Brazilian Journal of Biometrics **44**, 1-15 (2026)
doi:10.28951/bjb.v44i3.979

ARTICLE

Spline-Enhanced survival modelling of treatment effect heterogeneity in breast Cancer

Patrick Oni Adebayo*[1], Ibrahim Ahmed[2], Azeez Daramol Mustaph[3] and Abubakar Abdullahi Wakili[4]

[1]Department of Statistics, Federal University of Agriculture and Technology Okeho, Oyo State, Nigeria.
[2]Department of Statistics, Nasarawa State University, Keffi, Nasarawa State, Nigeria.
[3]National Bureau of Statistics, Nigeria.
[4]Department of Higher Education, Federal Capital Territory Adminstration, Abuja, Nigeria.
*Corresponding author. Email: patrick.adebayo@funato.edu.ng

(Received: August 31, 2025; Revised: February 9, 2026; Accepted: February 19, 2026; Published: September 28, 2026)
Section editor: João Domingos Scalon

Abstract
This study investigates breast cancer prognosis using the Rotterdam dataset by integrating traditional and advanced modelling approaches to examine treatment heterogeneity and non-linear predictor effects. Linear regression revealed tumor size (>50 mm) and hormonal therapy were strongly associated with higher nodal counts, the latter likely indicative of treatment selection bias. Logistic regression provided superior predictive performance for recurrence (AUC = 0.71) compared with linear models (AUC = 0.61). Cubic spline functions were applied to Cox proportional hazards models to capture non-linear relationships, uncovering a U-shaped effect of age on recurrence risk and a pronounced hazard increase (HR = 38.3) for patients with 1–3 positive nodes, indicating a strong threshold effect. While chemotherapy showed a modest overall benefit (HR = 0.89), its interaction with nodal status was not significant (p = 0.78). Key predictors, including tumor size and grade, violated the proportional hazards assumption, indicating time-varying effects. Methodologically, logistic regression offered strong predictive performance, whereas spline-based Cox models provided deeper clinical insight into complex risk patterns. These findings underscore the value of flexible, non-linear modelling in oncology and suggest that risk stratification based solely on linear assumptions or nodal thresholds may require refinement.

Keywords: Breast Cancer Prognosis, Spline Regression, Nodal Status, Precision Oncology, Recurrence-free Survival.

1. Introduction

Breast cancer management faces a critical challenge in understanding why treatment responses vary across patient subgroups. While adjuvant therapies such as chemotherapy and hormonal treatments have improved overall survival, their benefits are not uniformly distributed (Sparano *et al.*, 2018). Traditional prognostic models often assume linear relationships between clinical variables and outcomes, which can obscure important non-linear patterns in treatment efficacy (Harrell, 2015). Flexible modelling techniques like spline regression are uniquely suited to address this, as they allow the data to reveal the true functional form of risk relationships without restrictive linear assumptions. For instance, they can identify if the risk from nodal burden plateaus after a certain count or if age

exhibits a U-shaped relationship with recurrence patterns that linear models would miss (Durrleman *et al.*, 1989; Harrell, 2015). This study addresses these limitations by performing an integrated analysis of the Rotterdam breast cancer cohort, combining flexible spline regression with interaction-term analysis to better characterize heterogeneity in treatment effects.

Current approaches to modelling recurrence risk frequently overlook two key aspects: (1) the non-linear dependence of treatment benefit on continuous predictors such as age and nodal count, and (2) how treatment effects differ across biologically distinct patient subgroups. While nodal status is well-established as a prognostic factor (Carter *et al.*, 1989), less attention has been paid to how nodal burden modifies specific treatment effects. Similarly, although menopausal status is routinely considered in therapeutic decisions, its interaction with chemotherapy efficacy remains incompletely understood (Early Breast Cancer Trialists' Collaborative Group, 2012).

The Rotterdam dataset provides an ideal platform for this investigation, offering detailed treatment records and long-term follow-up (Foekens *et al.*, 2000). Our analysis extends beyond conventional Cox models through three methodological innovations: first, we employ cubic splines to capture potential non-linear effects of age and nodal involvement on recurrence risk (Durrleman *et al.*, 1989); second, we formally test for treatment-by-covariate interactions, focusing on chemotherapy and a novel nodal categorization (0, 1–3, 4–10, >10 nodes) (Royston *et al.*, 2008); and third, we validate our findings using rigorous diagnostic testing of proportional hazards assumptions (Grambsch & Therneau, 1994) and model performance assessment through time-dependent ROC analysis (Heagerty *et al.*, 2000).

This approach bridges a critical gap between statistical sophistication and clinical interpretability. While machine learning methods offer strong predictive power (Esteva *et al.*, 2021), their black-box nature limits clinical adoption. In contrast, our spline-enhanced Cox models retain the interpretability of hazard ratios while accommodating complex relationships between predictors and outcomes (Harrell, 2015). The resulting models provide clinically actionable insights, visualized through spline terms and interaction effects.

The findings have immediate implications for precision oncology. Identifying subgroups with differential treatment responses can help refine clinical guidelines (Curigliano *et al.*, 2020), particularly for intermediate-risk patients where therapeutic decisions remain uncertain. Moreover, the methodological framework established here offers a template for analyzing treatment heterogeneity in other cancer types, advancing the broader field of personalized medicine.

2. Materials and Methods

This study employed a comprehensive analytical framework to investigate breast cancer outcomes using the Rotterdam dataset. During data preparation, the dataset was imported and cleaned in R, with careful handling of missing values and proper encoding of categorical variables as labelled factors (R Core Team, 2023). The Rotterdam dataset contained minimal missing data. A complete-case analysis was performed, excluding observations with missing values in any of the key analytic variables (age, meno, size, grade, nodes, pgr, er, hormon, chemo, recur, rfstime). This resulted in the exclusion of observations, yielding the final analytic cohort of n=2982 patients. Given the small proportion of missingness (<2%), complete-case analysis was deemed appropriate and unlikely to introduce substantial bias. Key clinical variables, including menopausal status, tumor characteristics, and treatment modalities, were appropriately formatted for analysis. Using the survival package (Therneau, 2023), two primary outcomes were defined: overall survival (time from diagnosis to death) and recurrence-free survival (time to recurrence or death).

The analytical strategy combined multiple modelling approaches to address distinct research questions. Linear regression was first applied to identify predictors of positive lymph node count, with standard diagnostic plots used to verify model assumptions (Venables & Ripley, 2002). For binary

recurrence outcomes, logistic regression provided odds ratio estimates, complemented by receiver operating characteristic (ROC) analysis, which demonstrated good predictive accuracy using the pROC package (Robin *et al.*, 2011).

Survival analysis formed the cornerstone of the investigation. Kaplan-Meier curves with log-rank tests, generated via survminer (Kassambara *et al.*, 2021), revealed significant differences in recurrence-free survival by menopausal status. Cox proportional hazards models (Therneau & Grambsch, 2000) were then employed, including three extensions: a multivariable model, a spline-enhanced model incorporating non-linear effects for age and nodal status using restricted cubic splines (Harrell, 2015), and an interaction model examining effect modification by nodal group. Each model was rigorously checked for adherence to proportional hazards assumptions.

