Am J Cancer Res 2024;14(4):1609-1621
www.ajcr.us /ISSN:2156-6976/ajcr0155349

Original Article

Development and validation of a random survival forest model for predicting long-term survival of early-stage young breast cancer patients based on the SEER database and an external validation cohort

Lin-Wei Li[1,2,3], Xin Liu[1,2,3], Meng-Lu Shen[1,2,3], Meng-Jun Zhao[1,2,3], Hong Liu[1,2,3]

[1]The Second Surgical Department of Breast Cancer, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin 300060, China; [2]Tianjin's Clinical Research Center for Cancer, Tianjin 300060, China; [3]Key Laboratory of Breast Cancer Prevention and Therapy, Tianjin Medical University, Ministry of Education, Tianjin 300060, China

Received January 9, 2024; Accepted March 10, 2024; Epub April 15, 2024; Published April 30, 2024

Abstract: Young breast cancer (YBC) patients often face a poor prognosis, hence it's necessary to construct a model that can accurately predict their long-term survival in early stage. To realize this goal, we utilized data from the Surveillance, Epidemiology, and End Results (SEER) databases between January 2010 and December 2020, and meanwhile, enrolled an independent external cohort from Tianjin Medical University Cancer Institute and Hospital. The study aimed to develop and validate a prediction model constructed using the Random Survival Forest (RSF) machine learning algorithm. By applying the Least Absolute Shrinkage and Selection Operator (LASSO) regression analysis, we pinpointed key prognostic factors for YBC patients, which were used to create a prediction model capable of forecasting the 3-year, 5-year, 7-year, and 10-year survival rates of YBC patients. The RSF model constructed in the study demonstrated exceptional performance, achieving C-index values of 0.920 in the training set, 0.789 in the internal validation set, and 0.701 in the external validation set, outperforming the Cox regression model. The model's calibration was confirmed by Brier scores at various time points, showcasing its excellent accuracy in prediction. Decision curve analysis (DCA) underscored the model's importance in clinical application, and the Shapley Additive Explanations (SHAP) plots highlighted the importance of key variables. The RSF model also proved valuable in risk stratification, which has effectively categorized patients based on their survival risks. In summary, this study has constructed a well-performed prediction model for the evaluation of prognostic factors influencing the long-term survival of early-stage YBC patients, which is significant in risk stratification when physicians handle YBC patients in clinical settings.

Keywords: Young breast cancer, random survival forest, the Surveillance, Epidemiology, and End Results program (SEER), prediction model

Introduction

Breast cancer (BC) stands as the most prevalent cancer and the foremost cause of cancer-related mortality among women in the globe [1]. Young breast cancer (YBC), the most common malignant tumor affecting young individuals, is characterized as a type of breast cancer diagnosed in individuals under the age of 40 [2, 3].

Age is a critical determinant for the long-term survival of BC patients. Young patients, when compared to patients in an older age group, typically have poorer prognoses [4-6]. An extensive number of research has demonstrated that YBC patients often exhibit more aggressive biological behaviors and less favorable phenotypes. These include a higher incidence of triple-negative breast cancer (TNBC), symptom manifestation in a later stage, larger tumors, increased lymph node involvement, higher histological grades, a positive family history, and a heightened risk of BRCA1/2 mutations [6-10]. Consequently, YBC patients, in contrast to their older counterparts, often have lower 5-year sur-

https://doi.org/10.62347/OJTY4008

vival rates, marked by a significantly elevated risk of overall tumor recurrence and distant metastasis [11-15]. Given the above characteristics, the traditional American Joint Cancer Committee (AJCC) staging system, a system that has wide acceptance, may not be adequate in predicting the survival time of YBC patients. Additionally, taking into consideration the unique challenges faced by YBC patients, such as fertility preservation and the long-lasting side effects from treatment, is essential. Therefore, accurately predicting the long-term survival time of early-stage YBC patients and thus dividing them into different risk subgroups are crucial for physicians to design the best treatment regimen for them.

Currently, several clinical prediction models have been developed for assessing the survival rate of YBC patients. In 2020, Sun et al. developed a nomogram containing 13 predictors through univariate and multivariate Cox analysis, which was utilized to predict the 3-year and 5-year overall survivals (OS) and breast cancer-specific survival (BCSS) of YBC patients [16]. In 2022, Huang et al. constructed a nomogram to predict the 3-year and 5-year OS of YBC patients through the Last Absolute Shrink and Selection Operator (LASSO) regression analysis [17]. However, as YBC patients often have a long survival time, predictions limited to 3 or 5 years may not suffice to meet clinical demands. Additionally, the models constructed in the prior studies have not been externally validated with certain sample sizes, whose performance was also constrained by the uniformity of prediction algorithms.

Recently, the Random Survival Forest (RSF), a novel machine learning algorithm, has emerged for predicting disease progression [18]. Recognized for its high performance and interpretability, the RSF algorithm is currently under development. However, its utilization in forecasting the prognosis of YBC patients remains unexplored.

In this study, we screened the characteristics of the data from the Surveillance, Epidemiology, and End Results (SEER) databases using the LASSO regression analysis and constructed a prediction model for the long-term survival prognosis (including 3-year, 5-year, 7-year and 10-year survivals) of early-stage YBC patients with the use of the RSF algorithm. The model is externally validated in a separate external dataset from the Tianjin Medical University Cancer Institute and Hospital.

Materials and methods

Data sources

In this study, we utilized the SEER*Stat version 8.4.1 software, encompassing datasets from seven centers. Given that the year 2010 was when the data on human epidermal growth factor receptor-2 (HER-2) in the SEER databases started to be collected, we focused on young patients diagnosed with breast cancer between 2010 and 2020.

Inclusion criteria: patients were eligible if (1) they were confirmed with primary BC by pathological examinations; (2) they were under the age of 40; (3) their BCs were primary. Exclusion criteria: patients were excluded if (1) they showed distant metastasis; (2) they didn't have follow-up records; (3) their information about pathological type, histological grade, AJCC stage, lesion size, lymph node status, surgical type, estrogen receptor (ER), progesterone receptor (PR), HER-2 status and treatment regimen for tumors were unknown.

