

Comparing Random Survival Forests and Cox Regression for Non-Responders to Neoadjuvant Chemotherapy in Breast Cancer Patients: A Multicenter Retrospective Cohort Study

Yudi Jin, Min Zhao, Tong Su, Yanjia Fan, Zubin Ouyang, Fajin Lv

Submitted to: Journal of Medical Internet Research
on: December 10, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
Supplementary Files.....	28
Figures	29
Figure 1.....	30
Figure 2.....	31
Multimedia Appendixes	32
Multimedia Appendix 1.....	33
Multimedia Appendix 2.....	33
Multimedia Appendix 3.....	33
Multimedia Appendix 4.....	33
Multimedia Appendix 5.....	33
Multimedia Appendix 6.....	33
Multimedia Appendix 7.....	33
Multimedia Appendix 8.....	33
Multimedia Appendix 9.....	33
Multimedia Appendix 10.....	33

Comparing Random Survival Forests and Cox Regression for Non-Responders to Neoadjuvant Chemotherapy in Breast Cancer Patients: A Multicenter Retrospective Cohort Study

Yudi Jin¹ MM; Min Zhao¹ MM; Tong Su¹ MM; Yanjia Fan² BMed; Zubin Ouyang¹ Prof Dr; Fajin Lv¹ Prof Dr

¹ Department of Radiology The First Affiliated Hospital of Chongqing Medical University Chongqing CN

² Department of Breast and Thyroid Surgery The First Affiliated Hospital of Chongqing Medical University Chongqing CN

Corresponding Author:

Fajin Lv Prof Dr

Department of Radiology
The First Affiliated Hospital of Chongqing Medical University
No 1, Youyi Road
Yuzhong District
Chongqing
CN

Abstract

Background: Breast cancer is one of the most common malignancies among women globally. Patients do not achieve a pathological complete response (pCR) post neoadjuvant chemotherapy (NAC) are associated with poorer survival outcomes.

Objective: This study aimed to develop and validate a Random Survival Forest (RSF) model to predict survival risk in breast cancer patients who do not achieve pCR after NAC.

Methods: We reviewed patients diagnosed with breast cancer at the First Affiliated Hospital of Chongqing Medical University from January 2019 to January 2023, focusing on those who underwent NAC. After treatment, patients who were evaluated for non-pCR were included in the study. We developed an RSF model and a Cox regression model. The models were validated using external datasets from Duke University and the SEER database.

Results: A total of 306 patients were included in the internal dataset, with a median disease-free survival (DFS) of 25.9 months. The RSF model demonstrated superior predictive performance compared to the Cox regression model, with AUC values of 0.811, 0.834, and 0.810 at 1-year, 3-year, and 5-year, respectively. The C-index was 0.803 (95%CI, 0.747-0.859). Validation in the Duke dataset yielded AUC values of 0.912, 0.803, and 0.776 at 1-year, 3-year, and 5-year, respectively. While in the SEER dataset showed AUC values of 0.823, 0.756, and 0.731 at 1-year, 3-year, and 5-year, respectively. The RSF model effectively stratified patients into high-risk and low-risk groups, with significant differences in survival outcomes.

Conclusions: The RSF model, based solely on clinicopathological variables, provides a promising tool for identifying high-risk breast cancer patients post-NAC. This approach may facilitate personalized treatment strategies and improve patient management in clinical practice.

(JMIR Preprints 10/12/2024:69864)

DOI: <https://doi.org/10.2196/preprints.69864>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in the JMIR Publications Preprint service, I will be able to make my accepted manuscript PDF available to anyone at any time.

No. Please do not make my accepted manuscript PDF available to anyone.



Original Manuscript

Title page

Title: Comparing Random Survival Forests and Cox Regression for Non-Responders to Neoadjuvant Chemotherapy in Breast Cancer Patients: A Multicenter Retrospective Cohort Study

Author: Yudi Jin¹, Min Zhao¹, Tong Su¹, Yanjia Fan², Zubin Ouyang¹, Fajin Lv^{1#}

Authors' affiliation

1. Department of Radiology, the First Affiliated Hospital of Chongqing Medical University, Chongqing, China
2. Department of Breast and Thyroid Surgery, the First Affiliated Hospital of Chongqing Medical University, Chongqing, China

#Corresponding author:

Fajin Lv, Department of Radiology, the First Affiliated Hospital of Chongqing Medical University.

No.1, Youyi Road, Yuzhong district, Chongqing 400016, China.

Fax: 0086-023-68811487

Phone: +86 02389012228

Email: fajinlv@163.com

Abstract

Background: Breast cancer is one of the most common malignancies among women globally.

Patients do not achieve a pathological complete response (pCR) or a clinical complete response (cCR) post neoadjuvant chemotherapy (NAC) typically have a worse prognosis compared to those who do achieve these responses.

Objective: This study aims to develop and validate a Random Survival Forest (RSF) model to predict survival risk in breast cancer patients who do not achieve pCR or cCR after NAC.

Methods: We analyzed patients with non-pCR/cCR post NAC treated at the First Affiliated Hospital of Chongqing Medical University (2019.1-2023.1), with external validation in Duke University and SEER cohorts. RSF and Cox regression models were compared using time-dependent AUC, C-index, and risk stratification.

Results: The study cohort included 306 breast cancer patients with most aged 40-60 years (204/306, 66.7%), predominantly invasive ductal carcinoma (290/306, 94.8%), and receptor profiles of ER+ (59.5%, 182/306), PR- (58.5%, 179/306), and HER2 positive (30.7%, 94/306). Most presented with T2 (185/306, 60.5%), N1 (142/306, 46.4%), and M0 (295/306, 96.4%) staging, with 17.6% (54/306) experiencing disease progression during median 25.9-month follow-up (IQR:17.2-36.3). External validation using Duke (n=94) and SEER (n=2,760) cohorts confirmed consistent patterns: 63% (59/94) vs. 53.6% (1,480/2,760) aged 40-60, HER2 positive rates 28% (26/94) vs 33.9% (935/2,760), and invasive ductal carcinoma prevalence 95% (89/94) vs 90.8% (2,506/2,760). In the internal cohort, the RSF achieved significantly higher time-dependent AUC values compared to Cox regression at 1-year (0.811 vs. 0.763), 3-year (0.834 vs. 0.783), and 5-year intervals (0.810 vs. 0.771), with an overall C-index of 0.803 (0.747-0.859) versus 0.736 (0.673-0.799). External validation confirmed robust generalizability: the Duke cohort showed 1/3/5-year AUCs of 0.912/0.803/0.776, while the SEER cohort maintained consistent performance with AUCs of 0.771/0.729/0.702. Risk stratification using RSF identified 25.8% (79/306) as high-risk patients, and significantly reduced survival time ($P<.001$). Notably, the RSF model maintained improved net benefit across decision thresholds in DCA analysis. Similar results were observed in external studies. Furthermore, when analyzing across different molecular subtypes, the RSF model achieved promising performance across all datasets. Based on the predicted scores from the RSF model, populations were stratified into high-risk and low-risk groups, with notably poorer survival outcomes observed in the high-risk group compared to the low-risk group.

Conclusions: The RSF model, based solely on clinicopathological variables, provides a promising

tool for identifying high-risk breast cancer patients post-NAC. This approach may facilitate personalized treatment strategies and improve patient management in clinical practice.

