

Random Survival Forest quantifies intra-stage heterogeneity of recurrence-free survival after curative gastrectomy: a single-center cohort study

Goshi Fujimoto

g_chimera_7@yahoo.co.jp

Kameda Medical Center <https://orcid.org/0000-0002-4793-5503>

Hiroshi Kusanagi

Kameda Medical Center

Research Article

Keywords: gastric cancer, machine learning, recurrence-free survival, intra-stage heterogeneity, random survival forest, ACTS-GC

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11213616/v1>

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

[Read Full License](#)

Abstract

Background

Pathological TNM staging classifies recurrence risk at the group level but cannot resolve intra-stage biological heterogeneity. We aimed to develop a Random Survival Forest (RSF) model predicting recurrence-free survival (RFS) after curative gastrectomy and to quantify the clinical significance of intra-stage risk heterogeneity.

Methods

This retrospective single-center study included patients who underwent R0 gastrectomy for pathological Stage IIA–IIIC gastric cancer (2008–2020). Ten clinicopathological predictors were modeled using RSF with multivariate iterative imputation. The development cohort (2008–2015; $n = 330$) was temporally validated (2016–2020; $n = 156$) using Harrell's C-index, time-dependent AUC (td-AUC), decision curve analysis (DCA), and net reclassification improvement (NRI).

Results

Among 486 patients, RSF achieved a validation C-index of 0.729 (95% confidence interval 0.653–0.809) and td-AUC of 0.764 and 0.777 at 3 and 5 years, respectively, compared with 0.687, 0.711, and 0.731 for pathological stage alone. The predicted 3-year cumulative recurrence probability showed clinically meaningful intra-stage heterogeneity for Stages II–IIIB (interquartile range, 10–15 percentage points; log-rank $p < 0.001$ for each). Stage IIIC showed a narrow predicted-risk distribution (coefficient of variation 0.18), reflecting uniformly high recurrence burden. DCA demonstrated net benefit over a stage-based strategy at 20–50% threshold probability.

Conclusions

RSF quantifies clinically meaningful intra-stage heterogeneity in Stages II–IIIB gastric cancer after curative resection. Stage IIIC represents a uniformly high-risk subgroup already well-captured by TNM staging. These findings provide a risk-stratification framework for individualized postoperative surveillance planning and a benchmark for multicenter validation.

Trial registration:

UMIN Clinical Trial Registry, UMIN000061288 (registered 1 May 2026).

Introduction

Gastric cancer, one of the most prevalent malignancies worldwide, is a leading cause of cancer-related mortality in Japan [1]. Despite curative resection, recurrence occurs in 20%–50% of patients and substantially compromises long-term prognosis [2, 3]. Current risk stratification relies primarily on pathological TNM staging; however, demonstrating that this classification inadequately separates patients with varying prognoses within the same stage, Sano et al. [4] proposed a revised TNM stage grouping to address intra-stage outcome heterogeneity, underscoring the need for more precise risk stratification tools.

Accurate prediction of postoperative recurrence could directly influence clinical practice by individualizing postoperative surveillance frequency, supporting perioperative chemotherapy selection, and facilitating shared decision-making [5, 6]. Furthermore, as biomarker-driven perioperative therapies based on HER2 expression, microsatellite instability (MSI), and programmed death-ligand 1 (PD-L1) status continue to evolve [7, 8], recurrence risk models may add another layer of precision to guide treatment intensity.

Several machine learning (ML) approaches have been developed to predict gastric cancer recurrence. A multicenter study of 11,615 patients from 10 Chinese institutions developed 10 ML models to predict early recurrence, reporting areas under the receiver operating characteristic curves (AUROCs) of 0.80–1.00 [9], although the study lacked an independent validation cohort. A single-center comparison of five ML algorithms achieved an AUROC of up to 0.962 [10]; however, calibration and decision curve analysis (DCA) were not evaluated. A systematic review of ML survival prediction models in gastric cancer found that most published studies lacked sufficient independent validation and had inadequately handled missing data [11]. Importantly, previous ML studies predominantly adopted binary outcome modeling without incorporating censoring information, which violates the assumptions required for valid clinical risk prediction. Furthermore, these studies were predominantly conducted in Western and Chinese cohorts.

To address these limitations, we followed the TRIPOD + AI statement [12] and designed the present study to: (1) develop an RSF model predicting recurrence-free survival (RFS) after curative R0 gastrectomy for pathological Stage IIA–IIIC gastric cancer in the ACTS-GC era, using a time-to-event framework with appropriate censoring handling; (2) temporally validate model performance using Harrell's C-index, time-dependent AUC (td-AUC), calibration, and DCA; and (3) quantify the extent of intra-stage recurrence risk heterogeneity and determine whether it is of clinically meaningful magnitude.

Patients and Methods

Study Design and Participants

This was a single-center retrospective cohort study. Patients aged ≥ 20 years who underwent curative (R0) resection for gastric adenocarcinoma (pathological Stage IIA, IIB, IIIA, IIIB, or IIIC per the Japanese Classification of Gastric Carcinoma, 15th edition) between January 2008 and December 2020 were eligible. Eligible surgical procedures included total gastrectomy (TG), distal gastrectomy (DG), or

proximal gastrectomy (PG) with lymph node dissection [13]. Patients with poor functional status or those who underwent emergency procedures were excluded. Additional exclusion criteria included: (1) concurrent non-gastric malignancy, (2) neoadjuvant chemotherapy, and (3) a follow-up duration of < 90 days without a recurrence event. Pathological Stage IIA encompasses T3N0, T2N1, and T1bN2 disease; per Japanese guidelines and institutional practice, adjuvant S-1 was not routinely administered for T3N0 or T1bN2 disease (reflecting guideline-concordant decision-making in the ACTS-GC era rather than protocol non-compliance). The Institutional Review Board of Kameda Medical Center approved this study (approval number: 25-003), and the requirement for individual informed consent was waived because of the retrospective study design.

