Computational Discrimination of Breast Cancer for Korean Women Based on Epidemiologic Data Only

Chiwon Lee,†  Jung Chan Lee,‡# Boyoung Park,§ Jonghee Bae,|| Min Hyuk Lim,¶ Dahee Kang,a,b,c,d Keun-Young Yoo,a,b,c,d Sue K. Park,a,b,c,d Youdan Kim,a,b,c,d Sungwan Kim,a,b,c,d*†

Original Article

INTRODUCTION

Breast cancer is one of leading cancers for the Korean women and its incidence rate has been rapidly increasing every year (1). An early diagnosis of breast cancer is a crucial step in reducing the mortality rate because the breast cancer survival rate rises to 95.6% (2) if breast cancer is detected during stage I. Various researches on the early diagnosis of breast cancer have been conducted and the National Cancer Institute (NCI) in the United States has released a Web-based Breast Cancer Risk Assessment Tool based on Gail model. However, it is inapplicable directly to Korean women since breast cancer risk is dependent on race. Also, it shows low accuracy (58%-59%). In this study, breast cancer discrimination models for Korean women are developed using only epidemiological case-control data (n = 4,574). The models are configured by different classification techniques: support vector machine, artificial neural network, and Bayesian network. A 1,000-time repeated random sub-sampling validation is performed for diverse parameter conditions, respectively. The performance is evaluated and compared as an area under the receiver operating characteristic curve (AUC). According to age group and classification techniques, AUC, accuracy, sensitivity, specificity, and calculation time of all models were calculated and compared. Although the support vector machine took the longest calculation time, the highest classification performance has been achieved in the case of women older than 50 yr (AUC = 64%). The proposed model is dependent on demographic characteristics, reproductive factors, and lifestyle habits without using any clinical or genetic test. It is expected that the model could be implemented as a web-based discrimination tool for breast cancer. This tool can encourage potential breast cancer prone women to go the hospital for diagnostic tests.

Keywords: Breast Neoplasms; Support Vector Machines; Neural Networks; Computers

© 2015 The Korean Academy of Medical Sciences. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
models are helpful for clinical diagnosis and breast cancer treatment planning (9). Hybrid models of ANN and BN using clinical data (especially mammography) have been previously used in the risk analysis and diagnosis of breast cancer (10, 11). BN models based on a number of risk factors have been previously investigated for Korean women (8, 12). However, these studies have limitations in their generalizability because of small sample sizes. However, a discrimination of breast cancer using only epidemiological risk factors based on case-control study data set has not been performed previously.

Breast cancer risk assessment requires the analysis of different characteristics that depend on race (13). Particularly, the incidence of breast cancer in Korean women is lower than in Western women, and the incidence in Korean women begins to decrease after the age of 50, whereas the incidence in Western women increases gradually with age until 80 yr old, as shown in Fig. 1 (14, 15). The Gail model based on western women data would have to be recalibrated to rates of breast cancer in Korea to make it applicable. Therefore, a Korean breast cancer risk assessment tool (KoBCRAT) which is more applicable to Korean women than the Gail model based on western populations has been developed using a case-control study (Seoul Breast Cancer Study, SeBCS) with 4,647 cases and 4,601 controls (16). Although the same raw data from SeBCS has been used to develop each model, there are two main differences between the KoBCRAT and the breast cancer risk classification algorithm proposed in this research. First, unlike the KoBCRAT, the proposed algorithms in this research were developed and validated with a rich data set, which were 2,291 cases and 2,283 controls through different selection criteria. Breast cancer risk classification algorithms for Korean women were developed considering those rates as well as augmenting risk factors for Korean women. Data were separated with respect to age groups and risk factors were selected according to the methods. Secondly, KoBCRAT has been developed based upon Gail model’s equation while the three different classifiers such as the support vector machine (SVM), ANN, and BN are being considered in this research. Further, the results of three different methods were compared for two different reasons; 1) to find out which method is the best and 2) to validate a method comparing with the other two methods. Results showed that area under the curve (AUC) values of receiver operating characteristic (ROC) curve are calculated as 60%-64%. Accuracy, sensitivity, specificity, single iterative calculation time, and contribution of a specific risk factor to models are presented in detail. Although the true diagnosis of breast cancer cannot be replaced by proposed model for predicting the risk of breast cancer, the proposed model can encourage potential breast cancer prone women to go the hospital for diagnostic tests.

