

Original Paper

Comparing Random Survival Forests and Cox Regression for Nonresponders to Neoadjuvant Chemotherapy Among Patients With Breast Cancer: Multicenter Retrospective Cohort Study

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

¹Department of Radiology, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China

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

Corresponding Author:

Fajin Lv, Prof Dr

Department of Radiology

The First Affiliated Hospital of Chongqing Medical University

No 1, Youyi Road

Yuzhong District

Chongqing, 400016

China

Phone: 86 02389012228

Fax: 86 023 68811487

Email: fajinlv@163.com

Abstract

Background: Breast cancer is one of the most common malignancies among women worldwide. Patients who 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 aimed to develop and validate a random survival forest (RSF) model to predict survival risk in patients with breast cancer who do not achieve a pCR or cCR post-NAC.

Methods: We analyzed patients with no pCR/cCR post-NAC treated at the First Affiliated Hospital of Chongqing Medical University from January 2019 to 2023, with external validation in Duke University and Surveillance, Epidemiology, and End Results (SEER) cohorts. RSF and Cox regression models were compared using the time-dependent area under the curve (AUC), the concordance index (C-index), and risk stratification.

Results: The study cohort included 306 patients with breast cancer, with most aged 40-60 years (204/306, 66.7%). The majority had invasive ductal carcinoma (290/306, 94.8%), with estrogen receptor (ER)+ (182/306, 59.5%), progesterone receptor (PR)- (179/306, 58.5%), and human epidermal growth factor receptor 2 (HER2)+ (94/306, 30.7%) profiles. Most patients presented with T2 (185/306, 60.5%), N1 (142/306, 46.4%), and M0 (295/306, 96.4%) staging (TNM meaning “tumor, node, metastasis”), with 17.6% (54/306) experiencing disease progression during a median follow-up of 25.9 months (IQR 17.2-36.3). External validation using Duke (N=94) and SEER (N=2760) cohorts confirmed consistent patterns in age (40-60 years: 59/94, 63%, vs 1480/2760, 53.6%), HER2+ rates (26/94, 28%, vs 935/2760, 33.9%), and invasive ductal carcinoma prevalence (89/94, 95%, vs 2506/2760, 90.8%). In the internal cohort, the RSF achieved significantly higher time-dependent AUCs compared to Cox regression at 1-year (0.811 vs 0.763), 3-year (0.834 vs 0.783), and 5-year (0.810 vs 0.771) intervals (overall C-index: 0.803, 95% CI 0.747-0.859, vs 0.736, 95% CI 0.673-0.799). External validation confirmed robust generalizability: the Duke cohort showed 1-, 3-, and 5-year AUCs of 0.912, 0.803, and 0.776, respectively, while the SEER cohort maintained consistent performance with AUCs of 0.771, 0.729, and 0.702, respectively. Risk stratification using the RSF identified 25.8% (79/306) high-risk patients and a significantly reduced survival time ($P<.001$). Notably, the RSF maintained improved net benefits across decision thresholds in decision curve analysis (DCA); similar results were observed in external studies. The RSF model also showed promising performance across different molecular subtypes in all datasets. Based on the RSF predicted scores, patients were stratified into high- 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 patients with breast cancer post-NAC. This approach may facilitate personalized treatment strategies and improve patient management in clinical practice.

(*J Med Internet Res* 2025;27:e69864) doi: [10.2196/69864](https://doi.org/10.2196/69864)

KEYWORDS

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 of developing effective prognostic tools to identify patients at high risk for adverse outcomes. However, there are limited studies focusing on developing predictive models for patients with breast cancer 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 use Cox regression to predict the prognosis of patients with cancer. 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, nonlinear relationships between independent variables and survival time [7]. Furthermore, previous studies have confirmed that some other models outperform the Cox regression model. Especially, many studies have shown that the random survival forest (RSF) model can manage high-dimensional data, improve the accuracy of survival predictions, and support personalized treatment strategies, which typically yields the best performance [8-13].

This study aimed to develop and validate an RSF model to predict survival risk in patients with breast cancer who fail to achieve a 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 Design

Patients diagnosed with breast cancer at the First Affiliated Hospital of Chongqing Medical University from January 2019 to 2023 were comprehensively reviewed. We selected patients who underwent NAC for subsequent analysis. After administering 4-8 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 examination, lasting for a minimum of 4 weeks [14]. A pathological complete response (pCR) was defined as the absence of any residual invasive tumor in both the breasts and axillary lymph nodes (ypT0ypN0) [15]. Patients without a 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]. Pathological assessment was conducted according to the American Society of Clinical Oncology guidelines (Multimedia Appendix 2) [18-20].

Initially, we established an 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 used as validation cohorts.

Ethical Considerations

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: 2020-59). Compensation and informed consent were waived due to the retrospective nature of the study and the use of deidentified patient data.

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 the patient survival duration. The 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 (R Foundation for Statistical Computing). 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 2 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

In this study, external validation datasets were obtained from 2 sources: the Duke breast cancer dataset and the SEER dataset. For the Duke dataset, the inclusion criterion was that patients had undergone NAC. The exclusion criteria were (1) any missing or unknown information and (2) patients who achieved a CR to NAC. For 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 were (1) patients with more than 1 primary tumor, (2) any missing information or records marked as “Unknown,” (3) patients who did not undergo NAC, and (4) patients who achieved a CR 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 and the Cox regression model was assessed using the 2 validation cohorts. Key metrics, including the concordance index (C-index), 95% CIs, and the integrated Brier score, were calculated to evaluate the models' predictive accuracy and calibration. Survival curves were generated using the Kaplan-Meier (K-M) method, stratified by risk group identified using the RSF and the Cox regression model. Additionally, variable importance plots were created to illustrate the contribution of each variable to the models. Finally, the models were validated in the Duke and the SEER dataset.

