Journal of Breast Cancer

J Breast Cancer 2012 June; 15(2): 230-238 http://dx.doi.org/10.4048/jbc.2012.15.2.230

ORIGINAL ARTICLE

Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine

Woojae Kim[1,*], Ku Sang Kim[1,2,*], Jeong Eon Lee[3], Dong-Young Noh[4], Sung-Won Kim[4], Yong Sik Jung[2], Man Young Park[1], Rae Woong Park[1]

Departments of [1]Biomedical Informatics and [2]Surgery, Ajou University School of Medicine, Suwon; [3]Department of Surgery, Samsung Medical Center, Seoul; [4]Department of Surgery, Seoul National University College of Medicine, Seoul, Korea

Purpose: The prediction of breast cancer recurrence is a crucial factor for successful treatment and follow-up planning. The principal objective of this study was to construct a novel prognostic model based on support vector machine (SVM) for the prediction of breast cancer recurrence within 5 years after breast cancer surgery in the Korean population, and to compare the predictive performance of the model with the previously established models. **Methods:** Data on 679 patients, who underwent breast cancer surgery between 1994 and 2002, were collected retrospectively from a Korean tertiary teaching hospital. The following variables were selected as independent variables for the prognostic model, by using the established medical knowledge and univariate analysis: histological grade, tumor size, number of metastatic lymph node, estrogen receptor, lymphovascular invasion, local invasion of tumor, and number of tumors. Three prediction algorithms, with each using SVM, artificial neural network and Cox-proportional hazard regression model, were constructed and compared with one another. The resultant and most effective model based on SVM was compared with previously established prognostic models, which included Adjuvant! Online, Nottingham prognostic index (NPI), and St. Gallen guidelines. **Results:** The SVM-based prediction model, named 'breast cancer recurrence prediction based on SVM (BCRSVM),' proposed herein outperformed other prognostic models (area under the curve=0.85, 0.71, 0.70, respectively for the BCRSVM, Adjuvant! Online, and NPI). The BCRSVM evidenced substantially high sensitivity (0.89), specificity (0.73), positive predictive values (0.75), and negative predictive values (0.89). **Conclusion:** As the selected prognostic factors can be easily obtained in clinical practice, the proposed model might prove useful in the prediction of breast cancer recurrence. The prediction model is freely available in the website (http://ami.ajou.ac.kr/bcr/).

Key Words: *Artificial intelligence, Breast neoplasms, Neural networks, Recurrence, Risk factors*

INTRODUCTION

Although 5-year survival rate of breast cancer is relatively high, the recurrence rate of it is also high (about 20% to 30%, depending on stage) [1]. One of the major challenges in breast cancer management is to classify patients into correct risk groups, for better treatment and follow-up planning. Appropriate risk assessment is critically important, not only to avoid breast cancer recurrence, but also to optimize patient's health and the use of medical resources. A variety of prediction models for breast cancer prognosis have been developed and utilized. These can be categorized as international treatment guidelines, gene expression profiles and computer-based risk calculators [2]. However all of these approaches have their own strength and weakness.

International treatment guidelines, including St. Gallen guidelines, were prepared by the clinical expert panels. Since 1978, the St. Gallen international expert consensus proposed St. Gallen guidelines for the selection of the optimal adjuvant systemic treatments for each specific patient group [3,4]. By the guideline, adjuvant chemotherapy is the recommended treatment for patients with lymph node-negative breast cancer. In spite of its simplicity to be applied in clinical setting, ethnic differences have been noted in the prognosis of lymph node-negative breast cancer. Iwamoto et al. [5] demonstrated that

Correspondence: **Rae Woong Park**
Department of Biomedical Informatics, Ajou University School of Medicine, 206 Worldcup-ro, Yeongtong-gu, Suwon 443-721, Korea
Tel: +82-31-219-5342, Fax: +82-31-219-4472
E-mail: veritas@ajou.ac.kr

*These authors contributed equally to this work.

This research is supported by the Korea Breast Cancer Foundation, and by National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (2011- 0018258 to R.W.P.).

Received: November 30, 2011 **Accepted:** April 27, 2012

the prognoses of patients with lymph node-negative breast cancer in Japan tended to be more positive than those of their Western counterparts. Thus, they concluded that the use of St. Gallen guidelines may result in overtreatment [5-8].

Microarray technology is increasingly contributing to our understanding of cancer biology, specifically via the study of gene expression. Gene expression profiles, including Oncotype DX (Genomic Health Inc., Redwood City, USA) and MammaPrint (Agendia, Amsterdam, The Netherlands) allow for a more quantitative and rationalized approach to individualized breast cancer treatments, by identifying gene activity patterns. MammaPrint is a commercial gene-expression diagnostic test that employs a 70-gene prognostic signature to classify the recurrence of breast cancer, as low-risk or high-risk [9,10]. However, MammaPrint is rather expensive, and is constrained to women with age 61 years or younger with primary invasive breast cancer of tumor size <5 cm, stage 1 or 2, and up to 3 positive lymph nodes. Both Oncotype DX and MammaPrint are also limited in that they assign almost all estrogen receptor (ER)-negative patients into high-risk group [11].

The Nottingham prognostic index (NPI) and Adjuvant! Online are computer-based models used for the prognosis of breast cancer recurrence [12-18]. The NPI is based on multivariate analysis, and has been employed broadly in clinical practice. However, the NPI employs only three prognostic factors (tumor size, tumor grade, and lymph node status) [12]. Adjuvant! Online is a web-based software application that calculates patients' 10-year survival probability, which is based on the patient's age, tumor size, grade, ER status, and nodal status [13].

