

Machine learning-based survival modeling reveals distinct surgical benefits across molecular subtypes in young women with de novo metastatic breast cancer

Zhiqiong Wang

Wuhan University of Science and Technology

Cheng Zheng

Wuhan University of Science and Technology

Yu Xi

Wuhan University of Science and Technology

Huchen Yu

Wuhan University of Science and Technology

Yingdan Zhu

Wuhan University of Science and Technology

Lifang Huang

huanglifang627@163.com

Huazhong University of Science and Technology

Ling Zhang

Wuhan University of Science and Technology

Research Article

Keywords: De novo metastatic breast cancer, Machine learning, SHAP analysis Surgery

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10816106/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background:

Survival prediction in de novo metastatic breast cancer (dnMBC) remains challenging due to its heterogeneous molecular subtypes, diverse metastatic patterns, and varied treatment responses. This study aimed to develop accurate prognostic models for overall survival (OS) and cancer-specific survival (CSS) prediction in childbearing-age women (≤ 49 years) diagnosed with dnMBC, utilizing advanced machine learning (ML) techniques and robust statistical methodologies.

Methods:

A total of 2,605 women diagnosed with dnMBC from 2010 to 2019 were extracted from the SEER database and randomly divided into training ($n = 1,833$; 70%) and validation ($n = 772$; 30%) cohorts. Eight ML models—Random Survival Forest (RSF), Gradient Boosting Machine (GBM), Lasso-Cox, CoxBoost, Survival SVM, XGBoost, plsRcox, and superpc—were developed and validated. Model performance was evaluated through area under the receiver operating characteristic curve (AUC), concordance index (C-index), calibration plots, Brier scores, and decision curve analysis (DCA). Inverse probability of treatment weighting (IPTW) was applied to assess the independent survival impact of surgical interventions. Shapley Additive Explanations (SHAP) were employed for interpretability.

Results:

RSF and GBM demonstrated superior predictive accuracy compared to other ML models. In the validation cohort, RSF achieved the highest C-index (0.789), along with favorable AUC values at 1-year (0.799), 3-year (0.783), and 5-year (0.749) survival intervals, and a low Brier score of 0.134 at 3 years. SHAP analysis consistently identified molecular subtype, metastatic burden, surgical intervention, and chemotherapy as the most influential predictive factors. Surgical intervention significantly improved OS (weighted hazard ratio [HR]: 0.59; 95% CI: 0.52–0.66, $p < 0.001$) and CSS (weighted hazard ratio [HR]: 0.59; 95% CI: 0.52–0.66, $p < 0.001$), particularly among HR+/HER2+ and HR-/HER2+ subtypes. Notably, patients with combined bone and liver metastases exhibited superior survival outcomes after breast-conserving surgery compared to mastectomy (median OS: 63 vs. 65 months; both superior to non-surgical management at 37 months; $p < 0.001$).

Conclusions:

This ML-based study highlights the prognostic value of molecular subtype, metastatic patterns, and surgical intervention in dnMBC among young women, emphasizing the importance of personalized treatment approaches. Our findings provide robust evidence supporting the integration of ML models in clinical practice, laying a foundation for precision oncology and individualized management of dnMBC. Future prospective trials integrating multi-omics data are warranted to validate these results further.

1 INTRODUCTION

Breast cancer (BC) is the most commonly diagnosed cancer and remains a leading cause of cancer-related mortality among women worldwide[1]. Despite advances in early detection and systemic therapies, approximately 3–8% of breast cancer patients initially present with de novo stage IV disease (dnMBC)[2, 3]. dnMBC is characterized by distant metastases at initial diagnosis, with a poor prognosis and median overall survival (OS) of only 2–3 years[4, 5]. Since dnMBC remains incurable, current treatments primarily aim to relieve symptoms, enhance quality of life, and prolong survival. However, the survival benefit of primary tumor resection remains controversial due to conflicting evidence[6].

Retrospective studies have suggested that primary tumor removal may improve survival outcomes for select dnMBC patients[7, 8], but evidence from randomized controlled trials (RCTs) remains inconclusive. While the MF07-01 trial reported a survival benefit associated with surgery, the Tata Memorial and ABCSG-28 POSYTIME trials failed to demonstrate a significant advantage[9–11]. Additionally, systemic therapies—including chemotherapy, endocrine therapy, and targeted therapy—remain the cornerstone of dnMBC treatment, with surgery often considered only in well-selected cases[12–14]. Some studies propose that surgery may accelerate disease progression by altering tumor microenvironments and facilitating metastatic growth[15, 16]. These conflicting findings highlight the necessity for an individualized, multidisciplinary approach that integrates systemic therapy, local interventions, and patient-specific factors such as tumor biology, metastatic burden, and treatment response [17, 18].

Conventional prognostic models, including TNM staging, provide valuable clinical insights but often fail to account for patient-specific factors such as molecular subtype, metastatic pattern, and response to systemic therapy, all of which significantly influence survival outcomes in dnMBC[19]. Recent advances in machine learning (ML) algorithms have demonstrated superior predictive performance in clinical oncology by leveraging high-dimensional data to provide more precise, individualized survival estimates[20]. Interpretable ML techniques, including LASSO regression, gradient boosting machines (GBM), random survival forests (RSF), and Cox-based models, have shown promise in risk stratification and prognostic modeling, offering greater clinical applicability than conventional statistical methods[21–23].

This study utilizes a large, population-based cohort of childbearing-age female patients with dnMBC (≤ 49 years) from the Surveillance, Epidemiology, and End Results (SEER) database (2010–2019). Our objectives are threefold: (1) develop and validate ML-based prognostic models for OS prediction, (2) assess the independent impact of surgical intervention through the inverse probability of treatment weighting (IPTW)-adjusted survival analyses, and (3) apply Shapley Additive Explanations (SHAP) to identify key predictive features influencing survival outcomes[21].

Additionally, we perform survival stratification based on molecular subtype and metastatic patterns, aiming to generate clinically relevant insights for personalized treatment decision-making[24]. By integrating advanced ML modeling with causal inference methodologies, this study seeks to refine

evidence-based treatment strategies for dnMBC patients, facilitating data-driven, patient-specific clinical decision-making.

2 MATERIALS AND METHODS

2.1 Data source and study design

We extracted data from the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute, a population-based cancer registry covering approximately 30% of the U.S. population. The SEER database is an open-access resource that provides detailed information on tumor characteristics, patient demographics, treatment modalities, and survival outcomes. Data extraction was performed using SEER*Stat software (version 8.3.4; <http://www.seer.cancer.gov/seerstat>), selecting female patients diagnosed with breast cancer from 2010 to 2019. The study period was chosen as SEER began recording detailed human epidermal growth factor receptor-2 (HER2) status and distant metastatic sites in 2010. A detailed overview of the study design and analytical process is illustrated in Fig. 1. As the SEER database contains publicly available, de-identified data for research purposes, ethical approval and informed consent were not required.

Eligible patients were identified using the International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) site codes C50.0 to C50.6, C50.8, and C50.9, along with histological codes 8070 to 8078, representing primary breast tumors. The inclusion criteria were: (1) female patients aged 18 to 49 years at diagnosis; (2) histological diagnosis of infiltrating duct carcinoma (IDC); (3) breast cancer as the first and only primary tumor; (4) stage IV breast cancer based on the American Joint Committee on Cancer (AJCC) TNM staging system; and (5) presence of at least one site of de novo distant metastasis. Patients were excluded if they had incomplete information regarding marital status, race, differentiation grade, tumor size, T stage, N stage, metastatic site, Estrogen receptor (ER), progesterone receptor (PR) status, molecular subtype, or follow-up.

Demographic variables included age at diagnosis (≤ 30 , 31–40, > 40 years), race (White, Black, Other), and marital status (married, unmarried). Unmarried status included divorced, separated, widowed, single/never married, and domestic partners. Tumor characteristics encompassed histology (IDC, invasive lobular carcinoma [ILC], and mixed IDC and ILC), grade (I, II, and III/IV), molecular subtype, tumor size, T stage (T1, T2, T3, and T4), N stage (N0, N1, N2, and N3), and site of metastasis (bone only, brain only, liver only, lung only and Multiple sites). Treatment variables included chemotherapy (yes or no/unknown), radiotherapy (yes or no/unknown), and surgery to the primary tumor (none, mastectomy, or breast-conserving surgery [BCS]). ER status (Negative or positive) and PR statuses were combined as hormone receptor (HR) status. Molecular subtype classifications were defined as HR+/HER2-, HR+/HER2+, HR-/HER2+, and HR-/HER2- based on joint HR and HER2 status.

2.2 Data preprocessing and feature selection

As part of the feature selection process, we employed the LASSO Cox regression model to identify the most relevant predictors while minimizing overfitting. We assessed potential collinearity among variables using correlation analysis. Highly correlated features, such as ER status and molecular subtype, were managed to prevent multicollinearity. This penalized regression approach minimizes prediction error by applying a regularization term that shrinks the coefficients of less important variables toward zero. Variables with nonzero coefficients were retained as significant predictors, while those with coefficients shrunk to zero were excluded. LASSO regression was performed using the “glmnet” package in R. Ten-fold cross-validation was applied to determine the optimal tuning parameter (λ) by selecting the value that minimized the partial likelihood deviance. This method reduced model complexity, enhancing generalizability and minimizing overfitting. The final selected variables were used in subsequent univariate and multivariable Cox regression analyses as well as machine learning model development.

2.3 Model establishment and performance In this study, we randomly partitioned the dataset into a training set (70%) and a validation set (30%) using stratified sampling. The training set was used to develop machine learning (ML) models, while the validation set was exclusively reserved for evaluating their predictive performance. We applied eight widely used ML algorithms, including Random Survival Forest (RSF), Gradient Boosting Machine (GBM), Lasso-Cox Regression (Lasso-Cox), CoxBoost, Survival Support Vector Machine (Survivalsvm), Extreme Gradient Boosting (XGBoost), Partial Least Squares for Cox Regression (plsRcox), and Supervised Principal Component Analysis (superpc), to predict survival outcomes in female patients with de novo stage IV breast cancer of childbearing age. To determine the optimal hyperparameters for these models, ten-fold cross-validation was conducted. The validation set remained unused during model training and hyperparameter tuning, serving solely for independent performance evaluation.

