

Does the amount of Lipolysis-stimulated lipoprotein
receptors in breast cancer cells affect the survival
probability?

Consultant: Zhaoqi Liu

Advisors : Dr. Steve Marron and Dr. Perry Haaland

Client: Dr. Jodie Fleming, North Carolina Central University

March 9, 2021

Summary

In breast cancer, Lipolysis-stimulated lipoprotein receptor (LSR) is a protein that possibly correlates with tumor-initiating features. In this study, we investigated the effect of the LSR amount in breast cancer cells on survival probability by performing survival analyses. We found that patients with higher LSR amounts in cells have higher risk of death, without controlling for the other variables. When other variables were taken into consideration, such as patient's age, cancer stage and cancer cell grade, the evidence is not strong to call the effect of LSR amount as statistically significant, but it's close. For a fitted Lasso Cox model, the LSR amount variable was kept in the model, which made it possible to predict the patient's survival time using the LSR amount. As per client's request, we checked the effect of LSR amount while considering the effect of race and ER status on survival probability, none of the variables were significant.

Introduction

Breast cancer is the most common cancer in American women, except for skin cancer.¹ About 1 in 8 American women will develop invasive breast cancer over the course of their life, and 2.6% of the patients will die from the disease.² Many risk factors and prognostic factors of breast cancer were found and investigated. Lipolysis-stimulated lipoprotein receptor (LSR) is a protein that functions in lipoprotein endocytosis and tight junctions, and it was found to have multifaceted roles in directing breast cancer cell behavior.³ In vivo xenograft studies revealed that LSR expression enhances tumorigenesis.⁴ The goal of this report is to assess the effect of LSR amount in cancer cells, along with other prognostic factors, on the overall survival time of female breast cancer patients. Specifically, we are most interested in the survival by LSR localization, with respect to race and ER status.

Data

The data consist of the clinical information and LSR data. The clinical information contains the patients' demographics, tumor subtypes, and survival and relapse status. The LSR data are the LSR amount in the nucleus, in the cytoplasm or membrane, and the total of the two. The values were determined by using the count of cells with LSR in those compartments normalized by the total cell counts. Each patient had 2 to 3 samples of the LSR data, which were taken from various regions of the same tumor. There were 775 patients in the study, but only 348 had the follow-up information. Thus, we took the average of the LSR data for each patient and only kept the patients with follow-up information. We also eliminated the patients whose race was neither Black nor White (n=5) or grade value was missing (n=18). After cleaning, the final dataset has 323 observations and 19 variables.

Variables of our interests are:

- **Race:** subject's race (0 as Black and 1 as White)
- **Age:** subject's age at breast cancer diagnosis

- **ER:** estrogen receptor (ER) status of the subject's biopsy specimen at diagnosis (0 as negative and 1 as positive)
- **PR:** progesterone receptor (PR) status of the subject's biopsy specimen at diagnosis (0 as negative and 1 as positive)
- **HER2:** human epidermal growth factor receptor 2 (HER2) status summary at diagnosis (0 as negative and 1 as positive)
- **Grade:** the overall grade of the subject's tumor specimen at definitive surgery (0 as low grade and 1 as high grade)
- **Stage:** the overall stage of the subject's tumor specimen at definitive surgery (0 as low stage and 1 as high stage)
- **Menopause:** subject's menopause status at diagnosis (0 as post-menopause and 1 as pre-menopause)
- **AveNCount:** average count of LSR in the nucleus
- **AveCMCount:** average count of LSR in the cytoplasm and membrane
- **AveTCount:** sum of AveNCount and AveCMCount

The response variables are:

- **Survival:** subject's survival status (0 as alive and 1 as dead)
- **Survival: Survival_month:** the number of months survived since breast cancer diagnosis
- **Relapse:** subject's relapse status (0 as no relapse and 1 as local or distant cancer recurrence or died of disease)
- **Relapse_month:** Number of months that the subject is free from any relapse (NA if the subject's relapse is 0)

Methods & Results

The initial data analysis plan was to examine the effect of LSR localization, along with other prognostic factors, on the overall survival time and time to relapse. However, since only a few patients had relapse and metastasis in this dataset, when the event is relapse, no significant results were found. None of the LSR covariates were statistically significant in the simple Cox model and the proportional hazard assumption did not hold. To make the report clear and easy to read, we only show the relapse time in the exploratory data analysis (part 1) and choose to focus on the overall survival analysis in the rest parts of this Method & Result section. In this report, our goal is to assess the effect of LSR amount in cancer cells on overall survival time, and examine whether the effect is consistent when other factors are considered.

1. Exploratory Data Analysis (EDA)

We first looked at the patients' survival time and time to relapse. The results are summarized in Table 1. For the overall survival analysis, the event is death. There are 323 patients in the dataset. 232 of them were censored for the survival time, which means they were alive until the end of follow up, and 91 of them died of breast cancer or other reasons. The longest censored survival time was 225 months and the longest uncensored survival time was 210 months. Median survival time is the time that survival probability drops to 0.5. In the death group, 50% of the patients survived 63 months. For the relapse, the event is the

observed relapse. only 62 patients had observed relapse and 48 of them were dead. The median time relapse is 36.5 months, which means the smallest time that half of the relapsed patients had recurrence was 36.5 months. We can see that in this dataset, about 70% of patients survived at the time that the data was collected and only about 20% of them had observed relapse. However, the patients with observed relapse have relatively shorter time to relapse compared with the overall survival time.

| | Group | n(%) | Median Survival Time(months) | 95% Confidence Interval |
|----------|----------|------------|--------------------------------|-------------------------|
| Survival | Censored | 232(71.8%) | - | - |
| | Death | 91(28.2%) | 63 | [55,90] |
| | Group | n(%) | Median Time to Relapse(months) | 95% Confidence Interval |
| Relapse | Censored | 261(80.8%) | - | - |
| | Relapse | 62(19.2%) | 36.5 | [21,54] |

Table 1. Table of the number of cases in each group, median survival time and the median time to relapse.