Finally, a total of 6884 eligible patients were included. Following precedent studies, these patients were randomly allocated into a training set (n=4818) and an internal validation set (n=2066) in a 7:3 ratio, with which the prediction model were constructed and verified, respectively. Additionally, an external validation set was formed, comprising 966 YBC patients diagnosed in the Tianjin Medical University Cancer Institute and Hospital from January 2006 to December 2021, to further assess the model's performance. Patients in the validation set were selected based on the same inclusion and exclusion criteria as the training set. The last follow-up for patients was finished in May 2023. Approval for this study was obtained from the Institutional Review Committee of the Tianjin Medical University Cancer Institute and Hospital.

Variables

This study included the following 17 variables: marital status, age, race, lesion sites, AJCC stage, T stage, N stage, histological grade,

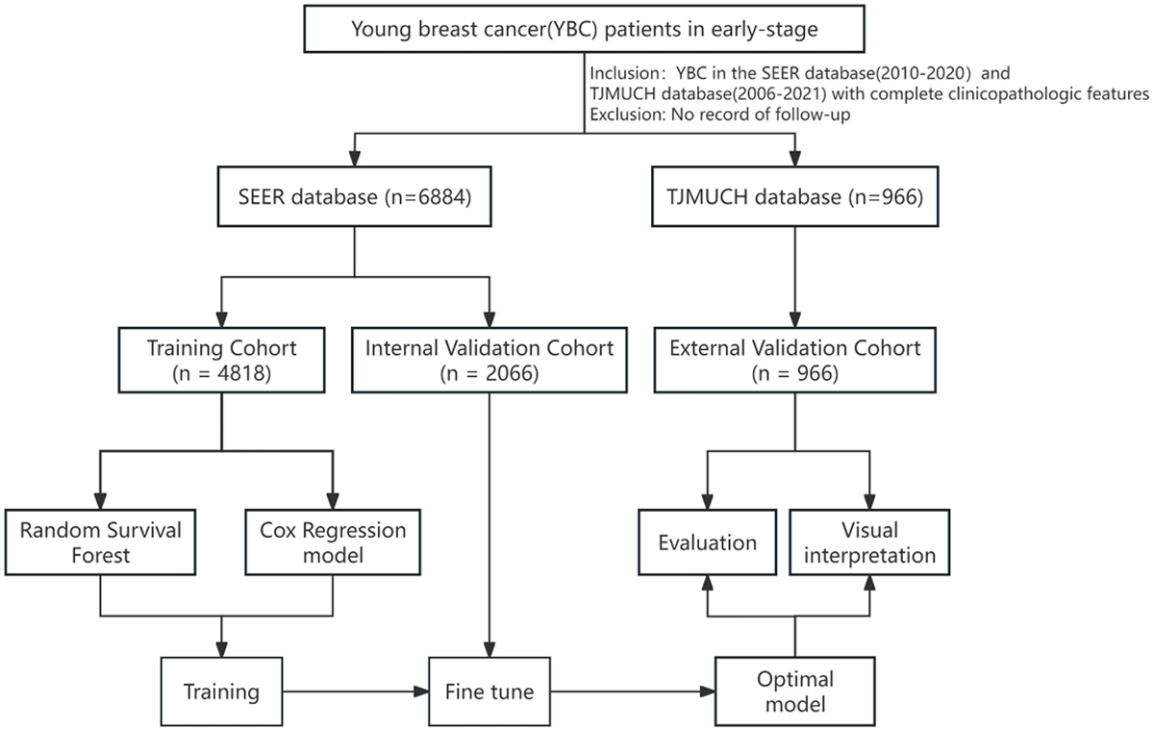

Figure 1. The flowchart of the models' development. SEER: The Surveillance, Epidemiology, and End Results; TJMUCIH: Tianjin Medical University Cancer Institute and Hospital.

pathological type, molecular typing, lesion size, the number of positive lymph nodes, the number of surgically dissected lymph nodes, the proportion of positive lymph nodes, surgical type, chemotherapy, and radiotherapy.

Modeling

The LASSO algorithm, suitable for high-dimensional data regression, was employed to select and rank statistically significant clinical variables from the training dataset [19]. These features extracted by the LASSO algorithm assist in building subsequent RSF and Cox prediction models, providing robust predictive capabilities. In this study, the RSF algorithm was utilized to construct a RSF model. To adjust the parameters of the RSF model, the grid search method was adopted. The hyperparameters and range of grid search are as follows: the number of estimators (10, 100, 500, 1000); minimum samples split (1, 3, 5, 10, 15, 20); minimum samples leaf (1, 3, 4, 10, 15, 20). The Cox regression model was a regular survival prediction model. In the Cox regression model, based on the prognosis-relevant characteristics selected by the LASSO regression, a training-set-based nomogram survival prediction model was established using the "rms" package in the R language to predict the 3-, 5-, 7-, and 10-year OS for non-metastatic YBC patients, which was subsequently employed for the comparison with the RSF model. Both models were built on a development dataset using the 10-fold cross-validation. **Figure 1** illustrates the flowchart of the models' development.

Evaluation and interpretation of the models

The RSF model was evaluated using the internal validation set and the external validation set, respectively, to assess its performance. The primary outcome was the 3-, 5-, 7- and 10-year OS of the early-stage YBC patients.

The model was evaluated using the following indicators: The C-index, a correlation coefficient between anticipated survival risks and actual survival times, was used to assess the model's accuracy. A C-index value of 0.5 denoted for a random prediction. In contrast, a C-index value of 1.0 denoted for an accurate forecasting. Additionally, the model's calibration was evaluated by the Brier scores. The Brier scores - which range from 0 to 1, with 0 being the best outcome - were obtained. They

represented the mean square difference between the observed status of patients and their expected survival time. In practice, a model is deemed helpful if its Brier score is less than 0.25. Receiver operating characteristic (ROC) curves were produced, and area under the curves (AUC) were computed for the 3-, 5-, 7-, and 10-year OS to evaluate the model's time-dependent sensitivities and specificities. The decision curve analysis (DCA) was applied to calculate the clinical net benefit of the model [20].

The interpretation of the prediction model was essential for supporting medical decision-making process, in which physicians could simply understand how the models predict the postoperative prognosis transparently. The Shapley Additive Explanations (SHAP) plot, which was a game-theoretic approach to explain the output of the model, demonstrated the contribution of the variables to the outcome [21].

The RSF risk stratification of patients

The computed risk score from the RSF model: a higher risk score indicating a higher chance of early occurrence of the event of interest (in this case, death). The RSF risk stratification based on the risk score, which was computed by the expected number of events for a particular terminal node in the RSF model, could quantify patients' survival hazards. The critical value was defined according to the risk score and by X-tile software version 3.6.1 (Yale University, New Haven, CT). All early-stage YBC patients were divided into a high-risk group (risk score ≥17.83) and a low-risk group (risk score <17.83).

