Clinics in Surgery

Research Article
Published: 26 Feb, 2019

# Deep Survival Model Identifies Prognostic Subgroups of Triple-Negative Breast Cancer Patients

*Isaac Kim[1], Sung Wook Seo[2]*, Jeong Eon Lee[1]*, Hee Jun Choi[1], Jai Min Ryu[1], Se Kyung Lee[1], Jong Han Yu[1], Seok Won Kim[1] and Seok Jin Nam[1]*

[1]Department of Surgery, Sungkyunkwan University School of Medicine, Republic of Korea

[2]Department of Orthopedic Surgery, Sungkyunkwan University School of Medicine, Republic of Korea

## Abstract

**Objective:** The aim of this study was to develop predict model for survival of BC patients using deep learning method and to apply our model for classifying the prognostic subgroups of TNBC patients.

**Methods:** We enrolled 7915 patients above pathologic stage 1 among 16254 patients who underwent surgery for breast cancer between May 1996 and January 2015. The enrolled patients were randomly sorted patients into a training group (80%, *n*=6288) and a test group (20%, *n*=1627). The Survival Recurrent Neural Network (SRN) model was constructed based on logistic regression function and long short-term memory recurrent neural network.

**Results:** The mean area under the receiver operating characteristics curve (AUC) of validation sets was 0.780 at 1st year, 0.840 at the 2nd year, 0.857 at 3rd year, 0.869 at 4th year, and 0.834 at 5th year. The mean AUC of test set was 0.945 at 1st year, 0.861 at 2nd year, 0.834 at 3rd year, 0.825 at 4th year, and 0.805 at 5th year. The c-index of the final model was 0.810 in the test group. In addition, the precision of SRN model was evaluated for classifying TNBC patients in test group. TNBC patients were classified into subgroups by SRN-suggested survival probability. Especially in pathologic stage 3 patients, SRN guided subgroups A, B and C were more precisely classified (*P value <0.0001*) than conventional stage IIIA, IIIB and IIIC (*P value =0.292*).

**Conclusion:** Our SRN model provides a useful tool for survival prediction of BC patients and also divides subgroup according to survival in TNBC patients.

**Keywords: Breast Neoplasm; Machine Learning; Survival**

OPEN ACCESS

***Correspondence:***
*Sung Wook Seo, Department of Orthopedic Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, 81 Irwon-ro, Gangnam-gu, Seoul 06351, Republic of Korea, Tel: 82-2-3410-1229; Fax: 82-2-3410-1754; E-mail: sungwseo@skku.edu*

*Jeong Eon Lee, Department of Surgery, Division of Breast Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, 81 Irwon-ro, Gangnam-gu, Seoul 06351, Republic of Korea, Tel: 82-2-3410-3479; Fax: 82-2-3410-6982; E-mail: jeongeon.lee@samsung.com*

**Received Date**: *24 Jan 2019*
**Accepted Date:** *22 Feb 2019*
**Published Date:** *26 Feb 2019*

***Citation:***
*Kim I, Wook Seo S, Eon Lee J, Jun Choi H, Min Ryu J, Kyung Lee S, et al. Deep Survival Model Identifies Prognostic Subgroups of Triple-Negative Breast Cancer Patients. Clin Surg. 2019; 4: 2344.*



## Introduction

Breast Cancer (BC) is the second leading cause of cancer-related death in females [1]. Among BCs, Triple-Negative Breast Cancers (TNBCs) are characterized by the lack of expression of Estrogen Receptor (ER), Progesterone Receptor (PR), and Human Epidermal Growth Factor Receptor 2 (HER-2) [2]. TNBCs comprise 15% to 20% of breast cancers and carry with poor prognosis [3] despite its higher chemo-sensitivity compared to hormone receptor–positive BC [4]. Various prediction models for BC prognosis have been developed and utilized [5,6]. However, prediction model using machine learning is uncommon and suitable prediction model for TNBC is also hard to find.

Machine learning technology is an emerging field of medical data analysis because of its ability of transforming complex data into knowledge [7]. As Obermeyer et al. [8] mentioned in the New England Journal of Medicine that "machine learning will become an indispensable tool for clinicians", machine learning has been applied to many clinical field such as diagnosis and prognosis modeling [8].

In the past, medical information was limited to a few expert clinicians, thus medical decision used to be dependent on expert's opinion. In 21st century, medical data became digitalized and available to clinical researchers, paradigm of medical decision was shifted to data driven medicine, which is good marriage with machine learning technology [9].

Deep learning algorithm is a part of machine learning technology composed of multi-layered neural network and it outperformed other algorithm for pattern recognition and classification in the recent decade [10].