MATERIALS AND METHODS

Acquisition of epidemiological data and study subjects
This study is based on cases and controls recruited from 1994 to 2007 in the SeBCS, which is a community-based breast cancer case-control study. The cases (n = 4,647) consisted of female patients diagnosed with histologically confirmed breast cancer, who were admitted to three major general hospitals in Korea. The controls (n = 4,601) were selected from a cancer-free health examinee population. Only non-invasive information was obtained through survey questionnaire for the controls. From these groups, frequency matching by 5-yr age groups and enrollment years (1994-1997, 1998-2002, 2003-2007) was performed and 2,370 patients and 2,370 controls were selected as a trial set. A total of 166 subjects were excluded in the trial set because they did not have a sufficient number of risk factor variables. Finally,
4,574 subjects were used in this study, including 2,291 cases and 2,283 controls. All trial sets were divided into training set and test set. A training set was used to train the three classification algorithms. Then, a test set was evaluated by trained classification algorithms for receiver operating characteristic (ROC) curve. From a test set, cases correctly classified into the case category were considered as true positives (TP), controls correctly classified into the control category were considered as true negatives (TN), controls classified into the case category were considered as false positives (FP), and cases classified into the control category were considered as false negatives (FN).

**Major risk factor selection**

A variety of risk factors have been associated with breast cancer risk. Demographic characteristics, reproductive factors, and lifestyle habits (17, 18) were assessed by trained interviewers after receiving informed consent from the patients.

The trend of incidence rate for Korean women differs from that for American white women, as shown in Fig. 1. The incidence rate of breast cancer has a different aspect, depending on the age group, compared with the incidence rate of other cancers (19). Therefore, we divided our data into two the datasets to select major risk factors for Korean women, with one dataset including subjects under the age of 50 and the other including subjects at and over the age of 50, because the trends of incidence rates differ around the age of 50 among Korean women. The major risk factors of breast cancer were selected by computing the odds ratios (ORs) and 95% confidence intervals (95% CIs) by repeating univariate and stepwise multivariate logistic regression analyses. The risk factors included in the final model were based on Wald tests for individual parameters (20) and on information from previously established breast cancer risk factors.

The age of first full-term pregnancy (AFFP), the number of children (NOC), the age of menarche (AOMn), body mass index (BMI), family medical history of breast cancer (FMH), menopausal status (MS), regular mammography (RM), and regular exercise (RE) were considered as major risk factors in the trial set. In addition to these factors, lifetime exposure to estrogen (estrogen exposure duration, ED), the age at menarche and gestation period for premenopausal and postmenopausal women, and the age at menopause for postmenopausal women were considered. In case of AFFP, age of subjects was considered as numeric data. Input value of ‘No children’ subjects was set to be 0. Numeric data were used for input of NOC, AOMn, BMI, and ED risk factors. Several risk factors in forms of non-numeric data such as FMH, MS, RM, and RE, were assigned as 0, 1, or 2 in forms of numeric data.

In total, 4,599 calculations were performed in this study. The raw data obtained from the epidemiological data set were separated by three age group models: under the age of 50 (U50), equal to or over the age of 50 (O50), and all ages (ALL). The analysis methods were developed according to the classification algorithms (SVM, ANN, and BN). These methods were configured differently depending on the selected combination of risk factors to find the optimal combination of selected risk factors. Based on all of the risk factor selection combinations (selecting from one to nine risk factors), the performances of a total of 511 combinations (Sum of \(sC_i\) [1 = 1, 2, 3, … , 9]) were analyzed by each classification algorithm. One thousand iterations were performed for each calculation to yield a total of 4,599,000 iterations (3 classification algorithms × 3 age groups × 511 risk factor selection combinations × 1,000 iterations).

