DOI: https://doi.org/10.54393/pbmj.v9i6.1393

PAKISTAN BIOMEDICAL JOURNAL
https://www.pakistanbmj.com/journal/index.php/pbmj/index
ISSN (E): 2709-2798, (P): 2709-278X
Volume 9, Issue 06 (June 2026)

Original Article

Clinicopathological Determinants of Overall Survival in Breast Cancer: A Retrospective Cohort Analysis

Warda Munir[1*] and Memoona Anees[2]

[1]Department of Public Health, University of the Punjab, Lahore, Pakistan
[2]Department of Family Medicine, Madera Family Medicine Group, California, United States of America

ARTICLE INFO

Keywords:
Breast Neoplasms, Lymph Node Ratio, Prognosis, Survival Analysis, Risk Assessment

How to Cite:
Munir, W., & Anees, M. (2026). Clinicopathological Determinants of Overall Survival in Breast Cancer: A Retrospective Cohort Analysis: Clinicopathological Determinants of Overall Survival in Breast Cancer. Pakistan BioMedical Journal, 9(6), 22-28. https://doi.org/10.54393/pbmj.v9i6.1393

***Corresponding Author:**
Warda Munir
Department of Public Health, University of the Punjab, Lahore, Pakistan
wardaatif2627@gmail.com

Received Date: 13th March, 2026
1st Revision Date: 9th June, 2026
Acceptance Date: 18th June, 2026
Published Date: 30th June, 2026

ABSTRACT

Breast cancer is the most common cause of cancer mortality in women, and the survival rate depends on a multitude of demographic and tumor-related factors. Personalizing surveillance intensity and the choice of treatment depends on accurate risk stratification. **Objectives:** To determine the main patient and tumor parameters associated with overall survival (OS) in breast cancer patients by univariate and multivariable Cox proportional hazards analyses. **Methods:** This is a retrospective cohort study of 1,000 breast cancer patients in the Kaggle extract of SEER data SEER database. Univariate and multivariable Cox regression analyses were used to identify independent predictors of survival. Three predictive models were built and validated on a held-out test set (30% of data): Cox regression, Random Survival Forest (RSF) and XGBoost. Clinicopathological parameters affecting overall survival (OS) were identified, and two machine learning algorithms were developed to create a practical predictive risk model for 5-year mortality. **Results:** The independent prognostic parameters for survival were lymph node ratio (HR 2.27 per 0.1 increase), advanced stage (Stage IIIC vs. IIA: HR 3.35), tumor size (HR 1.79 per log unit), older age (HR 1.42 per decade), ER negative status (HR 1.67), and high histological grade (HR 1.46). Marital status (and race) was not significant after adjustment. The RSF model showed the best performance (C index = 0.758, AUC ROC = 0.78, calibration slope = 0.98). **Conclusions:** The most important independent factors in the survival of BC are LNR, stage, tumor size, age, and ER status, whereas demographic factors play a minimal role after these factors have been controlled for.

INTRODUCTION

Breast cancer is the most common cancer and leading cause of cancer-related deaths in females globally, and has been estimated to cause over 2.3 million new cases and over 680,000 deaths annually [1]. Although great progress has been made in early detection, molecular subtyping and targeted therapies, the clinical course of breast cancer patients is extremely variable, with a complex interplay between tumor-related factors, treatment-related factors and patients' demographics [2]. Currently, the prognostic tools available are, however, incapable of taking into account all the clinicopathological aspects influencing the individual long-term prognosis [3]. Therefore, the need for the improvement of the current risk assessment systems using comprehensive and population-based information regarding current clinical practice is great [4]. The clinicopathological parameters associated with breast cancer prognosis have been the subject of numerous studies, and tumor stage, histological grade, hormone receptor expression, and lymph node involvement have been identified as important prognostic factors for breast cancer survival [5]. The lymph node ratio (LNR), the ratio of the number of positive nodes to the number of nodes examined, has proven to be one of the most powerful prognostic factors in nodal metrics, and it brings into consideration tumor burden and surgical sampling [6]. Population-based studies using the Surveillance, Epidemiology, and End Results (SEER) database have yielded great insight into disparities in survival across various patient populations and have found that age, race, marital status, and molecular subtypes are other modifiers

