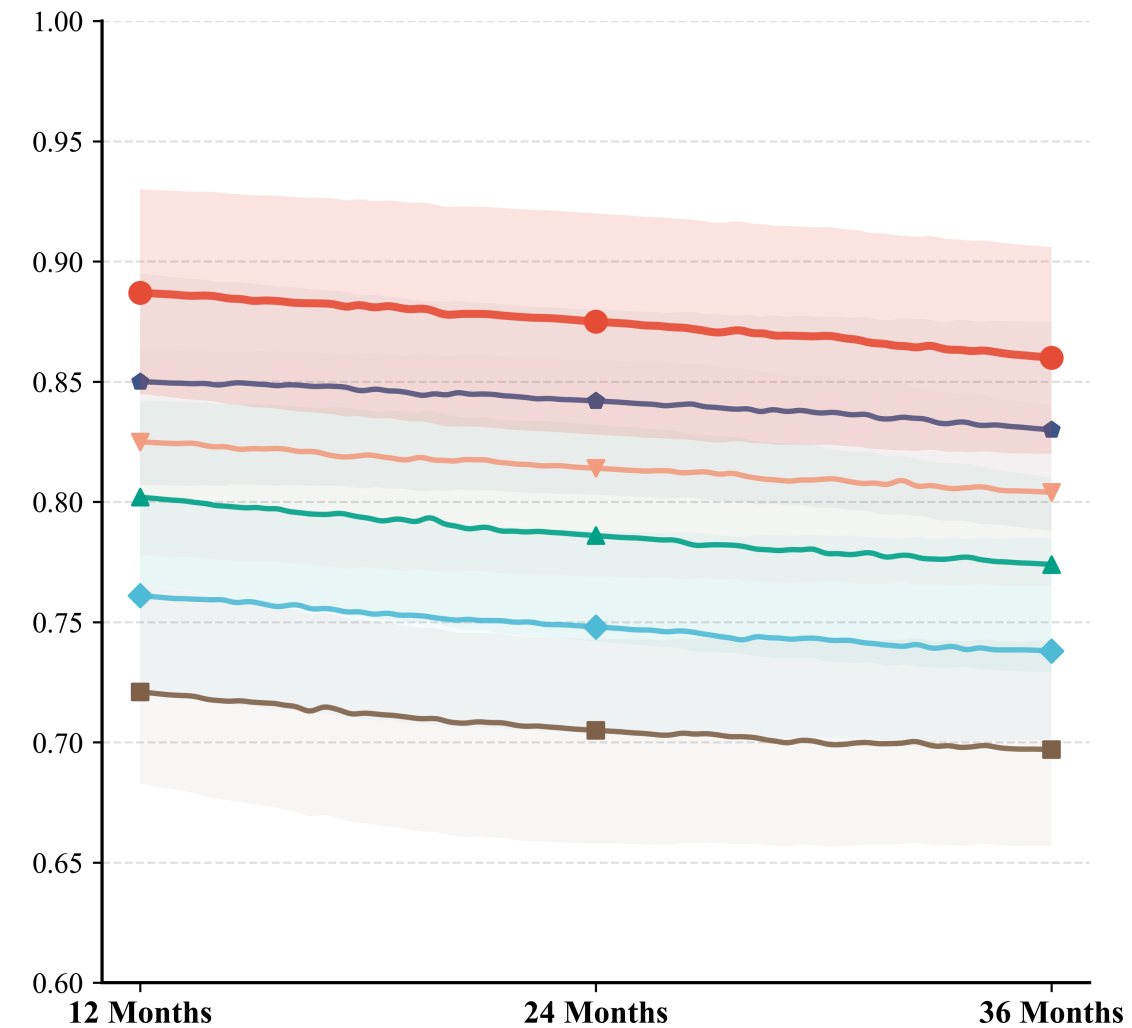
Training Cohort



Internal Validation Cohort

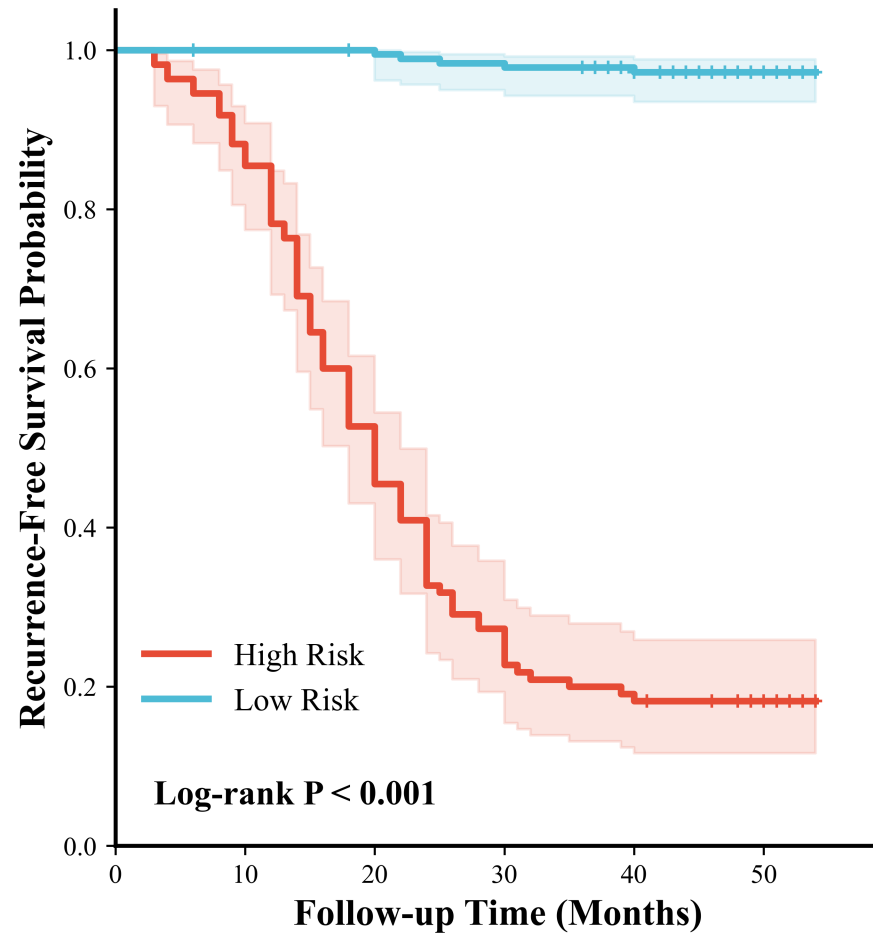


External Validation Cohort

