Detection of Genes Associated with the Survival Rate of Breast Cancer Patients

Yuan Wu

B00748226

Supervisor: Dr.Lam Ho

Abstract

In this paper, we study the association between genes and the survival rate of cancer patients by analyzing the Metabric data from the database Cbioportal. We focus on triple-negative patients who have negative status on their expression of ER, HER2, and PR. We used the Cox proportional hazard model and the Glmnet package as two main materials and methods for this analysis. Our result shows that there are 21 genes associated with triple-negative breast cancer patients' survival rates.

1 Introduction

Breast cancer is cancer that forms in the cells of the breasts. This type of cancer can occur in both males and females, but it's far more common in females.

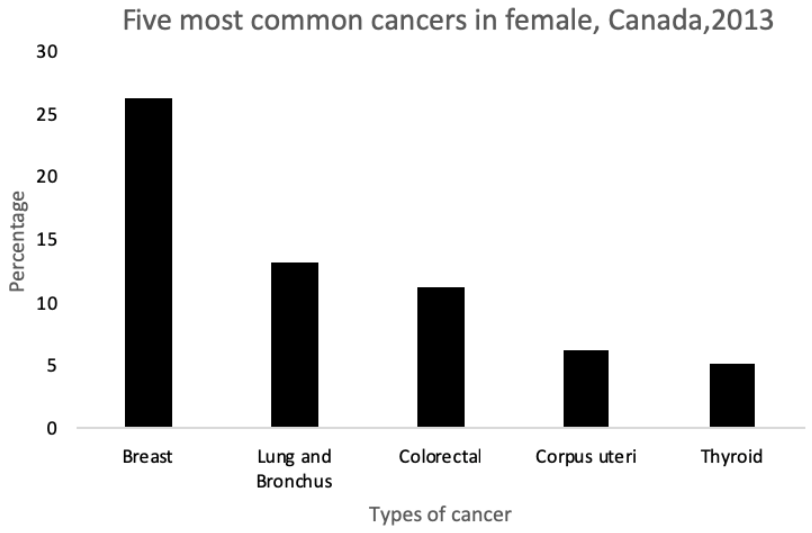

Figure 1: *Five most common cancers in female, Canada, 2013 (Canadian Cancer Registry).*

According to Figure 1 above, it's easy to find that Breast cancer is the most frequently diagnosed cancer among females in Canada. Statistics Canada (2016) indicates that 26.2% of new females' cancer cases in 2013 are breast cancer. Lung and bronchus cancers ranked second (13.2%), especially during the woman's age above 25 years old. There was 33.9% of females have breast cancer with age between 25 to 49 years old. And 25.3% of patients with age above 50 years old. Also, this phenomenon is similar to the USA. "There were about 232,340 new cases of invasive breast cancer and 39620 deaths are expected among US women aged 50 and older in 2013" (DeSantis et al., 2014). Even worldwide, the breast cancer mortality rate is pretty high.

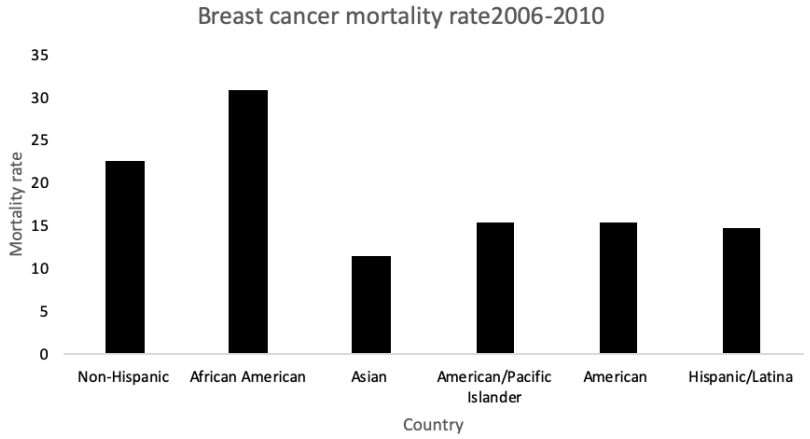

Figure 2: *Breast cancer mortality rate between year 2006 to 2010 (DeSantis et al., 2014).*