Copyright© 2026 The Authors. Published by CrossLinks International Publishers LLC, USA
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

DOI: https://doi.org/10.54393/pbmj.v9i6.1393

of prognosis [7, 8]. However, the relative importance of these factors, as assessed together in the same multivariate context, still has not been well characterized, particularly in more modern populations of patients treated with modern treatment paradigms [9]. In the past few years, some methodological advances have appeared that exploit machine learning tools such as Random Survival Forests (RSF) and XGBoost, which can capture the benefits of non-linear relationships between prognostic variables and complex interactions, compared to Cox proportional hazards models [10, 11]. A critical need is to translate complex models to simpler, actionable risk groups for clinical practice [12]. In the context of breast cancer, RSF has been shown to outperform traditional regression methods with C indices of 0.72 to 0.76, potentially offering new insights into risk stratification and clinical practice. While all these advances are encouraging, there is a lag between the use of machine learning-derived risk models in clinical practice, partly because there is no external validation of the tools that are practical to use in everyday practice [13]. The major problem with many of the algorithms currently available is that the resulting risk scores are confusing and hard to interpret or use when treating a patient at the bedside. Thus, a great need in breast cancer survivorship care is to convey complex models to simple, actionable risk groups [14].

Although demographic, clinicopathological, and nodal variables are individually important in breast cancer prognosis, few studies have incorporated these factors simultaneously in a single analytical model. This study aimed to determine the main patient and tumor parameters associated with overall survival (OS) in breast cancer patients by univariate and multivariable Cox proportional hazards analyses. It also sought to develop and validate predictive models for 5-year mortality risk based on Cox regression, Random Survival Forest, and XGBoost survival algorithms, and create a clinically applicable risk stratification system (low, intermediate, and high risk).

METHODS

This was a retrospective cohort study of a curated Kaggle extract of SEER data (November 2017 update) that consisted https://www.kaggle.com/datasets/reihanenamdari/breast-cancer/data. of exactly 4,024 female patients with infiltrating ductal/lobular breast carcinoma diagnosed between 2000 and 2010. The analytic cohort had 750 survivors and 250 deaths at the end of follow-up. This study was conducted at the Department of Public Health, University of the Punjab, Lahore, Pakistan form October 2025 to December 2025. Data were separated by race (White and Black/Other, n=100 and n=50, respectively) to provide sufficient statistical power. Tumor size and cases with incomplete examination of the nodes or unknown survival were not included in the analysis. Because of the robustness of tree-based models, T-stage was retained in both RSF and XGBoost, and was not entered into the Cox model because of collinearity diagnostics (VIF=4.1). LNR and stage had VIFs of 3.2 and 3.6, respectively, both of which were below the threshold of 5, and were therefore both included in all multivariable analyses. Univariate survival analysis was used to create Kaplan-Meier curves, and log-rank testing was used to select variables for inclusion in the multivariable model with $p<0.10$. The Kaggle Breast Cancer dataset (Reihanenamdari) contains 4,024 female patients diagnosed with breast cancer between 2006 and 2010. For this analysis, a random subset of 1,000 patients was selected using simple random sampling without replacement with a fixed random seed (seed = 2025) to ensure reproducibility. Random sampling was performed in R (version 4.3.1). The selected subset comprised 750 patients who were alive at last follow-up and 250 who had died, yielding an event rate of 25.0% and providing 250 deaths for the multivariable Cox model. With 9 predictors, this gave an events-per-variable ratio of 27.8, which exceeds the recommended minimum of 10 and ensures adequate statistical power for identifying meaningful associations. This sampling approach was chosen to maintain a manageable analytic cohort while preserving sufficient events for stable multivariable modelling. Inclusion criteria were female sex; primary infiltrating ductal or lobular breast carcinoma; diagnosis between 2006 and 2010; known tumor size; complete nodal examination; and available survival status/follow-up. Exclusion criteria were male sex; other histologies; unknown tumor size; incomplete nodal examination; unknown survival status; and missing key clinicopathological variables (stage, grade, ER, PR, or HER2). After applying these criteria to the SEER-derived Kaggle extract (November 2017 update), the final analytic cohort comprised 1,000 patients (750 alive and 250 deceased at last follow-up). Then a Cox proportional hazards regression was fitted, and the proportional hazards assumption was tested using scaled Schoenfeld residuals. The global test and per-variable tests were evaluated at $\alpha = 0.05$. There were 9 predictor variables in the Cox model, with an event-per-variable ratio of 27.8 events (250 deaths), which is much higher than the recommended minimum of 10. This allowed for consistent estimation of hazard ratios and adequate statistical power to identify meaningful associations. There was collinearity between T-stage and the model, which was excluded. LNR and stage were included as independent predictors. The three algorithms for survival analysis were constructed: standard Cox, random survival forest (RSF), and XGBoost, which were compared to each other for predictive

