Cancer Management and Research

Elevate Series

Open Access Full Text Article

ORIGINAL RESEARCH

Machine Learning-Based Predictive Model for Mortality in Female Breast Cancer Patients Considering Lifestyle Factors

Meixin Zhen[1,*], Haibing Chen[1,*], Qing Lu[1], Hui Li[2], Huang Yan[2], Ling Wang[2]

[1]Xiangya College of Nursing, Central South University, Changsha, Hunan, 410013, People's Republic of China; [2]Nursing Department, The Third Xiangya Hospital, Central South University, Changsha, Hunan, 410013, People's Republic of China

*These authors contributed equally to this work

Correspondence: Ling Wang, Nursing department, Department of Thyroid Breast Surgery, The third Xiangya hospital of Central South University, 138 Tong Zi Po Road, Changsha, Hunan, 410013, People's Republic of China, Tel +86-15274940253, Email 132788113@qq.com

Purpose: To construct a free and accurate breast cancer mortality prediction tool by incorporating lifestyle factors, aiming to assist healthcare professionals in making informed decisions.

Patients and Methods: In this retrospective study, we utilized a ten-year follow-up dataset of female breast cancer patients from a major Chinese hospital and included 1,390 female breast cancer patients with a 7% (96) mortality rate. We employed six machine learning algorithms (ridge regression, k-nearest neighbors, neural network, random forest, support vector machine, and extreme gradient boosting) to construct a mortality prediction model for breast cancer.

Results: This model incorporated significant lifestyle factors, such as postsurgery sexual activity, use of totally implantable venous access ports, and prosthetic breast wear, which were identified as independent protective factors. Meanwhile, ten-fold cross-validation demonstrated the superiority of the random forest model (average AUC = 0.918; 1-year AUC = 0.914, 2-year AUC = 0.867, 3-year AUC = 0.883). External validation further supported the model's robustness (average AUC = 0.782; 1-year AUC = 0.809, 2-year AUC = 0.785, 3-year AUC = 0.893). Additionally, a free and user-friendly web tool was developed using the Shiny framework to facilitate easy access to the model.

Conclusion: Our breast cancer mortality prediction model is free and accurate, providing healthcare professionals with valuable information to support their clinical decisions and potentially promoting healthier lifestyles for breast cancer patients.

Keywords: breast cancer, machine learning, predict model, mortality, lifestyle, SHAP

Introduction

Breast cancer (BC) has emerged as the predominant global malignancy, surpassing even lung cancer, and stands as the foremost cause of cancer-related mortality among women, representing a significant global public health challenge.[1] China accounted for 24% of newly diagnosed cases and 30% of cancer-related deaths worldwide in 2020,[2] imposing a heavy burden on the finance and healthcare systems.[3] Cancer prognosis prediction is critical for determining appropriate treatment strategies. A variety of clinicopathological features, such as tumor size, lymph node status and histological grading, as well as molecular markers, are commonly used to describe the specific characteristics and prognosis of BC.[4] However, traditional clinicopathological factors do not accurately identify low-risk individuals. To achieve more accurate prognostic predictions, a range of prognostic models add genomic information to clinicopathological factors and molecular markers. Unfortunately, the application of polygenic testing is still limited due to its high cost, lack of technology and poor reproducibility.[5] Several studies have shown that in addition to clinical and genomic information, lifestyle factors such as exercise, weight, and diet can affect the prognosis of a BC patient.[6–8] Physical activity influences breast cancer prognosis and survival significantly by regulating biological mechanisms including sex hormones, metabolic hormones, the immune system, and oxidative stress. This direct impact on tumors and their microenvironment supports breast cancer treatment, improves quality of life, and affects related parameters.[9] Meanwhile, obesity leads to poorer prognosis in breast cancer patients by promoting adipocyte dysfunction and secretion

Received: 4 March 2024
Accepted: 12 August 2024
Published: 14 September 2024

of substances that enhance cancer cell proliferation and invasion, alter gene expression, induce inflammation and hypoxia, inhibit apoptosis, and increase chemotherapy resistance.[10] Prolonged sedentary behavior and poor dietary habits lead to excessive caloric intake, eventually resulting in obesity. Therefore, in addition to exercise, a balanced diet is also crucial. Besides exercise, weight, and diet, there are numerous other lifestyle-related variables that have not received sufficient attention in previous studies. Information about the patient's lifestyle can be easily obtained from the hospital's follow-up system. We found that prognostic models of BC developed in previous studies have rarely considered the impact of lifestyle on patient prognosis.[11,12] Considering the potential impact of lifestyle on mortality, we believe it is essential to develop a mortality prediction model that incorporates relevant lifestyle factors to guide personalized treatment and lifestyle advice.

Evidence-based conventional models like Adjuvant! Online and PREDICT are widely employed for predicting breast cancer prognosis in women. These models primarily utilize clinicopathological features and, in some cases, genetic information to estimate risk. However, they often do not incorporate lifestyle factors, despite evidence showing their significant impact on patient outcomes. Moreover, both Adjuvant! Online and PREDICT are predominantly based on Western datasets, limiting their direct applicability to Asian populations, including Chinese patients, due to genetic, environmental, and healthcare system differences. Therefore, there is a growing need to develop breast cancer prognosis models using domestic data specifically tailored for Asian populations, particularly for Chinese individuals.

In recent years, machine learning (ML) has been applied to predict cancer prognosis and optimize treatment strategies.[13,14] However, ML is often considered a black-box model, meaning that humans cannot fully understand the processes that lead to the model's output.[15] With the development of local interpretation methods, we can peer into the black box and explain how predictions are made for each observation. SHapley Additive exPlanations (SHAP) is the most popular method in the local interpretation toolbox, calculating the SHAP value through which the contribution of each feature to the disease outcome can be well explained, making the model more readable. Currently, few BC prognostic models use the SHAP algorithm to identify the feature importance of the models and their impact on the prediction outcomes.[11,12,16] Hence, we used the SHAP algorithm to gain a deeper understanding of the mechanisms behind the model and make them more interpretable.

