Breast Cancer Survival Analysis with Molecular Subtypes : An Initial Step

1st Lingli Zhang
College of Media Engineering
Communication University of Zhejiang
Hangzhou, China
0000-0001-7406-3600

2nd Jiajun Wu
College of Media Engineering
Communication University of Zhejiang
Hangzhou, China
0000-0002-1870-5020

3rd Youbing Zhao *
College of Media Engineering
Communication University of Zhejiang
Hangzhou, China
0000-0002-8619-084X

4th Wenxian Hu *
Division of Surgical Oncology
Run Run Shaw Hospital, Zhejiang University
Hangzhou, China
3309020@zju.edu.cn

5th Aihong Qin
College of Media Engineering
Communication University of Zhejiang
Hangzhou, China
qinaih@cuz.edu.cn

6th Feng Dong
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, UK
feng.dong@strath.ac.uk

7th Enjie Liu
School of Computer Science and Technology
University of Bedfordshire
Luton, UK
enjie.liu@beds.ac.uk

8th Hao Zeng
College of Media Engineering
Communication University of Zhejiang
Hangzhou, China
hao.zeng@cuz.edu.cn

9th Hao Xie
College of Media Engineering
Communication University of Zhejiang
Hangzhou, China
xiehao@cuz.edu.cn

10th Hui Du
College of Media Engineering
Communication University of Zhejiang
Hangzhou, China
duhui@cuz.edu.cn

***Abstract*—As a predominant threat to women's health worldwide, breast cancer has become increasingly important in oncology research. The discovery of molecular subtypes of breast cancer has led to more subtype oriented treatment and prognosis prediction. Effective prognosis models help to estimate the recurrence as well as the quality and duration of survival, leading to more personalized treatments. However, most traditional prognostic models either ignore molecular subtypes or only make limited use of them. The roles of molecular subtypes in the development and treatment of breast cancer have not been fully revealed. With the over 1200 cases collected by Sir Run Run Shaw Hospital of Zhejiang University in the past two decades, we aim to improve understanding of molecular subtypes and their impacts on the prognosis via data analysis in the long run. As the initial stage, this short paper presents our preliminary work of logistic regression experiments with the data. Though molecular subtypes have not been included the tentative model, they are to be explored further in following stages.**

***Index Terms*—breast cancer, prognostic models, survival analysis, logistic regression, molecular subtypes**

This paper is supported by National Student Innovation Training Program "Survival analysis and prediction of breast cancer based on molecular subtypes" of Communication University of Zhejiang as well as Zhejiang Provincial Public Welfare Fund (LGF21F020004, LGF21F020002, LGF22F020015). (Corresponding authors: Youbing Zhao, Wenxian Hu)

I. Introduction

Breast cancer has long been the most prevalent cancer threatening women's health worldwide [1] and women suffering from breast cancer has kept increasing worldwide. In 2020, the incidence of breast cancer in women has reached 2.26 million globally, not only accounting for 24.5% of new cancer diagnoses in women and about 11.7% of all new cancer diagnoses, but also making it surpassing lung cancer as the most prevalent cancer worldwide for the first time [2], [3].

In China, the economic and social developments has led to improvements of living standards and more westernized lifestyles. It can be observed that the incidence of breast cancer in China follows a similar rising pattern. In 2020, breast cancer is in the first place among new cancer patients in women and ranks fourth among all cancers for the whole Chinese population.

The prognosis of breast cancer is a direct predictor of the quality and duration of survival. Highly correlated with the prospect of recurrence and the duration of survival, the molecular subtype [4], [5] has been adopted in the standard process of breast cancer diagnosis and treatment. However, the relationship between the molecular subtype and the prognosis

has not been thoroughly studied. More detailed study of the role of molecular subtypes in prediction of prognosis for breast cancer is desired as most traditional prognostic models either ignore molecular subtypes or only make limited use of them.

Sir Run Run Shaw Hospital of Zhejiang University is a top-ranked 3A hospital located in Hangzhou China. In the past two decades the hospital has collected over 1200 cases of breast cancer patients with follow-ups.