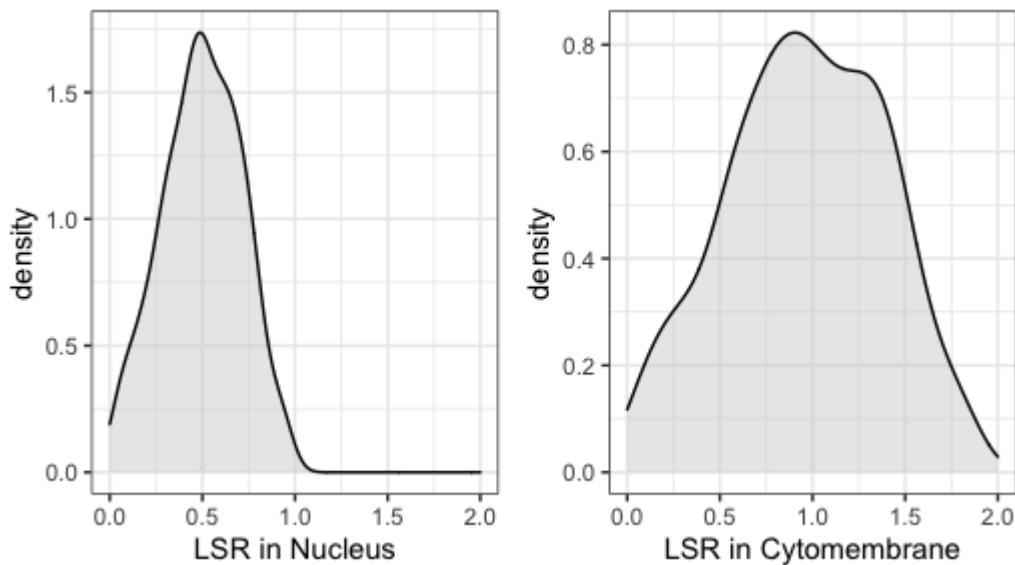


Figure 1. Density Plots of LSR in nucleus and cytomembrane. The x scales are set to be the same and the grey area under the curve is 1. The distribution seems to be normal. LSR in the nucleus is about half the size of LSR in the cytomembrane.

| | groups | mean LSR in nucleus | p-value | mean LSR in cytomembrane | p-value |
|------|----------|---------------------|---------|--------------------------|---------|
| Race | black | 0.555 | 0.00537 | 1.075 | 0.0047 |
| | white | 0.477 | | 0.921 | |
| ER | negative | 0.579 | <0.0001 | 1.122 | <0.0001 |
| | positive | 0.465 | | 0.587 | |
| PR | negative | 0.562 | <0.0001 | 1.090 | <0.0001 |
| | positive | 0.459 | | 0.884 | |

Table 2. The table summarizes the mean LSR count difference between Race, ER, and PR groups in both nucleus and cytomembrane. Black group, ER negative group and PR negative group have higher mean LSR count in both nucleus and cytomembrane.

Then, we took a glance at the LSR data. Figure 1. are the density plots of LSR in the nucleus and cytomembrane, which suggests the distribution of LSR. We can see that the range of LSR amount is 1 in the nucleus and 2 in the cytomembrane. Most of the concentration of values lies around the middle. They both seem to follow a normal distribution. We also did one-way ANOVA tests and found that the mean LSR count is different in Race, ER and PR, and the difference is statistically significant. The normality assumption is also held based on the result of the shapiro-wilk test. Table 2. summarizes the result. The boxplots can be found in the appendix. The mean LSR count is higher in the black group, ER negative group and PR negative group in both nucleus and cytomembrane. In fact, the LSR count in the nucleus and the LSR count in the cytomembrane are highly correlated (pearson correlation $\rho=0.99$). Thus, putting both variables into a linear model may cause the collinearity issue because it violates the independence assumption. Figure 2. visualizes the relationship between total LSR and different times. The upper panel shows that the censored data have an overall longer survival length than uncensored data since most of the red points appear on the middle right side, but there is no apparent visual relationship between LSR and survival time. The bottom panel shows the LSR distribution in the time to relapse. The plot only shows 62 patients who had observed relapse. It seems that there is a weak negative trend, which means that a patient with high level LSR seems to have a shorter length of time to relapse.

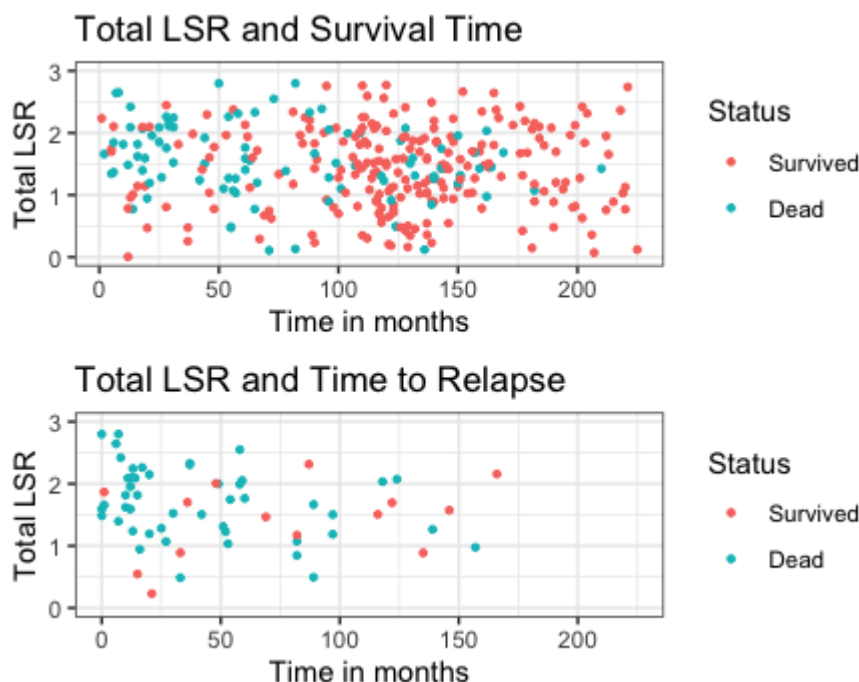


Figure 2. Total LSR in cells vs Survival Time and Time to Relapse. Each point represents a patient; blue means the patient was dead and red means alive at the end of follow up time. The plots show that LSR could have a negative effect on the time to relapse but it's hard to find statistically significant results using this dataset.