According to Figure 2 above, it compares the breast cancer mortality rate between the year 2006 to 2010 in different countries. For African Americans, the mortality rate was higher than in other countries (30.8%), and Non-Hispanic White's average mortality rate was 22.7% between the years 2006 to 2010. For the remaining countries, the average mortality rate caused by breast cancer was around 12%. So, we can know that breast cancer is the most invasive cancer for women age above 25 years old, and it can lead to a high mortality rate worldwide. Breast cancer patients' survival rate is getting a lot of attention and concern around the world.

Many factors could affect the survival rate of breast cancer patients, such as the health status of the patient, age of the patient, and patient's tumor size and grade. However, these factors alone are inadequate for predicting the prognosis of breast cancer patients. In recent years, "lots of researches have reported the relationship between gene expression and cancer prognosis" (Zhang et al., 2018). The goal of this thesis is to detect such genes that

have an association with triple-negative breast cancer prognosis.

Triple-negative breast cancer is a type of cancer with high recurrence and poor survival rates. Triple-negative breast cancer means patients' expression of ER, PR, and HER2 are all negative. ER-negative means "patients lack expression of the estrogen receptor, which is a ligand-activated enhancer protein that is a member of the nuclear receptor superfamily" (Foulkes et al., 2010). ERα is an acknowledged prognostic and predictive factor for established breast cancer Locker et al. (2002). PR negative means patients lack expression of the progesterone receptor, which is "ligand-activated transcription factor members of the steroid hormone receptor subfamily of nuclear receptors" (Daniel et al., 2011). PRs function works as critical regulators of transcription and helps to activate signal transduction pathways. Many of them are involved in pro-proliferative signaling in the breast as well. HER2 negative means patients lack expression of the human epidermal growth factor receptor 2. The HER2 is used to control a protein on the surface of cells, which can help cells grow. If the HER2 gene changes, it may lead to a tumor grow(Canadian Cancer Society, 2020).

2 Material And Methods

2.1 Data

The data for this report comes from the cBio Cancer Genomics Portal, which is an open-access resource that helps researchers get multiple cancer genomics data sets. Currently, there are more than 5000 tumor samples from 20 different types of cancer research are available on the cBio Cancer Genomics Portal

website for researchers. Those huge amounts of data significantly help to reduce barriers between complex genomic data and cancer researchers (Cerami et al., 2012). In this report, we only focus on Breast Cancer, in particular, the Metabric (Molecular Taxonomy of Breast Cancer International Consortium) data set. Metabric is a Canada-UK project which aims for classifying breast tumors into further subcategories based on molecular signatures, which can help researchers determine the optimal treatment for different types of breast cancer (El-Naggar et al., 2020). Also, Metabric is the worldwide most extensive study of breast cancer tissue and the culmination of decades of research into the disease. Metabric project gathered robust clinical date from over 1904 breast cancer samples and 18543 types of genes. This comprehensive genomic data helps us to explore the association between gene expression and survival rate of breast cancer patients. Here, we used Metabric data to identify genes that have a significant association with the survival rate of triple-negative breast cancer patients. We focus on 320 sets of triple-negative breast cancer samples from the Metabric data set for further analysis.

2.2 Method

We used the Cox proportional hazards model and the Glmnet package as two main methods and materials for this research.