**Configuration of classifiers: Support vector machine**

SVM is a classification algorithm based on a statistical machine learning technique. This method was first introduced in 1995 and has since rapidly spread to be used in various applications (21-23). Breast cancer diagnosis using clinical data was also developed (24). Given a set of training data, SVM finds hyperplanes in a higher dimensional feature space, and the optimal hyperplane with the largest distance to the nearest training data provides a high generalization of the performance of the classification or regression. The non-linear kernel functions can be applied to the algorithm to classify linearly inseparable problems. After a preliminary comparison test (linear function, radial basis function, and polynomial function), we used a nonlinear third-order polynomial kernel function in this study since all SVM models (SVM-ALL, SVM-U50, and SVM-O50) showed a better performance with polynomial, linear, and radial basis function, respectively. Among the polynomial functions, third-order polynomial function which showed highest mean values was chosen for three SVM models.

**Configuration of classifiers: Artificial neural network**

ANN is a classical artificial intelligence algorithm that has a structure of simulated neurons that are connected together in a similar manner as neurons in the biological brain. Each neuron of the ANN receives inputs either from a number of other neurons or from an external stimulus. The output of a neuron is a function of the weighted sum of the connected neurons, and the weighting factors between neurons are iteratively trained to output an optimal result by the back-propagation procedure. For a preliminary test on layer structures, single layer structures (number of nodes: 5, 9, 18) were simulated, but the result was slightly lower than the one of double layer structures. Therefore, several simulations were conducted for the double layer structures to find the optimal layer structure. A layer structure with 3 or over hidden layers did not improve the classification performance. As a result, a double hidden layer neural network was established with 9 nodes in the input layer and 5 nodes in the hidden layers, as shown in Fig. 2. A hyperbolic tangent sigmoid function was applied for the hidden and output layers as the trans-
Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)

Sensitivity = \frac{TP}{(TP+FN)} \quad (2)

Specificity = \frac{TN}{(TN+FP)} \quad (3)

The performances of a computational discrimination method can be evaluated by the ROC analysis. ROC analysis provides criteria for selection of the optimal model and clinical decision. The ROC curve is presented at varying thresholds to balance sensitivity and specificity. For plotting the ROC curve, simulation results of ANN and BN which were evaluated as varying threshold values were calculated by Pprop & msereg and probability function, respectively (25, 26). In SVM, threshold controlled by bias which was intercept of the hyperplane. The prediction validity can be estimated by the AUC of the ROC curve. The AUC was calculated using a numerical integration of the ROC curve.

To compare the performance between the methods, a single iterative calculation time was measured from the total iterative processing time divided by 1,000 according to the classification algorithms, age division models and risk factor selections. All calculations were conducted in Matlab (The Mathwork, Inc., Natick, MA, USA) using a Windows 7 64-bit operating system on an Intel Xeon E5405@2.00 GHz CPU with 8 GB RAM.

Ethics statement
This retrospective analysis study has been approved by the Seoul National University institutional review board (IRB No. C-0909-048-295). All participants gave their informed consent about survey questionnaire.

