Linking Clinical Outcomes to Genetics in Breast Cancer

Foundations of Machine Learning (DS3001) Final Paper

By: Lily Thomson and Alysha Akhtar

1. Executive Summary:

It is estimated that 1 in 8 women (13%) in the United States will develop invasive breast cancer in their lifetime.[2] Breast cancer is a very complex disease with many genetic factors contributing to a patient's prognosis with the disease. The dataset we used is from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database that contains clinical attributes and Z-scores of gene expression of 1905 patients with breast cancer.[3] We wanted to see if we could train a model to depict the relationship between these clinical and genetic attributes and survival in months. Specifically, could we predict a breast cancer patient's overall survival in months based on clinical attributes, gene Z-scores, and cancer cell surface receptor statuses using partial least squares regression (PLSR)? We chose to use 13 components for our PLSR model after seeing that the percentage of variance in survival time began to level off after doing cross-validation. We calculated the VIP scores for features in the model and retrained the PLSR model only using features with a VIP score greater than 1. After retraining the model, the predicted survival time was plotted against the observed survival time using the training data. The R^2 of this model on testing data is only 0.01, which means that the model does not explain 99% of the variance. The R^2 of the predicted survival time plotted against the observed survival time using the training dataset is 0.32, which is better, but still does not exhibit strong predictive power. Thus, we could not effectively predict a breast cancer patient's overall survival time based on clinical attributes, gene expression, and receptor statuses using the features that we selected for the PLSR model.

2. Introduction:

Breast cancer is a difficult disease to treat because there are so many subtypes with different genetic characteristics that require different treatment paradigms. We want to investigate how gene expression, the presence of different surface receptors, and clinical attributes affect the overall survival of a patient with breast cancer. The dataset we are looking at is from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database that contains clinical attributes and Z-scores of gene expression of 1905 patients with breast cancer.[3] We want to see how specific clinical attributes like age at diagnosis, tumor size, and mutation count play a role in the overall survival of a patient with breast cancer. In 2024, approximately 310,720 women will be diagnosed with invasive breast cancer. About 16% of these women will be younger than 50 years old.[2] A study at the National University of Ireland School of Medicine concluded that age at diagnosis has a strong association with overall survival, with lower survival being seen in patients diagnosed under 50 years old. The lowest survival was seen for patients over 70.[9] Tumor mutation burden (TMB) is the measure of genetic

mutations in the DNA of a tumor. TMB-high cancer cells are defined as having ten mutations per megabase of the tumor's genome.[11] In a clinical study on patients with HER2+ metastatic breast cancer, the median overall survival for patients with low and high TMB groups was 44.9 months and 85.8 months, respectively. This was a statistically significant difference in overall survival, likely explained by the better response to immunotherapy.[8] Additionally, it is generally accepted that tumor size is related to overall survival, with a smaller tumor corresponding with better overall survival.[12]

The presence of cancer cell surface receptors is an important factor in determining what treatment patients are eligible for. ER status (er_status) and PR status (pr_status) in the dataset indicate whether the breast cancer has estrogen or progesterone receptors on its surface. This is important because if a patient is estrogen receptor positive or progesterone receptor positive, they are eligible for endocrine therapy that can be used to block cancer growth. HER2 status refers to the presence of human epithelial growth factor receptor 2. While healthy breast cells produce some HER2, some breast cancer cells produce extra HER2. If someone is classified as HER2-positive or HER2-low, they are eligible for separate treatments that can target the HER2 protein itself.[1]

These are just a few characteristics of the potential makeup of breast cancer patients. This dataset contains 498 numerical variables, including the clinical attributes discussed above, gene expression data, and dummy variables like receptor status. We want to investigate if we can predict a breast cancer patient's overall survival in months based on these features using partial least squares regression (PLSR). We began by doing exploratory data analysis (EDA) to understand the distribution of certain variables in the dataset. We plotted a histogram of the age at diagnosis and duration from the time of intervention to the time of death to gather insights into the data's spread. We also looked at the gene expression of CDH1 and BRCA1, whose mutations are linked to a higher risk of breast cancer.[6] Finally, we looked at a histogram of er_status, pr_status, and her2_status (positive or negative) to understand the presence of surface receptors in patients in the dataset.

Given the complexity of breast cancer, patients often require different treatment plans. A model that could identify critical predictors or estimate survival times could be used by healthcare professionals to aid in treatment decisions. We want to model a response variable (overall survival in months) given a large number of highly correlated predictor variables, so we decided to use PLSR for our analysis.

