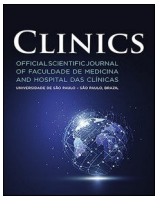
Clinics 80 (2025) 100818

CLINICS
OFFICIAL SCIENTIFIC JOURNAL OF FACULDADE DE MEDICINA AND HOSPITAL DAS CLÍNICAS
UNIVERSIDADE DE SÃO PAULO - SÃO PAULO, BRAZIL

Contents lists available at ScienceDirect

Clinics

journal homepage: www.elsevier.com/locate/clinsp



Original articles

Machine learning predictor to investigate treatment modalities and overall survival in HER2+ patients with early-stage breast cancer

Kai Wang [a], Jianing Liu [b,*]

[a] Medical School, Southeast University, Nanjing, China
[b] Medical Faculty, Ulm University, Ulm, Germany

ARTICLE INFO

Editor: José Maria Soares Junior

Keywords:
Breast cancer
Treatment
Machine learning
Predictions
Nomogram

ABSTRACT

Purpose: This study aimed to explore the impact of treatment modalities on the survival outcomes of HER2-positive patients with early-stage invasive ductal breast cancer.
Methods: Hierarchical clustering analysis was used to identify distinct subgroups based on treatment modalities. Comparative analysis between the clusters identified significant treatment-related variables. Cox regression analysis was performed to construct a survival prediction model and nomogram incorporating these variables. Random Survival Forest (RSF) and SHapley Additive exPlanations (SHAP) analysis were employed for further validation and interpretation.
Results: A total of 9569 patients with early-stage HER2+ invasive ductal breast cancer were included, and five treatment-related clusters were identified using hierarchical clustering. Post-clustering analysis revealed that survival outcomes were influenced by various treatment factors, including time length from diagnosis to treatment, surgery approach, response to neoadjuvant therapy, combination with radiation, chemotherapy and/or systemic therapy, and treatment sequence. A prediction model and nomogram were developed, demonstrating good discriminatory ability and excellent predictive performance at 3-, 5-, and 8-years.
Conclusions: The study highlighted the importance of an aggressive and comprehensive treatment approach for patients with early-stage HER2-positive breast cancer. It emphasized the multifaceted nature of treatment outcomes and the need to consider multiple treatment factors beyond surgery alone. The developed survival prediction model provided valuable insights into the contribution of different treatment modalities to survival outcomes.

Introduction

Breast cancer is the most common cancer globally, accounting for 12.5 % of all new annual cancer cases, and in the United States, it is the most frequently diagnosed cancer among women, representing about 30 % of new cancer diagnoses in women each year. Approximately 13 % of women are predicted to develop invasive breast cancer in their lifetime, and 0.5 %–1 % of breast cancer cases occur in men[1,2] Since breast cancer is a formidable health concern, treatment interventions for early-stage patients play crucial roles in improving outcomes and survival rates[3,4] In the management of early-stage breast cancer, treatment modalities such as surgery, radiation therapy, chemotherapy, hormone therapy, and targeted therapy have been widely employed. While these approaches have shown efficacy, understanding the specific impact of each modality and combination strategy on survival outcomes remains an ongoing research endeavor.

Traditional statistical analyses have contributed valuable insights into the relationship between treatment modalities and survival outcomes. However, the complexity and heterogeneity of breast cancer and treatment-related data necessitate innovative approaches to unravel the intricate interplay between treatments and outcomes comprehensively. Recently, machine learning algorithms have garnered increasing attention in clinical research due to their capability to process extensive clinical data and implement performance metrics, enhancing the ability to model straightforward relationships between exposures and outcomes. Classical machine learning methods are categorized into supervised and unsupervised approaches, distinguished by the presence or absence of labeled outcome variables[5] On one hand, supervised machine learning, particularly the utilization of random forests in survival prediction, facilitates predictive modeling and validation based on

* Corresponding author.
E-mail address: jianing-1.liu@uni-ulm.de (J. Liu).

https://doi.org/10.1016/j.clinsp.2025.100818
Received 22 April 2024; Received in revised form 30 August 2025; Accepted 17 October 2025
Available online 31 October 2025
1807-5932/© 2025 HCFMUSP. Published by Elsevier España, S.L.U. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

annotated training data with known survival outcomes. Further integration of SHAP (SHapley Additive exPlanations) analysis enhances the interpretability of the model by revealing the individual influence of each predictive variable on the overall survival performance. On the other hand, unsupervised machine learning algorithms process unlabeled data to elucidate hidden structures. In detail, it can identify distinct subgroups based on inherent similarities and differences in treatment characteristics, without any preconceived knowledge or guidance from survival outcomes. Notably, Hierarchical Clustering (HC), compared with other clustering methods, does not assume specific cluster shapes, making it adept at capturing diverse and non-uniform patterns in treatment-related data. The hierarchical nature of HC allows for the discovery of nested or overlapping subgroups within larger clusters, offering deeper insights into the organization of complex treatment-related data[6] Hence, the integration of unsupervised and supervised machine learning approaches holds significant potential for elucidating the complex relationships between treatment modalities and survival outcomes.

In this study, the authors focused on early-stage invasive ductal breast cancer HER2+ patients as the study population due to several reasons. Firstly, Invasive Ductal Carcinoma (IDC), also called infiltrating ductal carcinoma, is the most common and aggressive type of breast cancer, with a higher likelihood of lymph node involvement and metastasis. Secondly, HER2+ breast cancer is a distinct subtype characterized by overexpression of the Human Epidermal Growth Factor Receptor 2 (HER2) protein, which is associated with aggressive tumor behavior and poorer prognosis[7] Hence, the treatment of HER2+ breast cancer has its unique considerations. HER2-targeted therapy, such as trastuzumab and pertuzumab, is often considered the cornerstone of systemic therapy and enriched the neoadjuvant therapy for HER2+ breast cancer[8], reflecting the specific molecular characteristics of this subtype. Thirdly, directing our attention towards patients in the early stages enables a meticulous exploration of the efficacy of diverse therapeutic approaches within the framework of the initial disease manifestation, a phase where interventions are anticipated to exert the most profound influence on survival outcomes[9].