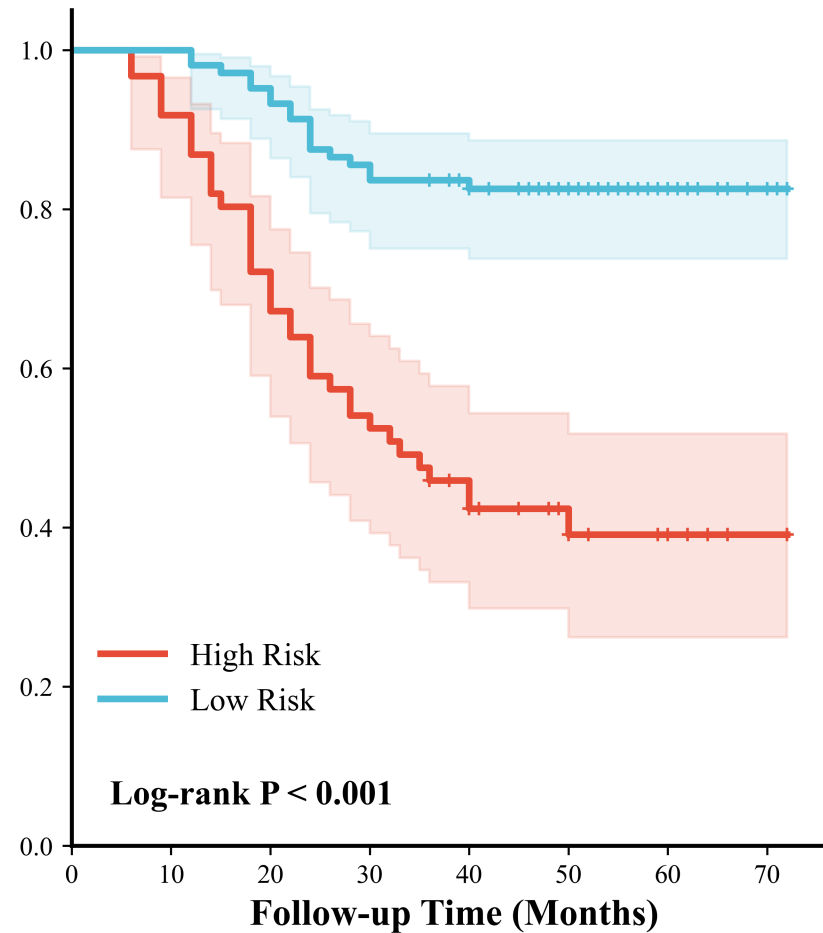


— COX
 — Elastic Net-Cox
 — RSF
 — DeepSurv
 — SVR
 — Meta

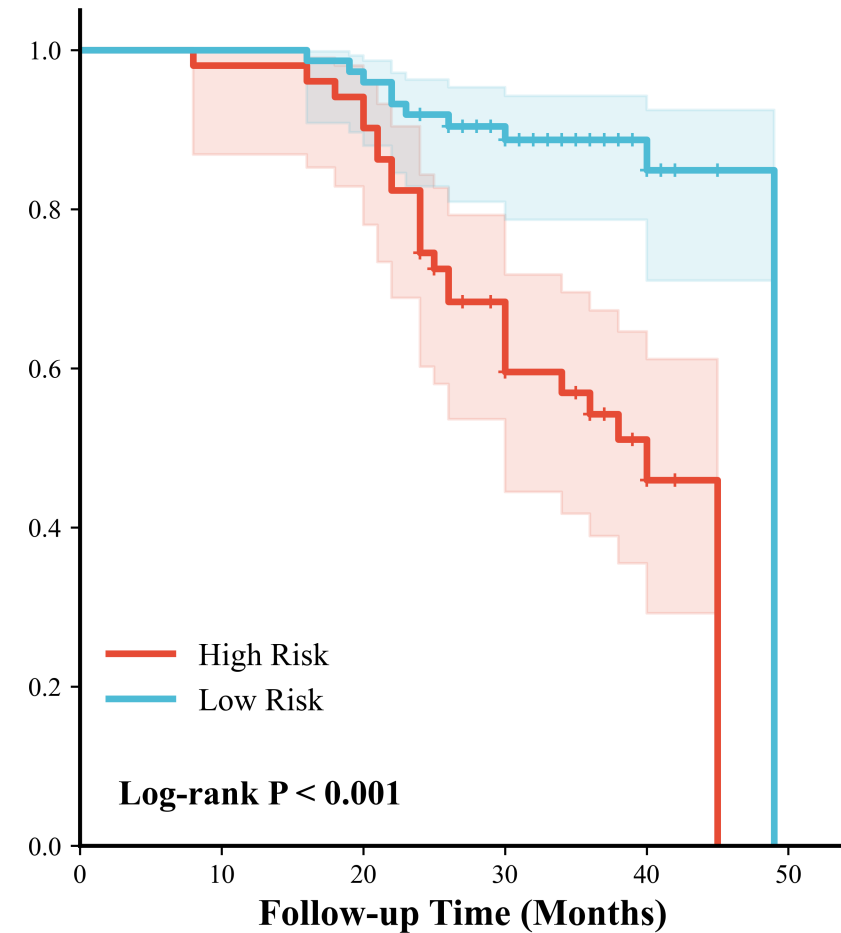