Key words: Breast cancer, neoadjuvant chemotherapy, Pathological complete response, survival risk, random survival forest

Introduction

Breast cancer remains one of the most prevalent malignancies among women worldwide, accounting for a significant proportion of cancer-related morbidity and mortality [1, 2]. Despite advancements in treatment modalities, including neoadjuvant chemotherapy (NAC), a substantial number of patients do not achieve a complete response (CR). These patients usually have a worse prognosis compared to those who do achieve CR (Multimedia Appendix 1) [3]. This underscores the necessity for developing effective prognostic tools to identify patients at high risk for adverse outcomes. However, there are limited studies focused on developing predictive models for breast cancer patients who do not attain a CR following NAC.

Currently, machine learning has emerged as a powerful tool for survival analysis, providing significant advantages over traditional statistical methods [4-6]. Traditional methods used Cox regression to predict the prognosis of cancer patients. However, it is important to note that if the proportional hazards assumption is violated, the results of the Cox regression model may be biased. Additionally, Cox regression may struggle to capture complex, non-linear relationships between independent variables and survival time [7]. Besides, previous studies have confirmed that some other models have outperformed the Cox regression model. Especially, lots of researches have shown that the random survival forest (RSF) model is capable of managing high-dimensional data, improving the accuracy of survival predictions, and supporting personalized treatment strategies, which typically yield the best performance [8-13].

This study aims to develop and validate a RSF model to predict survival risk in breast cancer patients who fail to achieve CR after NAC, comparing its performance with traditional Cox regression. We hypothesized that the RSF model would offer a reliable tool for clinicians to stratify patients based on predicted survival risks, ultimately supporting more informed treatment decisions and enhancing patient management outcomes.

Methods

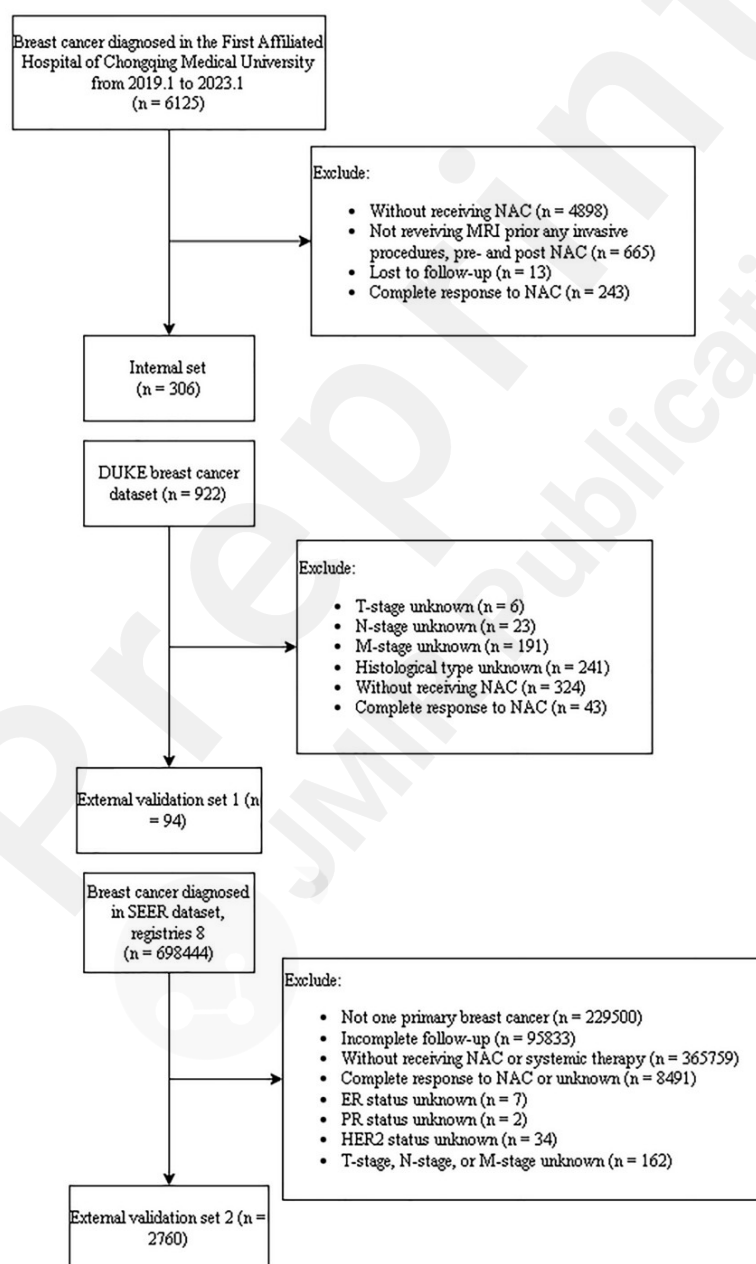
Recruitment and study designed

Patients diagnosed with breast cancer at the First Affiliated Hospital of Chongqing Medical University from January 2019 to January 2023 were comprehensively reviewed. We selected those who underwent NAC for following analysis. After administering 4 to 8 cycles of NAC, clinicians, radiologists, and pathologists evaluated the treatment response. A clinical complete response (cCR) was defined as the complete disappearance of all tumor lesions, as confirmed by imaging examinations, lasting for a minimum of four weeks [14]. A pathological complete response (pCR) was defined as the absence of any residual invasive tumor in both the breast and axillary lymph nodes (ypT0ypN0) [15]. The patients without achieving cCR or pCR were enrolled in this study. The adjuvant treatment process was determined according to the guidelines outlined by the Chinese Society of Clinical Oncology (CSCO) and the National Comprehensive Cancer Network (NCCN) [16, 17]. The pathological assessment was according to the American Society of Clinical Oncology guidelines (Multimedia Appendix 2) [18-20].

Initially, we established the RSF model using the entire patient cohort. Subsequently, we examined the relationship between clinicopathological variables and survival outcomes using univariate and multivariate Cox regression analyses. We then constructed a Cox regression model using the variables selected based on $P < .05$ from the multivariate analysis. Furthermore, the Duke

University Breast Cancer dataset [21] and data from the Surveillance, Epidemiology, and End Results (SEER) [22] database were utilized as validation cohorts. The study flowchart and patient inclusion process are illustrated in Figure 1. This study was approved by the ethics committee of the First Affiliated Hospital of Chongqing Medical University (ID: No. 2020–59).

Figure 1. Flowchart of this study.



Follow-up

All patients included in this study were interviewed through either outpatient visits or telephone consultations. The follow-up period extended from discharge until April 30, 2024. Disease-free survival (DFS) was selected as the primary metric for assessing patient survival duration. DFS was defined as the time from surgery to the occurrence of breast cancer recurrence, the diagnosis of a new primary cancer, or death from any cause, whichever comes first.

Model Development (RSF and Cox Regression)

The RSF model was constructed using the randomForestSRC package in R. A starting model was trained on the entire cohort. The following hyperparameters were configured: number of trees (ntree), node size, mtry, and independent variables. The model underwent two key adjustments: (1) hyperparameter tuning via cross-validation to optimize performance, and (2) permutation-based variable importance scoring to identify significant predictors. Following optimization, the final model was built. Additionally, variables selected based on $P < .05$ from the multivariate Cox regression analysis were incorporated into the Cox regression model.

External validation set

In this study, external validation datasets were obtained from two sources: the DUKE breast cancer dataset and the SEER dataset. For the DUKE breast cancer dataset, the inclusion criteria required that patients had received NAC. The exclusion criteria included: 1) any missing or unknown information; and 2) patients who achieved a complete response to NAC. In the SEER dataset, the inclusion criteria were: 1) patients registered in registry 8; and 2) patients diagnosed with a malignant breast tumor. The exclusion criteria for this dataset included: 1) patients with more than

one primary tumor; 2) any missing information or records marked as “Unknown”; 3) patients who did not receive NAC; and 4) patients who achieved a complete response to NAC. In the Duke dataset, survival time was evaluated as recurrence-free survival (RFS), whereas the SEER dataset used overall survival (OS) as its measure.

