Liver Cancer Patient Subtypes and their Respective Survival Rates

Author: Shirley Wang

Question

Are there distinct biological "subtypes" of patients with liver cancer that have different survival times? What impact does age at diagnosis, tumor stage, and sex have on the relationship between survival and subgroup? For this report, I will be using hierarchical clustering to find subtypes of patients through their gene counts, and then conducting survival regression to see differences in survival and what factors have a significant effect on survival within the subgroups.

Data

The data is gathered from the Genomic Data Commons website using The Cancer Genome Atlas (TCGA) transcriptome sequencing data for liver cancer. There are 424 cases recorded, each with 60483 genes counted. There is also 742 cases of clinical information regarding the patients in these cases with genome data. The gene counts data is all positive and right-skewed, so I've added 1 to all of it and then taken the log to transform it to be more normal.

The variables of interest for the survival analysis are days_to_death, days_to_last_follow_up, and vital_status. If a patient has died, their vital status is Dead, and I take their days_to_death as days survived, which records the number of days they lived after they were diagnosed. If a patient has been censored, their vital status is Alive, and I take their days_to_last_follow_up as days survived, which records the last time the study checked the status of the patient and they were still alive. There are twelve cases where both days_to_death and days_to_last_follow_up are missing, so I count those as outliers and removed them for the analysis.

The possible predictors of survival are age at diagnosis, tumor stage, and sex. When viewing the clinical data in terms of only these predictors, the survival variables, plus the patient id they correspond to, their are duplicates of each case, so those have been removed. Tumor stage has stages 1 to 4, with substages a, b, and c. Due to the severe decrease in data in higher stages, with only 1 person at stage 4, I have collapsed all substages down into just the numbered stages. There are also 24 cases of tumor stage not reported, so those have been removed for the survival regression portion.

Methods

Part 1: Cluster Analysis. I will be using Principal Component Analysis to reduce the number of gene features and dimensionality of the data. After that, using the PCA-transformed gene counts data, I will use hierarchical clustering to locate distinct clusters in the data. Then using the measures of survival defined above, I will use a multivariate logrank test to compare the survival distributions between the clusters to see if they are significantly different or not.

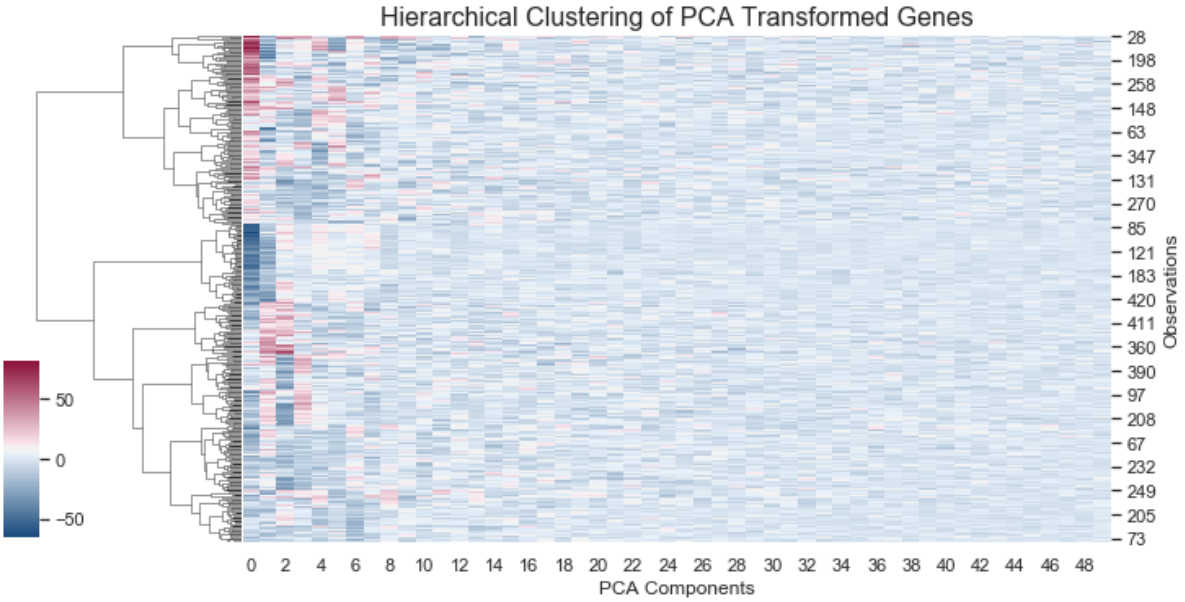
Part 2: Survival Analysis From the clusters found in part 1, I will use a Cox regression model for each group with the predictors DiagnosisAge, Sex, and TumorStage to see if any of the predictors are significant, and if the effects of the predictors are different from group to group. I will also view the distributions of these variables in each group to see if the same type of people tend to end up in the same groups.