2. Kaplan-Meier Estimator and Log-rank Test

In survival analysis, there are different types of events. The time to the occurrence of the event of interest is called survival time. In this report, our event of interest is death. We used the Kaplan-Meier estimator to estimate the survival function, and used the log-rank test to compare the survival distribution of two groups. The null hypothesis is that there is no difference in survival between the two groups. If p-value is less than a certain threshold value α , which is referred to as the level of significance, we would like to reject the null hypothesis and say that groups differ significantly in survival at the level of α . The cutoff point is commonly set to be 0.05. Kaplan-Meier plots are drawn for each categorical variable to visually compare the survival function in different groups. Here, we took Age and ER variables as examples, and the Kaplan-Meier plots for other variables can be found in the appendix. ER is the variable that the client wants to see specifically. Age is the variable that the difference in survival between two groups is strongly significant.

Figure 2. shows the overall survival curves. The patients were divided into two groups based on whether they are older than 54 years old, which was the median age of the 323 patients. The median survival time for the older group is 164 months, as opposed to no median survival time for the younger group since the survival probability only drops to 0.75. There appears to be a survival advantage for younger patients with breast cancer compared to the olds. The log rank test for difference in survival gives a p-value of $p = 0.0043$, which is a very strong result. It indicates that the age groups differ significantly in survival. Older age reduces overall survival probability.

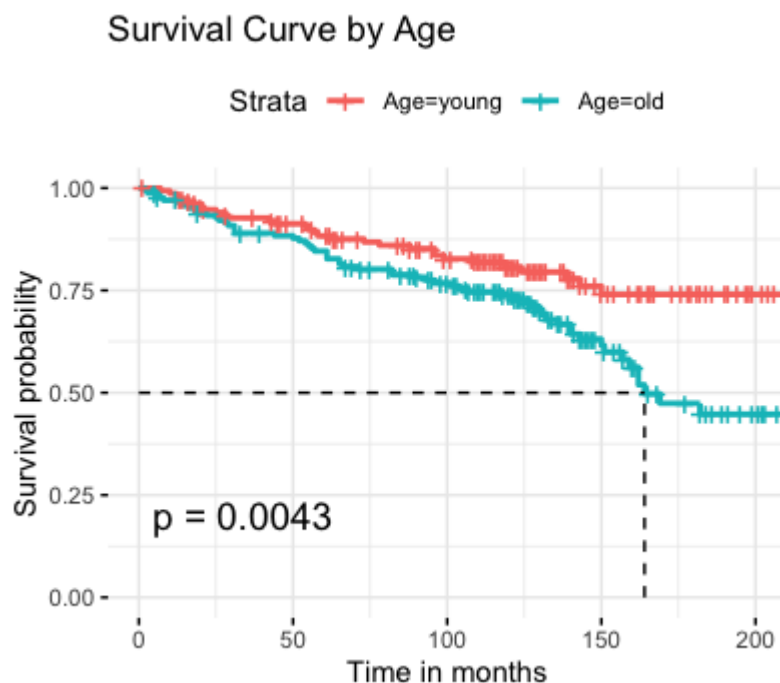


Figure 3. Kaplan-Meier survival curve estimates of two age groups for overall survival. The red curve is the Kaplan-Meier estimate of survival distribution conditional on age < 54 . The cyan curve is the Kaplan-Meier estimate of survival distribution conditional on age ≥ 54 . The value at the bottom

left is the p-value of the log-rank test. The black dashed line indicates the median survival time in that group. For overall survival, the younger group has better survival probability.

The estrogen receptor (ER) positive of the biopsy specimen means the cancer has receptors for estrogen, which suggests that the cancer cells could receive signals from estrogen.⁵ Similarly, the progesterone receptor (PR) positive suggests that the cancer cell may receive signals from progesterone.⁵ The testing results are vital because they directly determine the treatment.⁵ If the cancer cells have hormone receptors, patients can undergo hormonal therapy.⁵ In Figure 3., the ER positive group seems to have a survival advantage over the ER negative group, since the ER negative group has the median survival time but the ER positive does not. The log-rank test gives the p-value of 0.057, even though it's not strongly significant, which means that it's not statistically significant at the 5% level of significance, but it's very close. What's more, we also tried to do the stratified log-rank test on ER, which compares two groups while adjusting for the PR status. We find that the survival difference may exists when PR is positive, since the p-value = 0.067 is close to be regarded as strongly significant, but there is no statistically significant difference in survival when PR is negative (p-value = 0.5).

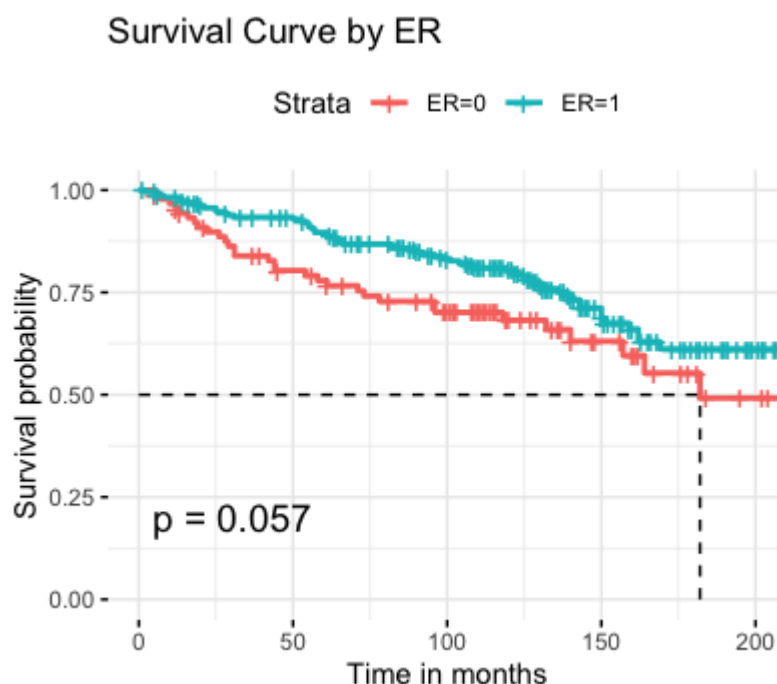


Figure 4. Kaplan-Meier survival curve estimate for overall survival by ER status. Even though the log-rank test gives the p-value 0.057, which is not strongly statistically significant at the 5% significance level, it's very close.