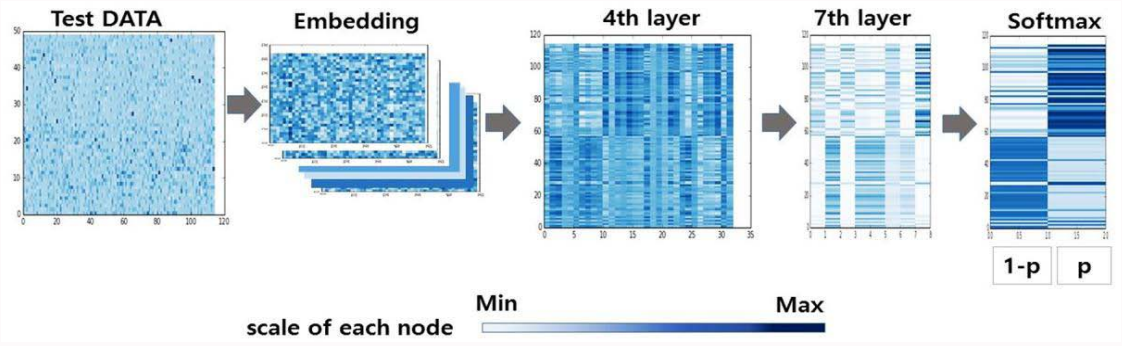


**Figure 1:** Schematic structure of Survival Recurrent Neural Network (SRN) model. The network architecture of the basic unit comprised eight hidden layers with a single RNN layer (LSTM). The input layer comprised 35 nodes that represented 33 input features and two survival features. The output layer comprised two nodes implementing the Soft max function, which represents live/dead probability. The individual features were embedded in a 35 × 32 vector. The other four layers comprised fully connected nodes implementing the rectified linear-unit function. The number of nodes was gradually reduced across the hidden layers.

Recent studies attempted to generate deep learning model for predicting survival of cancer patients. Lee et al. [11] developed Survival Recurrent Network (SRN) which predict prognosis of patients suffering from gastric cancer with high accuracy [11]. SRN learned from surgical findings also showed better performance in classifying prognostic subgroup compared to conventional staging system [12].

In this study, we hypothesized that SRN may be equally capable of identifying independent prognostic subgroups in BC patients and may perform better to identify prognostic subgroup of TNBCs.

## Material and Methods

### Study population

We performed a retrospective chart review of 16,774 BC patients who underwent surgery between May 1996 and January 2015 at Samsung Medical Center (SMC). All electronic medical records and pathology reports were reviewed. A total of 520 patients were excluded from analysis due to incomplete medical records. Of the remaining 16,254 patients, 7,915 with pathologic stage greater than 1 were included in the analysis. The pathologic stage was according to American Joint Committee on Cancer (AJCC) 7$^{th}$ edition. Their median follow-up period was 50.8 months. This study was approved by the Institutional Review Board of SMC in Seoul, Korea (IRB number: 2016-09-075). It adhered to the tenets of the Declaration of Helsinki.

### Baseline characteristics of patients analyzed

Clinicopathologic characteristics of these 7,915 patients analyzed in this study are summarized in Table 1. Their median age was 48.1 years. Most of these patients were women. Most of these patients had invasive ductal carcinoma, luminal A-like subtype, T2, N1, and stage 2 diseases. More than half of these patients were treated by radiation therapy. Chemotherapy and hormonal therapy were also performed for 90.6% and 71.2% of patients, respectively. Local recurrence and distant metastasis occurred in 8.1% and 12.6% of patients, respectively. Among all patients, 11.6% had died by the end of 5-year follow-up.

### Data preprocessing

Variables evaluated in this study were as follows. Categorical variables included sex, type of surgery, contra-laterality, histopathologic subtype, ER, PR, HER2 receptor status, and therapy history (radiation, chemotherapy, and hormonal therapy). Discrete ordinal variables included Histologic Grade (HG) and TNM stage converted by one hot encoding process. Continuous variables including age and Body Mass Index (BMI) were preprocessed with normalization. All missing data were imputed by the k-nearest neighbor algorithm. Event cases in a time window were ranked by month and rank scores were inserted for event cases. Nonparametric rank scores (0<r<1) were calculated as follows.

Let $S$ be an $N \times N$ score matrix. Its element $S_{ij}$, $S_{ij}$

$$= \begin{cases} 1, & for\ t_i > t_j\ when\ e_j = 0, or\ for\ t_i = t_j\ when\ e_i = 2\ and\ e_j = 0 \\ 0, & for\ t_i < t_j\ when\ e_j = 0, or\ for\ t_i = t_j\ when\ e_i = 2\ and\ e_j = 1 \\ -1, & for\ t_i < t_j\ when\ e_i = 2,\ for\ t_i = t_j\ when\ e_i = 0\ and\ e_j = 0, or\ e_i = e_j = 0 \end{cases}$$

where i,j€{1, ..., $N$}, $t_i$ is the follow-up months of each subject or patient and $e_i$ is its corresponding event (1 for survived, 0 for died, and 2 for censored). Score of each subject/patient was then rescaled using min and max values.

