Original Research Article

Development of a nomogram for predicting recurrence in breast cancer patients using a machine learning method

Abhishek O. Tibrewal*

Department of Epidemiologist and Biostatistician, Independent Consultant, New Delhi, India

*Correspondence:
Dr. Abhishek Tibrewal,
E-mail: dr.abhishek08@hotmail.com

ABSTRACT

Background: Current breast cancer (BC) recurrence models do not account for treatment modalities, one of the strongest prognostic factors. This analysis was conducted to apply machine learning (ML) algorithm to identify BC patients at a higher recurrence risk.
Methods: It is based on a downloadable BC Wisconsin dataset, containing 9 independent (socio-demographic, tumor and treatment-related) and a dependent (recurrence) variable(s). Using training dataset (70% sample), a multivariate LR model was developed using univariate analysis identified variables (p<0.2). The model performance was assessed on test dataset (remaining 30%) using standard statistical measures. A nomogram was developed using model identified variables (p<0.05), and its cut-off score categorized BC patients into a high/low recurrence risk.
Results: 277 patients (recurrence (n=81)) were included. In univariate analysis, tumor size (p=0.002), invasive nodes number (p<0.001), node capsule (p<0.001), degree of malignancy (p<0.001) and irradiation (p<0.001) were associated with recurrence. After balancing, both groups included 243 patients. Using training dataset (n=342), invasive nodes (p<0.05), degree of malignancy (p<0.05) and irradiation (p=0.0009) were significant in a multivariate model. The model's accuracy and area under curve (AUC) were 74% (66-81%) and 0.74 (0.67-0.81), respectively in the test dataset (n=144). The nomogram's cut-off score of 55 has an AUC of 0.73 (0.66-0.80) for recurrence prediction, indicative fair discriminating ability.
Conclusions: The developed nomogram can be a valuable tool in guiding appropriate treatment based on recurrence risk. ML and data mining methods can be the future of clinical decision process.

Keywords: Breast cancer, Machine learning, Nomogram, Recurrence

INTRODUCTION

Breast cancer (BC) is the most commonly diagnosed cancer and second leading cause of cancer death among women. In India, the age-adjusted incidence rate of BC is 25.8/100,000 women and mortality is 12.7/100,000 women.[1] The treatment modalities include surgery (breast conserving/mastectomy), radiotherapy, chemotherapy, and hormonal therapy. However, despite these treatments, BC can recur, which can be local (same place as the original cancer), regional (chest wall or lymph nodes under the arm or in the chest), or distant metastasis (another place, including distant organs such as the bones, lungs, liver, and brain). For instance, in an Indian study among early BC patients with 1-3 positive axillary lymph nodes treated with modified radical mastectomy and adjuvant systemic therapy, a total of 38 (15.7%) patients had recurrent disease [only loco-regional recurrence in 10 patients (4.1%), only distant recurrence in 22 patients (9.1%) and simultaneous loco-regional and distant

recurrence in 6 patients (2.5%) at a median follow-up of 5 years.[2]

It has been a major challenge for oncologists to determine which BC patients will have a recurrence. For their assistance, a couple of risk prediction models have been developed. For example, Nottingham prognostic index (NPI) is an old computer-based prognostic model (developed in 1980-90s) based on three variables (tumor size, tumor grade, and lymph node status). According to NPI score, BC patients are divided into low or high risk of recurrence.[3,4] Adjuvant online is another computer-based prognostic model (developed in 2001) to estimate 10-year survival and recurrence for BC patient based on six variables (age at diagnosis, comorbidity, estrogen receptor (ER), tumor size, tumor grade and lymph node status).[5,6] Both these models, do not take into account the treatment modalities which were observed to be one of the strongest prognostic factors. Therefore, the objective of this analysis was to apply machine learning algorithm and build a nomogram for identifying BC patients at a higher risk of recurrence based on some additional variables.