In this study, we did not find significant differences in survival distribution between black and white groups. The p-value for its log-rank test is 0.64. The difference between pre and post Menopause groups in overall survival curves was significant, using the log-rank test (p-value = 0.031). However, the menopause status is highly correlated with age. Thus, we performed a stratified log-rank test, which compared the pre and post menopause groups while adjusting for the categorical age. After stratifying the age variable, the difference

between pre and post menopause was no longer significant (p-value = 0.35). With a significance level of 0.05, we found that Age, PR, Grade, and Stage variables have statistically significant group differences for the overall survival distribution. In other words, the young group, the PR positive group, the low-grade group, and the low-stage group seem to have higher survival probabilities.

3. Cox Proportional Hazards Model

The Cox proportional hazards model allows us to fit regression models to censored survival data and examine how specific factors influence the rate of a particular event happening at a particular point in time. The coefficient estimate is the estimate of log hazard ratio, which measures the effect size of the covariate. If coefficient is greater than 0, the hazard ratio will be greater than 1. Thus, for a quantitative variable, as the value of the covariate increases, the event hazard increases and the length of survival decreases. The Wald test, score test, and likelihood ratio test are derived using partial likelihood, which usually yields the same results for large enough samples.

We first studied the effect of LSR count on the overall survival time. We fit simple Cox models using each one of the covariates: LSR in nucleus, LSR in cytomembrane and LSR in total. Efron method was used to handle tied survival times. Table 3. is a result summary of these three univariate Cox models. All three covariates have positive coefficient estimates and the proportional hazard assumption holds. The three p-values are very close. All of them are right around the 0.05 threshold, the p-value of the coefficient of LSR in the nucleus is slightly above 0.05, which fails to make it statistically significant, but it's close. The other two are slightly below the threshold, so they are statistically significant. What's more, the LSR in the nucleus has the biggest hazard ratio even though it's not strictly statistically significant.

| | Coefficient | SE | Hazard Ratio | Wald Test | P-value |
|--------------|-------------|------|--------------|-----------|---------|
| Nucleus | 0.97 | 0.51 | 2.64 | 1.915 | 0.055 |
| Cytomembrane | 0.51 | 0.25 | 1.67 | 2.019 | 0.043 |
| Total | 0.33 | 0.17 | 1.40 | 1.987 | 0.047 |

Table 3. Univariate Cox models result summary for each one of the LSR variables. The table shows the coefficient estimates, standard error of coefficient, hazard ratios, the Wald test statistics of overall significance and the corresponding p-values. The three p-values are very close, and the LSR in cytomembrane and in total have statistically significant coefficient. Higher LSR in cytomembrance and higher total LSR in the cells are associated with poorer survival. Even though the effect of LSR in Nucleus on survival is not statistically significant, it has the largest coefficient with a relatively large standard error.

LSR amount in cytomembrane and LSR amount in total have statistically significant coefficients. Higher LSR amount in cytomembrance and higher total LSR amount in the cells are associated with poorer survival. Specifically, A one unit increase in the LSR amount in cytomembrance results in a 67% additional risk of death, and one unit increase in the total LSR amount in cells increases 40% risk of death.

We were also interested in how the LSR data affect the overall survival after controlling the other variables. We used Akaike Information Criterion to do stepwise model selection. AIC balances the goodness of fit and the number of parameters in the model, so that the selection will not end in keeping all variables in the final model. A smaller value of AIC indicates a better model. The model selected based on AIC includes LSR count in nucleus, Age, HER2, Grade and Stage, even though the coefficient of HER2(p-value = 0.13) and AveNCount(p-value = 0.14) were not statistically significant. The coefficients are shown in the forest plot (Figure 4.).

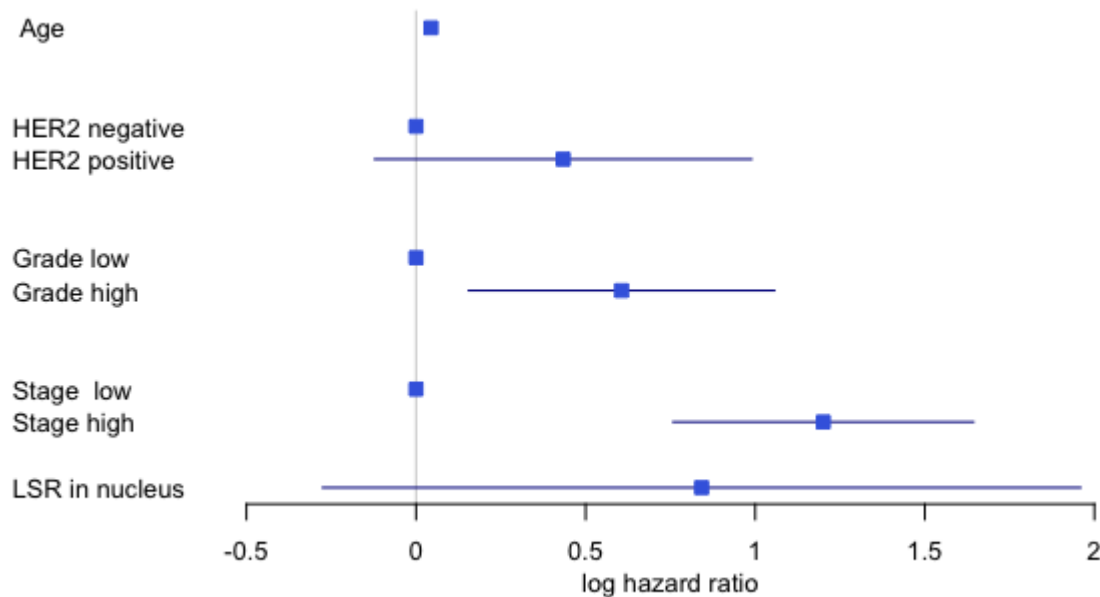


Figure 5. Forest plot of parameter estimates of log hazard ratios for the Cox model selected based on AIC. The square denotes the coefficient estimate and the horizontal line denotes the 95% confidence interval. Younger patients, patients in low grade, and patients in low stage have a better survival rate.