The regression analysis is one of the most widely used multivariate analysis method, assuming linear relationships between the independent and dependent variables. However, it has been demonstrated that much of biomedical variables are non-linear in nature. Thus, regression method cannot be readily adapted to non-linear problems [19]. The support vector machine (SVM) method was recently suggested by Cortes and Vapnik [20]. It has been well established that SVM evidences superior prediction performance in both linear and non-linear problems [21]. The SVM is a firmly established data mining algorithm, which is widely used in a variety of fields, not only in the biomedical area, but also in the fields of engineering [22,23]. Despite its superior prediction performance, the SVM is relatively unfamiliar to the prognostic model field for cancer. Brief descript of SVM is provided in the methods section. However, detailed description of it is quite complex and requires long lists of sequential equations and notations, which are quite beyond the scope of this journal.

The aims of this study were to develop a novel prognostic model, which is based on SVM named 'breast cancer recurrence prediction based on SVM (BCRSVM)' for the prediction of Korean breast cancer recurrence within 5 years after breast cancer surgery, and to compare the predictive performance of the model with the previously established models, including Adjuvant! Online, NPI, and St. Gallen guidelines. We also identified relevant prognostic factors in breast cancer patients after surgical interventions, and calculated the importance of the prognostic factors by normalized mutual information [24].

METHODS

Study population

This study used a longitudinal observation data of 733 patients, whose information was maintained in a breast cancer center of a Korean tertiary teaching hospital. A subset of 808 patients, out of the total 1,541 diagnosed, was excluded in the study group because there was no clinical data in the research registry, as a result of unidentified and/or incomplete follow-up. Identifiable personal data of the patients were removed from the data before analysis. The protocol of this study were reviewed and approved by the Ajou University Hospital Institutional Review Board (AJIRB-MED-MDB-10-226). These data were relevant to the cohorts of breast cancer patients, who underwent breast cancer surgery, between July 1994 and April 2002, with a follow-up period of at least 60 months, and a median follow-up period of 86 months. The disease-free survival was 79.9% at the 5 years follow-up from the surgery. The mean disease-free survival was 93.3 ± 1.6 months for patients who developed recurrent breast cancer. Recurrent breast cancer includes any of ipsilateral breast tumor recurrence, contralateral breast tumor recurrence, regional lymph node metastasis and distant metastasis. Of the 733 study population, 54 subjects were excluded from the study participants, through the following exclusion criteria: male patients (11), with other multiple cancers (14), stage IV cancer (7), and unidentified follow-up time (22). Thus the resulting 679 subjects with invasive breast cancer were included in the study population (Figure 1).

Prognosis factor selection

Previously established clinical knowledge and univariate analyses were used to select relevant variables for independent variables to the prediction model. Of the 193 available variables in the data set, which were composed of administrative, epidemiologic, clinical and pathologic data, 38 clinically relevant variables were preliminarily selected by one of the authors. Through second-rounds of consensus meeting between the authors including physician, surgical pathologist and breast

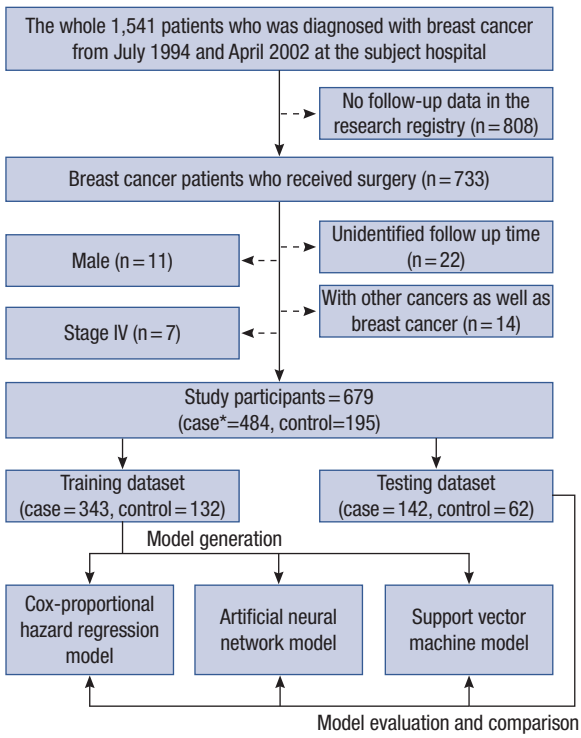

Figure 1. Patient cohort. Patient cohort fulfilled the criteria as data. *Recurrence of breast cancer within 5 years after the primary breast cancer surgery

surgeon, 14 variables were selected. Although the use of established clinical knowledge is one of the most representative methods for preliminary screening of independent variable selection, however, this would introduce a significant bias in the selection process. Therefore, as a final stage of variable selection, univariate analyses based on Kaplan-Meier analysis for categorical variables and univariate Cox regression for continuous variables were applied. Resulting statistically significant (p-value <0.05) variables included histological grade, local invasion of tumor, number of tumors, tumor size, lymphovascular invasion (LVI), ER status, and the number of metastatic lymph. 'Local invasion of tumor' was defined as not only immovability of tumor at palpation, due to chest wall fixation through pathological direct tumor invasion, but also radiologically suspicious invasion to pectoralis muscle or skin. LVI was defined as the unequivocal presence of tumor cells within any of endothelial-lined space in the breast tissue around the invasive carcinoma. All of the 7 variables were employed to the construction of the prediction model.