2.2.1 Cox Proportional Hazards Model

The Cox proportional hazards model is a type of statistical regression model for modeling the relationship of covariates to survival or other censored outcome in medical research. There are two main parts to compose survival

models, the underlying baseline hazard function, which denote as $\lambda_0(t)$ and the effect parameters. The underlying baseline hazard function explains how the risk is changed for each event when per time unit changed at baseline levels of covariates. The effect parameters explain how the hazard varies are changed when covariates change. The Cox Proportional Hazards Model expresses as:

$$\begin{aligned}\lambda(t|X_i) &= \lambda_0(t)exp(\beta_1 X_{i1} + ... + \beta_p X_{ip}) \\ &= \lambda_0(t)exp(X_i * \beta),\end{aligned}$$

where $\lambda_0(t)$ is the underlying baseline hazard function, and X_i is a vector of possibly time-dependent covariables, and β is a vector of unknown parameters.

Here is the basic idea of Cox partial likelihood: At time Y_i, the likelihood of the event for subject i is:

$$\begin{aligned}L_i(\beta) &= \frac{\lambda(Y_i|X_i)}{\sum_{j:Y_j>=Y_i}\lambda(Y_i|X_j)} \\ &= \frac{\lambda_0(Y_i)\theta_i}{\sum_{j:Y_j>=Y_i}\lambda_0(Y_i)\theta_j} \\ &= \frac{\theta_i}{\sum_{j:Y_j>=Y_i}\theta_j},\end{aligned}$$

where $\theta_j = exp(X_j * \beta)$ and the summation is over the set of subjects j where the event has not occurred before time Y_i. and $0 < L_i(\beta) <= 1$.This is a partial likelihood: the effect of the covariates can be estimated without the need to model the change of the hazard over time.

Cox (1972) indicates that if we assume each subject is independent of others, the following partial likelihood expresses the joint probability of all realized

events. $C_i = 1$ means the occurrence of the events:

$$L(\beta) = \prod_{i:C_i=1} L_i(\beta).$$

The log partial likelihood is:

$$l(\beta) = \sum_{i:C_i=1} (X_i * \beta - log \sum_{j:Y_j>=Y_i} \theta_j).$$

To maximize the estimates of the parameters in the partial likelihood model, we can maximize over β in corresponding log partial likelihood function. The partial score function is:

$$l'(\beta) = \sum_{i:C_i=1} (X_i - \frac{\sum_{j:Y_j>=Y_i} \theta_j X_j}{\sum_{j:Y_j>=Y_i} \theta_j}),$$

the Hessian matrix of the partial log likelihood is:

$$l''(\beta) = \sum_{i:C_i=1} ((\frac{\sum_{j:Y_j>=Y_i} \theta_j X_j X_j'}{\sum_{j:Y_j>=Y_i} \theta_j}) - (\frac{[\sum_{j:Y_j>=Y_i} \theta_j X_j][\sum_{j:Y_j>=Y_i} \theta_j X_j']}{[\sum_{j:Y_j>=Y_i} \theta_j]^2})).$$

We can maximize the partial likelihood by using the Newton-Raphson algorithm when we are using the partial score function and Hessian matrix function, which are showed before. The estimated value of β can be used to evaluate the inverse of the Hessian matrix. Then, to evaluate the regression coefficients' approximated standard errors, the inverse of the Hessian matrix can be used to approximate the variance-covariance matrix of the estimate.

2.2.2 Glmnet Package

Glmnet is a package that fits a generalized linear model via penalized maximum likelihood. This algorithm fits many kinds of regression models, such as

linear regression models, multinomial and logistic regression models, passion regression models, the Cox model, multiple response Gaussian and grouped multinomial regression. This package can be used to do a variety of predictions from the fitted models and fit muti-response linear regression. In this project, we only used the function "cv.glmnet"(cross-validation for Glmnet) to do the prediction of survival rate test. This function is used to determine "does the k-fold cross-validation for the Glmnet" (Friedman et al., 2019). Cross-validation is usually used to determine the optimal values for hyper-parameters in a model by choosing the parameter value that minimizes the cross-validated error. K-fold cross-validation split the data into K parts. $K-1$ of these parts is defined as a new training part, and K^{th} partition is defined as the validation part. This method is using those new training parts to build the model and validate the K^{th}. The cross-validation process will be cycled K times, and the cross-validation error is the average or sum of the k different deviances (Hastie et al., 2009). The formula of k-fold cross-validation can be expressed as below:

For each $k = 1, 2, ...k$, fit the model with parameter λ to the other $K-1$ parts, giving $\hat{\beta}^{-k}(\lambda)$ and compute its errors in predicting the
K^{th} part:

$$E_k(\lambda) = \sum_{i \in k^{th} part} (y_i - x_i \hat{\beta}^{-k}(\lambda))^2.$$

This gives the cross-validation:

$$CV(\lambda) = \frac{1}{K} \sum_{K=1}^{K} E_k(\lambda).$$

We do this for many values of λ and choose the value of λ that makes $CV(\lambda)$smallest.

As I mentioned before, in the Lasso penalized regression analysis, the data set will be randomly allocated into two sets, training set, and test set. The process of "cv.glmnet" is to run the Glmnet n folds plus one time and get the sequence of λ. Then, the rest part will compute the fits with each of the folds omitted and accumulate all errors to compute the average error and standard deviation over the folds. There is one thing to note. Since the data are split randomly, the results for function "cv.glmnet" are random. To reduce the randomness, users can run function "cv.glmnet" more times and average the error curves.

3 Analysis

This section explains how the methods and materials described previously will be used for our research. In this project, we only focus on patients with Triple-negative breast cancer, which means their ER_status, PR_status, and HER2_status need to show all negative. However, the original data set included all types of breast cancer patients. Thus, we first use R to extract the data for triple-negative patients. The 320 sets of samples were identified and allocated to two parts randomly, training part and test part. Then, we applied the LASSO Cox proportional hazards model to build a survival rate model with training part. 21 out of 18543 genes are detected to have association to the survival rate of triple negative breast cancer patients (Figure3&4)