Therefore, the present study aims to employ both supervised and unsupervised machine learning approaches to investigate the impact of treatment modalities on patient survival outcomes. Leveraging the Surveillance, Epidemiology, and End Results (SEER) database, a comprehensive repository housing extensive clinical and demographic data from a large cohort of early-stage breast cancer patients, the authors seek to uncover valuable insights. By integrating supervised and unsupervised machine learning algorithms, the authors' objective is to identify significant predictors, reveal latent treatment-related subgroups, and construct a practical nomogram based on the association between treatment modalities and survival. The findings gleaned from the investigation hold the potential to provide valuable perspectives on the efficacy of various treatment modalities and combination strategies, thereby facilitating improvements in personalized therapeutic interventions for early-stage breast cancer.

Methods

Data collection

The study data were obtained from the Surveillance, Epidemiology, and End Results (SEER) database, which comprises comprehensive tumor registries across various regions in the United States. This extensive database contains detailed information on demographics, socioeconomic factors, cancer characteristics, and treatment approaches. The authors extracted clinical data of patients diagnosed with early-stage breast cancer from 1975 to 2019 (November 2021 Submission) from the publicly available SEER data. The collected data included a wide range of variables: 1) Demographic information: age, gender, race, marital status at diagnosis, household income, and residence area; 2) Tumor-related characteristics: AJCC stage, distant lymph node metastasis, tumor size, statuses of Estrogen Receptor (ER) and Progesterone Receptor (PR), total number of in-situ or malignant tumors, and primary site of cancer; 3) Treatment-related variables: surgery, radiation therapy, radiation and surgery sequence, chemotherapy, systemic therapy and surgery sequence, response to neoadjuvant therapy, nonprimary surgical procedure, reason for no cancer-directed surgery, months between diagnosis and treatment; 4) Outcome variables, including survival time and overall survival status. It is important to note that in the SEER dataset, neoadjuvant therapy is defined as systemic therapy provided before curative surgery with the intention of reducing tumor size prior to surgery. However, the systemic treatment administered before surgery might not have had neoadjuvant intent or might not have been administered long enough to expect a relevant tumor response[10]

Participant selection criteria

To be included in the study, participants had to meet the following criteria: 1) Diagnosis of primary breast cancer in American Joint Committee on Cancer (AJCC) stages IA, IB, IIA to IIB; 2) HER2 status positive; 3) Histologic type of infiltrating duct carcinoma, according to the International Classification of Diseases for Oncology, Third Edition (ICD-O-3). Exclusion criteria encompassed: 1) Advanced stage, HER2-negative, non-infiltrating duct cancer; 2) Incomplete information pertaining to demographic characteristics, treatment modalities, or other tumor-related variables; 3) Missing follow-up data.

Unsupervised machine learning analysis

To explore the heterogeneity of treatment modalities and identify distinct patient subgroups, the authors employed Hierarchical Clustering (HC) for the data of treatment-related variables. This technique groups patients based on similarities in treatment characteristics. Firstly, the authors used Multiple Correspondence Analysis (MCA) for dimensionality reduction, capturing underlying relationships between categorical variables. MCA reduced the dataset's dimensionality while retaining informative features. To ensure comparability and meaningful interpretation, MinMax standardization scaled numerical variables to a common range, preventing bias from variable scales. Next, the authors used the Silhouette Score to determine the optimal number of clusters. Subsequently, the authors employed the Agglomerative Clustering algorithm (in Python) with Euclidean distance and Ward's linkage to construct the hierarchical clustering dendrogram. In the post hoc validation for the quality of the clustering scheme, the authors used two clustering metrics: the Calinski-Harabasz index and the Davies-Bouldin Index. The Calinski-Harabasz index measures the ratio of between-cluster dispersion to within-cluster dispersion. Higher values of the Calinski-Harabasz index indicate better-defined clusters. The Davies-Bouldin index measures the average similarity between each cluster and its most similar cluster while considering the distance between the cluster centroids. The index ranges from 0 to positive infinity, where lower values indicate more distinct clusters. The closer the index is to 0, the better the clustering result. These metrics quantified clustering performance, assessing the algorithm's effectiveness in capturing meaningful patterns.

Analysis of the relationship between treatment modalities and survival outcomes for the clusters identified by HC

Statistical tests were conducted to explore differences in the distribution of treatment-related variables among the identified clusters. Chi-Square tests were used for categorical variables, while *t*-tests were employed for continuous variables with normal distribution; otherwise, the Kruskal-Wallis test was used; p-values below 0.05 were considered significant. Kaplan-Meier (KM) survival curves and log-rank tests were used to further investigate outcome differences between the clusters,

visualizing overall survival outcomes and assessing their significance.

Cox proportional hazard regression analysis, construction of a prediction model and nomogram

To further investigate the association between treatment-related clusters and survival outcomes while controlling for potential confounding factors, the authors performed univariate and multivariate Cox regression analysis. Significant co-variables ($p < 0.05$) were integrated into a prediction model. The dataset was split into training and testing sets (7:3 ratio), with the model fitted to training data and validated on testing data. A nomogram was generated for individualized predictions and evaluated through the Concordance index (C-index), calibration curve, and Receiver Operating Characteristic (ROC) curve.

Supervised machine learning and SHAP analysis

To validate the impact of these important variables on survival prediction, the authors conducted supervised machine learning under random forest algorithms and SHAP analysis. C-index was employed to assess the model's performance. Additionally, SHAP values quantified each variable's contribution to the prediction for each patient, providing a unified measure of feature importance (Supplementary Table 2).

Statistical analysis

The design, analysis and reporting within this study adhered to the STROBE Statement. All statistical analyses were performed using *R* and Python software packages. Appropriate statistical tests, models, and algorithms were chosen based on the nature of the data and research questions.

Results

Study population

Based on the inclusion criteria, the authors included 9569 patients diagnosed with early-stage HER2+ invasive ductal breast cancer in this analysis. The demographic characteristics of the study population are shown in Table 1.

Table 1
Demographic characteristics of the study population.

