Original Article
Middle East Journal of Cancer; April 2019; 10(2): 132-138

A Model to Predict Breast Cancer Survivability Using Logistic Regression

Mehdi Nourelahi*, Ali Zamani*♦, Abdolrasoul Talei, Sedigheh Tahmasebi****

**Department of Biomedical Physics and Biomedical Engineering,Shiraz University of Medical Sciences, Shiraz, Iran*
***Breast Disease Research Center, Shiraz University of Medical Sciences, Shiraz, Iran*

Abstract

Background: Breast cancer is the most common type of cancer amongst women worldwide. Considering its high incidence, effective detection and prognosis of this type of cancer may have a significant effect on reducing expenditures. In this study, we propose a model to predict the 60-month survivability in patients with breast cancer and investigate the effects of each feature on the obtained model.

Methods: We base this model on the information gathered by the Breast Disease Research Center, Shiraz University of Medical Sciences, Shiraz, Iran from 5673 patients with breast cancer. The goal of this study was to predict breast cancer survivability at early diagnosis, so the features used in the research are among those considered affordable, specifically at the initial steps of diagnosis. After preprocessing all of the cases and features, we constructed this model based on 1930 cases and 16 of their associated features using logistic regression method. The model then was evaluated with 10-fold cross validation.

Results: Based on all subsets of the 16 features, we evaluated numerous models. We selected a model that achieved the best sensitivity and specificity, and used fewer features as the best model. We considered this model for further analysis, which is consisted of following features: age at the time of diagnosis, type of invasion, HER2, size of the tumor, in situ component, lymph node involvement ratio, progesterone receptor status, and the total number of dissected lymph nodes. The best model obtained overall accuracy, specificity and sensitivity of 72.49%, 72.83%, and 71.85%, respectively.

Conclusion: The performance of model is quite satisfactory due to the fact that we only used features, which could be obtained at the initial steps of diagnosis. Even though, the effect of patient's age is controversial, we concluded that ageing would decrease the 60-month survivability. Our model indicated that having all type of invasions (i.e. vascular, lymphatic, etc.) would result in poorer chance of survival compared to other features effect.

Keywords: Statistical learning, Data mining, Breast cancer, Logistic regression, Survivability

♦Corresponding Author:
Ali Zamani, PhD
Zand Av., Imam Hossein Square, Department of biomedical physics and biomedical engineering, School of medicine, Shiraz university of medical sciences, Shiraz, Iran
Tel: +98 71 3234 9332
Email: zamani_a@sums.ac.ir

Received: June 06, 2018; Accepted: November 10, 2018

Introduction

Cancer, the second cause of death worldwide, caused 8.8 million deaths in 2015. Among all cancer types, breast cancer is the most frequent in women globally. According to the World Health Organization (WHO), breast cancer caused 571,000 deaths in 2015. In Iran, according to GLOBOCAN, breast cancer has the highest incidence amongst all cancers and is the leading cause of death among women.[1-3]

Detection and prognosis of a disease is an important challenge for health care management and researchers in order to make better decisions and obtain a deeper understanding. Researchers apply knowledge discovery in database (KDD) and various statistical and machine learning algorithms to solve this problem. Knowledge discovery in database[4,5] is a method to ascertain patterns and relationships between variables in datasets and consequently build a prognosis model using the extracted knowledge. The process of KDD attempts to extract a higher level of knowledge from raw data. Considering the definition of KDD proposed by Fayyad et al.,[5] data mining is one of the key parts of the KDD process, but both of these terms are interchangeable.

Typically, researchers perform the following steps within KDD. First step is the choice of a dataset related to the subject of interest that includes information about the question to be addressed. Next, the data is preprocessed and transformed into a desirable form to improve the results. In the next step, researchers use various data mining techniques considered suitable for the problem, such as clustering or classification. As the final step, researchers should present an understandable model to the expert in that particular field. Most often, the presentation is a predicting black box model or a model that describes the relationships between variables or illustrates the hidden patterns inside the dataset.[5,6]

Bruke et al.[7] began the first steps of building prognosis models by comparing the TNM staging system with artificial neural networks that used using the same features of tumor size, nodal status, and distant metastasis. The results have indicated that neural networks outperformed the TNM staging system.

In our survey, we found a number of papers that applied KDD techniques to predict the survivability of patients with breast cancer. Delen et al.[8] examined 3 classification techniques (decision tree, logistic regression, and artificial neural networks) on the surveillance, epidemiology, and end results (SEER) program dataset to predict the 5-year survivability of breast cancer. Their results indicated that the decision tree outperformed the other 2 methods. They also determined the most descriptive features among the available features by using artificial neural networks. Sensitivity analysis showed that grade, stage of cancer, and radiation were the most important features. Ahmad et al.[9] used 3 data mining methods to predict breast cancer recurrence. They built their model based on information gathered from 549 cases and a 2-year follow-up. They reported that the support vector machine (SVM) outperformed the decision tree and artificial neural networks; however, they did not clearly report the most important specific features. In other words, they did not extract any rules to show the effect of the available features, which has been the main problem of most studies in this field.