The forest plot is a plot of coefficient estimates with 95% confidence intervals. Each categorical variable has a reference group. We can see that the Age variable is highly significant with a small magnitude of coefficient estimate. High grade group has a higher risk of death compared with low grade group, and high stage group has worse results than low stage group. Usually, higher stage cancers are typically more aggressive. High grade cancer cells are usually poorly differentiated and much more different from normal cells than low grade cells. Specifically, the coefficient estimate of high-grade is 0.606, which means that holding other variables constant, patients in high grade have 1.833 ($=e^{0.606}$) times the risk of death as patients in low grades do. Similarly, patients in high stage have 3.32 times the risk of death as patients in low stage do. A one-year increase in age adds 4.5% risk of death. However, we cannot draw any conclusion on the other variables. Even though the LSR count in nucleus has a positive coefficient, which suggests that higher LSR amount in nucleus has a higher hazard of death, but the effect is not statistically significant (p-value = 0.14 based on the wald test). The uncertainty is very high after adjusting for the other variables, and only some trend toward significance.

4. The Lasso Method for Variable Selection in the Cox Model

The Lasso Cox model is a good approach to predict a patient's survival prospects. Lasso stands for Least Absolute Shrinkage and Selection Operator and it is a penalized regression method. Lasso shrinks the regression coefficients and constrains the number of variables in the model by penalizing the regression model with a penalty term. The penalty can be tuned using a tuning parameter λ . In this study, we used 10-fold cross-validation to pick the optimal value of λ . We used the “penalized” package in R to build the model.

Since the three LSR amount variables are highly correlated, which can make the parameter estimates unreliable, we fitted two models to check whether the results were consistent. We first fit a model using all clinical variables and LSR in total only. We used a 10-fold cross-validation approach to select a value that optimizes the predictive ability of the Lasso model, as defined by maximizing the cross-validated partial log-likelihood (CVL). Figure 6. shows the optimal CVL lambda and corresponding lambda. The number and magnitude of the coefficient estimates depends on the penalty size, which is determined by λ . The plot confirms that the optimization procedure has selected the global maximum. Then, we can plot the coefficient profile (Figure 7.).

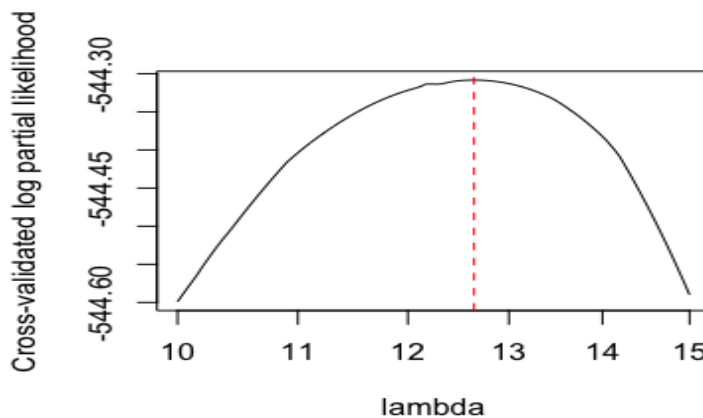


Figure 6. Cross-validated log partial likelihood for a range of values of lambda for the model with all clinical variables and LSR total account in cells. The red dashed vertical line shows the global maximum at $\lambda = 12.65$.

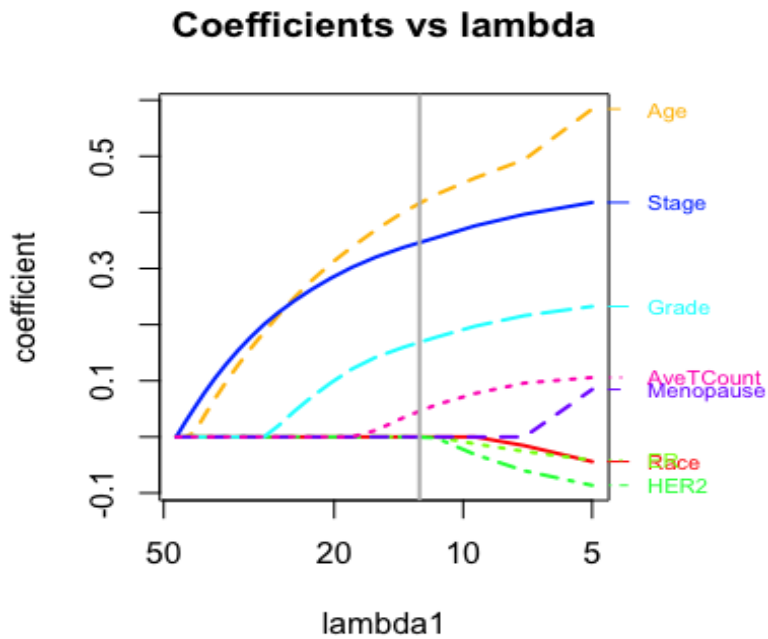


Figure 7. Paths of the standardized coefficients estimate over a range of values of λ for the model only using all clinical variables and LSR count in total. The gray vertical line is the optimal value of λ , the same value in Figure 4.

In Figure 7., four curves intersect with the gray line and all the intersects are positive coefficients. Thus, in this Lasso Cox model, four variables are left after variable selection using the λ selected before. The coefficient for Age is 1.517, the coefficient for Stage is 1.184, the coefficient for Grade is 1.414, and the coefficient for AveTCount is 1.04.

In the second model, where all variables are put in, the results are very similar. The optimal λ is 12.84. Age, Stage, and Grade variables are also selected with roughly similar coefficients 1.513, 1.182 and 1.412. The only difference is that, instead of selecting the total LSR amount, the model selects two separate variables: LSR in the nucleus and LSR in the cytomembrane. The coefficients profile is shown in Figure 8.

Age, Stage and Grade are the variables that have significant group differences in overall survival (Methods and Results 2 & 3). And we are happy to see that the LSR total is selected in the first model, and LSR in the nucleus and LSR in the cytomembrane are selected in the second model. Unfortunately, the magnitudes of these parameter estimates are not intended to be interpreted in terms of hazard ratios, because of shrinked coefficients and standardization with standard deviation equal to 1. We can use these coefficients to predict the survival for a new patient. That's to say, the LSR amount can be useful in prediction and

some association can be further analyzed.