Training Cohort



Internal Validation Cohort



External Validation Cohort

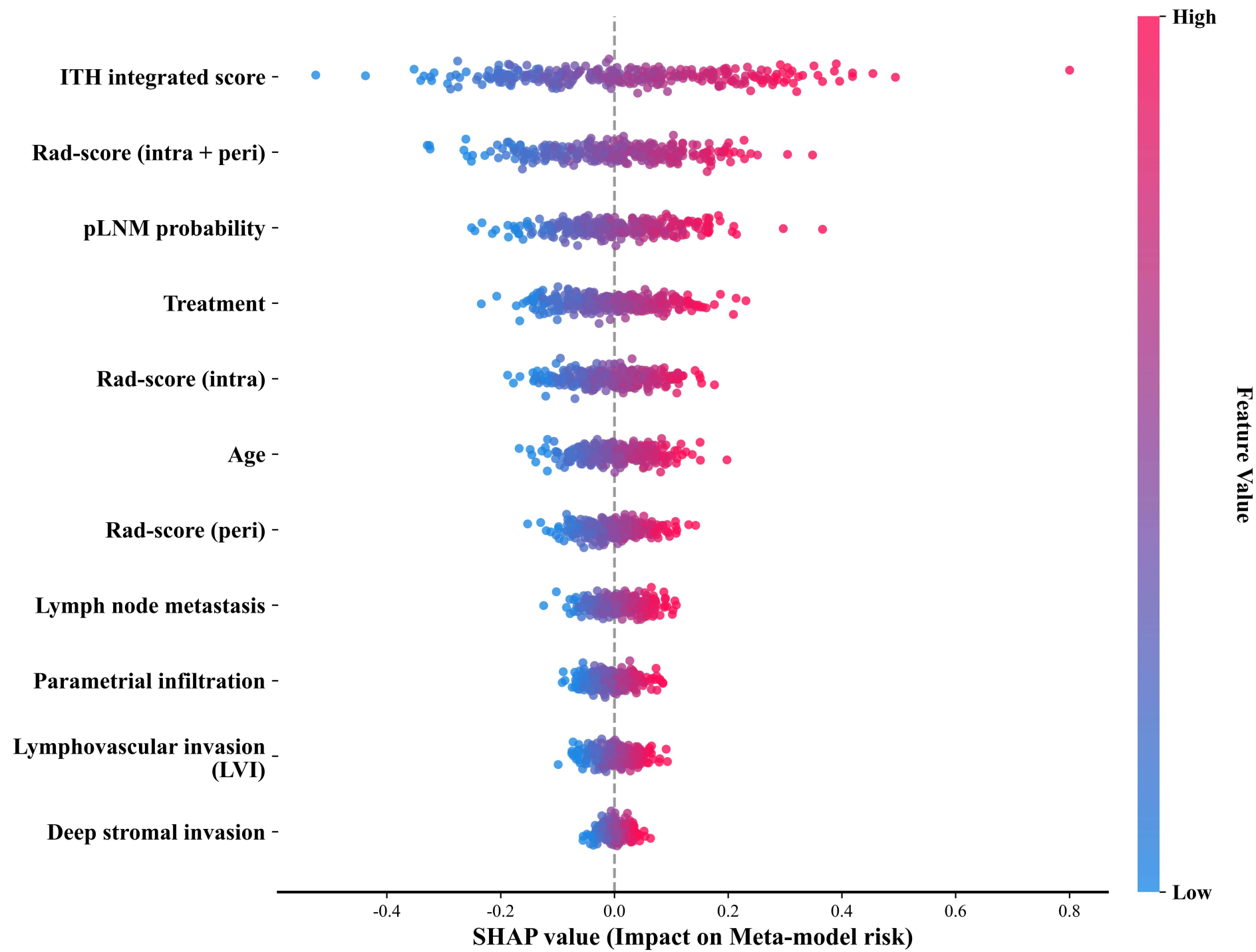


High Risk						
At risk	110	94	50	25	20	9
Censored	0	0	0	0	0	11
Events	0	16	60	85	90	90
Low Risk						
At risk	183	182	180	177	161	65
Censored	0	1	2	2	17	113
Events	0	0	1	4	5	5