### Data separation and cross-validation

Subjects were sorted randomly into training set (80%, $n$=6288) and test set (20%, $n$=1627). The test set was separated for final test. Five-fold cross validation was performed to fine tune an optimal condition of neural networks.

### Performance evaluation and statistics

The performance of the SRN was evaluated using Receiver Operating Characteristics (ROC) curves, Areas under ROC Curve (AUCs), and concordance index (c-index). The AUCs were compared using a non-parametric Mann-Whitney $U$ test (MedCalc Software, Seoul, Korea). The algorithms were programed using Keras and Theano library in Python (https://keras.io/). Data preprocessing was processes Scikit-learn library was used for other data management and (http://scikit-learn.org/). Pearson's Chi-square test was performed to find related factors using SAS 9.4 (SAS Institute, Cary, NC, USA). This study was developed according to Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis model development guidelines [13].

### Basic concept of survival recurrent neural network (SRN) model

SRN algorithm has been well described in other studies and showed excellent performance for predicting prognosis of gastric cancers [11,12]. Briefly, SRN model was constructed based on logistic regression at a discrete time window (t).

$$Survival\ Probability\ at\ time\ (t):\ ft(x) = \sigma(wt * Xt) = \frac{1}{1 + e^{-\omega t Xt}}$$

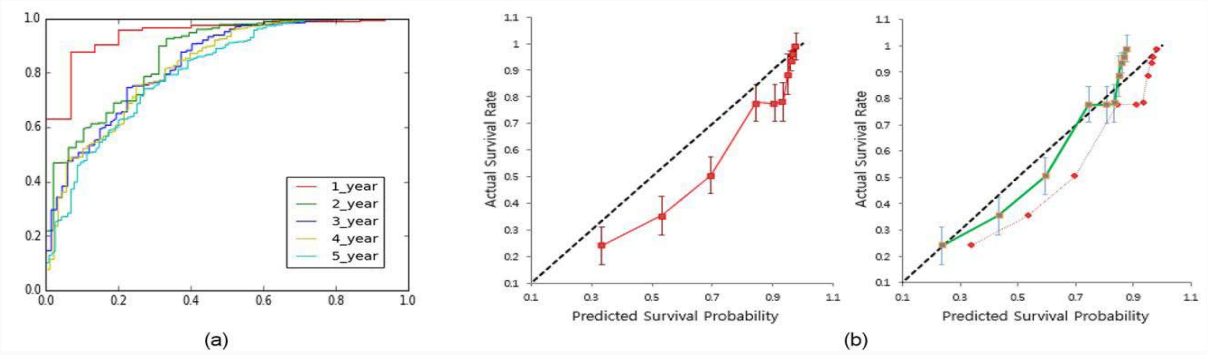


**Figure 2:** The ROC curves and the calibration plot of the SRN for predicting survival of a sample test group. (a) The ROC curves and the mean AUC of the final model was evaluated using the test group at each 5-time point. (0.945 at 1st year, 0.861 at 2nd year, 0.834 at 3rd year, 0.825 at 4th year, and 0.805 at 5th year) (b) Calibration curve of the SRN in the test dataset. The X-axis represents the probability of 5-year survival, and the Y axis represents the actual survival proportion. The Red line represents the calibration plot of the SRN. The actual survival proportion and 95% Cis were calculated using the Kaplan-Meier method. The prediction scores were within a 15% margin of the perfect-prediction.

$$HRt(x) = e^{Wt * Xt}$$

$$HRt(x) = e^{\theta 1x + \theta 2t}$$

Once parameters ($W_t$: $\theta_1 X + \theta_2 t$) were optimized with the patient group (features $X_t$ and target $Y_t$) for the first time point (t), Long Short-Term Memory (LSTM) cells could memorize $W_t$. The model was then retrained with target value ($Y_{t+1}$) of the next time point, yielding $W_{t+1}$. For example, if patient (X) died of disease two years after the first visit, the model should learn the target value ($Y_1$=1) at the first year and the target value ($Y_2$=0) at the second year. Since LSTM can memorize and optimize W for each target value, our RNN-based model is able to infer a target value at a certain time point.

Patient factors (X) were not updated at every time point because we could not collect all data without loss. Moreover, the purpose of the survival model was to predict long-term survival based on the first visit information. Thus, we generated $X_t$ with the assumption that patient features would be constant during the observation time. Instead, there should be latent features dependent on sequential time point that can indicate patient status at a discrete time point. We defined these latent features as time-dependent life values:

$$Time - dependent\ life\ value = \theta 2t + \theta 3 \Delta S$$

$$Time - dependent\ hazard\ function\ e^{WtXt} = e^{\theta 1X + \theta 2t + \theta 3 \Delta S}$$

Life value dimensions were embedded in constant patient features (X) to generate time-dependent features ($X_t$). Life value features consisted of time (t) and prior life expectancy ($S_t$) that was updated using gradient descent equation (ΔS).