We proposed a model by that used logistic regression to predict 60-month survivability of patients diagnosed with breast cancer. In addition, we determined the essential features of this model to predict the outcome of breast cancer and attempted to describe the effect of each of the aforementioned variables on survivability based on the proposed method.

Materials and Methods

Data and its preparation

The Breast Disease Research Center, Shiraz University of Medical Sciences, Shiraz, Iran gathered the data used in this study. The dataset consisted of 5673 cases and 41 features. In the first step, we thoroughly reviewed the features of the dataset. In order to reach a prognosis in the first steps, we selected those features that could be

Table 1. Set of features used for building the model.

Features	Notes	Abbreviation
Age at diagnosis	Continuous value (year)	Age
Involved Breast	Left or right	Breast
Type of invasion	Perineural Lymphatic Lymphovascular Vascular	Invasion
Progesterone receptor	Positive Negative	Pr
Estrogen receptor	Positive Negative	Er
HER2	Positive Negative	HER2
Node involvement ratio	0-1	Nrat
Tumor size Number of involved lymph nodes Total number of dissected lymph nodes	Millimeter	Tsize Ninv Ntot
Type of surgery	Mastectomy Quadrantectomy	Operation
Grade of tumor	1,2,3	Grade
In situ component	Yes No	Insitu
Tumor necrosis	Yes No	Tnecros
Type of dissected axillary lymph node	SLNB ALND Both	Axillary
Type of breast cancer	Invasive ductal carcinoma Medullary Invasive lobular carcinoma In situ Other	Invasion

HER2: Human epidermal growth factor receptor 2; SNLB: Sentinel lymph node biopsy; ALND: Axillary lymph node biopsy

obtained specifically at the initial steps of the breast cancer diagnosis. Consequently, we deleted features that required time to be gathered (e.g., recurrence, number of times receiving chemotherapy and radiotherapy).

In the next step, we omitted cases according to the following rules:[10] i) cause of death other than breast cancer and ii) lack of follow-up for at least 60 months.

After imposing these rules, 1930 cases remained along with 16 features (Table 1). Additional comments about the selected features

Table 2. The confusion matrix.

		Predicted class	
		Yes	No
Actual class	Yes	True positive (TP)	False negative (FN)
	No	False positive (FP)	True negative (TN)

include:

Human epidermal growth factor receptor 2 (HER2) is the result of merging immunohistochemistry (IHC) and fluorescence in-situ hybridization (FISH) tests.

We obtained the node involvement ratio by dividing the number of involved nodes and the total number of dissected nodes.

Sentinel lymph node biopsy (SLNB) and axillary lymph node dissection (ALND) are the same in all aspects. However, SLNB is performed to determine if the cancer cells are present in a lymph node and if more lymph nodes should be dissected.[11]

Mathematical Background

Logistic regression

Logistic regression is a generalization of linear regression.[12] In multiple variable linear regression, inputs are usually represented as a vector like $X^T=[x_1\ x_2 \cdots x_p]$ and the output Y is calculated as in Equation (1):

$$Y = \beta_0 + \sum_1^p \beta_j\, x_j \qquad (1)$$

where β_j represents model coefficients and are estimated using the least squares method. The aforementioned linear regression can also be used for classification problems. However, in order to obtain a probability for each vector, we define the logistic function as in Equation (2):

$$P(\boldsymbol{X}) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} \qquad (2)$$

Equation (2) can be rewritten as:

$$\frac{P(x)}{1-P(x)} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p} \qquad (3)$$

Equation (3) is defined as the odds ratio and can take any values between 0 and ∞. The odds ratio illustrates the chance of event x divided by the alternative event.