Statistical analysis

The difference between the demographic and the clinical information was compared using the Wilcoxon test for continuous variables, while the χ^2 test or Fisher's exact test for categorical variables in the training set and the validation set. Two-tailed *p*-values less than 0.05 were considered statistically significant. Python (Version 3.8, Van Rossum, Scotts Valley, CA, USA) was implemented to derive the models. The Cox model and RSF model were based on the scikit-survival module (Version 0.17.2, Sebastian P). The fundamental data analysis was conducted by the R software (Version 4.1.2, RCoreTeam, Vienna, Austria).

Results

The characteristics of patients

A total of 7850 patients were enrolled in the study. **Table 1** summarizes the demographic and clinical information of the enrolled YBC patients in the training set, the internal validation set and the external validation set.

In terms of clinicopathological characteristics, there were significant differences in molecular classification, histological grade, T stage, N stage and AJCC stage between the Tianjin Medical University Cancer Institute and Hospital cohort and the SEER cohort (P<0.001). Compared with the SEER cohort, YBC patients of the Tianjin Medical University Cancer Institute and Hospital cohort had a higher proportion of HR+HER-2-, HR-HER-2+ and TNBC, and a lower proportion of HR+HER-2+. The histological grade of YBC patients in the Tianjin Medical University Cancer Institute and Hospital cohort was mainly grade II, late T stage and late N stage, and AJCC stage was given priority to stage II. The histological grade of YBC patients in the SEER cohort was primarily grade III/IV, and the AJCC stage was mainly stage I. When forming treatment methods, YBC patients from the Tianjin Medical University Cancer Institute and Hospital received total mastectomy and modified radical mastectomy as surgeries, with their proportion of breast conservation and reconstruction lower than that in the SEER cohort. A higher proportion of YBC patients in the Tianjin Medical University Cancer Institute and Hospital cohort received chemotherapy, which may be related to a higher proportion of TNBC and a later AJCC stage. However, there was no significant difference in lymph node metastasis rate between the two cohorts.

Selection of characteristic variables

The LASSO regression analysis was employed in the study. Only one variable, the lesion site, was eliminated, and the remaining 16 variables were included in the construction of the RSF model. These remaining variables included marital status, age, race, AJCC stage, T stage, N stage, histological grade, pathological type, molecular typing, lesion size, the number of positive lymph nodes, the number of surgically dissected lymph nodes, the proportion of positive lymph nodes, surgical methods, chemo-

Table 1. Demographics and clinicopathologic characteristics of the training and validation cohorts

Characteristic	Training Set (n=4818)	Internal validation cohort (n=2066)	External validation cohort (n=966)	*p* Value
Age	37 (33, 39)	37 (33, 39)	37 (33, 39)	0.419
Marital status				<0.001
Divorced	290 (6.0%)	137 (6.6%)	0 (0.0%)	
Married	3079 (63.9%)	1313 (63.6%)	761 (78.8%)	
Separated	3080 (1.0%)	20 (1.0%)	0 (0.0%)	
Single (never married)	3081 (28.7%)	591 (28.6%)	205 (21.2%)	
Widowed	3082 (0.3%)	5 (0.2%)	0 (0.0%)	
Race				<0.001
American Indian/Alaska Native	59 (1.2%)	17 (0.8%)	0 (0.0%)	
Asian or Pacific Islander	815 (16.9%)	339 (16.4%)	966 (100.0%)	
Black	606 (12.6%)	295 (14.3%)	0 (0.0%)	
White	3338 (69.3%)	1415 (68.5%)	0 (0.0%)	
Primary Site				0.586
Nipple	11 (0.2%)	2 (0.1%)	1 (0.1%)	
Central portion of breast	163 (3.4%)	67 (3.2%)	32 (3.3%)	
Upper-inner quadrant of breast	635 (13.2%)	238 (11.5%)	123 (12.7%)	
Lower-inner quadrant of breast	265 (5.5%)	116 (5.6%)	57 (5.9%)	
Upper-outer quadrant of breast	1546 (32.1%)	735 (35.6%)	330 (34.2%)	
Lower-outer quadrant of breast	442 (9.2%)	182 (8.8%)	87 (9.0%)	
Axillary tail of breast	32 (0.7%)	14 (0.7%)	7 (0.7%)	
Overlapping lesion of breast	1112 (23.1%)	441 (21.3%)	217 (22.5%)	
Other	612 (12.7%)	271 (13.1%)	112 (11.6%)	
Histology Type				0.001
Infiltrating duct	4440 (92.2%)	1906 (92.3%)	869 (90.0%)	
Lobular	151 (3.1%)	60 (2.9%)	26 (2.7%)	
Adenocarcinoma	112 (2.3%)	40 (1.9%)	18 (1.9%)	
Ductal	30 (0.6%)	12 (0.6%)	11 (1.1%)	
Medullary	22 (0.5%)	8 (0.4%)	11 (1.1%)	
Metaplastic	17 (0.4%)	12 (0.6%)	7 (0.7%)	
Paget	13 (0.3%)	10 (0.5%)	5 (0.5%)	
Inflammatory	8 (0.2%)	5 (0.2%)	2 (0.2%)	
Other	25 (0.5%)	13 (0.6%)	17 (1.8%)	
Subtype				<0.001
HR+/HER-2+	834 (17.3%)	361 (17.5%)	103 (10.7%)	
HR+/HER-2-	2897 (60.1%)	1230 (59.5%)	578 (59.8%)	
HR-/HER-2+	292 (6.1%)	127 (6.1%)	81 (8.4%)	
HR-/HER-2-	795 (16.5%)	348 (16.8%)	204 (21.1%)	
Histology Grade				<0.001
I	515 (10.7%)	218 (10.6%)	24 (2.5%)	
II	1784 (37.0%)	773 (37.4%)	747 (77.3%)	
III/IV	2519 (52.3%)	1075 (52.0%)	195 (20.2%)	
AJCC stage				<0.001
0	1 (0.0%)	0 (0.0%)	10 (1.0%)	
I	2089 (43.4%)	884 (42.8%)	257 (26.6%)	
II	1944 (40.3%)	854 (41.3%)	527 (54.6%)	
III/IV	784 (16.3%)	328 (15.9%)	172 (17.8%)	