The performance of the RSF model and the Cox regression model was assessed using the two validation cohorts. Key metrics, including concordance index (C-index), 95% confidence interval (CI), and integrated Brier score, were calculated to evaluate the model's predictive accuracy and calibration. Survival curves were generated using the Kaplan-Meier (K-M) method, stratified by risk groups identified by the RSF and the Cox regression model. Additionally, variable importance plots were created to illustrate the contribution of each variable to the model. Finally, the model was validated in the DUKE dataset, as well as in the SEER dataset.

Validate model performance in different molecular subtypes

We further employed the RSF model to assess the performance of various molecular subtypes across different datasets, with the results illustrated through time-dependent ROC curves. Patients with different subtypes in these datasets were categorized into high-risk and low-risk groups based on the model's predictions. The survival differences between these two groups were evaluated using the K-M curves.

Statistical analysis

We conducted the statistical analysis using RStudio (R version 4.4.2). All independent variables were categorized and presented as absolute counts. To compare categorical data, we employed the chi-square test. The survival time was continuous variable and was presented as median and interquartile range (IQR). The Cox proportional hazards model was utilized to identify prognostic

factors through both univariate and multivariate analyses. For the development of the RSF model, we used the “randomForestSRC” package, while the “cph” function was employed to construct the Cox regression model. The receiver operating characteristic curve (ROC) and the decision curve analysis (DCA) were used to display the performance of the models. Survival analysis was performed using K-M curves, with differences between groups evaluated using the Log-rank test. In this study, only $P < .05$ was considered indicative of a statistically significant difference.

Results

Demographic and Clinical Characteristics

The internal cohort comprised 306 breast cancer patients with the following characteristics: a majority (204/306, 66.7%) were aged 40-60 years, 94.8% (290/306) had invasive ductal carcinoma, and receptor status analysis revealed 59.5% (182/306) estrogen receptor (ER)-positive, 58.5% (179/306) progesterone receptor (PR)-negative, and 30.7% (94/306) human epidermal growth factor receptor 2 (HER2)-positive cases. The most prevalent staging was T2 (185/306, 60.5%), N1 (142/306, 46.4%), and M0 (295/306, 96.4%). Over a median follow-up of 25.9 months (IQR:17.2-36.3), 17.6% (54/306) experienced disease-related events. External validation cohorts demonstrated comparable patterns, with the Duke dataset (n=94) showing similar age distribution (59 out of 94, 63% aged 40-60), HER2-positivity rate (26 out of 94, 28%), and hormone receptor (HR) profiles, while the SEER cohort (n=2,760) maintained consistent invasive ductal carcinoma predominance (90.8%, 2,506/2,760) and staging trends. Full demographic details are presented in Table 1.

Table 1. Baseline characteristic of the internal set, DUKE dataset, and the SEER dataset, respectively.

	Internal set (N=306)	DUKE dataset (N=94)	SEER dataset (N=2760)
Age (years), n(%)			

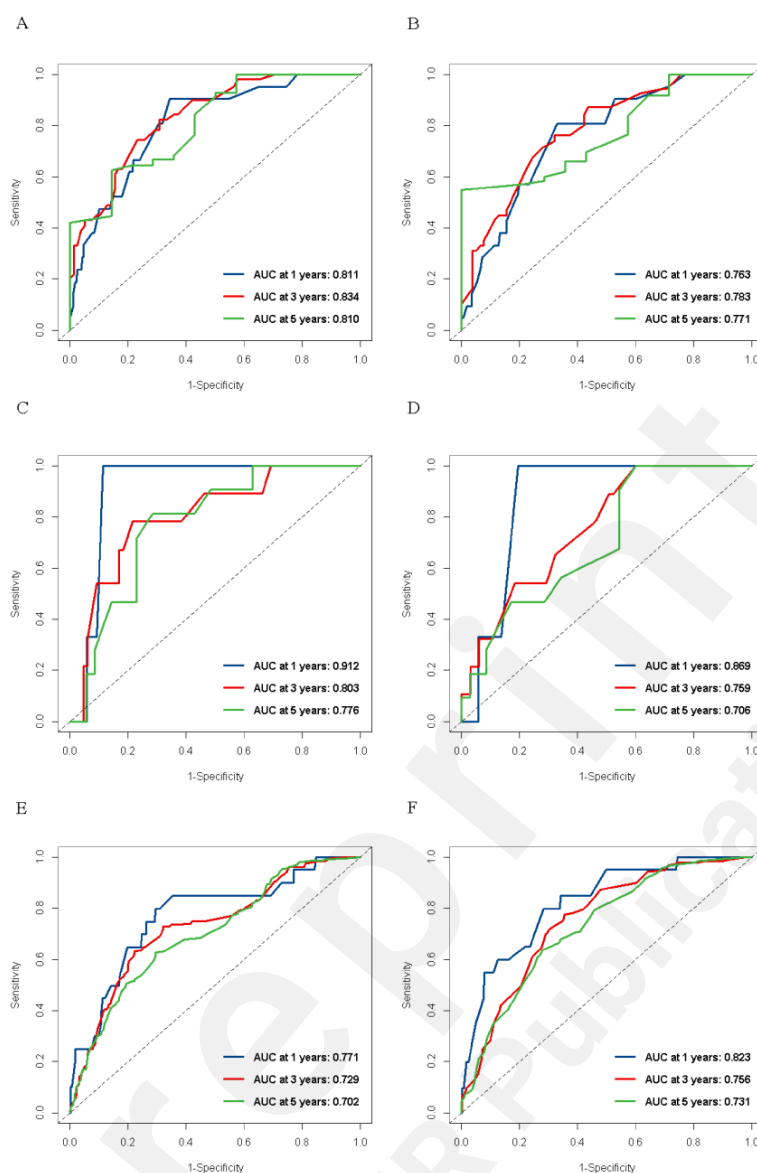
		Internal set (N=306)	DUKE dataset (N=94)	SEER dataset (N=2760)
	> 60	42 (13.7%)	14 (14.9%)	732 (26.5%)
	≤ 40	60 (19.6%)	21 (22.3%)	548 (19.9%)
	40-60	204 (66.7%)	59 (62.8%)	1480 (53.6%)
Histological type, n(%)				
	Invasive ductal carcinoma	290 (94.8%)	89 (94.7%)	2506 (90.8%)
	Other	16 (5.2%)	5 (5.3%)	254 (9.2%)
ER, n(%)				
	Negative	124 (40.5%)	40 (42.6%)	686 (24.9%)
	Positive	182 (59.5%)	54 (57.4%)	2074 (75.1%)
PR, n(%)				
	Negative	179 (58.5%)	52 (55.3%)	1093 (39.6%)
	Positive	127 (41.5%)	42 (44.7%)	1667 (60.4%)
HER2, n(%)				
	Negative	212 (69.3%)	68 (72.3%)	1825 (66.1%)
	Positive	94 (30.7%)	26 (27.7%)	935 (33.9%)
Molecular subtype, n(%)				
	HR-/HER2-	73 (23.9%)	29 (30.9%)	466 (16.9%)
	HR-/HER2+	50 (16.3%)	11 (11.7%)	182 (6.6%)
	HR+/HER2-	139 (45.4%)	39 (41.5%)	1359 (49.2%)
	HR+/HER2+	44 (14.4%)	15 (16%)	753 (27.3%)
T, n(%)				
	1	36 (11.8%)	16 (17%)	449 (16.3%)
	2	185 (60.5%)	58 (61.7%)	1458 (52.8%)
	3	51 (16.7%)	17 (18.1%)	578 (20.9%)
	4	34 (11.1%)	3 (3.2%)	275 (10%)
N, n(%)				
	0	43 (14.1%)	41 (43.6%)	929 (33.7%)
	1	142 (46.4%)	40 (42.6%)	1232 (44.6%)
	2	61 (19.9%)	7 (7.4%)	343 (12.4%)
	3	60 (19.6%)	6 (6.4%)	256 (9.3%)
M, n(%)				
	0	295 (96.4%)	93 (98.9%)	2677 (97%)
	1	11 (3.6%)	1 (1.1%)	83 (3%)
Status, n(%)				
	0	252 (82.4%)	83 (88.3%)	2364 (85.7%)
	1	54 (17.6%)	11 (11.7%)	396 (14.3%)
Survival time, median (IQR)		25.9 (17.2 to 36.3)	50.2 (29.3 to 66.2)	35.0 (15.5 to 64.0)