The model was retrained at the following sequential time with $X_{t+1}$ where $S_{t+1}$ was embedded.

$$S_{t+1} = S_t + \Delta S$$

$$S_{t+1} = S_t + \alpha (Y_t - \hat{Y}_t) \frac{ft'(x)}{\left| \frac{ft(x)}{e^{-wx}} \right|} * \hat{Y}_t$$

$$\frac{dft(x)}{dx} = -W * \frac{e^{-wx}}{(1+e^{-wx})^2} \approx \alpha * ft(x)(1 - ft(x))$$

where $\hat{Y}_t = ft(x)$

$$S_{t+1} = S_t + \alpha (Y_t - \hat{Y}_t)(1 - \hat{Y}) \hat{Y}_t$$

## Results

### Validation of the SRN model

The SRN was composed of three learning-system parts: 1) information on patient status (covariates, X); 2) time-dependent life values (θ2t+θ3ΔS); and 3) nonparametric rank scores (R) of events that occurred during the interval (0<R<1). R was inserted into the target value instead of binary values. $X_t$ was inputted into the SRN. The SRN was then sequentially retrained with updated $X_n$. Network architecture of the SRN consisted of five hidden layers with two RNN layers [Long Short-Term Memory (LSTM)]. The input layer was composed of 35 nodes that represented 33 input features and two life value features: time and a prior survival probability feature. The output layer comprised of two nodes implementing the Soft max function that represented live/dead probability. A total of 35 features were preprocessed before accessing input layer. Each feature was preprocessed using a standard scaler and each value was encoded as an integer within 10,000 scores. Clinical variables of an individual were embedded in 35 × 32 vector for dimensionality reduction. The other two layers consisted of fully connected nodes implementing rectified linear-unit function. Gaussian dropout was performed to prevent over fitting problem. The number of nodes was gradually reduced across hidden layers (Figure 1).

At each time point, predicted survival probability of the model was compared to actual survival data. Mean AUC of the five-fold validation sets was 0.780 (95% Confidence Interval (CI): 0.646-0.909) at the 1st year, 0.840 (95% CI: 0.784-0.895) at the 2nd year, 0.857 (95% CI: 0.831-0.895) at the 3rd year, 0.869 (95% CI: 0.822-0.916) at the 4th year, and 0.834 (95% CI: 0.784-0.885) at the 5th year.

### Performance of the final model for predicting survival of the test group

Mean AUC of the test set was 0.945 at the 1st year, 0.861 at the 2nd year, 0.834 at the 3rd year, 0.825 at the 4th year and 0.805 at 5th year (Figure 2a). The c-index of the final model was 0.810. A calibration curve was obtained for the test set (Figure 2b). In this curve, SRN-predicted survival probability was compared to the actual survival proportion which was calculated by the Kaplan–Meier method. The intercept indicates that the extent that the calibration curve is systematically too low or too high ('calibration-in-the-large'). It was 0.1 which was within 10% margin of a perfect line. The slope of the calibration curve was 0.95 (95% CI: 0.85–1.08).

### Evaluation of subgroups according to SRN-predicted survival probability in TNBC patients of the test group

We evaluated the precision of the SRN model for classifying TNBC group according to their prognosis. Our SRN model yielded SRN

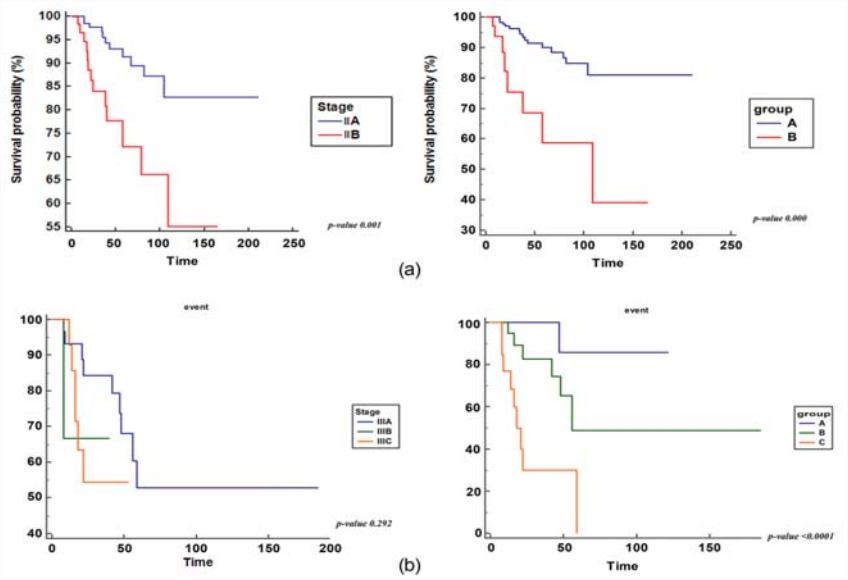


**Figure 3:** (a) Kaplan-Meier survival curve according to anatomic stage IIA and IIB and SRN guided sub group A and B respectively. X axis means time (months) and Y axis means survival probability (%). (b) Kaplan-Meier survival curve according to anatomic stage IIIA, IIIB and IIIC and SRN guided sub group A, B and C respectively.