RESULTS
The characteristics of major risk factors from the trial set of the SeBCS are summarized in Table 1. Breast-feeding per child has been found to reduce the breast cancer risk, but it is a controversial issue because of the different results among the subjects with...
Table 1. Major risk factors selected from the trial set of the Seoul Breast Cancer Study (SeBCS)

| Risk factors | Total set (n = 4,574) | U50 (n = 2,622) | O50 (n = 1,952) |
|--------------|----------------------|----------------|----------------|
|              | Cases (n = 2,291)    | Controls (n = 2,283) | Cases (n = 1,314) | Controls (n = 1,308) | Cases (n = 977) | Controls (n = 975) |
| Age of first full-term pregnancy, yr (No. %) | | | | | | |
| No children  | 208 (8.08)           | 183 (8.02)       | 157 (11.95)     | 161 (12.31)       | 51 (5.22)       | 22 (2.26)         |
| < 24 (early pregnancy) | 459 (20.03) | 579 (25.36) | 154 (11.72)     | 231 (17.66)       | 305 (31.22)     | 348 (35.69)       |
| 24 to 28     | 1,191 (51.99)       | 1,157 (50.68)    | 708 (53.88)     | 667 (50.99)       | 483 (49.44)     | 490 (60.26)       |
| > 28 (late pregnancy) | 433 (18.9) | 364 (15.94) | 295 (22.45)     | 249 (19.04)       | 138 (14.12)     | 115 (11.79)       |
| Number of children, (No. %) | | | | | | |
| 0 (no childbirth) | 239 (10.43) | 204 (8.94) | 178 (13.55)     | 176 (13.45)       | 61 (6.24)       | 28 (2.87)         |
| 1 to 2       | 1,458 (63.64)       | 1,392 (60.97)    | 993 (75.57)     | 965 (73.78)       | 465 (47.6)      | 427 (43.8)        |
| ≥ 3 (many childbirths) | 594 (25.93) | 687 (30.09) | 143 (10.88)     | 102 (12.77)       | 451 (46.16) | 520 (53.33) |
| Age of menarche, yr (mean ± SD) | | | | | | |
| ≤ 15 (early menarche) | 1,492 (65.12) | 1,351 (59.18) | 1,006 (76.56) | 951 (72.71) | 486 (49.74) | 400 (41.03) |
| > 15 (late menarche) | 799 (34.88) | 932 (40.82) | 308 (23.44) | 357 (27.29) | 491 (50.26) | 575 (58.97) |
| Body mass index (mean ± SD) | | | | | | |
| < 25 (underweight woman) | 1,709 (74.6) | 1,753 (76.78) | 1,081 (82.27) | 1,087 (83.11) | 628 (64.28) | 666 (68.31) |
| 25 to 30 | 520 (22.77) | 486 (21.29) | 216 (16.44) | 202 (15.44) | 304 (31.11) | 284 (29.13) |
| ≥ 30 (overweight woman) | 62 (2.7) | 44 (1.93) | 17 (1.29) | 19 (1.45) | 45 (4.61) | 25 (2.56) |
| Family medical history of breast cancer (No. %) | | | | | | |
| Yes | 97 (4.23) | 52 (2.28) | 58 (4.41) | 30 (2.29) | 39 (3.99) | 22 (2.26) |
| No | 2,194 (95.77) | 2,231 (97.72) | 1,202 (91.48) | 1,143 (87.77) | 924 (96.01) | 953 (97.74) |
| Menopausal status (No. %) | | | | | | |
| Premenopausal | 165 (7.2) | 295 (12.92) | 112 (8.52) | 165 (12.61) | 53 (5.42) | 130 (13.33) |
| Postmenopausal | 2,126 (92.8) | 1,988 (87.08) | 1,202 (91.48) | 1,143 (87.37) | 924 (96.01) | 953 (97.74) |
| Regular mammography (No. %) | | | | | | |
| No | 1,275 (55.65) | 957 (41.92) | 709 (53.96) | 537 (41.06) | 566 (57.93) | 420 (43.08) |
| Regular | 1,016 (44.35) | 1,326 (58.08) | 605 (46.04) | 771 (58.94) | 411 (42.07) | 555 (56.92) |
| Regular exercise (No. %) | | | | | | |
| No | 1,429 (62.37) | 1,283 (56.2) | 801 (60.66) | 737 (56.33) | 628 (64.28) | 546 (56.68) |
| Regular | 862 (37.63) | 1,000 (43.8) | 513 (39.34) | 571 (43.66) | 349 (35.72) | 429 (44.4) |
| Estrogen duration, yr (mean ± SD) | 28.57 ± 7.46 | 27.80 ± 7.14 | 24.61 ± 5.54 | 24.40 ± 5.48 | 33.88 ± 6.32 | 32.35 ± 6.56 |