To evaluate model performance, we employed several metrics, including the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, concordance index (C-index), calibration curves, decision curve analysis (DCA), and Brier scores. The Brier score quantifies the deviation between predicted and actual outcomes, with lower values indicating superior predictive performance. By analyzing AUC values at 1-year, 3-year, and 5-year intervals alongside Brier scores, we identified the model with the highest predictive accuracy. Furthermore, feature ranking analysis was conducted to assess the relative importance of individual predictors. To enhance interpretability, Shapley Additive Explanations (SHAP), based on cooperative game theory, were applied to quantify the contribution of each input variable to model predictions. The global SHAP values were visualized as a bar chart to illustrate the average contribution of each feature to the predictive model.

2.4 Statistical analysis

To explore the relationship between clinical and pathological features and patient survival, we first conducted correlation analysis and LASSO regression to select relevant variables. Statistically significant variables were then incorporated into multivariable Cox regression to identify independent prognostic factors. The primary survival outcomes assessed were overall survival (OS) and cancer-specific survival (CSS). Kaplan–Meier curves and the log-rank test were used to estimate survival distributions.

Inverse probability of treatment weighting (IPTW) was used to balance baseline clinical characteristics among patients receiving BCS, mastectomy, or no surgery, ensuring comparability among treatment groups. For IPTW, stabilized weights were calculated by multiplying the original weight by the probability of receiving the treatment received, as derived from the propensity score. Covariate balance after IPTW was assessed using standardized differences, with values below 0.1 indicating acceptable balance. Kaplan–Meier survival analysis stratified by metastatic modality and molecular subtype was also performed on the IPTW-adjusted population. Subgroup survival analyses were further stratified by metastatic patterns to evaluate potential differential effects of treatment according to specific metastatic sites. All statistical analyses were conducted using R software (version 4.1.2) and Python (version 3.7.0). All tests were two-sided, with statistical significance set at $P < 0.05$. Confidence intervals (CIs) were reported at the 95% confidence level.

3 RESULTS

3.1 Baseline characteristics

From the SEER database, 3,922 patients were initially identified using strict inclusion and exclusion criteria, with 2,605 patients diagnosed with dnMBC ultimately included. These patients were divided into training ($n = 1,833$) and validation cohorts ($n = 772$) (Figure 1). Table 1 summarizes the demographic, pathological, and treatment characteristics of the study population. Baseline characteristics showed good balance between training and validation cohorts, with no significant differences ($P > 0.05$ for most variables).. The majority of patients were aged 40–50 years (63.22%), and 69.94% were white. In terms of tumor biology, HR+/HER2– was the most common subtype (52.13%), and 72.32% of patients were ER-positive. Regarding metastatic patterns, bone-only (39.31%) and multiple metastases (39.58%) were the most frequent, while liver-only (11.75%), lung-only (8.33%), and brain-only (1.04%) metastases were less common. For treatment, 81.15% of patients received chemotherapy, 39.62% underwent radiotherapy, and 45.41% underwent surgery (34.13% mastectomy, 11.29% breast-conserving surgery). These baseline characteristics provide essential context for subsequent survival analysis and model development.

3.2 Feature extraction

To ensure the robustness of feature selection, correlation analysis was first conducted to assess potential collinearity among variables (Figure 2A). This analysis revealed a strong correlation between ER status and molecular subtype ($r = 0.73$), leading to the exclusion of ER status to prevent multicollinearity. Following this, the LASSO Cox regression model was applied to identify the most relevant predictors. To determine the optimal tuning parameter (λ), ten-fold cross-validation was performed. The optimal λ value ($\log(\lambda) = -3.0$) was determined via ten-fold cross-validation, balancing model simplicity and predictive accuracy by minimizing mean squared error (Figure 2B). At this point, the model retained nine key predictive variables, including age, tumor grade, molecular subtype, T stage, N stage, and specific metastatic sites, which were subsequently used in survival analyses and machine learning model development (Figure 2C). The relative importance of these selected features, ranked by

their nonzero LASSO coefficients, is depicted in Figure 2D. The most influential predictors were PR status, molecular subtype, surgery, marital status, chemotherapy, race, tumor size, T stage, and grade, with PR status (-0.45), molecular subtype (-0.28), and surgery (-0.21) exhibiting the strongest impact. These features significantly enhanced model performance, improving generalizability and reducing overfitting.

3.3 Univariable and multivariable COX regression analyses

We conducted univariable Cox regression analysis to identify factors significantly associated with overall survival (OS) and cancer-specific survival (CSS) in patients with de novo stage IV breast cancer. The analysis revealed that race, marital status, tumor grade, molecular subtype, metastatic sites, and treatment modalities were significant prognostic factors (Table 2).

To account for potential confounders, multivariable Cox regression was performed, confirming that higher tumor grade and brain metastases were independently associated with worse OS and CSS. Patients with HR-/HER2- and HR-/HER2+ subtypes had significantly poorer survival outcomes than those with HR+/HER2-, while HR+/HER2+ patients exhibited slightly worse survival compared to HR+/HER2-. Regarding treatment, surgery, and chemotherapy were independently linked to improved OS and CSS, whereas radiotherapy did not confer a significant survival benefit. Both BCS and mastectomy significantly improved survival compared to no surgery, with chemotherapy remaining independently protective. Additionally, social determinants played a role in prognosis, with married patients demonstrating better survival outcomes compared to unmarried patients. These findings highlight the prognostic significance of tumor biology, metastatic burden, and treatment strategies, emphasizing the need for personalized therapeutic approaches to improve survival outcomes in this patient population

3.4 Development and comparison of prediction models

To develop a robust prognostic model for patients with de novo stage IV breast cancer, we implemented and compared eight machine learning (ML) models: Random Survival Forest (RSF), Gradient Boosting Machine (GBM), XGBoost, CoxBoost, SurvivalSVM, superPC, LASSO-Cox, and plsRcox. Hyperparameters for each model were fine-tuned using cross-validation, with detailed settings provided in Supplemental Table 1.

Model discrimination was evaluated using receiver operating characteristic (ROC) curves at 1-, 3-, and 5-year survival intervals in both cohorts (Figure 3). RSF and GBM consistently demonstrated the highest AUC values across all time points, with RSF achieving AUC_{12-month} = 0.896, AUC_{36-month} = 0.884, and AUC_{60-month} = 0.791 in the training set, and AUC_{12-month} = 0.799, AUC_{36-month} = 0.783, and AUC_{60-month} = 0.749 in the validation set. GBM exhibited similar performance, with AUC_{12-month} = 0.813, AUC_{36-month} = 0.798, and AUC_{60-month} = 0.773 in the training set, and AUC_{12-month} = 0.807, AUC_{36-month} = 0.782, and AUC_{60-month} = 0.742 in the validation set. Calibration plots (Supplemental Figure 2) were used to assess the agreement between predicted and observed survival probabilities. Calibration plots indicated strong agreement for 1-year survival predictions, though some

discrepancies emerged at longer intervals, suggesting potential model limitations or overfitting (Supplemental Figure 2).

To assess the clinical applicability of the models, Decision Curve Analysis (DCA) (Figure 4) was performed. RSF and GBM provided the greatest net benefit across different threshold probabilities, reinforcing their potential in clinical decision-making. For a comprehensive comparison, AUC, C-index, Brier Score, sensitivity, and specificity were evaluated in both training and validation sets (Supplemental Figure 3). Performance metrics were visualized using heatmaps, with values standardized to enhance comparability due to their close numerical range. Among all models, RSF and GBM consistently achieved the highest C-index and lowest Brier Score, indicating superior predictive accuracy. For example, in the validation cohort, RSF had a C-index of 0.789 and a Brier Score of 0.134 for 3-year survival, while GBM showed a C-index of 0.782 and a Brier Score of 0.141. Given that validation set performance better reflects real-world applicability, model selection primarily relied on validation metrics.

3.5 Model interpretation

To enhance the interpretability of our predictive models, Shapley Additive exPlanations (SHAP) analysis was employed to assess feature importance and visualize the impact of individual variables on survival predictions. Figure 6 provides a comprehensive overview of feature contributions across different modeling approaches. Figure 6A presents a radar plot comparing the relative importance of clinical features across RSF and GBM models for both OS and CSS. SHAP analysis consistently identified molecular subtype, metastatic burden, and surgical intervention as the most influential predictors across models, highlighting their robust contribution to survival predictions. Additionally, chemotherapy, PR status, and marital status contributed significantly, with slight variations in ranking between models. Figure 6B presents a SHAP dependence plot, illustrating the interaction between molecular subtype and metastatic burden. The color gradient represents different metastatic sites, revealing that multiple-site metastases exerted the most detrimental impact on survival, particularly among HR-/HER2- and HR+/HER2+ subtypes. This interaction highlights the complex interplay between tumor biology and metastatic patterns, underscoring the necessity for individualized treatment approaches.

To further explore the influence of key variables, SHAP summary plots for RSF and GBM (Figures 6C and 6D) illustrate the relative contributions of different clinical features. Molecular subtype, metastatic burden, and surgery consistently ranked among top predictors based on SHAP analysis, highlighting their robustness. The results confirm that the molecular subtype had the greatest impact on survival predictions, followed by surgery and metastatic burden. Higher SHAP values for surgery suggest a survival benefit, particularly for patients with HR+/HER2+ subtypes.

3.6 Benefits of surgical treatment in de novo breast cancer patients after IPTW

To evaluate the impact of surgical treatment on survival outcomes, inverse probability of treatment weighting (IPTW) was applied to balance baseline characteristics among patients undergoing BCS, mastectomy, and non-surgical management. Before IPTW adjustment, several covariates, including

marital status, tumor size, tumor grade, T stage, N stage, metastatic patterns, and receipt of chemotherapy or radiotherapy, differed significantly across treatment groups (Table 3). Supplemental Figure 3A - B illustrates the covariate balance before and after IPTW. Before weighting, multiple covariates exhibited absolute standardized mean differences exceeding 0.1, indicating significant baseline heterogeneity among treatment groups. After IPTW adjustment, covariate imbalances markedly improved, with standardized differences reduced to below 0.1, confirming comparability among treatment groups. Stratified comparisons (Supplemental Figure 3C) further demonstrate that IPTW effectively minimized imbalances across key clinical variables, including tumor burden, metastatic sites, and treatment modalities.