Results

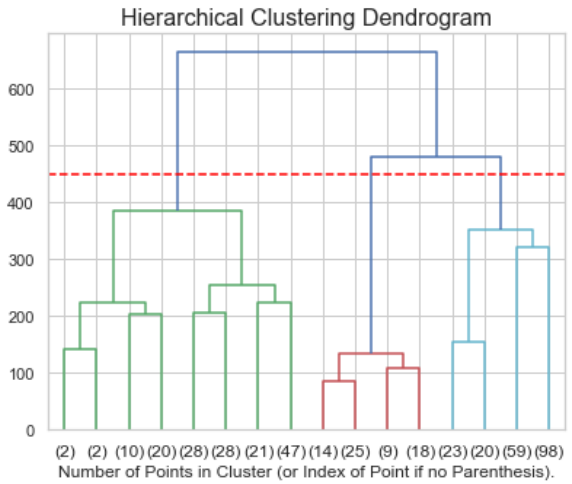
Principal Component Analysis

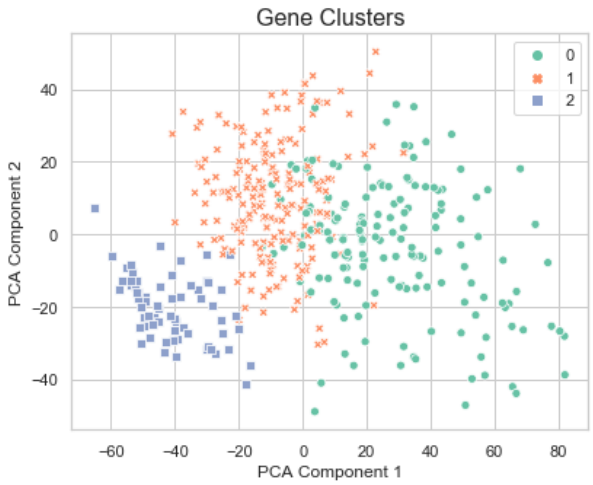
With just 50 components, 71.9% of the variation in the data is captured. I will use the 50 PCA component transformed gene data for the clustering analysis, since a large portion of the variance is already captured and this way I avoid problems that come with working with incredibly high dimensional data. I will also be using the 2 PCA component transformed gene data for graphing purposes, and with only 2 components, 25.7% of the variation in data is already captured, so the visualizations should be fair.

Hierarchical Clustering



From the heatmap of the PCA transformed genes, there appears to be a group with very high values, a group with very low values, and a group with moderately low values in the first component.