Model Performance Validation in Different Molecular Subtypes

We further used the RSF model to assess the performance of various molecular subtypes across different datasets, with the results illustrated through time-dependent receiver operating characteristic (ROC) curves. Patients with different subtypes in these datasets were categorized into high- and low-risk groups based on the model's predictions. The survival differences between these 2 groups were evaluated using K-M curves.

Statistical Analysis

We conducted statistical analysis using RStudio (R version 4.4.2). All independent variables were categorized and presented as absolute counts. To compare categorical data, we performed the chi-square test. The survival time was a continuous variable and was presented as the median (IQR). The Cox proportional hazards model was used 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 used to construct the Cox regression model. The ROC curve and 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, $P < .05$ was considered indicative of a statistically significant difference.

Results

Demographic and Clinical Characteristics

The study flowchart and patient inclusion process are illustrated in [Figure 1](#). The internal cohort comprised 306 patients with breast cancer with the following characteristics: the 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)+, 58.5% (179/306) progesterone receptor (PR)-, and 30.7% (94/306) human epidermal growth factor receptor 2 (HER2)+ 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) patients experienced disease-related events. External validation cohorts demonstrated comparable patterns, with the Duke dataset (N=94) showing a similar age distribution (59/94, 63%, aged 40-60 years), HER2-positivity rate (26/94, 28%), and hormone receptor (HR) profiles, while the SEER cohort (N=2760) maintained consistent invasive ductal carcinoma predominance (2506/2760, 90.8%) and staging trends. Full demographic details are presented in [Table 1](#).

Figure 1. Flowchart of this study. ER: estrogen receptor; HER2: human epidermal growth factor receptor 2; MRI, magnetic resonance imaging; NAC: neoadjuvant chemotherapy; PR: progesterone receptor; TNM: tumor, node, metastasis; SEER: Surveillance, Epidemiology, and End Results.