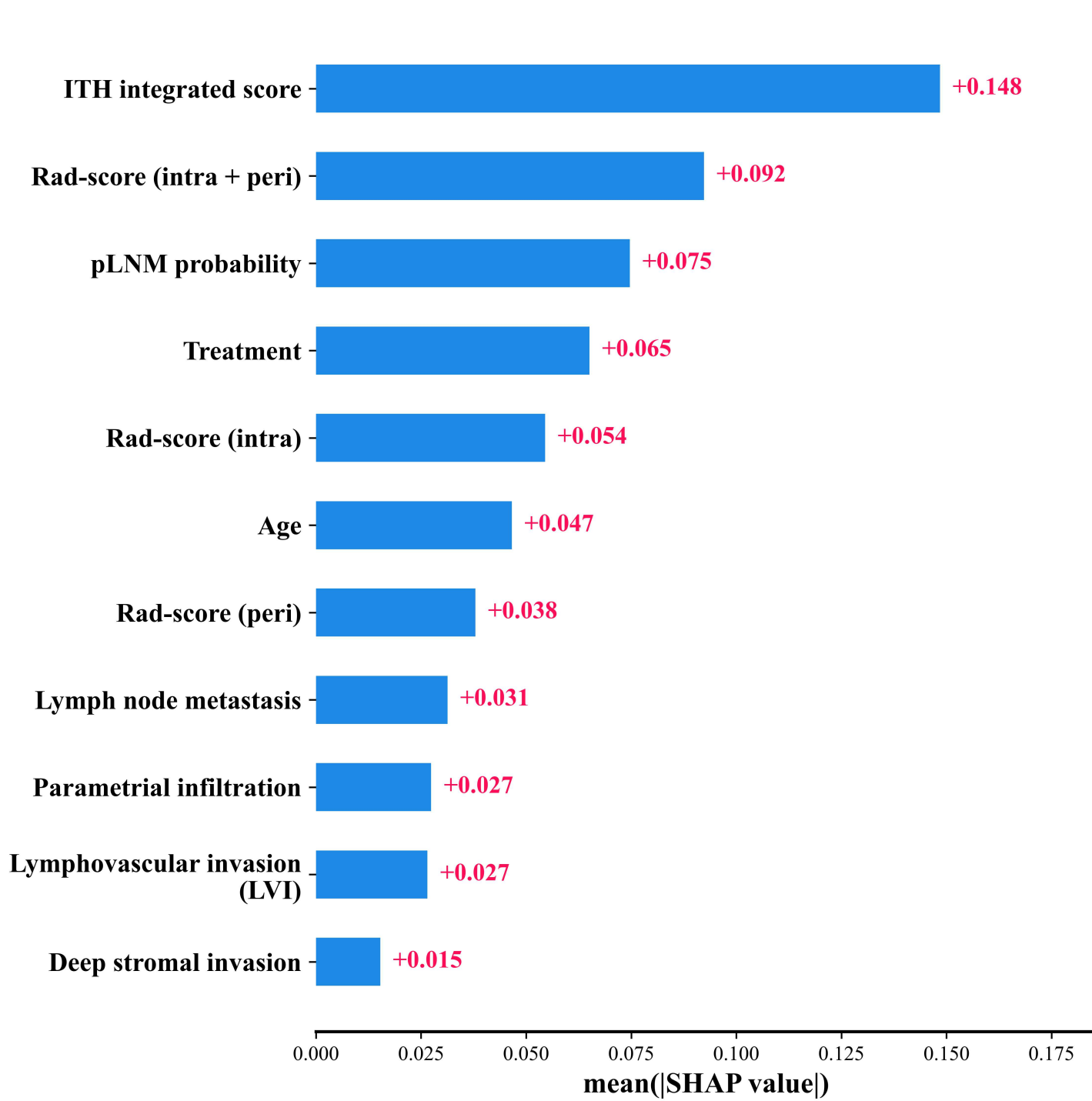
High Risk								
At risk	61	56	41	32	20	10	5	1
Censored	0	0	0	0	6	15	20	24
Events	0	5	20	29	35	36	36	36
Low Risk								
At risk	104	104	97	87	74	50	28	5
Censored	0	0	0	0	12	36	58	81
Events	0	0	7	17	18	18	18	18

High Risk						
At risk	51	50	46	23	5	0
Censored	0	0	0	8	22	26
Events	0	1	5	20	24	25
Low Risk						
At risk	74	74	71	52	14	0
Censored	0	0	0	14	51	64
Events	0	0	3	8	9	10