Construction of the RSF model

Initially, we integrated all independent variables into the RSF model, setting the number of trees (ntree) to 2000. Our findings indicated that the model stabilized when ntree reached 1000.

Additionally, the variable importance analysis revealed that the variable T-stage has significant negative effects on the model's performance (Multimedia Appendix 3). Consequently, we adjusted ntree to 1000 and included Age, Histological type, ER, PR, HER2, N-stage, and M-stage as independent variables to reconstruct the RSF model. Through hyperparameter tuning, we determined that the optimal model performance and generalization ability were achieved with a nodesize of 10 and an mtry of 2 (Multimedia Appendix 4). The ROC curves showed that in the training set the AUC of the model at 1-year, 3-year, and 5-year was 0.811, 0.834, and 0.810, respectively (Figure 2A). The C-index was 0.803 (95%CI, 0.747-0.859). The brier score was showed in Multimedia Appendix 5.

Figure 2. The ROC curves of the RSF model in the internal set (A), DUKE dataset (C), SEER dataset (E). And the ROC curves of the Cox regression model in the internal set (B), DUKE dataset (D), SEER dataset (F).



Construction of the Cox regression model

The results of the multivariate Cox regression analysis indicated that Age, PR, HER2, N-stage, and M-stage were significantly associated with survival risks (Table 2). These variables were incorporated into the Cox regression model, which yielded an AUC of 0.763, 0.783, and 0.771 at the 1-year, 3-year, and 5-year marks, respectively (Figure 2B). The C-index was calculated to be 0.736 (95% CI, 0.673-0.799). Additionally, the Brier score, which was relatively lower than that of the RSF model at each time point, was presented in Multimedia Appendix 5.

Table 2. Cox regression of the clinicopathological variables.

			Hazard Ratio (univariable)	Hazard Ratio (multivariable)
Age, n(%)				
	> 60	42 (13.7%)		
	≤ 40	60 (19.6%)	0.36 (0.15-0.87, $P = .02$)	0.45 (0.18-1.12, $P = .09$)
	40-60	204 (66.7%)	0.40 (0.20-0.77, $P = .007$)	0.48 (0.24-0.97, $P = .04$)
Histology, n(%)				
	Invasive ductal carcinoma	290 (94.8%)		
	Other	16 (5.2%)	0.31 (0.04-2.22, $P = .24$)	0.30 (0.04-2.25, $P = .24$)
ER, n(%)				
	Negative	124 (40.5%)		
	Positive	182 (59.5%)	0.74 (0.44-1.27, $P = .28$)	1.05 (0.53-2.09, $P = .88$)
PR, n(%)				
	Negative	179 (58.5%)		
	Positive	127 (41.5%)	0.54 (0.30-0.98, $P = .04$)	0.45 (0.21-0.96, $P = .04$)
HER2, n(%)				
	Negative	212 (69.3%)		
	Positive	94 (30.7%)	0.54 (0.29-1.02, $P = .06$)	0.45 (0.23-0.87, $P = .02$)
T, n(%)				
	1	36 (11.8%)		
	2	185 (60.5%)	1.42 (0.55-3.63, $P = .47$)	1.47 (0.57-3.82, $P = .42$)
	3	51 (16.7%)	1.35 (0.44-4.14, $P = .60$)	1.26 (0.40-3.93, $P = .69$)
	4	34 (11.1%)	1.73 (0.56-5.29, $P = .34$)	1.58 (0.49-5.08, $P = .45$)
N, n(%)				
	0	43 (14.1%)		
	1	142 (46.4%)	1.84 (0.63-5.38, $P = .26$)	1.65 (0.55-4.93, $P = .37$)
	2	61 (19.9%)	2.75 (0.89-8.55, $P = .08$)	2.06 (0.65-6.58, $P = .22$)
	3	60 (19.6%)	4.34 (1.45-12.96, $P = .009$)	3.36 (1.10-10.24, $P = .03$)
M, n(%)				
	0	295 (96.4%)		
	1	11 (3.6%)	3.30 (1.31-8.31, $P = .01$)	2.74 (1.03-7.25, $P = .04$)

Model validation

A total of 94 patients from the DUKE dataset and 2760 patients from the SEER dataset were included in the analysis, the selection process was displayed in Figure 1. For the RSF model, the AUC in the DUKE dataset was 0.912 at 1 year, 0.803 at 3 years, and 0.776 at 5 years, as illustrated in Figure 2C. In the SEER dataset, the AUC values were 0.771 at 1 year, 0.729 at 3 years, and 0.701 at 5 years, as shown in Figure 2E. For the Cox regression model, the AUC in the DUKE dataset was

0.869 at 1 year, 0.759 at 3 years, and 0.706 at 5 years, with the corresponding ROC curves presented in Figure 2D. In the SEER dataset, the AUC values were 0.823 at 1 year, 0.756 at 3 years, and 0.731 at 5 years, as depicted in Figure 2F.

Comparison of the RSF model and the Cox regression model

Multimedia Appendix 6 presented the DCA curves for the models at 1-year, 3-year, and 5-year intervals across all datasets. In both the training set and the DUKE dataset, patients derived greater benefits from the RSF model compared to the Cox regression model. Conversely, in the SEER dataset, the benefits for patients using both models were comparable. Furthermore, in the training set, the RSF model identified a predictive cut-off value of 8.70 to categorize patients into high-risk and low-risk groups (Multimedia Appendix 7). Survival analysis demonstrated that the prognosis for the high-risk group was significantly worse than that of the low-risk group (Multimedia Appendix 8). This cut-off value was also applied to classify patients from both the DUKE and SEER datasets into high-risk and low-risk groups. Survival analyses indicated that patients in the high-risk group had poorer prognoses compared to those in the low-risk group across both datasets (Multimedia Appendix 8). Similarly, the Cox regression model established a predictive cut-off value of 0.27 in the training set to differentiate between high-risk and low-risk groups (Multimedia Appendix 7). Survival analyses yielded results consistent with those obtained using the RSF model (Multimedia Appendix 8).