Selection of data mining algorithm

Although many data mining algorithms have been developed, this study entailed a comparison of the SVM, artificial neural network (ANN), and traditional Cox-proportional hazard regression model (Cox regression). The primary purpose of the SVM is to minimize the upper boundary of the generalization error by maximizing the margin by the decision boundary, called the hyperplane, which separates the subjects of one class (or group) from another, and by minimizing the empirical classification error by taking into consideration the inherent complexity of the model. The SVM employs a non-linear mapping to transform the original training data into higher-dimensional data and searches for the linear optimal separating hyperplane within this new dimension. With appropriate non-linear mapping to a sufficiently high dimension, a decision boundary can separate data into two classes (Figure 2). The SVM finds this decision boundary using support vectors and margins. In this study, the goal of SVM modeling is to classify patients who have high risk of breast cancer recurrence. The result of this classification shows recurrence probability of breast cancer within 5 years after breast cancer surgery. The ANN is a traditional data mining algorithm, and is employed extensively in a variety of clinical areas [25]. Usually SVM or ANN does not consider time-to-event. However, several approaches have been proposed to analyze data with time-to-event. The present study used single time-point approach for the output prediction of breast cancer recurrence within 5 years after the breast cancer surgery. This approach can be used to produce the estimates of outcome at a specific time of follow-up. The status (dependent) variable has either recurrent or non-recurrent within 5 years of follow-up. The Cox regression is a standard statistical model that reveals the relationships between different prognostic factors and patient survival on the basis of time-to-event. All of the selected 7 variables were entered into the Cox regression model. Clementine 12 (SPSS Inc., Chicago, USA) was employed for model construction

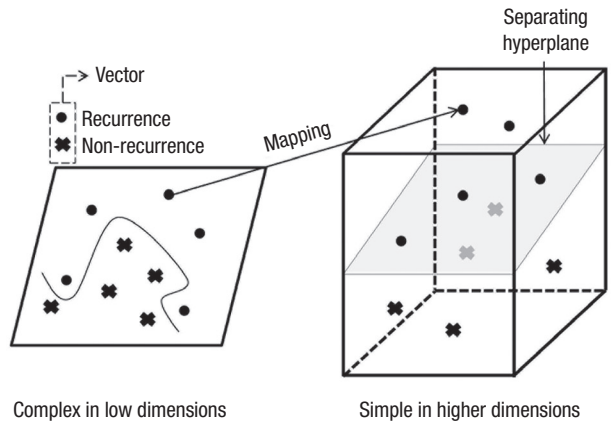

Figure 2. The basic idea of support vector machine. The data are specified as feature vectors, and then these feature vectors are mapped into a feature space. A hyperplane is computed in the feature space to optimally separate two groups of vectors.

http://dx.doi.org/10.4048/jbc.2012.15.2.230

and comparison.

Previously established recurrence prediction models

To compare the performance of the proposed models, 3 previously well-known prediction or classification models were selected: St. Gallen guidelines, NPI, and Adjuvant! Online. The International Consensus Panel, which was developed during the 2009 St. Gallen Conference, defines low clinical-risk factors, as node-negative, positive ER and positive progesterone receptors (PR), histological grade 1, low proliferation, no peritumoral vascular invasion, and tumor size of ≤2 cm [4]. Because PR, proliferation and peritumoral vascular invasion were unavailable in our data set, the 4 available factors were considered in this study.

The NPI is a prognostic model based on tumor size, histological grade, and lymph node status. The NPI point calculation equation is as follows: tumor size (cm) × 0.2 + histological grade + lymph node point (negative nodes = 1; 1-3 positive nodes = 2; ≥4 positive nodes = 3). The patients were divided, according to NPI points, into the low-risk (NPI point < 3.4) and high-risk groups (NPI point ≥ 3.4) [12].

Adjuvant! Online is a computer-based prognostic model that can be used to estimate the risk of breast cancer recurrence and death. Patients were divided into a low-risk (recurrence probability < 30%) and a high-risk group (recurrence probability ≥ 30%), using Adjuvant! Online for comparison with other prognostic models.

Prediction model validation and comparison

The holdout method was employed to reduce overfitting in the model and to derive a reliable estimate of the performance of the model. The holdout method randomly splits the entire data sample into two mutually exclusive training set (70%) and testing set (30%). The training set was utilized to generate the prediction model and the remaining 30% of the data (testing set) was employed to estimate the model's accuracy. The accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area under the curve (AUC), and Kaplan-Meier analysis of each of the models were calculated for performance comparison between the proposed BCRSVM, Adjuvant! Online, NPI, and St. Gallen.

Estimation of prognostic factor importance by normalized mutual information index

The normalized mutual information index (NMI), which is based on mutual importance, in addition to its role in calculating the correlation coefficient, is also used to determine the importance of an explanatory prognostic factor for the prediction of recurrence [24,26,27]. Mutual information is a quantitative measure for the mutual dependence of the variables. In the biomedical field, it is also employed to find functional genomics clusters in the RNA expression data [27]. We computed the entropy of the prediction results of breast cancer recurrence and the mutual information between prognostic factor patterns for prediction results, after which, the calculated mutual information was normalized. The expected fraction of uncertainty reduction, due to prognostic factors, is the NMI. This NMI ranges between 0 and 100%. If the NMI approaches to 100%, then the 2 variables are profoundly related in some form, either linearly or nonlinearly.

Statistical analysis

Disease-free survival (DFS) curves were estimated by the Kaplan-Meier method, and were compared using the log-rank test. The *p*-values of all statistical tests were two-tailed, and *p*-values equal to or less than 0.05 were employed to evaluate statistical significance. SPSS version 18.0 software (SPSS Inc., Chicago, USA) and R package (R Development Core Team, 2010) were used for all statistical analyses.