US0, under 50 yr old group; O50, equal to or over 50 yr old group.

differing age groups, control characteristics and yearly trends from 1994 to 2007. Risk factors, such as smoking, postmenopausal hormonal replacement therapy, non-steroidal anti-inflammatory drug use, oral contraceptives and ionizing radiation, were not associated with the risk factors of breast cancer, but must be further evaluated among Koreans. Finally, nine risk factors were observed to be statistically correlated with age in Korean women. The maximum AUCs by three classifier algorithms (SVM, ANN, and BN) and three age division models (ALL, U50, and O50) using specific risk factor combinations are shown in Table 2. The SVM-O50 model produced the highest AUC of 0.6415 among all of the combinations. The O50 models resulted in higher values than those of the other age division models; the AUC values from SVM, ANN, and BN were 0.6415, 0.6383, and 0.6290, respectively. The AUC values for ALL models were 0.6213, 0.6173, and 0.6101, respectively. U50 models resulted in lower values than those of the other age division models; 0.6076, 0.6027, and 0.6027, respectively. All models that were performed...
for one thousand iterations were demonstrated as normal distribution by Shapiro-Wilk Test (27). A comparison of statistical significance based on the classification algorithms, one-way Analysis Of Variance (ANOVA) of which significance level was set to be 0.05, was conducted for three age groups: ALL (SVM-ANN-BN), U50 (SVM-ANN-BN) and O50 (SVM-ANN-BN) using a SPSS software (IBM Corporation, Armonk, NY, USA). As a result, significance probabilities for the ‘Test of Homogeneity of Variances’ were estimated as 0.641 (ALL age group), 0.284 (U50 age group) and 0.704 (O50 age group), respectively, which means that the hypothesis for homogeneity of variances is satisfied. The P values, which were obtained by ANOVA, were indicated as < 0.001 for all age groups. This result stands for the fact that each model has not the same mean value. In addition, there was a significant difference founded by the Tukey’s Honestly Significant Difference in the most of results (Multiple comparisons: P < 0.05), except for the case of U50 (ANN-BN), whose significance probability was 0.278. Based on the above values, it was demonstrated that the values from SVM were highest, the values from ANN were second highest, and the values from BN were lowest. The corresponding ROC curves for the maximum AUC achieved in each method are presented in Fig. 4. The SVM algorithm provided the highest AUCs of the all-age division models.

Maximum accuracy was acquired from each ROC analysis by using the optimal combinations of risk factors that achieved the highest AUCs in Table 2. The maximum sensitivity and specificity were attained using the threshold that achieved the maximum accuracy. In contrast to the AUC results, the highest accuracy was achieved in ANN-O50 (0.6230). In the other age division models, ANN resulted in higher accuracy than the other classification algorithms. The highest sensitivity and specificity were observed in BN-U50 (0.6192) and BN-O50 (0.6833), respectively.

The single iterative calculation times are shown in Table 2. The calculation measurements were mainly influenced by the number of participating subjects. The ALL models required more calculation time than the other age division models, U50 and O50, because the subjects of the ALL models (n = 4,574) were 5.4916 s, and 3.9561 s, respectively. The ANN-ALL model required approximately twice as long a calculation time as that of ANN-U50 or ANN-O50. For example, to perform a single iterative calculation, ANN-ALL, ANN-U50, and ANN-O50 required 9.9536 s, 16.1134 s, and 2.9548 s, respectively.