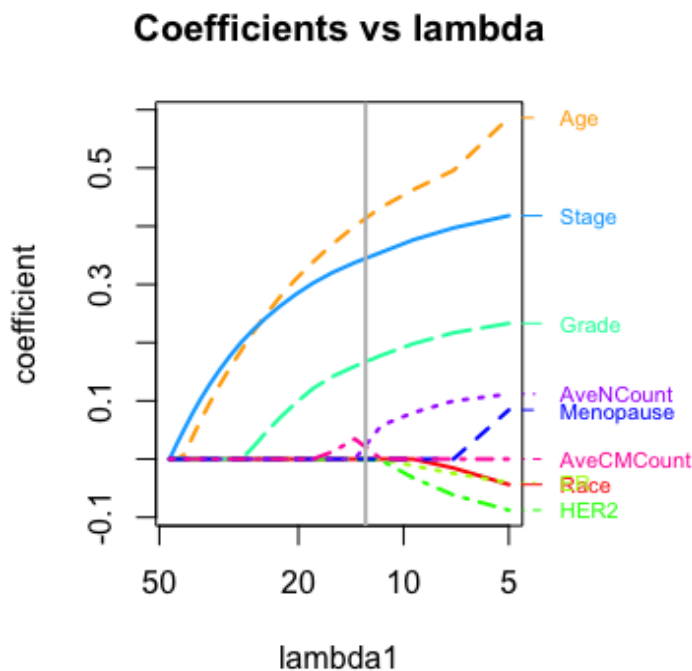


Figure 8. Paths of the standardized coefficients estimate over a range of values of λ for the model using all the variables. The gray vertical line is the optimal value of $\lambda = 12.84$.

Discussion

This is a preliminary study of the effect of LSR on survival time of breast cancer patients. For overall survival, the LSR amount in cytomembrance increases the hazard of dying, based on the result of the univariate Cox model. However, when considering along with other factors, the effect of LSR on survival is not quite statistically significant, although it's close. Because the results are very close to the boundary between significant or not, they can be sharpened by using a larger sample size study. The result of the multivariate Cox model shows that survival is relatively poor for high stages, high grades and older people. What's more, the fitted Lasso Cox model kept the LSR amount variables after penalization, but the coefficient estimates of the model are difficult to interpret. Anyway, we still can use the Lasso Cox model to predict new patients whose LSR amount data are known. For further study, introducing more samples can be useful. In addition, since the data of LSR amounts in nucleus and cytomembrane are highly correlated, we could not see much difference between them. In the models we fitted, the LSR amount in the nucleus always had a larger coefficient than the LSR amount in cytomembrane did, but also higher uncertainty, which made the results nearly but not quite statistically significant. If the localization of LSR is of interest, LSR amount in the nucleus may be a good way to start.

Reference

1. Breast Cancer. Centers for Disease Control and Prevention website. Updated September 14, 2020. Accessed March 5, 2021. https://www.cdc.gov/cancer/breast/basic_info
2. How Common is Breast Cancer. American Cancer Society website. Updated January 12, 2021. Accessed March 5, 2021. <https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer>
3. Reaves DK, Fagan-Solis KD, Dunphy K, Oliver SD, Scott DW, Fleming JM. The role of lipolysis stimulated lipoprotein receptor in breast cancer and directing breast cancer cell behavior. *PLoS One*. 2014;9(3):e91747. Published 2014 Mar 17. doi:10.1371/journal.pone.0091747
4. Reaves DK, Hoadley KA, Fagan-Solis KD, et al. Nuclear Localized LSR: A Novel Regulator of Breast Cancer Behavior and Tumorigenesis. *Mol Cancer Res*. 2017;15(2):165-178. doi:10.1158/1541-7786.MCR-16-0085-T
5. Hormone Receptor Status. Breast Cancer Organization website. Updated Sep 21, 2020. Accessed March 6, 2021. https://www.breastcancer.org/symptoms/diagnosis/hormone_status

Appendix

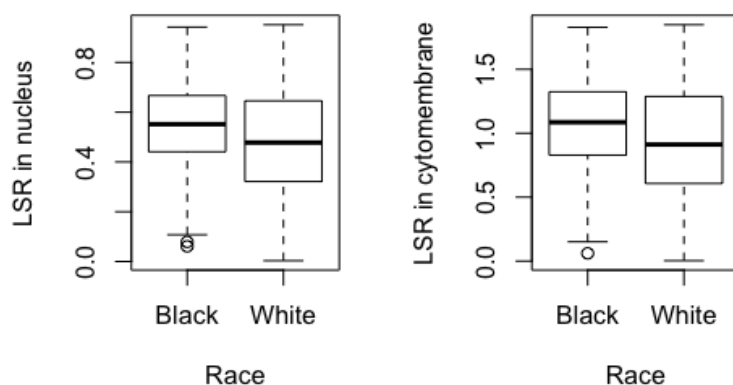


Figure 9. Boxplots of LSR count in the nucleus and in the cytomembrane by race. The mean difference of LSR count between black and white are statistically significant. The black group has higher mean LSR count in both nucleus and cytomembrane.

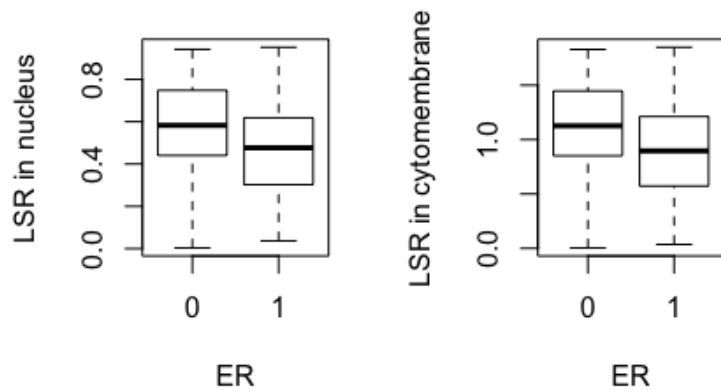


Figure 10. Boxplots of LSR count in the nucleus and in the cytomembrane by ER. The mean difference of LSR count between ER negative and ER negative are statistically significant. The ER negative group has higher mean LSR count in both nucleus and cytomembrane