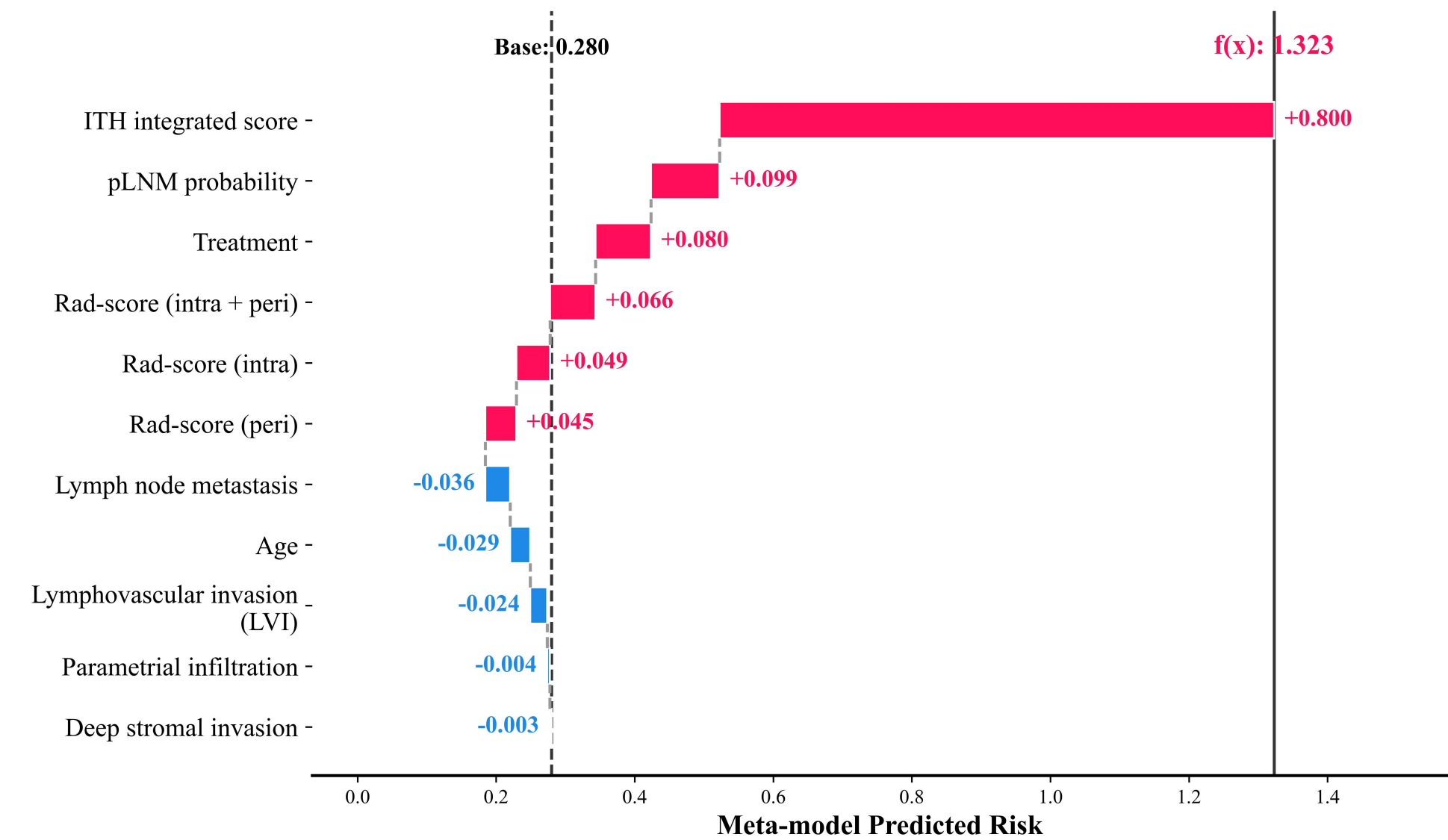
A. Global SHAP Summary



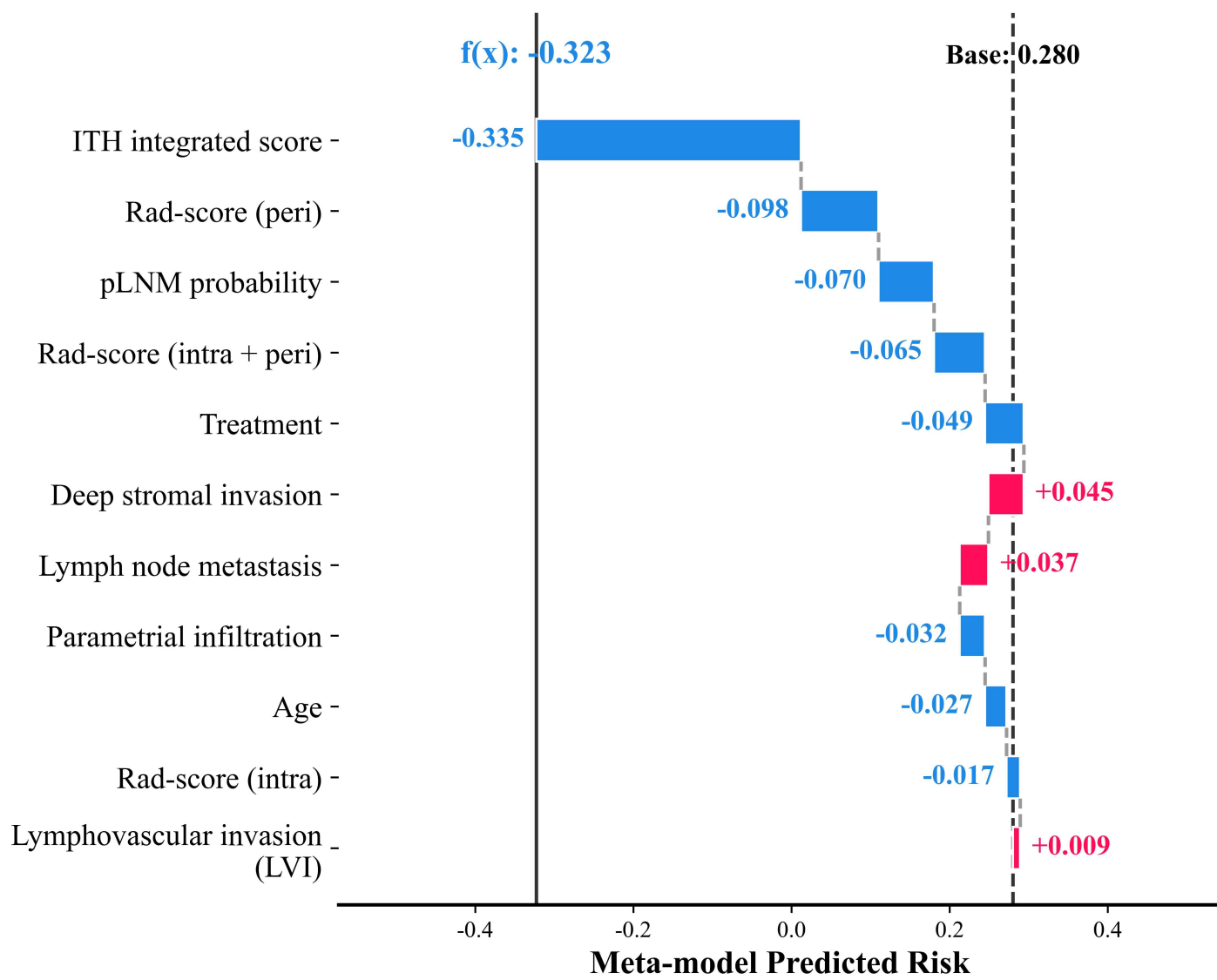