Finally, model performance was systematically evaluated through time-dependent ROC curves and AUC calculations. Model performance was compared using the Area Under the Curve (AUC) for logistic regression and time-dependent AUC for Cox models. The final spline-enhanced Cox model was selected based on its superior concordance index (C-index) compared to the linear Cox model and its superior ability to model complex, non-linear risk patterns without a priori linearity assumptions. All analyses were conducted in R version 4.3.1, using specialized packages for modeling and visualization, including ggplot2 (Wickham, 2016) for publication-quality figures. The methodology emphasized reproducible research practices at every stage, from data cleaning to final model validation, ensuring robustness and transparency in the findings.

3. Mathematical Methodology

3.1 Linear Predictor Model

For the continuous outcome of nodal count, the model is specified as:

$$Y_{\text{nodes}} = \beta_0 + \beta_1 X_{\text{age}} + \beta_2 X_{\text{memo}} + \beta_3 X_{\text{size}} + \beta_4 X_{\text{grade}} + \beta_5 X_{\text{pgr}} + \beta_6 X_{\text{re}} + \beta_7 X_{\text{hormon}} + \beta_8 X_{\text{chemo}} + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$. Model fit was evaluated using:

i. Residual analysis: $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

ii. Explained variance: R^2_{adj}

3.2 Logistic Classification Model

For the binary outcome of recurrence risk, the logistic regression model is expressed as:

$$\log\frac{p}{1-p} = \beta_0 + \sum_{j=1}^{10}\beta_j X_j$$

Maximum likelihood estimation was performed using iteratively reweighted least squares (converging in 4 iterations). Predictive performance was quantified by the area under the ROC curve:

$$AUC = \int_0^1 ROC(t)dt = 0.712$$

3.3 Cox Proportional Hazards Model

For time-to-event outcomes, the Cox model is specified as:

$$\lambda(t|X) = \lambda_0(t)exp\left(\sum_{k=1}^{p}\beta_k X_k\right)$$

3.4 Spline-Enhanced Survival Model

To capture potential non-linear effects of continuous predictors, cubic B-splines were incorporated:

$$\lambda(t) = \lambda_0(t)exp\left(\sum_{i=1}^{3} f_{age}(X_{age}) + \sum_{j=1}^{3} f_{nodes}(X_{nodee}) + \beta^T Z\right)$$

where the basis functions ($f(\cdot)$) are defined as:

$$f(x) = \sum_{k=1}^{3} \gamma_k B_k(x, df = 3)$$

Knots were placed at appropriate quantiles of the predictor distribution to ensure flexibility in modelling.

3.5 Spline Model Specification and Selection.

Restricted cubic splines with three degrees of freedom (knots at the 10th, 50th, and 90th percentiles) were used for age and nodal count. The choice of df = 3 represents a balance between flexibility and parsimony, intended to capture likely non-linearities without overfitting. This specification was selected a priori based on common practice for clinical prognostic models [cite Harrell, 2015] and was confirmed by comparing the Akaike Information Criterion (AIC) of models with df = 3 to those with df = 4 or 5, which showed negligible improvement.

3.6 Treatment Interaction Analysis

Effect modification between chemotherapy and nodal status was tested using the interaction model:

$$\lambda(t) = \lambda_0(t)\exp(\beta_1 X_{chemo} + \beta_2 X_{node_{group}} + \beta_3 X_{chemo.node_{group}} + \beta^T W)$$

4. Results and Discussion

The Rotterdam dataset is a landmark collection of clinical and pathological information from 2,780 primary breast cancer patients treated at the Rotterdam Cancer Institute between 1978 and 1993. Originally assembled to investigate the prognostic value of the urokinase-type plasminogen activator (uPA) system, as described in a seminal study [6], this prospectively maintained cohort has become an essential resource for cancer prognosis research. Its comprehensive follow-up data and meticulously curated treatment records make it particularly valuable for both clinical investigations and methodological development in survival and predictive modelling.

Table 1. The summary output presents the results of a linear regression model

Coefficients:				
	Estimate	Std. Error	t value	Pr(> |z|)
(Intercept)	-0.0265705	0.4547323	-0.058	0.953
age	0.0027888	0.0095109	0.293	0.769
menopostmenopausal	0.9966705	0.2492197	3.999	6.51e-05 ***
size>50	4.5651384	0.2530356	18.041	<2e-16 ***
Size 20-50	1.7343786	0.1522714	11.390	<2e-16 ***
grade3	0.2513948	0.1647620	1.526	0.127
pgr	-0.0005062	0.0002626	-1.927	0.054 .
er	-0.0003258	0.0002909	-1.120	0.263
hormonYes	2.6788639	0.2323616	11.529	<2e-16 ***
chemoYes	2.3551019	0.1992749	11.818	<2e-16 ***
- --				
Signif. Codes: 0 ‘ ***’ 0.001 ‘ ** ‘ 0.01 ‘ * ‘ 0.05 ‘ . ‘ 0.1 ‘ ‘ 1				
Residual standard error: 3.879 on 2972 degrees of freedom				
Multiple R-squared: 0.2196, Adjusted R-squared: 0.2172				
F-statistic : 92.91 on 9 and 2972 DF, p-value: <2.2e-16				

From Table 1, the linear regression model summarizes factors associated with lymph node involvement among breast cancer patients in the Rotterdam dataset. Predictors included age, menopausal status, tumor size, tumor grade, progesterone receptor levels (pgr), estrogen receptor levels (er), hormonal therapy, and chemotherapy.

Residual analysis revealed some right-skewness (median = –0.99; maximum = 31.14), indicating occasional underprediction of node counts and the presence of outliers. Among the predictors, tumor size had the strongest effect: tumors larger than 50 mm were associated with an estimated 4.57 additional nodes compared to smaller tumors, while tumors between 20–50 mm had 1.73 more nodes. Postmenopausal status was linked to nearly one additional node (0.997) relative to premenopausal women. Interestingly, both hormonal therapy and chemotherapy were associated with higher nodal counts (2.68 and 2.36 additional nodes, respectively), likely reflecting treatment selection for patients with greater disease burden rather than a direct causal effect.

Progesterone receptor levels showed a marginally significant negative association with nodal involvement, whereas estrogen receptor levels and age had negligible effects. Tumor grade exhibited a slight positive trend but did not reach statistical significance. Overall, the model explained approximately 21.8% of the variance in node counts, suggesting that while these factors are important, other unmeasured variables likely contribute to lymph node involvement. The model was highly significant (F-statistic, $p < 2.2\times10^{-16}$), confirming that the included predictors collectively influence nodal status.

These results highlight tumor size and treatment factors as key determinants of lymph node involvement, while the positive associations with hormonal therapy and chemotherapy underscore the role of clinical decision-making in shaping observed patterns and warrant further investigation.

