ISSN: 2508-7894 © 2018 KAIA. http://www.kjai.or.kr
Doi: http://dx.doi.org/10.24225/kjai.2018.6.2.11

A prediction of overall survival status by deep belief network using Python® package in breast cancer: a nationwide study from the Korean Breast Cancer Society

Dong-Won Ryu[1]

[1, First Author] Good Moon Hwa Hospital, Korea. E-mail: lovebreast@naver.com

Received: October 29, 2018. Revised: December 02, 2018. Accepted: December 05, 2018.

Abstract

Breast cancer is one of the leading causes of cancer related death among women. So prediction of overall survival status is important into decided in adjuvant treatment. Deep belief network is a kind of artificial intelligence (AI). We intended to construct prediction model by deep belief network using associated clinicopathologic factors. 103881 cases were found in the Korean Breast Cancer Registry. After preprocessing of data, a total of 15733 cases were enrolled in this study. The median follow-up period was 82.4 months. In univariate analysis for overall survival (OS), the patients with advanced AJCC stage showed relatively high HR (HR=1.216 95% CI: 0.011-289.331, p=0.001). Based on results of univariate and multivariate analysis, input variables for learning model included 17 variables associated with overall survival rate. output was presented in one of two states: event or cencored. Individual sensitivity of training set and test set for predicting overall survival status were 89.6% and 91.2% respectively. And specificity of that were 49.4% and 48.9% respectively. So the accuracy of our study for predicting overall survival status was 82.78%. Prediction model based on Deep belief network appears to be effective in predicting overall survival status and, in particular, is expected to be applicable to decide on adjuvant treatment after surgical treatment.

Keywords: Overall Survival Status, Deep Belief Network, Python®, Breast Cancer, Breast Cancer Society.

1. Introduction

Breast cancer is one of the most commonly diagnosed cancer in women. and the one of the leading causes of cancer related death among women (Lee et al., 2017; Park et al., 2017). Despite progress in the treatment of breast cancer, 20–30% of the patients will develop a distant recurrence (Park et al., 2017; Chang et al., 2016). Once distant recurrence has occurred the disease remains largely incurable and median survival of patients with metastatic breast cancer ranges from 2 to 3 years (Moon et al., 2013; Yoo et al., 2017). So the prediction of overall survival status is important in decision on continuity of adjuvant treatment (Park, Chae, Song, Jung, Han, & Nam, 2014). Several prognostic factors were known that associated with overall survival rate such as AJCC stage, estrogen receptor status (ER), progesteron receptor status (PgR), human epithelial receptor type-2(HER-2), Ki-67 and others. But the degree of influence in overall survival status among the factors were different in research (Yoo, Chae, Kim, Lee, Yoon, & Lee, 2017). However that were calculated automatically in deep belief network (DBN) based on machine learning (ML) (Lee et al., 2017; Kim et al., 2017). So Out study constructed the prediction model on overall survival status based on DBN using prognostic factors.

2. Method

The flowchart of this study is showed in Fig. 1. According to exclusion criteria, data was selected and cleaned. And then that was analyzed by statistics package to select variables to be used for deep belief network. The dataset for DBN were composed of factors associated with overall survival rate statistical importantly. The dataset for DBN was divided into a training set (70%) and a testing set (30%) (Shao, Jiang, Wang, & Wang, 2017). Learning procedure was done based on the DBN using the training set. Lastly the accuracy of the prediction on overall survival status during follow up period was measured by DBN (Kim, Kang, & Lee, 2017).

2.1. Dataset

The KBCR is a prospectively maintained, web-based database of the Korean Breast Cancer Society (Yoo et al., 2017; Ryu et al., 2017). Almost of them attended in KBCS database were associated with surgical treatment of breast cancer and worked in hospitals listed on 100 teaching hospitals throughout the Republic of Korea (Yoo et al., 2017; Ahn et al., 2011). About 65% of patients newly diagnosed as breast cancer in 2013 were estimated to included in this database (Yoo et al., 2017). The registry was composed of essential items and optional factors (Park et al., 2017; Kim et al., 2015). These were included in essential items that patients' sex, age, type of surgical treatment and cancer stage based on the American Joint Committee on Cancer classification (Park et al., 2017; Chang et al., 2016). The others were included in optional factors that status of ER, PgR, HER-2, pre operative imaging findings and treatment modality (Ryu et al., 2017; Gwak et al., 2012). Patients' survival status were obtained from the Korean Central Cancer Registry, Ministry of Health and Welfare, Korea, and were recently updated on December 31, 2014 (Moon, Han, Kim, Kim, Yoon, & Oh, 2013). Input variables for DBN were composed of 17 subunits : tumor size, the number of metastatic axillary lymphnode, pre operative metastasis status, proliferative index of Ki-67 based on post operative pathologic report, invasion of peritumoral lymphatic vessels, the number of tumor, histologic tumor grade, cancer stage based on AJCC, patients' age, types of surgical treatment, types of axillary lymphnode operation, status of estrogen receptor, progesterone receptor, HER-2 and survival months after operation (Park et al., 2017; Park et al., 2014; Kim et al., 2017). Output variable included overall survival status: event and cencored. 103881 experimental records were included in KBCR dataset. 88148 cases of these were excluded from the study because of records without follow up. The final dataset comprised 15733 records. After statistical analysis of these dataset, 2500 records is comprised of group for DBN (Kim et al., 2017).