T stage				<0.001
Tis/T0	2 (0.0%)	3 (0.1%)	11 (1.1%)	
T1	2033 (42.2%)	841 (40.7%)	377 (39.0%)	
T2	2124 (44.1%)	957 (46.3%)	497 (51.4%)	
T3	533 (11.1%)	223 (10.8%)	73 (7.6%)	
T4	126 (2.6%)	42 (2.0%)	8 (0.8%)	
N stage				<0.001
N0	2630 (54.6%)	1123 (54.4%)	499 (51.7%)	
N1	1389 (28.8%)	603 (29.2%)	261 (27.0%)	
N2	341 (7.1%)	133 (6.4%)	92 (9.5%)	
N3	182 (3.8%)	89 (4.3%)	65 (6.7%)	
N1mic	276 (5.7%)	118 (5.7%)	49 (5.1%)	
Surgery				<0.001
Biopsy only	155 (3.2%)	58 (2.8%)	0 (0.0%)	
Mastectomy	1275 (26.5%)	573 (27.7%)	647 (67.0%)	
BCS	1701 (35.3%)	720 (34.8%)	218 (22.6%)	
Reconstruction	1687 (35.0%)	715 (34.6%)	101 (10.5%)	
Radiation recode				0.052
No/Unknown	2136 (44.3%)	921 (44.6%)	389 (40.3%)	
Yes	2682 (55.7%)	1145 (55.4%)	577 (59.7%)	
Chemotherapy recode				<0.001
No/Unknown	1151 (23.9%)	486 (23.5%)	149 (15.4%)	
Yes	3667 (76.1%)	1580 (76.5%)	817 (84.6%)	
Positive lymph nodes	0 (0, 1)	0 (0, 1)	0 (0, 2)	0.002
Examined lymph nodes	4 (2, 10)	4 (2, 10)	19 (14, 23)	<0.001
Positive lymph nodes rate	0 (0.00, 0.18)	0 (0.00, 0.18)	0 (0.00, 0.11)	0.525
Tumor Size (mm)	24 (15, 36)	24 (15, 35)	35 (22, 35)	<0.001

HR: Hormone Receptors; HER-2: Human epidermal growth factor receptor-2; AJCC: American Joint Cancer Committee; BCS: Breast-conserving surgery.

therapy and radiotherapy. The procedures for selecting variables are shown in **Figure 2**. Taking advantage of the grid search, the optimal structure of the RSF model comprised 500 estimators, 10 minimum samples split, and 10 minimum samples leaf.

Evaluation and interpretation of the models

In this study, the models were evaluated with both the internal verification set and the external verification set. The performance of the models was shown in **Tables 2**-**4**. The Brier scores of both models were less than 0.25, showing their good calibration. The RSF model outperformed the Cox regression model, with the highest C index and AUC of the 3-, 5-, 7- and 10-year survival in both internal and external validation sets. In addition, the Brier scores of the 3-, 5-, 7-, and 10-year survival from the RSF model were the lowest in both internal and external validation sets. As shown in **Figure 3**, the DCA of the RSF model showed fair clinical net benefits for the 3-, 5-, 7- and 10-year survival prediction.

In addition, we visualized the RSF model in the form of a SHAP plot. In the SHAP plot shown in **Figure 4**, the variables in the model were arranged in a descending order of importance. AJCC stage was the most important variable, followed by T stage, histological grade, molecular typing, and N stage, etc.

Evaluation of risk stratification ability of the RSF model

The risk stratification of patients is of great significance for guiding patient management. We calculated the risk score of each patient through the RSF model and stratified the risk. Patients were divided into a high-risk group

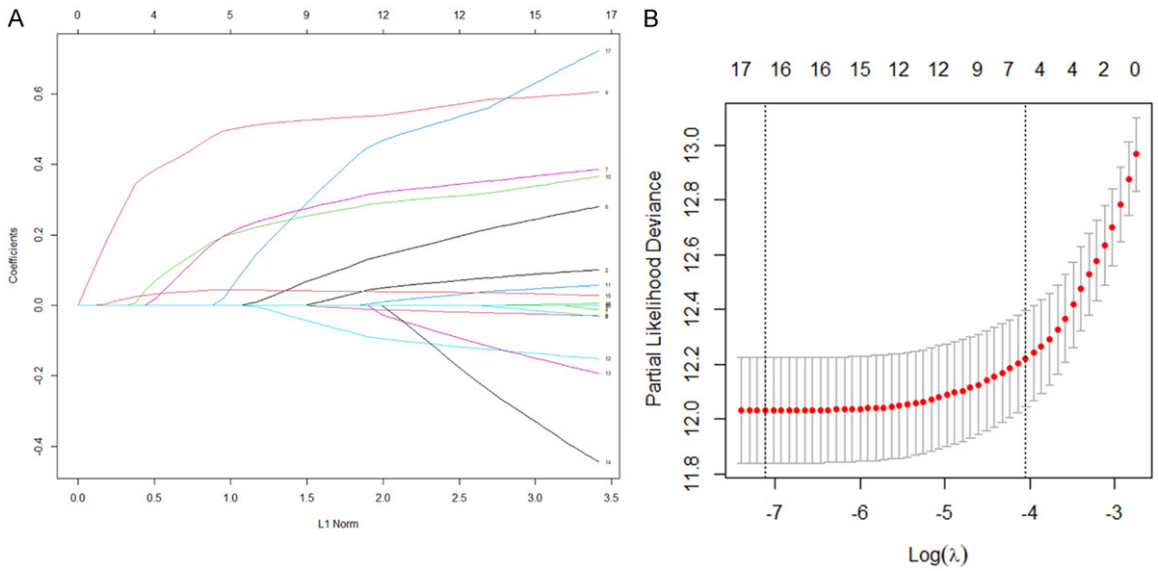

Figure 2. The results of the last absolute shrink and selection operator (LASSO) regression analysis for the random survival forest (RSF) model. A. The LASSO coefficient profiles of the expression of 17 variables. B. Selection of the λ in the LASSO regression analysis via 10-fold cross-validation. The dotted vertical lines are plotted at the optimal values following the minimum criteria (right) and "one standard error" criteria (left).

Table 2. Overall survival (OS) in the training cohort

Model	AUC					Brier score					C-Index (95% CI)	*p* Value
	3-Year	5-Year	7-Year	10-Year	*p* Value	3-Year	5-Year	7-Year	10-Year	*p* Value		
Cox Model	0.827	0.786	0.761	0.718	<0.001	0.015	0.03	0.044	0.061	0.01	0.785 (0.759, 0.811)	<0.001
RSF model	0.947	0.934	0.919	0.890		0.012	0.024	0.034	0.047		0.920 (0.912, 0.928)	

AUC: Area under the Curve; RSF: Random Survival Forest; CI: Confidence Intervals.