In summary, there is a current lack of cost-effective and interpretable models for postoperative breast cancer prognosis. Our study utilized 10 years of electronic medical records from Chinese breast cancer patients, incorporating lifestyle information to construct a breast cancer mortality prediction model. Our goal is to create a model that accurately predicts the mortality rate of breast cancer patients by integrating their lifestyle and clinical data, using the SHAP algorithm to identify the most important features and improve interpretability. We will perform external validation using real-world data from breast cancer patients to ensure the robustness of the model. Additionally, we plan to develop a free website based on the best model for use by patients and doctors, helping them assess the prognosis risk of breast cancer patients, choose the most suitable treatment plan, and provide beneficial lifestyle advice. This research aims to provide a convenient and cost-effective tool to improve postoperative breast cancer prognosis.

Material and Methods

Study Population

This was a retrospective study of patients who underwent surgical treatment at the breast surgery department of a large tertiary hospital in Hunan Province, China between January 1, 2010, and September 1, 2020. Although we included follow-up data to improve the accuracy and relevance of our model, all data collection and analysis were performed retrospectively. This approach allowed us to leverage existing data while maintaining the retrospective nature of the study. Eligible patients met the following criteria: (1) pathologically diagnosed with BC; (2) underwent BC-related surgery in the department; and (3) voluntarily participated in the study and signed informed consent. Patients were excluded if they (1) were diagnosed with BC that was not the primary tumor; (2) were concurrently diagnosed with other cancers; (3) died perioperatively; (4) did not receive adjuvant therapy in the department; (5) had missing data on death outcomes; or (6) were male. In total, 1390 female BC patients were included as the primary cohort for developing and internally validating the prediction model. The mean age was 49 years (range: 44–61 years). Using the same inclusion

and exclusion criteria, we selected another 66 female BC patients treated from October 2021 to July 2023 as a validation cohort, with a mean age of 52 years (range: 40–62 years). The study design and use of clinical data were approved by the Ethics Committee of Xiangya Hospital, Central South University. As this was a retrospective study, the requirement for informed consent was waived, and all patient data were anonymized. Our study complies with the Declaration of Helsinki. The flow chart of the study is shown in Supplementary Material eFigure 1.

Measurements

We obtained medical, nursing, and follow-up data from the hospital's electronic medical record system and follow-up system. From the uploaded electronic medical records, we extracted information on patient demographics, tumor characteristics, and treatment details. Demographic data included age, height, sex, education level, and marital status. Comorbidities included diabetes, hypertension, trauma, and others. Menstrual and reproductive factors included age at menarche, menstrual cycle, menstrual duration, parity, number of children, menopause status, and age at menopause. Tumor information comprised pathological classification, size, lymph node metastasis, number of metastatic lymph nodes, pathological staging, and TNM staging. Treatment details included surgery type, surgeon, site of surgical resection, number of resected lymph nodes, endocrine therapy, chemotherapy, radiotherapy, targeted therapy, and catheterization type. Follow-up system data provided information on patients' diet, exercise, mental health, sexual activity, and social support. In total, 246 variables were included in the analysis. TNM staging followed the American Joint Committee on Cancer (AJCC) tumor-node-metastasis (TNM) system.[4] Sexual activity was defined as whether the patient had had sex after surgery. Breast prosthesis was defined as whether the patient ever had a breast prosthesis after surgery. Functional exercises for the affected limb were defined as the number of times the exercises were performed on a weekly basis. Pain was defined as whether the patient considered the pain from the disease to be the greatest difficulty they faced postoperatively. Overall survival (OS) was defined as the time from diagnosis to death from any cause.

Follow-Up

To reduce information bias, we trained professional case managers to follow up with patients. The case manager made the first follow-up visit to the patient before the surgery. One week after surgery, the case manager followed up with the patient a second time. During adjuvant therapy, the case manager followed up with the patient once a month. The patient was followed every three months for two years after surgery, every six months for two to five years after surgery, and annually after five years. The case manager followed the patient until the patient died or was lost to follow-up.

Data Processing

For data preprocessing, the comma-separated values (CSV) dataset was imported. Outliers were identified and handled using z scores. Variables with missing rates exceeding 15% were removed, and the remaining missing values were imputed using multiple imputation with the miceforest package. Since the number of death events (96) was very low compared to the total sample size (1390) (only 7%), there was an imbalance problem in the data. To address this problem, we used the weighting method to correct the bias in the data. Higher weights were assigned to the minority class (mortality events) during the model training phase. This weighting scheme was designed to ensure that the model treated the minority class with greater importance, thus improving its ability to predict mortality events accurately.

Feature Selection

First, a research team of breast surgeons, head nurses, and patients was consulted to remove predict factors that were deemed to be redundant or to have minimal influence on BC survival. Next, variables with 30% missing data were removed. To examine associations between various clinical and pathological characteristics and patient survival, univariate Cox regression models were utilized. Multivariate Cox analysis was further conducted to assess patient mortality risk and To ensure model stability, we employed 10-fold cross-validation to confirm key hyperparameters and generate the optimal model. Additionally, we enhanced performance evaluation by repeating 10-fold cross-validation three times. Each iteration involved randomly dividing the dataset into 10 subsets, using 9 for training and 1 for testing.

Identify independent predict markers. We then selected the most important variables for predicting mortality using the Boruta algorithm.[17]

Validation of the Machine Learning Model

Six different ML algorithms were applied independently to develop predict models in patients with BC, as follows: ridge regression (ridge), k-nearest neighbors (KNN), neural network (NNet), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGB). All ML algorithms were implemented using the R "Caret" package,[18] and each algorithm was validated using three repeated tenfold cross-validations. The "pROC" package for R was used to obtain the receiver operating characteristic (ROC) curve area,[19] and the "survival ROC" package was used for independent time ROC curve analysis.[20] Finally, the best-performing model was designed as a web-based tool for predicting the prognosis of BC patients.

Model Visualization

Machine learning (ML) models often exhibit the characteristics of a "black-box model", meaning that the internal processes and logic leading to predictions are not easily understood. To address this, the SHAP (SHapley Additive exPlanations) algorithm was employed.[21] SHAP quantifies the importance of each input variable in making predictions and explains how each feature affects the overall model. This approach ensures the rationality of the prediction results and enhances the interpretability of the model.

Shiny-Web

The Shiny framework is a web application development framework based on the R language. It allows for the rapid and convenient construction of interactive web applications without the need for frontend technology knowledge, enabling developers to write backend code using only R. The Shiny framework employs a reactive programming approach to achieve interactive web applications. This method allows developers to define the application's UI and backend logic using declarative language. Additionally, Shiny provides many reusable components and tools to help developers build web applications more efficiently. Using the Shiny framework allows R developers to easily present their analysis results to users and quickly build an interactive web application, facilitating user data analysis and interactive exploration.