We conducted a PLSR model using 13 components. After calculating the VIP scores for model features, we retrained the model, only using features with a VIP score greater than one. The features with the highest VIP scores were age_at_diagnosis (2.54055) and tumor_size (2.310454), which aligns with our predictions based on the literature. However, after plotting the predicted survival time vs the observed survival time on the training data, the R^2 is only 0.01. The R^2 of the predicted survival time versus the observed survival time for the testing dataset is 0.32. The model performed very poorly on the unseen data and performed relatively poorly with the seen data (poor predictive power). While this exploration allowed us to gain a deeper

understanding of the complexity of breast cancer and the factors that contribute to the overall survival of a patient with this disease, we were unable to effectively predict a breast cancer patient's overall survival time based on the features chosen for the PLSR model in the METABRIC dataset.

3. Data:

Our dataset is from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database. This is a Canada-UK project that looks at a cohort of 1905 women for targeted sequencing data.[3] The dataset contains clinical attributes of individuals with breast cancer. Some examples of this are age_at_diagnosis, cellularity, chemotherapy, and her2_status. It also contains genetic attributes of these patients including m-RNA level z-scores for 331 genes and mutations for 175 genes. mRNA or messenger RNA acts as a code for DNA to be transcribed off of. Based on the documentation of this dataset, $(\frac{expression\ in\ tumor\ sample - mean\ expression\ in\ normal\ sample}{standard\ deviation\ of\ expression\ in\ normal\ sample} = mRNA\ z\ score)$ which means that the data is already normalized.[5] Essentially, these scores tell us if a gene is upregulated or downregulated relative to the expression of normal tissue.

According to the American Cancer Society, breast cancer accounts for 1 in 3 of all new female cancers each year. Additionally, incidence rates have increased by 0.6% per year.[4] Although genetic links in breast cancer have been thoroughly studied, we want to further understand correlations between clinical variables and expression of various genes. Further study in this area can aid in early detection, risk assessment, and treatment selection. This is a very robust dataset that we have confidence in to be able to draw meaningful conclusions.

Given that this is a very large dataset though, we need to be able to cast a wide net to find the correlation between expression of certain genes and clinical variables. We want to use our methods to validate already established correlations, while also discovering new, less studied ones. Additionally, given how complex of a disease breast cancer is, it is difficult to completely isolate variables without taking others into account. Our plan is to look at combinations of variables side by side in order to perform effective analysis and draw conclusions on our hypotheses.

We looked at both clinical variables and gene expression for our EDA. The distribution of patient age at diagnosis, overall survival in months, tumor size, and mutation counts were looked at by creating a histogram for each variable and looking at the descriptive statistics tables. This was helpful to better understand patient characteristics in this dataset.

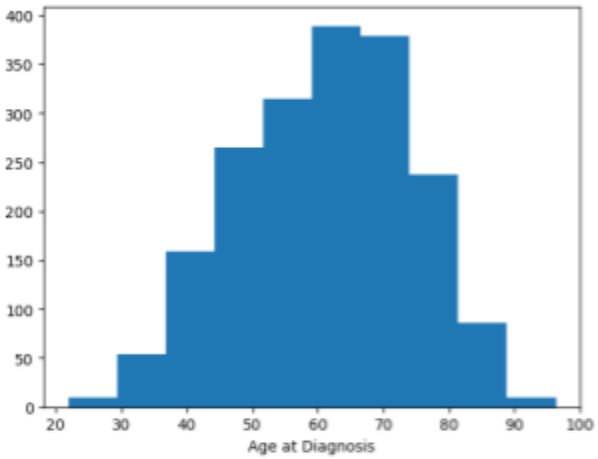

Figure 1. Histogram of age at diagnosis of breast cancer patients in the METABRIC dataset.

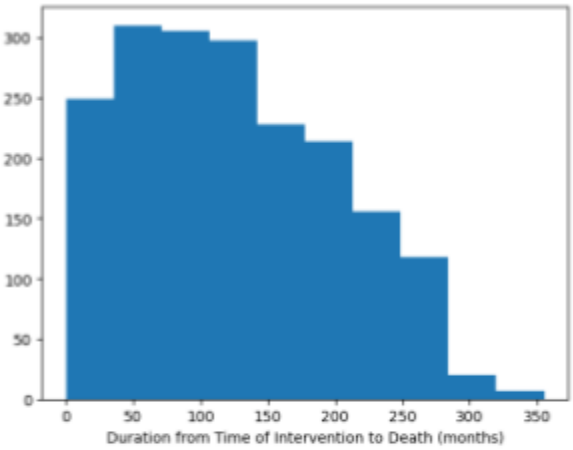

Figure 2. Histogram of duration from time of intervention to death in months of breast cancer patients in the METABRIC dataset.