The contribution of a specific risk factor on the AUC was demonstrated that the values from SVM were highest, the corresponding ROC curves for the maximum AUC achieved in each method are presented according to classification algorithms and age division models.
AUC among all risk factors. NOC, AOMn, BMI, MS, and RE had no observable effect on increasing the AUC. FMH, ED, and AFFP exhibited different results depending on the classifier algorithms and age division models. FMH exhibited the greatest effect on decreasing the AUC of the BN-O50 model. ED exhibited only negative effects on the AUCs of the BN-U50 models for all of the classification algorithms, and AFFP exhibited a remarkable negative effect on the AUCs of the SVM-O50 and ANN-O50 models.

**DISCUSSION**

The discrimination of breast cancer for Korean women was performed with computational classifier algorithms based on an epidemiological data set and demonstrated that the SVM algorithm produced the best AUC among the three classifier algorithms. In the SVM-O50 model, an AUC of 0.6415 was achieved when the ED risk factor was excluded. Although ED negatively affected AUC in this model, the combinations that included most of the other risk factors generally elevated the AUC and accuracy. RM contributed to a remarkable increase in the AUC compared with other risk factors. Furthermore, RM was included in all the optimal combinations of risk factors according to the classification methods and age division models. These observations imply that RM is the most important risk factor in...
The O50 model exhibited better performance than the other study. If genetic information, mammographic diagnosis, and the mortality rate of breast cancer will decrease. Addition of risk factors and using a proper classifier in a certain discrimination problem could be considered as an efficient, handy method in predicting breast cancer risk because the model is developed in this study. Therefore, the early diagnosis of breast cancer will be more effective, and the mortality rate of breast cancer will decrease. Fig. 5. Contribution of a specific risk factor on the area under curve (AUC). AFFP, age of first full-term pregnancy; NOC, number of children; AOMn, age of menarche; BMI, body mass index; FMH, family medical history of breast cancer; MS, menopausal status; RM, regular mammography; RE, regular exercise; ED, estrogen duration; SVM, support vector machine; ANN, artificial neural network; BN, Bayesian network; U50, under 50 yr old group; O50, equal to or over 50 yr old group.

Increasing the prediction performance of breast cancer discrimination. ED and AFFP contributed considerably to increasing the performance as well. These two risk factors are commonly related to the effect of the estrogen hormone. This observation of the connection to estrogen is in accordance with previous literature (28). Although certain risk factors are viewed as negative contributions to the AUC, the optimal AUC was achieved when most of the risk factors considered in this study were included. Therefore, although a certain risk factor may have an undesirable effect, the combination of risk factors provides a synergistic effect that increases the AUC. This observation implies that when the AUC is not improved by a particular risk factor, combining several risk factors and using a proper classifier algorithm can improve the prediction validity. The AUC value between 0.6027 and 0.6415 seems lower considering other studies, but the proposed method in this research is a prediction technique without using any clinical data while others are mainly dependent on clinical data. Because of the fact that SVM is applied to epidemiological case-control data concerning the breast cancer, these AUC values are improved comparing with the conventional method, logistic regression. Therefore, the developed model in this study could be considered as an efficient, handy method in predicting breast cancer risk because the model is purely dependent on the survey questionnaire. The better results are expected when the genetic information, mammographic image, or biopsy are added to the model developed in this study.

The O50 model exhibited better performance than the other age division models. This result was due to a characteristic of our epidemiological data. Some risk factors considered in the classifiers were affected by age. For example, AFFP, NOC, MS, and ED might not yet be determined for the subjects in U50. Therefore, more ascertainable risk factors could be collected from O50.

O50 is also the lowest time-consuming age model, as indicated in Table 2. This result was due to a smaller sample size. The O50 dataset included only 1,952 subjects, whereas the U50 and ALL datasets included 2,622 and 4,574, respectively. Because the ALL models consisted of a large size of subjects, they required the longest computation time among all the classification methods.

