Development of Breast Cancer Survival Prediction Models Based on Real-world Data and Machine Learning

Chun-Jung Wang[1], Min-Huei Hsu[2], Ruo-Kai Lin[3], Chin-Sheng Hung[4,5], Nei-Hui Kuo[6], Yu-Wen Cheng[1,7], Phung-Anh Nguyen[8], Phan Thanh Phuc[9], Chi-Tsun Cheng[10], Jason C. Hsu[11,12]*


Chun-Jung Wang


Jason C. Hsu


Min-Huei Hsu

1. School of Pharmacy, Taipei Medical University, Taipei, Taiwan.
2. Graduate Institute of Data Science, College of Management, Taipei Medical University, Taipei, Taiwan.
3. Graduate Institute of Pharmacognosy, College of Pharmacy, Taipei Medical University, Taipei, Taiwan.
4. Department of Surgery, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan.
5. Division of General Surgery, Department of Surgery, Taipei Medical University Hospital, Taipei, Taiwan.
6. Cancer Center Division Head, Taipei Medical University Hospital, Taipei, Taiwan
7. Ph.D. Program in Drug Discovery and Development Industry, College of Pharmacy, Taipei Medical University, Taipei, Taiwan.
8. Clinical Data Center, Office of Data Science, Taipei Medical University, Taipei, Taiwan
9. International Ph.D. Program in Biotech and Healthcare Management, College of Management, Taipei Medical University, Taipei, Taiwan
10. Research Center of Health Care Industry Data Science, College of Management, Taipei Medical University, Taipei, Taiwan
11. International PhD Program in Biotech and Healthcare Management, College of Management, Taipei Medical University, Taipei, Taiwan.
12. Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei Medical University, Taipei, Taiwan.

Background

Background

Breast cancer is the cancer with the highest incidence and mortality among women in most countries[1]. There are approximately 2.3 million newly diagnosed cases worldwide each year, and approximately 680,000 deaths annually. With the rise of artificial intelligence, in the past, many researchers used various clinical big data and machine learning algorithms to establish prediction models for breast cancer diagnosis[2] and prognosis[3], respectively, to assist medical decision-making and improve treatment outcomes. However, the parameters and accuracy of such prediction models may vary due to differences in race, geographic location, or other ethnic or individual factors, so it is necessary to use various data sources to develop various prediction models.

Objectives

This study aims to use clinical real-world data with multiple attributes and multiple machine learning algorithms to determine the key factors that affect overall survival when the patient be diagnosed with breast cancer and establish a prediction model which can act as a supporting decision aid for physician by modifying the magnitude of treatment.

Methods

Taipei Medical University Clinical Research Database (TMUCRD) was the data source of this study, which contains the electronic medical records of three hospitals in Taiwan, including Taipei Medical University Hospital (TMUH), Wan-Fang Hospital (WFH), and Shuang-Ho Hospital (SHH). All the data was mapped to OHDSI OMOP CDM. We selected breast cancer female patients whose ICD-O-3 code was C50.0-C50.9 from 2000 to 2019 as the study cohort, and non-primary breast cancer cases or cases with insufficient information on personal medical background and treatment were excluded. Neither do the patient whose follow-up period less than one year. Patients from TMUH and WFH were the training dataset, and patients from SHH were used for external testing. The percentage of alive and death is around 87% and 13% in training data, and 84% and 16% in external testing data, respectively. The date of diagnosis of breast cancer for each patient was used as the index date, and death within five years after diagnosis was used as the outcome. All the information could be gathered on the index data. Totally, there were 45 features involved, including the patient's basic demographic information, cancer condition, comorbidity, current medication, laboratory test result, were selected and used in the model generation. The comorbidities which occurred prior to breast cancer diagnosis were collected. And the lab values which recording within one year from diagnostic date were remained. Missing value of categorical data were classified as a new group. As for continuous data, were imputed with mean value. If the percentage of missing value more than 75%, the feature will be removed. The excluded features are BUN, CA153, CEA, creatine kinase, Ki67, AIDS/HIV, clinical differentiation and lymph vessels or vascular invasion. Finally, there are 37 features be retained. Machine learning algorithms such as logistic regression (LR), support vector machine (SVM), decision tree (DT), gradient boosting (GB), random forest (RF), and artificial neural network (ANN) were applied to build prediction modules. Based on the external test results, the model with the largest area under the receiver operating characteristic curve (AUC) is the best model.

Results

A total of 5,503 patients were included (4,071 for the training dataset and 1,432 for the testing dataset). Based on the external test results, neural network model had the highest AUC (0.880), following by GB (AUC=0.864), RF (AUC=0.863), SVM (AUC=0.804), LR (AUC=0.798) and DT (AUC=0.675). The accuracy of all models is above 85%. In addition, this study also found that, according to the results of the best model (NN), tumor clinical stage, clinical lymph node stage, primary site, Charlson-Deyo Comorbidity Index, and dementia played the most important role in predicting the five-year survival of breast cancer.

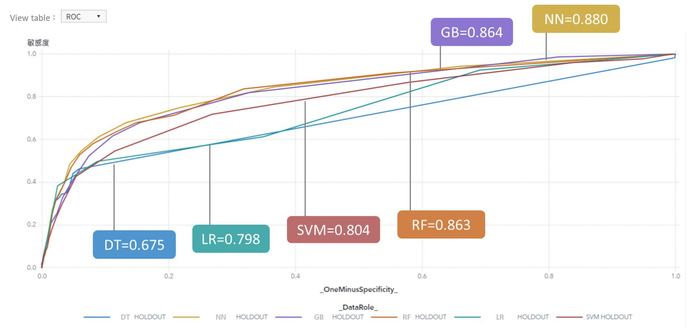

Figure1. ROC Curves of five years survival prediction model when diagnosed with breast cancer.

Model	AUC	Accuracy	Sensitivity	Specificity	Precision	F1-score
Logistic Regression	0.798	0.881	0.383	0.975	0.744	0.506
SVM	0.804	0.862	0.300	0.968	0.636	0.407
Decision Tree	0.675	0.865	0.348	0.963	0.637	0.450
Gradient Boosting	0.864	0.854	0.123	0.992	0.737	0.211
Random Forest	0.863	0.873	0.344	0.973	0.703	0.462
Neural Network	0.880	0.871	0.273	0.983	0.756	0.401

Table1. Performance of Survival Prediction Models.

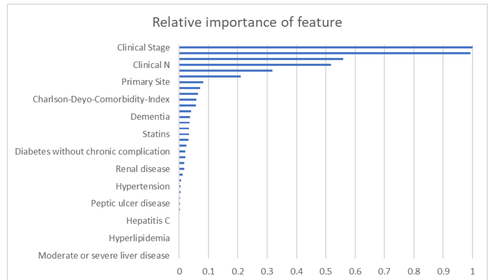

Figure2. Relative importance of feature in neural network model.

Conclusions

This study successfully established an accurate 5-year survival predictive model for breast cancer patients. Furthermore, this study also found many key factors that may affect the survival of breast cancer patients in Taiwanese patients. The results of the study can be used as a reference for clinical practice of breast treatment.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. 2021;71(3):209-249.
2. He T, Puppala M, Ezeana CF, et al. A Deep Learning–Based Decision Support Tool for Precision Risk Assessment of Breast Cancer. JCO Clin Cancer Inform. 2019(3):1-12.
3. Hernandez RK, Wade SW, Reich A, Pirolli M, Liede A, Lyman GH. Incidence of bone metastases in patients with solid tumors: analysis of oncology electronic medical records in the United States. BMC Cancer. 2018;18(1):44-44.