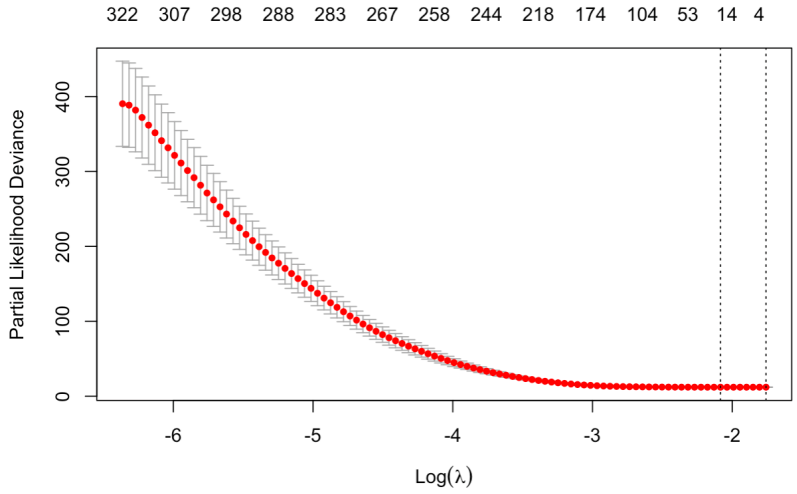

Figure 3: *Values of lambda(best value of λ is 0.1723218 in this analysis).*

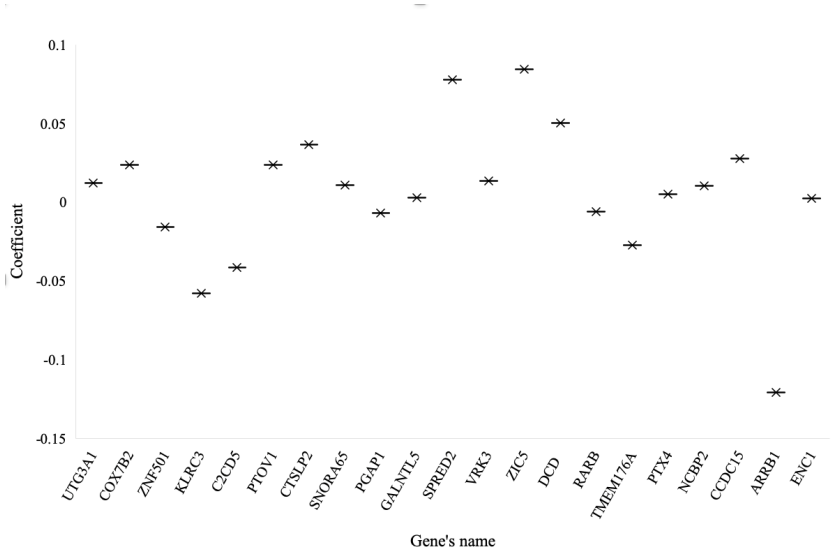

Figure 4: *Significant genes' coefficients.*

According to the final model, 14 genes have positive coefficients which are UGT3A1, COX7B2, PTOV1, CTSLP2, SNORA65, GALNTL5, SPRED2, VRK3, ZIC5, DCD, PTX4, NCBP2, CCDC15, and ENC1. In those genes, gene ZIC5 has the highest positive coefficient, which is 0.084591467. Seven genes have a negatively coefficient which are ZNF501, KLRC3, C2CD5, PGAP1, RARB, TMEM176A, and ARRB1. In those genes, KLRC3 has the lowest coefficient which is -0.0575580908. The final model indicates that those 14 genes may have a negative impact on the survival rate of triple negatives patients while the other 7 genes may have a positive impact.

References

Canadian Cancer Society. (2020). *Her2 status test - canadian cancer society.* Retrieved from https://www.cancer.ca/en/cancer-information/diagnosis-and-treatment/tests-and-procedures/her2-status-testing/?region=on

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., ... Schultz, N. (2012). The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, *2*(5), 401–404. Retrieved from https://cancerdiscovery.aacrjournals.org/content/2/5/401 doi: 10.1158/2159-8290.CD-12-0095

Cox. (1972). Each failure contributes to the likelihood function. , 191.

Daniel, A. R., Hagan, C. R., & Lange, C. A. (2011). Progesterone receptor action: defining a role in breast cancer. *Expert review of endocrinology & metabolism*, *6*(3), 359–369.

DeSantis, C., Ma, J., Bryan, L., & Jemal, A. (2014). Breast cancer statistics, 2013. *CA: a cancer journal for clinicians*, *64*(1), 52–62.

El-Naggar, Negri, & Rousseau. (2020, Mar). *2020 lifesciences bc annual award winners announced.* Retrieved from https://www.bccrc.ca/dept/mo/

Foulkes, W. D., Smith, I. E., & Reis-Filho, J. S. (2010). Triple-negative breast cancer. *New England journal of medicine*, *363*(20), 1938–1948.

Friedman, Hastie, Tibshirani, Narasimhan, Simon, & Qian. (2019). Package 'glmnet'. *Lasso and Elastic-Net Regularized Generalized Linear Models*.

Hastie, Trevor, Tibshirani, Robert, Friedman, & Jerome. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Locker, G., Sainsbury, R., & Cuzick, J. (2002). Breast surgery in the atac trial: women from the united states are more likely to have mastectomy. In *Breast cancer research and treatment* (Vol. 76, pp. S35–S35).

Statistics Canada. (2016, Mar). *Health fact sheets cancer incidence in canada, 2013 health fact sheets cancer incidence in canada, 2013*. Retrieved from `https://www150.statcan.gc.ca/n1/pub/82-625-x/2016001/article/14363-eng.htm`

Zhang, C.-D., Yang, Y., Chen, H.-H., Zhang, T., Wang, Q., Liang, Y., ... Zhou, Y. (2018). Rtpdb: a database providing associations between genetic variation or expression and cancer prognosis with radiotherapy-based treatment. *Database*, *2018*.