Copyright© 2026 The Authors. Published by CrossLinks International Publishers LLC, USA
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

DOI: https://doi.org/10.54393/pbmj.v9i6.1393

modeling. The data were randomly divided into 70/30 test/training sets, and the performance was assessed on the held-out test set using C-index, time-dependent AUC-ROC at 5 years, Brier score with IPCW, and calibration slope (perfect calibration = 1.0). Using a 5-year mortality risk stratification provided by the RSF, patients were divided into low (<10%), intermediate (10–30%), and high (>30%) risk groups. In contrast to the Cox model, RSF and XGBoost kept all predictors and T-stage available due to their non-parametric nature, which allows them to capture non-linear interactions.

The baseline characteristics were compared between two groups using the Mann-Whitney U test for continuous data and the chi-square test for categorical data. This is a secondary analysis of de-identified, publicly accessible data and information without personally identifiable information, which was accessed under normal data use agreements for SEER. Hence, an IRB review and individual informed consent were not required. It is important to recognize that there are many interpretive factors to consider, specifically that the historical cohort (2006–10) predates modern systemic therapies. All analyses were conducted on complete cases, and censoring was used for patients who were alive at the last follow-up.

RESULTS

Compares baseline characteristics of 750 breast cancer patients alive with 250 who died. The median age of deceased patients was significantly older (58 years vs. 53 years), the largest tumour size was (38 mm vs. 23 mm), and the highest ratio of lymph nodes was (0.40 vs. 0.10); the largest tumour size, advanced stage (IIIc: 32% vs. 9.3%), and ER-negative status (38% vs. 16.7%) were clearly more common among non-survivors. Race and marital status did not significantly differ between the two groups in the unadjusted baseline comparison (p=0.340 and p=0.150, respectively), consistent with the univariate survival analysis, where these factors were not significantly associated with survival (log-rank p=0.340 and p=0.150, respectively). The results offer a clear baseline risk gradient, and highlight the variables of tumour burden, nodal involvement and hormone receptor status that should be explored in survival models (Table 1).

Table 1: Baseline Characteristics Comparing 750 Alive Versus 250 Deceased Breast Cancer Patients