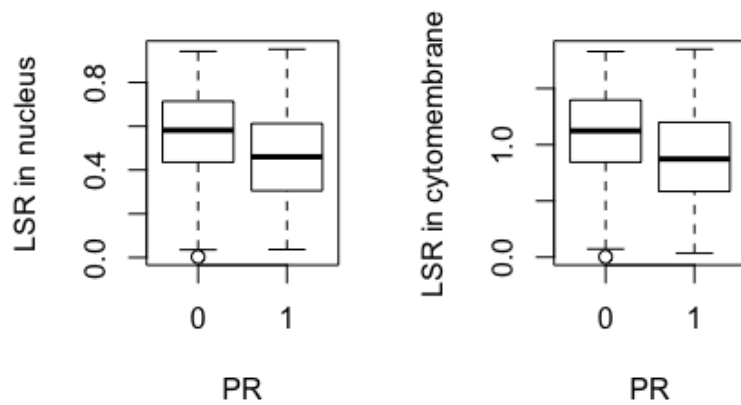


Figure 11. Boxplots of LSR count in the nucleus and in the cytomembrane by PR. The mean difference of LSR count between PR negative and PR negative are statistically significant. The PR negative group has higher mean LSR count in both nucleus and cytomembrane

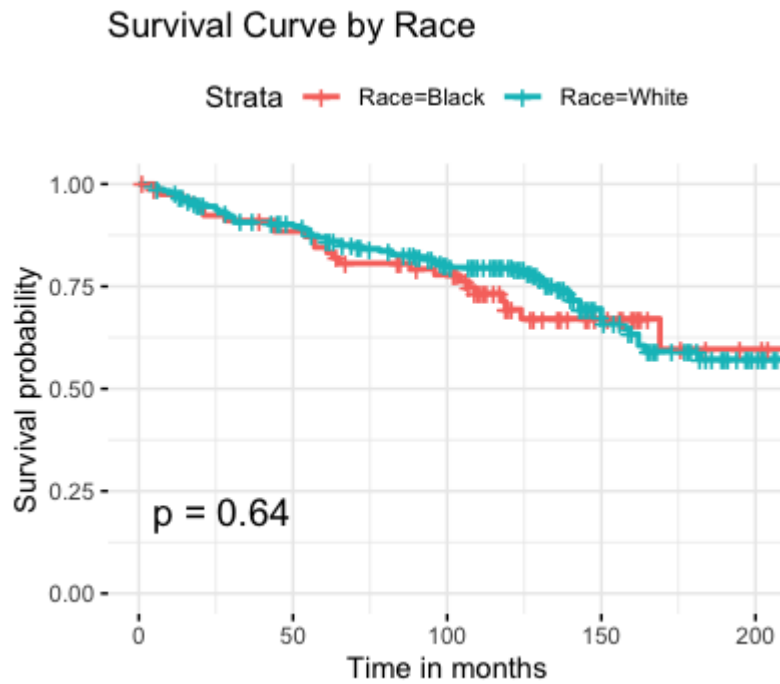


Figure 12. Kaplan-Meier survival curve estimates of two race groups for overall survival. There is no difference in survival between the black and white groups. The log-rank test has p -value = 0.64.

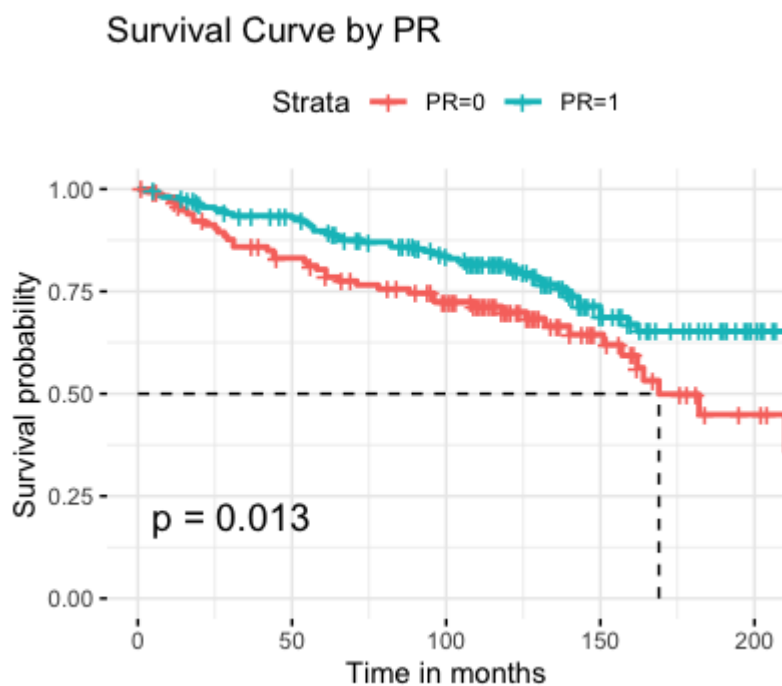


Figure 13. Kaplan-Meier survival curve estimates of two PR-status groups for overall survival. The PR groups differ significantly in survival. The log-rank test has p -value = 0.013.

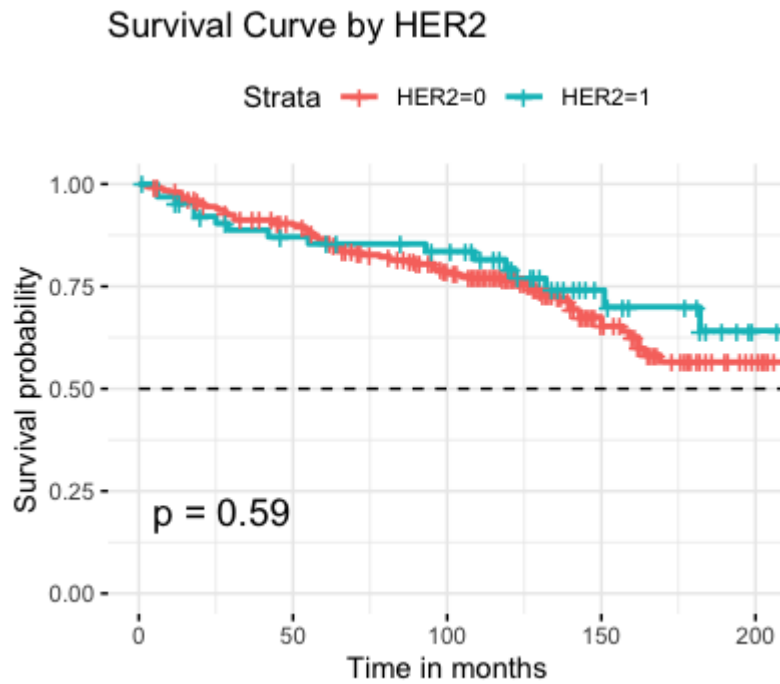


Figure 14. Kaplan-Meier survival curve estimates of two HER2 groups for overall survival. There is no difference in survival between HER2 negative and HER2 positive. The log-rank test has p -value =0.59.