Performance of the RSF model among different molecular subtypes

We conducted a performance evaluation of the RSF model across various molecular subtypes. The ROC curves indicated that, within the internal dataset, the RSF model achieved an AUC of 1.000 for both the 1-year and 3-year survival rates, and 0.748 for the 5-year survival rate in the HR+/

HER2+ subtype. In the HR+/HER2- subtype, the AUC values were 0.872 for the 1-year, 0.699 for the 3-year, and 0.778 for the 5-year survival rates. For the HR-HER2+ subtype, the AUC values were 0.639 for the 1-year, 0.845 for the 3-year, and 0.698 for the 5-year survival rates. In the HR-HER2- subtype, the AUC values were 0.681 for the 1-year and 0.832 for the 3-year survival rates. Consistent trends were observed in the SEER and DUKE validation datasets (Multimedia Appendix 9). The K-M curves further demonstrated the model's ability to stratify patients into distinct high-risk and low-risk groups across all four molecular subtypes in both the internal and SEER datasets. In the DUKE dataset, while no statistically significant difference was observed in the HR-/HER2- subgroup, low-risk patients still exhibited higher RFS compared to high-risk patients (Multimedia Appendix 10).

Discussion

In this study, our key findings demonstrated that the RSF model outperformed Cox regression in predicting survival risk for non-responders post NAC, with validated generalizability across external cohorts. The RSF model also demonstrated consistent effectiveness when analyzing various molecular subgroups. The findings highlighted the potential of machine learning techniques, particularly RSF, in enhancing prognostic accuracy and guiding clinical decision-making in oncology.

Previous research has demonstrated that achieving a pCR following NAT is associated with significantly improved event-free survival and overall survival [3]. Consequently, numerous studies have concentrated on predicting tumor responses to NAC. For instance, Zhao et al. constructed machine learning models to predict pCR to NAC based on clinicopathological variables [23]. Similarly, Zhang et al. developed machine learning models that incorporated clinicopathological features, radiomic features, and pathomic features to forecast pCR following NAC [24-30]. Additionally, Sammut et al. created models utilizing multi-omics data [31, 32]. These studies

highlighted the substantial value of machine learning models in predicting patients' responses to NAC.

However, as previous research has confirmed that patients with non-pCR tend to have a worse prognosis and a higher risk of adverse events, it is important to note that there are few studies exploring risk stratification for those patients. In one of our earlier studies, we have tried to develop a random forest model to predict event occurrence among breast cancer patients with no response to NAC [33]. At the same time, we aimed to create a model that directly predicted patients' survival risk. The RSF algorithm integrates random forests with survival analysis, enabling the prediction of individual event probabilities and survival time. Compared to the traditional Cox regression, it offers several advantages: it is not limited by the proportional hazards assumption, can effectively handle high-dimensional data, and demonstrates strong generalization capabilities [11]. Liao et al. developed a prediction model using 10 machine learning algorithms across 101 combinations to forecast cancer related mortality in patients with gastric neuroendocrine neoplasms. They have finally found that the RSF model obtained the highest AUC value [9]. Similar results have been found in other studies [8, 10-13]. In our study, we also found that the RSF model outperformed the Cox regression model, a finding that was further validated using the DUKE dataset.

One of the strengths of our study is the robust validation of the RSF model using two independent validation cohort. The C-index demonstrated the model's predictive accuracy, suggesting that it can reliably stratify patients based on their survival risk. Additionally, the survival time metrics varied across datasets: the training set employed DFS, whereas the DUKE dataset was based on RFS, and the SEER dataset utilized OS. The DFS status encompassed disease recurrence, new primary diseases, and death. Consequently, the model demonstrated strong adaptability across datasets that utilize these endpoints as status indicators. Therefore, we suggested using DFS to construct time-to-event predictive models in future studies. Another issue pertains to the selection of independent variables. Initially, we have tried to develop the RSF model using variables selected

through multivariate analysis in the Cox regression; however, the results were not satisfactory. Subsequently, we employed the variable importance metrics from the RSF algorithm to identify the variables that were ultimately included in the model. Hence, we also suggested to utilize a variable selection method that aligns with the specific model being used.

To the best of our knowledge, our study demonstrated that a more complex algorithm could effectively predict the survival risk of breast cancer patients without CR post NAC, based solely on the clinicopathological variables. Nevertheless, there were limitations in our study. First, the sample of the training set was relative small, which might affect the variables selection and the parameters set. Second, the follow-up duration in our study was relatively short, with only a limited number of patients followed for more than 5 years. This might have constrained the model's ability to accurately predict long-term prognosis. Third, the retrospective nature of the data collection might introduce biases, and the findings should be interpreted with caution. Forth, we included only a limited subset of clinicopathological variables, and further exploration is needed to assess the potential inclusion of additional variables. Additionally, we noted that the performance of the two models was comparable in the SEER dataset, with the Cox regression model performing slightly better. This might be attributed to the limited training sample size, which could restrict the performance of the RSF model. Nevertheless, our findings confirm the potential of RSF model for predicting prognosis using clinicopathological variables in patients. And further prospective study is necessary to confirm its applicability. Moreover, the RSF model was particularly well-suited for handling higher-dimensional variables. Thus, future studies should also explore the integration of radiomics, pathomics, as well as the molecular data to enhance the predictive power of the model [34-36].

In conclusion, our study highlighted the feasibility and effectiveness of utilizing a RSF model based exclusively on clinicopathological variables to predict survival risk in breast cancer patients who did not achieve CR after NAC. This approach could enhance clinical practice by assisting oncologists in identifying high-risk patients who might benefit from more aggressive treatment

strategies or closer monitoring. As we continued to refine and validate this model, we anticipated that it would significantly contribute to the advancement of personalized medicine in breast cancer care.

Acknowledgements

We thank all the participants of the SEER database, as well as the DUKE-Breast-cancer-dataset for contributing their works in supporting our studies. This study was funded by the First-class Discipline Construction Project of Clinical Medicine in the First Clinical College of Chongqing Medical University (Grant no. CYYY-BSYJSKYCXXM202412). The funders did not participate in the study's design, data collection, management, analysis, interpretation, preparation, review of this paper, or the decision to publish.

Ethical approves and Consent to participate

This study was conducted in accordance with the principles of the Declaration of Helsinki and was approved by the ethics committee of the First Affiliated Hospital of Chongqing Medical University (ID: No. 2020–59). Informed consent was waived due to the retrospective nature of the study and the use of de-identified patient data.

Data Availability

The data of the First Affiliated Hospital of Chongqing Medical University are available and can be requested from the corresponding author. The data of the Duke-Breast-Cancer dataset could be searched from the following website: <https://www.cancerimagingarchive.net/collection/duke-breast-cancer-mri/>. And the data of the SEER dataset could be retrieved from SEERStat (<https://seer.cancer.gov/>).

Authors' contributions

YJ conceived the study concept; YJ and TS designed the study protocol; YF, MZ and YJ retrieved and selected the data; YJ and MZ completed the draft writing or manuscript revision; YJ and TS performed statistical analysis and interpretation; YJ and FL oversaw the integration of the entire study; FL and ZO edited

the manuscript and approved the final version. All authors approved the submission of the final manuscript.

Conflicts of Interest

None declared.

Abbreviations

cCR: clinical complete response

pCR: pathological complete response

CR: complete response

NAC: neoadjuvant chemotherapy

RSF: random survival forest

DFS: disease-free survival

RFS: recurrence-free survival

OS: overall survival

ER: estrogen receptor

PR: progesterone receptor

HR: hormone receptor

HER2: human epidermal growth factor receptor 2

SEER: Surveillance, Epidemiology, and End Results

IQR: interquartile range

ROC: receiver operating characteristic curve

AUC: area under the receiver operating characteristic curve

CI: confidence interval

K-M: Kaplan-Meier

DCA: decision curve analysis

FISH: fluorescence in situ hybridization

IHC: Immunohistochemistry

Multimedia Appendix 1

Survival analysis comparing patients who achieved pCR or cCR after NAC with those who did not achieve pCR or cCR, across four groups: (A) all patients, (B) HR+ patients, (C) HR-/HER2+ patients, and (D) HR-/HER2- patients, using the internal dataset.