The oncologists in Sir Run Run Shaw Hospital are strongly interested in using the collected data to search for possible hidden relationships between molecular subtypes and prognosis. Consequently the long term aim of our study is to gain better understanding of molecular subtypes and their impacts on the prognosis via analysis of the dataset. In the initial stage we are trying to build enhanced prognostic models through logistic regression and survival analysis with our dataset.

This short paper presents our initial investigation of logistic regressions with the dataset. First, Kendall analysis is employed to discover the main variables affecting survival, aiming at reducing data dimensionality and simplifying the temporal complexity of model training. Secondly, the optimal subset selection with AIC is used to choose the optimal logistic regression model. Finally a K-fold cross-validation is performed to evaluate the performance of the model. Preliminary results show that the tentative model can be employed to provide predictions with high accuracy.

Though in the experiments the molecular subtype has not been chosen into the tentative model, they are to be explored further in following stages so that we can gain more knowledge of the role of molecular subtypes and their impacts on the survival of breast cancer.

II. Related Work

Long been the most threatening cancer to women worldwide [1], breast cancer has also become the most prevalent cancer worldwide from 2020 [2], [3]. In China, breast cancer is the No. 1 cancer affecting women and ranks fourth in all cancers affecting the whole Chinese population.

Effective prognosis models help to estimate the quality and duration of survival, enabling more personalized treatments. However, breast cancer is a highly heterogeneous disease, the widely adopted TNM staging of cancers without genetic information is not capable enough to reflect the innate difference of breast cancer patients.

Though the diversity of breast cancer and the connection between oophorectomy and the remission of breast cancer have been observed a long time ago, it is not until in the late 1950s that Jensen et al. in their pioneering work [6], [7] discovered estrogen-receptor (ER) and found its overexpression was related to some breast cancers. About one third of breast cancer patients are ER+ and are more responsive to treatments while ER- patients usually have unoptimistic prognosis. Oncologists discovered later more related hormone receptors and group them into different molecular subtypes. From two decades ago, incorporating related tumor genotypes such as the human epidermal growth factor receptor 2 (HER2), Perou et al. [4], [5] further classified molecular subtypes into 4 major categories: Luminal A(ER+/HER-), Luminal B(ER+/HER-), HER2+, basal-like(triple-negative) [ER-/HER2-], which had been adopted by the breast cancer treatment consensus [8]. Molecular subtypes have become indispensable in guiding clinical treatment, assessing efficacy and predicting prognosis of breast cancer [9]. The study of roles of molecular subtypes in breast cancer treatment and prediction has become an attractive field in breast cancer research.

However, despite substantial efforts made in studies of molecular subtype related prognosis analysis [10]–[15], most of popular prognostic models of breast cancer either ignores molecular subtypes (such as NPI [16]) or only makes limited use of them (such as Predict [17]). Phung et al [18] present a detailed survey of prognostic models for breast cancer.

This short paper reports our preliminary results of prognostic modeling based on 1200 breast cancer cases as the initial attempt to discover more underlying mechanisms related to molecular subtypes and to establish more effective predictive models for breast cancer diagnosis and treatment.

III. Dataset and Preprocessing

A. The Dataset

Our breast cancer dataset is a database containing clinical records and follow-ups of 1208 breast cancer patients collected by Run Run Shaw Hospital of Zhejiang University in the past 20 years. First time visits span from 2004 to 2020 and follow-ups span from 2009 to 2020, with an average of five years between the diagnosis and the follow-ups. There are 136 deaths during the follow-up period, 123 deaths from breast cancer-related neoplastic deaths, and 13 deaths from other neoplastic deaths caused by breast cancer.

The dataset contains 78 attributes, including patient information, examination and diagnosis, surgery and postoperative pathology, treatment, follow-ups, recurrence, and metastasis, etc. However, most patients have missing data for certain attributes. Among them, missing entries on surgery and postoperative pathology are fewer.

B. Data Preprocessing

1) Processing of Missing Values: Most of our breast cancer cases suffer from missing values. In the 1208 rows and 78 columns of the case data, 60 columns have missing values, accounting for 5.0% of all data values. As missing values are deleted by default machine processing, they affect the result of analysis. Accordingly missing value preprocessing is required to either remove data records with missing values or fill in the missing entry with estimated values.