**Table 1:** Baseline characteristics of patients.

| Characteristics | | No. | (%) |
|---|---|---|---|
| **Sex** | female | 7878 | 99.5 |
| | male | 37 | 0.5 |
| **Age (yr) (Mean ± SD)** | 48.1 ± 2.1 | | |
| **BMI (kg/m$^2$) (Mean ± SD)** | 23.9 ± 4.8 | | |
| **Operation type** | TM+SLNB | 475 | 6 |
| | TM+ALND | 3028 | 38.3 |
| | BCS+SLNB | 1786 | 22.6 |
| | BCS+ALND | 2393 | 30.2 |
| | SSM or NSM + SLNB | 59 | 0.7 |
| | SSM or NSM + ALND | 30 | 0.4 |
| | others | 144 | 1.8 |
| **Contralaterality** | (-) | 7709 | 97.4 |
| | (+) | 206 | 2.6 |
| **Histopathology** | IDC | 6944 | 87.7 |
| | ILC | 340 | 4.3 |
| | Others | 631 | 8 |
| **Histologic Grade** | 1 | 1187 | 15 |
| | 2 | 3326 | 42 |
| | 3 | 2955 | 37.3 |
| | unknown | 447 | 5.6 |
| **Subtype** | Luminal A like | 4302 | 54.4 |
| | Luminal A like | 1105 | 14 |
| | HER2 | 897 | 11.3 |
| | TNBC | 1382 | 17.5 |
| | unknown | 229 | 2.9 |
| **pathologic T stage** | Tis | 27 | 0.3 |
| | T1 | 2015 | 25.5 |
| | T2 | 5081 | 64.2 |
| | T3 | 709 | 9 |
| | T4 | 48 | 0.6 |
| | unknown | 35 | 0.4 |
| **pathologic N stage** | N0 | 2526 | 31.9 |
| | N1 | 3526 | 44.5 |
| | N2 | 1139 | 14.4 |
| | N3 | 698 | 8.8 |
| | unknown | 26 | 0.3 |
| **Distant metastases** | M0 | 7821 | 98.8 |
| | M1 | 92 | 1.2 |
| | unknown | 2 | 0 |
| **Final pathologic stage** | S2 | 5804 | 73.3 |
| | S3 | 2019 | 25.5 |
| | S4 | 92 | 1.2 |
| **Radiation therapy** | undone | 2254 | 28.5 |
| | done | 5443 | 68.8 |
| | unknown | 218 | 2.8 |
| **Chemotherapy** | undone | 600 | 7.6 |
| | done | 7172 | 90.6 |
| | unknown | 143 | 1.8 |
| | Neo adjuvant | 690 | 8.7 |
| | Adjuvant | 6482 | 81.9 |
| **Regimen** | AC based | 5143 | 65 |
| | non AC based | 1566 | 19.8 |
| | unknown | 463 | 5.8 |
| **Hormonal therapy** | undone | 2278 | 28.8 |
| | done | 5637 | 71.2 |
| **Regimen** | Tamoxifen | 3915 | 49.5 |
| | Toremifen | 151 | 1.9 |
| | Aromatase Inhibitor | 1767 | 22.3 |
| **Local recurrence** | (-) | 7275 | 91.9 |
| | (+) | 640 | 8.1 |
| **Distant metastasis** | (-) | 6918 | 87.4 |
| | (+) | 997 | 12.6 |
| **Overall survival** | Survivor | 6996 | 88.4 |
| | Death | 919 | 11.6 |
| **months(Mean, range)** | 50.8 , 1~237 | | |
| **Total** | | **7915** | **100** |

BMI: Body Mass Index; TM: Total Mastectomy; BCS: Breast Conserving Surgery; SLNB: Sentinel Lymph Node Biopsy; ALND: Axillary Lymph Node Dissection; SSM: Skin Sparing Mastectomy; NSM: Nipple Sparing Mastectomy; IDC: Invasive Ductal Carcinoma; ILC: Invasive Lobular Carcinoma; HER2: Human Epidermal Growth Factor Receptor2; TNBC: Triple Negative Breast Cancer; AC: Adriamycin + Cyclophosphamide