Table 3. Overall survival (OS) in the internal validation cohort

Model	AUC					Brier score					C-Index (95% CI)	*p* Value
	3-Year	5-Year	7-Year	10-Year	*p* Value	3-Year	5-Year	7-Year	10-Year	*p* Value		
Cox Model	0.838	0.782	0.741	0.738	0.025	0.016	0.032	0.047	0.067	0.589	0.771 (0.754, 0.788)	0.080
RSF model	0.84	0.802	0.771	0.777		0.016	0.032	0.045	0.063		0.789 (0.771, 0.807)	

AUC: Area under the Curve; RSF: Random Survival Forest; CI: Confidence Intervals.

Table 4. Overall survival (OS) in the external validation cohort

Model	AUC					Brier score					C-Index (95% CI)	*p* Value
	3-Year	5-Year	7-Year	10-Year	*p* Value	3-Year	5-Year	7-Year	10-Year	*p* Value		
Cox Model	0.817	0.763	0.720	0.696	0.028	0.012	0.028	0.044	0.064	0.169	0.697 (0.669, 0.726)	0.845
RSF model	0.821	0.768	0.733	0.696		0.011	0.028	0.043	0.060		0.701 (0.673, 0.730)	

AUC: Area under the Curve; RSF: Random Survival Forest; CI: Confidence Intervals.

(risk score ≥17.83) and a low-risk group (risk score <17.83) with the assistance of X-tile. The results of the Kaplan-Meier analysis and log-rank test between the high-risk group and the low-risk group were presented in **Figure 5**, which demonstrated significant differences between the two groups.

Discussion

The incidence of breast cancer in young women is relatively low [22]. However, the prognosis of YBC patients is generally worse compared with that of older patients [4-6]. Hence it is important to accurately predict the long-term survival

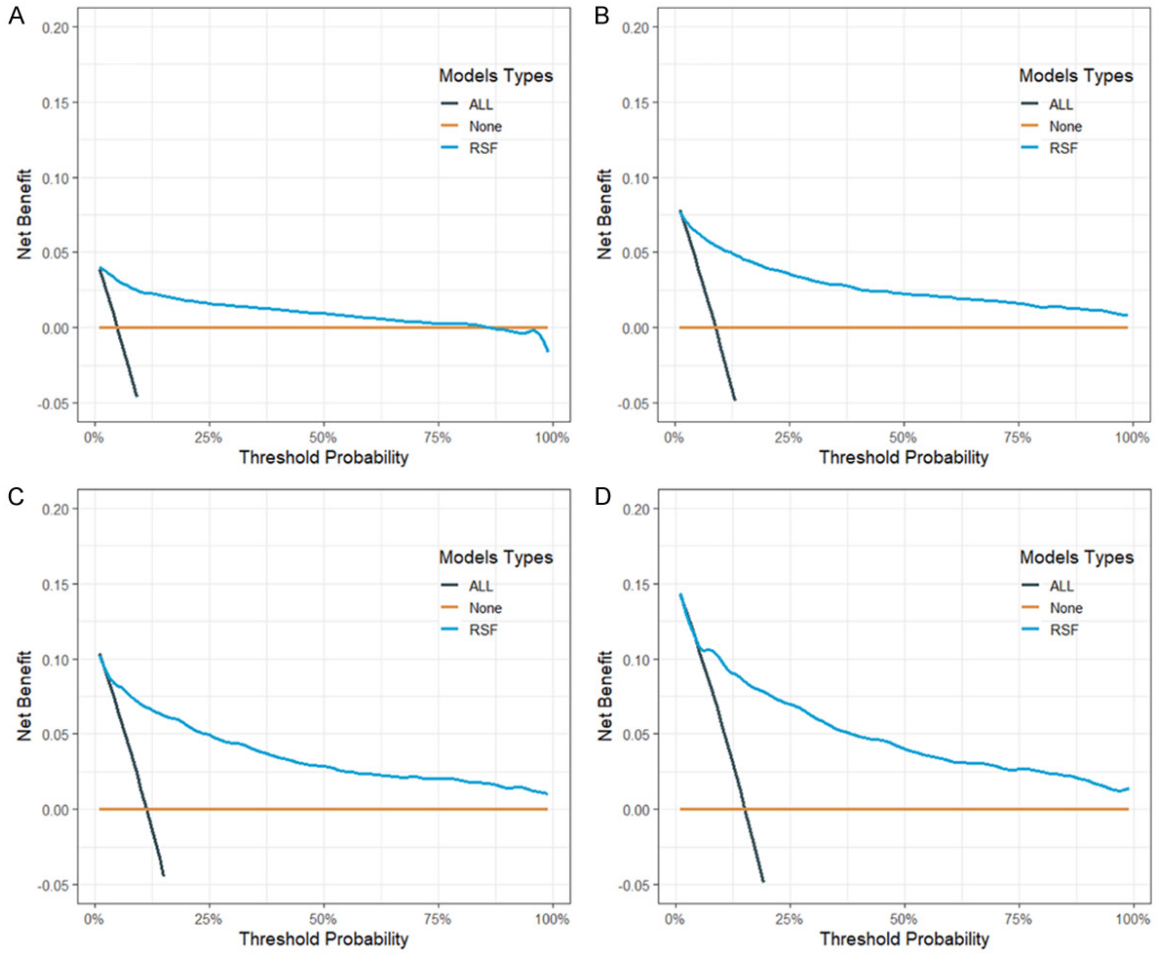

Figure 3. The decision curve analysis (DCA) of the Random Survival Forest (RSF) model. A. The 3-year decision analysis curve of the RSF model. B. The 5-year decision analysis curve of the RSF model. C. The 7-year decision analysis curve of the RSF model. D. The 10-year decision analysis curve of the RSF model. In the decision analysis curve, the x-axis represented the threshold probability, while the y-axis represented the clinical net benefits. The black line in the DCA plot reflects the strategy of "assume all patients have received the assessment of the RSF model", while the horizontal orange line demonstrates the strategy of "assume no patient has received the assessment of the RSF model".

of YBC patients for the benefits of their fertility counseling, long-term follow-up and customized treatment.