Statistical Analysis

The R 'CBCgreps' package was used for baseline analysis.[22] Continuous variables are presented as the mean ± standard deviation (SD) or the median for descriptive statistics. For categorical variables, the percentage of patients in each category was calculated. Comparisons between categorical data were conducted using the $\chi 2$ test, while comparisons between continuous variables were performed using the *t* test. A p value of less than 0.05 was considered statistically significant. Model building was performed using R language (version: 4.3.1) to analyze gradient boosting machine (GBM), ridge, KNN, NNet, RF, decision tree (DT), SVM, and XGB models with 10-fold cross-validation. The R "kernelshap" package was used to draw the SHAP interpretation of importance and contribution to the model and interpret the model results by calculating the contribution of each feature to the predicted results.

Results

Clinical Characteristics of BC Patients

In total, 1,390 female BC patients who underwent surgical treatment were included in the study. Seven percent (96) of the patients died, and 93% (1,294) of the patients survived. The median follow-up time was 34 months. Table 1 shows the characteristics of the patients analyzed in this study. The median patient age was 49 years. Fourteen percent of the patients had noninvasive BC, and only 1% of the patients had preoperative distant metastasis. T2 was the most common tumor size (53%), and N0 was the most common lymph node metastasis grade (58%), followed by N1 (23%), N2 (11%), and N3 (8%). Ninety-nine percent of patients did not develop distant metastases. Twenty-one percent of the patients underwent lymph node dissection. In terms of postoperative complications, 2% of the patients had wound dehiscence,

Table 1 Baseline Characteristics of BC Patients

Variables	Total (n = 1390)
Follow-up Time, Median (Q1, Q3)	34 (19.25, 53)
Age, Median (Q1, Q3)	49 (44, 56)
Body Mass Index (BMI), Median (Q1, Q3)	23.11 (21.23, 25.24)
Noninvasive Breast cancer, n (%)	
No	1201 (86)
Yes	189 (14)
Tumor stage (T stage), n (%)	
T1	520 (37)
T2	733 (53)
T3	103 (7)
T4	3 (0)
Tis	10 (1)
T0	4 (0)
Tx	17 (1)
Node stage (N stage), n (%)	
Nx	2 (0)
N0	810 (58)
N1	318 (23)
N2	147 (11)
N3	113 (8)
Metastasis stage (M stage), n (%)	
Mx	4 (0)
M0	1373 (99)
M1	13 (1)
Axillary lymph node dissection (ALND), n (%)	
No	1093 (79)
Yes	297 (21)
Wound dehiscence, infection or nonhealing, n (%)	
No	1361 (98)
Yes	29 (2)
Chemotherapy, n (%)	
No	1137 (82)
Yes	253 (18)
Radiation therapy, n (%)	
No	218 (16)
Yes	1172 (84)
Endocrine therapy, n (%)	
No	533 (38)
Yes	857 (62)
Targeted therapy, n (%)	
No	165 (12)
Yes	1225 (88)
Preoperative light diet, n (%)	
No	767 (55)
Yes	623 (45)
Wound drain placement, n (%)	
No	1350 (97)
Yes	40 (3)
Ports, n (%)	
No	700 (50)
Yes	690 (50)

(*Continued*)

Table 1 (Continued).

Variables	Total (n = 1390)
PICCs, n (%)	
No	1074 (77)
Yes	316 (23)
Functional exercise of affected limb, n (%)	
Everyday	422 (30)
Occasionally	943 (68)
Never	25 (2)
Postoperative sex life, n (%)	
Yes	476 (34)
No	914 (66)
Pain of the disease, n (%)	
No	595 (43)
Yes	795 (57)
Hairpiece (wig), n (%)	
No	1058 (76)
Yes	332 (24)
Prosthetic breast, n (%)	
No	976 (70)
Yes	414 (30)
Light makeup, n (%)	
No	1385 (100)
Yes	5 (0)

infection or nonhealing after surgery, and 7% of the patients had seroma formation. In terms of treatment, 18% of the patients received chemotherapy, 84% received radiotherapy, 88% received targeted therapy, and 38% received endocrine therapy. In terms of venous catheters, 3% of patients had wound drains, 50% of patients had ports placed postoperatively, and 23% of patients had peripherally inserted central catheters (PICCs) placed. Functional exercises for the affected limb were performed daily by 30% of the patients, occasionally by 68%, and never by 2%. Thirty-four percent of patients reported sexual activity after surgery. Fifty-seven percent of patients considered the pain of the disease as the greatest difficulty they faced after surgery. Twenty-four percent of patients wore wigs, 30% wore prosthetic breasts, and 5% wore light makeup.

Univariate and Multivariate Cox Regression Analysis and Boruta Algorithm

We performed univariate Cox regression analysis to identify variables that significantly affect OS in BC patients (Supplementary Material eTable 1). Then, we conducted multivariate Cox regression analysis on the statistically significant variables to eliminate confounding factors and identify the independent factors affecting OS (Supplementary Material eTable 1). The results showed that 12 variables were statistically significant. These variables included noninvasive BC, N stage, M stage, axillary lymph node dissection (ALND), wound dehiscence, infection or nonhealing, endocrine therapy, wound drain placement, ports, functional exercise of the affected limb, postoperative sex life, pain from disease, and prosthetic breast use. Noninvasive BC patients who underwent ALND, received endocrine therapy, had port placement postoperatively, were sexually active, and wore prosthetic breasts had better OS. In contrast, postoperative wound dehiscence or infection and retained wound drains were associated with worse OS. Patients with N3 lymph node metastases had worse OS than patients with N0 and N1 lymph node metastases. Patients with M1 had a worse prognosis than patients with M0. Patients who perceived pain from the disease as the greatest difficulty they faced postoperatively had a worse prognosis than those who did not. Patients who never performed functional exercises for the affected limb had a worse prognosis than those who performed them occasionally. We further screened the variables using Boruta's algorithm, and the results are shown in Supplementary Material eFigure 2. N stage, M stage,

ALND, endocrine therapy, ports, sexual life, and pain from the disease were considered important in predicting prognosis. The variable noninvasive BC was indeterminate. Finally, nine variables were used as candidates for subsequent machine learning model development, including noninvasive BC, N stage, M stage, ALND, endocrine therapy, ports, sex life, prosthetic breast use and pain of disease.