Data Partitioning

Patients treated between January 2008 and December 2015 comprised the development cohort (n = 330), and those treated between January 2016 and December 2020 formed the temporal internal validation cohort (n = 156). This chronological split was pre-specified to reflect a realistic deployment scenario while ensuring a minimum events-per-variable (EPV) ratio of 10.2 in the development cohort [14]. The validation cohort was excluded entirely from model development.

Outcome

The primary outcome was recurrence-free survival (RFS), defined as the time from surgery to confirmed first recurrence at any site. Death without prior recurrence was treated as a competing event [15]. Patients alive without recurrence at the last follow-up visit were censored. Recurrence was confirmed via imaging (computed tomography), histopathological diagnosis, or both.

Predictors

Ten clinicopathological predictors available at four to six weeks after surgery (when pathological results are finalized) were pre-specified: age, serum albumin (Alb), C-reactive protein (CRP), carcinoembryonic antigen (CEA), carbohydrate antigen 19-9 (CA19-9), pathological T-category (pT; ordinal encoding 1a=1 to 4b=6), number of pathologically positive lymph nodes (pN_num), maximum tumor diameter (size), lymphovascular invasion grade (ly; ordinal 0–3), and administration of adjuvant S-1 chemotherapy (binary). Adjuvant S-1 was included as a covariate to adjust for guideline-concordant treatment differences, not as a treatment effect estimator. The Charlson Comorbidity Index (CCI) was excluded because, in a competing-risks framework where death without recurrence is handled separately, CCI primarily predicts the competing event rather than cancer-specific recurrence.

Missing Data

Missing values were handled using multivariate iterative imputation (IterativeImputer; estimator: Bayesian Ridge, max_iter = 20), which sequentially imputes each variable from all others and preserves the intervariable correlation structure better than single-value imputation [11]. To prevent data leakage,

the imputer was fitted exclusively on the development cohort and applied via transformation to the validation cohort.

Model Development

The primary model was a Random Survival Forest (RSF) [16] implemented in scikit-survival (`n_estimators = 500`, `min_samples_leaf = 15`, `max_features = "sqrt"`, random seed 42). The null comparator was a Cox proportional hazards model with pathological stage as the sole ordinal predictor, used to quantify the incremental value of the RSF. No class-resampling (SMOTE) or post-hoc probability calibration was applied, because the RSF directly estimates survival functions that accommodate censoring.

Model Evaluation

Discrimination was assessed using Harrell's C-index and td-AUC at 3 and 5 years [17]. Calibration was assessed using quintile-based Kaplan–Meier observed versus model-predicted cumulative incidence at 3 and 5 years. Bootstrap 95% confidence intervals (CIs) for all discrimination metrics were derived from 1,000 validation-cohort resamples. Clinical utility was evaluated using DCA [18] at the 3-year horizon. Net Reclassification Improvement (NRI) and Integrated Discrimination Improvement (IDI) were calculated using continuous reclassification, with bootstrap 95% CIs from 1,000 iterations [19].

Intra-Stage Heterogeneity Analysis

For each stage group (Stage II [IIA+IIB combined; observed recurrence rates 15.3% and 17.9%, respectively], Stage IIIA, Stage IIIB, and Stage IIIC), the distribution of RSF-predicted 3-year cumulative recurrence probability was visualized as violin plots. The interquartile range (IQR) of predicted risk within each stage group was the primary heterogeneity metric, with ≥ 10 percentage points pre-specified as clinically meaningful. The coefficient of variation (CV) characterized relative spread. Patients were classified as low or high risk by the development-cohort stage-specific median predicted probability, and Kaplan–Meier RFS curves were compared by log-rank test in the combined cohort.

Model Interpretability

Variable importance was quantified by permutation importance (30 repeats; decrease in C-index) using the development cohort. The direction of predictor effects was confirmed using SHapley Additive exPlanations (SHAP) values [20] computed with a gradient-boosting classifier proxy model.

Statistical Analyses

Continuous variables were compared using the Mann–Whitney *U* test, and categorical variables were compared using the chi-squared test (two-sided, $\alpha = 0.05$). All analyses were conducted in Python 3.13 (scikit-learn, scikit-survival, lifelines, shap). The analysis code is publicly available at https://github.com/GoshiFujimoto/gc_recurrence_rsf. The random seed was fixed at 42 throughout all

computations. This study was conducted and reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement.

Results

Patient Characteristics

A total of 486 patients met the inclusion criteria (development $n = 330$, temporal validation $n = 156$; Table 1). Recurrence occurred in 102 of 330 development patients (30.9%) and 51 of 156 validation patients (32.7%). Competing events (death without recurrence) occurred in 63 and 19 patients, respectively. The cohorts were broadly similar in age, sex, tumor markers, lymph node count, and tumor size.

Lymphovascular invasion grade was significantly higher in the development cohort (median 2.0 vs. 1.0; $p < 0.001$), and censored follow-up was longer (median 7.3 vs. 5.0 years; $p < 0.001$)—both structural consequences of the chronological split.

Temporal Validation Performance

In the temporal validation cohort, RSF achieved a C-index of 0.729 (95% CI 0.653–0.809) versus 0.687 (0.616–0.765) for pStage-only Cox (Table 2; Fig. 1). Time-dependent AUC at 3 and 5 years was 0.764 (0.673–0.851) and 0.777 (0.688–0.863) for RSF and 0.711 (0.619–0.802) and 0.731 (0.637–0.823) for pStage-only Cox, respectively. All 95% CIs overlapped between models, reflecting the statistical limitations of the 156-patient validation cohort. Calibration plots at 3 and 5 years demonstrated good agreement at lower risk quintiles, with modest over-prediction at the highest quintile (Fig. 2).

DCA, NRI, and IDI

DCA demonstrated that the RSF model yielded higher net benefit than a pStage-based strategy across the clinically relevant threshold range of 20%–50% predicted 3-year recurrence probability (Fig. 3). Continuous NRI was 0.065 (95% CI –0.126 to 0.242) and IDI was 0.017 (95% CI –0.005 to 0.039), neither reaching statistical significance, reflecting limited statistical power at the current sample size.