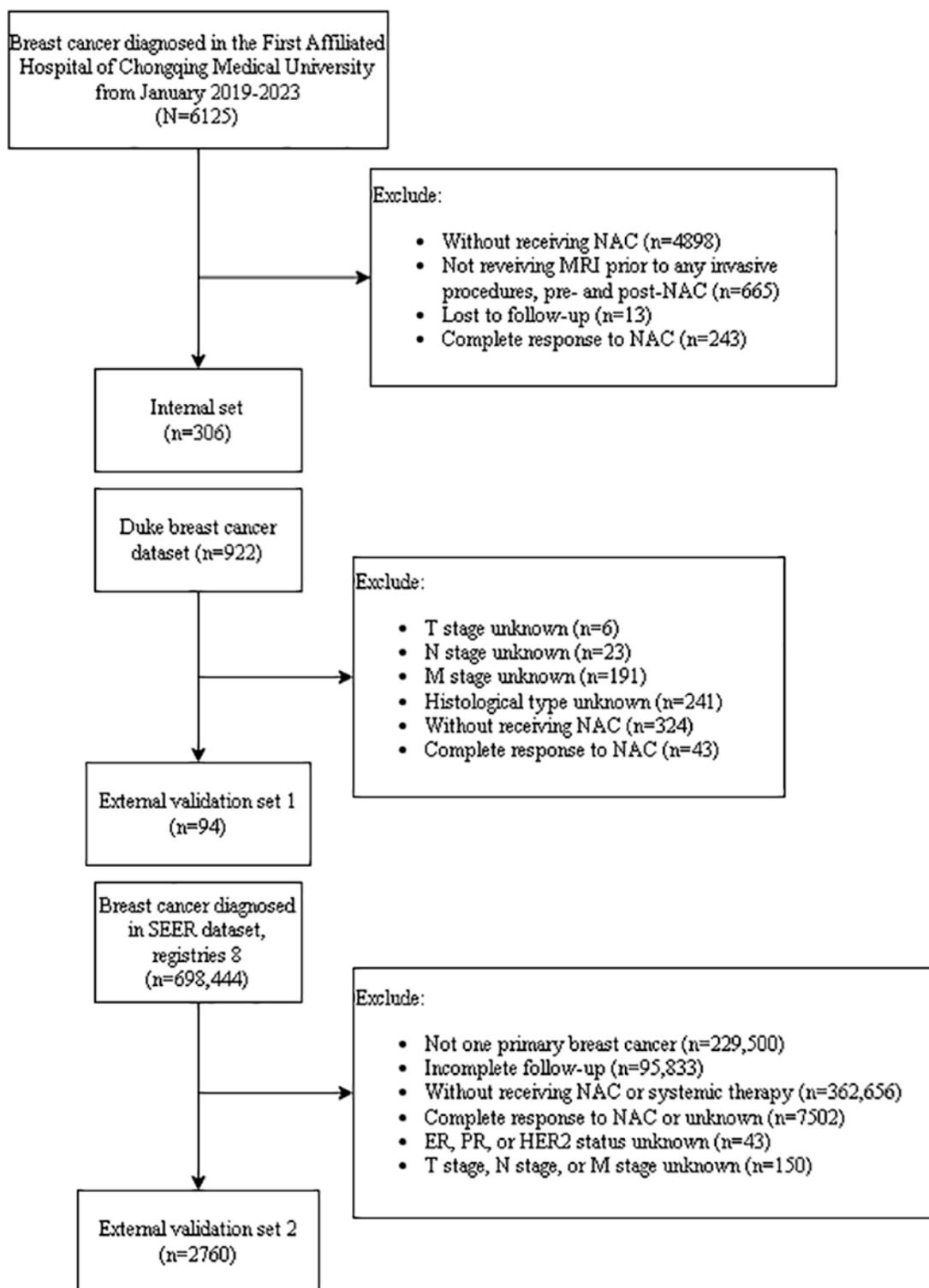


Table 1. Baseline characteristics of patients in the internal, Duke, and SEER^a datasets.

Characteristics	Internal dataset (N=306)	Duke dataset (N=94)	SEER dataset (N=2760)
Age (years), n (%)			
>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^b, n (%)			
Negative	124 (40.5)	40 (42.6)	686 (24.9)
Positive	182 (59.5)	54 (57.4)	2074 (75.1)
PR^c, n (%)			
Negative	179 (58.5)	52 (55.3)	1093 (39.6)
Positive	127 (41.5)	42 (44.7)	1667 (60.4)
HER2^d, 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 ^e -/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 (16.0)	753 (27.3)
T^f, n (%)			
1	36 (11.8)	16 (17.0)	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 (10.0)
N^f, 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^f, n (%)			
0	295 (96.4)	93 (98.9)	2677 (97.0)
1	11 (3.6)	1 (1.1)	83 (3.0)
S tatus , 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 (17.2-36.3)	50.2 (29.3-66.2)	35.0 (15.5-64.0)

^aSEER: Surveillance, Epidemiology, and End Results.^bER: estrogen receptor.^cPR: progesterone receptor.

^dHER2: human epidermal growth factor receptor 2.

^eHR: hormone receptor.

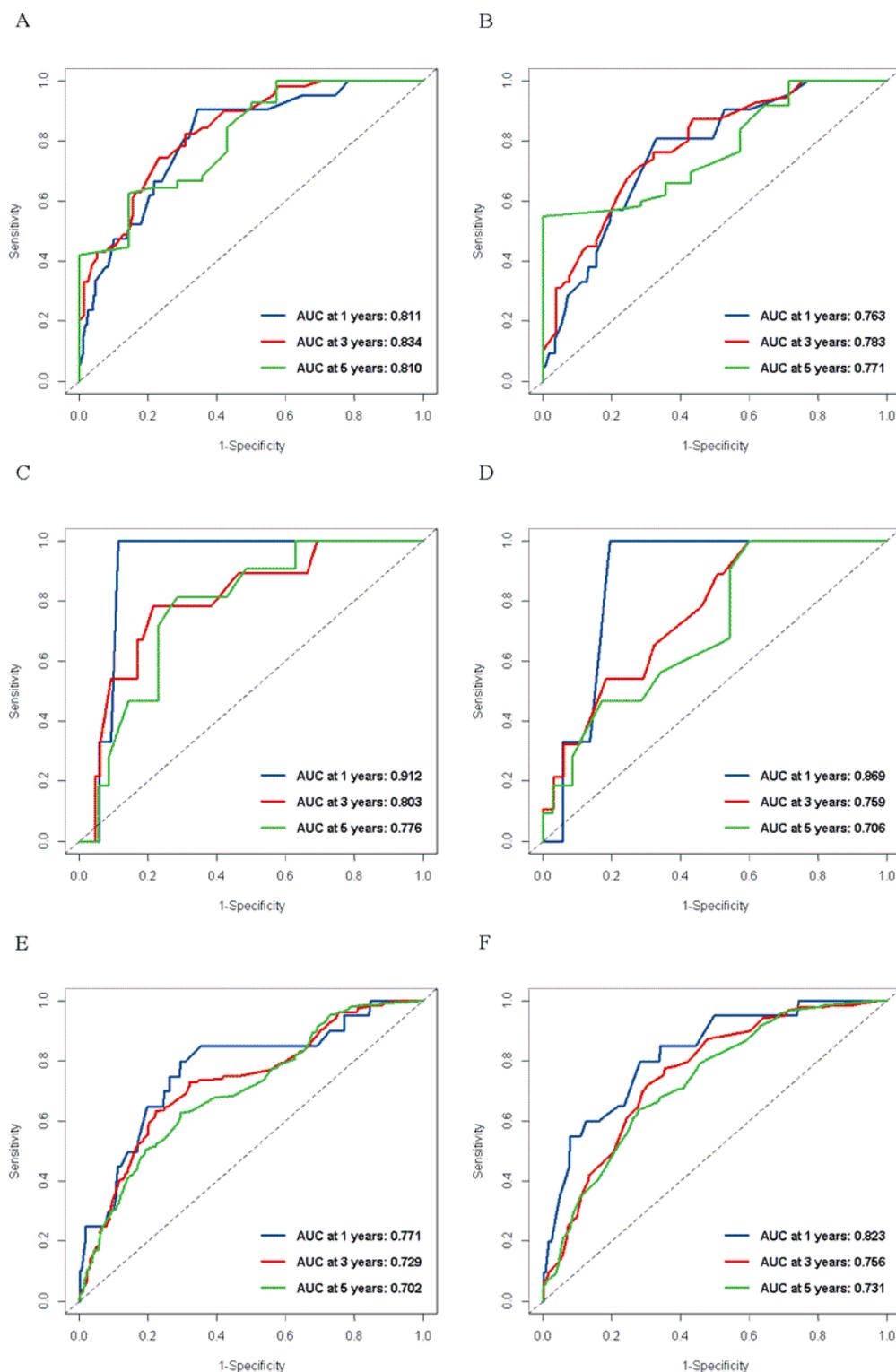
^fTNM: tumor, node, metastasis.