Characteristic	Overall (n = 9569)
Age	57 (48, 67)[a]
Sex	
Female	9512 (99 %)[b]
Male	57 (0.6 %)
Race	
American Indian or Alaska Native	72 (0.8 %)
Asian or Pacific Islander	1637 (17 %)
Black	901 (9.4 %)
White	6959 (73 %)
Residence area	
Counties in metropolitan areas of >1 million population	5447 (57 %)
Counties in metropolitan areas of 250,000 to 1 million population	2656 (28 %)
Counties in metropolitan areas of <250 thousand population	538 (5.6 %)
Nonmetropolitan counties adjacent to a metropolitan area	448 (4.7 %)
Nonmetropolitan counties not adjacent to a metropolitan area	480 (5.0 %)
Marital status when diagnosed	
Divorced	1033 (11 %)
Married	5942 (62 %)
Separated	93 (1.0 %)
Single (never married)	1537 (16 %)
Unmarried or Domestic Partner	55 (0.6 %)
Widowed	909 (9.5 %)
Median household income (Yearly)	
< $35,000	36 (0.4 %)
$35,000 – $49,999	411 (4.3 %)
$50,000 – $69,999	3412 (36 %)
> $70,000	5710 (60 %)
Stage	
IA	5133 (54 %)
IB	746 (7.8 %)
IIA	2368 (25 %)
IIB	1322 (14 %)
Hormone Receptor (HR) status	
Negative	1329 (14 %)
Positive	8240 (86 %)
Progesterone Receptor (PR) status	
Borderline/Unknown	10 (0.1 %)
Negative	3510 (37 %)
Positive	6049 (63 %)
Oestrogen Receptor (ER) Status	
Borderline/Unknown	1 (<0.1 %)
Negative	1522 (16 %)
Positive	8046 (84 %)
Distant Lymph Nodes Metstatsis	1648 (17 %)

[a] Median (IQR).
[b] n (%).

Difference in treatment modalities of clusters assigned by the hierarchical clustering (HC) algorithm

Based on the hierarchical clustering algorithm, the study population was divided into five clusters (labeled 0, 1, 2, 3, and 4) comprising 3945, 1020, 1438, 2501, and 665 patients, respectively. The clustering results demonstrated good quality, as indicated by the Calinski-Harabasz index of 10,424.25 and the Davies-Bouldin index of 0.77, better than other clustering schemes using non-optimal cluster numbers (Supplementary Table 3).

Cluster characteristics were visualized in Fig. 1, highlighting the differences in treatment modalities. Cluster 0 predominantly underwent partial mastectomy (88 %) and rarely received neoadjuvant therapy. Radiation therapy was administered after surgery in the majority of patients (99 %), with beam radiation being the most common form (95 %). Chemotherapy was given to 74 % of patients, and systemic therapy was typically administered after surgery (93 %). Cluster 2 also favored partial mastectomy (64 %), along with beam radiation and chemotherapy, but there are two differences from Cluster 0. Firstly, in terms of neoadjuvant therapy response, over 80 % of patients in Cluster 2 achieved Partial Response (PR) or Complete Response (CR), with CR accounting for 33 %. Secondly, more than half (56 %) of patients received systemic therapy both before and after surgery.

In contrast, Cluster 1 featured a higher proportion of total mastectomy (43 %) and nearly no utilization of radiation, chemotherapy, and neoadjuvant therapy. Cluster 3 predominantly underwent total mastectomy (61 %) and most of the patients received chemotherapy and systemic therapy after surgery, but not neoadjuvant therapy and radiation. Cluster 4, another total mastectomy group, received neoadjuvant therapy, exhibited a high response rate (80 %), and had higher utilization of chemotherapy and a different systemic therapy sequence compared to Cluster 3.

Post-clustering analysis: survival outcome of identified clusters

Having gained insight into the treatment-related differences between clusters identified by hierarchical clustering, the authors analyzed the survival outcomes and explored the relationship between treatment modalities and overall survival. Fig. 2 shows the overall survival curves of the different clusters. Cluster 4 exhibited the best overall survival rate, with a slightly fluctuating curve reflecting varying outcomes over time. Comparatively, Cluster 0 and Cluster 2 had lower survival rates but showed only a gradual decline, while Cluster 3 demonstrated a remarkable decline over time. Notably, Cluster 1 had the lowest overall survival rate. The log-rank test showed significant differences among