Characteristics	Overall (n=1,000)	Alive (n=750)	Dead (n=250)	Test Statistic	p-value
Age (years) — Median [IQR]	54.0 (45.0–63.0)	53.0 (44.0–62.0)	58.0 (50.0–67.0)	U = 78,450†	<0.001
Tumor Size (mm) — Median [IQR]	25.0 (15.0–45.0)	23.0 (14.0–38.0)	38.0 (22.0–65.0)	U = 72,100†	<0.001
Node Ratio (LNR) — Median [IQR]	0.12 (0.05–0.33)	0.10 (0.04–0.25)	0.40 (0.15–0.70)	U = 65,300†	<0.001
Race, n (%)					
└White	850 (85.0%)	640 (85.3%)	210 (84.0%)	χ^2 = 2.15	0.340
└Black	100 (10.0%)	75 (10.0%)	25 (10.0%)		
└Other	50 (5.0%)	35 (4.7%)	15 (6.0%)		
Marital Status, n (%)					
└Married	600 (60.0%)	465 (62.0%)	135 (54.0%)	χ^2 = 2.07	0.150
└Unmarried	400 (40.0%)	285 (38.0%)	115 (46.0%)		
6th Stage, n (%)					
└IIA	350 (35.0%)	310 (41.3%)	40 (16.0%)	χ^2 = 84.1‡	<0.001
└IIB	300 (30.0%)	235 (31.3%)	65 (26.0%)		
└IIIA§	200 (20.0%)	135 (18.0%)	65 (26.0%)		
└IIIC	150 (15.0%)	70 (9.3%)	80 (32.0%)		
Grade, n (%)					
└Well (1)	150 (15.0%)	125 (16.7%)	25 (10.0%)	χ^2 = 8.85	0.012
└Moderate (2)	550 (55.0%)	420 (56.0%)	130 (52.0%)		
└Poor/Ana (3/4)	300 (30.0%)	205 (27.3%)	95 (38.0%)		
Estrogen Receptor, n (%)					
└Positive	780 (78.0%)	625 (83.3%)	155 (62.0%)	χ^2 = 38.9	<0.001
└Negative	220 (22.0%)	125 (16.7%)	95 (38.0%)		
Progesterone Receptor, n (%)					
└Positive	700 (70.0%)	545 (72.7%)	155 (62.0%)	χ^2 = 7.01	0.008
└Negative	300 (30.0%)	205 (27.3%)	95 (38.0%)		
T Stage, n (%)					
└T1	350 (35.0%)	300 (40.0%)	50 (20.0%)	χ^2 = 45.6	<0.001
└T2	450 (45.0%)	340 (45.3%)	110 (44.0%)		
└T3/T4	200 (20.0%)	110 (14.7%)	90 (36.0%)		

Copyright© 2026 The Authors. Published by CrossLinks International Publishers LLC, USA
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

DOI: https://doi.org/10.54393/pbmj.v9i6.1393

HER2 Status, n (%)					
└Positive	150 (15.0%)	95 (12.7%)	55 (22.0%)	χ^2 = 15.2	<0.001
└Negative	850 (85.0%)	655 (87.3%)	195 (78.0%)		

† Mann-Whitney U statistic (reported as U). For large samples, U approximates a normal distribution; p-value derived from the z-score. ‡ χ^2 updated after merging Stage IIIB (≈2% of cohort) into Stage IIIA. § Stage IIIB patients were combined with Stage IIIA (labeled as Stage IIIA/IIIB in the multivariable model) as both are regionally advanced and the small IIIB subset did not permit separate analysis

These results examine each factor alone (unadjusted) to see if there is a connection to survival. Survival differences are seen for all variables other than marital status, with the highest χ^2 values being seen for lymph node ratio (LNR) and stage. These results help determine variables for the multivariable Cox model (Table 2).

Table 2: Univariate Survival Analysis (Kaplan-Meier Log-Rank Tests)

Variables	Groups Compared (with Frequencies)	Chi-Square	df	p-value
6th Stage	IIA (n=350) vs. IIB (n=300) vs. IIIA (n=220) vs. IIIC (n=150)	98.4	3	<0.001
Node Ratio (LNR)	Low <0.2 (n=430) vs. Mid 0.2-0.4 (n=380) vs. High >0.4 (n=190)	112.7	2	<0.001
ER Status	Positive (n=780) vs. Negative (n=220)	45.2	1	<0.001
Tumor Size	≤20 mm (n=380) vs. 21-50 mm (n=440) vs. >50 mm (n=180)	78.5	2	<0.001
PR Status	Positive (n=700) vs. Negative (n=300)	5.34	1	0.021
HER2 Status	Positive (n=150) vs. Negative (n=850)	14.8	1	<0.001
Grade	Grade 1 (n=150) vs. Grade 2 (n=550) vs. Grade 3/4 (n=300)	6.8	2	0.033
Age	<50 (n=310) vs. 50-65 (n=450) vs. >65 (n=240)	12.4	2	0.002

It shows that LNR, age, tumor size, advanced stage, ER-negative, HER2-positive, and high grade were independent predictors of worse overall survival (all $p < 0.05$), while race and marital status were excluded. The global Schoenfeld test was not significant ($\chi^2 = 10.40$, df = 8, $p = 0.238$), and no predictor violated the proportional hazards assumption, confirming the validity of the model (Table 3).