The average age at diagnosis is 61 years. The distribution is bell-curve shaped with many patients being diagnosed between 45 and 80 years (Figure 1). The distribution of the survival time is right skewed, with most values being lower. Most patients survive between 0 and 150 months following intervention. It is important to consider whether this death is caused by the cancer itself or some other reason which can be further explored (Figure 2).

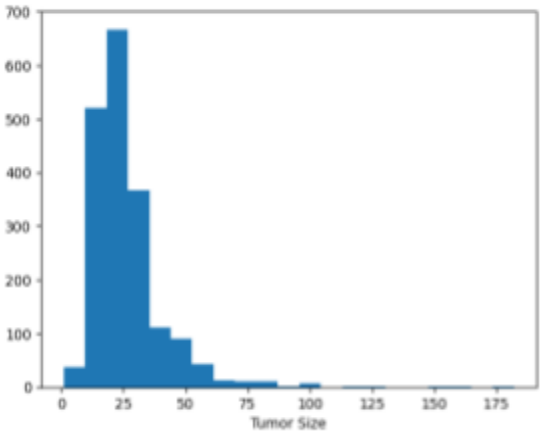

Figure 3. Histogram of tumor size of breast cancer patients in the METABRIC dataset.

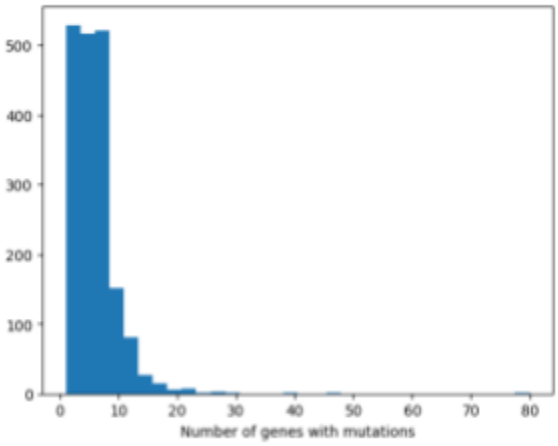

Figure 4. Histogram of number of genes with mutations of breast cancer patients in the METABRIC dataset.

The size of the tumor is on average around 26 (Figure 3). Most tumors are around this size, but the distribution is right skewed, so there are some unusual abnormally large tumors. Additionally, Most patients have mutations in between 1 and 7 genes, but a few patients have a very large number of mutated genes (Figure 4).

Two histograms were created of BRCA2 expression and CDH1 expression, two genes that play an important role in developing breast cancer.[7]

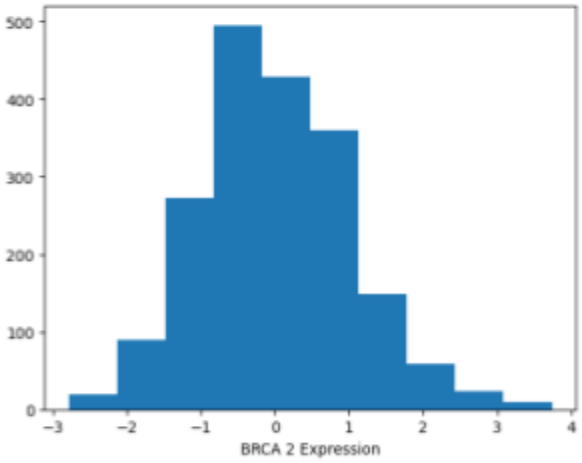

Figure 5. Histogram of BRCA2 gene expression in breast cancer patients in the METABRIC dataset.