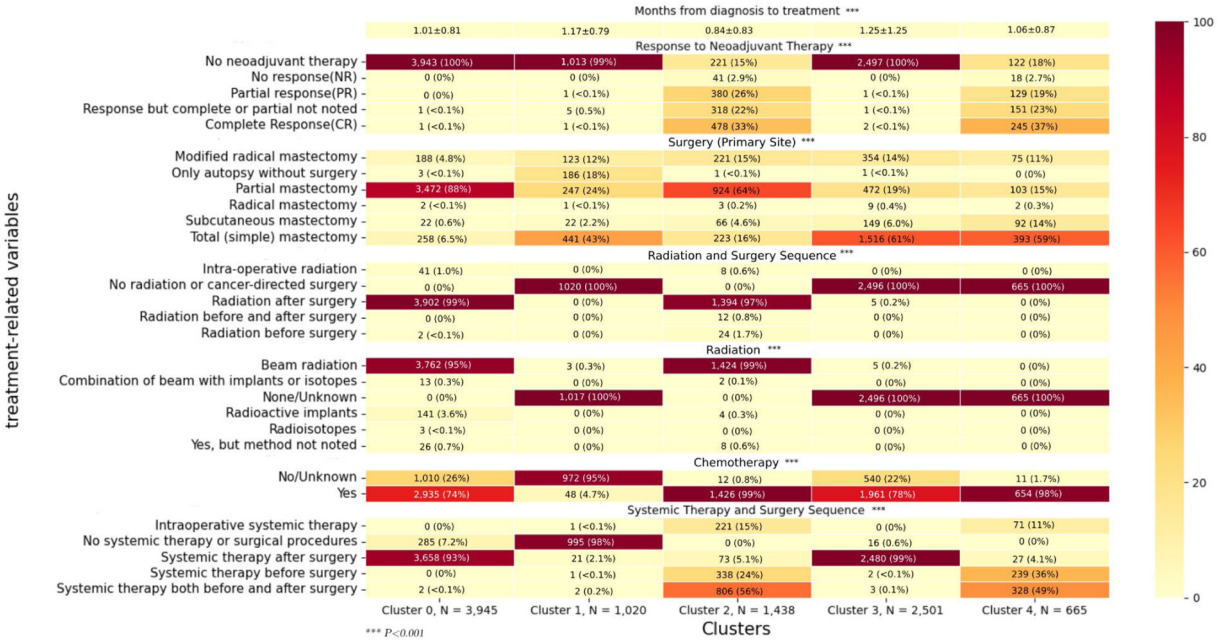


Fig. 1. Heatmap of different percentages of treatment modalities among the clusters assigned by hierarchical clustering.

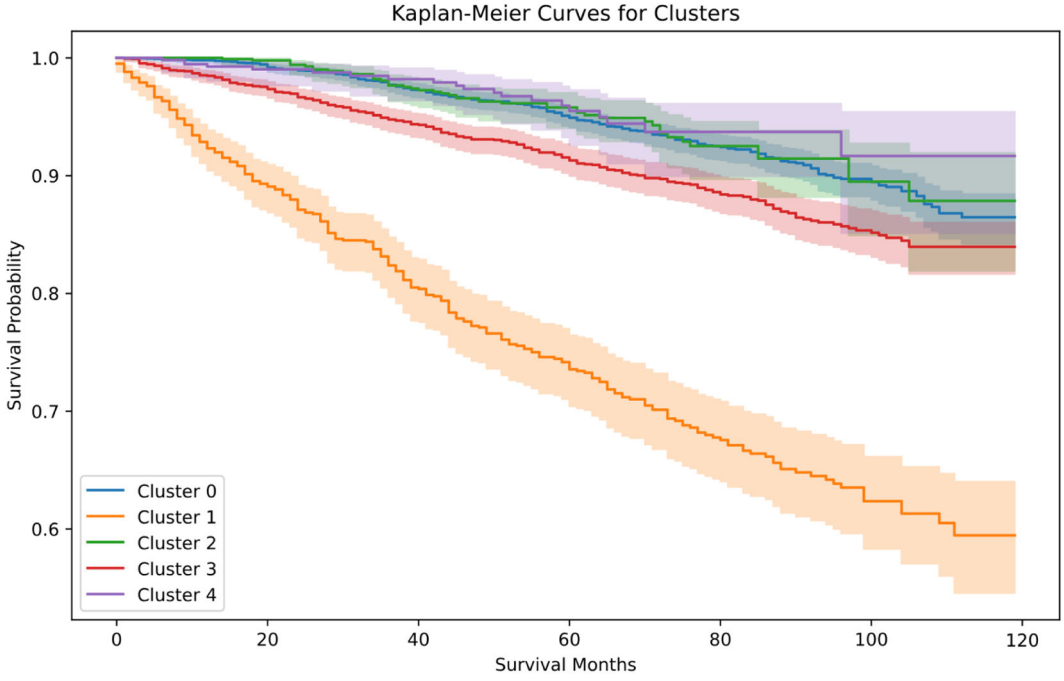


Fig. 2. Overall survival curve of different clusters assigned by hierarchical clustering.

these clusters (effect size: 413.89, p-value < 0.005).

In addition, post-clustering analysis revealed that survival outcomes were not solely determined by surgical procedures but also influenced by other treatment conditions. Cluster 4, with total mastectomy as the dominant treatment, exhibited the highest overall survival rate, while Cluster 1, despite a similar surgical approach, demonstrated the poorest survival. This observation highlights the potential impact of factors such as neoadjuvant therapy response, combination with chemotherapy and/or systemic therapy, and treatment sequence. Interestingly, variations in systemic therapy and surgery sequence, and neoadjuvant therapy response, did not significantly affect overall survival for patients in Clusters 0 and 2, who underwent partial mastectomy.

Adjustment of potential confounding factors and identification for other significant covariates

To eliminate interference from demographic characteristics and tumor-related variables, univariate and multivariate Cox Proportional Hazard Regression analyses were used to examine the correlation between treatment modality clusters and overall survival status (Supplementary Table 1), under the premise that the cluster variable is not highly correlated with other clinical variables. In the fully adjusted model (multivariate model 2), Cluster 1 remained significantly associated with a higher HR of 3.04 (95 % CI: 2.53, 3.66), and Cluster 3 also retained its significant association with a higher HR of 1.42 (95 % CI:

1.18, 1.70). It suggested that the treatment-related clusters identified through hierarchical clustering analysis are independently associated with overall survival in early-stage breast cancer, even after adjusting for various potential confounding factors. Moreover, covariates significantly associated with overall survival were identified, including sex, race, age, marital status, stage, tumor size, and total number of in-situ or malignant tumors.

Prediction model and nomogram

To comprehensively understand the impact of treatment modalities on overall survival, the authors developed a prediction model and nomogram to estimate individualized 3-, 5-, and 8-year survival probabilities for early-stage HER2+ invasive ductal breast cancer (Fig. 3). The model includes treatment-related variables (time from diagnosis to treatment, surgery approach, response to neoadjuvant therapy, chemotherapy, radiation and surgery sequence, and systemic therapy and surgery sequence) and other significant factors (age, race, marital status, stage, total number of in-situ or malignant tumors, and tumor size). The nomogram assigns scores to each factor, enabling clinicians to calculate survival probabilities at specific time points.

Further evaluation demonstrated the nomogram's good discriminatory ability, with C-index values of 0.821 (training set) and 0.793 (test set), indicating accurate predictions. ROC curves (Fig. 4A and Fig. 4C) displayed excellent performance at 3-, 5-, and 8-years, with AUC values of 0.816, 0.805, and 0.801, respectively. The results in the test set (AUC values of 3-, 5-, and 8-year ROC curves: 0.805, 0.784, and 0.763) were consistent with those in the training set. Calibration curves confirmed a strong alignment between predicted and observed survival outcomes (Fig. 4B and Fig. 4D).