RESULTS

The clinicopathologic findings of the study participants are listed in Table 1. The mean age was 46.5 ± 11.5. The mean tumor size and number of tumors were 3.22 ± 2.50 cm and 1.07 ± 0.51 cm, respectively. There was a total of 197 of the 679

Table 1. Comparison of clinicopathologic characteristics between the case (recurrent) and control (non-recurrent) group

Variable	Total (n = 679)	Non-recurrent (n = 484)	Recurrent (n = 195)	*p*-value
Age*	46.5 ± 11.5	46.6 ± 11.2	46.2 ± 12.2	0.341†
Histological grade				< 0.001‡
Grade 1	126 (18.6)	113 (23.4)	13 (6.70)	
Grade 2	266 (39.2)	192 (39.7)	74 (38.0)	
Grade 3	287 (42.3)	179 (36.9)	108 (55.4)	
Local invasion of tumor				< 0.001‡
Yes	84 (12.4)	30 (6.2)	54 (27.7)	
No	595 (87.6)	454 (93.8)	141 (72.3)	
No. of tumor*	1.07 ± 0.51	1.01 ± 0.33	1.21 ± 0.78	< 0.001†
Tumor size (cm)*	3.22 ± 2.50	2.99 ± 2.12	3.79 ± 3.19	0.007†
Lymphovascular invasion				< 0.001†
Yes	320 (47.1)	198 (40.9)	122 (62.6)	
No	359 (52.9)	286 (59.1)	73 (37.4)	
Estrogen receptor				0.018‡
Positive	452 (66.6)	337 (69.6)	115 (59.0)	
Negative	227 (33.4)	147 (30.4)	80 (41.0)	
No. of metastatic lymph nodes	3.57 ± 7.50	2.19 ± 5.73	7.03 ± 9.90	< 0.001†

Data are presented as mean ± SD or number (%).
*Mean ± SD; †Univariate Cox regression; ‡Kaplan-Meier analysis.

cases (28.6%) recurred during the study period. Statistically significant difference between the training data set (n=475) and test data set (n=204) was not found (Table 2).

The selected prognostic factors were as follows: histological grade, local invasion of tumor, number of tumors, tumor size, LVI, ER, and number of metastatic lymph nodes. The clinico-

Table 2. Comparison of clinicopathologic characteristics between training & testing dataset

Characteristic	Total dataset (n=679)	Training dataset (n=475)	Testing dataset (n=204)	*p*-value
Age*	46.5±11.5	46.4±11.8	46.7±10.6	0.70†
Recurrence				0.45‡
Yes	195 (28.7)	133 (28.0)	62 (30.4)	
No	484 (71.3)	342 (72.0)	142 (69.6)	
Histological grade				0.99‡
Grade 1	126 (18.6)	88 (18.5)	38 (18.6)	
Grade 2	266 (39.2)	187 (39.4)	79 (38.7)	
Grade 3	287 (42.3)	200 (42.1)	87 (42.6)	
Local invasion of tumor				0.48‡
Yes	84 (12.4)	56 (11.8)	28 (13.7)	
No	595 (87.6)	419 (88.2)	176 (86.3)	
No. of tumor*	1.07±0.51	1.09±0.56	1.03±0.36	0.17†
Tumor size (cm)*	3.22±2.50	3.16±2.37	3.37±2.77	0.31†
Lymphovascular invasion				0.72‡
Yes	320 (47.1)	226 (47.6)	94 (46.1)	
No	359 (52.9)	249 (52.4)	110 (53.9)	
Estrogen receptor				0.46‡
Positive	452 (66.6)	312 (65.7)	140 (68.6)	
Negative	227 (33.4)	163 (34.3)	64 (31.4)	
No. of metastatic lymph node*	3.57±7.50	3.54±7.60	3.63±7.24	0.88†
Chemotherapy				0.78‡
Yes	377 (55.5)	263 (69.0)	114 (69.2)	
No	273 (40.2)	189 (30.7)	84 (30.0)	
Hormone therapy				0.43‡
Yes	201 (29.6)	142 (29.5)	59 (30.7)	
No	469 (69.1)	336 (69.7)	133 (69.3)	

Data are presented as mean±SD or number (%).
*Mean±SD; †Student's t-test; ‡Pearson's chi-square test.

Table 3. The importance of prognostic factors by normalized mutual information index

Variable	Normalized mutual information index	
	SVM (%)	ANN (%)
Local invasion of tumor	55.3	21.5
No. of tumor	23.2	21.5
No. of metastatic lymph node	10.5	17.1
Histological grade	5.9	12.8
Estrogen receptor	2.4	11.5
Lymphovascular invasion	1.8	10.8
Tumor size	1.0	4.8

SVM=support vector machine; ANN=artificial neural network.

Table 4. Adjusted hazard ratios (HRs) considering the risk factors listed by Cox-proportional hazard regression model for recurrence prediction of breast cancer

Variable	β	*p*-value	HR*	95% CI for exp β	
				Lower	Upper
Local invasion of tumor					
No			1†		
Yes	2.27	<0.001	9.69	6.63	14.16
Lymphovascular invasion					
No			1†		
Yes	0.45	0.01	1.57	1.12	2.20
Histological grade					
Grade 1			1†		
Grade 2	0.82	0.01	2.27	1.21	4.25
Grade 3	1.22	<0.001	3.38	1.79	6.40
Estrogen receptor					
Positive			1†		
Negative	0.20	0.25	1.22	0.87	1.70
No. of tumor	0.51	<0.001	1.67	1.41	1.98
No. of metastatic lymph node	0.04	<0.001	1.04	1.02	1.05
Tumor size	-0.03	0.21	0.97	0.93	1.02

CI=confidence interval.
*Adjusted HR considering all the risk factors listed in the table by Cox-proportional hazard regression model; †Reference.