**Table 2:** Baseline characteristics of TNBC patients according to SRN guided subgroup.

| Characteristics (*n*=46 patients ) | | subgroup | | | |
|---|---|---|---|---|---|
| | | A | B | C | *P value* |
| **Age (yr)** | **≤ 35** | 4(66.7%) | 1(16.7%) | 1(16.7%) | 0.237 |
| | **35< ≤ 55** | 9(31.0%) | 12(41.4%) | 8(27.6%) | 0.532 |
| | **55<** | 3(27.3%) | 7(63.6%) | 1(9.1%) | 0.337 |
| **Mean ± SD (yr)** | | | | | |
| **Body Mass Index (kg/m²)** | **<25** | 12(34.3%) | 15(42.9%) | 8(22.9%) | 1 |
| | **25 ≤ <30** | 4(40.0%) | 5(50.0%) | 1(10.0%) | 0.727 |
| | **30 ≤** | 0 | 0 | 1(100%) | 0.217 |
| **Mean ± SD (kg/m²)** | | | | | |
| **Operation type** | **TM+SLNB** | 0 | 2(66.7%) | 1(33.3%) | 0.431 |
| | **TM+ALND** | 10(41.7%) | 10(41.7%) | 4(16.7%) | 0.515 |
| | **BCS+SLNB** | 0 | 2(66.7%) | 1(33.3%) | 0.431 |
| | **BCS+ALND** | 6(35.3%) | 8(47.1%) | 3(17.6%) | 0.925 |
| | **SSM or NSM + SLNB** | 0 | 0 | 1(100%) | 0.217 |
| | **SSM or NSM + ALND** | 0 | 0 | 0 | N-C |
| | **others** | 0 | 0 | 1(100%) | 0.217 |
| **Contra-laterality** | **(-)** | 15(34.1%) | 19(43.2%) | 10(22.7%) | 1 |
| | **(+)** | 1(50.0%) | 1(50.0%) | 0 | |
| **Histopathology** | **IDC** | 15(36.6%) | 19(46.3%) | 7(17.1%) | 0.134 |
| | **ILC** | 0 | 0 | 1(100%) | 0.217 |
| | **Others** | 1(25.0%) | 1(25.0%) | 2(50.0%) | 0.415 |
| **Histologic Grade** | **1** | 0 | 1(100%) | 0 | 1 |
| | **2** | 6(66.7%) | 1(11.1%) | 2(22.2%) | 0.041 |
| | **3** | 10(28.6%) | 18(51.4%) | 7(20.0%) | 0.184 |
| **pathologic T stage** | **Tis** | 0 | 0 | 0 | N-C |
| | **T1** | 3(23.1%) | 9(69.2%) | 1(7.7%) | 0.103 |
| | **T2** | 12(57.1%) | 8(38.1%) | 1(4.8%) | 0.005 |
| | **T3** | 0 | 3(33.3%) | 6(66.7%) | 0.001 |
| | **T4** | 1(33.3%) | 0 | 2(66.7%) | 0.092 |
| **pathologic N stage** | **N0** | 0 | 0 | 1(100%) | 0.217 |
| | **N1** | 1(50.0%) | 0 | 1(50.0%) | 0.314 |
| | **N2** | 14(48.3%) | 11(37.9%) | 4(13.8%) | 0.029 |
| | **N3** | 1(7.1%) | 9(64.3%) | 4(28.6%) | 0.029 |
| **Distant metastases** | **M0** | 16(34.8%) | 20(43.5%) | 10(21.7%) | N-C |
| | **M1** | 0 | 0 | 0 | |
| **Radiation therapy** | **undone** | 1(33.3%) | 1(33.3%) | 1(33.3%) | 1 |
| | **done** | 15(34.9%) | 19(44.2%) | 9(20.9%) | |
| **Chemotherapy** | **undone** | 0 | 0 | 0 | N-C |
| | **done** | 16(42.1%) | 20(50.0%) | 10(7.9%) | 0 |
| | **Neo adjuvant** | 0 | 3(23.1%) | 10(76.9%) | 0 |
| | **Adjuvant** | 16(48.5%) | 17(51.5%) | 0 | 0 |
| **Regimen** | **AC based** | 15(39.5%) | 16(42.1%) | 7(18.4%) | 0.321 |
| | **non AC based** | 1(12.5%) | 4(50%) | 3(37.5%) | |

TM: Total Mastectomy; BCS: Breast Conserving Surgery; SLNB: Sentinel Lymph Node Biopsy; ALND: Axillary Lymph Node Dissection; SSM: Skin Sparing Mastectomy; NSM: Nipple Sparing Mastectomy; IDC: Invasive Ductal Carcinoma; ILC: Invasive Lobular Carcinoma; HER2: Human Epidermal Growth Factor Receptor2; TNBC: Triple Negative Breast Cancer; AC: Adriamycin + Cyclophosphamide

guided subgroups according to SRN predicted survival probability. Kaplan-Meier survival curve was generated in each subgroup and compared to those in each anatomical stage.

For stage 2 patients (Figure 3a), Kaplan Meier survival curves were statistically meaningful not only in conventional stage of IIA and IIB, but also in SRN guided subgroups (SRN group A: probability $P<0.5$; SRN group B: $P \geq 0.5$). Meanwhile, SRN guided subgroups were more precisely classified for stage 3 TNBC patients (Figure 3b). There was no significant difference in survival among anatomical stages IIIA, IIIB, and IIIC ($P$ value =0.292). However, cumulative survival rate was decreased significantly in the order of SRN guided subgroup A, B, and C (SRN group A: $0.7<P \leq 1.0$, SRN group B: $0.3<P \leq 0.7$, SRN group C: $P \leq 0.3$) ($P$ value <0.0001).