Establishing and Evaluating Predictive Models for Estimating the Prognosis of Patients with BC

Six predictive models for predicting the prognosis of BC patients were developed using the training set data. The average area under the curve (AUC) of the six models determined by three 10-fold cross-validations is shown in Figure 1A, with the RF model (AUC=0.918) performing best. When the models were subjected to external validation (Figure 1B), the RF model also achieved the best performance in terms of prediction (AUC = 0.782) and was therefore selected for the design of the web-based prediction tool. We generated predicted ROC curves and calculated the corresponding AUCs. The predicted results are shown in Figure 2. Our RF model performed exceptionally well in predicting the prognosis of BC patients at 1 year (training set: AUC = 0.914; validation set AUC = 0.809), 2 years (training set: AUC = 0.867; validation set AUC = 0.785) and 3 years (training set: AUC = 0.883; validation set AUC = 0.893). Figure 3 contains the confusion matrix and other model performance for the random forest model. The RF model of the training data (Figure 3A), with 0.914 (95% CI, 0.898, 0.929) accuracy, 0.792 sensitivity, 0.923 specificity, and 0.561 F1-score. And in the external validation data (Figure 3B), with 0.883 (95% CI, 0.774, 0.952) accuracy, 0.4 sensitivity, 0.98 specificity, and 0.533 F1-score. In general, our models behaved efficiently and successfully. Additionally, the model hyperparameters can be found in eTable 2.

Explanation of the RF Model with the SHAP Method

The SHAP algorithm was used to determine the importance of each variable to the outcome predicted by the RF model. Figure 4A shows nine important feature lines. The attributions of all patients to the results are plotted with different colored dots, where yellow dots represent high risk values and purple dots represent low risk values. High N stage, high M stage, and the pain of the disease have a positive impact and push the prediction toward mortality, whereas no sexual activity, no endocrine therapy, no ports, no ALND, no prosthetic breast, and invasive BC have a negative impact and push the prediction toward survival. The variable importance plot lists the most significant variables in descending order. Figure 4B shows that sex activity has the strongest predictive value among all predictive variables, followed by N stage, endocrine therapy, pain of disease, prosthetic breast, ALND, ports, and other variables.

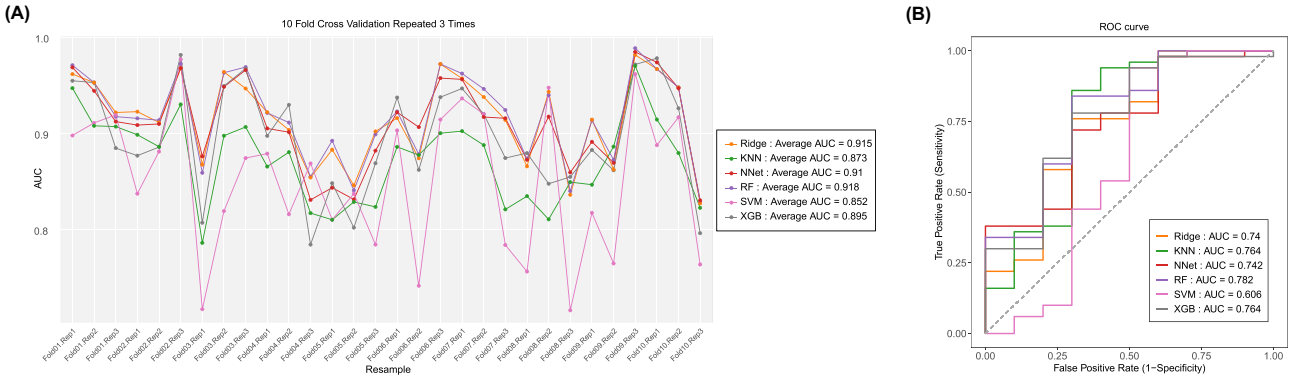

Figure 1 The result of 10-fold cross-validation (training data) and external validation results. (**A**) Average area under the curve (AUC) values of 10-fold cross-validation. Ridge, ridge regression model; KNN, k-nearest neighbors model; NNet, neural network model; RF, random forest predictive model; SVM, support vector machine; XGB, extreme gradient boosting model. (**B**) AUCs of external validation sets.