Finally, by taking the logarithm of Equation (3),

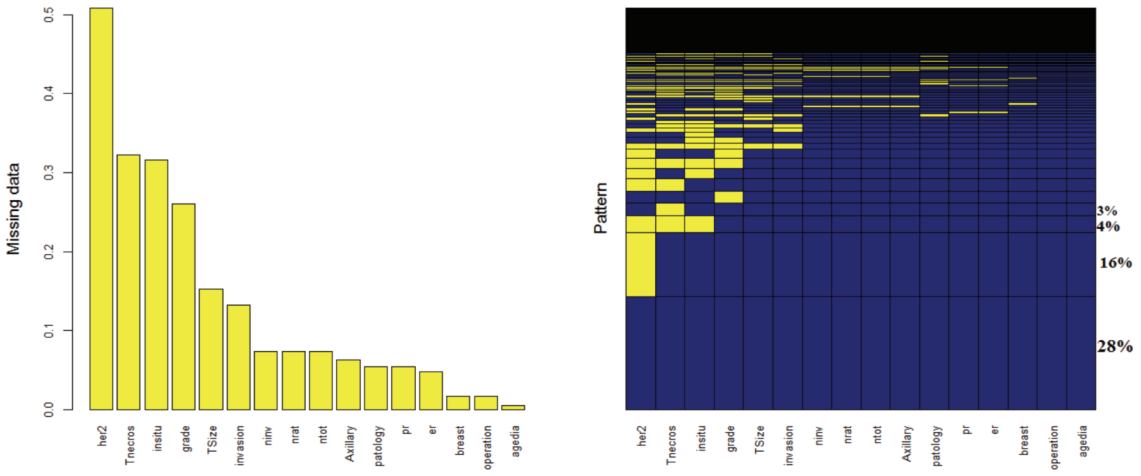

Figure 1. Bar plot and the pattern of missing values for all of the features.

we obtain the logit equation in Equation (4), which gives us a linear relationship for the logistic regression model.[12,13]

$$\mathrm{Log}\left(\frac{P(x)}{1-P(x)}\right) = \beta_0 + \beta_1 x + \cdots + \beta_p \qquad (4)$$

The coefficients were obtained by the method of maximum likelihood. To build our model we used the R programming language.[14]

Model evaluation

Evaluation criteria

In order to estimate the model's accuracy, we calculated sensitivity and specificity as follows:[6,15]

$$\text{Accuracy} = \frac{(TP+TN)}{(\text{Number of all cases})}$$

$$\text{Sensitivity} = \frac{TP}{(TP+FN)}$$

$$\text{Specificity} = \frac{TN}{(TN+FP)}$$

True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are illustrated in table 2 in a confusion matrix.

Of note, sensitivity and specificity are equal to the accuracy of the survived and not alive groups, respectively.

Evaluation method

We have used 10-fold cross-validation to reduce the error caused by bias and variance.[13] In this method, the data is divided into 10 divisions; in every step, the model is built based on 9 portions and the remaining part is used to evaluate the model. At last, to obtain each criterion, we consider the average of the estimated values for accuracy, specificity, and sensitivity of these 10 models.

Results

There were some missing values in the dataset after preprocessing. In figure 1, by using VIM16 as a package in R, we obtained the bar plot for the missing values of the different features and the pattern of these missing values in our dataset. The right panel of figure 1 illustrates that only 0.28 of cases do not have missing values; removal of HER2 added 0.16 to the usable cases. We considered these circumstances and built all possible models by using all of the possible subsets of the 16 features. For each subset of the features (216-1 subsets) we evaluated the models and, as the dataset, we omitted the cases with missing values. We calculated the accuracy, sensitivity and specificity for all of these models.

Table 3. Distribution of dependent variables for the proposed model.

	Frequency	Percentage
Did not survive (0)	152	21.84
Survived (1)	544	78.16
Total	696	100

Finally, we selected the model that simultaneously had the highest specificity and sensitivity values and fewer numbers of features as the best model.

We obtained the best model by selecting eight features: age at the time of diagnosis, type of invasion, HER2, size of the tumor, in situ component, lymph node involvement ratio, progesterone receptor status, and the total number of dissected lymph nodes. In our dataset there were 696 cases that had complete information about all of these features. We defined this model as the proposed model. Table 3 illustrates the distribution of dependent variables.

For the proposed model, the averages for 100 iterations of a 10-fold cross validation were 72.49±0.55 (accuracy), 72.83±0.62 (specificity), and 71.85±1.3 (sensitivity). Table 4 summarizes the coefficients of the proposed model and the odds ratios.

Discussion

We built this model with the aforementioned features. The small value of the standard deviation validated the stability of our selected model. According to table 4, lymph node involvement ratio, age at the time of diagnosis, progesterone receptor, and invasion were statistically significant. We extracted the presented *P*-value based on the

Table 4. Coefficients of the logit equation, related *P*-values and odds ratios for the proposed model.

	Estimate	Pr(>|z|)	Odds ratio
(Intercept)	4.417	5.72e-12 ***	82.825
invasionvascular	1.215	0.15	3.372
invasionperineural	-0.240	0.55	0.787
invasionlympvas	-0.794	0.0096 **	0.452
invasionall	-0.938	0.0016 **	0.391
invasionnone	-0.086	0.46	0.917
prnegative	-0.766	0.0003**	0.465
Age	-0.029	0.0039 **	0.971
HER2neg	0.181	0.22	1.199
Ntot	-0.017	0.13	0.983
Insituno	-0.057	0.79	0.944
TSize	-0.077	0.22	0.925
Nrat	-0.070	7.98e-10 ***	0.980

hypothesis that omitting the feature would have no effect on the model. Of note, even though the other variables did not have a small *P*-value, they are important for a better performance. We evaluated the performance of this model although these features were omitted in our exhaustive search to find the best subset of features.