Table 5. The performance comparison of three data mining algorithms and four prognostic models for the prediction of breast cancer recurrence within 5 years of breast cancer surgery

		Sensitivity	Specificity	PPV	NPV	Accuracy (%)	AUC (95% CI)
Algorithms	SVM	0.89	0.73	0.75	0.89	84.58	0.85 (0.79-0.91)
	ANN	0.95	0.52	0.80	0.82	81.37	0.80 (0.74-0.87)
	Cox	0.24	0.94	0.63	0.74	72.55	0.73 (0.66-0.81)
Prognostic models	BCRSVM	0.89	0.73	0.75	0.89	84.58	0.85 (0.79-0.91)
	Adjuvant!	0.95	0.38	0.69	0.83	71.43	0.70 (0.59-0.81)
	NPI	0.74	0.07	0.65	0.10	53.73	0.71 (0.61-0.81)
	St. Gallen	1.00	0.01	0.13	1.00	13.25	

PPV=positive predictive value; NPV=negative predictive value; AUC=area under the curve; CI=confidence interval; SVM=support vector machine; ANN=artificial neural network; Cox=Cox-proportional hazard regression model; BCRSVM=breast cancer recurrence prediction based on SVM; Adjuvant!=Adjuvant! Online; NPI=Nottingham prognostic index.

http://dx.doi.org/10.4048/jbc.2012.15.2.230

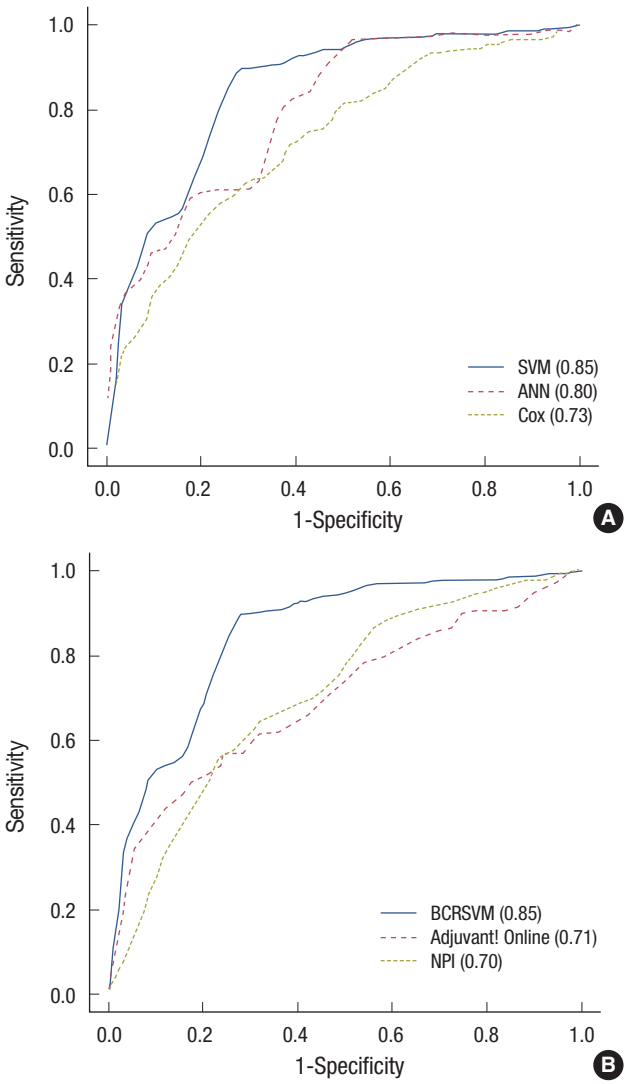

Figure 3. The receiver operating characteristic (ROC) curves of the algorithms and prognostic models at 5 years. (A) The area under the ROC (AUC) was 0.73, 0.8, and 0.85 for the Cox regression, artificial neural network (ANN), and support vector machine (SVM), respectively. (B) AUC was 0.85, 0.71, and 0.7 for breast cancer recurrence prediction based on SVM (BCRSVM), Adjuvant! Online, and Nottingham prognostic index (NPI), reprectively.

pathologic features differed significantly between the recurrent and the non-recurrent groups (Table 1).

The significance of the selected prognostic factors was compared. Local invasion of tumor was identified unanimously among the 3 algorithms, as the most important factor in the prediction of recurrence (Tables 3, 4). In the SVM and ANN algorithms, local invasion of tumor (NMI: 55.3%, 21.5%), number of tumors (NMI: 23.2%, 21.5%), number of metastatic lymph nodes (NMI: 10.5%, 17.1%), the histological grade (NMI: 5.9%, 11.5%), ER (NMI: 2.4%, 11.5%), LVI (NMI: 1.8%,