### Characteristics of stage 3 TNBC patients according to probability group

We analyzed characteristics according to SRN guided subgroups among stage 3 TNBC patients in the test set (Table 2). Age, BMI, operation type, contra-laterality, histopathologic type, and HG were not correlated with difference of subgroups. T2 stage accounted for more than half of group A while T3 or T4 accounted for more than half of group C. In groups A and B, most cases were N2 (14/16 cases and 11/20 cases, respectively). However, N3 was more frequent in group B than that in other groups. In group C, majority of patients were treated with neo adjuvant chemotherapy whereas majority of patients in group A and B were treated with adjuvant chemotherapy.

## Discussion

There are several web-based calculators for predicting survival of BC patients. For example, PREDICT (http://www.predict.nhs.uk) [5] and Breast Cancer Treatment Outcome Calculator (http://www.cancermath.net) [6] are used worldwide. These tools were developed from clinical data (age, TNM stage, ER, HER2, Ki-67 status, therapy options) and analyzed by the Cox hazard model, one of the most popular methods for predicting cancer survival [14]. This model assumes that the hazard function is constant throughout patient lifespan. Therefore, it cannot represent changing weight of covariates during time intervals. We suggest that LSTM, a Recurrent Neural Network (RNN) proposed in 1997 by Sepp Hochreiter et al. [15], represents the optimal choice for a serial learning system.

In our study, SRN predicted survival probability as accurately as anatomical stage grouping for TNBC patients. TNBC is known to be a subtype of breast cancer with poor prognosis. There are many studies on survival of patients with TNBC. In a study of Liedkte et al. [16], Pathologic Complete Remission (PCR) rates after Neo Adjuvant Chemotherapy (NAC) were related to excellent survival while Residual Disease (RD) was related to worse survival [16]. As shown by Gui et al. [17], BRCA1 methylation and Ki67 are also potential prognostic markers of TNBC [17]. In other studies, IGF-I can predict survival [18] and low BCL2 expression is associated with good outcome of TNBC patients treated with both adjuvant and neo adjuvant anthracycline-based chemotherapy [19]. In addition, a higher level of Tumor Infiltrating Lymphocytes (TILs) at diagnosis is significantly associated with decreased distant recurrence rates in primary TNBC [20]. Although there are various factors related to TNBC survival, currently there is no prediction tool for this subtype. Our SRN model provides a useful tool for TNBC cases.

TNBC is a heterogeneous disease [21]. TNBC has been categorized into six subtypes (basal-like 1, basal-like 2, immunomodulatory, mesenchymal, mesenchymal stem-like, and luminal androgen receptor) [22] or Basal-Like Breast Cancer (BLBC) and Quintuple-Negative Breast Cancer (QNBC) according to additional gene expression data for EGFR or CK5/6 [23]. Basal markers are associated with worse survival [24,25]. If we had added CK5/6 and EGFR data to our analysis, the predicted survival rate by SRN could have been shown to be correlated with BLBC or non-basal TNBC.

Our SRN model has advantages in that our model is applicable for all type of breast cancer and we can use accessible clinical data and obtain simple and fast results. Our results showed reliable mean AUC of the test set containing all subtype BC (0.945 at the 1st year, 0.861 at the 2nd year, 0.834 at the 3rd year, 0.825 at the 4th year, and 0.805 at 5th year) and the c-index of the final model (0.810). Our study also supported that SRN model is applicable to TNBC especially. The stage 3 TNBC which was known for poor prognosis could be classified by SRN. In addition, our SRN model algorithm gives optimal results for each center. The findings of a predictive study with large data conducted at one center may not always be suitable for use by other institutions due to differences in race or variable values. Variables such as ER, PR, $HER_2$ or HG were measured according to the official international standard; however, there could be minimally differences among centers. Our predictive model can provide proper results for each center because the results can vary depending on entering data from each center, so any disparity among centers could be diminished.

Our study also has several limitations. First, our study had insufficient numbers of patients, especially in the test set. Initially, we performed SRN analysis including stage 1 patients ($n$=6,220) but it was unsuitable for SRN training due to high survival rate of stage 1 patients. The number of patients was decreased remarkably after excluding stage 1 patients. Second, we did not have data for EGFR, CK 5/6, or Ki-67 known to be important factors in TNBC survival [26,27] because we used data from 1996 whereas these data were available only after 2004. Additional analysis including EGFR, CK 5/6, and Ki-67 data could provide a more accurate prediction. Third, only internal validation was done. Prospective and multicenter validations are required for widespread use of our SRN model.

## Conclusion

In conclusion, our SRN model provides a useful tool for survival prediction of BC patients and also divides subgroup according to survival in TNBC patients. After additional evaluation with large data and external validation, worldwide use of our SRN model could be expected.