2.2. Statistical Analysis

Incidence rates were expressed as percent (%) per total individuals. Kaplan–Meier curves with the corresponding results of log rank tests were constructed for overall survival (OS). Univariate and multivariate analyses for OS were conducted with a Cox proportional hazards model to estimate hazard ratios (HR) and 95% confidence intervals (CI). Patients with any missing or unknown data were excluded from analysis by Cox models. OS was defined as the time between date of surgery and date of death from any cause. Data analysis was performed using SPSS Statistics version 24.0 (SPSS Inc., Chicago, USA). The primary endpoint was overall survival, defined as the time from the first diagnosis of breast cancer to death from any cause, which was censored at December 31, 2016. The limit of significance for all tests was $p < 0.05$. This study adhered to the ethical tenets of the Declaration of Helsinki and was approved by the Institutional Review Board of OOO university in OOO, Korea (IRB number: 2017-09-021). The need for informed consent was waived because of the low risk posed by this investigation.

2.3. Deep Belief Network (DBN)

DBN is one of the methods for deep learning, composed of multiple layers of restricted Boltzmann machines (RBM) (Lee et al., 2017; Kim et al., 2017). Python^{R} was used for the DBN in this research. The Keras was used for the DBN library. The RBM, which is based on the Hopfield network, can give a weighted value number to unit (Huang, Dong, Duan, & Liu, 2017). The RBM is shown in figure 4 (Kim et al., 2017). The RBM is composed of a visible unit layer and a hidden unit layer, and its internal connection intensity is 0 (Lee, Jun, Cho, Lee, Kim, & Seo, 2017). As shown in figure 5, The DBN was comprised of several RBM which are connected sequentially (Zhang & An, 2017). Among the RBM layers, the first hidden unit layer was known as the previous visible unit layer (Zhang & An, 2017). Learning process in DBN is started from configuring the each one of first visible layer and hidder layer into a single RBM (Lee et al., 2017). After learning process is finished, the first and second hidden layers are trained via the RBM by giving a new input as a weighted value number of the first hidden layer (Wang, Li, Glicksberg, Israel, Dudley, & Ma'ayan, 2017). As such, learning is sequential up to the last layer (Suzuki, 2017). A supervised learning-based classification technique using the DBN is the back propagation algorithm, which is configured in the uppermost layer in the DBN (Wang et al., 2017; Suzuki, 2017). A classification prediction model using the back propagation-DBN was created for this paper.

3. Results

3.1. Characteristics

The process of patient selection is shown in figure 1. Firstly 103881 cases with breast cancer were found in the Korean Breast Cancer Registry between January 2007 and December 2014 (Park et al., 2014). But cases were excluded that had not results of post operative AJCC stage, ER, PR, HER-2 and survival status during follow up period. So a total of 15733 cases were enrolled in this study (Fig. 2). The mean age was 45 years ranged from 15years old to 114 years old. The most prevalent age group was that from 40 to 49 years old (6226 cases, 39.6% of total cases) followed by the 50 to 59 years old age group (4204 cases, 26.7% of total cases). Among the methods of operative treatment, the breast conserving surgery was the most prevalent (9612 cases, 61.1% of total cases), followed by total mastectomy (6085 cases, 38.7% of total cases) (Table 1). According to pathologic report based on AJCC system, the most cases were diagnosed at stage Ia (31.3% of total cases), followed by stage IIa (28.3%), stage 0 (10.7%), stage IIb (12.5%) and stage III (13.5%). Stage IV only accounted for a small portion of total cases (194cases, 1.2% of total cases). In the result of Immunohistochemical (IHC) stains for breast cancer subtype, the proportions of cases that were positive for ER and PgR were 70.0% and 61.5% respectively. When HER-2 expression was identified, 41.4% of all patents had negative IHC staining in the tumor, and 21.4%, 17.5%, and 19.6% of all cases had an IHC stain rating of 1+, 2+, and 3+, respectively (Table 1). The median follow-up period was 82.4 months (range, 1–216 months). In univariate analysis for OS, the patients with advanced AJCC stage showed relatively high HR (HR=1.216 95% CI: 0.011-289.331, *p*=0.001) (Table 2). And it was identified as independent risk factors for overall survival in univariate analysis that younger age, Mastectomy, higher BMI, family history with breast cancer, axillary lymph node dissection, negative for hormonal receptor, negative for HER-2 , Triple negative breast cancer, peritumoral lymphatic invasion, peritumoral vascular invasion, higher Ki-67, mutated p53 , history of adjuvant chemotherapy with regimens containing taxane or adriamycin and no history of adjuvant anti hormonal therapy. Also it was identified as independent risk factors in multivariate analysis that axillary lymph node dissection, advanced T-stage, advanced N-stage, advanced AJCC stage, peritumoral lymphatic invasion, no history of adjuvant anti hormonal therapy (Table 2).