Multimedia Appendix 2

Treatment protocols and pathological assessment of the patients in the internal dataset.

Multimedia Appendix 3

The relationship between the number of trees in the model and the out-of-bag (OOB) error rate (A).

The relative contribution of each predictor variable to the model's overall predictive accuracy (B).

Multimedia Appendix 4

The out-of-bag (OOB) error rate for node size and mtry selection.

Multimedia Appendix 5

The brier score of the RSF model and the Cox regression model, respectively.

Multimedia Appendix 6

The DCA curves presented for 1-year, 3-year, and 5-year intervals for the internal set (A-C), DUKE dataset (D-F), and SEER dataset (G-I), respectively.

Multimedia Appendix 7

The risk group setting by the RSF model (A) and the Cox regression model (B).

Multimedia Appendix 8

The RSF model stratified patients into high-risk and low-risk groups across the internal set (A), DUKE dataset (C), and SEER dataset (E). Similarly, the Cox regression model categorized patients into high-risk and low-risk groups in the internal set (B), DUKE dataset (D), and SEER dataset (F).

Multimedia Appendix 9

The ROC curves for the HR+/HER2+ (A), HR+/HER2- (B), HR-/HER2+ (C), and HR-/HER2- (D) subtypes were analyzed at 1-year, 3-year, and 5-year time points in the internal dataset, with the exception of the HR-/HER2- subtype, which was assessed only at 1-year and 3-year points. In the SEER dataset, the ROC curves for HR+/HER2+ (E), HR+/HER2- (F), HR-/HER2+ (G), and HR-/HER2- (H) were analyzed at 1-year, 3-year, and 5-year time points, except for the HR-/HER2+ subtype, which was evaluated only at 3-year and 5-year points. The DUKE dataset included ROC curves for HR+/HER2- (I) and HR-/HER2- (J) subtypes at all three time points: 1-year, 3-year, and 5-year. (Owing to the limited number of HR+/HER2+ and HR-/HER2+ subtypes in the DUKE dataset, ROC curves could not be plotted.)

Multimedia Appendix 10

The K-M curves of the RSF model stratified HR+/HER2+, HR+/HER2-, HR-/HER2+, and HR-/HER2- subtype patients into high-risk and low-risk groups within the internal (A-D, respectively) and SEER datasets (E-H, respectively). Similarly, the RSF model divided HR+/HER2- and HR-/HER2- subtype patients into risk groups within the DUKE dataset (I-J). (Owing to the limited number of HR+/HER2+ and HR-/HER2+ subtypes in the DUKE dataset, K-M curves could not be plotted.)

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 Cancer J Clin.* 2024. 74(3): 229-263.
2. Giaquinto AN, Sung H, Newman LA, et al. Breast cancer statistics 2024. *CA Cancer J Clin.* 2024. 74(6): 477-495.
3. Spring LM, Fell G, Arfe A, et al. Pathologic Complete Response after Neoadjuvant Chemotherapy and Impact on Breast Cancer Recurrence and Survival: A Comprehensive Meta-analysis. *Clin Cancer Res.* 2020. 26(12): 2838-2848.
4. Xiao J, Mo M, Wang Z, et al. The Application and Comparison of Machine Learning Models for the Prediction of Breast Cancer Prognosis: Retrospective Cohort Study. *JMIR Med Inform.* 2022; 10(2): e33440. Medline: 35179504 DOI: 10.2196/33440
5. Janbain A, Farolfi A, Guenegou-Arnoux A, et al. A Machine Learning Approach for Predicting Biochemical Outcome After PSMA-PET-Guided Salvage Radiotherapy in Recurrent Prostate Cancer After Radical Prostatectomy: Retrospective Study. *JMIR Cancer.* 2024; 10e60323. Medline: 39303279 DOI: 10.2196/60323
6. Zheng Y, Zhao A, Yang Y, et al. Real-World Survival Comparisons Between Radiotherapy and Surgery