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, variable importance analysis revealed that the variable T stage had 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 optimal model performance and generalization ability were achieved with a node size 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, 3, and 5 years 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 is shown in [Multimedia Appendix 5](#).

Figure 2. ROC curves of the RSF model in the internal (A), Duke (C), and SEER (E) datasets. ROC curves of the Cox regression model in the internal (B), Duke (D), and SEER (F) datasets. AUC: area under the curve; ROC: receiver operating characteristic; SEER: Surveillance, Epidemiology, and End Results.



Construction of the Cox Regression Model

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-, 3-, and 5-year marks, respectively (Figure 2B). The C-index was calculated to be 0.736 (95% CI 0.673-0.799). The Brier score, which was relatively lower than that of the RSF model at each time point, is presented in Multimedia Appendix 5.

Table 2. Cox regression of clinicopathological variables.

Variables	Patients (N=306), n (%)	Univariable hazard ratio (95% CI); <i>P</i> value	Multivariable hazard ratio (95% CI); <i>P</i> value
Age (years)			
>60	42 (13.7)		
≤40	60 (19.6)	0.36 (0.15-0.87); <i>P</i> =.02	0.45 (0.18-1.12); <i>P</i> =.09
40-60	204 (66.7)	0.40 (0.20-0.77); <i>P</i> =.01	0.48 (0.24-0.97); <i>P</i> =.04
Histology			
Invasive ductal carcinoma	290 (94.8)		
Other	16 (5.2)	0.31 (0.04-2.22); <i>P</i> =.24	0.30 (0.04-2.25); <i>P</i> =.24
ER^a			
Negative	124 (40.5)		
Positive	182 (59.5)	0.74 (0.44-1.27); <i>P</i> =.28	1.05 (0.53-2.09); <i>P</i> =.88
PR^b			
Negative	179 (58.5)		
Positive	127 (41.5)	0.54 (0.30-0.98); <i>P</i> =.04	0.45 (0.21-0.96); <i>P</i> =.04
HER2^c			
Negative	212 (69.3)		
Positive	94 (30.7)	0.54 (0.29-1.02); <i>P</i> =.06	0.45 (0.23-0.87); <i>P</i> =.02
T^d			
1	36 (11.8)		
2	185 (60.5)	1.42 (0.55-3.63); <i>P</i> =.47	1.47 (0.57-3.82); <i>P</i> =.42
3	51 (16.7)	1.35 (0.44-4.14); <i>P</i> =.60	1.26 (0.40-3.93); <i>P</i> =.69
4	34 (11.1)	1.73 (0.56-5.29); <i>P</i> =.34	1.58 (0.49-5.08); <i>P</i> =.45
N^d			
0	43 (14.1)		
1	142 (46.4)	1.84 (0.63-5.38); <i>P</i> =.26	1.65 (0.55-4.93); <i>P</i> =.37
2	61 (19.9)	2.75 (0.89-8.55); <i>P</i> =.08	2.06 (0.65-6.58); <i>P</i> =.22
3	60 (19.6)	4.34 (1.45-12.96); <i>P</i> =.01	3.36 (1.10-10.24); <i>P</i> =.03
M^d			
0	295 (96.4)		
1	11 (3.6)	3.30 (1.31-8.31); <i>P</i> =.01	2.74 (1.03-7.25); <i>P</i> =.04

^aER: estrogen receptor.

^bPR: progesterone receptor.

^cHER2: human epidermal growth factor receptor 2.

^dTNM: tumor, node, metastasis.

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 is displayed in [Figure 1](#). For the RSF model, the AUC for 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](#), while the AUC values for the SEER dataset 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 for 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](#). For 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 and Cox Regression Model

[Multimedia Appendix 6](#) presents DCA curves for the 2 models at 1-, 3-, 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- and low-risk groups (Multimedia Appendix 7). Survival analysis demonstrated that the prognosis for the high-risk group was significantly worse than that for 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- and low-risk groups. Survival analysis 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- and low-risk groups (Multimedia Appendix 7). Survival analysis 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 in the internal dataset, the RSF model achieved an AUC of 1.000 for both 1- and 3-year survival rates and 0.748 for the 5-year survival rate for the HR+/HER2+ subtype. For 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 rate. 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 rate. For the HR-/HER2- subtype, the AUC values were 0.681 for the 1-year and 0.832 for the 3-year survival rate. Consistent trends were observed in the SEER and Duke validation datasets (Multimedia Appendix 9). The K-M curves further demonstrated the RSF model's ability to stratify patients into distinct high- and low-risk groups across all 4 molecular subtypes in both internal and SEER datasets. In the Duke dataset, although no statistically significant difference was observed in the HR-/HER2- subtype, low-risk patients still exhibited higher RFS compared to high-risk patients (Multimedia Appendix 10).