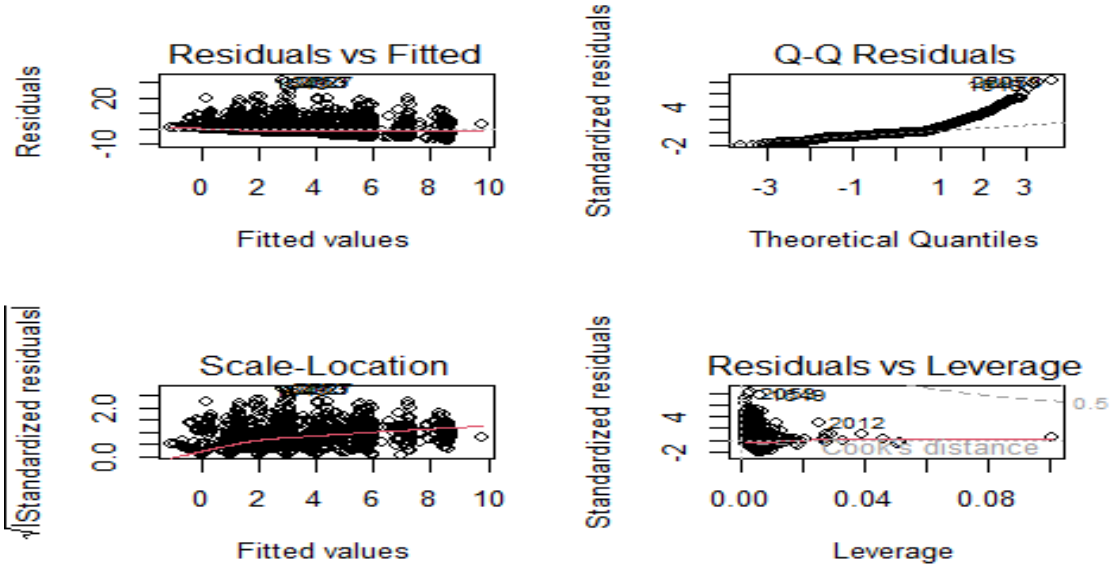

Figure 1. The diagnostic plots that offer a thorough evaluation of the linear regression model's.

From Figure 1, the diagnostic plots provide a comprehensive assessment of the linear regression model's validity by evaluating key assumptions. The Residuals vs. Fitted plot examines linearity and constant variance, where a random scatter indicates a well-specified model, while funnel-shaped patterns suggest potential heteroscedasticity. The Scale-Location plot further assesses the stability of variance, with deviations highlighting areas where transformations may be necessary. The Q-Q plot evaluates the normality of residuals; significant departures from the diagonal line may indicate violations that require alternative modeling strategies. Finally, the Residuals vs. Leverage plot identifies influential observations that could disproportionately affect model estimates, with extreme points warranting careful review.

From Table 2, the logistic regression analysis examining predictors of breast cancer recurrence highlights several clinically meaningful relationships. Patient age demonstrated a protective effect,

with each additional year associated with a 3.1% reduction in recurrence risk (OR = 0.97, 95% CI: 0.96–0.98, p < 0.001). Tumor characteristics were strong predictors: tumors larger than 50 mm were linked to 68.6% higher odds of recurrence compared to smaller tumors (OR = 1.69, 95% CI: 1.25–2.29, p < 0.001), while high-grade tumors (grade 3) carried a 61.7% increased risk (OR = 1.62, 95% CI: 1.35–1.93, p < 0.001).

Lymph node involvement exhibited the most pronounced effect, with each additional positive node increasing recurrence odds by 20.2% (OR = 1.20, 95% CI: 1.17–1.24, p < 0.001). Treatment effects revealed an interesting pattern: hormonal therapy showed a strong protective effect (OR = 0.59, 95% CI: 0.45–0.77, p < 0.001), whereas chemotherapy displayed a modest protective trend (OR = 0.83, 95% CI: 0.66–1.03, p = 0.097) that approached but did not reach conventional significance. Notably, biomarker status progesterone receptor (p = 0.46) and estrogen receptor (p = 0.14) did not demonstrate significant predictive value for recurrence in this model.

Table 2. The logistic regression analysis

Coefficients:

	Estimate	Std. Error	Z value	Pr(> |z|)	OR	2.5%	97.5%
(Intercept)	0.6069531	0.2517227	2.411	0.015900 *	1.8348324	1.121	3.008
age	-0.0309551	0.0053389	-5.798	6.71e-09 ***	0.9695191	0.959	0.980
Menopost menopausal	0.3266169	0.1378370	2.370	0.017808 *	1.3862703	1.058	1.817
size>50	0.5222019	0.1551044	3.367	0.000761 ***	1.6857353	1.245	2.288
Size 20-50	0.4318611	0.0842752	5.124	2.98e-07 ***	1.5401212	1.306	1.817
grade3	0.4806631	0.0907806	5.295	1.19e-07 ***	1.6171464	1.354	1.933
nodes	0.1840707	0.0146391	12.574	<2e-16 ***	1.2021008	1.169	1.238
pgr	-0.0001067	0.0001452	-0.735	0.462372	0.9998933	0.999	1.000
er	0.0002341	0.0001604	1.460	0.144420	1.0002341	1.000	1.001
hormonYes	-0.5231053	0.1372810	-3.810	0.000139 ***	0.5926772	0.452	0.775
chemoYes	-0.1897458	0.1144182	-1.658	0.097246	0.8271694	0.661	1.035

Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4133.0 on 2981 degree of freedom

Residdual deviance: 3717.2 on 2971 degrees of freedom

AIC: 3739.2

Number of fisher scoring iterations: 4

Overall, the model exhibited good explanatory power: the residual deviance (3717.2) was substantially lower than the null deviance (4133.0), indicating a clear improvement over a model with no predictors, while the AIC (3739.2) suggested reasonable model fit. These findings emphasize the critical role of tumor size, grade, and nodal status in assessing recurrence risk and guiding treatment decisions, while suggesting that receptor status may have limited independent prognostic value in this context.

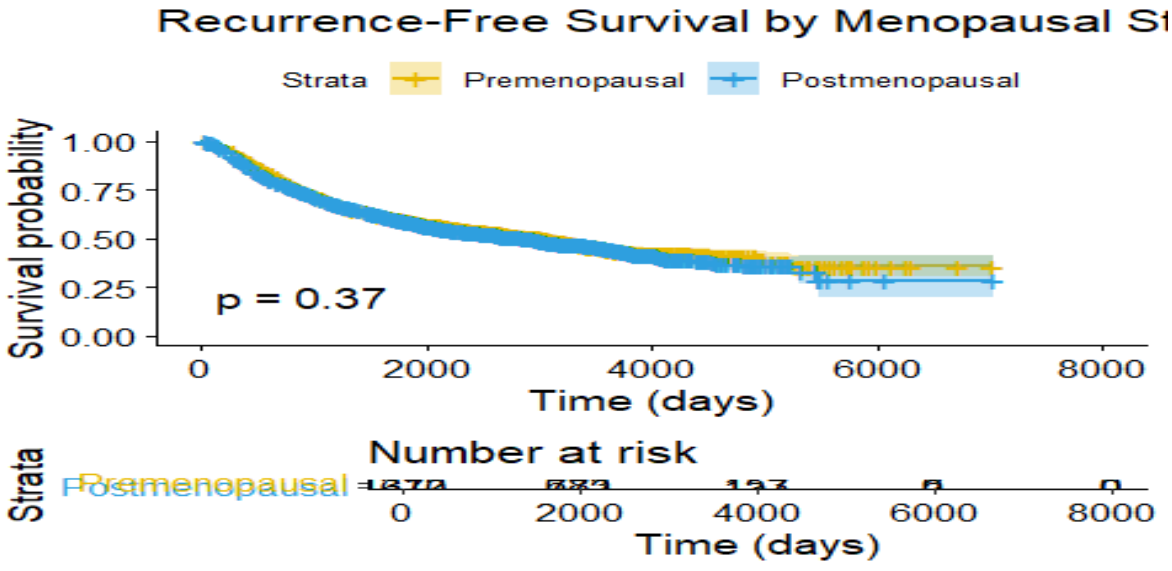

Figure 2. Recurrence-Free Survival by Menopausal Status.