A strategy of missing value filling is employed when missing values only occur in a low proportion of a column. We choose to fill in the missing entries with the mean value of the attribute when missing values are within 5%, and with random values when they are within 15% [19]. However, if the missing entries exceed 15% of the total case number in a column, to preserve the accuracy of the data, they should not be filled.

2) Inference of Molecular Subtypes: Since our original data have no field for the molecular subtype, we inferred the molecular subtype according to the criteria defined in [8].

3) Implementation: We use R 4.1.2 to perform missing value filling and molecular type inference for all 1208 rows of the dataset, based on which new data are generated and used as the basis of data analysis in the next section.

IV. Design and Implementation of Breast Cancer Survival Analysis

A. Overall Process

After data preprocessing, a experimental breast cancer survival analysis which comprises of the following steps is performed. First, Kendall correlation coefficient analysis is performed on the data to extract significant variables which are then used for logistic regression in the second step. To choose the optimal regression model, best subset selection based on AIC is used. Finally, a K-fold cross-validation is performed on the selected model to evaluate its accuracy.

The following subsections introduce each of the steps in detail. R is used in the implementation of all steps.

B. Step 1: Selection of significant variables

There are 78 attributes in our data but they are not equally important to the prognosis prediction. Accordingly we first employ the Kendall correlation coefficient [20] to extract attributes significantly correlated with the patient survival status. Variables with T value (correlation coefficient) greater than 0.15 are selected for logistic regression in the second step. Through several experiments, a total of 15 variables are found eligible, i.e. "latest follow-up time", "lump length", "clinical stage T", "clinical stage N", "surgery purpose"(radical/palliative), "surgery type", "lesion length", "lesion width", "number of metastatic lymph nodes", "number of lymph nodes removed", "sentinel lymph node biopsy(SLB)" (Yes/No), "Ki67 expression", "pathological stage T", "pathological stage N", and "molecular subtype". While these variables are highly correlated with the survival status, the highest correlations happen for "clinical stage N", "surgery purpose", "surgery type" and "pathological stage N".

C. Step 2: Logistic regression

Logistic regression is a generalized regression model for dealing with dichotomous and multicategorical data.

Breast cancer survival prediction can be viewed as a dichotomous logistic regression problem represented by the following equation:

$$P = \frac{1}{1 + e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}} \tag{1}$$

where x_i are variables, βi are regression coefficients and P is the binary outcome of survival.

D. Assessment of the "histological grading" variable

The "histological grading" variable is singled out because there are many missing data in this attribute. In order to assess the impact of "histological grading", we made two logistic regression experiments with and without the attribute to decide if it should be included in the logistic regression.

Logistic regression results with "histological grading" show high significance for "histological grading"($p < 0.001$) and "pathological stage N"($p < 0.001$). Logistic regression results without "histological grading" show high significance for "surgery type" ($p < 0.001$), "number of lymph nodes removed", "molecular subtype" ($p < 0.01$), general significance for "number of metastatic lymph nodes" and "Ki67 expression" ($p < 0.05$). The logistic regression model without "histological grading" are more understandable, so we choose to opt out "histological grading" in logistic regression.

E. Step 3: Logistic variable selection with best subset selection and AIC

AIC (Akike Information Criterion), proposed by Hiroji Akike [21] in 1974, is a criterion for measuring the fitness of a statistical model. Based on the concept of entropy, it can weigh the complexity of the estimated model and the fitness of the model to the data. In this paper, AIC is employed to measure the fitness of logistic regression models built with selected variables.

The bestglm() function in R is used to obtain the best subset of logistic regression models with AIC in the following way:

```
bestglm(Xy, family, IC = "AIC")
```

where `Xy` denotes the data frame, `family` denotes the regression distribution, and `IC` denotes the information criterion, here is AIC.