Figure 5A presents the adjusted median overall survival (OS) across different clinical subgroups, stratified by surgical treatment. Patients undergoing BCS or mastectomy had significantly longer OS compared to those who did not receive surgery. Specifically, the median OS was 63 months for the BCS group, 65 months for the mastectomy group, and 37 months for the non-surgical group. A similar trend was observed for cancer-specific survival (CSS), reinforcing the survival advantage associated with surgical intervention.

Kaplan-Meier survival curves for OS and CSS before and after IPTW adjustment are depicted in Figure 5B. Before weighting, both surgical groups exhibited a statistically significant survival advantage compared to the non-surgical group. After IPTW adjustment, this survival benefit remained evident, confirming the independent prognostic impact of surgical treatment. The weighted hazard ratios (HRs) for OS were 0.59 (95% CI: 0.50–0.71, $p < 0.001$) for BCS and 0.59 (95% CI: 0.52–0.66, $p < 0.001$) for mastectomy, while the weighted HRs for CSS were 0.59 (95% CI: 0.49–0.71, $p < 0.001$) for BCS and 0.58 (95% CI: 0.52–0.66, $p < 0.001$) for mastectomy.

3.7 Stratified Survival Outcomes by Molecular Subtypes and Metastatic Patterns After IPTW Adjustment

To further examine the impact of surgical intervention across molecular subtypes, we conducted IPTW-adjusted Kaplan-Meier survival analyses, stratified by hormone receptor (HR) and HER2 status (Figure 7). The results demonstrated significant variations in overall survival (OS) and cancer-specific survival (CSS) among subgroups. Surgical treatment conferred a substantial survival benefit in HR+/HER2+ and HR-/HER2+ patients, with both BCS and mastectomy groups achieving significantly prolonged OS and CSS compared to the non-surgical cohort ($p < 0.05$). However, in HR+/HER2- patients, surgical intervention did not yield a significant survival advantage, suggesting a limited benefit in this subgroup.

In contrast, the HR-/HER2- subgroup exhibited heterogeneous survival outcomes following surgery. While mastectomy showed a trend toward improved OS, the difference in CSS was not statistically significant, indicating that the role of surgery in this group may be influenced by factors such as disease burden and response to systemic therapy. Further, stratification by metastatic patterns (Figure 8) revealed that patients with bone-only and liver metastases derived the greatest survival benefit from surgery, whereas those with brain or lung metastases showed minimal or no significant improvement. Patients with multiple metastatic sites generally had a poorer prognosis, though some subgroups still

demonstrated a survival advantage with surgical intervention. Notably, patients with combined bone and liver metastases specifically showed improved survival with BCS compared to mastectomy, distinct from other metastatic patterns (Figure 8).

4 | DISCUSSION

Survival prediction in dnMBC remains challenging due to its diverse molecular subtypes, varying metastatic patterns, and inconsistent treatment responses. In this study, we integrated ML algorithms alongside traditional survival analysis to construct a prognostic model that offers improved risk stratification and clinical applicability. Our analysis demonstrated that RSF and GBM outperformed other ML models in both the training and validation cohorts, exhibiting superior calibration accuracy, discrimination ability, and greater clinical net benefits. Additionally, SHAP analysis identified key predictive features, including molecular subtype, metastatic burden, surgical intervention, and chemotherapy, which were consistently ranked among the most influential variables across multiple analytical approaches. The stability of these predictors across different modeling techniques suggests their robustness in survival prediction and justifies further subgroup analyses based on these characteristics. Given the significant heterogeneity observed in dnMBC, we conducted stratified analyses to explore survival differences across molecular subtypes, metastatic patterns, and treatment modalities, aiming to refine personalized treatment strategies.

To further refine prognostic accuracy beyond traditional parametric models, we compared eight machine learning–based survival models. Parametric methods, which rely on stringent assumptions, are commonly used in survival data analysis. However, capturing nonlinear effects often requires transforming variables or expanding the design matrix to incorporate specialized basis functions[25, 26]. Building upon these traditional approaches, various machine learning–based survival models have been developed to flexibly handle complex data structures without relying on strong parametric assumptions. In this study, we evaluated eight such models: RSF, GBM, Lasso-Cox, CoxBoost, Survival SVM, XGBoost, plsRcox, and superpc. These models differ in their ability to capture nonlinear relationships, manage high-dimensional data, and handle censored survival outcomes. Specifically, RSF and GBM excel at capturing interactions and nonlinear relationships due to their ensemble learning framework, which does not require the proportional hazards assumption (Supplemental Fig. 2). RSF constructs numerous decision trees through bootstrap sampling and random feature selection, enabling it to capture intricate relationships and higher-order interactions that traditional parametric models might overlook[27]. Similarly, GBM iteratively fits weak learners to the residuals along the negative gradient of a loss function, allowing it to adaptively focus on complex patterns and difficult-to-predict samples[28]. These properties make both RSF and GBM particularly robust when dealing with heterogeneous and high-dimensional data.

Cox-based methods, however, often perform suboptimally when significant nonlinearities and complex interactions are present due to their reliance on the proportional hazards assumption.[29]. CoxBoost and plsRcox are designed to address high dimensionality yet still operate under the proportional hazards

assumption, which can limit their flexibility in modeling complex interactions[30]. Superpc, although useful for dimensionality reduction by extracting principal components most correlated with survival outcomes, similarly faces challenges when nonlinearity plays a significant role[31]. Thus, while Cox-based approaches provide valuable insights, their reliance on restrictive assumptions may reduce their predictive performance in settings where the data exhibit substantial nonlinear behavior. Notably, XGBoost, which often excels in disease prediction, demonstrated signs of overfitting, achieving high Integrated AUC but elevated Brier scores in both training and validation sets (Supplemental Fig. 6). These findings underscore the importance of selecting models that accommodate nonlinear effects while maintaining robustness in high-dimensional settings. Rigorous parameter tuning and validation remain critical, as even the most advanced methods can underperform if not properly optimized.

In addition to evaluating various machine learning models, our analyses identified three key clinical factors—molecular subtype, metastatic burden, and surgical intervention—that consistently emerged as significant predictors of survival in dnMBC. These factors were confirmed through univariate and multivariate analyses, as well as SHAP-based feature importance (Table 2, Figs. 6A–6D). The impact of surgical intervention on survival in dnMBC remains a subject of ongoing debate. While retrospective studies have indicated a potential survival advantage, particularly in patients with limited metastatic burden, randomized controlled trials have reported no significant improvement in OS, highlighting the complexity of surgical decision-making in this setting[32, 33]. In our study, we observed that both BCS and mastectomy were associated with significantly improved OS and CSS compared to no surgery, a trend that persisted even after adjusting for confounding variables using IPTW weighting (Table 2, Fig. 5). However, no significant difference was observed between BCS and mastectomy, suggesting that the extent of surgical resection may not be the primary determinant of survival once a patient undergoes surgery.

Several hypotheses may explain this finding. First, previous studies have proposed that surgical removal of the primary tumor may reduce the overall tumor burden, potentially enhancing systemic therapy efficacy by limiting tumor-induced immunosuppression[34, 35]. Additionally, patients who undergo surgery are more likely to receive multimodal therapy, including systemic treatments such as chemotherapy, endocrine therapy, or targeted therapy, which could contribute to their improved survival outcomes[36]. The absence of a survival advantage between BCS and mastectomy aligns with emerging evidence suggesting that tumor biology and systemic disease control play a more decisive role than the type of surgery itself in dnMBC[36]. This finding highlights the need for an individualized treatment approach that prioritizes systemic therapy while considering surgery as a complementary rather than a primary intervention.

Our results are consistent with prior studies using the SEER database and other population-based cohorts, which have reported mixed findings regarding the benefit of surgery in dnMBC. A previous SEER-based analysis (2010–2015) suggested that surgery improved survival in stage IV breast cancer patients but did not differentiate between BCS and mastectomy[33, 37]. Similarly, a propensity-score-matched study from the NCDB database found that surgical intervention was associated with better OS (HR =

0.68, 95% CI: 0.63–0.72, $p < 0.001$), yet no significant difference was observed between different surgical modalities[38]. However, randomized controlled trials such as the MF07-01 trial and ECOG 2108 have challenged this notion, reporting no clear survival advantage of surgery over systemic therapy alone[39, 40]. These discrepancies underscore the complexity of surgical decision-making in dnMBC and the limitations of retrospective analyses in accounting for selection bias.

Our stratified analysis revealed notable variations in surgical benefits across molecular subtypes and metastatic patterns in patients with dnMBC. Specifically, in patients with HR+/HER2 + disease, BCS was superior to mastectomy, a phenomenon not observed in other subtypes. Conversely, in patients with HR-/HER2- and HR+/HER2- subtypes, both BCS and mastectomy demonstrated a significant survival advantage over no surgery, but there was no significant difference between the two surgical approaches. Meanwhile, in the context of metastatic burden, only patients with bone + liver metastases derived a survival benefit from BCS over mastectomy, whereas other metastatic patterns (e.g., bone + brain, bone + lung) showed no statistically significant difference.

These findings partially align with previous studies on the role of surgery in multifocal and multicentric breast cancer (MIBC) and dnMBC. Prior research suggested that the survival benefits of surgery vary depending on tumor biology and metastatic burden[41]. The superiority of BCS over mastectomy in HR+/HER2 + patients may be attributable to the well-documented favorable prognosis of this subtype, which responds effectively to systemic therapy, particularly HER2-targeted and endocrine therapy[42]. Our results complement prior studies indicating that HER2 + patients, despite having a higher incidence of brain and visceral metastases, still derive significant survival benefits from systemic therapy when treated optimally[41]. Conversely, the lack of significant difference between BCS and mastectomy in other molecular subtypes suggests that the extent of local therapy may be less critical than systemic control in certain patients. Prior studies have demonstrated that BCS followed by radiotherapy may be sufficient for local control in many cases, leading to comparable outcomes between the two surgical modalities [43, 44]. The findings that both surgical approaches significantly improve survival over non-surgical management support the concept that locoregional control is still important, albeit with differential benefits depending on tumor subtype.