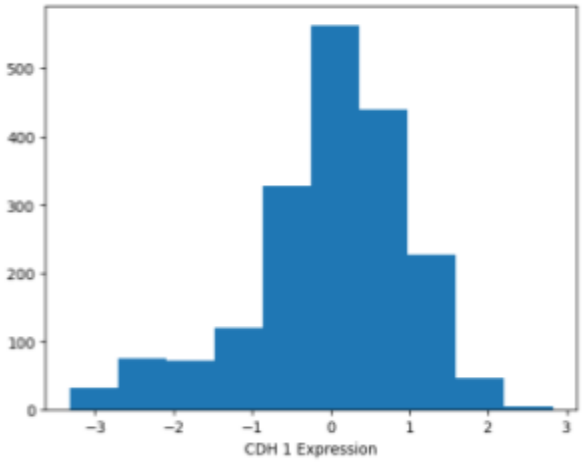

Figure 6. Histogram of CDH1 gene expression in breast cancer patients in the METABRIC dataset.

BRCA 2 expression appears to be relatively normally distributed which makes sense since these values are z-scored (Figure 5). It could be beneficial to also look at raw values to better understand how expression values are actually spread. CDH1 expression is somewhat left skewed even after the z-score normalization (Figure 6). It could be helpful to also look at raw CDH1 expression values to better understand the distribution of these values and how they affect cancer prognosis.

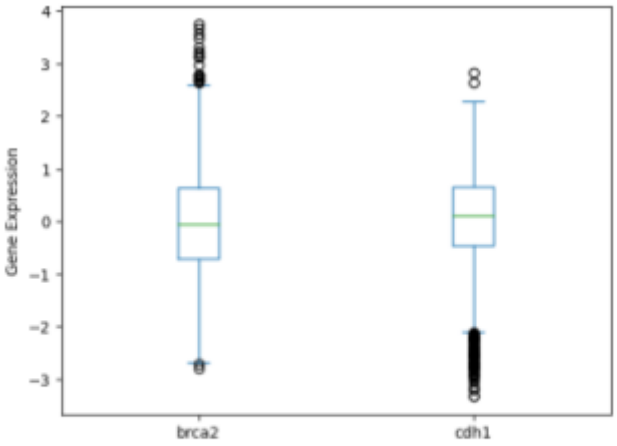

Figure 7. Boxplot of BRCA2 and CDH1 gene expression in breast cancer patients in the METABRIC dataset.

The boxplots for the z-scored BRCA2 and CDH1 expression are centered around 0 which is expected for z-scores (Figure 7). However, BRCA2 expression has a somewhat larger spread of expected values in the center box compared to CDH1, so BRCA2 expression is expected to have somewhat more variability.

The relationship between these two gene expressions and survival in months was studied using scatter plots. There does not appear to be a very clear relationship between BRCA2 expression and survival in months, likely because breast cancer is heterogeneous and can present very differently (Figure 8). Thus, we need to consider how we can better segment the data to find relationships between gene expression and cancer prognosis. Similarly, there does not appear to be a very clear relationship between CDH1 expression and survival in months, due to the

heterogeneity of breast cancer cases (Figure 9). Thus, we once again need to consider how we can better segment the data to find relationships between CDH1 expression and cancer development.

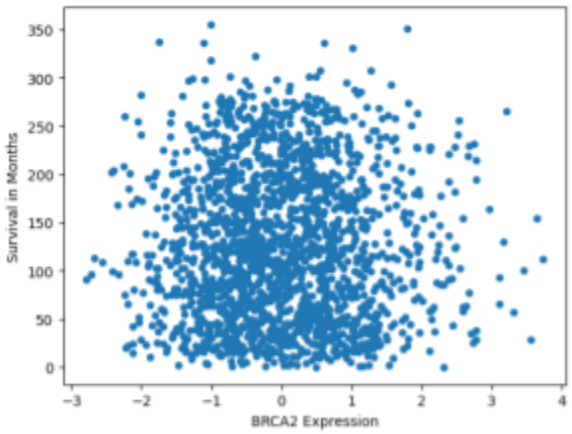

Figure 8. Scatterplot of BRCA2 expression in breast cancer patients in the METABRIC dataset.

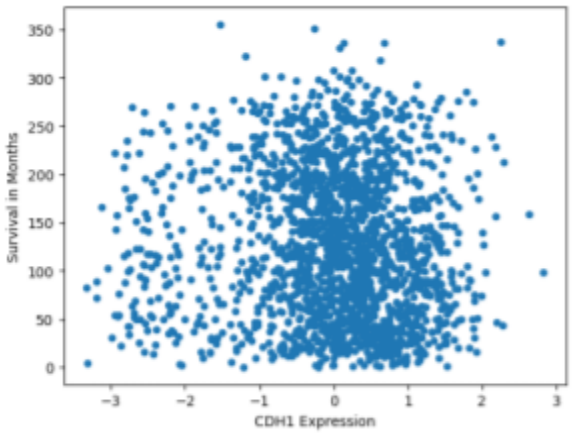

Figure 9. Scatterplot of CDH1 expression in breast cancer patients in the METABRIC dataset.