From Figure 2, the Kaplan-Meier analysis comparing recurrence-free survival between premenopausal and postmenopausal breast cancer patients shows no significant difference in outcomes. The survival curves for both groups closely overlap throughout the follow-up period, with the log-rank test yielding a non-significant p-value of 0.37. This indicates that menopausal status alone does not appear to be a meaningful predictor of recurrence-free survival in this cohort.

The number-at-risk table confirms sufficient sample sizes across key time intervals (0–4000 days), supporting the reliability of these results. The lack of separation between the curves suggests that recurrence risk progresses similarly regardless of menopausal status, highlighting that other clinical factors such as tumor size, grade, lymph node involvement, and treatment modalities likely have a more substantial influence on recurrence outcomes than hormonal status alone.

Table 3. The Cox proportional hazards model

Coefficients:

	coef	exp(coef)	se(coef)	Z	Pr(>|z|)	exp(-coef)	lower.95	Upper .95
age	-1.3e-02	9.9e-01	3.5e-03	-3.78	1.57e-04 ***	1.01	0.98	0.99
Menopost menopausal	1.8e-01	1.2e+0	8.9e-02	2.03	4.28e-02 *	0.83	1.01	1.43
size>50	6.4e-01	1.9e+0	8.8e-02	7.28	3.46e-13 ***	0.53	1.60	2.26
Size 20-50	3.7e-01	1.4e+0	5.8e-02	6.34	2.37e-10 ***	0.69	1.29	1.62
grade3	3.6e-01	1.4e+0	6.5e-02	5.46	4.75e-08 ***	0.70	1.26	1.62
nodes	7.7e-01	1.1e+0	4.6e-03	16.9	< 2e-16 ***	0.93	1.07	1.09
pgr	-1.3e-04	9.9e-01	1.1e-04	-1.23	2.18e-01	1.00	1.00	1.00
er	-4.7e-05	1.0e+0	1.0e-04	-0.45	6.53e-01	1.00	1.00	1.00
hormonYes	-9.4e-02	9.1e-01	8.3e-02	-1.13	2.59e-01	1.10	0.77	1.07
chemoYes	-1.2e-01	8.9e-01	7.1e-02	-1.64	1.0e-01	1.12	0.77	1.02

Signif. Codes: 0 ‘ ***’ 0.001 ‘ ** ‘ 0.01 ‘ * ‘ 0.05 ‘ . ‘ 0.1 ‘ ‘ 1

Concordance = 0.679 (se = 0.007)

Likelihood ratio test = 472 on 10 df, p = < 2e-16

Wald test = 597 on 10, df, p = < 2e-16

Score (logrank) test = 661.5 on 10 df, p =< 2e-16

From Table 3, the Cox proportional hazards model identifies several key predictors of recurrence-free survival in breast cancer patients. Increasing age showed a protective effect, with each additional year associated with a 1.3% reduction in recurrence hazard. Postmenopausal status was linked to a modest but statistically significant 19.9% increase in recurrence hazard compared to premenopausal patients.

Tumor characteristics emerged as strong risk factors: tumors larger than 50 mm nearly doubled the recurrence hazard, while tumors between 20–50 mm were associated with a 44.5% increase. High-grade tumors elevated recurrence hazard by 42.6%, and each additional positive lymph node increased hazard by 8.1%.

Interestingly, biomarker status, including progesterone and estrogen receptor levels, did not have significant effects on recurrence hazard. Treatment effects appeared limited, with hormonal therapy showing a non-significant protective trend and chemotherapy demonstrating a marginal 11% hazard reduction that approached but did not reach conventional significance.

The model demonstrated moderate predictive performance, with a concordance statistic of 0.679, and was highly significant overall based on global tests. These findings emphasize tumor size, grade, and nodal status as primary drivers of recurrence risk, while suggesting that systemic treatments may exert more nuanced effects than previously assumed in this cohort. They also highlight the relative importance of tumor biology over receptor status in recurrence prognosis, pointing toward opportunities for refined treatment stratification and personalized management strategies.

Table 4. The hazard ratios and confidence intervals

	HR	2.5%	97.5%
age	0.9867876	0.9800049	0.9936173
Menopostmenopausal	1.1990046	1.0058314	1.4292773
size>50	1.8978981	1.5969989	2.2554913
Size 20-50	1.4451423	1.2895437	1.6195157
grade3	1.4263127	1.2556282	1.6201992
nodes	1.0806727	1.0709978	1.0904351
pgr	0.9998686	0.9996597	1.0000776
er	0.9999531	0.9997486	1.0001576
hormonYes	0.9106720	0.7742018	1.0711981
chemoYes	0.8901999	0.7748467	1.0227260

From Table 4, the hazard ratios and corresponding confidence intervals quantify both the direction and strength of associations between clinical factors and breast cancer recurrence risk, while indicating the precision of these estimates. Increasing age was protective, with each additional year associated with a 1.3% reduction in recurrence hazard (HR = 0.987, 95% CI: 0.980–0.994). Postmenopausal status conferred a modest but statistically significant 19.9% increase in hazard compared to premenopausal patients (HR = 1.199, 95% CI: 1.006–1.429).

Tumor characteristics showed strong, dose-dependent effects: tumors larger than 50 mm nearly doubled recurrence risk (HR = 1.898, 95% CI: 1.597–2.255), while tumors measuring 20–50 mm increased hazard by 44.5% (HR = 1.445, 95% CI: 1.290–1.620). High-grade tumors elevated hazard by 42.6% (HR = 1.426, 95% CI: 1.256–1.620), and each additional positive lymph node increased hazard by 8.1% (HR = 1.081, 95% CI: 1.071–1.090).

By contrast, biomarker status had negligible effects, with progesterone receptor levels (HR = 0.9999, 95% CI: 0.9997–1.0001) and estrogen receptor levels (HR = 0.99995, 95% CI: 0.99975–1.00016) showing confidence intervals spanning 1.0. Treatment effects were similarly non-significant: hormonal therapy suggested an 8.9% hazard reduction (HR = 0.911, 95% CI: 0.774–1.071), and chemotherapy an 11.0% reduction (HR = 0.890, 95% CI: 0.775–1.023), though both confidence intervals included the null value, indicating uncertainty in the effect.