Discussion

Principal Findings

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

Previous research has demonstrated that achieving a pCR following NAC is associated with significantly improved event-free survival and OS [3]. Consequently, numerous studies have concentrated on predicting tumor responses to NAC. For instance, Zhao et al [23] constructed machine learning models to predict the pCR to NAC based on clinicopathological variables. Similarly, Zhang and coworkers [24-30] developed machine learning models that incorporated clinicopathological features, radiomic features, and pathomic features to forecast

the pCR following NAC. Additionally, Sammut et al [31] and Chen et al [32] created models using multi-omics data. These studies have highlighted the substantial value of machine learning models in predicting patients' responses to NAC.

However, as previous research has confirmed that patients with no pCR tend to have a worse prognosis and a higher risk for adverse events, it is important to note that few studies have explored risk stratification among those patients. In one of our earlier studies, we tried to develop a random forest model to predict event occurrence among patients with breast cancer 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, the RSF model offers several advantages: It is not limited by the proportional hazards assumption, it can effectively handle high-dimensional data, and it demonstrates strong generalization capabilities [11]. Liao et al [9] developed a prediction model using 10 machine learning algorithms across 101 combinations to forecast cancer-related mortality in patients with gastric neuroendocrine neoplasms. They finally found that the RSF model obtains 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 outperforms the Cox regression model, a finding that was further validated using the Duke dataset.

Strengths

One of the strengths of our study is the robust validation of the RSF model using 2 independent validation cohorts. 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 used DFS, whereas the Duke dataset was based on the RFS, and the SEER dataset used OS. The DFS status encompassed disease recurrence, new primary diseases, and death. Consequently, the model demonstrated strong adaptability across datasets that use 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 tried to develop the RSF model using variables selected through multivariate analysis in the Cox regression; however, the results were not satisfactory. Subsequently, we used the variable importance metrics from the RSF algorithm to identify variables that were ultimately included in the model. Hence, we also suggested using a variable selection method that aligns with the specific model being used.

Limitations

To the best of our knowledge, our study demonstrated that a more complex algorithm could effectively predict the survival risk of patients with breast cancer without a CR post-NAC, based solely on clinicopathological variables. Nevertheless, there are limitations of our study. First, the sample of the training set was relative small, which might affect variable 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 models' 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. Fourth, 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 2 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 the RSF model for predicting prognosis using clinicopathological variables in patients. Further prospective study is necessary to confirm its applicability.

Moreover, the RSF model is particularly well suited for handling higher-dimensional variables. Thus, future studies should also explore the integration of radiomics, pathomics, and molecular data to enhance the predictive power of the model [34-36].

Conclusion

Our study highlighted the feasibility and effectiveness of using an RSF model based exclusively on clinicopathological variables to predict survival risk in patients with breast cancer who do not achieve a 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 continue to refine and validate this model, we anticipate that it would significantly contribute to the advancement of personalized medicine in breast cancer care.

Acknowledgments

We thank all the participants in the SEER database, as well as the Duke breast cancer dataset, for contributing their works to support our study. 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 the paper, or the decision to publish.

Data Availability

The data of the First Affiliated Hospital of Chongqing Medical University are available from the corresponding author upon request. The data of the Duke breast cancer dataset can be searched online [21]. The data of the SEER dataset can be retrieved from SEERStat [22].

Authors' Contributions

YJ conceived the study concept; YJ and TS designed the study protocol and performed statistical analysis and interpretation; YF, MZ, and YJ retrieved and selected the data; YJ and MZ completed the draft writing and manuscript revision; YJ and FL oversaw the integration of the entire study; FL and ZO edited the manuscript and approved the final version. All authors have approved the submission of the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Survival analysis comparing patients who achieved a pCR or a cCR after NAC with those who did not, using the internal dataset across 4 groups: (A) all patients, (B) HR+ patients, (C) HR-/HER2+ patients, and (D) HR-/HER2- patients. cCR: clinical complete response; HER2: human epidermal growth factor receptor 2; HR: hormone receptor; NAC: neoadjuvant chemotherapy; pCR: pathological complete response.

[\[PNG File , 28 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

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

[\[DOCX File , 12 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

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

[\[PNG File , 14 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

OOB error rate for node size and mtry selection. OOB: out of bag.

[\[PNG File , 21 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Brier scores of the RSF and Cox regression models. RSF: random survival forest.

[\[PNG File , 11 KB-Multimedia Appendix 5\]](#)

Multimedia Appendix 6

DCA curves presented for 1-, 3-, and 5-year intervals for the internal set (A-C), the Duke dataset (D-F), and the SEER dataset (G-I). DCA: decision curve analysis; SEER: Surveillance, Epidemiology, and End Results.

[\[PNG File , 64 KB-Multimedia Appendix 6\]](#)

Multimedia Appendix 7

Risk group setting using RSF (A) and Cox regression (B) models. RSF: random survival forest.

[\[PNG File , 33 KB-Multimedia Appendix 7\]](#)

Multimedia Appendix 8

Stratification of patients into high- and low- groups across the internal (A), Duke (C), and SEER (E) datasets using the RSF model and across the internal (B), Duke (D), and SEER (F) datasets using the Cox regression model. RSF: random survival forest; SEER: Surveillance, Epidemiology, and End Results.