B. Mean Absolute Impact



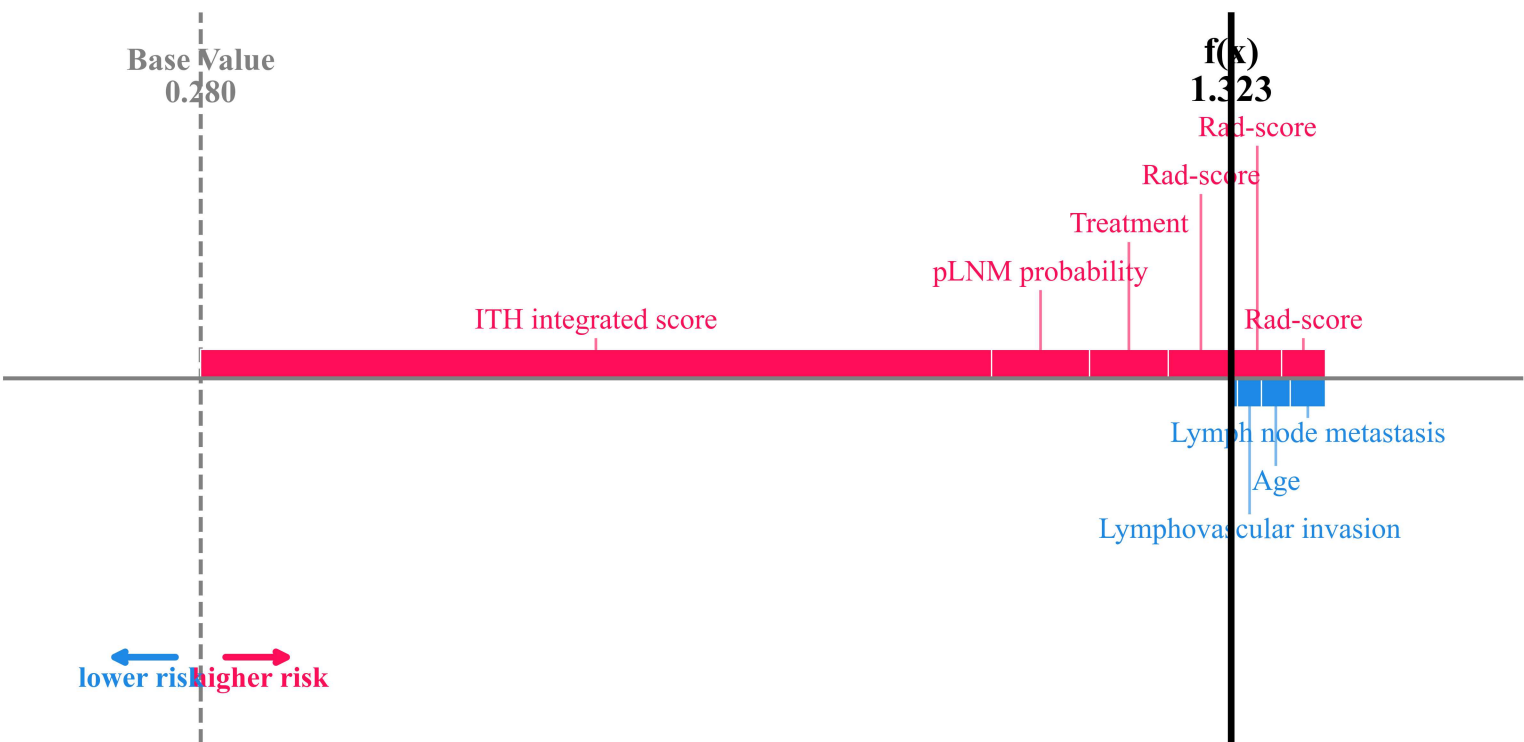
C. Waterfall (Typical High-Risk Patient)