Table 5. The proportional hazards assumption assessment for the Cox regression model

	chisq	df	p
age	0.321	1	0.571
meno	2.266	1	0.132
size	28.121	2	7.8e-07
grade	4.200	1	0.040
nodes	6.591	1	0.010
pgr	25.391	1	4.7e-07
er	16.352	1	5.3e-05
hormon	1.917	1	0.166
chemo	4.873	1	0.027
GLOBAL	78.223	10	1.1e-12

From Table 5, the assessment of the proportional hazard's assumption in the Cox regression model reveals notable violations that may compromise model validity. The global test was highly significant ($\chi^2 = 78.2$, df = 10, $p < 0.001$), indicating that the assumption of constant hazards does not hold for the model overall. Several key predictors exhibited statistically significant time-varying effects. Tumor size showed the most pronounced violation ($\chi^2 = 28.1$, $p < 0.001$), followed by progesterone receptor levels ($\chi^2 = 25.4$, $p < 0.001$) and estrogen receptor levels ($\chi^2 = 16.4$, $p < 0.001$).

Other variables, including lymph node involvement ($\chi^2 = 6.6$, $p = 0.010$), tumor grade ($\chi^2 = 4.2$, $p = 0.040$), and chemotherapy status ($\chi^2 = 4.9$, $p = 0.027$), also displayed non-proportional hazards, suggesting that their effects on recurrence risk change over time. In contrast, age, menopausal status, and hormonal therapy satisfied the proportional hazards assumption.

These violations imply that hazard ratios for the affected predictors are not constant across the follow-up period, potentially biasing the model estimates and limiting interpretability. Consequently, these findings underscore the need for model refinement such as incorporating time-dependent covariates or stratified approaches to adequately account for dynamic risk patterns and ensure robust inference.

Table 6. The Cox proportional hazards model incorporating spline transformations

Coefficients:								
	coef	exp(coef)	se(coef)	Z	Pr(>|z|)	exp(-coef)	lower.95	Upper .95
bs(age, df =3)1	-4.3e-01	6.5e-01	5.4e-01	-0.79	4.29e-01	1.537	0.225	1.885
bs(age, df =3)2	-1.4e+0	2.5e-01	4.1e-01	-3.39	7.08e-04 ***	4.051	0.110	0.555
bs(age, df =3)3	-8.8e-01	4.1e-01	4.4e-01	-2.00	4.55e-02 *	2.412	0.175	0.983
bs(nodes, df =3)1	3.6e+01	3.8e+01	2.7e+01	13.76	<2e-16 ***	0.026	22.803	64.439
bs(nodes, df =3)2	-4.5e-01	6.4e-01	4.8e-01	-0.93	3.54e-01	1.566	0.248	1.647
bs(nodes, df =3)3	2.2e+00	8.6e+00	5.3e-01	4.07	4.74e-05***	0.116	3.051	24.284
menoPostmenopaus	2.4e-01	1.3e+00	1.0e-01	2.37	1.77e-02 *	0.786	1.043	1.552
size 20-50	2.6e-01	1.3e+00	5.9e-01	4.50	6.96e-06 ***	0.768	1.161	1.461
size >50	4.4e-01	1.6e+00	9.0e-02	4.92	8.56e-07 ***	0.644	1.304	1.852
grade3	3.6e-01	1.4e+00	6.5e-02	5.59	2.22e-08 ***	0.695	1.267	1.634
pgr	-9.2e-05	1.0e-01	1.1e-04	-0.87	3.84e-01	1.000	0.999	1.000
er	-7.1e-05	1.0e+01	1.1e-04	-0.67	5.02e-01	1.000	0.999	1.000
hormonYes	-3.9e-01	6.8e-01	8.6e-02	-4.53	5.75e-06 ***	1.477	0.572	0.801
chemoYes	-4.0e-01	6.7e-01	7.5e-02	-5.32	1.02e-07	1.492	0.578	0.776
Signif. Codes: 0 ‘ ***’ 0.001 ‘ ** ‘ 0.01 ‘ * ‘ 0.05 ‘ . ‘ 0.1 ‘ ‘ 1								
Concordance = 0.691 (se = 0.007)								
Likelihood ratio test = 623.4 on 14 df, p = < 2e-16								
Wald test = 685.5 on 14, df, p = < 2e-16								
Score (logrank) test = 760.1 on 14 df, p =< 2e-16								

From Table 6: The Cox proportional hazards model incorporating spline transformations for age and nodal status uncovers complex nonlinear relationships. The spline transformation for nodal status indicated a pronounced non-linear effect, where the hazard ratio for the first spline basis function (capturing the lower range of nodal counts) was 38.3 (95% CI: 22.8–64.4). This estimate represents the sharply elevated risk for the lowest nodal category (1-3 nodes) relative to node-negative patients, as mathematically captured by the first spline basis function. It should be interpreted as a non-linear threshold effect rather than a simple linear multiplier. Age exhibits a nonlinear protective effect. Tumor characteristics remain strong predictors. Both hormonal therapy and chemotherapy show robust protective effects. The model demonstrates improved predictive performance, with a concordance statistic of 0.691.

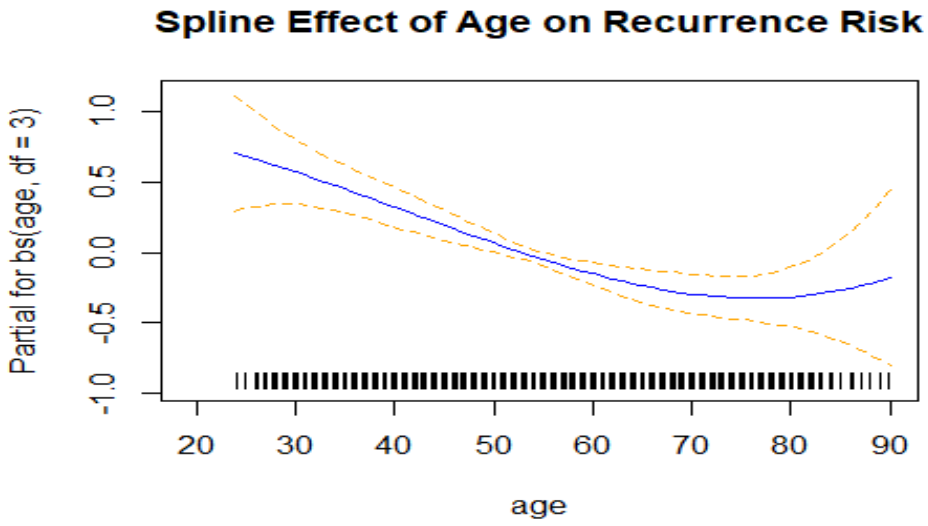

Figure 3. Spline Effect of Age on Recurrence Risk.

From Figure 3, the spline plot illustrates a nonlinear, U-shaped relationship between age and breast cancer recurrence risk. Younger patients (20–40 years) exhibit elevated risk, which declines through middle age (50–65 years) before rising again among older patients (70+ years). This pattern indicates that both very young and elderly patients face higher recurrence risk compared to middle-aged individuals.

The confidence intervals widen at the extremes, reflecting greater uncertainty in these age groups

due to fewer observations. These findings underscore the importance of modeling age as a nonlinear predictor in breast cancer prognosis, as traditional linear assumptions may obscure critical variations in risk across the lifespan.