3.2. DBN Model

The DBN was composed of two subunits: training set and testing set (Lee et al., 2017). Training set comprised 70% of the DBN. A learning model was constructed by a training set (Table 3). Input variables for learning model included tumor size, the number of metastatic axillary lymph node, status of distant metastasis, status of Ki-67, status of peritumoral lymphatic invasion, the number of tumor, histologic grade of tumor, pathologic AJCC stage, Age, methods of surgical treatment, methods of axillary operation , status of ER, PR, HER2, the Allred scores of ER, PR and periods of follow up (Fig. 3). In conclusion, output was presented in one of two states: event or cencored. The process of DBN is composed of two steps: forward and back. The epoch, batch size and momentum in first phase were setted at 1500, 50 and 1 respectively (Lee et al., 2017). The first phase is forward propagation (Suzuki, 2017). But the second phase is back propagation in setting of epoch and batch size at 1500 and 50 respectively (Fig. 4).

3.3. Experimental Results

Individual sensitivity of training set and test set for predicting overall survival status were 89.6% and 91.2% respectively. And specificity of that were 49.4% and 48.9% respectively. So the accuracy of our study for predicting overall survival status was 82.78% (Table 4).

4. Discussion

The prediction of recurrence is important in deciding adjuvant treatment of breast cancer (Ryu, Yu, Kim, Kim, Moon, & Choi, 2017). Mostly genomic tests were used to help make decisions about whether more treatments after surgery (Kallenberg, Petersen, Nielsen, Ng, Pengfei, & Igel, 2016). While their names sound similar, genomic testing and genetic testing are time-consuming and very different. DBN known as deep learning has been all over the news lately (Shao et al., 2017). Deep belief network (DBN) is a kind of machine learning that resembles the several layered human learning system (Kim et al., 2017). DBN known as deep learning has recently received attention because of its utilization with big healthcare data (Shao et al., 2017; Suzuki, 2017). DBN is useful for predicting the risk factors associated with disease from the medical big data (Huang et al., 2017). Additionally, DBN is useful for detecting lesion from radiologic images (Lee et al., 2017). DBN is used to predict the status of overall survival using clinicopathologic factors in our study (Li, Giger, Huynh, & Antropova, 2017). The methods of learning are composed of supervised learning and unsupervised learning (Lee et al., 2017). In supervised learning, the output is reproduced by inferring from training data (Suzuki, 2017; Romo-Bucheli et al., 2017). So training data is composed of two parts: the input characteristics and the corresponding output data. In our study, the prognostic factors associated with overall survival were acted as input data (Li et al., 2017; Shen et al., 2017). the corresponding output data were overall survival status: event or cencored. But unsupervised learning does not consider of output data, only infers the hidden structures known as output result from unlabeled input data. The DBN appears to be effective for the risk prediction of recurrent breast cancer and is expected to be particularly applicable to recurrent breast cancer prediction in Koreans. Future research will focus on deep learning research to improve the performance of DBN node optimization and predixction.

References

Lee, J. G., Jun, S., Cho, Y. W., Lee, H., Kim, G.B., & Seo, J. B. (2017). Deep Learning in Medical Imaging: General Overview. *Korean J Radiol, 18*(4), 570-584.

Park, E. H., Min, S. Y., Kim, Z., Yoon, C. S., Jung, K. W., & Nam, S. J. (2017). Basic Facts of Breast Cancer in Korea in 2014: The 10-Year Overall Survival Progress. *J Breast Cancer, 20*(1), 1-11.