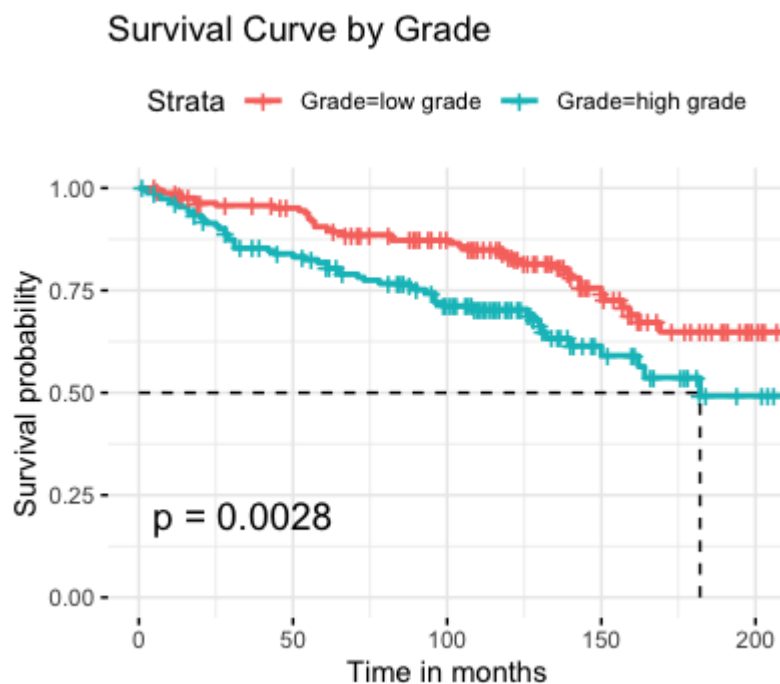


Figure 15. Kaplan-Meier survival curve estimates of two grade groups for overall survival. The grade groups differ significantly in survival. The log-rank test has p -value =0.0028.

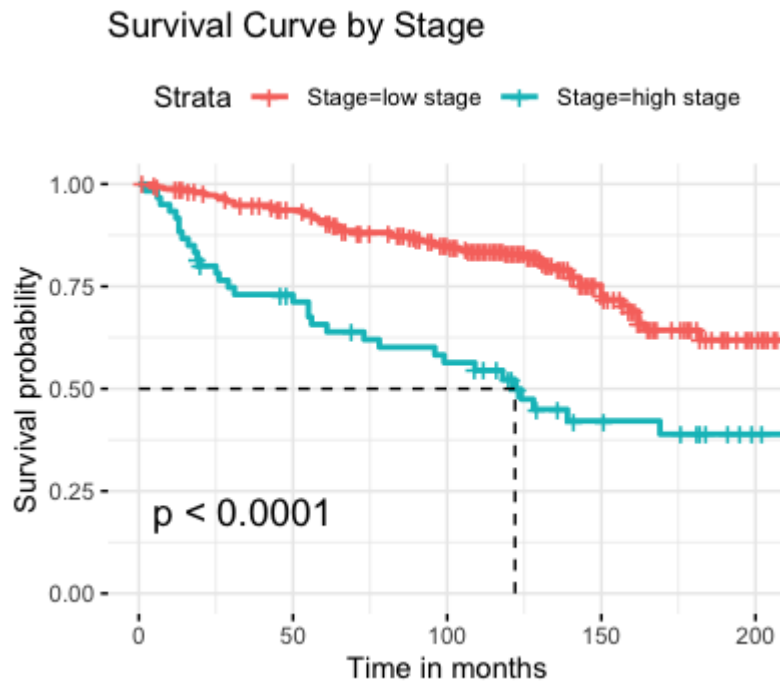


Figure 16. Kaplan-Meier survival curve estimates of two stage groups for overall survival. The stage groups differ significantly in survival. The log-rank test has p -value < 0.0001 .

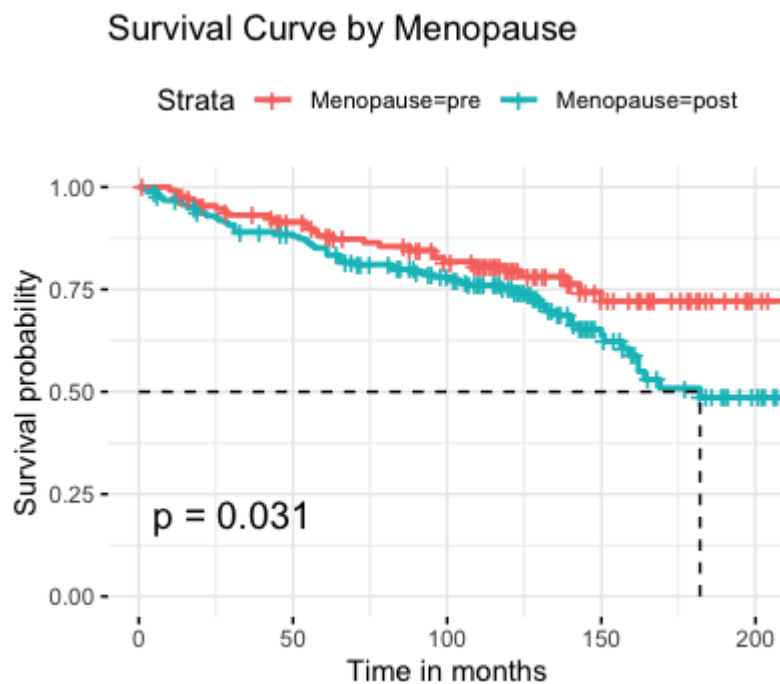


Figure 17. Kaplan-Meier survival curve estimates of two menopause groups for overall survival. The menopause groups differ significantly in survival. However, since menopause status are highly correlated with age, the group difference disappears after controlling the effect of age.