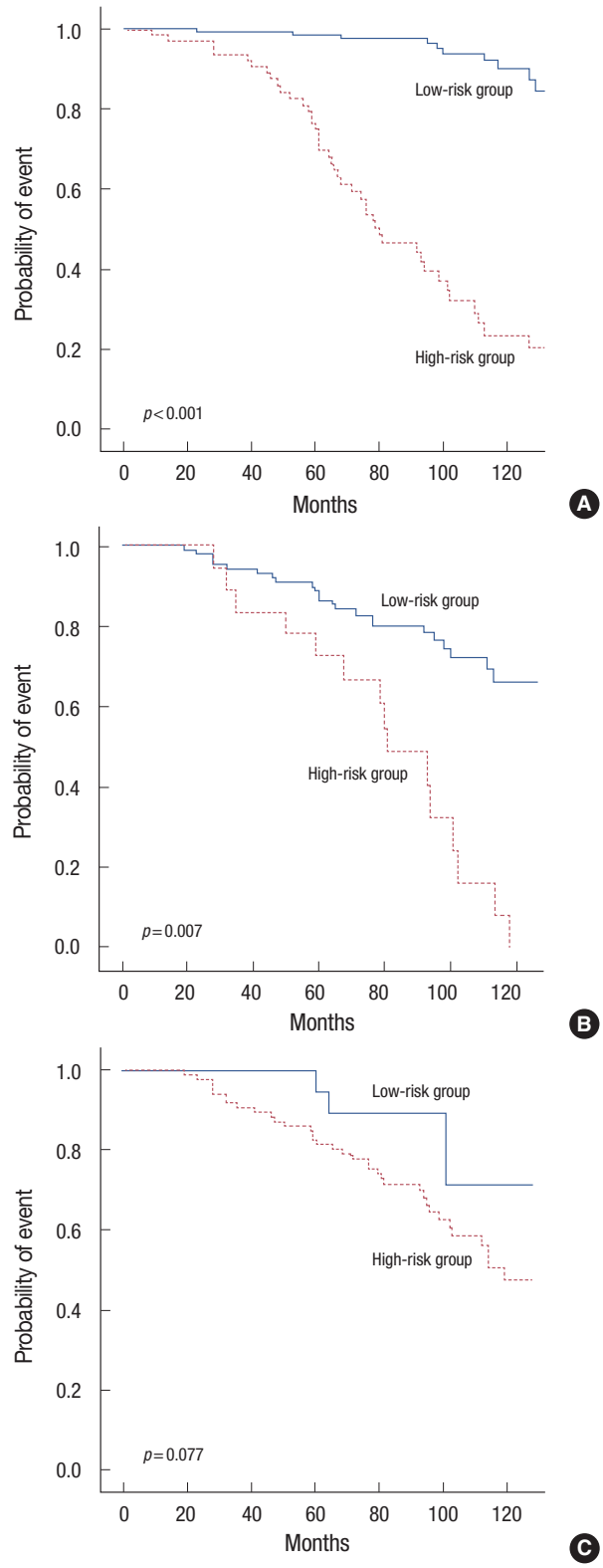

Figure 4. Prediction of disease-free survival in breast cancer patients using the three prognostic models. (A) Breast cancer recurrence prediction based on SVM (BCRSVM). (B) Adjuvant! Online. (C) Nottingham prognostic index. The log-rank test was applied for each comparison.

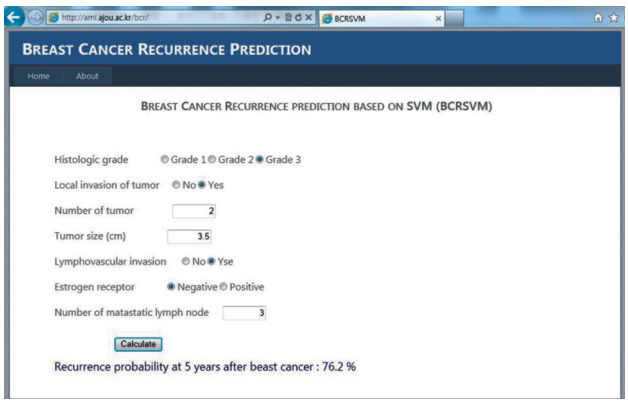

Figure 5. Website for the 'breast cancer recurrence prediction based on SVM (BCRSVM)' for easy use of the model in the clinical practice.

10.8%), and tumor size (NMI: 0.9%, 4.8%) affected the models for the prediction of breast cancer recurrence. For the Cox regression, histological grade, local invasion of tumor, number of tumors, LVI, and number of metastatic lymph nodes were associated with higher likelihoods of recurrence (Table 4).

On the basis of accuracy and AUC, SVM outperformed the ANN and Cox regression algorithms (accuracy = 84.6%, 81.4%, and 72.6%; AUC = 0.85, 0.80, and 0.73, respectively) (Table 5, Figure 3A).

Owing to its superb performance, we developed a prediction model based on SVM for predicting the recurrence of breast cancer, and named it as BCRSVM. The BCRSVM model was then compared with other well-established prognostic models. The BCRSVM proved superior to other models (AUC = 0.85). Adjuvant! Online and NPI evidenced similar AUC (0.70, 0.71, respectively) (Table 5). As the St. Gallen guidelines could divide patients only into low- and high-risk groups, the AUC could not be calculated. Its sensitivity and NPV were both 1, but its specificity and PPV were quite low (0.01, 0.13, respectively). Receiver operating characteristic (ROC) curves for each model, except for St. Gallen, are plotted in Figure 3B. The DFS estimated by the Kaplan-Meier curve revealed better discrimination of the high-risk group from the low-risk group in the BCRSVM than in the Adjuvant! Online or NPI models (Figure 4). Since the low-risk group identified via the St. Gallen guidelines included only 2 patients, the DFS curves constructed via St. Gallen guidelines could not be plotted.