Chang, J. S., Choi, J. E., Park, M. H., Jung, S. H., Choi, B. O., & Park, H. S. (2016). Trends in the Application of Postmastectomy Radiotherapy for Breast Cancer With 1 to 3 Positive Axillary Nodes and Tumors </=5 cm in

the Modern Treatment Era: A Retrospective Korean Breast Cancer Society Report. *Medicine (Baltimore), 95*(19), e3592.

Moon, H. G., Han, W., Kim, J. Y., Kim, S. J., Yoon, J. H., & Oh, S. J. (2013). Effect of multiple invasive foci on breast cancer outcomes according to the molecular subtypes: a report from the Korean Breast Cancer Society. *Ann Oncol, 24*(9), 2298-2304.

Yoo, T. K., Chae, B. J., Kim, S. J., Lee, J., Yoon, T. I., & Lee, S. J. (2017). Identifying long-term survivors among metastatic breast cancer patients undergoing primary tumor surgery. *Breast Cancer Res Treat, 165*(1), 109-118.

Park, H. S., Chae, B. J., Song, B. J., Jung, S. S., Han, W., & Nam, S. J. (2014). Effect of axillary lymph node dissection after sentinel lymph node biopsy on overall survival in patients with T1 or T2 node-positive breast cancer: report from the Korean Breast Cancer Society. *Ann Surg Oncol, 21*(4), 1231-1236.

Kim, J., Kang, U., & Lee, Y. (2017). Statistics and Deep Belief Network-Based Cardiovascular Risk Prediction. *Healthc Inform Res, 23*(3), 169-175.

Shao, H., Jiang, H., Wang, F., & Wang, Y. (2017). Rolling bearing fault diagnosis using adaptive deep belief network with dual-tree complex wavelet packet. *ISA Trans, 69*, 187-201.

Ryu, J. M., Yu, J., Kim, S. I., Kim, K. S., Moon, H. G., & Choi, J.E. (2017). Different prognosis of young breast cancer patients in their 20s and 30s depending on subtype: a nationwide study from the Korean Breast Cancer Society. *Breast Cancer Res Treat, 166*(3), 833-842.

Ahn, S. H., Kim, H. J., Lee, J. W., Gong, G. Y., Noh, D. Y., & Yang, J.H. (2011). Lymph node ratio and pN staging in patients with node-positive breast cancer: a report from the Korean breast cancer society. *Breast Cancer Res Treat, 130*(2), 507-515.

Kim, K. S., Kim, Z., Shim. E. J., Kim, N. H., Jung, S. Y., & Kim, J. (2015). The reality in the follow-up of breast cancer survivors: survey of Korean Breast Cancer Society. *Ann Surg Treat Res, 88*(3), 133-139.

Gwak, G., Lee, H. K., Kim, H. J., Lee, S. Y., Park, Y. L., & Lee, J. W. (2012). Survey of the application of the korean clinical practice recommendations on breast cancer treatment: the utility of the korean breast cancer society guidelines. *J Breast Cancer, 15*(2), 239-243.

Huang, Z., Dong, W., Duan. H., & Liu, J. (2017). A regularized deep learning approach for clinical risk prediction of acute coronary syndrome using electronic health records. *IEEE Trans Biomed Eng 2017, 65*(5), 956-968.

Zhang, Y., & An, M. (2017). Deep Learning- and Transfer Learning-Based Super Resolution Reconstruction from Single Medical Image. *J Healthc Eng,* 5859727.

Wang, Z., Li, L., Glicksberg, B. S., Israel, A., Dudley, J. T., & Ma'ayan, A. (2017) Predicting age by mining electronic medical records with deep learning characterizes differences between chronological and physiological age. *J Biomed Inform, 76,* 59-68.

Suzuki, K. (2017). Overview of deep learning in medical imaging. *Radiol Phys Technol, 10*(3), 257-273.

Kallenberg, M., Petersen, K., Nielsen, M., Ng, A. Y., Pengfei, D., & Igel, C. (2016). Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring. *IEEE Trans Med Imaging, 35*(5), 1322-1331.

Li, H., Giger, M. L., Huynh, B. Q., & Antropova, N. O. (2017). Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms. *J Med Imaging (Bellingham), 4*(4), 041304.

Romo-Bucheli, D., Janowczyk, A., Gilmore, H., Romero, E., & Madabhushi, A. (2017). A deep learning based strategy for identifying and associating mitotic activity with gene expression derived risk categories in estrogen receptor positive breast cancers. *Cytometry A, 91*(6), 566-573.

Shen, D., Wu, G., & Suk, H. I. (2017). Deep Learning in Medical Image Analysis. *Annu Rev Biomed Eng, 19*, 221-248.