Most of the previous breast cancer risk analysis studies were based on Western population data. These methods were inappropriate for Koreans because of the different characteristics of risk factors, in particular the lower incidence rates, different age-specific incidences, and different mortality patterns. Among the entire models, the highest AUC was achieved using the SVM method. We observed an even prediction validity with the AUC of 60%-64% compared with the AUC of 59%-64% achieved by logistic regression based on identical epidemiological data (29). In particular, the following models, SVM-U50, ANN-U50, BN-U50, and SVM-O50 produced a higher AUC than those of models using logistic regression analysis.

Classification algorithms for pattern recognition and discrimination problems are usually regarded as ‘black-boxes.’ Therefore, the consideration of all possible risk factors and transfer functions in an algorithm is a difficult and tremendously time-consuming process. Optimization of all of the possible approaches would take more than several months, even with high performance personal computers. However, after the final training of an algorithm with the optimal selection of risk factor combinations, the assessment of an individual woman would take under several seconds. For further enhancement of the classifier performance, ensemble methods using SVM, ANN, BN, and other classification algorithms could be employed. The main idea of the ensemble method is that the performance of combining the classifiers is superior to that of each individual classifier in a certain discrimination problem (30).

The expected benefits of the present method are described below. Only simple interviews or direct inputs of questionnaire regarding a subject’s status are required to perform a discrimination of breast cancer risk. However, regular mammography inspections are required for the detection of a newly developed cancer. The proposed methodology does not determine the onset of breast cancer, which can be performed through mammographic diagnosis. However, it can encourage potential breast cancer-prone women to go the hospital for diagnostic tests. Therefore, the early diagnosis of breast cancer will be more effective, and the mortality rate of breast cancer will decrease.
ally, if the present method is designed in the form of a web-based or smartphone application, women who want to know their own risk of breast cancer will be able to access this information easily in daily life.

DISCLOSURE

The authors have no conflicts of interest to disclose.

AUTHOR CONTRIBUTION

Manuscript conception and preparation: Lee C, Lee JC, Park B, Lim MH, and Kim S. Simulation of algorithms: Lee C, Lee JC, and Bae SK. Data collection and provision: Kang D, Yoo KY, and Park SK. Internal review for draft: Kim Y and Kim S. Manuscript approval: All authors.

ORCID

Chiwon Lee http://orcid.org/0000-0001-7342-6190
Jung Chan Lee http://orcid.org/0000-0002-4546-0962
Boyoung Park http://orcid.org/0000-0003-1902-3184
Jonghee Bae http://orcid.org/0000-0002-2069-0386
Min Hyuk Lim http://orcid.org/0000-0003-1547-2804
Dahee Kang http://orcid.org/0000-0003-4031-5878
Keun-Young Yoo http://orcid.org/0000-0002-3726-5390
Sue K. Park http://orcid.org/0000-0001-5002-9707
Youdan Kim http://orcid.org/0000-0001-5041-8243
Sungwan Kim http://orcid.org/0000-0002-9318-849X