METHODS

This cross-sectional ML based analysis was conducted from October 2019 to February 2020. It is based on the BC dataset downloaded from https:// www. openml.org/d/13.[7] This dataset was provided by M. Zwitter and M. Soklic from the University Medical Centre, Institute of Oncology, Yugoslavia. Since the data was downloaded from the public domain, ethical clearance consent was not required for conducting this analysis. It included 9 independent variables and 1 dependent variable. The independent variables were: age, patient's age (in years) at the time of diagnosis, reported as 20-29, 30-39, 40-49, 50-59, 60-69, and 70-79, menopause (menopausal status of the patient at the time of diagnosis, reported as premenopause, lt40 and ge40 (further details were not provided for lt40 and ge40), tumor size (the size of the tumor (in mm), reported as 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,45-49, 50- 54, and 55-59), invasive nodes (the number of lymph nodes showing BC at the time of histological examination, reported as 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, and 36-39), node-caps (the penetration (yes or no) of the tumor in the lymph node capsule), degree of malignancy.

The histological grade of the tumor, where grade 1: looks most like normal breast cells and is usually slow-growing; grade 2: looks less like normal cells and is growing faster and grade 3: looks different to normal breast cells and is usually fast-growing), breast (the breast (left or right) affected with BC), breast quadrant (the specific location of the breast affected with BC, reported as left-upper, left-lower, right-upper, right-lower and central); irradiation (the radiation therapy history of the patient (yes or no)). The dependent variable was class (the recurrence status (yes or no) of the patient).

The statistical analysis of the extracted data was performed using R Project for Statistical Computing https://www.r-project.org/8. For univariate analysis, the effect of each variable was tested for statistical significance using the chi-square test. All comparisons were two-tailed and p<0.05 were considered to be statistically significant. A multivariable analysis (predictive model development) was carried out using a logistic regression model where all the variables with p<0.2 in the univariate analysis were analyzed using an enter method. Before applying this multivariate model, the number of patients in the recurrence and non-recurrence group was balanced using Synthetic Minority Over-sampling Technique (SMOTE), one of the most popular algorithm for balancing the dataset9. The balanced dataset was split into training and test data in 70:30 ratios. The training dataset included both independent variables and a dependent variable (class) and was used to train the logistic regression model. While, the test dataset was used to assess how well the model was trained.

The performance of the model was assessed using standard statistical measures such as accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and AUC - Receiver Operating Characteristics (ROC) curve, where an AUC of 1.0 indicates perfect predictive ability, whereas 0.5 represents no predictive discrimination. The variables found to be statistically significant (p<0.05) in the multivariate logistic regression model were then used for a nomogram development, which provided the probability of BC recurrence. The total scores obtained from the nomogram were used to identify a cut-off score to categorize the BC patients into a high or low risk of recurrence. The performance of this cut-off score was assessed using accuracy, sensitivity, specificity, and AUC-ROC. Values ranging from 0.7 to 0.8 represent reasonable discrimination, and values exceeding 0.8 represent good discrimination.

RESULTS

The dataset included 286 patients with 9 variables for classifying whether the patient had a recurrence or not. Some of the variables had multiple groups, so those were clubbed into fewer groups for better comparison. That is, the age were clubbed into two groups (20-49 and 50-79 years), the tumor size into four groups (0-9, 10-19, 20-29 and >30), and the number of invasive nodes into four groups (0-2, 3-5, 6-8 and >8). For 8 patients, the data for node capsule was missing, while for another 1 patient, the data for breast quadrant was missing, and therefore these 9 patients were excluded from the analysis. Finally, a total of 277 patients (recurrence (n=81) and no recurrence (n=196)) were included for the analysis.

Univariable analysis

Table 1 showed the univariate analysis of different variables among 277 BC patients. The age (p=0.2169), menopausal status (p=0.2178), breast (p=0.5788) and breast quadrant (p=0.5073) were not significantly different among patients with and without recurrence. However, the tumor size (p=0.002), number of invasive nodes (p<0.001), node capsule (p<0.001), degree of malignancy (p<0.001) and irradiation (p<0.001) were significant factors influencing the recurrence of BC.

Multivariable analysis

The number of patients in the recurrence group (n=81) was almost half that in the non-recurrence (n=196) group, so, balancing was conducted using SMOTE technique. After balancing, there were 243 patients in both the groups with a total of 486 patients. The balanced dataset was split into training (n=342) and test (n=144) dataset in the 70:30 ratio. The model was developed using the

training dataset with the variables having p<0.2 in the univariate analysis (i.e. invasive nodes, tumor size, node capsules, irradiation and degree of malignancy) as shown in Table 2.

The tumor size (p>0.05) and node capsule (p=0.6744) were not statistically significant, while, the invasive nodes (p<0.05), degree of malignancy (p<0.05) and irradiation (p=0.0009) were significant factors influencing the recurrence of BC in the multivariate logistic regression model.