The differential impact of metastatic burden on survival following surgery was also evident in our study. Only patients with bone and liver metastases significantly benefited from BCS compared to mastectomy. No clear advantage was observed for other metastatic patterns, such as bone and brain or bone and lung metastases. This is consistent with previous studies suggesting that bone-dominant disease, particularly in HR+/HER2 + patients, has a more favorable prognosis compared to visceral-dominant disease[7, 45]. Prior research has also highlighted that the number of metastatic sites plays a crucial role in determining surgical benefit—patients with limited metastatic disease may achieve prolonged survival with aggressive local therapy, whereas those with extensive visceral metastases derive little benefit from surgery[9, 46]. Additionally, our results contrast with some previous findings that suggested mastectomy could be superior to BCS in dnMBC[47]. However, this discrepancy may stem from differences in patient selection, molecular subtypes, and systemic treatment approaches. Studies have shown that HER2 +

patients with well-controlled systemic disease and bone-only metastases may benefit from BCS, whereas those with visceral involvement may not derive a significant advantage from surgery[48]. Moreover, our findings reinforce that the impact of surgical intervention must be interpreted within the context of modern systemic therapy, which has significantly altered survival outcomes across different subtypes [9].

Beyond molecular subtypes and metastatic burden, several demographic and clinical factors independently influenced survival outcomes in our cohort. Consistent with prior studies, tumor differentiation, marital status, and race emerged as significant prognostic determinants[2, 15, 49]. Tumor differentiation remains a key prognostic factor in breast cancer, with our findings consistent with previous SEER-based studies reporting increased mortality in poorly differentiated tumors. This finding aligns with a prior SEER-based study, which reported that patients with high-grade tumors had nearly double the mortality risk compared to those with low-grade tumors[50]. Beyond tumor characteristics, several menstrual and reproductive factors are well-established in influencing breast cancer risk. Variables such as age at menarche, menstrual cycle regularity, cumulative menstrual cycles, menopausal status, and age at menopause, along with reproductive factors like parity, number of full-term pregnancies, age at first birth, and breastfeeding duration, have been shown to modulate risk. For example, a recent meta-analysis reported that delaying menarche by one year reduces breast cancer risk by 4–9%, each additional full-term pregnancy decreases risk by about 7%, and every extra 12 months of breastfeeding is associated with a 4% risk reduction, whereas a higher age at first birth is linked to a 3–5% increased risk[51]. Moreover, earlier age at menarche has been correlated with poorer survival outcomes and is significantly associated with more aggressive tumor features, such as higher grade and greater lymph node involvement[52]. Unfortunately, due to the absence of detailed reproductive histories in the SEER database, we could not further explore how these reproductive factors affect breast cancer prognosis. Future research focusing on these reproductive factors is needed to better understand their impact on breast cancer progression. In our study, race remained a significant prognostic factor, with White patients demonstrating a survival advantage compared to Black and other races. This aligns with a retrospective analysis indicating that Black male breast cancer patients experienced significantly worse 5-year survival compared to other races [53]. However, a separate study reported higher age-adjusted incidence rates among Black patients but did not identify race as a predictor of survival [54]. Similarly, an analysis from the California Cancer Registry showed no significant difference in survival after adjusting for age and stage [55]. These inconsistent findings underscore the complexity of racial disparities in MBC outcomes and highlight the need for more comprehensive, stratified analyses.

Limitations and Future Directions

While our study benefits from a large, population-based dataset and advanced ML modeling, several limitations should be acknowledged. First, as a retrospective SEER-based analysis, our findings are inherently subject to selection bias, despite adjustments using IPTW. Second, the SEER database lacks critical treatment details, such as chemotherapy regimens, HER2-targeted therapies, and endocrine

treatment duration, which may significantly influence survival outcomes. Future prospective trials integrating molecular profiling and treatment data will be essential to refine these predictive models.

Conclusion

Our study provides a comprehensive, ML-driven analysis of dnMBC among childbearing-age women, demonstrating superior prognostic accuracy of RSF and GBM models. SHAP analysis underscored molecular subtype, metastatic burden, surgical intervention, and chemotherapy as key predictive factors influencing survival. Stratified analyses further identified differential surgical benefits across molecular subtypes, particularly in HR+/HER2 + and HR-/HER2 + patients. Additionally, patients with bone and liver metastases uniquely benefited from breast-conserving surgery compared to mastectomy. These findings emphasize the inherent heterogeneity of dnMBC and highlight the critical role of personalized, data-driven therapeutic strategies. Future prospective studies integrating multi-omics data are necessary to validate these models and facilitate personalized clinical decision-making.

Declarations

Ethical Approval and Consent to participate

The present study was based on data obtained from the Surveillance, Epidemiology, and End Results (SEER) database. SEER is a publicly available, de-identified database supported by the National Cancer Institute. The authors obtained access to the data through the official SEER data access procedure and complied with the SEER data-use agreement. Because this study involved secondary analysis of anonymized existing data, it did not require additional ethical approval or informed consent. The authors obtained permission to access the SEER data through the official data access procedure and complied with the SEER data-use agreement. This study was conducted in accordance with the Declaration of Helsinki.

Clinical trial number

not applicable

Consent for publication

This study was based on anonymized data obtained from the Surveillance, Epidemiology, and End Results (SEER) database, which is a publicly available, de-identified database. No individual patient data, images, or other personally identifiable information are presented in this manuscript. Therefore, consent for publication was not required. The publication of this study does not compromise patient confidentiality or privacy.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This work was supported by Research and Innovation Team Project for Scientific Breakthroughs at Shanxi Bethune Hospital, 2024ZHANCHI12 and Hubei Provincial Public Health Leading Talent Program [Ewei Tong (2021) No. 73]. The funder played no role in the design, conduct, or reporting of this study.

Acknowledgements

The interpretation of the data is the sole responsibility of the authors. The authors acknowledge the efforts of the National Cancer Institute and the Surveillance, Epidemiology, and End Results (SEER) Program tumor registries in the creation of the SEER database.

Table

Tables 1- 3 are available in the Supplementary Files section.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209–49.
2. Youlden DR, Cramb SM, Dunn NA, Muller JM, Pyke CM, Baade PD. The descriptive epidemiology of female breast cancer: an international comparison of screening, incidence, survival and mortality. *Cancer Epidemiol*. 2012;36(3):237–48.
3. Barber LE, Maliniak ML, Moubadder L, Johnson DA, Miller-Kleinhenz JM, Switchenko JM, Ward KC, McCullough LE. Neighborhood Deprivation and Breast Cancer Mortality Among Black and White Women. *JAMA Netw open*. 2024;7(6):e2416499.
4. Schroeder MC, Rastogi P, Geyer CE Jr., Miller LD, Thomas A. Early and Locally Advanced Metaplastic Breast Cancer: Presentation and Survival by Receptor Status in Surveillance, Epidemiology, and End Results (SEER) 2010–2014. *Oncologist*. 2018;23(4):481–8.
5. Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, Jemal A, Kramer JL, Siegel RL. Cancer treatment and survivorship statistics, 2019. *CA Cancer J Clin*. 2019;69(5):363–85.
6. Wang X, Wang SS, Huang H, Cai L, Zhao L, Peng RJ, Lin Y, Tang J, Zeng J, Zhang LH, et al. Effect of Capecitabine Maintenance Therapy Using Lower Dosage and Higher Frequency vs Observation on Disease-Free Survival Among Patients With Early-Stage Triple-Negative Breast Cancer Who Had Received Standard Treatment: The SYSUCC-001 Randomized Clinical Trial. *JAMA*. 2021;325(1):50–8.
7. Dawood S, Haaland B, Albaracin C, Gupta S, Cortes J, Sim YY, Dent RA. Is the Proportion of Patients Diagnosed with Synchronous Stage IV Breast Cancer Who Survive More than Two Years Increasing over Time? *Oncology* 2015, 89(2):79–87.

8. Corona SP, Sobhani N, Ianza A, Roviello G, Mustacchi G, Bortul M, Zanconati F, Generali D. Advances in systemic therapy for metastatic breast cancer: future perspectives. *Med Oncol (Northwood Lond Engl)*. 2017;34(7):119.
9. Khan SA, Zhao F, Goldstein LJ, Cella D, Basik M, Golshan M, Julian TB, Pockaj BA, Lee CA, Razaq W, et al. Early Local Therapy for the Primary Site in De Novo Stage IV Breast Cancer: Results of a Randomized Clinical Trial (EA2108). *J Clin oncology: official J Am Soc Clin Oncol*. 2022;40(9):978–87.
10. Warschkow R, Güller U, Tarantino I, Cerny T, Schmied BM, Thuerlimann B, Joerger M. Improved Survival After Primary Tumor Surgery in Metastatic Breast Cancer: A Propensity-adjusted, Population-based SEER Trend Analysis. *Ann Surg*. 2016;263(6):1188–98.
11. Soran A, Ozmen V, Ozbas S, Karanlik H, Muslumanoglu M, Igci A, Canturk Z, Utkan Z, Ozaslan C, Evrensel T, et al. Randomized Trial Comparing Resection of Primary Tumor with No Surgery in Stage IV Breast Cancer at Presentation: Protocol MF07-01. *Ann Surg Oncol*. 2018;25(11):3141–9.
12. Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, Ruddy K, Tsang J, Cardoso F. Breast cancer. *Nat reviews Disease primers*. 2019;5(1):66.
13. Cardoso F, Paluch-Shimon S, Senkus E, Curigliano G, Aapro MS, André F, Barrios CH, Bergh J, Bhattacharyya GS, Biganzoli L, et al. 5th ESO-ESMO international consensus guidelines for advanced breast cancer (ABC 5). *Annals oncology: official J Eur Soc Med Oncol*. 2020;31(12):1623–49.
14. Bevers TB, Niell BL, Baker JL, Bennett DL, Bonaccio E, Camp MS, Chikarmane S, Conant EF, Eghtedari M, Flanagan MR, et al. NCCN Guidelines® Insights: Breast Cancer Screening and Diagnosis, Version 1.2023. *J Natl Compr Cancer Network: JNCCN*. 2023;21(9):900–9.
15. Fitzal F, Bjelic-Radisic V, Knauer M, Steger G, Hubalek M, Balic M, Singer C, Bartsch R, Schrenk P, Soelkner L, et al. Impact of Breast Surgery in Primary Metastasized Breast Cancer: Outcomes of the Prospective Randomized Phase III ABCSG-28 POSYITIVE Trial. *Ann Surg*. 2019;269(6):1163–9.
16. Badwe R, Hawaldar R, Nair N, Kaushik R, Parmar V, Siddique S, Budrukhar A, Mittra I, Gupta S. Locoregional treatment versus no treatment of the primary tumour in metastatic breast cancer: an open-label randomised controlled trial. *Lancet Oncol*. 2015;16(13):1380–8.
17. Al-Sahaf O, Wang JH, Browne TJ, Cotter TG, Redmond HP. Surgical injury enhances the expression of genes that mediate breast cancer metastasis to the lung. *Ann Surg*. 2010;252(6):1037–43.
18. Retsky M, Bonadonna G, Demicheli R, Folkman J, Hrushesky W, Valagussa P. Hypothesis: Induced angiogenesis after surgery in premenopausal node-positive breast cancer patients is a major underlying reason why adjuvant chemotherapy works particularly well for those patients. *Breast cancer research: BCR*. 2004;6(4):R372–374.
19. Park C, Yoon KA, Kim J, Park IH, Park SJ, Kim MK, Jang W, Cho SY, Park B, Kong SY, et al. Integrative molecular profiling identifies a novel cluster of estrogen receptor-positive breast cancer in very young women. *Cancer Sci*. 2019;110(5):1760–70.
20. Ye Q, Chai X, Jiang D, Yang L, Shen C, Zhang X, Li D, Cao D, Hou T. Identification of active molecules against *Mycobacterium tuberculosis* through machine learning. *Brief Bioinform* 2021, 22(5).