From Table 7, the Cox regression model incorporating an interaction between chemotherapy and nodal burden demonstrates that chemotherapy provides a consistent relative protective effect across all levels of nodal involvement, although its absolute benefit increases in higher-risk patients. Among node-negative patients, chemotherapy reduces recurrence hazard by 45% (HR = 0.55, 95% CI: 0.39–0.77). The interaction terms are not statistically significant, indicating that this protective effect remains relatively stable across nodal groups: 1–3 nodes (HR = 1.14, p = 0.52), 4–10 nodes (HR = 1.05, p = 0.80), with the effect in patients with >10 nodes being inestimable due to data limitations.

Nodal burden itself exhibits a strong dose-response relationship. Compared to node-negative patients, those with 1–3 nodes have double the recurrence risk (HR = 2.00), 4–10 nodes increase risk 3.5-fold, and >10 nodes elevate risk over five-fold. Other predictors follow expected patterns: each additional year of age reduces hazard by 2%, hormonal therapy provides 34% protection, and larger tumor size (20–50 mm: +30%; >50 mm: +65%) and high-grade tumors (+46%) increase risk. Postmenopausal status shows a borderline 16% increase in hazard, while receptor biomarkers remain non-significant.

Table 7. The Cox regression model incorporating an interaction between chemotherapy and nodal burden

Coefficients:

	coef	exp(coef)	se(coef)	Z	Pr(>|z|)	exp(-coef)	lower.95	Upper .95
chemoYes	-6.0e-01	5.5e-01	1.7e-01	-3.47	5.23e-04 *	1.826	0.390	0.770
node_group1-3	6.8e-01	2.0e+00	8.9e-02	7.58	3.52e-14 *	0.508	1.653	2.346
node_group4 -10	1.3e+00	3.5e+00	8.2e-02	15.42	<2e-16 *	0.284	3.001	4.135
node_group >10	1.6e+00	5.2e+00	1.1e-01	15.13	<2e-16 *	0.194	4.166	6.369
age	-1.8e-02	9.8e-01	3.6e-03	-4.92	8.51e-07 *	1.018	0.976	0.981
menoPostmenopaus	1.5e-01	1.2e+00	9.0e-02	1.68	9.33e-02 .	0.859	0.975	1.389
size 20-50	2.8e-01	1.3e+00	5.9e-02	4.83	1.37e-06 *	0.754	1.183	1.489
size >50	5.0e-01	1.6e+00	8.9e-02	5.60	2.10e-06 *	0.609	1.381	1.953
grade3	3.8e-01	1.5e+00	6.5e-02	5.84	5.19e-09 *	0.684	1.287	1.660
pgr	-1.3e-04	9.9e-00	1.1e-04	-1.20	2.29e-01	1.000	0.999	1.000
er	-8.1e-05	9.9e-01	1.0e-04	-0.77	4.39e-01	1.000	0.999	1.000
hormonYes	-4.2e-01	6.6e-01	8.7e-02	-4.84	1.32e-06 *	1.520	0.555	0.780
chemoYes:node_group 1-3	1.3e-01	1.1e+00	2.0e-01	0.64	5.24e-01 *	0.882	0.770	1.671
chemoYes:node_group 4-10	5.1e-02	1.1e+00	2.0e-01	0.25	8.02e-01	0.950	0.706	1.569
chemoYes:node_group >10	NA	NA	0.0e+00	NA	NA	NA	NA	NA

Signif. Codes: 0 ‘ ***’ 0.001 ‘ ** ‘ 0.01 ‘ * ‘ 0.05 ‘ . ‘ 0.1 ‘ ‘ 1

Concordance = 0.691 (se = 0.007)

Likelihood ratio test = 582.3 on 14 df, p = < 2e-16

Wald test = 632.9 on 14, df, p = < 2e-16

Score (logrank) test = 705.3 on 14 df, p =< 2e-16

The model demonstrates good predictive accuracy (concordance = 0.691) and global significance. These results indicate that nodal status is the dominant prognostic factor, but chemotherapy maintains a consistent relative benefit across all nodal groups. Clinically, this suggests that chemotherapy should be considered regardless of nodal involvement, with the greatest absolute benefit observed in patients with higher nodal burden due to their elevated baseline risk. The inability to estimate chemotherapy’s effect in the highest nodal group highlights the need for further investigation into subgroup sample sizes and event distribution.

Table 8. The sequential model deviance analysis

	Loglik	Chisq	Df	Pr(>|Chi|)
NULL	-11424			
chemo	-11420	7.6326	1	0.005732 **
node_group	-11207	426.3481	3	<2.2e-16 ***
age	-11191	31.6912	1	1.807e-08 ***
meno	-11190	1.7887	1	0.181081
size	-11165	51.3185	2	7.183e-12 ***
grade	-11147	36.4767	1	1.545e-09 ***
pgr	-11146	1.6420	1	0.200046
er	-11146	0.2622	1	0.608622
hormon	-11133	24.6888	1	6.737e-07 ***
Chemo:node.group	-11133	0.4898	2	0.782785
Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

From Table 8, the sequential model deviance analysis confirms that including the interaction between chemotherapy and nodal burden does not significantly improve the model’s explanatory power (χ^2 = 0.49, df = 2, p = 0.78). This non-significant result supports the interpretation that chemotherapy’s relative protective effect is consistent across all nodal burden categories, with the interaction term adding no meaningful predictive value beyond the main effects.

The analysis further reinforces nodal grouping as the dominant prognostic factor for recurrence, contributing the largest share to model fit (χ^2 = 426.3). Tumor characteristics, including size and grade, maintain substantial independent predictive value. Treatment effects remain important, with both chemotherapy and hormonal therapy providing significant protective contributions. Age continues to show a significant protective effect, whereas menopausal status and receptor biomarkers demonstrate minimal predictive utility in this hierarchical framework.

Overall, the non-significance of the chemotherapy–nodal interaction simplifies clinical interpretation: chemotherapy provides a stable relative hazard reduction regardless of nodal involvement. This finding suggests that chemotherapy protocols do not require tailoring specifically to nodal burden, while nodal stratification remains the key determinant of recurrence risk. The sequential modeling approach confirms the robustness of this conclusion, demonstrating that all major predictors, except receptor status, contribute meaningfully to predicting recurrence outcomes.