Model performance assessment (internal validation)

The performance of the model was assessed using the test dataset (n=144). The sensitivity, specificity, PPV and NPV of the model were 61.1%, 87.5%, 83.0% and 69.2% respectively. The accuracy of the model was 74.3% (95% CI: 66.4% - 81.2%) and the AUC of the model was 0.74 (95% CI: 0.67-0.81) (Figure 1).

Table 1: Univariate analysis of different variables among patients with and without recurrence.

| Variables		Recurrence		P value
		No (n=196)	Yes (n=81)	
Age (years)	20-49	84 (42.9%)	42 (51.9%)	0.2169
	50-79	112 (57.1%)	39 (48.1%)	
Menopausal status	premenopause	101 (51.5%)	48 (59.3%)	0.2178
	ge40	90 (45.9%)	33 (40.7%)	
	lt40	5 (2.6%)	0 (0%)	
Breast	Left	100 (51%)	45 (55.6%)	0.5788
	Right	96 (49%)	36 (44.4%)	
Breast quadrant	Central	17 (8.7%)	4 (4.9%)	0.5073
	Left lower	73 (37.2%)	33 (40.7%)	
	Left upper	69 (35.2%)	25 (30.9%)	
	Right lower	17 (8.7%)	6 (7.4%)	
	Right upper	20 (10.2%)	13 (16%)	
Tumor size (mm)	0-9	11 (5.6%)	1 (1.2%)	0.002184*
	10-19	50 (25.5%)	7 (8.6%)	
	20-29	67 (34.2%)	32 (39.5%)	
	≥30	68 (34.7%)	41 (50.6%)	
Invasive nodes	0-2	166 (84.7%)	43 (53.1%)	0.0000004766*
	03-05	17 (8.7%)	17 (21%)	
	06-08	7 (3.6%)	10 (12.3%)	
	>8	6 (3.1%)	11 (13.6%)	
Node capsule	No	171 (87.2%)	50 (61.7%)	0.000003392*
	Yes	25 (12.8%)	31 (38.3%)	
Degree of malignancy	1	57 (29.1%)	9 (11.1%)	0.00000002593*
	2	101 (51.5%)	28 (34.6%)	
	3	38 (19.4%)	44 (54.3%)	
Irradiation	No	164 (83.7%)	51 (63%)	0.0003142*
	Yes	32 (16.3%)	30 (37%)	

mm: millimetre, further details were not provided for lt40 and ge40, *statistically significant (p<0.05)

Table 2: Multivariate analysis of different variables among patients with and without recurrence (training dataset, n=342).

Variables	β coefficient	Odds ratio (95% CI)	P value
Intercept	-1.5369	0.22 (0.04 - 1.05)	0.0572
Tumor size (0-9)	1		
Tumor size (10-19)	-0.2165	0.81 (0.15 - 4.26)	0.7989
Tumor size (20-29)	0.507	1.66 (0.33 - 8.23)	0.5348
Tumor size (≥30)	0.0743	1.08 (0.22 - 5.32)	0.9274
Invasive nodes (0-2)	1		
Invasive nodes (3-5)	1.3807	3.98 (1.77 - 8.94)	0.0008*
Invasive nodes (6-8)	2.1877	8.91 (3.26 - 24.4)	<0.0001*
Invasive nodes (>8)	1.374	3.95 (1.46 - 10.71)	0.0069*
Node capsule (Yes)	0.1432	1.15 (0.59 - 2.25)	0.6744
Degree malignant (1)	1		
Degree malignant (2)	0.0841	1.09 (0.53 - 2.23)	0.8186
Degree malignant (3)	1.3376	3.81 (1.79 - 8.11)	0.0005*
Irradiation (Yes)	1.0696	2.91 (1.55 - 5.48)	0.0009*

CI: Confidence Interval, β: Beta, *statistically significant (p<0.05)

Table 3: Confusion matrix showing actual and predicted data (applying the model on test dataset).