This study reported the construction of a prediction model for forecasting the long-term survival of early YBC patients based on the RSF algorithm after variable screening by the LASSO regression. This study showed that the constructed RSF prediction model was capable of better calibration and discrimination in predicting the 3-, 5-, 7-, and 10-year OS of early YBC patients, both in the internal validation cohort based on SEER database and the external validation cohort based on the data from the Tianjin Medical University Cancer Institute and Hospital, in comparison to the Cox regression model. Therefore, the RSF algorithm is rendered with potential to improve the accuracy of individualized survival prediction.

In addition, this study also found that YBC patients in the Tianjin Medical University Cancer Institute and Hospital cohort showed a later AJCC stage, more axillary lymph node metastases, and more conservative surgical options than those in the SEER cohort. In the comparison of lymph node metastasis, the number of lymph node metastases was higher in the Tianjin Medical University Cancer Institute and Hospital cohort than that in the SEER cohort, but there was no significant difference in lymph

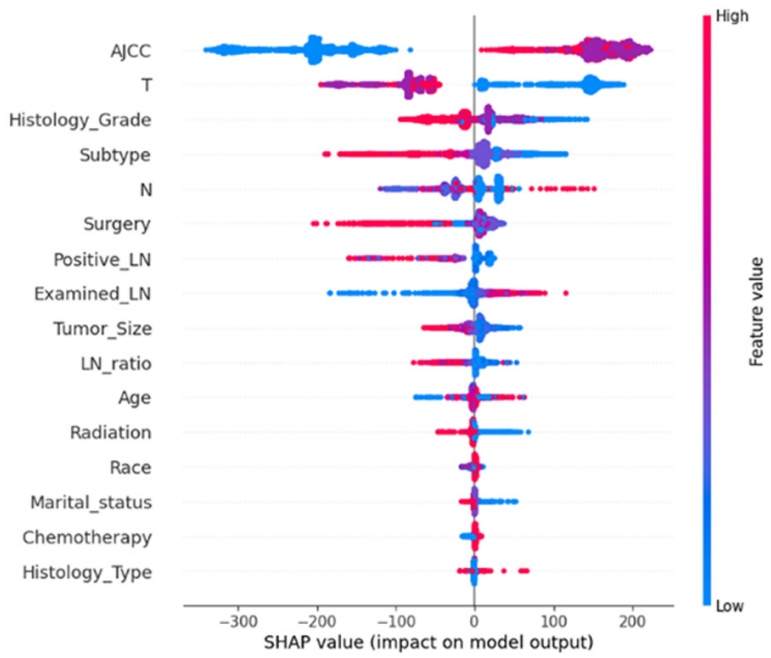

Figure 4. The Shapley Additive Explanations (SHAP) plot of the random survival forest (RSF) model. For each feature, one point corresponds to a single patient. A point's position along the x-axis (i.e., the actual SHAP value) represents the impact that the feature had on the model's output for that specific patient. Mathematically, this corresponds to the (logarithm of the) mortality risk relative across patients (i.e., a patient with a higher SHAP value has a higher mortality risk relative to a patient with a lower SHAP value). Features are arranged along the y-axis based on their importance, which is given by the mean of their absolute Shapley values. The higher the feature is positioned in the plot, the more important it is for the model.

node metastasis rates, which may be attributed from the higher proportion of YBC patients in the Tianjin Medical University Cancer Institute and Hospital cohort who underwent lymph node dissection rather than sentinel lymph nodes biopsy. The distribution of molecular typing in the Tianjin Medical University Cancer Institute and Hospital cohort was also significantly different from that in the SEER cohort. This is consistent with the previous study published by Guo and his colleagues on the comparison of clinicopathological characteristics and treatment methods of YBC patients between China and the West [23].

The LASSO regression analysis was employed for the selection of the necessary characteristics to build the models, as it's considered better for variable selection than the traditional multivariable regression. Moreover, the LASSO regression approach can minimize over-fitting and reduce the complexity of the model by using a loss function or penalty term in addition to the objective function [24].

The RSF algorithm was firstly proposed in 2008 and has become a universal tool for predicting patients' prognosis [25]. Compared with the Cox regression analysis, the RSF algorithm can develop models with better performance, especially when dealing with high-dimensional data [26]. Meanwhile, the application of the Cox regression analysis was limited due to the restriction of the proportional hazard assumption. However, as the structure of the RSF algorithm is non-parametric, it has no such restrictions. Although the models developed by neural networks have always demonstrated impressive performance, the "black box" nature remains an obstacle [27]. The RSF algorithm can achieve a balance between model fitting and interpretation, as shown in this study. To the best of our knowledge, so far, the RSF algorithm has not been applied to predict the prognosis of YBC patients.

In addition, the importance of the characteristics included in the RSF prediction model was visually displayed in the SHAP plot, suggesting that the AJCC stage was identified as the most significant risk variable, followed by T stage, histological grade, molecular typing, N stage, etc. This is consistent with prognostic factors identified by previous studies.

A study on 7665 women aged <40 years reported that the 10-year BCSS of patients diagnosed with stage I or II BC after breast-conserving surgery was 87.7%, the 10-year OS was 85.9%, the 10-year BCSS after mastectomy was 85.4%, and the 10-year OS rate was 83.5% [28]. According to previous reports, early-stage YBC patients had a longer survival time, regardless, the reported models for predicting the 3-year and 5-year survivals could not meet the needs of patients and physicians. Identifying high-risk groups in early-stage YBC patients and making accurate predictions on their long-term survivals are of great clinical significance. The risk score produced by the RSF model con-