21. Sun MS, Yun YY, Liu HJ, Yu ZH, Yang F, Xu L. Brain metastasis in de novo breast cancer: An updated population-level study from SEER database. *Asian J Surg.* 2022;45(11):2259–67.
22. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med.* 2019;380(14):1347–58.
23. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J biomedical health Inf.* 2018;22(5):1589–604.
24. Yu Y, Tran H. An XGBoost-Based Fitted Q Iteration for Finding the Optimal STI Strategies for HIV Patients. *IEEE transactions on neural networks and learning systems* 2022, Pp.
25. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Annals Appl Stat.* 2008;2(3):841–60.
26. Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. *Biostatistics.* 2014;15(4):757–73.
27. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *J Thorac Oncol Official Publication Int Association Study Lung Cancer.* 2008;2(12):841–60.
28. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. *Ann Stat.* 2001;29(5):1189–232.
29. Tibshirani R. The LASSO method for variable selection in the Cox Model. *Stat Med.* 1997;16(4):385–95.
30. Binder H, Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics.* 2008;9:14.
31. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.* 2004;2(4):E108.
32. Winters ZE, Horsnell J, Elvers KT, Maxwell AJ, Jones LJ, Shaaban AM, Schmid P, Williams NR, Beswick A, Greenwood R, et al. Systematic review of the impact of breast-conserving surgery on cancer outcomes of multiple ipsilateral breast cancers. *BJS open.* 2018;2(4):162–74.
33. Yerushalmi R, Tyldesley S, Woods R, Kennecke HF, Speers C, Gelmon KA. Is breast-conserving therapy a safe option for patients with tumor multicentricity and multifocality? *Annals oncology: official J Eur Soc Med Oncol.* 2012;23(4):876–81.
34. Lin Y, Huang K, Zeng Q, Zhang J, Song C. Impact of breast surgery on survival of patients with stage IV breast cancer: a SEER population-based propensity score matching analysis. *PeerJ.* 2020;8:e8694.
35. Wang K, Shi Y, Li ZY, Xiao YL, Li J, Zhang X, Li HY. Metastatic pattern discriminates survival benefit of primary surgery for de novo stage IV breast cancer: A real-world observational study. *Eur J Surg oncology: J Eur Soc Surg Oncol Br Association Surg Oncol.* 2019;45(8):1364–72.
36. Thomas A, Khan SA, Chrischilles EA, Schroeder MC. Initial Surgery and Survival in Stage IV Breast Cancer in the United States, 1988–2011. *JAMA Surg.* 2016;151(5):424–31.

37. Blanchard DK, Shetty PB, Hilsenbeck SG, Elledge RM. Association of surgery with improved survival in stage IV breast cancer patients. *Ann Surg*. 2008;247(5):732–8.
38. Gnerlich J, Jeffe DB, Deshpande AD, Beers C, Zander C, Margenthaler JA. Surgical removal of the primary tumor increases overall survival in patients with metastatic breast cancer: analysis of the 1988–2003 SEER data. *Ann Surg Oncol*. 2007;14(8):2187–94.
39. Rashid OM, Nagahashi M, Ramachandran S, Graham L, Yamada A, Spiegel S, Bear HD, Takabe K. Resection of the primary tumor improves survival in metastatic breast cancer by reducing overall tumor burden. *Surgery*. 2013;153(6):771–8.
40. Danna EA, Sinha P, Gilbert M, Clements VK, Pulaski BA, Ostrand-Rosenberg S. Surgical removal of primary tumor reverses tumor-induced immunosuppression despite the presence of metastatic disease. *Cancer Res*. 2004;64(6):2205–11.
41. Wolters R, Wöckel A, Janni W, Novopashenny I, Ebner F, Kreienberg R, Wischnewsky M, Schwentner L. Comparing the outcome between multicentric and multifocal breast cancer: what is the impact on survival, and is there a role for guideline-adherent adjuvant therapy? A retrospective multicenter cohort study of 8,935 patients. *Breast Cancer Res Treat*. 2013;142(3):579–90.
42. De Mattos-Arruda L, Bottai G, Nuciforo PG, Di Tommaso L, Giovannetti E, Peg V, Losurdo A, Pérez-García J, Masci G, Corsi F, et al. MicroRNA-21 links epithelial-to-mesenchymal transition and inflammatory signals to confer resistance to neoadjuvant trastuzumab and chemotherapy in HER2-positive breast cancer patients. *Oncotarget*. 2015;6(35):37269–80.
43. Janeva S, Krabbe E, Parris TZ, Nasic S, Sundquist M, Karlsson P, Audisio RA, Olofsson Bagge R, Kovács A. Clinical evaluation of molecular surrogate subtypes in patients with ipsilateral multifocal primary breast cancer. *Breast cancer research: BCR*. 2023;25(1):36.
44. Mann GB, Skandarajah AR, Zdenkowski N, Hughes J, Park A, Petrie D, Saxby K, Grimmond SM, Murugasu A, Spillane AJ, et al. Postoperative radiotherapy omission in selected patients with early breast cancer following preoperative breast MRI (PROSPECT): primary results of a prospective two-arm study. *Lancet (London England)*. 2024;403(10423):261–70.
45. Dawood S, Broglio K, Buzdar AU, Hortobagyi GN, Giordano SH. Prognosis of women with metastatic breast cancer by HER2 status and trastuzumab treatment: an institutional-based review. *J Clin oncology: official J Am Soc Clin Oncol*. 2010;28(1):92–8.
46. Curigliano G, Burstein HJ, Winer EP, Gnant M, Dubsy P, Loibl S, Colleoni M, Regan MM, Piccart-Gebhart M, Senn HJ et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. *Annals of oncology: official journal of the European Society for Medical Oncology* 2017, 28(8):1700–1712.
47. Liang Y, Zhang H, Song X, Yang Q. Metastatic heterogeneity of breast cancer: Molecular mechanism and potential therapeutic targets. *Sem Cancer Biol*. 2020;60:14–27.
48. Kennecke H, Yerushalmi R, Woods R, Cheang MC, Voduc D, Speers CH, Nielsen TO, Gelmon K. Metastatic behavior of breast cancer subtypes. *J Clin oncology: official J Am Soc Clin Oncol*.

2010;28(20):3271–7.

49. Lobbezoo DJ, van Kampen RJ, Voogd AC, Dercksen MW, van den Berkmortel F, Smilde TJ, van de Wouw AJ, Peters FP, van Riel JM, Peters NA, et al. Prognosis of metastatic breast cancer subtypes: the hormone receptor/HER2-positive subtype is associated with the most favorable outcome. *Breast Cancer Res Treat.* 2013;141(3):507–14.
50. Kast K, Link T, Friedrich K, Petzold A, Niedostatek A, Schoffer O, Werner C, Klug SJ, Werner A, Gatzweiler A, et al. Impact of breast cancer subtypes and patterns of metastasis on outcome. *Breast Cancer Res Treat.* 2015;150(3):621–9.
51. Breast cancer and. breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50 302 women with breast cancer and 96 973 women without the disease. *Lancet (London England)* 2002.
52. Orgéas CC, Hall P, Rosenberg LU, Czene K. The influence of menstrual risk factors on tumor characteristics and survival in postmenopausal breast cancer. *Breast cancer research: BCR.* 2008;10(6):R107.
53. O'Malley CD, Prehn AW, Shema SJ, Glaser SL. Racial/ethnic differences in survival rates in a population-based series of men with breast carcinoma. *Cancer.* 2002;94(11):2836–43.
54. Simon MS, McKnight E, Schwartz A, Martino S, Swanson GM. Racial differences in cancer of the male breast—15 year experience in the Detroit metropolitan area. *Breast Cancer Res Treat.* 1992;21(1):55–62.
55. Chavez-Macgregor M, Clarke CA, Lichtensztajn D, Hortobagyi GN, Giordano SH. Male breast cancer according to tumor subtype and race: a population-based study. *Cancer.* 2013;119(9):1611–7.