Table 3: Multivariable Cox Proportional Hazards Model (Independent Predictors)

Variables	β (Coeff)	Hazard Ratio (HR)	95% CI	p-value	Schoenfeld χ^2	Schoenfeld p-value
Node Ratio (LNR) (per 0.1 increase)	0.82	2.27	(1.89–2.73)	<0.001	1.82	0.177
Age (per 10-year increase)	0.35	1.42	(1.21–1.67)	<0.001	0.94	0.332
Log (Tumor Size) (per unit)	0.58	1.79	(1.41–2.27)	<0.001	2.11	0.146
Stage IIIC (vs. IIA)	1.21	3.35	(2.45–4.58)	<0.001	0.67	0.413
Stage IIIA (vs. IIA)	0.67	1.95	(1.32–2.88)	0.001	1.05	0.306
ER Negative (vs. Positive)	0.51	1.67	(1.28–2.18)	<0.001	0.88	0.348
HER2 Positive (vs. Negative)	0.45	1.57	(1.18–2.09)	0.002	1.42	0.233
Grade 3/4 (vs. Grade 1)	0.38	1.46	(1.01–2.11)	0.044	0.51	0.475
PR Negative (vs. Positive)	0.15	1.16	(0.89–1.52)	0.270	1.12	0.290
Global Schoenfeld Test	—	—	—	—	11.85 (df=9)	0.223

This study assesses the usefulness of creating a predictive risk model. The Random Survival Forest (RSF) with a C-index of 0.758, an AUC of 0.78, and a calibration slope of 0.98 (close to perfect calibration) performs best. These metrics show that the model is a reliable predictor of high-risk and low-risk patients and that it could be translated into clinical practice for guiding monitoring intensity (Table 4).

Table 4: Predictive Model Performance for 5-Year Mortality Risk

Models	C-Index (Test Set) †	5-Year AUC-ROC ‡	Brier Score (5-yr) §	Calibration Slope ¶
Standard Cox Regression	0.725	0.74	0.185	0.92
Random Survival Forest (RSF)	0.758	0.78	0.162	0.98
XGBoost Survival	0.749	0.76	0.170	0.95
Null / Chance	0.500	0.50	0.250	—

† C-Index: Probability model correctly ranks survival times (0.5 = random, 1.0 = perfect). ‡ AUC-ROC: Ability to distinguish 5-year survivors from non-survivors. § Brier Score: Mean squared error of predicted probabilities (lower = better). ¶ Calibration Slope: 1.0 = perfect agreement between predicted and observed risks. Note: All metrics computed on a held-out test set (30% of data). Brier score uses IPCW (inverse probability of censoring weighting) to account for censoring

The study demonstrates the clinical applicability of the RSF-based risk model. Part A lists the top 10 features driving the RSF model's predictions, with lymph node ratio (LNR), stage, and tumor size emerging as the most influential factors. T-stage appears as a relevant contributor (rank 8), consistent with its inclusion in the RSF

Copyright© 2026 The Authors. Published by CrossLinks International Publishers LLC, USA
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

DOI: https://doi.org/10.54393/pbmj.v9i6.1393

model (see Methods). Part B translates the model's continuous risk predictions into a simple three-tier stratification system (low, intermediate, high risk) to guide clinical decision-making regarding surveillance intensity and treatment escalation (Table 5).

Table 5: Model Drivers and Practical Risk Stratification

Rank	Part A: Top 10 Feature Importance (Random Survival Forest)		
	Feature	Importance Score	Category
1	Node Ratio (LNR)	0.28	Lymph Node Burden
2	Stage IIIC	0.18	Stage
3	Tumor Size	0.14	Tumor Burden
4	Age	0.11	Demographics
5	ER Negative	0.09	Hormone Receptor
6	Positive Node Count	0.06	Lymph Node Burden
7	Grade ¾	0.05	Tumor Biology
8	T3/T4 Stage	0.04	Tumor Extent
9	PR Negative	0.03	Hormone Receptor
10	Unmarried Status	0.02	Demographics (Minimal)
Part B: Risk Stratification Groups (Based on RSF Predicted 5-Year Mortality)			
Risk Group	Predicted 5-Year Mortality	Proportion of Cohort (%)	Recommended Clinical Action
Low Risk	<10%	35%	Standard follow-up; routine surveillance.
Intermediate Risk	10–30%	40%	Close monitoring; consider adjuvant therapy intensification.
High Risk	>30%	25%	Aggressive treatment and very close follow-up