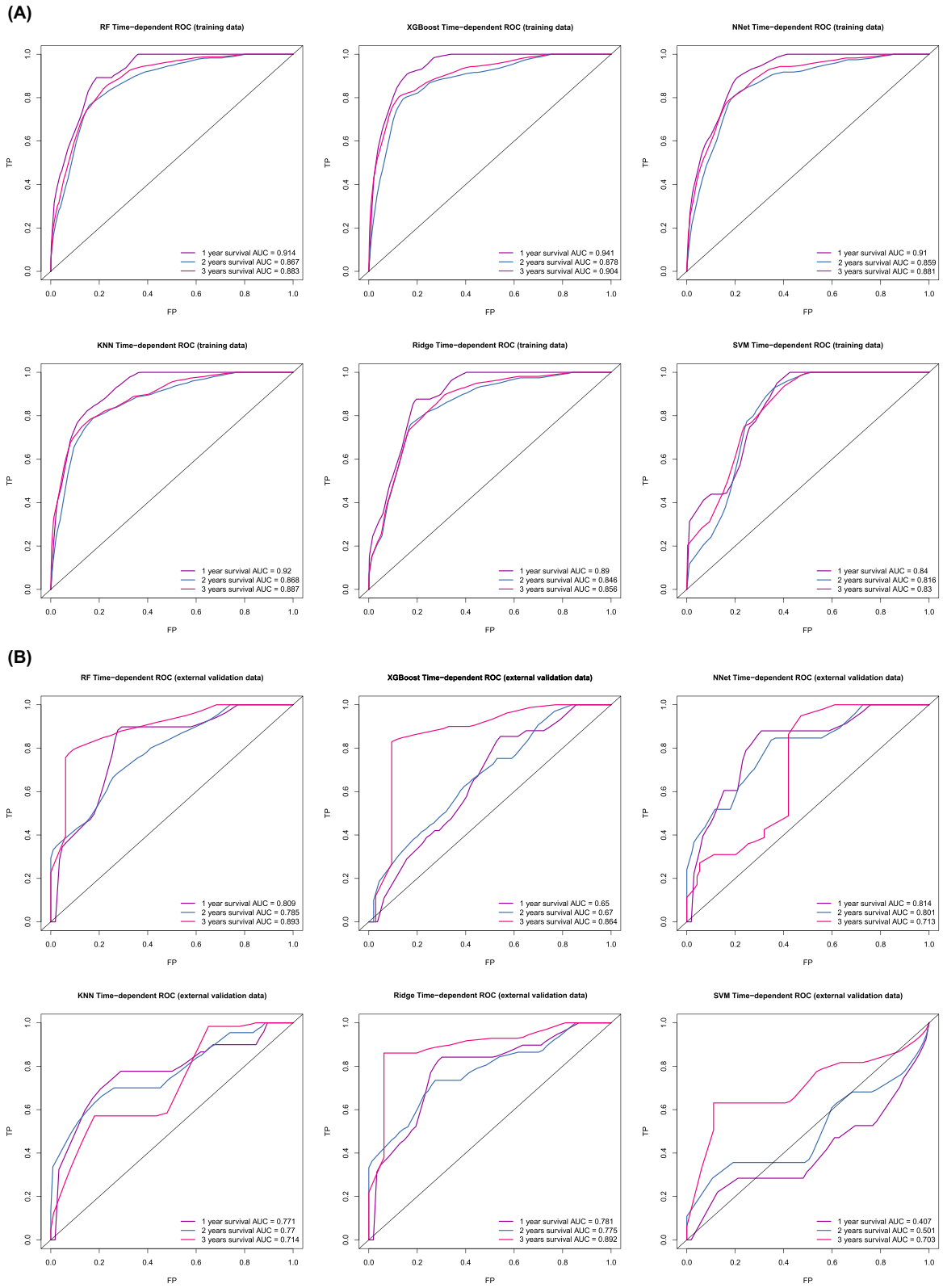

Figure 2 Independent time ROC curve. (**A**) ROC curve for the 1-year, 2-year, and 3-year prognostic models (training data). (**B**) ROC curve for the 1-year, 2-year, and 3-year prognostic models (external validation data).

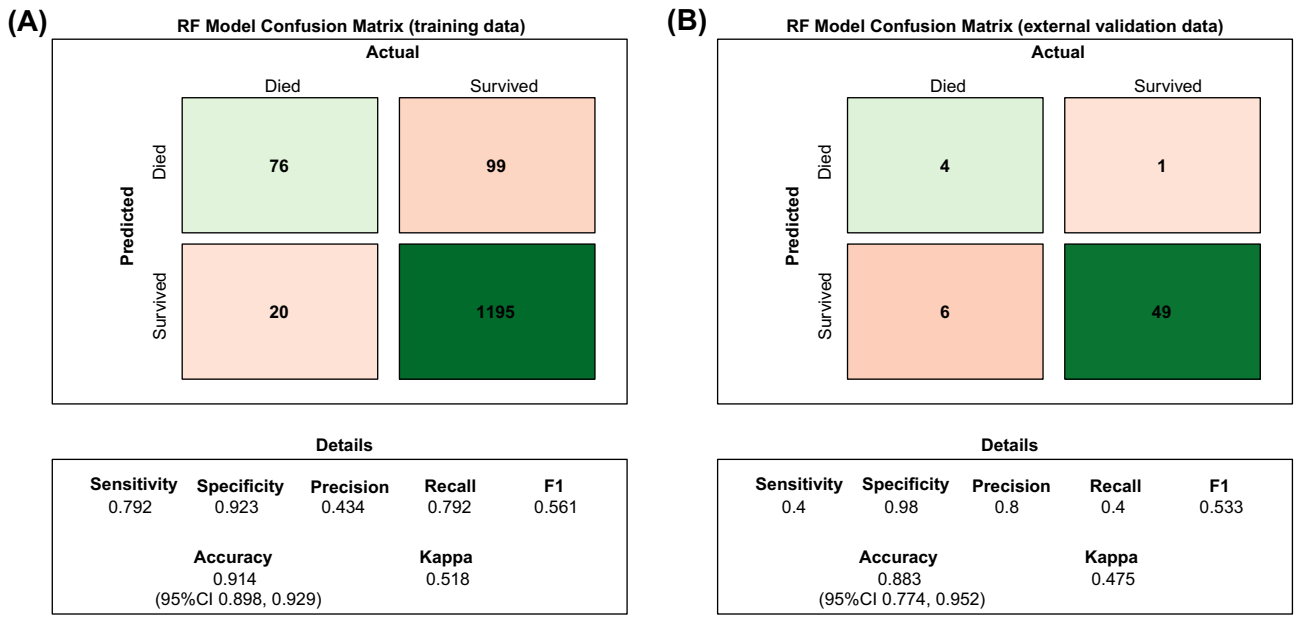

Figure 3 Confusion matrix of the Random Forest model's predicted results and model performance in the training data and external validation data. (**A**) RF model confusion matrix of training data. (**B**) RF model confusion matrix of external validation data.

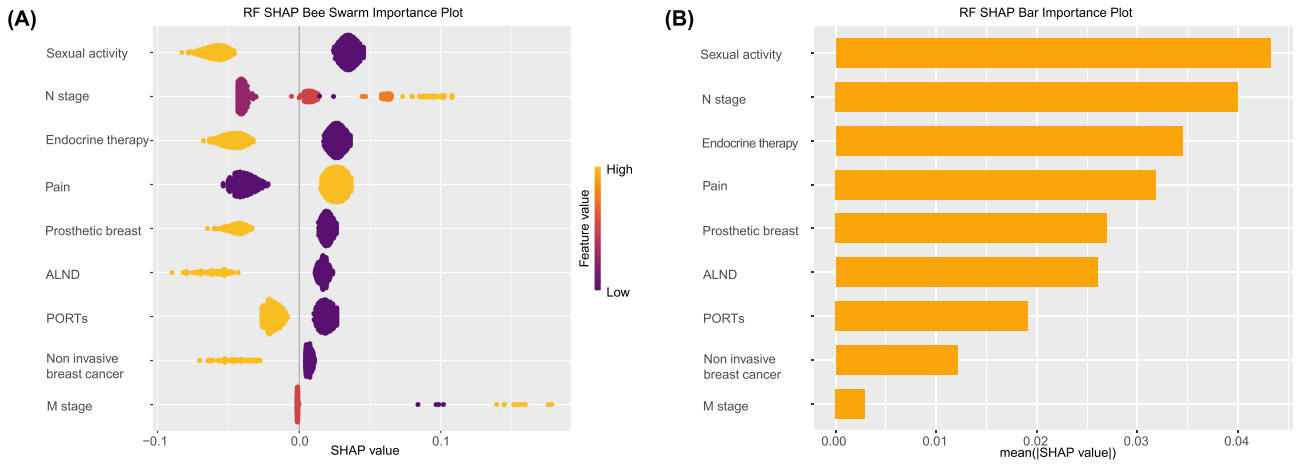

Figure 4 SHAP algorithm results. (**A**) Attributes of characteristics in SHAP. Each line represents a feature, and the abscissa is the SHAP value. Yellow dots represent higher eigenvalues, and purple dots represent lower eigenvalues. (**B**) Feature importance ranking as indicated by SHAP. The matrix diagram describes the importance of each covariate in the development of the final prediction model.