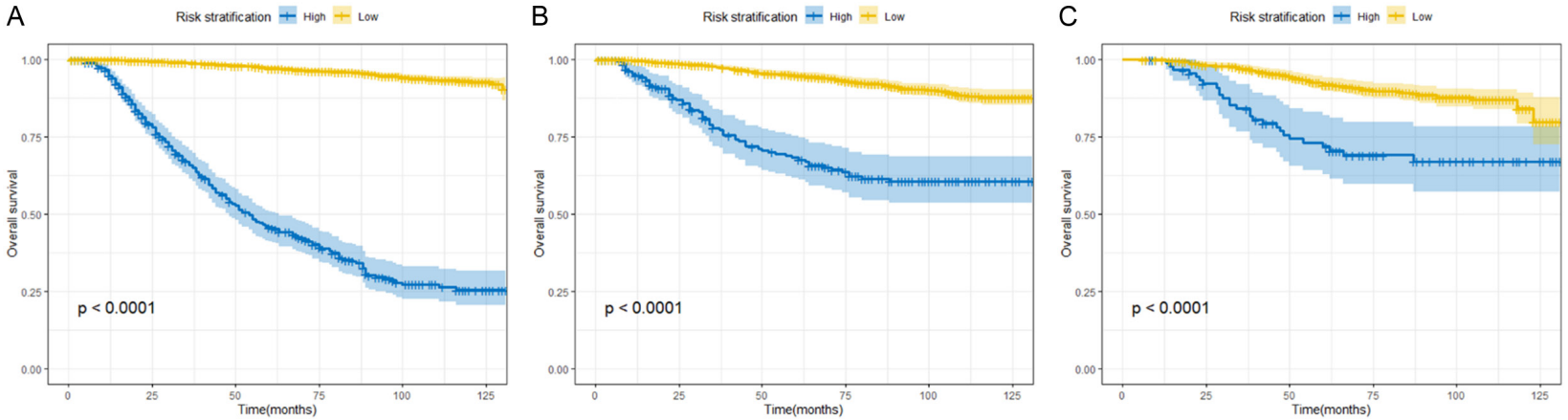

Figure 5. The random survival forest (RSF) risk stratification of patients. A. The RSF risk stratification of patients in the training set. B. The RSF risk stratification of patients in the internal validation cohort. C. The RSF risk stratification of patients in the external validation cohort.

structed in this study can stratify the risks for early-stage YBC patients. By using the RSF risk stratification, doctors can evaluate the survival cycle of patients and correspondingly provide suggestions on fertility and customized treatment regimen. The RSF model is more flexible than the nomogram in predicting clinical prognosis, and is widely used as a prognostic tool in medical field, particularly in the field of oncology [29]. Nomogram, however, is disadvantageous since it can only predict survival at accurate time points and lacks an intuitive representation of the impact of risk factors on the prediction results.

Compared to previous studies that utilized the nomogram construction approaches to build prediction models for YBC patients, our RSF prediction model performed better for some datasets. In the article by Sun et al., a nomogram was constructed using the SEER database to predict the OS of YBC patients. In comparison to our RSF prediction model, the 3-year and 5-year AUC values of the nomogram were relatively lower in the training set (the 3-year OS: 94.7% vs. 85.14%; the 5-year OS: 93.4% vs. 81.92%) [16]. Besides, an external validation set was employed in our study to verify the prediction accuracy of the RSF prediction model, while Sun and his colleagues did not apply this extra validation, thus the generalizability of their model might need to be further investigated. Huang et al. also built a nomogram based on the SEER database for predicting the OS of YBC patients, which was validated with an external validation set of 351 patients. Nevertheless, our RSF prediction model still demonstrated better 3-year and 5-year AUC values in both the training set (the 3-year OS: 94.7% vs. 83.40%; the 5-year OS: 93.4% vs. 77.80%) and the external validation set (the 3-year OS: 84.0% vs. 82.80%; the 5-year OS: 80.20% vs. 77.90%) [17]. Additionally, our study has included the biggest external validation set ever (966 cases), and our prediction model also showed acceptable predictive power over 7-year and 10-year survival predictions. Moreover, this study has the following advantages. Firstly, we have constructed a prediction model based on a machine learning algorithm for the first time for predicting long-term survival time of early-stage YBC patients, which has been compared with the Cox regression model. Secondly, compared to previous papers that have listed predictors, our RSF prediction model has included more predictors, which normally can be obtained from patients' medical records in real time in structured format [16, 17]. Finally, the follow-up time of YBC patients in the external validation cohort included in this study exceeded 10 years, which is longer than any other follow-up time in previous studies [17].

There are still some limitations in this study. First, this is a retrospective study based on the SEER database and the Tianjin Medical University Cancer Institute and Hospital database. Therefore, the selection of data is inevitably biased. Second, some important clinicopathological information, such as the Ki-67 index, endocrine therapy, specific chemotherapy regimen, specific chemotherapy cycle, neoadjuvant therapy, neoadjuvant treatment effect, etc., are not available in the SEER database. Lacking such information may reduce the predictive power of the model over individual prognosis of early YBC patients. Third, young age is associated with a high risk of recurrence [30]. Unfortunately, the SEER database does not provide information about disease recurrence, and thus the recurrence risk of early-stage YBC patients hasn't been evaluated in this study. Fourth, this study lacks relevant information about genetic characteristics. Some YBC patients with genetic mutations (such as PIK3CA, BRCA1/2, ESR1, etc.) may have different prognostic outcomes [31].

Conclusion

By using the LASSO regression to screen characteristics, we developed a high-performance model for predicting the long-term survival prognosis of early-stage YBC patients based on the RSF algorithm. The RSF model has better discrimination and calibration ability than the traditional Cox regression model. In addition, we conducted risk stratification for these populations, which will help doctors identify high-risk patients. Our research holds that deep learning algorithms, especially RSF algorithms, have great potential in future clinical research and practice.

Acknowledgements

This work was supported by Tianjin Key Medical Discipline (Specialty) Construction Project (TJYXZDXK-009A).

Disclosure of conflict of interest

None.

Address correspondence to: Hong Liu, The Second Surgical Department of Breast Cancer, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, The Intersection of Xinjiayuan North Road and Xinjin Road in Binhai New Area, Tianjin 300060, China. Tel: +86-022-23340123; E-mail: liuhong_submit@126.com

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A and 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: 209-249.

[2] Paluch-Shimon S, Cardoso F, Partridge AH, Abulkhair O, Azim HA Jr, Bianchi-Micheli G, Cardoso MJ, Curigliano G, Gelmon KA, Harbeck N, Merschdorf J, Poortmans P, Pruneri G, Senkus E, Spanic T, Stearns V, Wengström Y, Peccatori F and Pagani O. ESO-ESMO 4th international consensus guidelines for breast cancer in young women (BCY4). Ann Oncol 2020; 31: 674-696.

[3] Fidler MM, Gupta S, Soerjomataram I, Ferlay J, Steliarova-Foucher E and Bray F. Cancer incidence and mortality among young adults aged 20-39 years worldwide in 2012: a population-based study. Lancet Oncol 2017; 18: 1579-1589.

[4] Tzikas AK, Nemes S and Linderholm BK. A comparison between young and old patients with triple-negative breast cancer: biology, survival and metastatic patterns. Breast Cancer Res Treat 2020; 182: 643-654.