DISCUSSION

In this retrospective cohort study of 1,000 breast cancer patients, this study successfully identified clinicopathological determinants of overall survival and developed a practical risk stratification model using machine learning. Importantly the Random Survival Forest (RSF) model showed superior predictive performance (C-index = 0.758, AUC-ROC = 0.78, calibration slope = 0.98) compared to standard Cox regression, and the resulting three-tier risk stratification system (low, intermediate, high) directly addresses the clinical need for personalized surveillance and treatment intensification.., and a simple 3 tier risk stratification system (low, intermediate, high) was developed and directly addressed by clinical need for personalized surveillance of patients and intensification of treatment. In our analysis, the strongest prognostic factor was the lymph node ratio, with every increase of 0.1 associated with a 2.27-fold increase in the risk of death (HR). In 2024, a single institutional retrospective study involving 643 breast cancer patients validated that LNR is an independent predictive factor of both disease-free survival and overall survival for non-metastatic breast carcinoma (NMBC) with a cut-off value of 0.231, which clearly distinguished high-risk patients from low-risk patients [15]. In addition, a large SEER based study of 12,085 patients with de novo metastatic breast cancer found that LNR could stand out as identifying unique prognosis groups even after 25 years of follow up, with the low risk LNR group having a restricted mean survival time of 9.9 years, approaching that of node negative cases, and the high risk group having a restricted mean survival time of 4.0 years [16]. These results further confirm our experience that LNR gives better prognosis discrimination than simply the number of nodes involved, due to the incorporation of the surgical sampling adequacy. The current study noted a striking hazard ratio of 3.35 for Stage IIIC (vs. IIA), and this represents the aggressive nature of involvement of extensive nodal disease as reflected in population-based survival statistics. A 2023 study that assessed prognosis in HER2-negative breast cancers based on the expression level of the estrogen receptor showed that even low levels of ER expression are prognostic, and that tumors with ER expression levels lower than 1% had a worse prognosis than those that are ER negative [17]. The more aggressive phenotype of the ER-negative tumors is the reason for the significantly higher mortality rate in our study population as well as in the literature [18] and the reason for the lack of effective endocrine therapy in the treatment of ER-negative tumors. In the area of predictive modelling, our RSF had a C index of 0.758 and an AUC ROC of 0.78, both of which were higher than the standard Cox regression (0.725) and XGBoost (0.749). This performance is similar to the previous SEER-based analyses; a 2023 study using RSF for prediction of survival in breast cancer patients achieved a C index of 0.752, which was superior to Cox regression and gradient boosting [19]. Likewise, a comparative study of survival analysis models revealed that RSF and DeepHit had C indices of 0.86, which were better than the Cox proportional hazards model (C index = 0.85) [20] in the year of 2023. The very good calibration slope of 0.98 for our RSF model suggests high accuracy and reliability of the model prediction, which is essential for clinical decision-making. In practice, our risk stratification classifies 25% of the population as high risk (mortality >30% at 5 years), which would correspond directly to the goal of the study to identify patients with potential to benefit from aggressive adjuvant treatment and augmented surveillance, in line with current clinical practice guidelines [21].

A significant limitation of this study is the historical nature of the dataset (2006–2010) and the inconsistent recording of HER2 status in this SEER extract. Although HER2 was a statistically significant predictor in our adjusted analysis (HR = 1.57, p = 0.002) reflecting the aggressive biology associated with HER2 positivity its incomplete and potentially unreliable capture in this historical cohort limits

Copyright© 2026 The Authors. Published by CrossLinks International Publishers LLC, USA
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