| Predicted	Actual		Total
	Recurrence (n=72)	No recurrence (n=72)	
Recurrence	44	9	53
No recurrence	28	63	91

p=0.000000002222
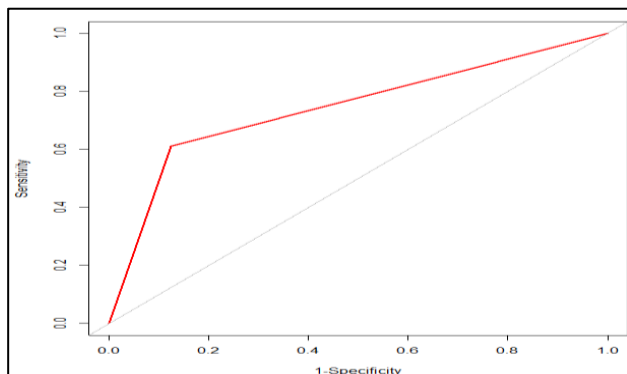
Figure 1: ROC curve for the model (applying on test dataset).

AUC of the model was 0.74 (95% CI: 0.67-0.81), ROC: receiver operating characteristics, Nomogram development.

The invasive nodes, degree of malignancy and irradiation were identified to be the significant predictors of BC recurrence in the multivariate logistic regression model. These three variables were considered for nomogram development. Invasive nodes were scored from 0 to 100 starting where 0-2 (0 points), 3-5 (75 points), 6-8 (88 points) and >8 (100 points). Similarly, the range of points for degree of malignancy was from 0 to 67 and for the irradiation it was 0 to 50.Figure 2 depicts the nomogram showing the predicted risk of BC recurrence.

Table 4: Confusion matrix showing actual and predicted data (based on the nomogram cut-off score; n=277).

Predicted	Actual		Total
	Recurrence (n=81)	Non-recurrence (n=196)	
Recurrence	57	62	119
Non-Recurrence	24	134	158

p<0.0001
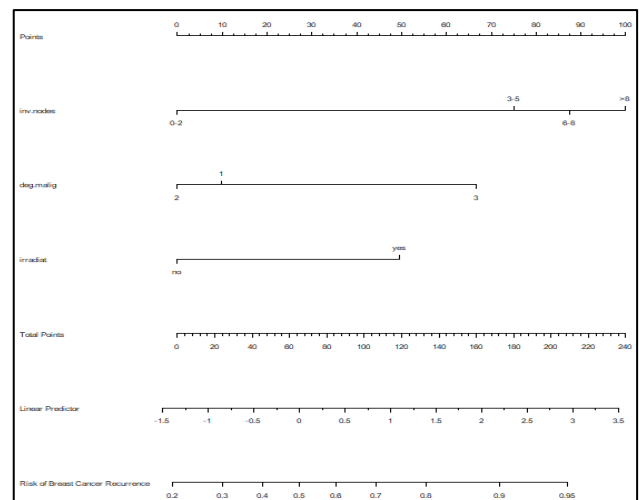
Figure 2: Nomogram showing predicted risk of breast cancer recurrence.

inv.nodes: Invasive nodes (0-2,3-5, 6-8 and >8), deg.malig: degree of malignancy (1,2,3), irradiat: irradiation (no or yes), As an example if patient with invasive nodes 3-5 (75 points), degree of malignancy 3 (66 points), and received irradiation (48 points) will have a total point of 189 and predicted risk of BC recurrence between 90% and 95%.

A cut-off score of 55 has a sensitivity of 70%, specificity of 68% and accuracy of 69% (95% CI: 63% - 74%). The AUC was 0.73 (95% CI 0.66-0.80) for the prediction of BC recurrence indicative of fair discriminating ability (Figure 3). The predicted probability for BC recurrence decreased from 95% to 30% as the total point score decreased from 209 to 24. Thus, based on prediction nomogram, patients could be discriminated into 2 groups with a significant difference (p <0.0001, chi-square test).
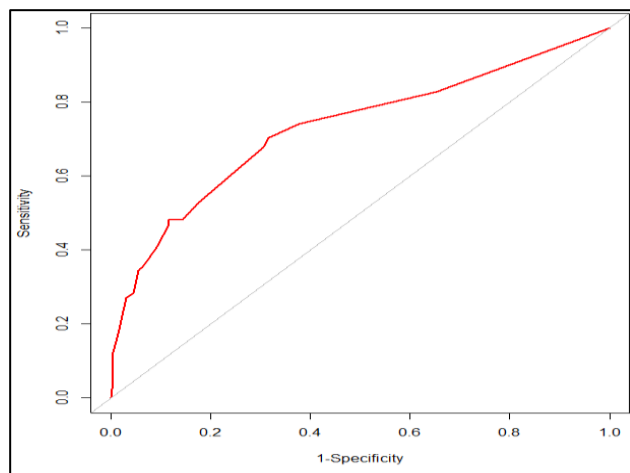
Figure 3: ROC for the nomogram cut-off score.
AUC was 0.73 (95% CI 0.66-0.80), ROC: receiver operating characteristics.