Notably, beyond treatment modality, treatment delay itself also emerged as a key determinant of survival. While the nomogram incorporated time to treatment as a continuous variable, the model's linear assumption may overlook potential threshold effects. To address this limitation, an additional analysis was conducted to explore whether delays beyond specific timepoints were associated with worsened prognosis. The authors performed a series of stratified Cox models using monthly cutoff points (1–8 months). The smoothed curve (Supplementary Fig. 1) indicated that hazard ratios for timely versus delayed treatment became statistically significant beyond 3 months (90-days), as the 95 % Confidence Interval dropped entirely below 1. Consistently, Kaplan-Meier survival curves using the 3-month threshold demonstrated a significant survival disadvantage for patients initiating treatment beyond 90-days from diagnosis ($p = 0.05$; Supplementary Fig. 2). These findings suggest that in HER2-positive early-stage breast cancer, delayed initiation of treatment was associated with poorer overall survival, with delays beyond 90-days emerging as a potential threshold for clinically significant treatment delay.

Supervised machine learning and SHAP analysis

To validate and explore the influence of treatment modalities and other covariates from the nomogram on survival outcomes, the authors employed Random Survival Forest (RSF), a supervised machine learning method designed for survival analysis. RSF demonstrated high performance with a concordance index of 0.844 in the training set and 0.812 in the test set. To provide insights into the model's predictions, the authors utilized SHAP analysis, quantifying the impact of each feature on survival predictions, including their interactions. Fig. 5 shows that age, cancer stage, and the total number of in-situ or malignant tumors significantly affected survival predictions, aligning with established clinical knowledge[11,12] Notably, specific treatment modalities also played a crucial role. The absence of systemic therapy, chemotherapy, and non-response to neoadjuvant therapy resulted in higher risk scores, while intraoperative and postoperative radiation lowered the risk score. The impact of partial and total mastectomy on survival risk showed no significant differences, suggesting other factors might be relevant in combination, consistent with previous results.

Discussion

This analysis is based on a large-scale survival cohort and employs an innovative analytical approach that combines supervised and unsupervised machine learning with traditional survival analysis. Utilizing hierarchical clustering, the authors successfully partitioned the dataset based on multiple treatment-related variables, enabling the identification of distinct clusters. Subsequently, the authors examined the differences between these clusters, with a focus on significant variations in treatment-related variables. Cox regression analysis revealed that the treatment-related clusters served as independent risk factors influencing survival outcomes and identified other significant covariates. A comprehensive survival prediction model incorporating specific treatment modalities factors identified from clustering analyses was constructed, providing an individualized prediction tool in the form of a nomogram. Additionally, the random survival forest model within supervised machine learning was utilized, followed by SHAP analysis to gain deeper insights into the importance of each treatment modality's impact on the survival of HER2+ patients with early-stage invasive ductal breast cancer.

Previous studies often focused on comparing the effect of single-treatment modalities, evaluating specific approaches in isolation[13–15] However, breast cancer treatment is a complex and multifaceted process, and the effectiveness of different treatment modalities can vary depending on various factors and their interaction. Understanding the

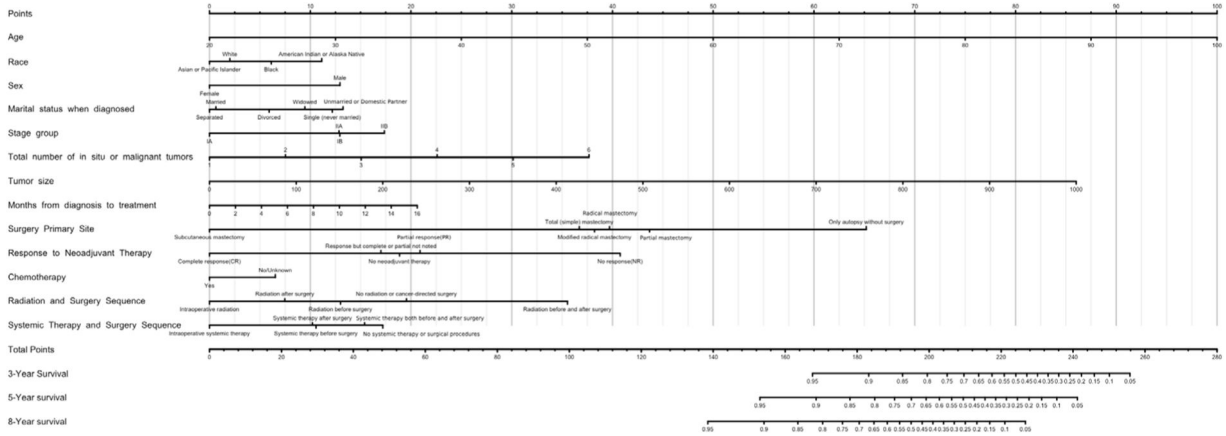


Fig. 3. The nomogram for the 3-year, 5-year, and 8-year overall survival prediction of patients with early-stage HER2+ invasive ductal breast cancer.