Figures

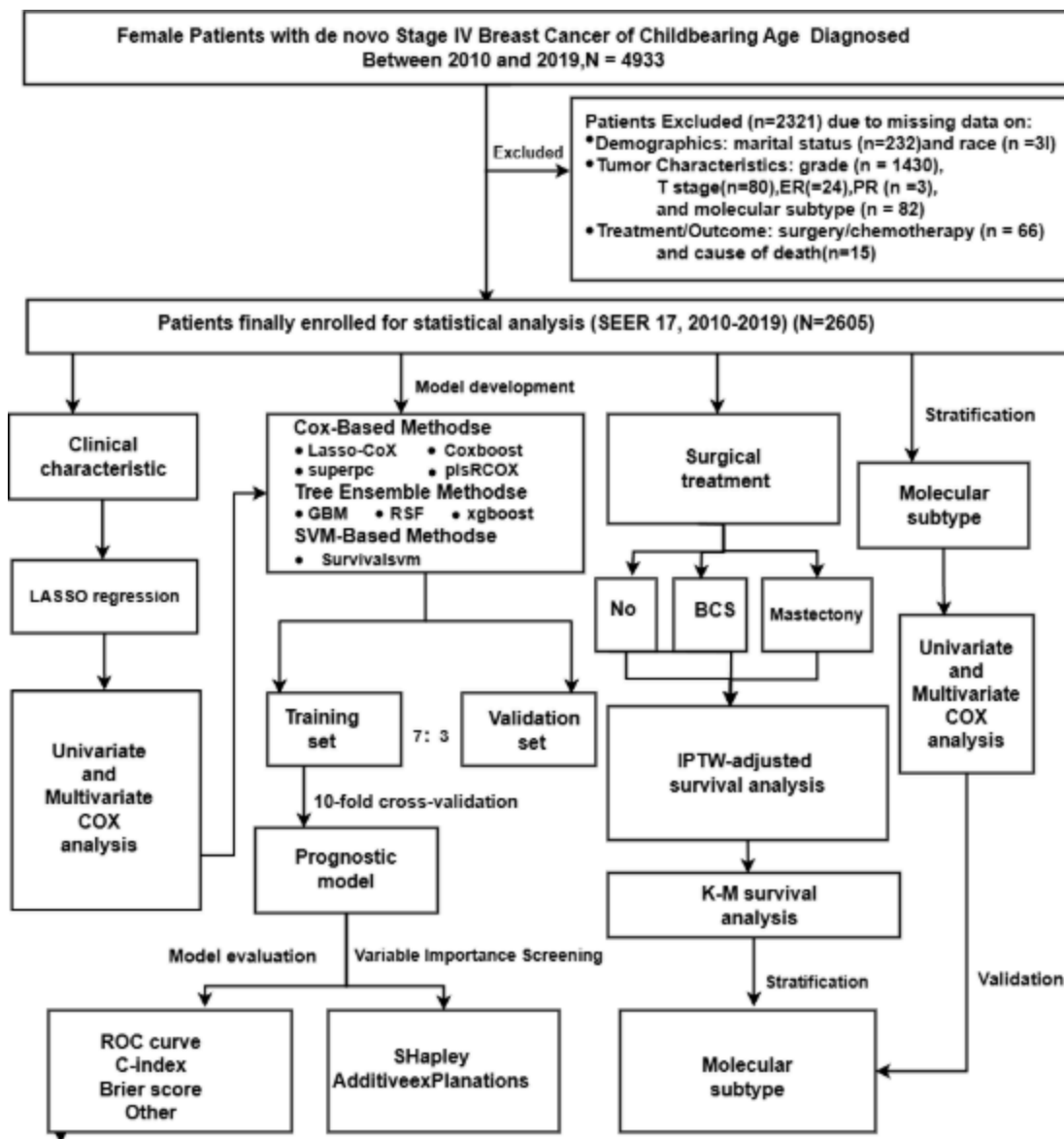


Figure 1

The flowchart of patient selection, model development, and analytical framework for female patients with de novo stage IV breast cancer of childbearing age (SEER 2010–2019).

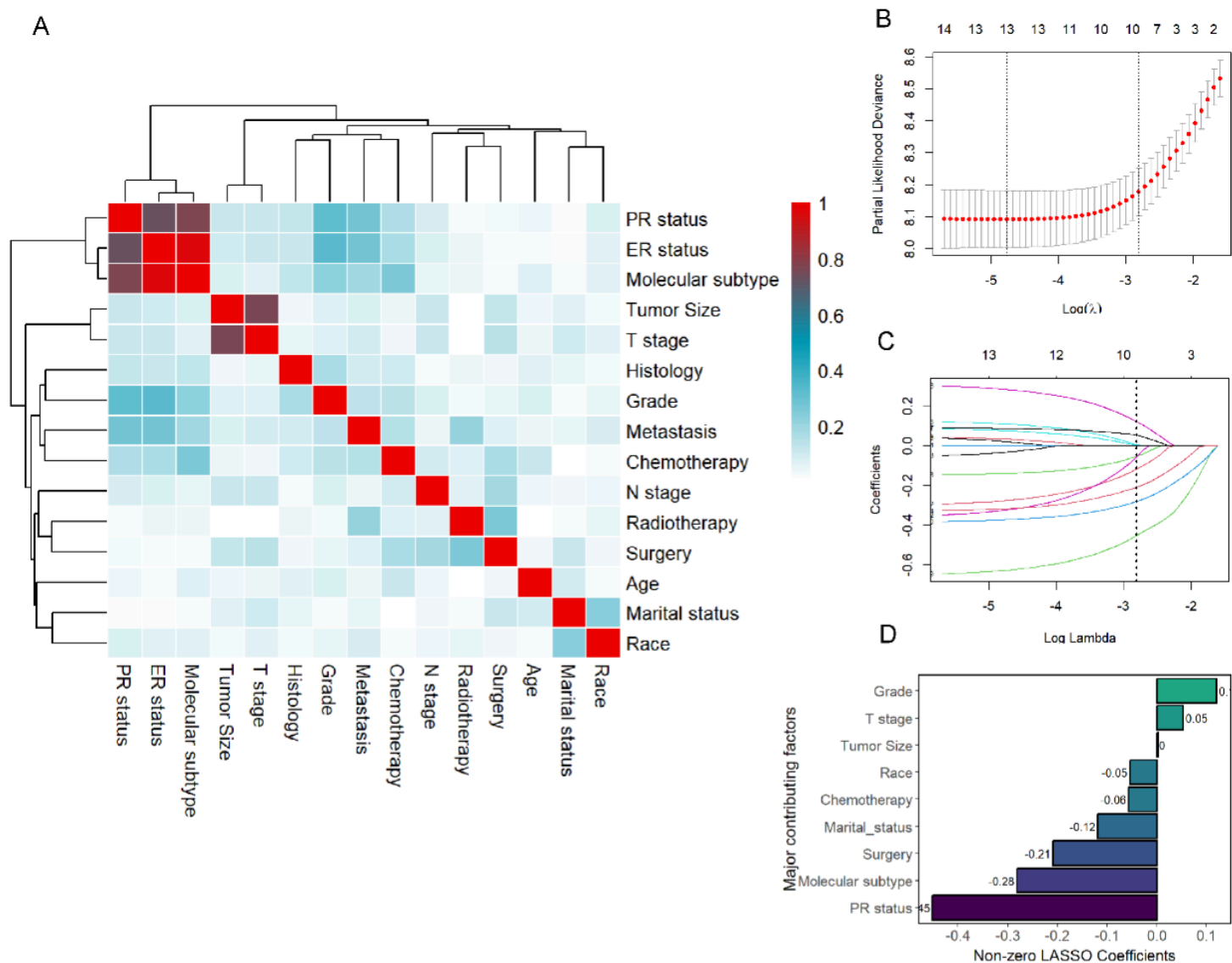


Figure 2

Correlation analysis was first conducted to assess potential collinearity among variables (Figure 2A). balancing model simplicity and predictive accuracy by minimizing mean squared error (Figure 2B). At this point, the model retained nine key predictive variables, including age, tumor grade, molecular subtype, T stage, N stage, and specific metastatic sites, which were subsequently used in survival analyses and machine learning model development (Figure 2C) The relative importance of these selected features, ranked by their nonzero LASSO coefficients—Figure 2D.

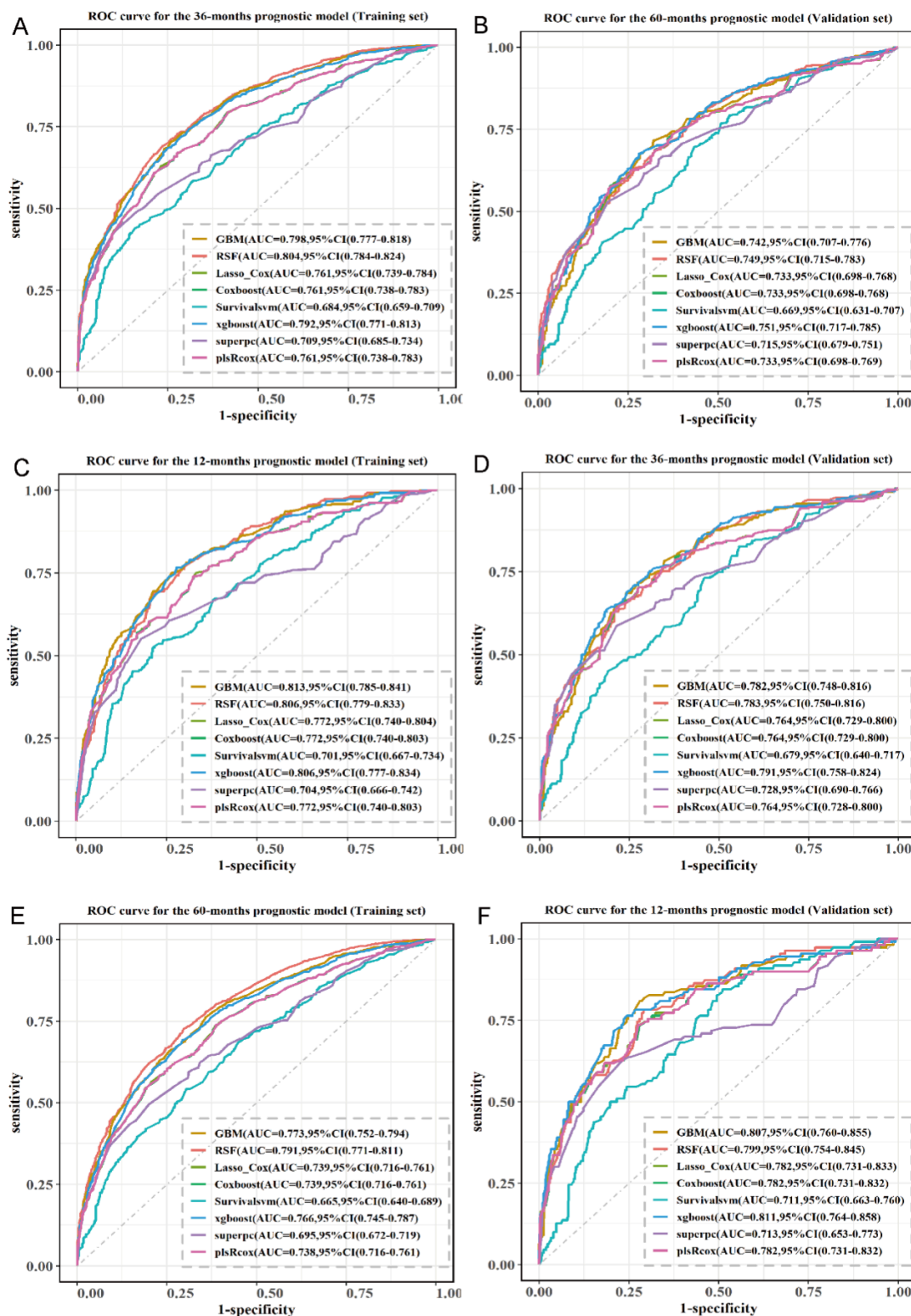


Figure 3

Model discrimination was evaluated using receiver operating characteristic (ROC) curves at 1-, 3-, and 5-year survival intervals in both cohorts. A. ROC curve for the 36-months prognostic model (Training set); B. ROC curve for the 60-months prognostic model (Validation set); C. ROC curve for the 12-months prognostic model (Training set); D. ROC curve for the 36-months prognostic model (Validation set); E.