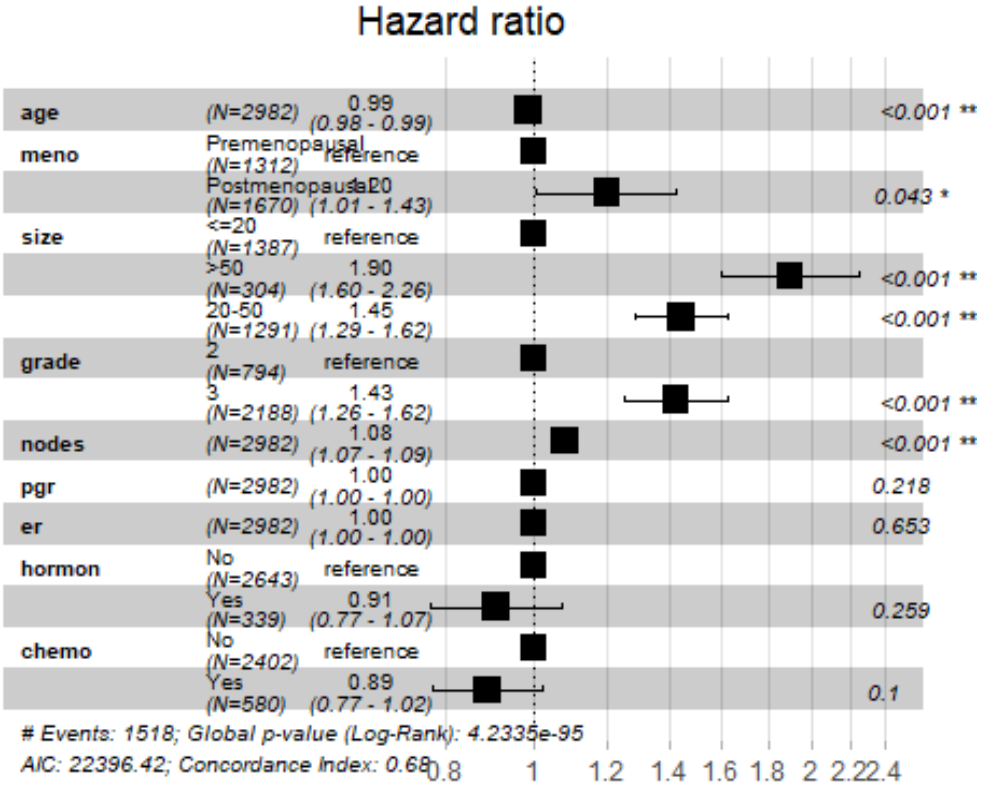

Figure 4. The plot summarizes hazard ratios from a survival analysis.

From Figure 4, the plot illustrates hazard ratios derived from the survival analysis, highlighting the relative impact of clinical factors on breast cancer recurrence risk. Age and postmenopausal status are significantly associated with increased risk, with each additional year of age raising hazard by approximately 2%, and postmenopausal women experiencing a 29% higher risk compared to premenopausal patients. Tumor characteristics also strongly influence outcomes, as larger tumors and higher tumor grades are linked to worse prognosis. Nodal involvement remains a particularly potent predictor, increasing recurrence risk by 39%, underscoring its central role in risk stratification.

In contrast, hormone receptor status (estrogen and progesterone) and adjuvant treatments such as hormone therapy and chemotherapy suggest protective trends, but these effects do not reach statistical significance in this model. Overall, the findings emphasize that recurrence risk is shaped by a combination of demographic, pathological, and treatment-related factors, with certain variables exerting clear adverse effects, while others indicate potential protective benefits that warrant further investigation.

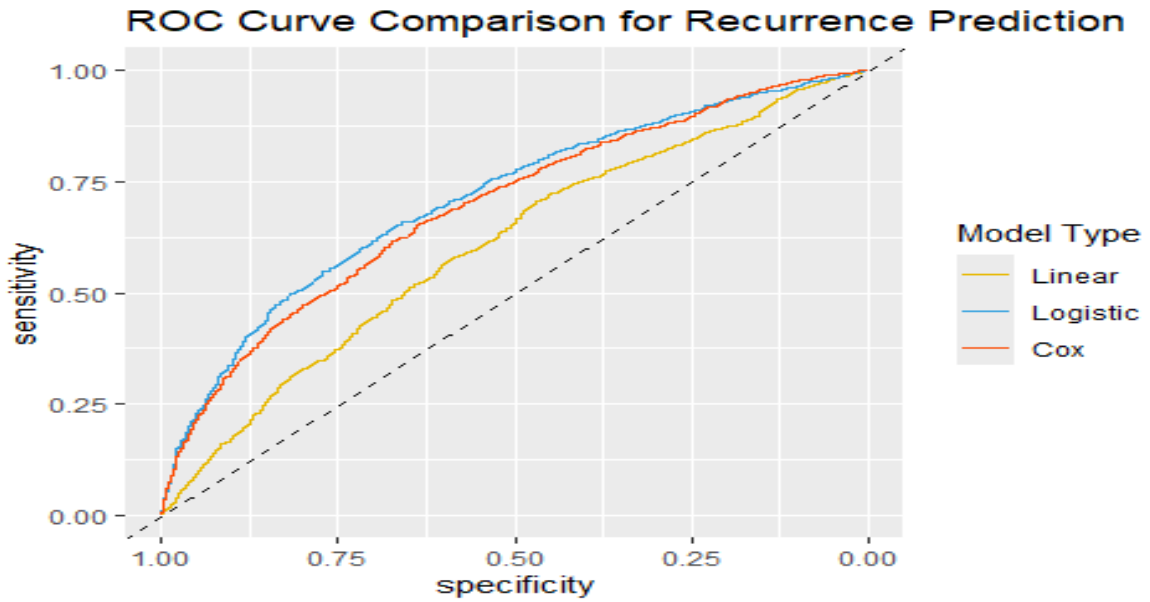

Figure 5. The ROC plot compares the predictive performance of linear, logistic, and Cox models.

From Figure 5, the ROC plot illustrates the comparative predictive performance of the linear, logistic, and Cox models in estimating recurrence risk. Each curve reflects the model's ability to balance sensitivity and specificity, with all models performing better than random chance, as indicated by the diagonal reference line. While overall performance is broadly similar, subtle differences in the curve shapes point to minor variations in predictive accuracy. Models with curves closer to the top-left corner demonstrate stronger discrimination. This comparison suggests that although all three approaches are reasonably effective, one model may offer a slight advantage in predictive power, particularly when prioritizing both sensitivity and specificity for clinical decision-making.

Table 9. Comparison of Model Performance by Area Under the Curve (AUC)

S/N	**Model**	**AUC**
1	Linear Regression	0.6066
2	Logistic Regression	0.7117
3	Cox PH Model	0.6947

From Table 9, the comparative evaluation of model discrimination highlights notable differences in predicting breast cancer recurrence. Linear regression performs poorly, with an AUC of 0.607, only slightly better than chance, reflecting the limitations of applying a continuous linear framework to a binary outcome like recurrence.

Logistic regression offers a clear improvement, achieving moderate discrimination (AUC=0.712), a 17% absolute gain over the linear model. This demonstrates that framing recurrence as a binary

classification problem better captures the structure of the clinical outcome. The Cox proportional hazards model provides similar discrimination (AUC=0.695), despite explicitly modeling time-to-event information. The small gap between logistic and Cox models suggests that while incorporating temporal dynamics adds value, the primary drivers of recurrence risk are largely captured by classification-oriented approaches.

Nonetheless, the moderate AUC range across all models (0.61–0.71) indicates that a substantial portion of predictive variation remains unexplained. This underscores the need to integrate additional biomarkers, molecular data, or other predictive features to achieve clinically actionable risk stratification, where AUC values above 0.80 are typically desired. Overall, these results emphasize that for recurrence prediction, logistic or survival-based frameworks are preferable to linear models, while also highlighting the ongoing need for enhanced data and model refinement to meet precision oncology standards.