[\[PNG File , 75 KB-Multimedia Appendix 8\]](#)

Multimedia Appendix 9

ROC curves for HR+/HER2+ (A), HR+/HER2- (B), HR-/HER2+ (C), and HR-/HER2- (D) subtypes were analyzed at 1-, 3-, and 5-year time points in the internal dataset, with the exception of the HR-/HER2- subtype, which was assessed only at 1- and 3-year time points. In the SEER dataset, ROC curves for HR+/HER2+ (E), HR+/HER2- (F), HR-/HER2+ (G), and HR-/HER2- (H) subtypes were analyzed at 1-, 3-, and 5-year time points, except for the HR-/HER2+ subtype, which was evaluated only at 3- and 5-year time points. The Duke dataset included ROC curves for HR+/HER2- (I) and HR-/HER2- (J) subtypes at all 3 time points (1, 3, and 5 years). (Owing to the limited number of HR+/HER2+ and HR-/HER2+ subtypes in the Duke dataset, ROC curves could not be plotted.) HER2: human epidermal growth factor receptor 2; HR: hormone receptor; ROC: receiver operating characteristic; SEER: Surveillance, Epidemiology, and End Results.

[\[PNG File , 83 KB-Multimedia Appendix 9\]](#)

Multimedia Appendix 10

K-M curves of the RSF model stratified patients with HR+/HER2+, HR+/HER2-, HR-/HER2+, and HR-/HER2- subtypes into high- and low-risk groups within the internal (A-D, respectively) and SEER (E-H, respectively) datasets. Similarly, the RSF model divided patients with HR+/HER2- and HR-/HER2- subtypes into high and low-risk groups within the Duke dataset (I-J, respectively). (Owing to the limited number of HR+/HER2+ and HR-/HER2+ subtypes in the Duke dataset, K-M curves could not be plotted.) HER2: human epidermal growth factor receptor 2; HR: hormone receptor; K-M: Kaplan-Meier; RSF: random survival forest; SEER: Surveillance, Epidemiology, and End Results.

[\[PNG File , 70 KB-Multimedia Appendix 10\]](#)

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* Apr 04, 2024;74(3):229-263. [[FREE Full text](#)] [doi: [10.3322/caac.21834](https://doi.org/10.3322/caac.21834)] [Medline: [38572751](https://pubmed.ncbi.nlm.nih.gov/38572751/)]
2. Giaquinto AN, Sung H, Newman LA, Freedman RA, Smith RA, Star J, et al. Breast cancer statistics 2024. *CA Cancer J Clin.* Oct 2024;74(6):477-495. [[FREE Full text](#)] [doi: [10.3322/caac.21863](https://doi.org/10.3322/caac.21863)] [Medline: [39352042](https://pubmed.ncbi.nlm.nih.gov/39352042/)]
3. Spring L, Fell G, Arfe A. 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. [doi: [10.1158/1078-0432.ccr-19-3492](https://doi.org/10.1158/1078-0432.ccr-19-3492)]
4. Xiao J, Mo M, Wang Z, Zhou C, Shen J, Yuan J, et al. The application and comparison of machine learning models for the prediction of breast cancer prognosis: retrospective cohort study. *JMIR Med Inform.* Feb 18, 2022;10(2):e33440. [[FREE Full text](#)] [doi: [10.2196/33440](https://doi.org/10.2196/33440)] [Medline: [35179504](https://pubmed.ncbi.nlm.nih.gov/35179504/)]