Intra-Stage Heterogeneity

The violin plots of predicted 3-year cumulative recurrence demonstrated clinically meaningful intra-stage heterogeneity for Stages II, IIIA, and IIIB (Fig. 4; Table 3). The IQR of predicted 3-year recurrence was: Stage II 0.101 (10.1 percentage points), Stage IIIA 0.136, and Stage IIIB 0.147, all exceeding the pre-specified 10 percentage-point threshold. Kaplan–Meier analysis (combined cohort) showed significant RFS separation between low- and high-risk groups in Stage II (log-rank $p < 0.001$; threshold 17.3%), Stage IIIA ($p < 0.001$; threshold 29.5%), and Stage IIIB ($p < 0.001$; threshold 34.5%). In Stage II, recurrence occurred in 10/118 low-risk patients (8.5%) versus 30/125 high-risk patients (24.0%); in Stage IIIA, 10/55 (18.2%) versus 30/53 (56.6%); in Stage IIIB, 7/36 (19.4%) versus 18/30 (60.0%).

Stage IIIC exhibited a notably narrow predicted-risk distribution (IQR 0.157; CV 0.180), substantially lower than Stages II–IIIA (CV 0.355–0.401), with a median predicted 3-year recurrence of 54.9%. Median-based two-group stratification showed no significant RFS difference ($p = 0.678$). Quartile-based analysis revealed a significant risk gradient (Q1 vs. Q4 log-rank $p = 0.019$; Q1 event rate 44.4% vs. Q4 94.1%).

Variable Importance

Permutation importance identified positive lymph node count as the dominant predictor (ΔC -index = 0.111), followed by pT-category (0.038), age (0.021), CA19-9 (0.019), and CEA (0.017) (Fig. 5). SHAP analysis confirmed that higher lymph node count and higher pT category were associated with increased predicted recurrence risk, while higher albumin was protective. Adjuvant S-1 showed the lowest importance (ΔC -index = 0.003), consistent with its role as a treatment covariate in a population receiving guideline-concordant care.

Discussion

In this retrospective study of Japanese patients who underwent curative resection for pathological Stage IIA–IIIC gastric cancer, the RSF model achieved a validation C-index of 0.729 and td-AUC of 0.764 and 0.777 at 3 and 5 years, demonstrating meaningful intra-stage risk heterogeneity in Stages II through IIIB.

Individualized Surveillance

The significant difference in RFS between the groups predicted to be at high and low risk (log-rank $p < 0.001$ for Stages II, IIIA, and IIIB) supports risk-stratified postoperative follow-up. However, the evidence that intensified surveillance improves survival in patients with gastric cancer remains inconclusive. A systematic review and meta-analysis reported that detecting asymptomatic recurrence was associated with improved post-recurrence survival [21], whereas another meta-analysis found no significant differences in recurrence detection rates or post-recurrence survival with planned surveillance [22]. In this context, the clinical role of the present model may lie less in intensifying standard follow-up and more in identifying high-risk patients for expedited intervention at recurrence or enrollment in clinical trials.

Intra-Stage Heterogeneity and Stage IIIC

The principal finding of this study is that a 10-variable RSF model quantifies clinically meaningful intra-stage heterogeneity that TNM staging cannot resolve for Stages II–IIIB. The Stage IIIC finding warrants specific interpretation: the narrow predicted-risk distribution (CV = 0.180 versus 0.355–0.401 for Stages II–IIIA) and absence of significant two-group separation do not indicate model failure. Rather, Stage IIIC constitutes the most advanced resectable category, with a cohort recurrence rate of 69.6% and near-universal early recurrence burden. In this subgroup, the marginal contribution of nutritional and inflammatory biomarkers is overwhelmed by the dominant pathological burden, and TNM staging alone adequately captures the recurrence signal.

Actionable Clinical Proposals

Based on the present findings, we propose two concrete clinical applications of the RSF model. First, patients in the high-risk stratum (predicted recurrence probability exceeding the stage-specific median) should be considered candidates for intensified imaging follow-up—specifically, switching from semi-annual to quarterly computed tomography during the first 2 postoperative years. This interval-based trigger is consistent with existing recommendations for individualized surveillance based on recurrence hazard rates [5,6] and aligns with the 20–50% threshold range used in our DCA. Second, the model supports shared prognostic communication: a clinician can present an individualized cumulative recurrence probability, enabling informed discussion regarding surveillance intensity preferences and participation in clinical trials [23]. This model estimates recurrence risk given the treatment actually administered and does not support causal inferences about adjuvant chemotherapy benefit in individual patients.

Comparison with Prior Prediction Methods

The current standard for postoperative risk stratification is the pathological TNM staging system, which does not adequately capture intra-stage biological heterogeneity or the complex interplay among nutritional status, comorbidities, and perioperative factors. TNM-based nomogram studies are often derived from single-center datasets and rarely report calibration or DCA. Compared with a prior ML study of 11,615 patients from 10 Chinese institutions reporting AUROCs of 0.80–1.00 [9] and a separate federated learning model across four Chinese institutions achieving AUROCs of 0.710–0.869 [24], the present RSF model achieved a validation C-index of 0.729 versus 0.687 for pStage-only Cox, using a smaller, single-center Japanese dataset. The 0.042 C-index advantage is of modest absolute magnitude, and the 95% CIs overlap; however, DCA showed consistent net benefit over the stage-based strategy at the clinically relevant 20%–50% threshold range, indicating that continuous risk scores carry decision-relevant information not fully captured by stage-level predictions.

Characteristics of the Japanese Cohort

The study period (2008–2020) spans the ACTS-GC era of adjuvant S-1 for pathological Stage II–III gastric cancer, following the landmark ACTS-GC trial [25]. This era is characterized by considerable evolution in surgical technique (laparoscopic gastrectomy) and adjuvant chemotherapy regimens. The dominant role of positive lymph node count—rather than categorical pN stage—likely reflects the continuous biological gradient that categorical classification necessarily discretizes, consistent with prior evidence that the absolute lymph node ratio provides incremental prognostic information beyond categorical pN staging. The low permutation importance of adjuvant S-1 (0.003) is consistent with the prognostic—rather than predictive—role of this covariate: all guideline-eligible patients received standard-of-care treatment, limiting the model's ability to estimate treatment effect.