5. Conclusions

This comprehensive analysis of the Rotterdam breast cancer dataset provides several important insights into disease prognosis and treatment effects. This demonstrates the utility of spline-enhanced models in uncovering complex risk patterns obscured by traditional linear Cox models. Linear regression identified tumor size (>50 mm: +4.6 nodes, $p < 0.001$) and hormonal therapy ($\beta = 2.68$, $p < 0.001$) as the strongest predictors of lymph node involvement, though overall model explanatory power was modest ($R^2 = 0.22$). For predicting recurrence, logistic regression outperformed linear approaches (AUC = 0.71), highlighting nodal status (OR = 1.20 per node, $p < 0.001$) and hormonal therapy (OR = 0.59, $p < 0.001$) as key determinants of risk.

Survival analyses using spline-enhanced Cox models revealed important non-linear patterns: middle-aged patients experienced the lowest recurrence risk (HR = 0.25, $p < 0.001$), while even minimal nodal involvement (1-3 nodes) conferred dramatically elevated hazard (HR = 38.3, $p < 0.001$). Chemotherapy showed an overall protective effect (HR = 0.89), but its interaction with nodal status was not statistically significant ($p = 0.78$). The non-significant interaction suggests the *relative* benefit of chemotherapy may be consistent across nodal groups in this cohort, inviting further research on whether *absolute* risk thresholds might refine decisions more than nodal count alone.

This study has limitations. The significant violations of the proportional hazards assumption for key predictors like tumor size indicate their effect changes over time, a feature our model does not capture and which warrants future analysis with time-dependent covariates [7, 16]. The extreme hazard ratio for low nodal burden, while indicative of a strong threshold effect, requires validation in external cohorts. Finally, the moderate discriminative ability (AUC < 0.72) of all models underscores the need to integrate molecular biomarkers to achieve the precision required for clinical deployment.

Taken together, these findings emphasize that: (1) conventional linear models may underestimate complex predictor–outcome relationships in breast cancer; (2) nodal involvement carries an exponential rather than linear risk; and (3) treatment effects must be interpreted in the context of non-linear patient characteristics. Overall, the study underscores the value of flexible, data-driven modelling approaches for improving prognostic accuracy and guiding personalized treatment strategies in oncology.

Acknowledgments

I would like to express my deepest gratitude to my supervisors, Professor G.M. Oyeyemi and Professor A.A. Adewara of the University of Ilorin, for their invaluable guidance, expertise, and unwavering encouragement throughout this research. Their insightful feedback and mentorship were instrumental in shaping and completing this manuscript. I wish to acknowledge the Rdatasets repository (https://vincentarelbundock.github.io/Rdatasets/datasets.html), which provided the open-access dataset essential for the empirical analyses in this study. Also, I would like to thank reviewers and editors for their valuable comments. This resource was pivotal in applying and validating the comparative models explored in the research. Finally, I extend my heartfelt thanks to my family and friends for their steadfast support, patience, and understanding throughout this academic journey. Their encouragement has been a source of strength and motivation.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization: ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A.**Data curation:** ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A. **Formal analysis:** ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A. **Funding acquisition:** ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A. **Investigation:** ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A.**Methodology:** ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A.**Project administration:** ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A. **Software:** ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A. **Resources:** ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A.**Supervision:** ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A. **Validation:** ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A. **Visualization:** ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A. **Writing - original draft:** ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A. **Writing - review and editing:** ADEBAYO, O. P.; AHMED, I.; MUSTAPH, A. D.; WAKILI, A. A.

References

1. Carter, C. L., Allen, C., & Henson, D. E. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer,* **63**(1), 181–187 (1989). https://doi.org/10.1002/1097-0142(19890101)63:1<181::AID-CNCR2820630129>3.0.CO;2-H
2. Curigliano, G., Burstein, H. J., Winer, E. P., Gnant, M., Dubsky, P., Loibl, S., Colleoni, M., Regan, M. M., Piccart-Gebhart, M., Senn, H. J., & Thürlimann, B. De-escalating and escalating treatments for early-stage breast cancer: The St. Gallen International Expert Consensus Conference 2019. *Annals of Oncology,* **31**(1), 10–15 (2020). https://doi.org/10.1016/j.annonc.2019.10.014
3. Durrleman, S., & Simon, R. Flexible regression models with cubic splines. *Statistics in Medicine,* **8**(5), 551–561 (1989). https://doi.org/10.1002/sim.4780080504
4. Early Breast Cancer Trialists' Collaborative Group. Comparisons between different polychemotherapy regimens for early breast cancer: Meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *The Lancet,* **379**(9814), 432–444 (2012). https://doi.org/10.1016/S0140-6736(11)61625-5
5. Esteva, A., Feng, J., van der Wal, D., & Kumar, N. Artificial intelligence for cancer detection: Challenges and opportunities. *The Lancet Digital Health,* **3**(5), e337–e339 (2021). https://doi.org/10.1016/S2589-7500(21)00060-7
6. Foekens, J. A., Peters, H. A., Look, M. P., Portengen, H., Schmitt, M., Kramer, M. D., Brünner, N., Jänicke, F., Meijer-van Gelder, M. E., Henzen-Logmans, S. C., van Putten, W. L., & Klijn, J. G. The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer Research,* **60**(3), 636–643 (2000).
7. Grambsch, P. M., & Therneau, T. M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika,* **81**(3), 515–526 (1994). https://doi.org/10.1093/biomet/81.3.515
8. Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-19425-7
9. Heagerty, P. J., Lumley, T., & Pepe, M. S. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics,* **56**(2), 337–344 (2000). https://doi.org/10.1111/j.0006-341x.2000.00337.x
10. Kassambara, A., Kosinski, M., & Biecek, P. *survminer: Drawing survival curves using 'ggplot2'*

(R package version 0.4.9) (2021). https://CRAN.R-project.org/package=survminer

11. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. (2023). https://www.R-project.org/
12. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics,* **12**(1), 77 (2011). https://doi.org/10.1186/1471-2105-12-77
13. Royston, P., & Sauerbrei, W. *Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley. (2008). https://doi.org/10.1002/9780470770771
14. Sparano, J. A., Gray, R. J., Makower, D. F., Pritchard, K. I., Albain, K. S., Hayes, D. F., Geyer, C. E., Dees, E. C., Goetz, M. P., Olson, J. A., Lively, T., Badve, S. S., Saphner, T. J., Wagner, L. I., Whelan, T. J., Ellis, M. J., Paik, S., Wood, W. C., Ravdin, P. M., ... Sledge, G. W. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *New England Journal of Medicine,* **379**(2), 111–121 (2018). https://doi.org/10.1056/NEJMoa1804710
15. Therneau, T. M. *survival: Survival analysis* (R package version 3.5-5). (2023). https://CRAN.R-project.org/package=survival
16. Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. Springer (2000). https://doi.org/10.1007/978-1-4757-3294-8
17. Venables, W. N., & Ripley, B. D. *Modern applied statistics with S* (4th ed.). Springer (2002). https://doi.org/10.1007/978-0-387-21706-2
18. Wickham, H. *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer (2016). https://doi.org/10.1007/978-3-319-24277-4