DISCUSSION

The main goal of this study was to solve a data science problem related with BC. Internet-based computational algorithms can serve as valuable aids for oncologists in the process of informed decision making. Identifying patients at a higher risk of BC recurrence has important implications not only for enabling the ability to provide accurate information to patients but also the potential to improve patient outcomes. Appropriate risk assessment of BC patients post primary treatment is critically important, not only to avoid BC recurrence, but also to optimize patient's health and the use of medical resources.

Our analysis showed that higher the number of invasive nodes involved (more the spread of BC) greater the risk of BC recurrence. Consistent with this finding, the adjusted 5-year risk of BC recurrence among pN2 (involvement of 4-9 axillary lymph nodes) and pN3 patients (≥10 axillary lymph nodes) was 2.47 (95% Confidence interval (CI) 1.72-3.56) and 2.42 (1.62-3.60) times higher, respectively, compared with pN1 patients (involvement of 1-3 axillary lymph nodes) (p<0.001).[10] Likewise, among early (stage I and II) BC patients, the adjusted 5 year loco-regional recurrence risk was observed to be higher among those having 1-3 nodes involved (OR: 1.64 (1.32-2.04); p<0.001), and >3 nodes involved (OR: 2.90 (2.14-3.94), p<0.001) as compared to those with no nodes involved.[11-16] Similarly, in the multivariate Cox regression analysis, node metastasis

(HR: 2.28 95% CI 1.5–3.45, p≤0.001) was observed to be an important factor for disease free survival in early BC patients.[11]

Also, the current analysis reported that the greater the degree of malignancy (faster the growth of BC) higher the risk of BC recurrence. In support with our finding, among early BC patients, the adjusted 5 year loco-regional recurrence risk was observed to be higher among those with grade 2 (OR: 1.92 (1.45–2.540, p<0.001) and grade 3 (OR: 2.96 (2.16-4.05), p<0.001) malignancy as compared to those with grade 1 malignancy16. Similarly, early BC patients with grade 2 (Hazard Ratio (HR): 1•44 (95% CI: 1•02-2•03), p=0.04) and grade 3 (HR: 2•15 (1•51-3•06), p<0.0001) malignancy had a significantly higher risk of recurrence as compared to those with grade 1 malignancy.[12]

Moreover, in our study, BC patients who had received irradiation had a greater risk of BC recurrence. Based on the current indications (axillary nodal involvement of ≥4 nodes, disease ≥5 cm in size, and positive surgical margins) for post-surgery radiotherapy, one can understand that those receiving irradiation are high-risk candidates.[13] For instance, a meta-analysis of the prospective clinical trials performed by the Early BC Trialists' Collaborative Group (EBCTCG) among >2,200 women has reported that breast radiation reduces the chance of a tumor recurrence within the breast after 20 years from 30.3% to 10.6%.[14] Also, for patients who had received locoregional radiotherapy for early BC, 10-year incidence of loco-regional recurrence is in the range of 3% - 5% versus ~35% for those who did not receive it.[15]

Tumor size, although an important predictor variable was not found to be significant in multivariate analysis in our study.[2,12] This finding suggested that risk factors for recurrence vary considerably among different study populations, and therefore, it is desirable to have careful selection criteria based on institutional data.

Nomograms serve as a useful, statistically based tool for decision making for clinicians as well as patients. These can be constructed on clinical data sets and used for decision making in an individual patient with respect to specific outcomes (e.g., recurrence of BC in our case). In the clinic, it is important to individualize treatment for which estimation of the overall risk of recurrence is advisable. The knowledge of conventional risk factors cannot accurately predict individual patient risk due to the significant interactions among them. The risk estimated from the nomogram predicted for individual patient has more practical implications as it accounts for the numerical and categorical data as well as its interaction. Moreover, the current nomogram was thoroughly validated internally to achieve consistent results.

Our study had the following limitations: the exact time (post-disease free period) and type (whether local, regional or distant) of recurrence was not reported in the

dataset, some of the important prognostic factors were not taken into consideration, due to lack of data.