Interpretation of DCA

The 20%–50% threshold range used in the DCA was selected based on two specific clinical decision points: the intensification of postoperative imaging surveillance (switching from semi-annual to quarterly CT during the first 2 postoperative years) and consideration of clinical trial enrollment for high-risk patients. This selection is consistent with DCA methodology, wherein net benefit is evaluated over a range of threshold probabilities that clinicians would consider plausible triggers for action [18]. The NRI of 0.065 and IDI of 0.017, while directionally positive, did not reach statistical significance at the current sample size. The DCA findings nonetheless suggest that continuous risk scores carry decision-relevant information in the moderate-risk threshold range.

Study Limitations

First, temporal validation within a single institution does not establish geographic generalizability. The transportability of the model to institutions with different surgical volumes or case-mix characteristics could not be established from the present study. Validation in geographically distinct Japanese cohorts is required before broad claims of generalizability can be made. Nevertheless, temporal validation is widely recognized as a necessary and accepted intermediate step preceding full external validation [26]. The present findings provide a methodological foundation upon which future multicenter external validation studies, including those leveraging the Japan Surgical Society National Clinical Database (NCD) [25], can be built.

Second, the binary recurrence definition does not distinguish early (< 2 years) from late recurrence, which carry different prognostic implications [3]. Survival analysis frameworks incorporating time-to-recurrence type and site as competing endpoints should be explored in future work. Third, lymphovascular invasion grade differed significantly between cohorts ($p < 0.001$), potentially attenuating validation performance; this likely reflects evolving pathological reporting standards across the study period. Fourth, SHAP values were computed using a gradient-boosting classifier proxy due to current incompatibility between RSF and tree-based SHAP explainers; the variable importance ranking was consistent between permutation importance and proxy SHAP. Fifth, the EPV of 10.2 is borderline; bootstrap optimism correction was not applied, and mild overfitting cannot be excluded.

Prospects

Future research should involve (1) independent validation in geographically distinct cohorts through the NCD [25], targeting EPV ≥ 10 ; (2) integration of CT-based imaging biomarkers and molecular markers (HER2, MSI-H, and PD-L1)—for which CT radiomics models enabling non-invasive simultaneous prediction have recently been developed [27]—to enhance model performance; and (3) a prospective interventional study evaluating whether risk-stratified surveillance guided by the model improves recurrence detection rates or survival. The TRIPOD+AI-compliant framework [12] and open-source code used in the present study are intended to serve as methodological templates for future ML prediction research in gastric cancer.

Conclusions

In a single-center Japanese cohort of patients undergoing curative resection for pathological Stage IIA–IIIC gastric cancer in the ACTS-GC era, the RSF model demonstrated that Stages II, IIIA, and IIIB harbor clinically meaningful intra-stage heterogeneity in RFS that TNM staging alone cannot resolve. Stage IIIC represents a uniformly high-risk subgroup already well-characterized by conventional staging. The model's continuous individual risk estimates provide a quantitative basis for risk-stratified postoperative surveillance planning, and the present findings serve as a methodological benchmark for future multicenter external validation.

Declarations

Acknowledgments: The authors thank Editage (www.editage.com) for the English language editing. Assistance from a large language model (LLM) was utilized in the development of the analysis code.

Ethics approval: This study was performed in line with the principles of the Declaration of Helsinki. Approval was obtained from the Institutional Review Board of Kameda Medical Center (approval number: 25-003). The requirement for individual informed consent was waived because of the retrospective study design; opt-out procedures were implemented and registered in the University Hospital Medical Information Network (UMIN) Clinical Trial Registry (UMIN000061288).

Conflict of interest: The authors have no competing interests to declare.

Funding: This study received no external funding.

Data availability: The analysis code is publicly available at https://github.com/GoshiFujimoto/gc_recurrence_rsf. Individual patient data cannot be shared publicly due to institutional privacy regulations.

Author contributions: GF: Conceptualization, Methodology, Formal Analysis, Writing (original draft). HK: Supervision, Writing (review and editing). Both authors read and approved the final manuscript.

References

1. Smyth EC, Nilsson M, Grabsch HI, van Grieken NC, Lordick F (2020) Gastric cancer. *Lancet* 396:635–648. [https://doi.org/10.1016/S0140-6736\(20\)31288-5](https://doi.org/10.1016/S0140-6736(20)31288-5)
2. Roviello F, Marrelli D, de Manzoni G, Morgagni P, Di Leo A, Saragoni L, De Stefano A, Italian Research Group for Gastric Cancer (2003) Prospective study of peritoneal recurrence after curative surgery for gastric cancer. *Br J Surg* 90:1113–1119. <https://doi.org/10.1002/bjs.4164>
3. D'Angelica M, Gonen M, Brennan MF, Turnbull AD, Bains M, Karpeh MS (2004) Patterns of initial recurrence in completely resected gastric adenocarcinoma. *Ann Surg* 240:808–816. <https://doi.org/10.1097/01.sla.0000143245.28656.15>
4. Sano T, Coit DG, Kim HH, Roviello F, Kassab P, Wittekind C, Yamamoto Y, Ohashi Y (2017) Proposal of a new stage grouping of gastric cancer for TNM classification: International Gastric Cancer