Finally, a bar plot of positive or negative status of ER, PR, and HER2 was made (Figure 10). It is important to consider the presence of clinical markers like ER, PR, and HER2 since these can guide breast cancer treatment protocols in seeing what treatment options patients are eligible for. It can be seen that most cases are ER positive, most cases are HER2 negative, and around half of cases are PR positive. It will be interesting to use these markers to better split up data points to consider relationships between various genes and cancer prognosis.

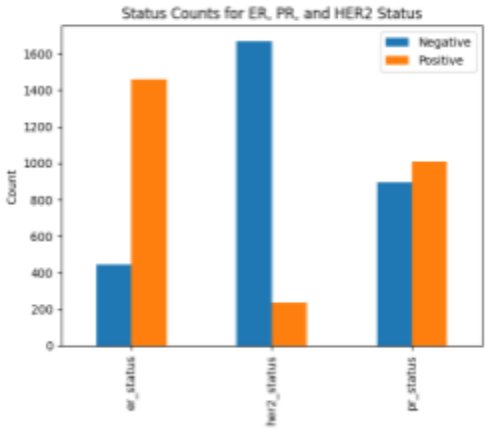

Figure 10. Histogram of status of ER, PR, and HER2 in breast cancer patients in the METABRIC dataset.

4. Methods:

Using the METABRIC database, an observation in our study is a given patient with breast cancer in our data set. Attributes include clinical variables like age at diagnosis, cellularity, chemotherapy, and ER, PR, and HER2 status, z-scores for gene expression, and survival time in months.

We will do supervised learning to make predictions or decisions based on labeled training data. In this case, we know the response variable we want is survival in months, and we have a multitude of known input data points. We hope to train a model that can depict the relationship between these clinical and gene attributes and survival in months. Since we have data on patients' clinical attributes as well as their survival in months, we already understand the relationship and want to create a model to represent this known relationship and predict outcomes in the future. Thus, in this case, we are doing a regression analysis because we have a

numerical output which is survival in months. We want our model to be able to use the input data about a patient in order to predict their survival in months.

We plan to use partial least squares regression (PLSR) for our analysis. Our predictor variable (X) will be gene expression data and clinical attributes like age at diagnosis, treatment method, and hormone receptor (ER, PR, HER2) characteristics. Our response variable (Y) will be survival in months. Since gene expression data is already z-scored, other forms of data cleaning can be conducted if needed like logarithmic conversion and handling of outliers. PLSR can be conducted using the corresponding algorithm in Python using many components and all gene data. The percentage of variance explained for Y versus the number of components can be plotted. Some threshold can be established for the amount of variance we want to be explained, and the number of components needed to explain this degree of variance can be determined. This cross-validation will be used to determine the optimal number of components. Based upon this value, we will repeat PLSR using this determined number of components. The weights obtained from the PLSR model can be normalized to ensure magnitudes do not influence VIP scores disproportionately. Then, VIP scores can be calculated to rank genes based on their importance, and genes with a VIP score greater than 1 will be selected for use in the model due to greater importance in predicting survival time. PLSR will be conducted once again, but this time using the determined number of components and the genes with high VIP scores. Then, PLSR results can be used to fit a model representing predicted survival time based on gene expression data. The conclusions made from the modeling will be visualized on multiple plots with the individual genes and their expression that were found to be most predictive of survival against the survival in months of the patients.

To know if the model developed using PLSR is accurate, the observed survival time can be plotted against the predicted survival time, and a line of best fit can be created to visualize the model's performance. The R^2 value will be computed to quantify how well the model fits the data and determine how effectively gene expression and clinical variables predict survival time. Leave-one-out validation can also be performed by removing one patient at a time from data sets, predicting survival time, and then recomputing R^2 in order to determine if any patients are disproportionately affecting the output of the model. These R^2 will be visualized in a matrix of different leave-on-out validation simulations.