Design of a Web-Based Tool for Predicting the Prognosis of Postoperative BC Patients

The best-performing random forest (RF) model was used to design a web-based tool to assist clinicians in predicting the mortality of postoperative breast cancer (BC) patients. The Shiny-Web interface (https://meixinzhen.shinyapps.io/BC_mortality_predict/) allows users to enter their information, and the application then predicts the probability of death based on the BC patient's information. The website is available in both Chinese and English and includes a user-friendly interface (Supplementary Material eFigure 3) designed to be intuitive and accessible. While formal usability testing and user experience studies were not conducted, the interface's design is based on established principles to maximize ease of use and accessibility. Future work will include comprehensive user testing and feedback to further refine the interface.

Discussion

In this study, we harnessed machine learning to craft an economical and interpretable model that predicts mortality from BC utilizing a dataset from the breast surgery department of a prominent tertiary care hospital. Our model integrates an array of variables, encompassing clinical and pathological features, details of surgical and adjuvant treatments, and lifestyle factors—such as diet, physical activity, sexual health, and the use of wigs, prosthetic breasts, and light makeup. To our knowledge, this research pioneers the inclusion of lifestyle elements in a predictive model for BC survival, alongside traditional clinical and pathological indicators. The implementation of our RF model offers several potential benefits for breast cancer patients. By accurately predicting mortality, the model supports healthcare professionals in tailoring personalized treatment plans and interventions, potentially improving patient outcomes. Furthermore, the integration of lifestyle factors underscores the importance of comprehensive patient care, encouraging healthier lifestyles that positively impact prognosis. The model demonstrated robust predictive power, as evidenced by an average cross-validation AUC of 0.918. In the external validation, the model still showed good performance, indicating the clinical utility of this model. Additionally, the web platform developed using the Shiny framework facilitates easy access to the predictive model for healthcare professionals, allowing them to input patient data and receive mortality risk assessments promptly. Moreover, the platform serves as a tool for data collection, enabling ongoing enhancement of the predictive model through the accumulation and analysis of patient data.

Our analysis identified several key variables for the predictive model: noninvasive BC, N stage, M stage, ALND, endocrine therapy, ports, sexual life, and the use of a prosthetic breast. Notably, noninvasive BC, a lower N stage, undergoing ALND, receiving endocrine therapy, ports, having a sexual life and the use of a prosthetic breast emerged as independent factors linked to a more favorable prognosis. Conversely, M1 stage and the pain of disease were identified as independent predictors of a poorer outcome for BC patients.

Similar to previous findings, lymphovascular infiltration (LVI) is associated with poor prognosis in BC.[23] Additionally, a Chinese study has demonstrated that intraductal carcinoma in situ, microinvasive BC, and mucinous carcinoma of the breast are associated with improved overall survival when compared to invasive ductal carcinoma.[24] In the TNM staging system, the N stage signifies whether the cancer has spread to nearby lymph nodes. Our study is consistent with earlier research, affirming that higher N-stages of BC are indeed associated with a worse prognosis.[25] Furthermore, the M stage is highly correlated with BC prognosis,[26] as supported by various studies. A Korean study reported a 5-year survival rate of only 29.3% for stage IV/M1, which is significantly lower than that for early-stage BC.[27] Within our study, patients with noninvasive BC exhibited a more favorable prognosis than those with invasive BC. Moreover, our analysis revealed that endocrine therapy was an independent protective factor in postoperative BC patients, consistent with prior research.[28,29] In our prediction model, different types of surgery did not emerge as independent predict factors for overall BC survival, corroborating findings from certain previous studies.[28,30] On the other hand, ALND emerged as an independent protective factor in our analysis. This result is substantiated by a meta-analysis demonstrating a significant benefit of ALND in terms of local control of axillary disease and overall survival in patients with invasive BC.[31] In our study, patients who perceived pain from their disease as the greatest difficulty they faced postoperatively tended to have a worse prognosis. Although there is no direct relationship between pain and prognosis, it has been shown that chronic pain after BC surgery may lead to reduced quality of life and additional psychosocial distress.[32] Therefore, we believe that pain relief for patients is necessary to improve their prognosis.

Several lifestyle-related variables were included in our study, including having sex after surgery, wearing prostheses, and retaining ports. In our study, patients who had sex after surgery had a better prognosis. Compared to healthy women, female breast cancer patients exhibit lower satisfaction with sexual activity and face greater difficulties in maintaining sexual life,[33] which affects not only their physiological health—manifesting as pain, reduced lubrication,[34] difficulty in intercourse, decreased libido, and vaginal atrophy[35]—but also brings psychological burdens, including disappointment in partners, sadness, reduced sexual desire, painful sexual experiences, and self-doubt.[36] However, sexual activity plays a significant role in marital relationships. It is not only an important way to express emotions but also a key method to enhance marital harmony and maintain marital health. During intercourse, the body releases β-endorphins, which can stabilize mood and behavior, and alleviate symptoms of anxiety and depression.[37,38] In most predictive models

established in our study, the importance of sexual life consistently ranks first. Therefore, the significance of sexual health in the overall health of BC patients should not be underestimated.[38] Due to the low rate of breast reconstruction surgery, postmastectomy patients usually require the use of an external breast prosthesis, which means wearing a prosthetic breast.[39,40] We found that wearing breast prostheses after surgery was an independent protective factor. A Chinese study showed that breast prostheses can enhance women's self-esteem and self-confidence, restore their social credibility and sense of belonging, and allow them to better participate in sports.[41] We identified ports as a protective factor. To our knowledge, this factor has not been included previously in the Prediction Model for Postoperative BC Patients. Totally implantable venous access ports (TIVAPs) are increasingly being used in patients undergoing chemotherapy and are considered a safe and feasible approach. Some studies have shown that ports tend to be associated with lower complication rates than PICCs,[42,43] particularly reduced thrombosis rates. Another study showed that fully implantable venous access ports (TIVAPs) were associated with higher cosmetic outcomes, increased awareness of body image protection in women, and higher patient satisfaction. While direct evidence linking TIVAPs to breast cancer prognosis is lacking, there is a belief that port implantation could potentially enhance survival in breast cancer patients. However, this hypothesis requires validation through large prospective studies.