ROC curve for the 60-months prognostic model (Training set); F. ROC curve for the 12-months prognostic model (Validation set).

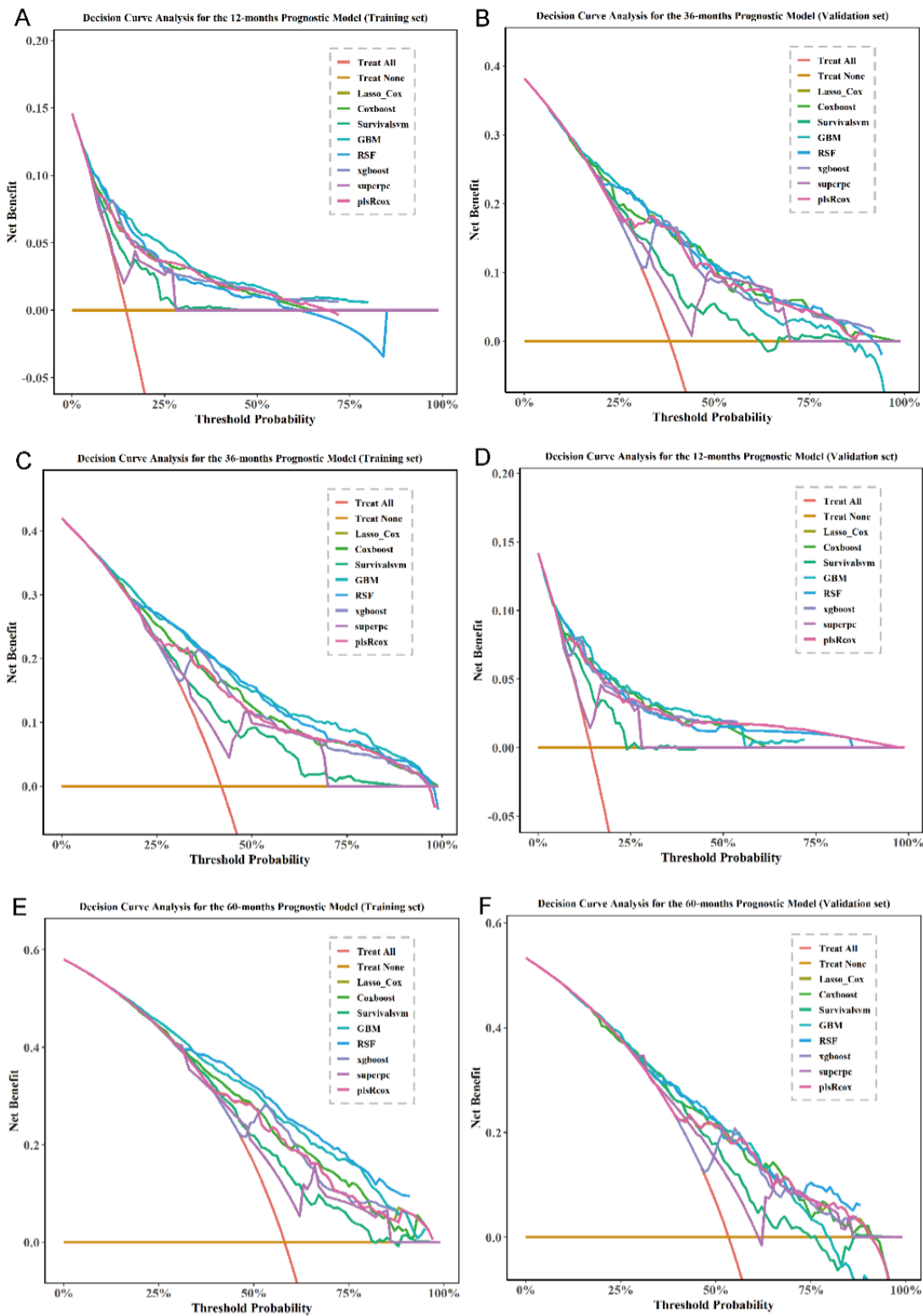


Figure 4

To assess the clinical applicability of the models, Decision Curve Analysis (DCA)

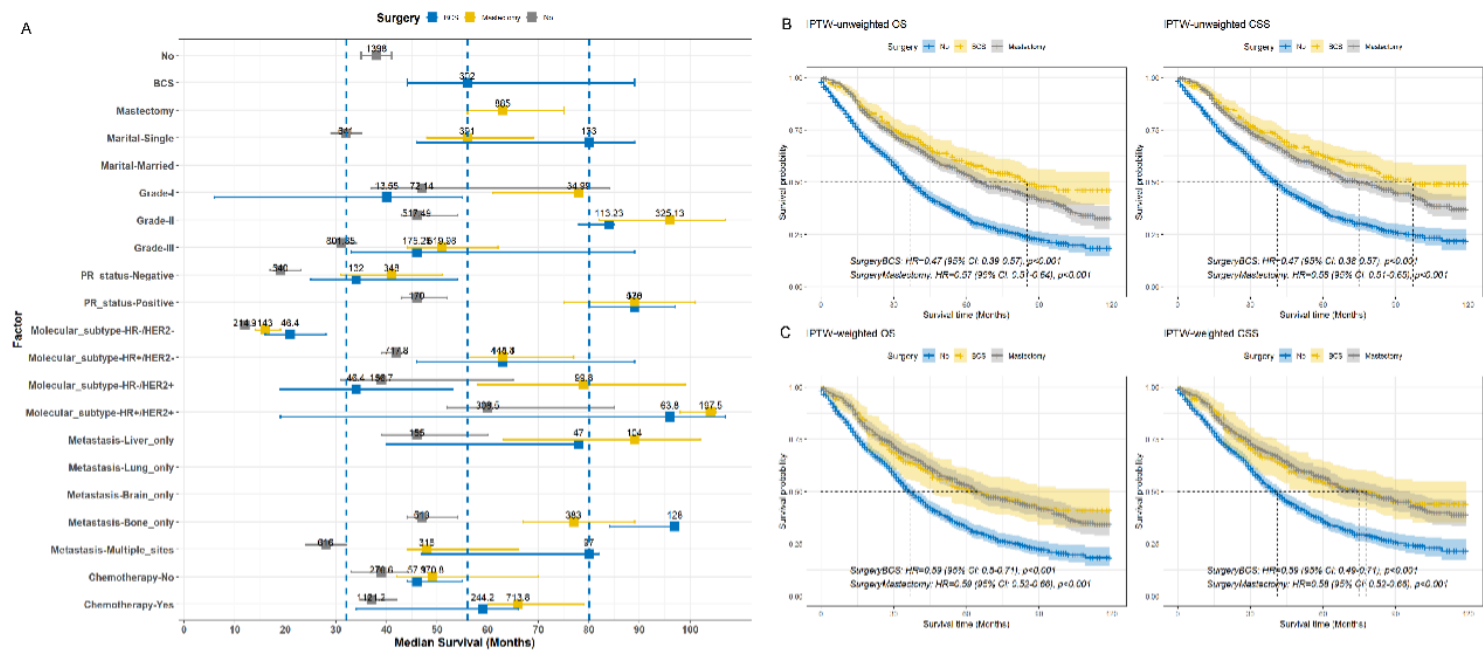


Figure 5

The adjusted median overall survival (OS) across different clinical subgroups is presented in Figure 5A. Kaplan-Meier survival curves for OS and CSS before IPTW adjustment and after IPTW adjustment are depicted in Figure 5B and Figure 5C.