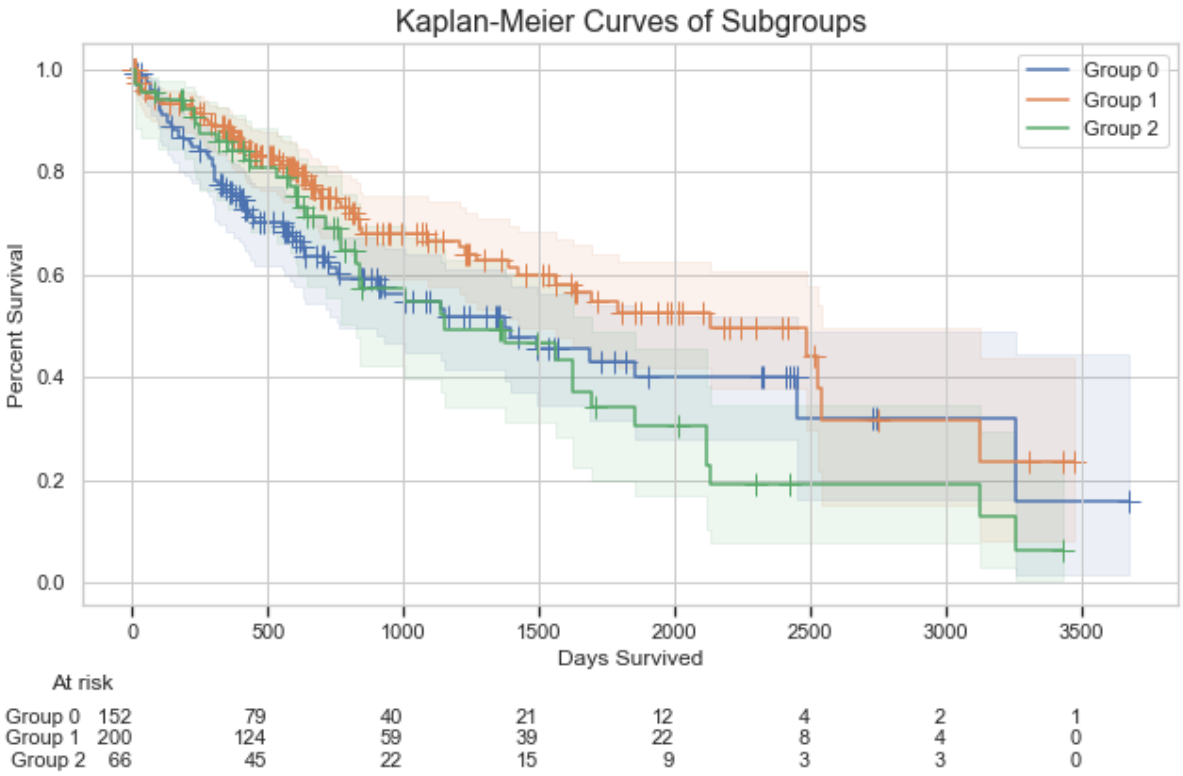




test_statistic	p
7.38122	0.024957

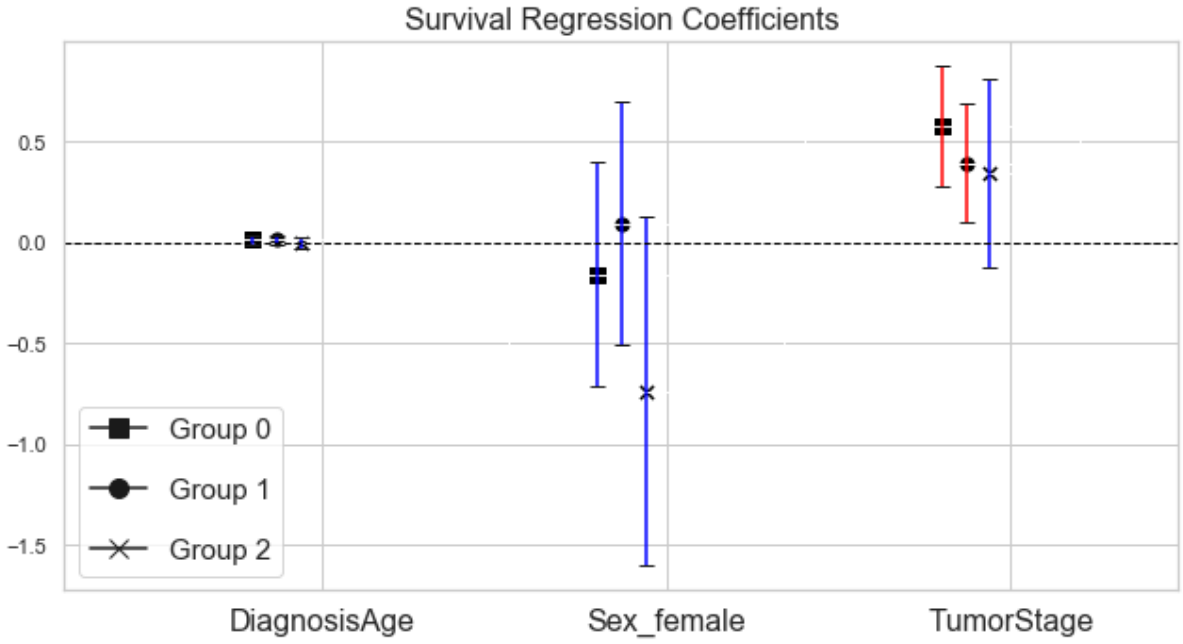
Using a multivariate logrank test, the p-value is less that 0.05, meaning we have significant evidence that at least two groups have different survival distributions.

Survival Time Comparison



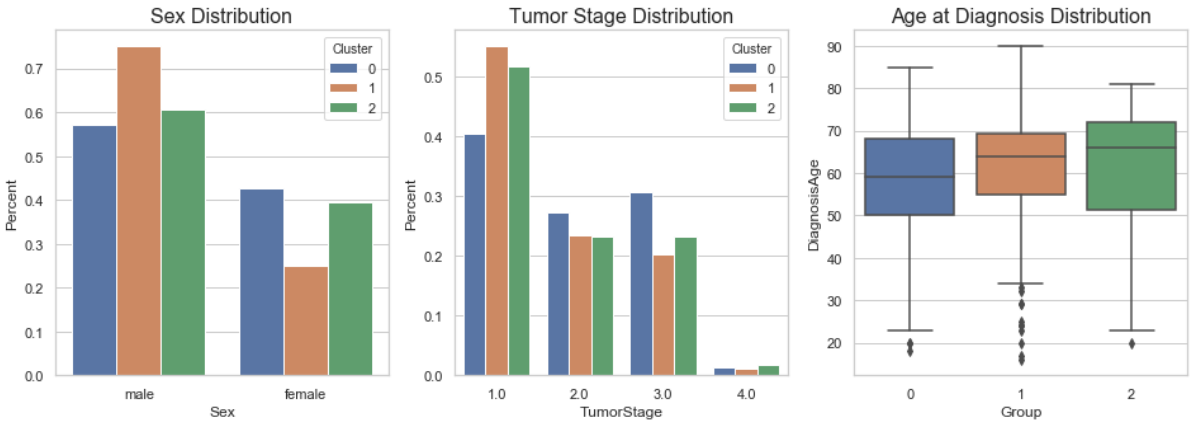
The Kaplan-Meier curves of the three groups all overlap in some places, although their shapes are somewhat different. Group 0 has the least survival early on, although after the 1000 day mark its survival rate descends slower. Group 1 has a better survival rate than Group 0 up until the 2500 day mark, when its survival rate becomes relatively the same as Group 0, so people in Group 1 tend to have a better short term survival rate than the other groups, although in the long term its relatively the same. Group 2 has a similar survival rate to Group 1 until the 800 day mark, when its survival rate decreases much faster than Group 0 and 1, and it has the lowest survival rate in the long term.