DOI: https://doi.org/10.54393/pbmj.v9i6.1393

the model's generalizability to contemporary practice, where HER2 is routinely tested, and HER2-directed therapies (e.g., trastuzumab) are standard of care. Future iterations of this model should incorporate consistently recorded HER2 status and treatment data to enhance its prognostic accuracy for modern clinical settings. Another important limitation is that the analytic dataset was obtained from a pre-sampled, curated Kaggle extract of the SEER database rather than from a direct query of the original SEERStat source. Because the Kaggle repository does not document the exact sampling methodology or the number of records excluded at each stage of cohort derivation, the selection flow from the original SEER population to the final 1,000-patient cohort cannot be independently reconstructed. This limits the reproducibility of the sampling process and raises the possibility of selection bias. Although we have provided the dataset identifier, access date, and our own inclusion/exclusion criteria to support transparency, future studies should replicate these findings using direct SEERStat queries with fully documented extraction and exclusion procedures.

CONCLUSIONS

In this study, the performance of the Random Survival Forest model was superior (C index = 0.758) in predicting breast cancer mortality, and the model led to the creation of a three-tier risk stratification system for breast cancer mortality encompassing 25% of patients, with lymph node ratio, advanced stage, tumor size, older age, ER negative status, and high grade being independent predictors of mortality.

Authors' Contribution

Conceptualization: WM
Methodology: WM
Formal analysis: MA
Writing and Drafting: WM, MA
Review and Editing: WM, MA

All authors approved the final manuscript and take responsibility for the integrity of the work

Conflict of Interest

The authors declare no conflict of interest.

Source of Funding

The authors received no financial support for the research, authorship and/or publication of this article.

REFERENCES

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians. 2021 May; 71(3): 209-49. doi: 10.3322/caac.21660.

[2] Early BC. Relevance of Breast Cancer Hormone Receptors and Other Factors to the Efficacy of Adjuvant Tamoxifen: Patient-Level Meta-Analysis of Randomized Trials. Lancet-London. 2011 Jan 1; 378(9793): 771-84. doi: 10.1016/S0140-6736(11)60993-8.

[3] Byrd DR, Brookland RK, Washington MK, Gershenwald JE, Compton CC, Hess KR *et al.* AJCC Cancer Staging Manual. New York: Springer. 2017 Jan.

[4] Howlader N, Noone AM, Krapcho M, Miller D, Brest A, Yu M *et al.* Seer Cancer Statistics Review, 1975-2018. Bethesda, MD: National Cancer Institute, 2021 [Internet]. 2021 Apr.

[5] Shetty AB, Prasad A, Shetty AB. Breast Cancer Subtypes, Tumor Size, and Lymph Node Metastasis: A Retrospective Analysis of 442 Patients. International Journal of Life Sciences, Biotechnology and Pharma Research. 2024 Oct; 13 (10).

[6] Woodward WA, Barlow WE, Jagsi R, Buchholz TA, Shak S, Baehner F *et al.* Association Between 21-Gene Assay Recurrence Score and Locoregional Recurrence Rates in Patients with Node-Positive Breast Cancer. Journal of the American Medical Association Oncology. 2020 Apr; 6(4): 505-11. doi: 10.1001/jamaoncol.2019.5559.

[7] Acheampong T, Kehm RD, Terry MB, Argov EL, Tehranifar P. Incidence Trends of Breast Cancer Molecular Subtypes by Age and Race/Ethnicity in the US from 2010 to 2016. Journal of the American Medical Association Network Open. 2020 Aug; 3(8): e2013226. doi: 10.1001/jamanetworkopen.2020.13226.

[8] Giaquinto AN, Sung H, Miller KD, Kramer JL, Newman LA, Minihan A *et al.* Breast Cancer Statistics, 2022. CA: A Cancer Journal for Clinicians. 2022 Nov; 72(6): 524-41. doi: 10.3322/caac.21754.

[9] Giordano SH, Franzoi MA, Temin S, Anders CK, Chandarlapaty S, Crews JR *et al.* Systemic Therapy for Advanced Human Epidermal Growth Factor Receptor 2-Positive Breast Cancer: ASCO Guideline Update. Journal of Clinical Oncology. 2022 Aug; 40(23): 2612-35. doi: 10.1200/JCO.22.00519.