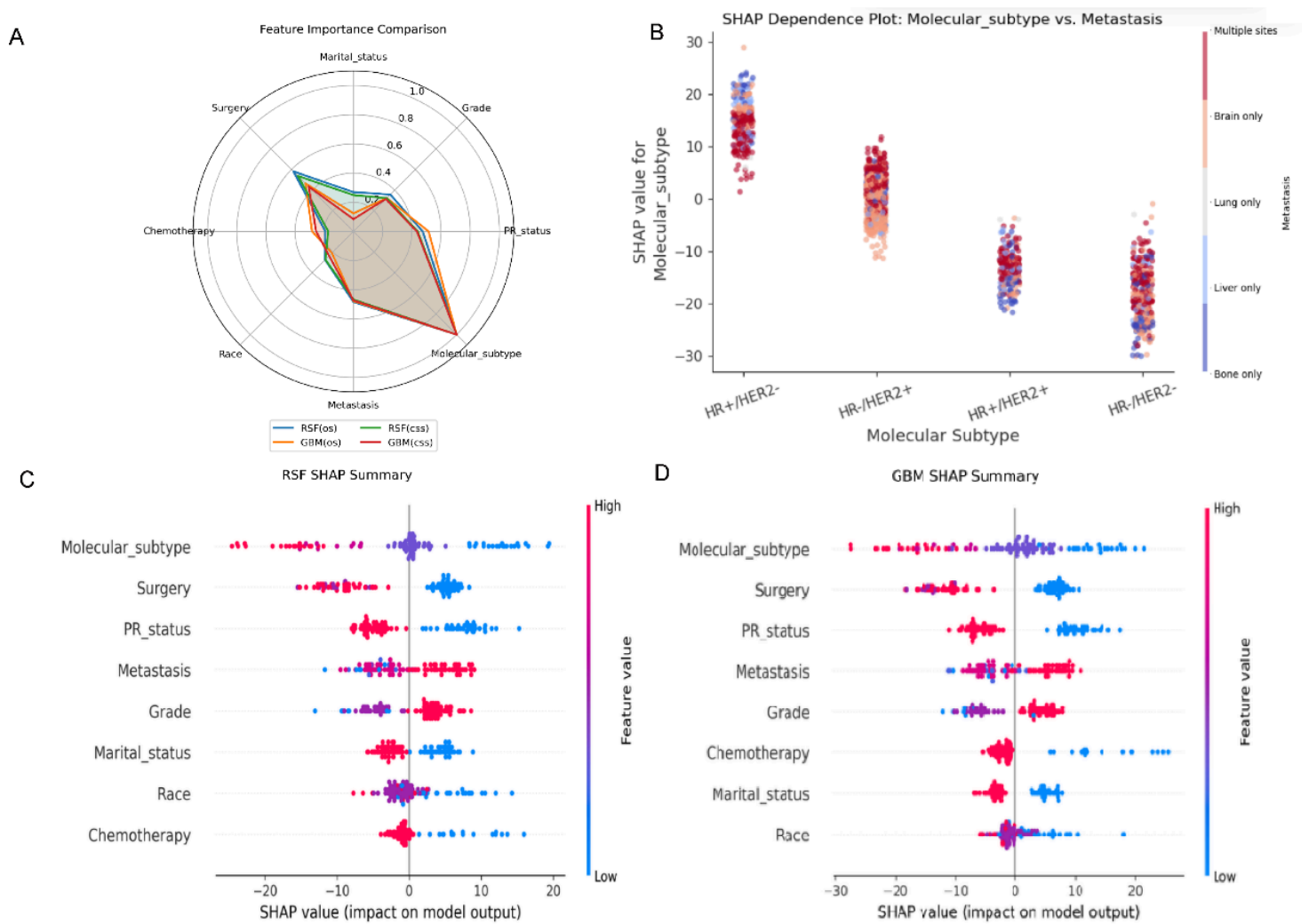


Figure 6

A comprehensive overview of feature contributions across different modeling approaches. Figure 6A presents a radar plot comparing the relative importance of clinical features across RSF and GBM models for both OS and CSS. Figure 6B presents a SHAP dependence plot, illustrating the interaction between molecular subtype and metastatic burden. The color gradient represents different metastatic sites, revealing that multiple-site metastases exerted the most detrimental impact on survival, to further explore the influence of key variables, SHAP summary plots for RSF and GBM (Figures 6C and 6D) illustrate the relative contributions of different clinical features.

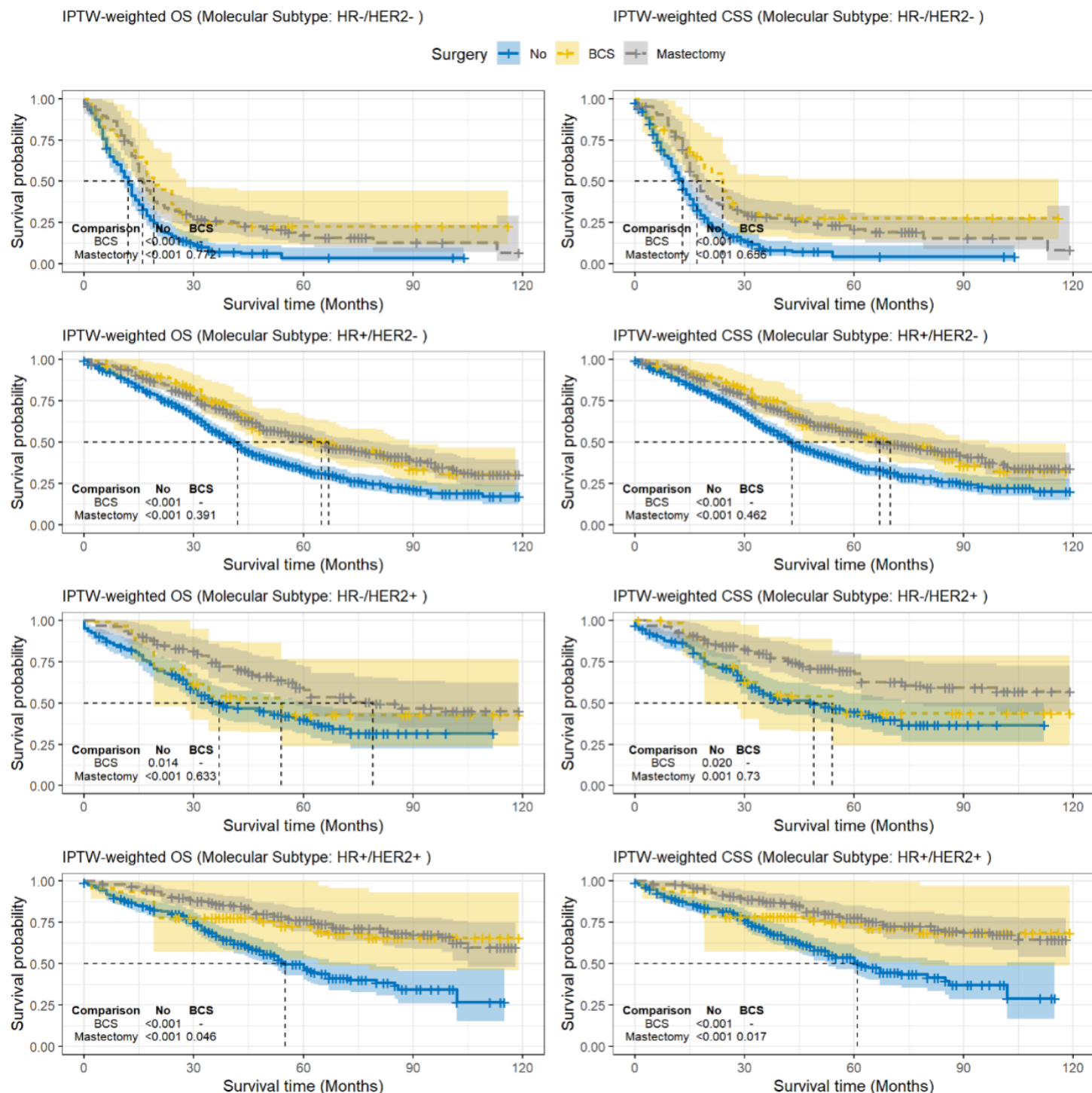


Figure 7

IPTW-adjusted Kaplan-Meier survival analyses, stratified by hormone receptor (HR) and HER2 status.

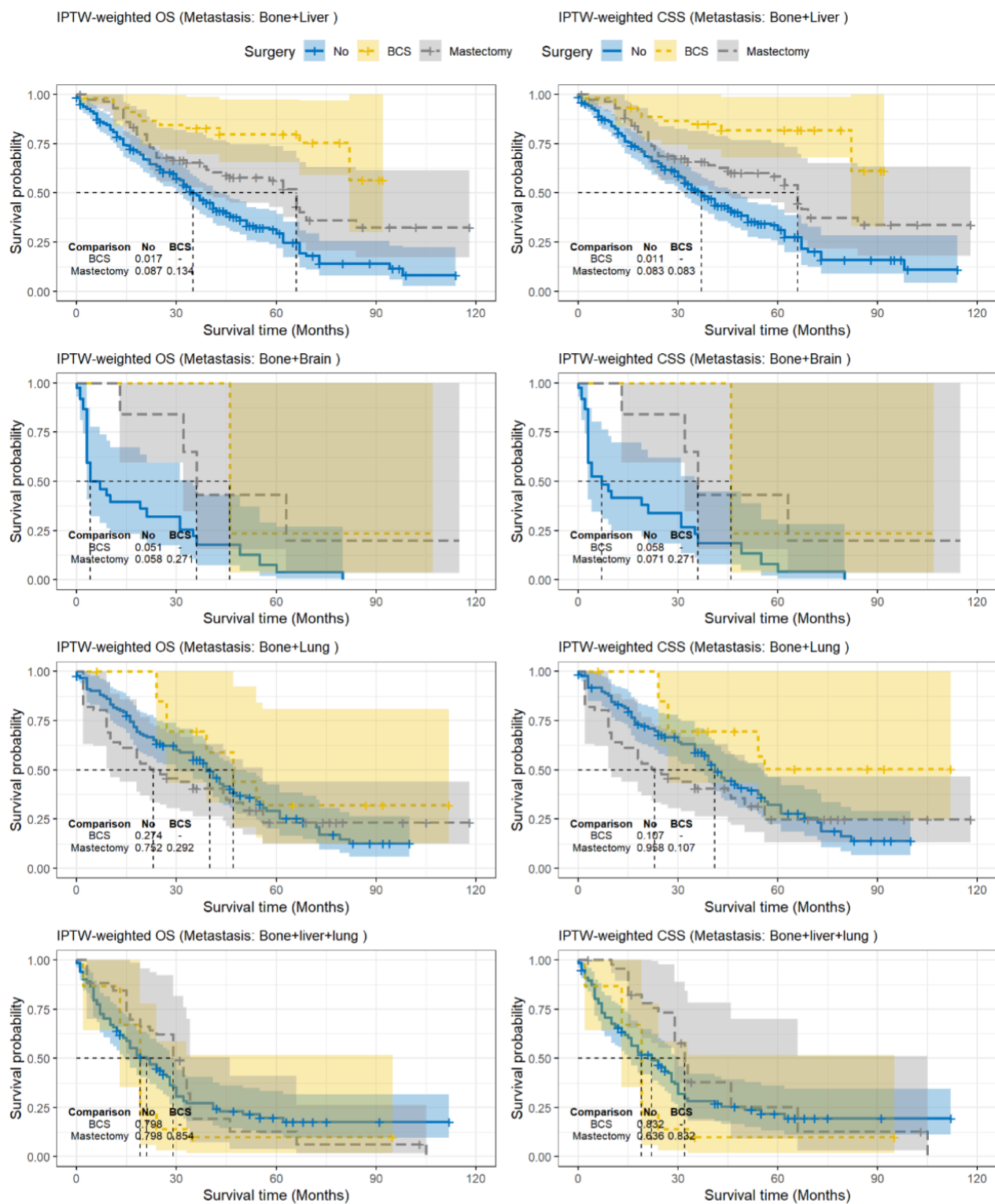


Figure 8

IPTW-adjusted Kaplan-Meier survival analyses, stratification by metastatic patterns.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SF13.doc](#)
- [Tables.docx](#)