- Association staging project. *Gastric Cancer* 20:217–225. <https://doi.org/10.1007/s10120-016-0601-9>
5. Aurello P, Petrucciani N, Antolino L, Giulitti D, D'Angelo F, Ramacciato G (2017) Follow-up after curative resection for gastric cancer: Is it time to tailor it? *World J Gastroenterol* 23:3379–3387. <https://doi.org/10.3748/wjg.v23.i19.3379>
 6. Kanematsu K, Nakayama Y, Tanabe M, Morita J, Nagasawa S, Yamada T, Ogata T, Oshima T (2025) Hazard rates of recurrence for gastric cancer after curative resection: implications for postoperative surveillance. *Gastric Cancer* 28:275–282. <https://doi.org/10.1007/s10120-024-01576-5>
 7. André T, Tougeron D, Piessen G, de la Fouchardière C, Louvet C, Adenis A, Jary M, Tournigand C, Aparicio T, Desrame J, Lièvre A, Garcia-Larnicol ML, Pudlarz T, Cohen R, Memmi S, Vernerey D, Henriques J, Lefevre JH, Svrcek M (2023) Neoadjuvant nivolumab plus ipilimumab and adjuvant nivolumab in localized deficient mismatch repair/microsatellite instability-high gastric or esophagogastric junction adenocarcinoma: The GERCOR NEONIPIGA Phase II Study. *J Clin Oncol* 41:255–265. <https://doi.org/10.1200/JCO.22.00686>
 8. Moehler M, Dvorkin M, Boku N, Özgüroğlu M, Ryu MH, Muntean AS, Lonardi S, Nechaeva M, Bragagnoli AC, Coşkun HS, Cubillo Gracian A, Takano T, Wong R, Safran H, Vaccaro GM, Wainberg ZA, Silver MR, Xiong H, Hong J, Taieb J, Bang YJ (2021) Phase III trial of avelumab maintenance after first-line induction chemotherapy versus continuation of chemotherapy in patients with gastric cancers: results from JAVELIN Gastric 100. *J Clin Oncol* 39:966–977. <https://doi.org/10.1200/JCO.20.00892>
 9. Zhang XQ, Huang ZN, Wu J, Liu XD, Xie RZ, Wu YX, Zheng CY, Zheng CH, Li P, Xie JW, Wang JB, He QC, Qiu WW, Tang YH, Zhang HX, Zhou YB, Lin JX, Huang CM (2025) Machine learning prediction of early recurrence in gastric cancer: a nationwide real-world study. *Ann Surg Oncol* 32:2637–2650. <https://doi.org/10.1245/s10434-024-16701-y>
 10. Zhou C, Hu J, Wang Y, Ji MH, Tong J, Yang JJ, Xia H (2021) A machine learning-based predictor for the identification of the recurrence of patients with gastric cancer after operation. *Sci Rep* 11:1571. <https://doi.org/10.1038/s41598-021-81188-6>
 11. Wang HN, An JH, Wang FQ, Hu WQ, Zong L (2025) Predicting gastric cancer survival using machine learning: A systematic review. *World J Gastrointest Oncol* 17:103804. <https://doi.org/10.4251/wjgo.v17.i5.103804>
 12. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix AL, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P (2024) TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385:e078378. <https://doi.org/10.1136/bmj-2023-078378>

13. Japanese Gastric Cancer Association (2023) Japanese Gastric Cancer Treatment Guidelines 2021 (6th edition). *Gastric Cancer* 26:1–25. <https://doi.org/10.1007/s10120-022-01331-8>
14. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KG, Collins GS (2019) Minimum sample size for developing a multivariable prediction model: Part I—Continuous outcomes. *Stat Med* 38:1262–1275. <https://doi.org/10.1002/sim.7993>
15. Fine JP, Gray RJ (1999) A Proportional Hazards Model for the Subdistribution of a Competing Risk. *J Am Stat Assoc* 94:496–509. <https://doi.org/10.1080/01621459.1999.10474144>
16. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. *Ann Appl Stat* 2:841–860. <https://doi.org/10.1214/08-AOAS169>
17. Uno H, Cai T, Pencina MJ, D’Agostino RB, Wei LJ (2011) On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 30:1105–1117. <https://doi.org/10.1002/sim.4154>
18. Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak* 26:565–574. <https://doi.org/10.1177/0272989X06295361>
19. Pencina MJ, D’Agostino RB, Sr, D’Agostino RB Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 27:157–172. <https://doi.org/10.1002/sim.2929>
20. Lundberg SM, Lee SI (2017) A Unified Approach to Interpreting Model Predictions. *Adv Neural Inf Process Syst* 30:4765–4774. <https://doi.org/10.48550/arXiv.1705.07874>
21. Pang HY, Yan MH, Chen LH, Chen XF, Chen ZX, Zhang SR, Sun H (2022) Detection of asymptomatic recurrence following curative surgery improves survival in patients with gastric cancer: a systematic review and meta-analysis. *Front Oncol* 12:1011683. <https://doi.org/10.3389/fonc.2022.1011683>
22. Chidambaram S, Sounderajah V, Maynard N, Markar SR (2022) Evaluation of post-operative surveillance strategies for esophageal and gastric cancers: a systematic review and meta-analysis. *Dis Esophagus* 35:doac034. <https://doi.org/10.1093/dote/doac034>
23. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, Cording E, Tomson D, Dodd C, Rollnick S, Edwards A, Barry M (2012) Shared decision making: a model for clinical practice. *J Gen Intern Med* 27:1361–1367. <https://doi.org/10.1007/s11606-012-2077-6>
24. Feng B, Shi J, Huang L, Yang Z, Feng ST, Li J, Chen Q, Xue H, Chen X, Wan C, Hu Q, Cui E, Chen Y, Long W (2024) Robustly federated learning model for identifying high-risk patients with postoperative gastric cancer recurrence. *Nat Commun* 15:742. <https://doi.org/10.1038/s41467-024-44946-4>
25. Suzuki S, Takahashi A, Ishikawa T, Akazawa K, Katai H, Isobe Y, Miyashiro I, Ono H, Tanabe S, Fukagawa T, Muro K, Nunobe S, Kadowaki S, Suzuki H, Irino T, Usune S, Miyata H, Kakeji Y (2021) Surgically treated gastric cancer in Japan: 2011 annual report of the national clinical database gastric cancer registry. *Gastric Cancer* 24:545–566. <https://doi.org/10.1007/s10120-021-01178-5>
26. Steyerberg EW, Vergouwe Y (2014) Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 35:1925–1931.