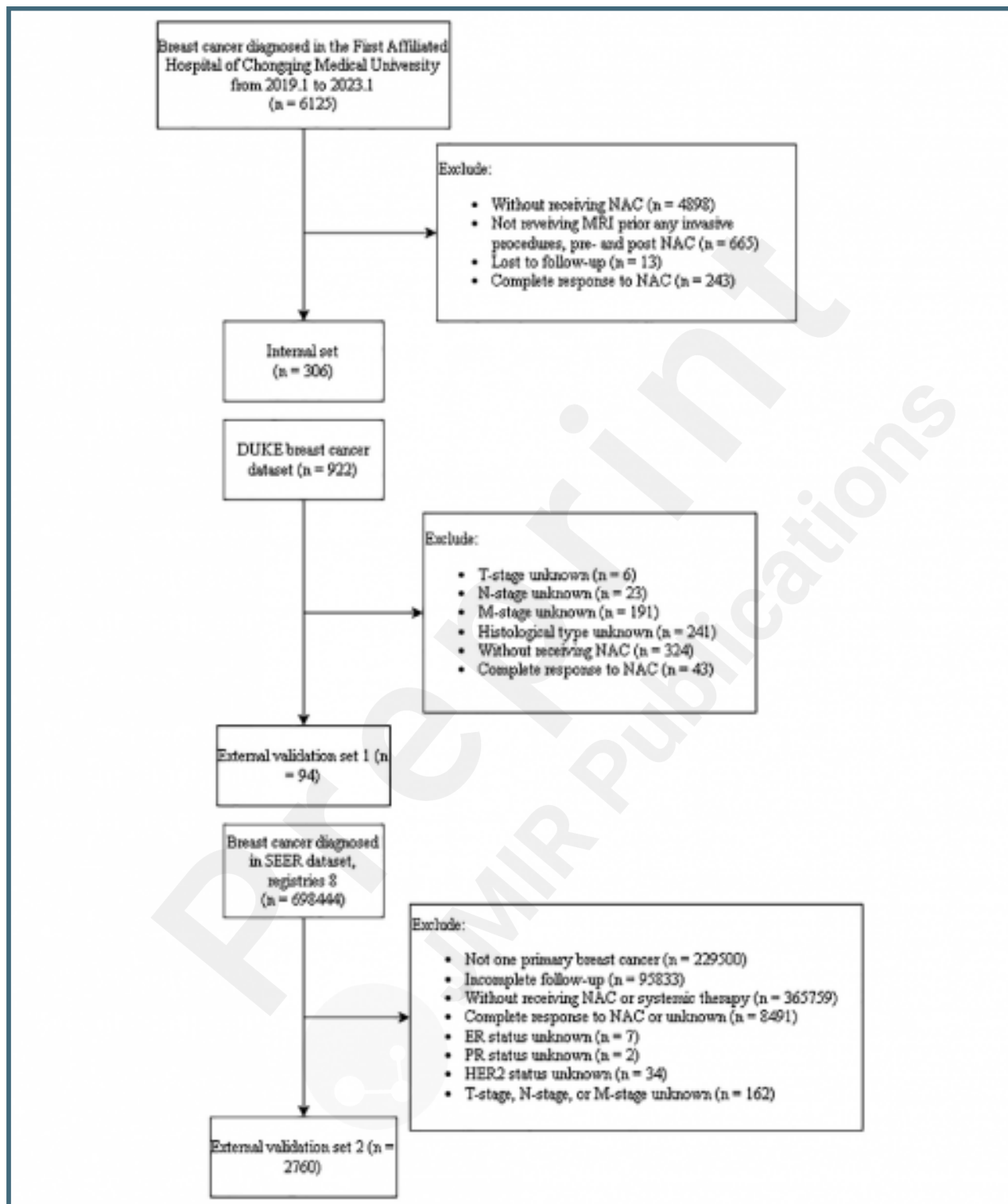
- for Metachronous Second Primary Lung Cancer and Predictions of Lung Cancer-Specific Outcomes Using Machine Learning: Population-Based Study. *JMIR Cancer*. 2024; 10e53354. Medline: 38865182 DOI: 10.2196/53354
7. Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. *Annu Rev Public Health*. 1999. 20: 145-57.
 8. Wu Y, Zhang Y, Duan S, Gu C, Wei C, Fang Y. Survival prediction in second primary breast cancer patients with machine learning: An analysis of SEER database. *Comput Methods Programs Biomed*. 2024. 254: 108310.
 9. Liao T, Su T, Lu Y, Huang L, Wei WY, Feng LH. Random survival forest algorithm for risk stratification and survival prediction in gastric neuroendocrine neoplasms. *Sci Rep*. 2024. 14(1): 26969.
 10. Tian D, Yan HJ, Huang H, et al. Machine Learning-Based Prognostic Model for Patients After Lung Transplantation. *JAMA Netw Open*. 2023. 6(5): e2312022.
 11. Yang X, Qiu H, Wang L, et al. Predicting Colorectal Cancer Survival Using Time-to-Event Machine Learning: Retrospective Cohort Study. *J Med Internet Res*. 2023; 25e44417. Medline: 37883174 DOI: 10.2196/44417
 12. Liu S, Chen YX, Dai B, Chen L. Development and Validation of a Novel Machine Learning Model to Predict the Survival of Patients with Gastrointestinal Neuroendocrine Neoplasms. *Neuroendocrinology*. 2024. 114(8): 733-748.
 13. Zhang YF, Shen YJ, Huang Q, Wu CP, Zhou L, Ren HL. Predicting survival of advanced laryngeal squamous cell carcinoma: comparison of machine learning models and Cox regression models. *Sci Rep*. 2023. 13(1): 18498.
 14. Schwartz LH, Litière S, de Vries E, et al. RECIST 1.1-Update and clarification: From the RECIST committee. *Eur J Cancer*. 2016. 62: 132-7.
 15. Korde LA, Somerfield MR, Carey LA, et al. Neoadjuvant Chemotherapy, Endocrine Therapy, and Targeted Therapy for Breast Cancer: ASCO Guideline. *J Clin Oncol*. 2021. 39(13): 1485-1505.
 16. Gradishar WJ, Moran MS, Abraham J, et al. Breast Cancer, Version 3.2022, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw*. 2022. 20(6): 691-722.
 17. Li JB, Jiang ZF. [Chinese Society of Clinical Oncology Breast Cancer Guideline version 2021: updates and interpretations]. *Zhonghua Yi Xue Za Zhi*. 2021. 101(24): 1835-1838.
 18. Allison KH, Hammond M, Dowsett M, et al. Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update. *J Clin Oncol*. 2020. 38(12): 1346-1366.
 19. Yu KD, Cai YW, Wu SY, Shui RH, Shao ZM. Estrogen receptor-low breast cancer: Biology chaos and treatment paradox. *Cancer Commun (Lond)*. 2021. 41(10): 968-980.
 20. Wolff AC, Hammond M, Allison KH, et al. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *J Clin Oncol*. 2018. 36(20): 2105-2122.
 21. Duke breast cancer dataset. <https://www.cancerimagingarchive.net/collection/duke-breast-cancer-mri/>.
 22. SEER database. <https://seer.cancer.gov/>.
 23. Zhao F, Polley E, McClellan J, Howard F, Olopade OI, Huo D. Predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer using a machine learning approach. *Breast Cancer Res*. 2024. 26(1): 148.
 24. Zhang J, Wu Q, Yin W, et al. Development and validation of a radiopathomic model for predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer patients. *BMC Cancer*. 2023. 23(1): 431.
 25. Mao N, Shi Y, Lian C, et al. Intratumoral and peritumoral radiomics for preoperative prediction of neoadjuvant chemotherapy effect in breast cancer based on contrast-enhanced spectral mammography. *Eur Radiol*. 2022. 32(5): 3207-3219.
 26. Roy S, Whitehead TD, Li S, et al. Co-clinical FDG-PET radiomic signature in predicting response to neoadjuvant chemotherapy in triple-negative breast cancer. *Eur J Nucl Med Mol Imaging*. 2022. 49(2): 550-562.
 27. Malhaire C, Selhane F, Saint-Martin MJ, et al. Exploring the added value of pretherapeutic MR descriptors in predicting breast cancer pathologic complete response to neoadjuvant chemotherapy. *Eur Radiol*. 2023. 33(11): 8142-8154.

28. Joo S, Ko ES, Kwon S, et al. Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. *Sci Rep*. 2021. 11(1): 18800.
29. Huang Y, Zhu T, Zhang X, et al. Longitudinal MRI-based fusion novel model predicts pathological complete response in breast cancer treated with neoadjuvant chemotherapy: a multicenter, retrospective study. *EClinicalMedicine*. 2023. 58: 101899.
30. Duanmu H, Bhattarai S, Li H, et al. A spatial attention guided deep learning system for prediction of pathological complete response using breast cancer histopathology images. *Bioinformatics*. 2022. 38(19): 4605-4612.
31. Sammut SJ, Crispin-Ortuzar M, Chin SF, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*. 2022. 601(7894): 623-629.
32. Chen J, Hao L, Qian X, Lin L, Pan Y, Han X. Machine learning models based on immunological genes to predict the response to neoadjuvant therapy in breast cancer patients. *Front Immunol*. 2022. 13: 948601.
33. Jin Y, Lan A, Dai Y, Jiang L, Liu S. Development and testing of a random forest-based machine learning model for predicting events among breast cancer patients with a poor response to neoadjuvant chemotherapy. *Eur J Med Res*. 2023. 28(1): 394.
34. Aswolinskiy W, Munari E, Horlings HM, et al. PROACTING: predicting pathological complete response to neoadjuvant chemotherapy in breast cancer from routine diagnostic histopathology biopsies with deep learning. *Breast Cancer Res*. 2023. 25(1): 142.
35. Fang C, Ji X, Pan Y, et al. Combining Clinical-Radiomics Features With Machine Learning Methods for Building Models to Predict Postoperative Recurrence in Patients With Chronic Subdural Hematoma: Retrospective Cohort Study. *J Med Internet Res*. 2024; 26e54944. Medline: 39197165 DOI: 10.2196/54944
36. Ji H, Kim S, Sunwoo L, et al. Integrating Clinical Data and Medical Imaging in Lung Cancer: Feasibility Study Using the Observational Medical Outcomes Partnership Common Data Model Extension. *JMIR Med Inform*. 2024; 12e59187. Medline: 38996330 DOI: 10.2196/59187