These included other treatment details (such as surgery, chemotherapy, hormonal therapy), disease-related variables such as receptor status, extensive intraductal component, lymphovascular invasion, circulatory tumor cells, disseminated tumor cells, tumor-related variables such as lymph node ratio, cancer in positive margins.[2] Nevertheless, the nomogram would be helpful in the prediction of individual risk and aid decision making after independent external validation in separate cohort with similar characteristics.

CONCLUSION

BC is a heterogeneous disease with great diversity in morphology and clinical behavior. The recurrence of BC after complete treatment is common; therefore, the prediction of BC recurrence is a crucial factor for successful treatment and follow-up planning. The number of invasive nodes, degree of malignancy, and irradiation were identified to be significantly associated with BC recurrence in the BC Wisconsin dataset. The nomogram developed based on this dataset can be a valuable tool in guiding appropriate treatment modalities based on the risk of recurrence. Machine learning and data mining methods can be the future of the clinical decision process.

REFERENCES

1. Malvia S, Bagadi SA, Dubey US, Saxena S. Epidemiology of breast cancer in Indian women. Asia Pac J Clin Oncol. 2017;13:289-95.
2. Wadasadawala T, Kannan S, Gudi S, Rishi A, Budrukkar A, Parmar V, et al. Predicting loco-regional recurrence risk in T1, T2 breast cancer with 1–3 positive axillary nodes postmastectomy: Development of a predictive nomogram. Indian J Cancer. 2017;54:352-7.
3. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham prognostic index in primary breast cancer. Breast Cancer Res Treat. 1992;22:207-19.
4. Hearne BJ, Teare MD, Butt M. Comparison of nottingham prognostic index and adjuvant online prognostic tools in young women with breast cancer: review of a single-institution experience. BMJ Open. 2015;5:e005576.
5. Hajage D, Rycke Y, Bollet M, Savignoni A, Caly M, Pierga JY. External validation of adjuvant online breast cancer prognosis tool. prioritising recommendations for improvement. PLoS One. 2011;6(11):e27446.
6. Wazir U, Mokbel K, Carmichael A, Mokbel K. Are online prediction tools a valid alternative to genomic profiling in the context of systemic treatment of ER-positive breast cancer? Cell Mol Biol Lett. 2017;22:20.
7. Vanschoren, J. Exploring machine learning better, together. Available at https://www.openml.org/d/13. Accessed on 17 April 2019.
8. The R project for statistical computing. Available at https://www.r-project.org/. Accessed on 14 August 2019.
9. Brownlee J. 8 tactics to combat imbalanced classes in your machine learning dataset, machine learning mastery. Available at https:// machinelearning mastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/. Accessed on 7 October 2019.
10. Tonellotto F, Bergmann A, Abrahão K, Aguiar S, Bello MA, Thuler L. Impact of number of positive lymph nodes and lymph node ratio on survival of women with node-positive breast cancer. Eur J Breast Health. 2019;15(2):76-84.
11. Nair N, Shet T, Parmar V, Havaldar R, Gupta S, Budrukkar A, et al. Breast cancer in a tertiary cancer center in India - an audit, with outcome analysis. Indian J Cancer. 2018;55:16-22.
12. Harris SR, Dahlstrom JE, Gupta R, Zhang Y, Craft P, Shadbolt B. Recurrence in early breast cancer: analysis of data from 3,765 Australian women treated between 1997 and 2015. The Breast. 2019;44:153-9.
13. Varela M, Chin YS, Makris A. Current indications for post-mastectomy radiation. Int Semin Surg Oncol. 2009;6:5.
14. Early Breast Cancer Trialists' Collaborative Group. Favourable and unfavourable effects on long-term survival of radiotherapy for early breast cancer an overview of the randomised trials. Lancet. 2000;355(9217):1757-70.
15. Wadasadawala T, Vadgaonkar R, Bajpai J. Management of isolated locoregional recurrences in breast cancer: a review of local and systemic modalities. Clin Breast Cancer. 2017;17(7):493-502.
16. Witteveen A, Vliegen IMH, Sonke GS. Personalisation of breast cancer follow-up: a time-dependent prognostic nomogram for the estimation of annual risk of locoregional recurrence in early breast cancer patients. Breast Cancer Res Treat. 2015;152:627-36.