<https://doi.org/10.1093/eurheartj/ehu207>

27. Wang S, Chai N, Xu J, Yu P, Huang L, Wang Q, Zhao Z, Yang B, Wei J, Wang X, Ji G, Zheng M (2025) Prediction of a panel of programmed cell death protein-1 (PD-1) inhibitor-sensitive biomarkers using multiphase computed tomography imaging textural features: retrospective cohort analysis. *JMIR Cancer* 11:e67379. <https://doi.org/10.2196/67379>

Tables

Table 1. Patient and tumor characteristics by cohort (pStage IIA–IIIC, R0 gastrectomy, 2008–2020)

Variable	Development 2008–2015 (n = 330)	Temporal validation 2016–2020 (n = 156)	p-value	Test
Events (recurrence), n (%)	102 (30.9%)	51 (32.7%)	—	—
Competing events (death w/o recurrence)	63	19	—	—
Age (years), median (IQR)	71.0 (64.0–77.8)	72.0 (65.0–78.0)	0.295	U
Male sex, n (%)	238 (72.1%)	110 (70.5%)	0.682	χ^2
Albumin (g/dL)	3.8 (3.4–4.2)	3.8 (3.6–4.1)	0.346	U
CRP (mg/dL)	0.2 (0.1–0.5)	0.1 (0.1–0.3)	0.017*	U
CEA (ng/mL)	2.2 (1.4–4.3)	2.4 (1.6–4.9)	0.149	U
CA19-9 (U/mL)	9.8 (5.5–20.9)	10.4 (6.1–25.4)	0.264	U
pStage, n (%)				
IIA	99 (30.0%)	38 (24.4%)	—	—
IIB	81 (24.5%)	25 (16.0%)	—	—
IIIA	61 (18.5%)	47 (30.1%)	—	—
IIIB	44 (13.3%)	22 (14.1%)	—	—
IIIC	45 (13.6%)	24 (15.4%)	—	—
Positive LN count	2.0 (1.0–6.0)	2.0 (0.0–6.0)	0.367	U
Tumor size (mm)	52.8 (37.2–70.0)	55.0 (40.0–70.0)	0.943	U
LVI grade (0–3)	2.0 (1.0–3.0)	1.0 (0.0–3.0)	<0.001*	U
S-1 adjuvant, n (%)	185 (56.1%)	86 (55.1%)	0.847	χ^2
Follow-up (years) [censored]	7.3 (3.8–9.7)	5.0 (3.3–5.8)	<0.001*	U

Values are median (interquartile range) unless otherwise stated. IQR, interquartile range; LN, lymph node; CRP, C-reactive protein; CEA, carcinoembryonic antigen; LVI, lymphovascular invasion. *p < 0.05. CRP and LVI grade differences reflect temporal trends in pathological reporting and surgical technique; the follow-up difference is a structural consequence of the chronological split. U = Mann–Whitney U test; χ^2 = chi-squared test.

Table 2. Discriminative performance of RSF versus pStage-only Cox proportional hazards model

Metric	RSF Development	RSF Validation (95% CI)	Cox (pStage) Development	Cox (pStage) Validation (95% CI)
Harrell's C-index	0.801	0.729 (0.653–0.809)	0.690	0.687 (0.616–0.765)
td-AUC 3-year	—	0.764 (0.673–0.851)	—	0.711 (0.619–0.802)
td-AUC 5-year	—	0.777 (0.688–0.863)	—	0.731 (0.637–0.823)
NRI (continuous)	—	0.065 (–0.126 to 0.242)	—	Reference
IDI	—	0.017 (–0.005 to 0.039)	—	Reference

95% CIs estimated from 1,000 bootstrap resamples of the validation cohort. RSF, random survival forest; td-AUC, time-dependent area under the ROC curve; NRI, net reclassification improvement; IDI, integrated discrimination improvement; CI, confidence interval. Note: all 95% CIs for C-index and AUC overlap between models, reflecting limited statistical power of the validation cohort (n = 156).

Table 3. Intra-stage heterogeneity of predicted 3-year cumulative recurrence probability (combined cohort, n = 486)

Stage group	N	Events n (%)	Median ci3	IQR (pp)	CV	Low-risk n / events	High-risk n / events	Log-rank p
Stage II (IIA+IIB)	243	40 (16.5%)	0.175	10.1	0.401	118 / 10 (8.5%)	125 / 30 (24.0%)	<0.001
Stage IIIA	108	40 (37.0%)	0.293	13.6	0.355	55 / 10 (18.2%)	53 / 30 (56.6%)	<0.001
Stage IIIB	66	25 (37.9%)	0.337	14.7	0.268	36 / 7 (19.4%)	30 / 18 (60.0%)	<0.001
Stage IIIC	69	48 (69.6%)	0.549	15.7	0.180	40 / 24 (60.0%)	29 / 24 (82.8%)	0.678

ci3, predicted 3-year cumulative recurrence probability; IQR, interquartile range (in percentage points, pp); CV, coefficient of variation. Low/high-risk groups defined by development-cohort stage-specific median ci3. Log-rank p-values from combined development + validation cohort.