The selected model variables include "latest follow-up time", "lump length ", "lesion width", "surgery purpose", "surgery type", "number of metastatic lymph nodes", "number of lymph nodes removed", "Ki67 expression", "pathological stage N" and "molecular subtype". The inclusion of "molecular subtype" shows that molecular type is one of the important variables affecting prognosis, which is consistent with the related work and physicians' medical experience.

F. Step 4: K-fold cross-validation

K-fold cross-validation is a method for model validation in machine learning. By dividing the data into K subsets and taking turns to use $K - 1$ subsets as training data and one subset as testing data, it can effectively evaluate the performance of a model on a limited dataset. Our experiments with cross-validation show that 5-fold and 10-fold cross-validation have better results. The average ROC curve (1) for 10-fold cross validation shows that the model has a high prediction accuracy of 96.7%.

V. The Result Model

The final logistic regression model we obtain is shown in the following equation:

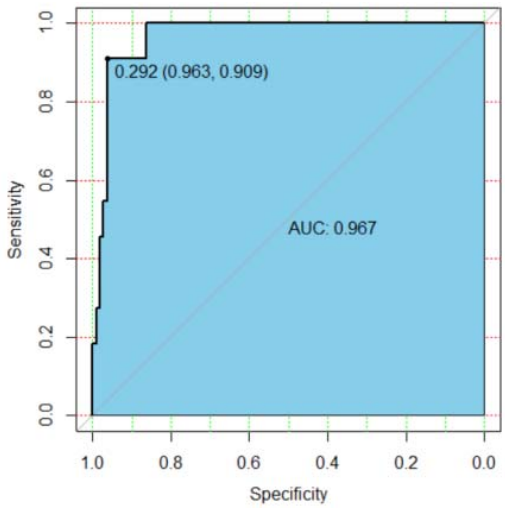

Fig. 1. The ROC Curve

$$ln\frac{p}{1-p} = -7.11 + 2.17 \times x_1 + 0.58 \times x_2 + 0.22 \times x_3 + 0.11 \times x_4 - 0.04 \times x_5 \quad (2)$$

where, x_1 is "surgery purpose", x_2 is "surgery type", x_3 is "lesion width", x_4 is "number of metastatic lymph nodes", and x_5 is "number of lymph nodes removed".

The 10-fold cross-validation shows that the accuracy, specificity, and sensitivity of the model are 96.7%, 96.3%, and 90.9%, respectively.

The preliminary model indicates that "surgery purpose" (radical/palliative) and "surgery type" have high impacts on the prognosis of patient survival. This confirms that early detection and early surgery of breast cancer is of great importance to improve the duration of survival [22]–[24].

VI. Conclusions and Future Work

In this short paper we present our initial work toward study of the roles of molecular subtypes in breast cancer prognosis analysis. Cross-validation results show that the preliminary model built via logistic regression has high credibility and accuracy, which in turn can be used to assist prognosis prediction and treatment of breast cancer.

Though the molecular subtype has been discovered a highly relevant variable in this study, it has not been included in the final preliminary model. Since molecular subtypes of breast cancer have great impacts on the prognosis of breast cancer in medical practice, we are planning to explore further on molecular subtype related aspects in future stages.

References

[1] M. P. Coleman, M. Quaresma, F. Berrino, J. michel Lutz, R. D. Angelis, R. Capocaccia, P. Baili, B. Rachet, G. Gatta, T. Hakulinen, A. Micheli, M. Sant, H. K. Weir, J. M. Elwood, H. Tsukuma, S. Koifman, G. A. E. Silva, S. Francisci, M. Santaquilani, A. Verdecchia, H. H. Storm, and J. L. Young, "Articles cancer survival in fivve continents: a worldwide population-based study (concord)," *The Lancet Oncology*, no. 8, pp. 730–56, Aug 2008.

[2] "World cancer report: Cancer research for cancer prevention," 2020.

[3] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020 globocan estimates of incidence and mortality," *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, May/June 2021.

[4] C. Perou, T. Sørlie, M. Eisen, and et al, "Molecular portraits of human breast tumours," vol. 406, pp. 747–752, Nov. 2000.