Survival Regression



I conducted a Cox regression using age at diagnosis, sex, and tumor stage as predictors for survival time and plotted the regression coefficients and their confidence intervals. Age at diagnosis and Sex both result in no significant predictors. Tumor stage appears to be significant for group 0 and 1, but not group 2. However, the confidence intervals for both group 0 and 1 overlap by quite a lot with eachother and with group 2, and it's possibly due to how group 2 has much less data in it than group 0 and 1 which resulted in a larger confidence interval and it being the only one whose coefficient for tumor stage wasn't significant. Since all the confidence intervals overlap greatly, it doesn't seem as though the effects of tumor stage vary group to group either. The significant coefficient implies that an increase in tumor stage results in an increased risk of death.

Distribution of Predictors within Groups



There appears to be higher percentage of males in group 1 than the other groups. There also appears to be a lower percentage of tumor stage one patients in group 0, with a corresponding larger percentage of tumor stage two and three patients. There doesn't appear to be any significant differences in distribution of age between the groups.

Conclusion

There appear to be three types of patients with liver cancer, identified by their similarity in gene counts. Group 0 appears to have relatively high values for the PCA-transformed component one, and a lower short term survival rate. Group 1 appears to have moderately low values for the PCA-transformed component one, and a higher short term survival rate but around the same long term survival rate as Group 0. Group 2 appears to have very low values for the PCA-transformed component one, and a high short term survival rate but a much lower long term survival rate than the other groups.

Both age at diagnosis and sex don't appear to have any significant effects on the survival rate of a patient with liver cancer. Tumor stage does seem to have an effect, with a higher tumor stage resulting in a higher death rate in groups 0 and 1. Group 2's coefficient for tumor stage isn't significant, but its confidence interval contains both group 0 and 1's coefficients, so its possible that tumor stage also affects it in the same way. Tumor stage having an effect makes sense, since a more severe tumor would seem to imply more serious symptoms and a higher chance of death. The effects of an increase in tumor stage doesn't seem to vary from group to group though, so none of the predictors seem to have different effects between groups.

There is a slightly higher percentage of males in group 1 than the other groups, and a slightly higher percentage of people with tumor stage three in group 0, which could be why group 0 has a lower short term survival rate. Maybe future analysis could try and see if the higher percentage of people with tumor stage three in group 0 is a coincidence, or if people with these types of genes really are more at risk of developing a tumor at stage three. It isn't a particularly large difference in percentage from the other groups though. Overall the type of people in each group doesn't appear to be significantly different, and perhaps more cases of liver cancer could help uncover if this is a trend or a coincidence.

General Considerations

There is an imbalance in the number of points in each group, with 158 in group 0, 200 in group 1, and 66 in group 2. I've taken the percentage of counts when viewing the distributions of predictors in each groups to alleviate this, although this might be what resulted in the coefficient for Tumor Stage in group 2 not being significant, when it was for group 0 and 1.

The clinical information and gene data is merged on the clinical information's submitter id and the gene data's case id. While submitter id in the clinical information is unique, the case id linked to the gene data has quite a few cases of ids appearing 2 times, so it is likely some pairs of patients end up with the same clinical information when merging the dataframes. It's unclear if this means one person had their gene counts taken multiple times with different results, or if it does refer to two different people but only one of their clinical information was logged in the database.

Since clustering was done on the PCA-transformed data, it's hard to interpret exactly what effect any gene has in creating the differences in the PCA transformed gene counts. Further analysis could be done with a biologist in viewing the genes most involved in creating the variation in clustering and seeing if the groups found in the gene counts appear to make sense. As well, there is much less data of patients with higher tumor stages, so adding more of those would help assess the effects of tumor stage between the groups.