Figures

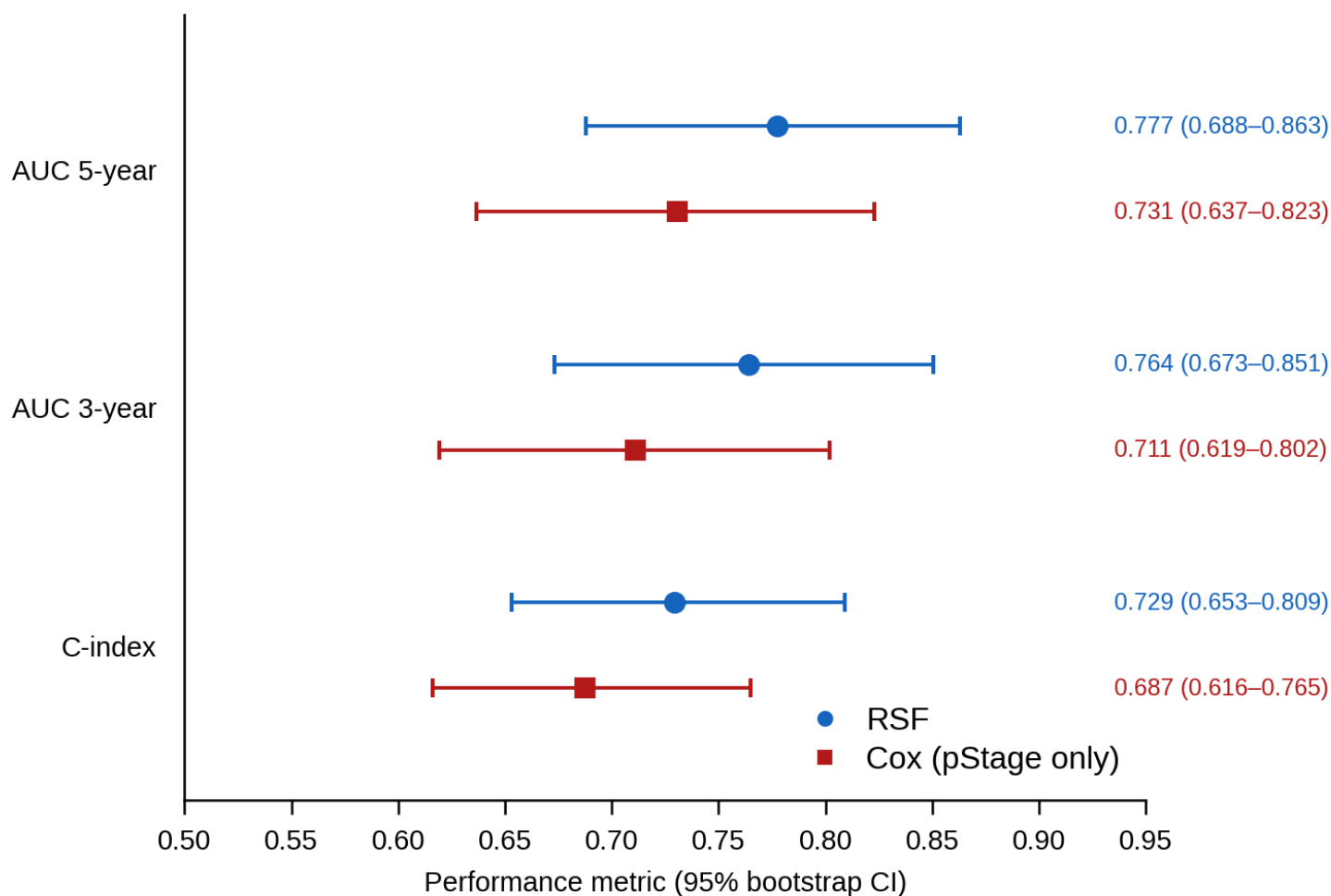


Figure 1

Discriminative performance comparison: Harrell’s C-index and time-dependent AUC at 3 and 5 years for RSF versus pStage-only Cox in the temporal validation cohort (n = 156). Points indicate point estimates; horizontal bars indicate 95% bootstrap confidence intervals (1,000 iterations). AUC, area under the receiver-operating-characteristic curve; RSF, Random Survival Forest