Potential issues are that our input data does not only contain one type of data, like just gene expression. Instead, there are a variety of other variables that will be included in addition to gene expression like age at diagnosis, treatment options, and receptor status. Thus, it is unclear whether our PLSR procedure will be able to handle these varied types of variables when predicting survival time. If there are issues with data compatibility in the input data, we can try to normalize the clinical attributes differently, or handle feature selection differently for clinical attributes and gene expression before creating the model. Worst case, we can choose to just use gene expression data to create our model. Another issue is the high dimensionality of gene expression data resulting in a model with a poor R^2. To remedy this issue, we can perform further cleaning or grouping of gene data through an algorithm like principal component analysis (PCA)

to reduce the dimensionality of the data. Given that the gene expression data are numerical variables that are highly correlated, using a PCA might be necessary for more accurate outcomes. If our approach fails, we will learn that the provided input or data set may just not provide any meaningful predictive value. We will also learn more about the importance of feature selection to produce meaningful results with true predictive power through exploring new methods to choose what variables will be important for our model.

In order to prepare the data specifically for our analysis, we want to ensure that the data is normalized and scaled correctly. The gene expression data is already Z-scored, so we want to ensure that the clinical variables are standardized. For example, we want to make sure that categorical variables are converted into binary format when appropriate. We will apply one-hot encoding for categorical clinical attributes like ER, PR, and HER2 status (hormone receptor status).

5. Results:

Using the METABRIC dataset, which contains clinical attributes and gene Z-scores of patients with breast cancer, we wanted to create a model that could predict a patient's overall survival in months. Our prediction question was: Can we predict a breast cancer patient's overall survival time in months based on clinical attributes, gene Z-scores, and receptor statuses using partial least squares regression (PLSR)? This could be used by healthcare professionals to identify critical predictors or estimate survival times to ultimately guide treatment decisions. We decided to use PLSR to analyze the data since we want to model a response variable (overall survival in months) given a large number of predictor variables that are highly correlated. The predictor variables are a mix of 498 numerical features like age_at_diagnosis, tumor_size, mutation count, gene expression data, and dummy variables like er_status, pr_status, and her2_status. Since the receptor statuses are categorical variables, they were converted into binary columns for our analysis. The data was split into an 80/20 training/testing set.

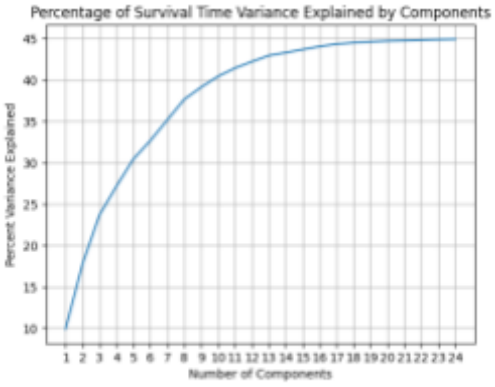

Figure 11. Percentage of response variance explained by the number of components in the PLSR

We chose to use 13 components for our PLSR model. The PLSR model constructs these new predictor variables, or components, as linear combinations of the original predictor variables.[10] Cross validation was used to evaluate the percentage of variance in survival time that was explained when a given number of components was used. As shown by Figure 15, as more components were added, the percentage of variance explained increases. We selected 13 components since this is where the graph started leveling off more steadily with 42.9% variance explained, and adding more components would not significantly improve the model's ability to explain the target variable. PLSR was conducted with 13 components using the training data to construct a model.

We then wanted to determine the Variable Importance in Projection (VIP) scores of the model. These scores represent the importance of each feature or predictor variable in the components created by the model. Now only using the 13 components and the 211 features calculated to have a VIP score above 1, the PLSR model was retrained. Figure 12 shows features with the highest VIP scores, with age_of_diagnosis seen to be the most important feature in the components and the pdpk1 gene seen to be the most important gene feature.

Index	VIP Score
age_at_diagnosis	2.54055
tumor_size	2.310454
pdpk1	1.753507
sf3b1	1.645189
hsd3b7	1.555044
acvr1b	1.532486
ep300	1.532028
zfp36l1	1.529072
gsk3b	1.52183

Figure 12. Highest VIP scores of features of PLSR model.

Using the retrained model, the predicted versus observed survival time in months was plotted on the test data as shown in Figure 13. The model is very inaccurate in predicting the test data. The R^2 is only 0.01, so the model only explains 1% of the variance in the target variable while the rest of the variance is unexplained by the model. As shown by the graph, there is no clear relationship between observed survival time and survival time as predicted by the model. The predicted versus observed survival time was then plotted on the training dataset as shown in Figure 14. The model performs better on the data that it is trained on as indicated by the R^2 of 0.32 which is greater. However, the performance is still not great and predictive power is limited.