The intercept term in table 4 corresponded to the reference level of the categorical features: positive HER2, lymphatic invasion, presence of an in situ component, and positive progesterone receptor. The odds ratio of each feature indicated how each change in the features could alter the chances for survival. For example, 0.97 was the odds ratio for lymph node involvement, which suggested that each 0.01 increase in this variable would decrease the chance of survival by 0.03 when all of the other features remained fixed. Likewise, for age, every unit of increment in age decreased the chance by 0.03. Our proposed model suggested that older patients would have a poorer chance of surviving breast cancer compared to younger patients.

The results of the current study differed from other studies. Rezaianzadeh et al. found no evidence of a relation between younger age and survival.[17] Chen et al. reported that middle-aged patients had a better overall survival rate than young and elderly patients.[18] In contrast, Alieldin et al. reported that young women were not found to have a poorer prognosis.[19] We omitted subjects that died due to reasons other than cancer and considered this to be the main reason for the differences in results between the studies.

The current study results had lower accuracy, specificity, and sensitivity compared to other studies. One explanation could be that some features were not used or the data did not contain those features, such as cancer stage, metastasis, and treatment (chemotherapy, radiotherapy, and hormonal treatment) despite the fact that these features were considered to be the most important predictive features of breast cancer survivability.[8,10,20] Hence, it seemed that the obtained performance could be presumed satisfactory.

Acknowledgement

The paper was extracted from the MSc thesis by Mehdi Nourelahi and supported by the Research Council at Shiraz University of Medical Sciences (95-01-01-13731).

Conflict of Interest

None declared.

References

1. World health organization (WHO). [Internet]. Cancer; [cited 2017]. Available from: http://www.who.int/en/news-room/fact-sheets/detail/cancer.
2. World health organization (WHO). [Internet]. Breast cancer; [cited 2017]. Available from: http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/.
3. Cancer today. [Internet]. Global Cancer Observatory; c2018. Population fact sheets; [cited 2017]. Available

from: http://gco.iarc.fr/today/fact-sheets-populations.
4. Piateski, G; Frawley, W. Knowledge discovery in databases. USA: MIT Press Cambridge; 1991.p.540.
5. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI Magazine*. 1996;17(3):37-54. DOI: https://doi.org/10.1609/aimag.v17i3.1230.
6. Han, J; Kamber, M; Pei, J. Data mining: concepts and techniques. 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.
7. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE Jr, et al. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer.* 1997;79(4):857-62.
8. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med*. 2005;34(2):113-27.
9. Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR. Using Three machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform.* 2013;4:124. doi: 10.4172/2157-7420.1000124.
10. BellaachiaA. Guven E. Predicting breast cancer survivability using data mining techniques. *Age.* 2006:58(13):10-110.
11. National cancer institute. [Internet]. Sentinel lymph node biopsy; [cited 2017 Oct]. Available from: https://www.cancer.gov/about-cancer/diagnosis-staging/staging/sentinel-node-biopsy-fact-sheet.
12. Hastie, T; Tibshirani, R; Friedman, J. The elements of statistical learning. 2nd ed. New York: Springer-Verlag; 2009.p.745.
13. James, G; Witten, D; Hastie, T; Tibshirani, R. An introduction to statistical learning with application in R. 1st ed. New York: Springer-Verlag. 2013. p.426.
14. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2017. Vienna, Austria. Available from: http://www.R-project.org.
15. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol.* 2008;56(1):45-50.
16. Kowarik A, Templ M. Imputation with the R Package VIM. *Journal of Statistical Software*. 2016;74(7):16. Doi: 10.18637/jss.v074.i07
17. Rezaianzadeh A, Peacock J, Reidpath D, Talei A, Hosseini SV, Mehrabani D. Survival analysis of 1148 women diagnosed with breast cancer in Southern Iran. *BMC Cancer*. 2009;9:168. doi: 10.1186/1471-2407-9-168
18. Chen HL, Zhou MQ, Tian W, Meng KX, He HF. Effect of age on breast cancer patient prognoses: A population-based study using the SEER 18 database. *PLoS One.* 2016;11(10):e0165409. doi: 10.1371/journal.pone.0165409.
19. Alieldin NH, Abo-Elazm OM, Bilal D, Salem SE, Gouda E, Elmongy M, et al. Age at diagnosis in women with non-metastatic breast cancer: Is it related to prognosis? *J Egypt Natl Canc Inst*. 2014;26(1):23-30. doi: 10.1016/j.jnci.2013.08.005.
20. Pill Choi J, Hwa Han T, Park RW. A hybrid Bayesian network model for predicting breast cancer prognosis. *J Korean Soc Med Informatics.* 2009;15(1):49-57.