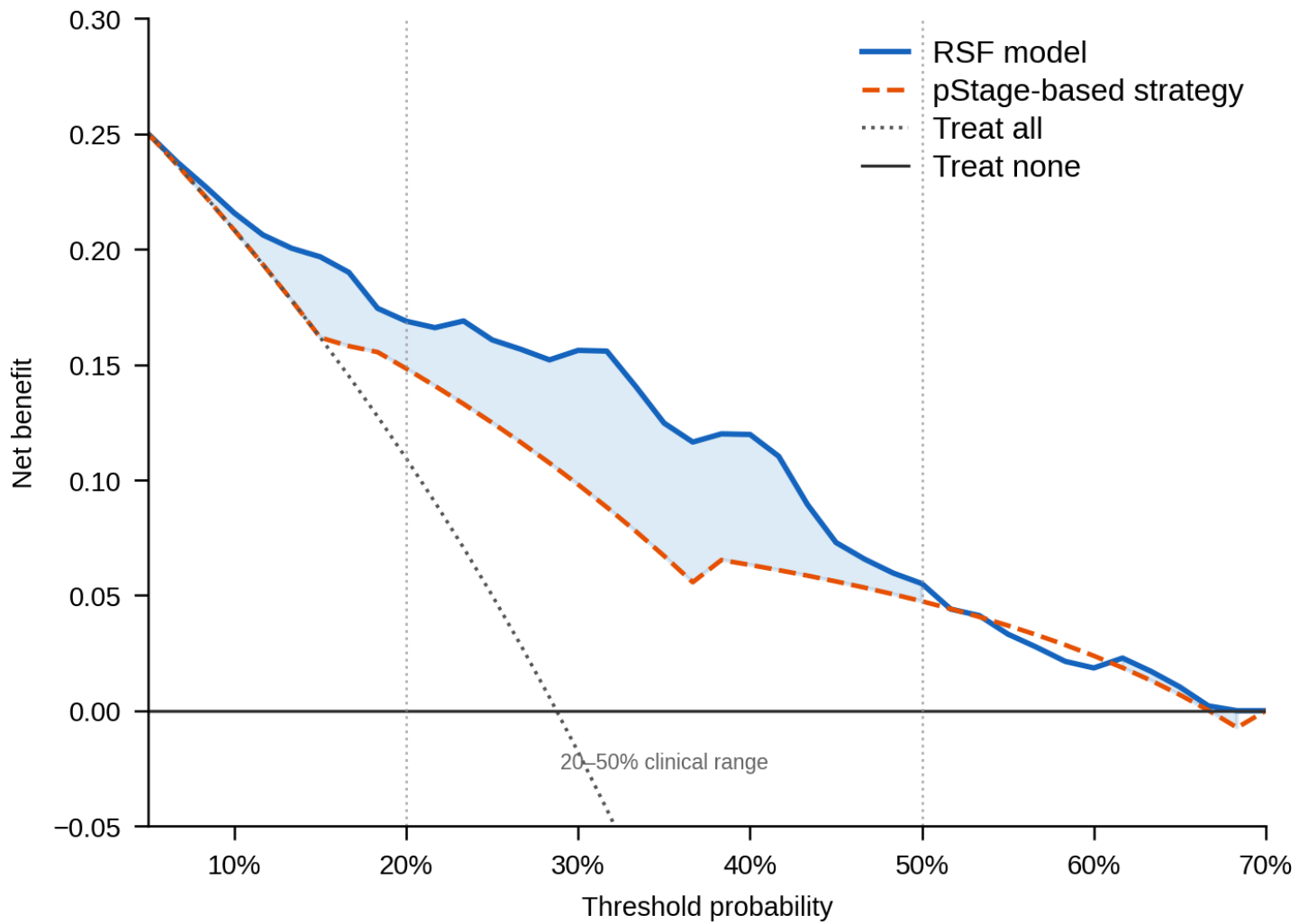


Figure 3

Decision curve analysis at the 3-year recurrence horizon (combined cohort, n = 486). Net benefit for the RSF model, pStage-based strategy (using stage-group observed recurrence rates), treat-all, and treat-none reference lines. The shaded area indicates the RSF net-benefit advantage over pStage at the 20%–50% threshold range. NB, net benefit. RSF, Random Survival Forest

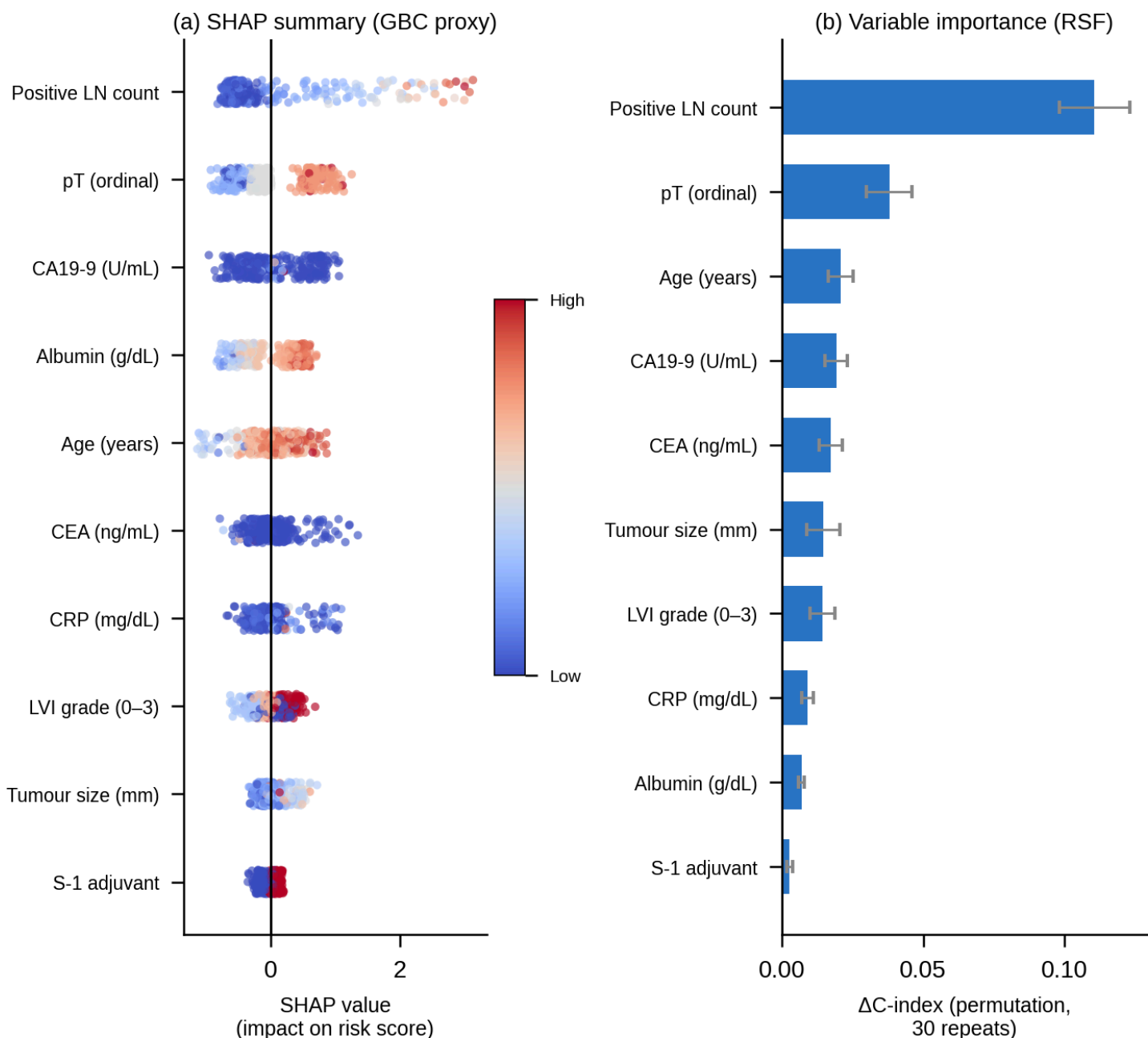


Figure 5

Model interpretability. Left: SHAP beeswarm plot showing directional impact of each predictor on the predicted recurrence risk score (gradient-boosting classifier proxy; development cohort). Each point represents one patient; color indicates feature value (red = high, blue = low). Right: permutation importance (decrease in C-index; $n = 30$ permutation repeats; RSF development cohort). Variables are ranked by importance magnitude. RSF, Random Survival Forest; LVI, lymphovascular invasion; LN, lymph node; CEA, carcinoembryonic antigen