Figure 13. Scatter plot of predicted vs observed survival time on testing dataset. R^2: 0.0104. MSE: 0.971.

Figure 14. Scatter plot of predicted vs observed survival time on training dataset. R^2: 0.325. MSE: 0.678.

The model does not do well on predicting survival time on unseen data. This is likely due to the complexity of breast cancer and the factors that contribute to it. The relationship between features and survival time may be too complicated for the PLSR model to capture well. Since the model performed poorly on training and testing data, this indicates underfitting which means the model is too simplistic to capture the underlying patterns in the response data. Thus, using the

predictor variables chosen, we were not effectively able to predict a breast cancer patient's overall survival time based on data for clinical attributes, gene expression, and receptor statuses.

6. Conclusion:

The findings from the analysis demonstrate that within this data set, a PLSR model created using gene Z-scores, receptor statuses for ER, PR, and HER2, and clinical attributes including age at diagnosis, tumor size, and mutation count is not able to effectively predict survival time. This is demonstrated by the poor performance of the model on both training and testing data as the R^2 values were 0.325 and 0.010 respectively. However, the VIP score analysis demonstrated that age at diagnosis and tumor size as well as gene expression data for genes including pdpk1, sf3b1, and hsd3b7 were the most important features in the model since they had the highest VIP scores. Thus, more investigation on these genes and how they contribute to cancer could be beneficial, and this could guide future genetic research.

Since the model performed poorly on both testing and training data, this indicates issues with the model created. The final PLSR model was created using 13 components and 211 features. This is still a very high level of dimensionality. The sample size was likely too low for this high dimensionality of data since the training data contained 1471 patients and the testing data contained 368 patients. Thus, this large number of variables could not be well fit to the training data in the first place. Further feature selection and gene filtering could help to create a more accurate model. Although a VIP score of 1 is typically used as the threshold for an important variable, the VIP score threshold for a relevant feature could be raised to include fewer genes. In addition, even with the 13 components used, only 42.9% of variance in the response variable was explained, and this did not increase much with the inclusion of more components. Thus, regardless of the number of components selected, the data selected was not effectively reflecting variability in the response variable and thus could not demonstrate much predictive power.

Another potential issue was that the three clinical variables included in the model which were age at diagnosis, tumor size, and mutation count were selected somewhat arbitrarily based on preconceptions regarding factors influencing breast cancer development. Other clinical variables like type of surgery, cancer type detailed, cellularity, and tumor stage were dropped. The rationale for this was that we wanted to evaluate whether survival time could be predicted based on variables that can be obtained prior to treatment. However, some of the clinical variables dropped could have still been predictive, so it was unreasonable to drop so many of them. Future analysis can be done to evaluate which of the clinical variables demonstrate the highest degree of variability and thus could be beneficial to include in the model.

Data preprocessing is also a very important step in analysis that could have been neglected. In our initial Exploratory Data Analysis, we looked at many variables that were included in the model. However, since the data set was so large, many of the variables were not carefully examined. Thus, there could have been outliers affecting the input data. Although

numerical variables were z-scored to normalize them and this likely does not explain the poor model performance, it is still an important consideration.

The inability to effectively predict survival time is likely because breast cancer is a highly heterogeneous disease, so using this particular combination of variables is not very predictive. In addition, the use of survival time as a response variable is not necessarily the best choice due to the lack of segmentation of data based on significant factors like cancer stage or type of surgery. Also, based on the data set, many of the patients are indicated to still be "Living" so the survival time data provided is likely not even accurate for the life-long survival time for patients since this is unknown for many of the patients. For accurate information, the data would have to be cleaned to only include patients who are known to have died. Since there are many confounding variables that could be influencing survival time, another response variable could provide a more meaningful analysis. For example, it could be interesting to apply a similar modeling method in order to predict how gene expression data influences tumor size so that clinicians can determine whether genetic information is indicative of tumor growth.

Furthermore, different regression models could be evaluated for their predictive ability in the future. PLSR constructs components as linear combinations of the original predictor variables.[10] Thus, it assumes that there are linear relationships between variables. However, for a disease as complex as breast cancer, it is unlikely that the relationships observed will be linear. As a result, PLSR struggled to capture meaningful patterns. It could be valuable to conduct feature selection and dimensionality reduction and then explore different types of regressions to evaluate whether any other patterns better fit the data.