In the Cox regression analysis, a preoperative light diet was not significantly associated with outcomes, and it may be that the reliability of the model estimates may be reduced due to the inclusion of more variables ($P>0.05$).[44] Cox regression showed that functional exercise of the affected limb was statistically significant, but functional exercise of the affected limb was considered unimportant according to the Boruta algorithm. This may be because the Boruta algorithm focuses on the predictive importance of the variables, whereas the emphasis in Cox regression focuses on the association between the variables and the outcome. To avoid overfitting and improve the generalization of the model, the Boruta algorithm may have excluded some variables.[17]

Limitations

Although our study produced many important findings, this investigation of ML-based models for predicting post-operative prognosis in BC patients had some limitations. (1) The retrospective nature of our study may have led to sample selection bias. (2) This model does not consider the molecular features of BC, such as HER2i67, Ki-67, ER, and PR, which limits the predictive value of the developed model. (3) Our study data used data from China and did not contain data from other countries and races. Nevertheless, our model is the first machine learning model to include lifestyle information to predict the prognosis of BC patients. Moreover, our model exhibited a high AUC. (4) The size of the external validation cohort (n = 66) was constrained by the availability of eligible patients within the specified timeframe. Despite this, the external validation provided valuable insights and demonstrated the model's robustness. Future studies with larger validation cohorts would be beneficial to further confirm the generalizability of our findings.

Conclusion

In conclusion, an affordable and interpretable random forest prediction model based on patients' lifestyle and clinical information was constructed. The model exhibited high performance in predicting the prognosis of patients after BC. The SHAP-based model can be used to accurately explore risk factors for BC patients and guide individualized treatment and lifestyle decisions. Additionally, the web platform developed using the Shiny framework facilitates easy access to the predictive model for healthcare professionals, allowing them to input patient data and receive mortality risk assessments promptly.

Data Sharing Statement

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

Ethical Approval

The current study was approved by the Institutional Review Board of Third Xiangya Hospital of Central South University (2021-S065), which waived the requirement for patient informed consent due to the anonymous nature of the data. All methods were carried out in accordance with relevant guidelines and regulations.

Acknowledgments

We would like to thank all the developers of the R programming package for selflessly sharing their code. Additionally, this paper is also available as a preprint on ResearchGate at https://www.researchgate.net/publication/376466145_Machine_Learning-based_Predictive_Model_for_Mortality_in_Female_Breast_Cancer_Patients_Considering_Lifestyle_Factors.

Funding

This research was supported by the Hunan Provincial Natural Science Foundation of China (2019JJ40473) and Hunan Traditional Chinese Medicine Research Project (B2022072).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209–249. doi:10.3322/caac.21660
2. Cao W, Chen HD, Yu YW, Li N, Chen WQ. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. *Chinese Med J*. 2021;134(7):783. doi:10.1097/CM9.0000000000001474
3. Lei S, Zheng R, Zhang S, et al. Breast cancer incidence and mortality in women in China: temporal trends and projections to 2030. *Cancer Biol Med*. 2021;18(3):900–909. doi:10.20892/j.issn.2095-3941.2020.0523
4. Giuliano AE, Edge SB, Hortobagyi GN. Eighth Edition of the AJCC Cancer Staging Manual: breast Cancer. *Ann Surg Oncol*. 2018;25(7):1783–1785. doi:10.1245/s10434-018-6486-6
5. Min N, Wei Y, Zheng Y, Li X. Advancement of prognostic models in breast cancer: a narrative review. *Gland Surg*. 2021;10(9):2815–2831. doi:10.21037/gs-21-441
6. Parada H, Sun X, Tse CK, Olshan AF, Troester MA. Lifestyle Patterns and Survival Following Breast Cancer in the Carolina Breast Cancer Study. *Epidemiology*. 2019;30(1):83–92. doi:10.1097/EDE.0000000000000933
7. Hamer J, Warner E. Lifestyle modifications for patients with breast cancer to improve prognosis and optimize overall health. *CMAJ*. 2017;189(7):E268–E274. doi:10.1503/cmaj.160464
8. Łukasiewicz S, Czeczelewski M, Forma A, Baj J, Sitarz R, Stanisławek A. Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review. *Cancers*. 2021;13(17):4287. doi:10.3390/cancers13174287
9. Ortega MA, Fraile-Martínez O, García-Montero C, et al. Physical Activity as an Imperative Support in Breast Cancer Management. *Cancers*. 2020;13(1):55. doi:10.3390/cancers13010055
10. Chu D-T, Nguyen Thi Phuong T, Tien NLB, et al. The Effects of Adipocytes on the Regulation of Breast Cancer in the Tumor Microenvironment: an Update. *Cells*. 2019;8(8):857. doi:10.3390/cells8080857
11. Wishart GC, Azzato EM, Greenberg DC, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res*. 2010;12(1):R1. doi:10.1186/bcr2464
12. Jonczyk MM, Fisher CS, Babbitt R, et al. Surgical Predictive Model for Breast Cancer Patients Assessing Acute Postoperative Complications: the Breast Cancer Surgery Risk Calculator. *Ann Surg Oncol*. 2021;28(9):5121–5131. doi:10.1245/s10434-021-09710-8
13. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2014;13:8–17. doi:10.1016/j.csbj.2014.11.005
14. Chiesa-Estomba CM, Graña M, Medela A, et al. Machine Learning Algorithms as a Computer-Assisted Decision Tool for Oral Cancer Prognosis and Management Decisions: a Systematic Review. *ORL J Otorhinolaryngol Relat Spec*. 2022;84(4):278–288. doi:10.1159/000520672
15. Li Z. Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGBoost. *Comput Environ Urban Syst*. 2022;96:101845. doi:10.1016/j.compenvurbsys.2022.101845
16. Wang G, Sun X, Ren X, et al. Establishment of prognostic model for postoperative patients with metaplastic breast cancer: based on a retrospective large data analysis and Chinese multicenter study. *Front Genet*. 2022;13:993116. doi:10.3389/fgene.2022.993116
17. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. *J Stat Softw*. 2010;36:1–13. doi:10.18637/jss.v036.i11
18. Kuhn M, Coffey CS, Muller KE. Building Predictive Models in R Using the caret Package. *J Stat Softw*. 2008;28(7):1–26. doi:10.18637/jss.v028.i05
19. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf*. 2011;12:77. doi:10.1186/1471-2105-12-77
20. Heagerty PJ, Zheng Y. Survival Model Predictive Accuracy and ROC Curves. *Biometrics*. 2005;61(1):92–105. doi:10.1111/j.0006-341X.2005.030814.x
21. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. Curran Associates, Inc.; 2017. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html. Accessed July 6, 2024.
22. Zhang Z, Gayle AA, Wang J, Zhang H, Cardinal-Fernández P. Comparing baseline characteristics between groups: an introduction to the CBCgrps package. *Ann Transl Med*. 2017;5(24):484. doi:10.21037/atm.2017.09.39
23. Nishimura R, Osako T, Okumura Y, et al. An evaluation of lymphovascular invasion in relation to biology and prognosis according to subtypes in invasive breast cancer. *Oncol Lett*. 2022;24(2):245. doi:10.3892/ol.2022.13366
24. Liu G, Kong X, Dai Q, et al. Clinical Features and Prognoses of Patients With Breast Cancer Who Underwent Surgery. *JAMA Netw Open*. 2023;6(8). doi:10.1001/jamanetworkopen.2023.31078