REFERENCES

1. Shin HR, Joubert C, Bonioli M, Hery C, Ahn SH, Won YJ, Nishino Y, Sobue T, Chen CJ, You SL, et al. Recent trends and patterns in breast cancer incidence among Eastern and Southeastern Asian women. Cancer Causes Control 2010; 21: 1777-85.
2. Survival analysis of Korean breast cancer patients diagnosed between 1993 and 2002 in Korea: a Nationwide Study of the Cancer Registry. J Breast Cancer 2006; 9: 214-29.
3. National Cancer Institute. Breast cancer risk assessment tool. Available at http://www.cancer.gov/bcrisktool/ [accessed on 8 December 2014].
4. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989; 81: 1879-86.
5. Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. J Natl Cancer Inst 2001; 93: 358-66.
6. Boyd CR, Tolson MA, Copes WS. Evaluating trauma care: the TRISS method. Trauma Score and the Injury Severity Score. J Trauma 1987; 27: 370-8.
7. Levy SM, Herberman RB, Maluash IM, Schlien B, Lippman M. Prognostic risk assessment in primary breast cancer by behavioral and immunological parameters. Health Psychol 1985; 4: 99-113.
8. Choi JP, Han TH, Park RW. A hybrid bayesian network model for predicting breast cancer prognosis. J Korean Soc Med Inform 2009; 15: 49-57.
9. Kiyan T, Yildirim T. Breast cancer diagnosis using statistical neural networks. IJ-IEEE 2004; 4: 1149-53.
10. Ayer T, Alagöz O, Chhatval J, Shavlik JW, Kahn CE Jr, Burnside ES. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. Cancer 2010; 116: 3310-21.
11. Burnside ES, Rubin DL, Fine JP, Shachter RD, Sisney GA, Leung WK. Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience. Radiology 2006; 240: 666-73.
12. Lee SM. Comparisons of predictive modeling techniques for breast cancer in Korean women. J Korean Soc Med Inform 2008; 14: 37-44.
13. Smigal C, Jemal A, Ward E, Cokkinides V, Smith R, Howe HL, Thun M. Trends in breast cancer by race and ethnicity: update 2006. CA Cancer J Clin 2006; 56: 168-83.
14. Centers for Disease Control and Prevention. United States Cancer Statistics: 1999-2011 Cancer Incidence and Mortality Data. Available at www.cdc.gov/uscs [accessed on 08 December 2014].
15. Jung KW, Park S, Kong HJ, Won YJ, Lee JY, Park EC, Lee JS. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2008. Cancer Res Treat 2011; 43: 1-11.
16. Park B, Ma SH, Shin A, Chang MC, Choi JY, Kim S, Han W, Noh DY, Ahn SH, Kang D, et al. Korean risk assessment model for breast cancer risk prediction. PLoS One 2013; 8: e76736.
17. McPherson K, Steel CM, Dixon JM. ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics. BMJ 2006; 321: 624-8.
18. Suzuki S, Kojima M, Tokudome S, Mori M, Sakauchi F, Fujino Y, Wakai K, Lin Y, Kikuchi S, Tamakoshi K, et al.; Japan Collaborative Cohort Study Group. Effect of physical activity on breast cancer risk: findings of the Japan collaborative cohort study. Cancer Epidemiol Biomarkers Prev 2008; 17: 3396-401.
19. Won YJ, Sung J, Jung KW, Kong HJ, Park S, Shin HR, Park EC, Ahn YO, Hwang IK, Lee DH, et al. Nationwide cancer incidence in Korea, 2003-2005. Cancer Res Treat 2009; 41: 122-31.
20. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49: 1373-9.
21. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995; 20: 273-97.
22. 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.
23. Rodriguez-Moguel L, Bega-Ramos B. Risk of breast cancer of low differentiation in tumors with estrogen-negative receptors. Ginecol Obstet Mex 1999; 67: 503-7.
24. Polat K, Güney S. Breast cancer diagnosis using least square support vector machine. Digit Signal Process 2007; 17: 694-701.
25. Hecht-Nielsen R. Theory of the backpropagation neural network. Proceedings of the International Joint Conference on Neural Networks; Washington, D.C.: IEEE Press, 1989, p.593-605.

http://dx.doi.org/10.3346/jkms.2015.30.8.1025
26. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Mateo, CA: Morgan Kaufmann Publishers Inc., 1988.
27. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965; 52: 591-611.
28. Clemons M, Goss P. Estrogen and the risk of breast cancer. N Engl J Med 2001; 344: 276-85.
29. Park B. Development of sporadic and hereditary breast cancer risk assessment model in Korean women. Seoul: Seoul National University, 2012. Dissertation.
30. Rokach L. Pattern classification using ensemble methods. Danvers, MA: World Scientific Pub. Co., 2010. (Series in Machine Perception and Artificial Intelligence; vol 75).