5. Janbain A, Farolfi A, Guenegou-Arnoux A, Romengas L, Scharl S, Fanti S, 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*. Sep 20, 2024;10:e60323. [[FREE Full text](#)] [doi: [10.2196/60323](https://doi.org/10.2196/60323)] [Medline: [39303279](#)]
6. Zheng Y, Zhao A, Yang Y, Wang L, Hu Y, Luo R, 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*. Jun 12, 2024;10:e53354. [[FREE Full text](#)] [doi: [10.2196/53354](https://doi.org/10.2196/53354)] [Medline: [38865182](#)]
7. Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. *Annu Rev Public Health*. May 1999;20(1):145-157. [doi: [10.1146/annurev.publhealth.20.1.145](https://doi.org/10.1146/annurev.publhealth.20.1.145)] [Medline: [10352854](#)]
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*. Sep 2024;254:108310. [doi: [10.1016/j.cmpb.2024.108310](https://doi.org/10.1016/j.cmpb.2024.108310)] [Medline: [38996803](#)]
9. Liao T, Su T, Lu Y, Huang L, Wei W, Feng L. Random survival forest algorithm for risk stratification and survival prediction in gastric neuroendocrine neoplasms. *Sci Rep*. Nov 06, 2024;14(1):26969. [[FREE Full text](#)] [doi: [10.1038/s41598-024-77988-1](https://doi.org/10.1038/s41598-024-77988-1)] [Medline: [39506090](#)]
10. Tian D, Yan H, Huang H, Zuo Y, Liu M, Zhao J, et al. Machine learning-based prognostic model for patients after lung transplantation. *JAMA Netw Open*. May 01, 2023;6(5):e2312022. [[FREE Full text](#)] [doi: [10.1001/jamanetworkopen.2023.12022](https://doi.org/10.1001/jamanetworkopen.2023.12022)] [Medline: [37145595](#)]
11. Yang X, Qiu H, Wang L, Wang X. Predicting colorectal cancer survival using time-to-event machine learning: retrospective cohort study. *J Med Internet Res*. Oct 26, 2023;25:e44417. [[FREE Full text](#)] [doi: [10.2196/44417](https://doi.org/10.2196/44417)] [Medline: [37883174](#)]
12. Liu S, Chen Y, Dai B, Chen L. Development and validation of a novel machine learning model to predict the survival of patients with gastrointestinal neuroendocrine neoplasms. *Neuroendocrinology*. May 6, 2024;114(8):733-748. [doi: [10.1159/000539187](https://doi.org/10.1159/000539187)] [Medline: [38710164](#)]
13. Zhang Y, Shen Y, Huang Q, Wu C, Zhou L, Ren H. Predicting survival of advanced laryngeal squamous cell carcinoma: comparison of machine learning models and Cox regression models. *Sci Rep*. Oct 28, 2023;13(1):18498. [[FREE Full text](#)] [doi: [10.1038/s41598-023-45831-8](https://doi.org/10.1038/s41598-023-45831-8)] [Medline: [37898687](#)]
14. Schwartz LH, Litière S, de Vries E, Ford R, Gwyther S, Mandrekar S, et al. RECIST 1.1-update and clarification: from the RECIST committee. *Eur J Cancer*. Jul 2016;62:132-137. [[FREE Full text](#)] [doi: [10.1016/j.ejca.2016.03.081](https://doi.org/10.1016/j.ejca.2016.03.081)] [Medline: [27189322](#)]
15. Korde LA, Somerfield MR, Carey LA, Crews JR, Denduluri N, Hwang ES, et al. Neoadjuvant chemotherapy, endocrine therapy, and targeted therapy for breast cancer: ASCO guideline. *J Clin Oncol*. May 01, 2021;39(13):1485-1505. [doi: [10.1200/jco.20.03399](https://doi.org/10.1200/jco.20.03399)]
16. Gradishar WJ, Moran MS, Abraham J, Aft R, Agnese D, Allison KH, et al. Breast cancer, version 3.2022, NCCN clinical practice guidelines in oncology. *J Natl Compr Canc Netw*. Jun 2022;20(6):691-722. [doi: [10.6004/jnccn.2022.0030](https://doi.org/10.6004/jnccn.2022.0030)] [Medline: [35714673](#)]
17. Li JB, Jiang ZF. [Chinese Society of Clinical Oncology Breast Cancer Guideline version 2021: updates and interpretations]. *Zhonghua Yi Xue Za Zhi*. Jun 29, 2021;101(24):1835-1838. [doi: [10.3760/cma.j.cn112137-20210421-00954](https://doi.org/10.3760/cma.j.cn112137-20210421-00954)] [Medline: [34619851](#)]
18. Allison KH, Hammond MEH, Dowsett M, McKernin SE, Carey LA, Fitzgibbons PL, et al. Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP guideline update. *J Clin Oncol*. Apr 20, 2020;38(12):1346-1366. [doi: [10.1200/jco.19.02309](https://doi.org/10.1200/jco.19.02309)]
19. Yu K, Cai Y, Wu S, Shui R, Shao Z. Estrogen receptor-low breast cancer: biology chaos and treatment paradox. *Cancer Commun (Lond)*. Oct 12, 2021;41(10):968-980. [[FREE Full text](#)] [doi: [10.1002/cac2.12191](https://doi.org/10.1002/cac2.12191)] [Medline: [34251757](#)]
20. Wolff AC, Hammond MEH, Allison KH, Harvey BE, Mangu PB, Bartlett JM, 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*. Jul 10, 2018;36(20):2105-2122. [doi: [10.1200/jco.2018.77.8738](https://doi.org/10.1200/jco.2018.77.8738)]
21. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations. National Cancer Institute. URL: <https://www.cancerimagingarchive.net/collection/duke-breast-cancer-mri/> [accessed 2025-04-04]
22. Surveillance, Epidemiology, and End Results program. National Cancer Institute. URL: <https://seer.cancer.gov/> [accessed 2025-04-04]
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*. Oct 29, 2024;26(1):148. [[FREE Full text](#)] [doi: [10.1186/s13058-024-01905-7](https://doi.org/10.1186/s13058-024-01905-7)] [Medline: [39472970](#)]
24. Zhang J, Wu Q, Yin W, Yang L, Xiao B, Wang J, et al. Development and validation of a radiopathomic model for predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer patients. *BMC Cancer*. May 12, 2023;23(1):431. [[FREE Full text](#)] [doi: [10.1186/s12885-023-10817-2](https://doi.org/10.1186/s12885-023-10817-2)] [Medline: [37173635](#)]
25. Mao N, Shi Y, Lian C, Wang Z, Zhang K, Xie H, et al. Intratumoral and peritumoral radiomics for preoperative prediction of neoadjuvant chemotherapy effect in breast cancer based on contrast-enhanced spectral mammography. *Eur Radiol*. May 23, 2022;32(5):3207-3219. [doi: [10.1007/s00330-021-08414-7](https://doi.org/10.1007/s00330-021-08414-7)] [Medline: [35066632](#)]