[5] C. G. A. Network, "The cancer genome atlas network. comprehensive molecular portraits of human breast tumours," *Nature*, no. 490, p. 61–70, 2012.

[6] E. Jensen, "H. 1. jacobson: Biological activities of steroids in relation to cancer," 1960.

[7] J. EV, J. HI, W. AA, and F. CA, "Estrogen action: a historic perspective on the implications of considering alternative approaches," *Physiol Behav*, vol. 99, no. 2, pp. 151–62, Feb 2010.

[8] A. Goldhirsch, E. Winer, A. Coates, R. Gelber, and et. al., "Personalizing the treatment of women with early breast cancer: highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2013," *Annals of Oncology*, vol. 24, no. 9, pp. 2206–2223, 2013.

[9] C. E. Park YH, Lee SJ and et al, "Clinical relevance of tnm staging system according to breast cancer subtypes," *Ann Oncol.*, vol. 22, no. 7, pp. 1554–1560, Jul 2011, erratum in: Ann Oncol. 2019 Dec 1;30(12).

[10] T. Cooke, J. Reeves, A. Lanigan, and P. Stanton, "Her2 as a prognostic and predictive marker for breast cancer," *Annals of oncology*, vol. 12, pp. S23–S28, 2001.

[11] Z. Yuan, S. Wang, Y. Gao, Z. Su, W. Luo, and Z. Guan, "Analysis of clinical characteristics and prognostic factors of 305 patients with triple-negative breast cancer," *Cancer*, vol. 27, no. 6, pp. 561–565, 2008, in Chinese.

[12] A. Masarwah, P. Auvinen, M. Sudah, V. Dabravolskaite, O. Arponen, A. Sutela, S. Oikari, V.-M. Kosma, and R. Vanninen, "Prognostic contribution of mammographic breast density and her2 overexpression to the nottingham prognostic index in patients with invasive breast cancer," *BMC cancer*, vol. 16, no. 1, pp. 1–9, 2016.

[13] Z. Liu, C. Chen, X. Yao, and S. Sun, "Clinicopathological characteristics and prognostic analysis of different molecular subtypes of breast cancer," *National Medical Journal of China*, vol. 96, no. 22, pp. 1733–1737, 2016, in Chinese.

[14] M. ZHU, "Nomograms to predict early response of neoadjuvant chemotherapy and disease-free survival for triple negative breast cancer," Ph.D. dissertation, Zhejiang University, mar 2021, in Chinese.

[15] Y. J, J. J, G. L, J. B, and et al., "Prediction of her2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning," *Comput Struct Biotechnol J.*, pp. 333–342, 2022, in Chinese.

[16] M. H. Galea, R. W. Blamey, C. Elston, and I. O. Ellis, "The nottingham prognostic index in primary breast cancer," *Breast Cancer Research and Treatment*, vol. 22, pp. 207–219, 1992.

[17] Predict. [Online]. Available: https://breast.predict.nhs.uk/

[18] E. J. Phung MT, Tin Tin S, "Prognostic models for breast cancer: a systematic review," *BMC Cancer*, p. 230, Mar 2019.

[19] J. Han, M. Kamber, and J. Pei, *Data mining concepts and techniques, third edition*. Waltham, Mass.: Morgan Kaufmann Publishers, 2012.

[20] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, Jun 1938.

[21] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[22] D. N. Krag, S. J. Anderson, T. B. Julian, and et. al., "Sentinel-lymph-node resection compared with conventional axillary-lymph-node dissection in clinically node-negative patients with breast cancer: overall survival findings from the nsabp b-32 randomised phase 3 trial," *The Lancet Oncology*, vol. 11, no. 10, pp. 927–933, 2010.

[23] W. AG, G. RL, and e. a. Turner BC, "A 10-year follow-up of treatment outcomes in patients with early stage breast cancer and clinically negative axillary nodes treated with tangential breast irradiation following sentinel lymph node dissection or axillary clearance," *Breast Cancer Res Treat*, vol. 125, p. 893–902, 2011.

[24] M. Pepels, M. Boer, M. Smidt, P. Diest, and G. Borm, "breast cancer: A systematic review," *Medicine*, 2010.