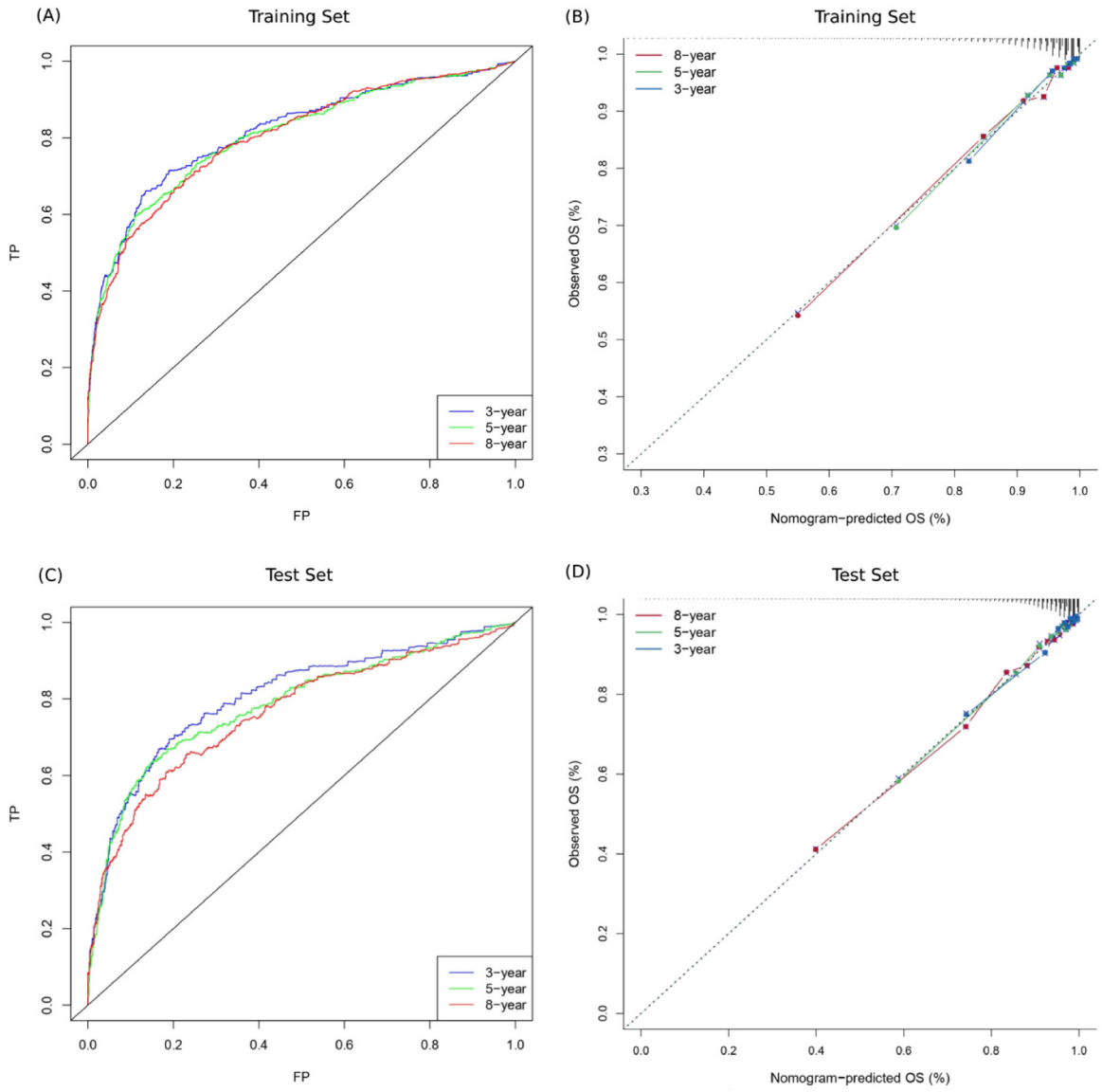


Fig. 4. ROC curves and Calibration curves of 3-year, 5-year, and 8-year overall survival. (A) ROC curve of 3-year, 5-year, and 8-year overall survival in the training set. (B) Calibration curve of 3-year, 5-year, and 8-year overall survival in the training set. (C) ROC curve of 3-year, 5-year, and 8-year overall survival in the test set. (D) Calibration curve of 3-year, 5-year, and 8-year overall survival in the test set.

importance of treatment combinations and sequencing can optimize patient outcomes and survival rates. The present study comprehensively examined various treatment-related factors, revealing associations between survival outcomes and multiple treatment modalities and combinations from a holistic perspective. The unsupervised clustering analysis uncovered distinct treatment-related subgroups without bias, providing unbiased insights into the relationships between treatment modalities and survival outcomes.

As the treatment strategies for patients with early breast cancer are mostly based on surgical treatment, there is extensive debate in academic circles about the impact of partial mastectomy plus radiotherapy or total mastectomy on survival outcomes. Many long-term follow-up studies observed that no significant difference between these two surgery approaches[16,17] aligning with these results. Interestingly, the present research observed favorable survival rates in two clusters of patients undergoing partial mastectomy with radiation therapy. However, the cluster predominantly treated with total mastectomy demonstrated slightly better survival outcomes when combined with effective neoadjuvant therapy, chemotherapy, and systemic treatment. On the other hand, among the two clusters predominantly treated with total mastectomy, one exhibited the worst survival curve, while the other performed the best. It further emphasizes the importance of considering other treatment modalities comprehensively. Existing research supports the present findings, indicating the benefits of a good response to neoadjuvant therapy[18], and application of systemic adjuvant therapies[19] for breast cancer patients' survival outcomes. The present results indicated the limitations of solely focusing on the superiority or inferiority of specific approaches to surgical procedures, as they fail to capture the broader context and complexity of treatment effects.

Additionally, although the superior survival outcomes observed in Cluster 4 are likely multifactorial, the high rate of Complete Response (CR) to neoadjuvant therapy might be a key contributing factor. Given that all patients in the present cohort were HER2-positive, this finding plausibly reflects the clinical benefit conferred by HER2-targeted agents such as trastuzumab and pertuzumab. These therapies are well-established for enhancing CR rates in neoadjuvant settings, which is strongly associated with improved long-term outcomes in early-stage HER2-positive breast cancer[20,21] Although the SEER database does not