Least Absolute Shrinkage and Selection Operator (LASSO) is an alternative approach that could be explored in the future. LASSO is often used to handle high dimensional data which is ideal for this data set since it performs variable selection and regularization to reduce dimensionality. While PLSR creates components from all features, LASSO identifies a sparse set of predictors which is helpful since it eliminates irrelevant features. Thus, this could potentially provide a more interpretable model and highlight critical genes and clinical features. After using LASSO for feature selection, other more complex models can be applied to investigate non-linear relationships between these relevant features and the response variable.

Overall, by more deeply investigating and analyzing the data through these alternative approaches, we can move towards answering the research question of how we can predict a breast cancer patient's survival time or another response variable based on data from the METABRIC dataset. This would be highly variable since breast cancer is a highly prevalent disease that will affect 1 in 8 women in the United States in their lifetime.[2] If clinicians are able to use known data about a patient to predict some aspect about the prognosis of the patient's breast cancer, this can be very valuable for guiding future treatment decisions. By conducting further analysis on the data through more careful response variable selection, feature selection, data preprocessing, and modeling method selection, a model with more predictive capability could ideally be built. This would allow for strides to be made towards better and more personalized breast cancer treatment.

7. References:

1. “Breast Cancer: Different Types, Different Treatments.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 31 Oct. 2024, www.mayoclinic.org/diseases-conditions/breast-cancer/in-depth/breast-cancer/art-20045654#:~:text=A%20breast%20cancer%20that’s%20estrogen,hormone%20therapy%20is%20endocrine%20therapy.
2. “Breast Cancer Facts and Statistics.” Breastcancer.Org - Breast Cancer Information and Support, Breastcancer.org, 14 Oct. 2024, www.breastcancer.org/facts-statistics.
3. “Breast Cancer Gene Expression Profiles (METABRIC).” *Kaggle*, www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric. Accessed 30 Sept. 2024.
4. “Breast Cancer Statistics: How Common Is Breast Cancer?” *Breast Cancer Statistics | How Common Is Breast Cancer? | American Cancer Society*, www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html. Accessed 30 Sept. 2024.
5. By, et al. “Z-Score: Definition, Formula, Calculation & Interpretation.” *Simply Psychology*, 6 Oct. 2023, www.simplypsychology.org/z-score.html.
6. DePolo, Jamie. “Test Results for BRCA1, BRCA2, and Other Gene Mutations.” Breastcancer.Org - Breast Cancer Information and Support, Breastcancer.org, 5 Dec. 2024, www.breastcancer.org/genetic-testing/getting-results.
7. “Inherited Cancer Risk: BRCA Mutation.” *Johns Hopkins Medicine*, 10 Oct. 2023, www.hopkinsmedicine.org/health/conditions-and-diseases/breast-cancer/inherited-cancer-risk-brca-mutation#:~:text=BRCA1%20and%20BRCA2%20account%20for,%2C%20PTEN%2C%20TP53%20and%20NF1.
8. Ke, Liyuan, et al. “The Prognostic Role of Tumor Mutation Burden on Survival of Breast Cancer: A Systematic Review and Meta-Analysis - BMC Cancer.” BioMed Central, BioMed Central, 17 Nov. 2022, bmccancer.biomedcentral.com/articles/10.1186/s12885-022-10284-1#:~:text=In%20a%20clinical%20study%20on,85.8%20months)%20%5B13%5D.
9. McGuire, Andrew, et al. “Effects of Age on the Detection and Management of Breast Cancer.” Cancers, U.S. National Library of Medicine, 22 May 2015,

pmc.ncbi.nlm.nih.gov/articles/PMC4491690/#:~:text=Notably%2C%20breast%20cancer%20survival%20is,70%20have%20the%20lowest%20survival.

10. “Partial Least Squares Regression and Principal Components Regression.” *MathWorks*, www.mathworks.com/help/stats/partial-least-squares-regression-and-principal-components-regression.html. Accessed 9 Dec. 2024.
11. Tumor Mutation Burden and Cancer Treatment | Oncology | JAMA Oncology | Jama Network, jamanetwork.com/journals/jamaoncology/fullarticle/2773840. Accessed 14 Dec. 2024
12. “Tumor Size.” Susan G. Komen®, 10 May 2024, www.komen.org/breast-cancer/diagnosis/factors-that-affect-prognosis/tumor-size/.