DISCUSSION

This study compared a variety of machine learning algorithms to develop a novel prognostic model that is superior to that of the previously employed models for the prognosis of breast cancer recurrence. Among the various machine learning algorithms, the SVM proved superior to that of the other algorithms utilized herein. Comparing the BCRSVM based on SVM with Adjuvant! Online, St. Gallen, and NPI, the BCRSVM demonstrated superior performance. These results reveal that the BCRSVM may prove to be an effective method for the prediction of breast cancer recurrence.

In spite of the superior performance of machine learning algorithms, use of such algorithms in daily clinical practice has been quite limited, because they cannot be easily calculated with traditional calculator. For the convenience of clinicians interested in the BCRSVM, we developed a tool realizing the BCRSVM and embedded it in the webpage (http://ami.ajou.ac.kr/bcr/), as shown in Figure 5.

For non-linear modeling, ANN was proposed as a supplement or alternative to the Cox regression [14-16]. Recently, SVM has been employed for non-linear modeling in a variety of fields, most notably bioinformatics [22,23]. The SVM has been well established in the field of machine learning, but is almost completely unknown, as a cancer predictive and prognostic method. Thus far, no prognostic models based on clinicopathologic data using SVM have been developed. The SVM was first proposed by Cortes and Vapnik [20], and was identified as a type of a universal feed forward network. It provides us with a mathematical understanding of the inputs, for which the learning method is employed. SVM also evidences the relatively of high recognition ability for practical problems. In particular, the SVM method is particularly well suited to problems of a non-linear nature. The SVM helps to create a high degree of feature space to linearize the non-linear input spaces, and suggests an optimal segregation aspect for each feature [22]. One important advantage of the SVM is that the computational complexity, which is inherent to SVM, can be reduced via a quadratic optimization problem. SVM tends to be less prone than ANN to over-fitting problems. Owing to these advantages, the BCRSVM based on SVM (AUC = 0.85) also evidenced performance superior to that of the ANN-based (AUC = 0.80) and the Cox regression-based model (AUC = 0.73).

Seven prognostic factors in the BCRSVM were selected, and their importance was calculated by NMI (Table 4). Histological grade, tumor size, and number of metastatic lymph nodes were employed for all other prognostic models, including NPI, Adjuvant! Online, St. Gallen guidelines and the BCRSVM. Thus, they appear to be important and consistent prognostic factors. ER was applied as a prognostic factor in the above three models, with the exception of NPI. Local invasion of tumor and number of tumors has not previously been included in other models, except for the BCRSVM. Based on the NMI results, 'local invasion of tumor' appears to be an important

http://dx.doi.org/10.4048/jbc.2012.15.2.230

prognostic factor for the prediction of recurrence, because it accounts for 55.3% and 21.5% of importance in the SVM and ANN models, respectively. Furthermore, the hazard ratio of 'local invasion of tumor' in the Cox regression model was also top-ranked (9.691). However, it was not used in other previous prognostic models, because the local invasion of tumor is a subjective measure, rather than an objective one. Thus, the variable may evoke controversy in deciding its' positivity. However, crude definition of the variable, such as defined in this study, may be possibly sufficient, as it is demonstrated in this study. The process of defining the local invasion of tumor, in a more precise and objective manner, seems to be yet another challenge that remains ahead, or another model using purely objective variables may be required. ER status and tumor size are well known prognostic factors, however they were not significant at the Cox regression model. Discrepancy of prognostic factors between prognostic score systems, based on multivariate analysis, are not unusual, because of the effects of other prominent covariates or multi-collinearities between the variables. Possible differences in the characteristics of study participants, between the studies, may be one of the causes.

We compared and validated the prognostic accuracy of the SVM and ANN models to those of the other models, including Adjuvant! Online, St. Gallen guidelines, and NPI. The St. Gallen guidelines evidenced the highest levels of sensitivity and NPV. However, via the application of the 2009 St. Gallen guidelines to test the dataset (n = 204), only 2 patients were allocated to the low-risk group. In the previous study of Ishitobi et al. [28], the proportion of low-risk patients, according to the 1998 and 2009 St. Gallen guidelines, were 0% and 7%, respectively. Additionally, in other studies, only 10% of patients were classified as low-risk [7,29]. This discrepancy in the population of the high-risk group could result in overtreatments in clinical practice [6-8,28]. Jung et al. [6] reported that only a few patients could avoid adjuvant chemotherapy via the strict application of St. Gallen guidelines. Although the NPI employs only three prognostic factors (tumor size, tumor grade, and lymph node status), its AUC was ranked similarly to that of Adjuvant! Online. Although the NPI evidenced an AUC similar to that of Adjuvant! Online, the discrepancy in the prediction values was due to the threshold value, which is used to demarcate the low-risk and the high-risk groups. It may be necessary, in future studies, to make efforts to readjust the threshold values. Adjuvant! Online is a well-known web-based prognostic model, and was validated using external data [13,18]. We also attempted to validate Adjuvant! Online using our dataset, as Adjuvant! Online has yet to be validated in Korea. The AUCs of Adjuvant! Online were 0.66 [9] and 0.66 [30], respectively. In our study, the AUC of Adjuvant! Online was 0.70, which is similar to or somewhat higher than the previous results. The AUC of the BCRSVM (0.85) was higher than that of Adjuvant! Online (0.7) or NPI (0.71). It also exhibited relatively high predictive values for other indicators (Table 5).