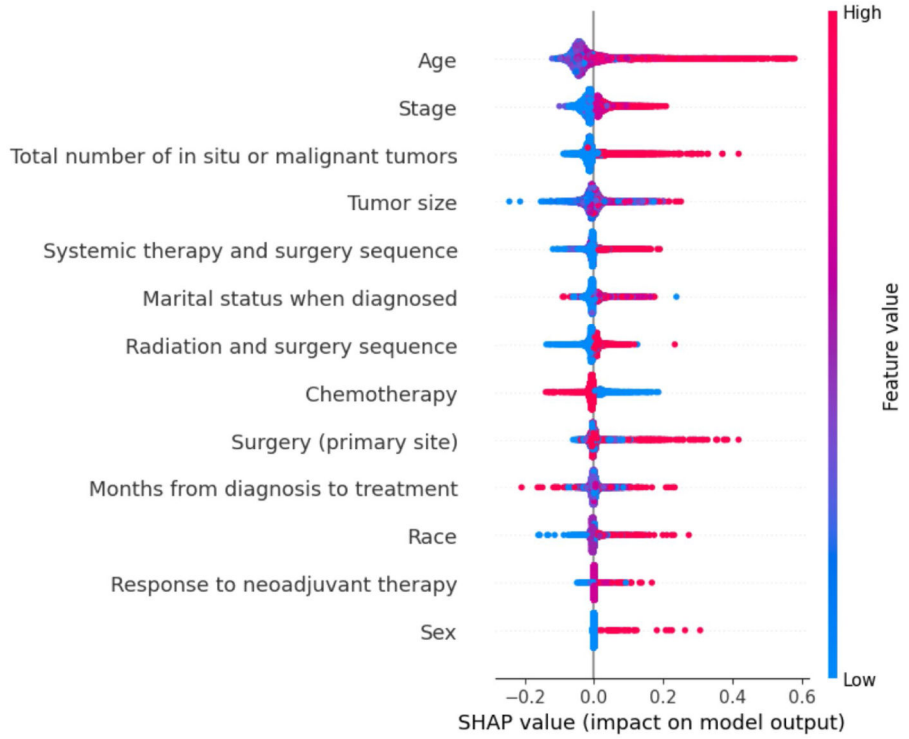


Fig. 5. SHAP summary plot of RSF model.

capture the use of specific biologic agents, the observed response pattern in Cluster 4 is consistent with outcomes expected from HER2 blockade. This evidence might collectively reinforce the pivotal role of HER2-targeted therapy in achieving pathological complete response and long-term survival benefits in HER2-positive breast cancer.

It is noteworthy that, beyond the effects of treatment combinations and sequencing mentioned earlier, this study also suggests that delayed treatment may result in a worse prognosis. This observation aligns with prior research findings. For instance, a modeling study in pregnant breast cancer patients indicates that treatment delays lead to a daily increased risk of axillary metastases by 0.028 % for tumors with moderate doubling times of 130 days and 0.057 % for tumors with rapid doubling times of 65 days[22] An observational, population-based study involving 24,843 patients with stage I to III invasive breast cancer found that delaying the initiation of adjuvant chemotherapy by 91 days or more is associated with adverse outcomes[23] Bleicher RJ observed that a delay of >90 days from diagnosis to surgery (in the non-neoadjuvant setting) occurred in <2 % of patients in America and was associated with a 3.1 %–4.6 % decrease in overall survival[24] Furthermore, it was reported that a delay of >90 days between surgery and the start of adjuvant chemotherapy was associated with an approximately 1.6 times higher risk of death[25] It is worth mentioning that the impact of delayed treatment on patient survival varies by the subtype and stage of breast cancer. Ho PJ et al. found that prolonged treatment delays led to poorer survival outcomes in patients with invasive breast cancer, while no significant impact was observed in patients with noninvasive breast cancer[26] Additionally, Jung SY et al. revealed that delaying treatment initiation for breast cancer in the metastatic stage beyond 12 weeks post-diagnosis was associated with a heightened risk of death[27] Hence, the present findings, along with those of previous studies, underscore the crucial importance of timely intervention following diagnosis for improving the prognosis of breast cancer patients. Delaying intervention is highly likely to worsen prognosis, particularly in patients with invasive breast cancer.

One of the key advantages of this research is the integration of supervised and unsupervised machine learning methods with traditional survival analysis. This innovative approach enhances the identification of hidden treatment-related clusters among diverse variables and their complex interactions, thereby assisting the construction of a prediction model and providing interpretability of the importance of each factor. Additionally, a nomogram with the inclusion of treatment data was developed for individualized prognostication showed good predictive accuracy in the following validation. The nomogram and clustering results suggest that patients undergoing more extensive surgery – such as total or modified mastectomy – may benefit from timely and integrated treatment strategies. These findings also indicate the potential value of prioritizing neoadjuvant therapy in this subgroup to optimize pathological response and long-term survival. Taken together, these results highlight important clinical implications for personalizing prediction therapy in early-stage HER2-positive breast cancer. Healthcare professionals can optimize treatment strategy and improve patient outcomes by incorporating these findings into clinical decision-making.

However, it is essential to acknowledge several limitations of this study. Firstly, as a retrospective analysis of a survival cohort, unmeasured confounding factors may exist. Secondly, the absence of external validation may limit the generalizability of the present findings beyond the studied population and healthcare setting. Future research should validate the present study's model using independent cohorts that capture more comprehensive treatment-related information. Additionally, the SEER database lacks details on the use and timing of HER2-targeted therapies such as trastuzumab and pertuzumab, which limits the ability to directly assess their contribution to treatment response and survival. Further investigation is warranted using datasets with more granular treatment and biomarker data to clarify the effects of specific therapeutic subtypes.

Conclusion

This study integrated novel data analysis methods, unsupervised and supervised machine learning, providing valuable tools to comprehensively interpret the impact of complex treatment modalities on the survival status of HER2+ patients with early-stage invasive ductal breast

cancer. The validated survival prediction model, incorporating detailed treatment factors, may guide treatment strategies in clinical practice for early-stage breast cancer.

Authors' contributions

All authors contributed to the study's conception and design. Data collection and analysis were performed by Kai Wang and Jianing Liu. The first draft of the manuscript was written by Jianing Liu. All authors reviewed this manuscript. All authors approved the final manuscript.

Statement of ethics

As the SEER database is a publicly available database of de-identified patient data, no ethics committee review was required for its use in this project.

Data availability statement

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgments

Open Access funding enabled and organized by Project DEAL.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.clinsp.2025.100818.