## Acknowledgment

This study was presented as abstract for poster session in European Breast Cancer Conference 2018.

This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (2018-0-00861, Intelligent SW Technology Development for Medical Data Analysis).

This study was supported by the Ministry of Education and Science Technology (MEST) (2016R1E1A1A0191433).

## References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin. 2016;66(1):7-30.

2. Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, et al. Triple-negative breast cancer: clinical features and patterns of recurrence. Clin Cancer Res. 2007;13(15 Pt 1):4429-34.

3. Irvin WJ, Carey LA. What is triple-negative breast cancer? Eur J Cancer. 2008;44(18):2799-805.

4. Carey LA, Dees EC, Sawyer L, Gatti L, Moore DT, Collichio F, et al. The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. Clin Cancer Res. 2007;13(8):2329-34.

5. Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence G, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res. 2010;12(1):R1.

6. Michaelson JS, Chen LL, Bush D, Fong A, Smith B, Younger J. Improved web-based calculators for predicting breast carcinoma outcomes. Breast Cancer Res Treat. 2011;128(3):827-35.

7. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood). 2014;33(7):1163-70.

8. Obermeyer Z, Emanuel EJ. Predicting the Future- Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375(13):1216-9.

9. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2006;2:59-77.

10. Arel I, Rose DC, Karnowski TP. Deep Machine Learning-A New Frontier in Artificial Intelligence Research. IEEE Computational Intelligence Magazine. 2010;5(4):13-8.

11. Lee J, An JY, Choi MG, Park SH, Kim ST, Lee JH, et al. Deep Learning–Based Survival Analysis Identified Associations Between Molecular Subtype and Optimal Adjuvant Treatment of Patients with Gastric Cancer. JCO Clinical Cancer Informatics. 2018;2:1-14.

12. Oh SE, Seo SW, Choi MG, Sohn TS, Bae JM, Kim S. Prediction of Overall Survival and Novel Classification of Patients with Gastric Cancer using the Survival Recurrent Network. Ann Surg Oncol. 2018;25(5):1153-9.

13. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMJ. 2015;350:g7594.

14. Bastien P. Cox model in high dimensional and low sample size settings. Proceeding of the Workshop on Knowledge Extraction and Modelling, Anacapri, Italy. 2006.

15. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;9(8):1735-80.

16. Liedtke C, Mazouni C, Hess KR, André F, Tordai A, Mejia JA, et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. J Clin Oncol. 2008;26(8):1275-81.

17. Gui Y, Xu S, Yang X, Gu L, Zhang Z, Luo X, et al. A meta-analysis of biomarkers for the prognosis of triple-negative breast cancer patients. Biomark Med. 2016;10(7):771-90.

18. Bahnassy A, Mohanad M, Ismail MF, Shaarawy S, El-Bastawisy A, Zekri AR. Molecular biomarkers for prediction of response to treatment and survival in triple negative breast cancer patients from Egypt. Exp Mol Pathol. 2015;99(2):303-11.

19. ouchalova K, Kharaishvili G, Bouchal J, Vrbkova J, Megova M, Hlobilkova A. Triple negative breast cancer - BCL2 in prognosis and prediction. Review. Curr Drug Targets. 2014;15(12):1166-75.

20. Loi S, Michiels S, Salgado R, Sirtaine N, Jose V, Fumagalli D, et al. Tumor infiltrating lymphocytes are prognostic in triple negative breast cancer and predictive for trastuzumab benefit in early breast cancer: results from the FinHER trial. Ann Oncol. 2014;25(8):1544-50.

21. Bianchini G, Balko JM, Mayer IA, Sanders ME, Gianni L. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. Nat Rev Clin Oncol. 2016;13(11):674-90.

22. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest. 2011;121(7):2750-67.

23. Hon JD, Singh B, Sahin A, Du G, Wang J, Wang VY, et al. Breast cancer molecular subtypes: from TNBC to QNBC. Am J Cancer Res. 2016;6(9):1864-72.

24. Cheang MC, Voduc D, Bajdik C, Leung S, McKinney S, Chia SK, et al. Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. Clin Cancer Res. 2008;14(5):1368-76.

25. Rakha EA, Elsheikh SE, Aleskandarany MA, Habashi HO, Green AR, Powe DG, et al. Triple-negative breast cancer: distinguishing between basal and nonbasal subtypes. Clin Cancer Res. 2009;15(7):2302-10.

26. Rakha EA, El-Sayed ME, Green AR, Lee AH, Robertson JF, Ellis IO. Prognostic markers in triple-negative breast cancer. Cancer. 2007;109(1):25-32.

27. Gamucci T, Pizzuti L, Sperduti I, Mentuccia L, Vaccaro A, Moscetti L, et al. Neoadjuvant chemotherapy in triple-negative breast cancer: A multicentric retrospective observational study in real-life setting. J Cell Physiol. 2018;233(3):2313-23.