Regarding the superior performance of the BCRSVM, relative to other models, several points could be considered. The BCRSVM, which involves the SVM algorithm, utilizes more factors than Adjuvant! Online or NPI. Additionally, as we employed only one hospital's data for model development and evaluation, the BCRSVM might be adjusted to our data. Therefore, in future studies, it will be necessary to validate the BCRSVM with external data, such as that acquired from other hospitals. The BCRSVM's parameters can be readily adjusted with different subject populations. It may prove beneficial to adjust prognostic models of breast cancer recurrence for each race or country, rather than imposing a universal predictive model for each. The study was also limited by the possible selection bias, which is related with the exclusion of 808 patients, who had no follow-up in the research registry.

In this study, the BCRSVM based on SVM for breast cancer recurrence was developed, and its performance was compared with that of the other prognostic models. The BCRSVM could be easily employed to assist clinicians and patients in making decisions, regarding breast cancer treatment through internet connection to the webpage (http://ami.ajou.ac.kr/bcr). The authors are currently preparing to conduct a study that would externally validate such results, with those from other hospitals.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

REFERENCES

1. Na KY, Kim KS, Lee JE, Kim HJ, Yang JH, Ahn SH, et al. The 70-gene prognostic signature for Korean breast cancer patients. J Breast Cancer 2011;14:33-8.
2. Muñoz M, Estévez LG, Alvarez I, Fernández Y, Margelí M, Tusquets I, et al. Evaluation of international treatment guidelines and prognostic tests for the treatment of early breast cancer. Cancer Treat Rev 2008;34:701-9.
3. Goldhirsch A, Wood WC, Gelber RD, Coates AS, Thürlimann B, Senn HJ, et al. Progress and promise: highlights of the international expert consensus on the primary therapy of early breast cancer 2007. Ann Oncol 2007;18:1133-44.
4. Goldhirsch A, Ingle JN, Gelber RD, Coates AS, Thürlimann B, Senn HJ, et al. Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009. Ann Oncol 2009;20:1319-29.
5. Iwamoto E, Fukutomi T, Akashi-Tanaka S. Validation and problems of St-Gallen recommendations of adjuvant therapy for node-negative

invasive breast cancer in Japanese patients. Jpn J Clin Oncol 2001;31: 259-62.
6. Jung SY, Han W, Lee JW, Ko E, Kim E, Yu JH, et al. Ki-67 expression gives additional prognostic information on St. Gallen 2007 and Adjuvant! Online risk categories in early breast cancer. Ann Surg Oncol 2009;16:1112-21.
7. Boyages J, Chua B, Taylor R, Bilous M, Salisbury E, Wilcken N, et al. Use of the St Gallen classification for patients with node-negative breast cancer may lead to overuse of adjuvant chemotherapy. Br J Surg 2002; 89:789-96.
8. Roila F, Ballatori E, Patoia L, Palazzo S, Veronesi A, Frassoldati A, et al. Adjuvant systemic therapies in women with breast cancer: an audit of clinical practice in Italy. Ann Oncol 2003;14:843-8.
9. Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 2006;98:1183-92.
10. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999-2009.
11. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. N Engl J Med 2009;360:790-800.
12. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham prognostic index in primary breast cancer. Breast Cancer Res Treat 1992;22: 207-19.
13. Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson N, et al. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol 2001;19: 980-91.
14. Jerez JM, Franco L, Alba E, Llombart-Cussac A, Lluch A, Ribelles N, et al. Improvement of breast cancer relapse prediction in high risk intervals using artificial neural networks. Breast Cancer Res Treat 2005;94: 265-72.
15. Jerez-Aragones JM, Gomez-Ruiz JA, Ramos-Jimenez G, Munoz-Perez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med 2003;27:45-63.
16. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer 2001;91(8 Suppl): 1636-42.
17. Lisboa PJ, Wong H, Harris P, Swindell R. A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. Artif Intell Med 2003;28:1-25.
18. Olivotto IA, Bajdik CD, Ravdin PM, Speers CH, Coldman AJ, Norris BD, et al. Population-based validation of the prognostic model Adjuvant! for early breast cancer. J Clin Oncol 2005;23:2716-25.
19. Aitkin M, Laird N, Francis B. A reanalysis of the Stanford heart transplant data. J Am Stat Assoc 1983;78:264-74.
20. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20: 273-97.
21. Meyer D, Leisch F, Hornik K. The support vector machine under test. Neurocomputing 2003;55:169-86.
22. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389-422.
23. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16: 906-14.
24. Estévez PA, Tesmer M, Perez CA, Zurada JM. Normalized mutual information feature selection. IEEE Trans Neural Netw 2009;20:189-201.
25. Moody J, Darken CJ. Fast learning in networks of locally-tuned processing units. Neural Comput 1989;1:281-94.
26. Kuncheva LI, Hadjitodorov ST. Using diversity in cluster ensembles. 2004 IEEE International Conference on Systems, Man and Cybernetics. 2004;2. p.1214-9.
27. Butte AJ, Kohane IS, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 2000;5:415-26.
28. Ishitobi M, Goranova TE, Komoike Y, Motomura K, Koyama H, Glas AM, et al. Clinical utility of the 70-gene MammaPrint profile in a Japanese population. Jpn J Clin Oncol 2010;40:508-12.
29. Sun JM, Han W, Im SA, Kim TY, Park IA, Noh DY, et al. A combination of HER-2 status and the St. Gallen classification provides useful information on prognosis in lymph node-negative breast carcinoma. Cancer 2004;101:2516-22.
30. Tutt A, Wang A, Rowland C, Gillett C, Lau K, Chew K, et al. Risk estimation of distant metastasis in node-negative, estrogen receptor-positive breast cancer patients using an RT-PCR based prognostic expression signature. BMC Cancer 2008;8:339.

 http://dx.doi.org/10.4048/jbc.2012.15.2.230