[5] Kataoka A, Iwamoto T, Tokunaga E, Tomotaki A, Kumamaru H, Miyata H, Niikura N, Kawai M, Anan K, Hayashi N, Masuda S, Tsugawa K, Aogi K, Ishida T, Masuoka H, Iijima K, Kinoshita T, Nakamura S and Tokuda Y. Young adult breast cancer patients have a poor prognosis independent of prognostic clinicopathological factors: a study from the Japanese Breast Cancer Registry. Breast Cancer Res Treat 2016; 160: 163-172.

[6] Kim J, Hong S, Lee JJ, Won YJ, Lee ES, Kang HS, Lee S, Han JH, Lee EG, Jo H, Kim HH and Jung SY. Analysis of the tumor characteristics in young age breast cancer patients using collaborative stage data of the Korea Central Cancer Registry. Breast Cancer Res Treat 2021; 187: 785-792.

[7] Maishman T, Cutress RI, Hernandez A, Gerty S, Copson ER, Durcan L and Eccles DM. Local recurrence and breast oncological surgery in young women with breast cancer: the POSH observational cohort study. Ann Surg 2017; 266: 165-172.

[8] Peto J, Collins N, Barfoot R, Seal S, Warren W, Rahman N, Easton DF, Evans C, Deacon J and Stratton MR. Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. J Natl Cancer Inst 1999; 91: 943-949.

[9] Anglian Breast Cancer Study Group. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group. Br J Cancer 2000; 83: 1301-1308.

[10] Crocetti E, Ravaioli A, Giuliani O, Bucchi L, Vattiato R, Mancini S, Zamagni F, Vitali B, Balducci C, Baldacchini F and Falcini F. Female breast cancer subtypes in the Romagna Unit of the Emilia-Romagna cancer registry, and estimated incident cases by subtypes and age in Italy in 2020. J Cancer Res Clin Oncol 2023; 149: 7299-7304.

[11] Fredholm H, Magnusson K, Lindström LS, Garmo H, Fält SE, Lindman H, Bergh J, Holmberg L, Pontén F, Frisell J and Fredriksson I. Long-term outcome in young women with breast cancer: a population-based study. Breast Cancer Res Treat 2016; 160: 131-143.

[12] Rudat V, El-Sweilmeen H, Fadel E, Brune-Erber I, Ahmad Nour A, Bushnag Z, Masri N and Altuwaijri S. Age of 40 years or younger is an independent risk factor for locoregional failure in early breast cancer: a single-institutional analysis in Saudi Arabia. J Oncol 2012; 2012: 370385.

[13] Bleyer A, Barr R, Hayes-Lattin B, Thomas D, Ellis C and Anderson B; Biology and Clinical Trials Subgroups of the US National Cancer Institute Progress Review Group in Adolescent and Young Adult Oncology. The distinctive biology of cancer in adolescents and young adults. Nat Rev Cancer 2008; 8: 288-298.

[14] Anders CK, Hsu DS, Broadwater G, Acharya CR, Foekens JA, Zhang Y, Wang Y, Marcom PK, Marks JR, Febbo PG, Nevins JR, Potti A and Blackwell KL. Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. J Clin Oncol 2008; 26: 3324-3330.

[15] Aalders KC, Postma EL, Strobbe LJ, van der Heiden-van der Loo M, Sonke GS, Boersma LJ, van Diest PJ, Siesling S and van Dalen T. Contemporary locoregional recurrence rates in young patients with early-stage breast cancer. J Clin Oncol 2016; 34: 2107-2114.

[16] Sun Y, Li Y, Wu J, Tian H, Liu H, Fang Y, Li Y and Yu F. Nomograms for prediction of overall and cancer-specific survival in young breast cancer. Breast Cancer Res Treat 2020; 184: 597-613.
[17] Huang X, Luo Z, Liang W, Xie G, Lang X, Gou J, Liu C, Xu X and Fu D. Survival nomogram for young breast cancer patients based on the SEER database and an external validation cohort. Ann Surg Oncol 2022; 29: 5772-5781.
[18] Rahman SA, Walker RC, Maynard N, Trudgill N, Crosby T, Cromwell DA and Underwood TJ; NOGCA project team AUGIS. The AUGIS survival predictor: prediction of long-term and conditional survival after esophagectomy using random survival forests. Ann Surg 2023; 277: 267-274.
[19] Simon N, Friedman J, Hastie T and Tibshirani R. Regularization paths for cox's proportional hazards model via coordinate descent. J Stat Softw 2011; 39: 1-13.
[20] Vickers AJ, Van Calster B and Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016; 352: i6.
[21] Lundberg S and Lee SI. A unified approach to interpreting model predictions. Arxiv 2017.
[22] Johnson RH, Anders CK, Litton JK, Ruddy KJ and Bleyer A. Breast cancer in adolescents and young adults. Pediatr Blood Cancer 2018; 65: e27397.
[23] Guo R, Si J, Xue J, Su Y, Mo M, Yang B, Zhang Q, Chi W, Chi Y and Wu J. Changing patterns and survival improvements of young breast cancer in China and SEER database, 1999-2017. Chin J Cancer Res 2019; 31: 653-662.
[24] Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc Series B Stat Methodol 2011; 73: 273-282.
[25] Ishwaran H, Kogalur UB, Blackstone EH and Lauer MS. Random survival forests. Arxiv 2008.
[26] Taylor JMG. Random survival forests. J Thorac Oncol 2011; 6: 1974-1975.
[27] Buhrmester V, Münch D and Arens M. Analysis of explainers of black box deep neural networks for computer vision: a survey. Arxiv 2019.
[28] Ye JC, Yan W, Christos PJ, Nori D and Ravi A. Equivalent survival with mastectomy or breast-conserving surgery plus radiation in young women aged <40 years with early-stage breast cancer: a national registry-based stage-by-stage comparison. Clin Breast Cancer 2015; 15: 390-397.
[29] Balachandran VP, Gonen M, Smith JJ and DeMatteo RP. Nomograms in oncology: more than meets the eye. Lancet Oncol 2015; 16: e173-180.
[30] Rosenberg SM and Partridge AH. Management of breast cancer in very young women. Breast 2015; 24 Suppl 2: S154-158.
[31] Xiao W, Zhang G, Chen B, Chen X, Wen L, Lai J, Li X, Li M, Liu H, Liu J, Han-Zhang H, Lizaso A and Liao N. Characterization of frequently mutated cancer genes and tumor mutation burden in Chinese breast cancer. Front Oncol 2021; 11: 618767.