26. Roy S, Whitehead TD, Li S, Ademuyiwa FO, Wahl RL, Dehdashti F, 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*. Jul 30, 2021;49(2):550-562. [doi: [10.1007/s00259-021-05489-8](https://doi.org/10.1007/s00259-021-05489-8)]
27. Malhaire C, Selhane F, Saint-Martin M, Cockenpot V, Akl P, Laas E, et al. Exploring the added value of pretherapeutic MR descriptors in predicting breast cancer pathologic complete response to neoadjuvant chemotherapy. *Eur Radiol*. Nov 15, 2023;33(11):8142-8154. [doi: [10.1007/s00330-023-09797-5](https://doi.org/10.1007/s00330-023-09797-5)] [Medline: [37318605](https://pubmed.ncbi.nlm.nih.gov/37318605/)]
28. Joo S, Ko ES, Kwon S, Jeon E, Jung H, Kim J, et al. Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. *Sci Rep*. Sep 22, 2021;11(1):18800. [FREE Full text] [doi: [10.1038/s41598-021-98408-8](https://doi.org/10.1038/s41598-021-98408-8)] [Medline: [34552163](https://pubmed.ncbi.nlm.nih.gov/34552163/)]
29. Huang Y, Zhu T, Zhang X, Li W, Zheng X, Cheng M, et al. Longitudinal MRI-based fusion novel model predicts pathological complete response in breast cancer treated with neoadjuvant chemotherapy: a multicenter, retrospective study. *EClinicalMedicine*. Apr 2023;58:101899. [FREE Full text] [doi: [10.1016/j.eclinm.2023.101899](https://doi.org/10.1016/j.eclinm.2023.101899)] [Medline: [37007742](https://pubmed.ncbi.nlm.nih.gov/37007742/)]
30. Duanmu H, Bhattarai S, Li H, Shi Z, Wang F, Teodoro G, et al. A spatial attention guided deep learning system for prediction of pathological complete response using breast cancer histopathology images. *Bioinformatics*. Sep 30, 2022;38(19):4605-4612. [FREE Full text] [doi: [10.1093/bioinformatics/btac558](https://doi.org/10.1093/bioinformatics/btac558)] [Medline: [35962988](https://pubmed.ncbi.nlm.nih.gov/35962988/)]
31. Sammut S, Crispin-Ortuzar M, Chin S, Provenzano E, Bardwell HA, Ma W, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*. Jan 07, 2022;601(7894):623-629. [FREE Full text] [doi: [10.1038/s41586-021-04278-5](https://doi.org/10.1038/s41586-021-04278-5)] [Medline: [34875674](https://pubmed.ncbi.nlm.nih.gov/34875674/)]
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*. Jul 22, 2022;13:948601. [FREE Full text] [doi: [10.3389/fimmu.2022.948601](https://doi.org/10.3389/fimmu.2022.948601)] [Medline: [35935976](https://pubmed.ncbi.nlm.nih.gov/35935976/)]
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*. Sep 30, 2023;28(1):394. [FREE Full text] [doi: [10.1186/s40001-023-01361-7](https://doi.org/10.1186/s40001-023-01361-7)] [Medline: [37777809](https://pubmed.ncbi.nlm.nih.gov/37777809/)]
34. Aswolinskiy W, Munari E, Horlings HM, Mulder L, Bogina G, Sanders J, et al. PROACTING: predicting pathological complete response to neoadjuvant chemotherapy in breast cancer from routine diagnostic histopathology biopsies with deep learning. *Breast Cancer Res*. Nov 13, 2023;25(1):142. [FREE Full text] [doi: [10.1186/s13058-023-01726-0](https://doi.org/10.1186/s13058-023-01726-0)] [Medline: [37957667](https://pubmed.ncbi.nlm.nih.gov/37957667/)]
35. Fang C, Ji X, Pan Y, Xie G, Zhang H, Li S, 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*. Aug 28, 2024;26:e54944. [FREE Full text] [doi: [10.2196/54944](https://doi.org/10.2196/54944)] [Medline: [39197165](https://pubmed.ncbi.nlm.nih.gov/39197165/)]
36. Ji H, Kim S, Sunwoo L, Jang S, Lee H, Yoo S. Integrating clinical data and medical imaging in lung cancer: feasibility study using the Observational Medical Outcomes Partnership Common Data Model Extension. *JMIR Med Inform*. Jul 12, 2024;12:e59187. [FREE Full text] [doi: [10.2196/59187](https://doi.org/10.2196/59187)] [Medline: [38996330](https://pubmed.ncbi.nlm.nih.gov/38996330/)]

Abbreviations

AUC: area under the curve
cCR: clinical complete response
C-index: concordance index
CR: complete response
DCA: decision curve analysis
DFS: disease-free survival
ER: estrogen receptor
HER2: human epidermal growth factor receptor 2
HR: hormone receptor
K-M: Kaplan-Meier
NAC: neoadjuvant chemotherapy
OS: overall survival
pCR: pathological complete response
PR: progesterone receptor
RFS: recurrence-free survival
ROC: receiver operating characteristic
RSF: random survival forest
SEER: Surveillance, Epidemiology, and End Results
TNM: tumor, node, metastasis

Edited by J Sarvestan; submitted 10.12.24; peer-reviewed by L Han, Z-L Cheng; comments to author 01.02.25; revised version received 19.02.25; accepted 25.03.25; published 08.04.25

Please cite as:

Jin Y, Zhao M, Su T, Fan Y, Ouyang Z, Lv F

Comparing Random Survival Forests and Cox Regression for Nonresponders to Neoadjuvant Chemotherapy Among Patients With Breast Cancer: Multicenter Retrospective Cohort Study

J Med Internet Res 2025;27:e69864

URL: <https://www.jmir.org/2025/1/e69864>

doi: [10.2196/69864](https://doi.org/10.2196/69864)

PMID:

©Yudi Jin, Min Zhao, Tong Su, Yanjia Fan, Zubin Ouyang, Fajin Lv. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 08.04.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.