References

1. American Cancer Society: Breast Cancer Facts & Figures 2022-2024. In: Annual Publication of the American Cancer Society. Atlanta, Georgia, U.S.; 2022.
2. American Cancer Society: Cancer Facts & Figures 2023. In: Annual Publication of the American Cancer Society. Atlanta, Georgia, U.S.; 2023.
3. Giaquinto AN, Sung H, Miller KD, Kramer JL, Newman LA, Minihan A, et al. Breast cancer statistics, 2022. *CA Cancer J Clin*. 2022;72(6):524–541.
4. Wilkinson L, Gathani T. Understanding breast cancer as a global health concern. *Br J Radiol*. 2022;95(1130), 20211033.
5. Eckhardt CM, Madjarova SJ, Williams RJ, Ollivier M, Karlsson J, Pareek A, et al. Unsupervised machine learning methods and emerging applications in healthcare. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(2):376–381.
6. Jordan JK M, Scholkopf B. *Pattern Recognition and Machine Learning*. Springer; 2006.
7. Menard S, Fortis S, Castiglioni F, Agresti R, Balsari A. HER2 as a prognostic factor in breast cancer. *Oncology*. 2001;61(2):67–72. Suppl.
8. Fazal F, Bashir MN, Adil ML, Tanveer U, Ahmed M, Chaudhry TZ, et al. Pathologic complete response achieved in early-stage HER2-positive breast cancer after neoadjuvant therapy with trastuzumab and chemotherapy vs. trastuzumab, chemotherapy, and pertuzumab: a systematic review and meta-analysis of clinical trials. *Cureus*. 2023;15(5), e39780.
9. Miller KD, Nogueira L, Devasia T, Mariotto AB, Yabroff KR, Jemal A, et al. Cancer treatment and survivorship statistics, 2022. *CA Cancer J Clin*. 2022;72(5):409–436.
10. Noone AM, Lund JL, Mariotto A, Cronin K, McNeel T, Deapen D, et al. Comparison of SEER treatment data with medicare claims. *Med Care*. 2016;54(9):e55–e64.
11. Paluch-Shimon S, Cardoso F, Partridge AH, Abulkhair O, Azim Jr HA, Bianchi-Micheli G, et al. ESO-ESMO 4th international consensus guidelines for breast cancer in young women (BCY4). *Ann Oncol*. 2020;31(6):674–696.
12. Vogt A, Schmid S, Heinimann K, Frick H, Herrmann C, Cerny T, et al. Multiple primary tumours: challenges and approaches, a review. *ESMO Open*. 2017;2(2), e000172.
13. Shumway DA, Corbin KS, Farah MH, Viola KE, Nayfeh T, Saadi S, et al. Partial breast irradiation compared with whole breast irradiation: a systematic review and meta-analysis. *J Natl Cancer Inst*. 2023;115(9):1011–1019.
14. Hasan MT, Hamouda M, Khashab MKE, Elsnhory AB, Elghamry AM, Hassan OA, et al. Oncoplastic versus conventional breast-conserving surgery in breast cancer: a pooled analysis of 6941 female patients. *Breast Cancer*. 2023;30(2):200–214.
15. Hassing CMS, Nielsen DL, Knoop AS, Tvedskov THF, Kroman N, Laenkholm AV, et al. Adjuvant treatment with trastuzumab of patients with HER2-positive, T1a-bN0M0 breast tumors: a systematic review and meta-analysis. *Crit Rev Oncol Hematol*. 2023;184, 103952.
16. Litiere S, Werutsky G, Fentiman IS, Rutgers E, Christiaens MR, Van Limbergen E, et al. Breast conserving therapy versus mastectomy for stage I-II breast cancer: 20 year follow-up of the EORTC 10801 phase 3 randomised trial. *Lancet Oncol*. 2012;13(4):412–419.
17. van Dongen JA, Voogd AC, Fentiman IS, Legrand C, Sylvester RJ, Tong D, et al. Long-term results of a randomized trial comparing breast-conserving therapy with mastectomy: european organization for research and treatment of cancer 10801 trial. *J Natl Cancer Inst*. 2000;92(14):1143–1150.
18. Cortazar P, Zhang L, Untch M, Mehta K, Costantino JP, Wolmark N, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet*. 2014;384(9938):164–172.
19. Early breast cancer trialists' collaborative G. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet*. 2005;365(9472):1687–1717.
20. Early breast cancer trialists' Collaborative G. Trastuzumab for early-stage, HER2-positive breast cancer: a meta-analysis of 13 864 women in seven randomised trials. *Lancet Oncol*. 2021;22(8):1139–1150.
21. von Minckwitz G, Procter M, de Azambuja E, Zardavas D, Benyunes M, Viale G, et al. Adjuvant Pertuzumab and Trastuzumab in early HER2-positive breast cancer. *N Engl J Med*. 2017;377(2):122–131.
22. Nettleton J, Long J, Kuban D, Wu R, Shaefffer J, El-Mahdi A. Breast cancer during pregnancy: quantifying the risk of treatment delay. *Obs Gynecol*. 1996;87(3):414–418.
23. Chavez-MacGregor M, Clarke CA, Lichtensztajn DY, Giordano SH. Delayed initiation of adjuvant chemotherapy among patients with breast cancer. *JAMA Oncol*. 2016;2(3):322–329.
24. Bleicher RJ. Timing and delays in breast cancer evaluation and treatment. *Ann Surg Oncol*. 2018;25(10):2829–2838.
25. Lohrisch C, Paltiel C, Gelmon K, Speers C, Taylor S, Barnett J, et al. Impact on survival of time from definitive surgery to initiation of adjuvant chemotherapy for early-stage breast cancer. *J Clin Oncol*. 2006;24(30):4888–4894.
26. Ho PJ, Cook AR, Binte Mohamed, Ri NK, Liu J, Li J, Hartman M. Impact of delayed treatment in women diagnosed with breast cancer: a population-based study. *Cancer Med*. 2020;9(7):2435–2444.
27. Jung SY, Sereika SM, Linkov F, Brufsky A, Weissfeld JL, Rosenzweig M. The effect of delays in treatment for breast cancer metastasis on survival. *Breast Cancer Res Treat*. 2011;130(3):953–964.