25. Zhao W, Wu L, Zhao A, et al. A nomogram for predicting survival in patients with de novo metastatic breast cancer: a population-based study. *BMC Cancer*. 2020;20(1):982. doi:10.1186/s12885-020-07449-1
26. Wu SG, Li H, Tang LY, et al. The effect of distant metastases sites on survival in de novo stage-IV breast cancer: a SEER database analysis. *Tumour Biol*. 2017;39(6):1010428317705082. doi:10.1177/1010428317705082
27. Son BH, Kwak BS, Kim JK, et al. Changing patterns in the clinical characteristics of Korean patients with breast cancer during the last 15 years. *Arch Surg*. 2006;141(2):155–160. doi:10.1001/archsurg.141.2.155
28. Ma Z, Huang S, Wu X, et al. Development of a Prognostic App (iCanPredict) to Predict Survival for Chinese Women With Breast Cancer: retrospective Study. *J Med Int Res*. 2022;24(3). doi:10.2196/35768
29. Davies C, Pan H, Godwin J, et al. Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial. *Lancet*. 2013;381(9869):805–816. doi:10.1016/S0140-6736(12)61963-1
30. Sinnadurai S, Kwong A, Hartman M, et al. Breast-conserving surgery versus mastectomy in young women with breast cancer in Asian settings. *BJS Open*. 2019;3(1):48–55. doi:10.1002/bjs5.50111
31. Joyce DP, Manning A, Carter M, Hill ADK, Kell MR, Barry M. Meta-analysis to determine the clinical impact of axillary lymph node dissection in the treatment of invasive breast cancer. *Breast Cancer Res Treat*. 2015;153(2):235–240. doi:10.1007/s10549-015-3549-2
32. Poleshuck EL, Katz J, Andrus CH, et al. Risk Factors for Chronic Pain Following Breast Cancer Surgery: a Prospective Study. *J Pain*. 2006;7(9):626–634. doi:10.1016/j.jpain.2006.02.007
33. Speer JJ, Hillenberg B, Sugrue DP, et al. Study of sexual functioning determinants in breast cancer survivors. *Breast J*. 2005;11(6):440–447. doi:10.1111/j.1075-122X.2005.00131.x
34. Fobair P, Stewart SL, Chang S, D'Onofrio C, Banks PJ, Bloom JR. Body image and sexual problems in young women with breast cancer. *Psychooncology*. 2006;15(7):579–594. doi:10.1002/pon.991
35. Ganz PA, Rowland JH, Desmond K, Meyerowitz BE, Wyatt GE. Life after breast cancer: understanding women's health-related quality of life and sexual functioning. *J Clin Oncol*. 1998;16(2):501–514. doi:10.1200/JCO.1998.16.2.501
36. Ussher JM, Perz J, Gilbert E. Changes to sexual well-being and intimacy after breast cancer. *Cancer Nurs*. 2012;35(6):456–465. doi:10.1097/NCC.0b013e3182395401
37. Veening JG, Barendregt HP. The effects of beta-endorphin: state change modification. *Fluids Barriers CNS*. 2015;12:3. doi:10.1186/2045-8118-12-3
38. Streicher L, Simon JA. Sexual Function Post-Breast Cancer. In: Gradishar WJ, editors. *Optimizing Breast Cancer Management. Cancer Treatment and Research*. Springer International Publishing; 2018:167–189. doi:10.1007/978-3-319-70197-4_11.
39. Metcalfe K, Gershman S, Ghadirian P, et al. Contralateral mastectomy and survival after breast cancer in carriers of BRCA1 and BRCA2 mutations: retrospective analysis. *BMJ*. 2014;348:g226. doi:10.1136/bmj.g226
40. Głowacka-Mrotek I, Tarkowska M, Nowikiewicz T, Hagner-Derengowska M, Goch A. Assessment of Postural Balance in Women Treated for Breast Cancer. *Medicina*. 2020;56(10):505. doi:10.3390/medicina56100505
41. Qiu J, Tang L, Huang L, Hou S, Zhou J. Physical and psychological effects of different temperature-controlled breast prostheses on patients with breast cancer during rehabilitation: a randomized controlled study (CONSORT). *Medicine*. 2020;99(13):e19616. doi:10.1097/MD.0000000000019616
42. Taxbro K, Hammarskjöld F, Thelin B, et al. Clinical impact of peripherally inserted central catheters vs implanted port catheters in patients with cancer: an open-label, randomised, two-centre trial. *Br J Anaesth*. 2019;122(6):734–741. doi:10.1016/j.bja.2019.01.038
43. Wang P, Soh KL, Ying Y, Liu Y, Huang X, Huang J. Risk of VTE associated with PORTs and PICCs in cancer patients: a systematic review and meta-analysis. *Thromb Res*. 2022;213:34–42. doi:10.1016/j.thromres.2022.02.024
44. Steyerberg EW. *Clinical Prediction Models*. Springer, 2009, doi:10.1007/978-0-387-77244-8