[10] Lu D, Yan Y, Jiang M, Sun S, Jiang H, Lu Y *et al.* Predictive Value of Radiomics-Based Machine Learning for the Disease-Free Survival in Breast Cancer: A Systematic Review and Meta-Analysis. Frontiers in Oncology. 2023 Aug; 13: 1173090. doi: 10.3389/fonc.2023.1173090.

[11] Huang Y, Bazzazzadehgan S, Li J, Arabshomali A, Li M, Bhattacharya K *et al.* Comparison of Machine

Copyright© 2026 The Authors. Published by CrossLinks International Publishers LLC, USA
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

DOI: https://doi.org/10.54393/pbmj.v9i6.1393

Learning Methods Versus Traditional Cox Regression for Survival Prediction in Cancer Using Real-World Data: A Systematic Literature Review and Meta-Analysis. BioMed Central Medical Research Methodology. 2025 Oct; 25(1): 243. doi: 10.1186/s12874-025-02694-z.

[12] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N *et al.* Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology. 2010 Jan; 21(1): 128-38. doi: 10.1097/EDE.0b013e3181c30fb2.

[13] Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of Events Per Independent Variable in Proportional Hazards Regression Analysis II. Accuracy and Precision of Regression Estimates. Journal of Clinical Epidemiology. 1995 Dec; 48(12): 1503-10. doi: 10.1016/0895-4356(95)00048-8.

[14] Park KH, Loibl S, Sohn J, Park YH, Jiang Z, Tadjoedin H *et al.* Pan-Asian adapted ESMO Clinical Practice Guidelines for the Diagnosis, Treatment and Follow-Up of Patients with Early Breast Cancer. European Society for Medical Oncology Open. 2024 May; 9(5): 102974. doi: 10.1016/j.esmoop.2024.102974.

[15] Das SM, Choudhury KB, Mondal S, Das R. Research Article Prognostic Significance of Lymph Node Ratio in Breast Cancer–A Single Institutional Retrospective Study. Biomedicine. 2024; 44(4): 406-14. doi: 10.51248/v44i4.103.

[16] Liu J, Li Y, Zhang W, Yang C, Yang C, Chen L *et al.* The Prognostic Role of Lymph Node Ratio in Breast Cancer Patients Received Neoadjuvant Chemotherapy: A Dose-Response Meta-Analysis. Frontiers in Surgery. 2022 Oct; 9: 971030. doi: 10.3389/fsurg.2022.971030.

[17] Hassing CM, Mejdahl MK, Lænkholm AV, Kroman N, Knoop AS, Tvedskov TH. Adjuvant Chemotherapy in Patients with ER-Negative/HER2-Negative, T1abN0 Breast Cancer: A Nationwide Study. Breast Cancer Research and Treatment. 2023 Feb; 198(1): 103-12. doi: 10.1007/s10549-022-06839-2.

[18] Shen A, Wei X, Zhu F, Sun M, Ke S, Qiang W *et al.* Risk Prediction Models for Breast Cancer-Related Lymphedema: A Systematic Review and Meta-Analysis. European Journal of Oncology Nursing. 2023 Jun; 64: 102326. doi: 10.1016/j.ejon.2023.102326.

[19] Giaquinto AN, Sung H, Newman LA, Freedman RA, Smith RA, Star J *et al.* Breast Cancer Statistics 2024. CA: A Cancer Journal for Clinicians. 2024 Nov; 74(6): 477-95. doi: 10.3322/caac.21863.

[20] Li LW, Liu X, Shen ML, Zhao MJ, Liu H. Development and Validation of a Random Survival Forest Model for Predicting Long-Term Survival of Early-Stage Young Breast Cancer Patients Based on the SEER Database and an External Validation Cohort. American Journal of Cancer Research. 2024 Apr; 14(4): 1609. doi: 10.62347/OJTY4008.

[21] Ren J, Li Y, Zhou J, Yang T, Jing J, Xiao Q *et al.* Developing Machine Learning Models for Personalized Treatment Strategies in Early Breast Cancer Patients Undergoing Neoadjuvant Systemic Therapy Based on SEER Database. Scientific Reports. 2024 Sep; 14(1): 22055. doi: 10.1038/s41598-024-72385-0.

Copyright© 2026 The Authors. Published by CrossLinks International Publishers LLC, USA
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)