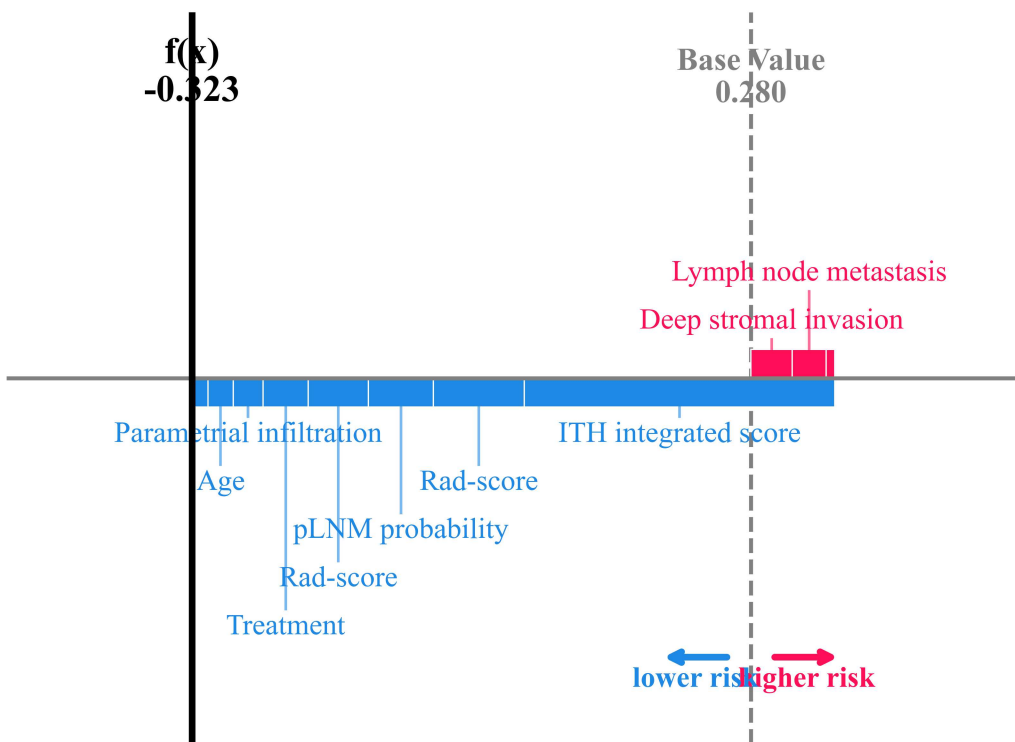
D. Waterfall (Typical Low-Risk Patient)



E. Force Plot (Typical High-Risk)



F. Force Plot (Typical Low-Risk)



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Tables.docx](#)
- [SupplementaryMaterial.docx](#)