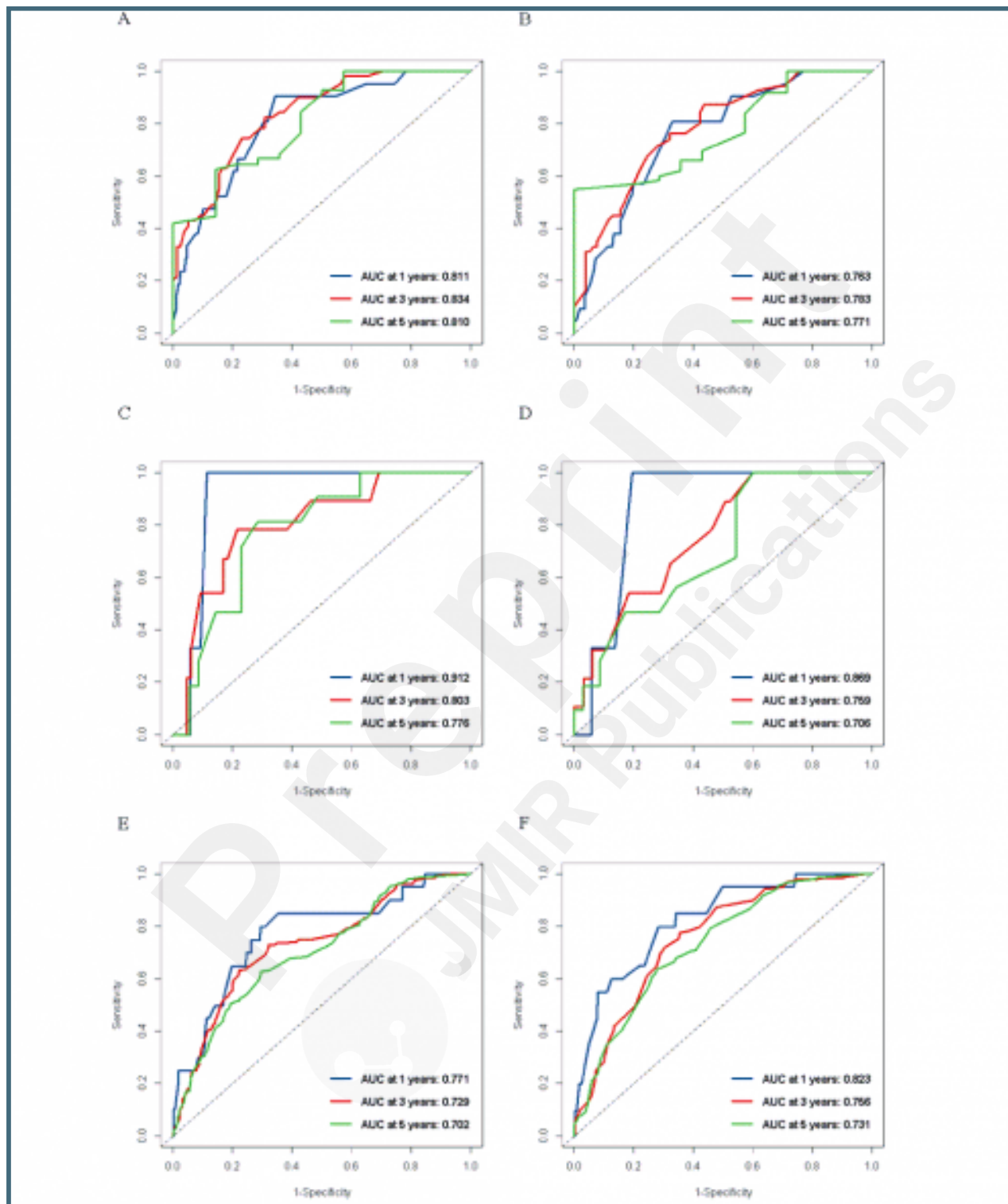
Supplementary Files

Figures

Flowchart of this study.



The ROC curves of the RSF model in the internal set (A), DUKE dataset (C), SEER dataset (E). And the ROC curves of the Cox regression model in the internal set (B), DUKE dataset (D), SEER dataset (F).



Multimedia Appendixes

Survival analysis comparing patients who achieved pCR or cCR after NAC with those who did not achieve pCR or cCR, across four groups: (A) all patients, (B) HR+/HER2- patients, (C) HR-/HER2+ patients, and (D) HR-/HER2- patients, using the internal dataset.

URL: <http://asset.jmir.pub/assets/6a491d9e3438a30c2dc4da83c9668acd.png>

Treatment protocols and pathological assessment of the patients in the internal dataset.

URL: <http://asset.jmir.pub/assets/8d3a8159f1a85508b333e1d875a3e068.docx>

The relationship between the number of trees in the model and the out-of-bag (OOB) error rate (A). The relative contribution of each predictor variable to the model's overall predictive accuracy (B).

URL: <http://asset.jmir.pub/assets/fabb868a8a98abea93156fd1f0a05b5b.png>

The out-of-bag (OOB) error rate for node size and mtry selection.

URL: <http://asset.jmir.pub/assets/a5b0f0808c32db104ecab43ad8b281b8.png>

The brier score of the RSF model and the Cox regression model, respectively.

URL: <http://asset.jmir.pub/assets/e0e1689a8e02a2a8e6ac163b602dcc81.png>

The DCA curves presented for 1-year, 3-year, and 5-year intervals for the internal set (A-C), DUKE dataset (D-F), and SEER dataset (G-I), respectively.

URL: <http://asset.jmir.pub/assets/0afe42b7c55e91864d8fa0ba2a115055.png>

The risk group setting by the RSF model (A) and the Cox regression model (B).

URL: <http://asset.jmir.pub/assets/6f33e3028fa2d1f80c9c157536a2f99c.png>

The RSF model stratified patients into high-risk and low-risk groups across the internal set (A), DUKE dataset (C), and SEER dataset (E). Similarly, the Cox regression model categorized patients into high-risk and low-risk groups in the internal set (B), DUKE dataset (D), and SEER dataset (F).

URL: <http://asset.jmir.pub/assets/8e768ff8f566cb02c0bb1f828f70dad5.png>

The ROC curves for the HR+/HER2+ (A), HR+/HER2- (B), HR-/HER2+ (C), and HR-/HER2- (D) subtypes were analyzed at 1-year, 3-year, and 5-year time points in the internal dataset, with the exception of the HR-/HER2- subtype, which was assessed only at 1-year and 3-year points. In the SEER dataset, the ROC curves for HR+/HER2+ (E), HR+/HER2- (F), HR-/HER2+ (G), and HR-/HER2- (H) were analyzed at 1-year, 3-year, and 5-year time points, except for the HR-/HER2+ subtype, which was evaluated only at 3-year and 5-year points. The DUKE dataset included ROC curves for HR+/HER2- (I) and HR-/HER2- (J) subtypes at all three time points: 1-year, 3-year, and 5-year. (Owing to the limited number of HR+/HER2+ and HR-/HER2+ subtypes in the DUKE dataset, ROC curves could not be plotted.?).

URL: <http://asset.jmir.pub/assets/3ea1062d559f441e19077aa8fe80ffdd.png>

The K-M curves of the RSF model stratified HR+/HER2+, HR+/HER2-, HR-/HER2+, and HR-/HER2- subtype patients into high-risk and low-risk groups within the internal (A-D, respectively) and SEER datasets (E-H, respectively). Similarly, the RSF model divided HR+/HER2- and HR-/HER2- subtype patients into risk groups within the DUKE dataset (I-J). (Owing to the limited number of HR+/HER2+ and HR-/HER2+ subtypes in the DUKE dataset, K-M curves could not be plotted.?).

URL: <http://asset.jmir.pub/assets/26556817198ee8f8df42164b9cb85127.png>