PM566/592 Final Project

Echo Tang

2022-12-13

# Introduction

Breast cancer is the most common malignant cancer in the world for women with over one million cases being diagnosed annually (Wang 2017). Although mortality rates for breast cancer are lower in the United States and Asia, other countries have unfortunately not seen the same trend, emphasizing the urgency of studying breast cancer and its associated risk factors (Azamjah et al. 2019). Like other cancers, breast cancer incidence increases with age with women above 50 years old being more likely to develop breast cancer than younger women. On the other hand, however, previous research has demonstrated that those who receive breast cancer diagnoses at younger ages also have higher reported mortality rates than those who are diagnosed later (McGuire et al., 2015). Previous studies have also shown that race may be a risk factor for breast cancer, as many women of color have higher mortality rates than non-Hispanic white women. Women of color are also associated with getting breast cancer diagnoses at later stages than white women, indicating that stage can confound the relationship between race and survival status (Ooi et al. 2011). Furthermore, women of color are less likely to have access to liquid biopsies as diagnosis methods, which have been seen to be associated with better survival outcomes compared to traditional diagnosis methods (Kim et al. 2017). Diagnosis methods, therefore, may also act as a confounder between vital status and race. Given that age and race are two risk factors, there is limited research done on how age at diagnosis can *interact* with stage of diagnosis and race specifically in breast cancer diagnosis, and how that in turn can contribute to and exacerbate these differences in mortality rates. Understanding differences between the clinical progression of breast cancer across these risk factors can contribute to more nuanced care that can potentially bridge inequities in survivability.

There has also been research suggesting that those with higher lymph node counts have lower survival rates of metastatic breast cancer for those in higher cancer stages (Yang et al. 2017). This also raises for concern on how lymph node counts, coupled with diagnosis method and cancer stage, can influence survival outcomes by interacting with stage of cancer diagnosis.

Given these risk factors, this study aims to address the question if race affects the age of diagnosis for breast cancer, if the effects of age and diagnosis stage on vital status differ by race, and if these relationships are complicated by lymph node count and diagnosis methods. To conduct the analysis, publicly available clinical data from The Cancer Genome Atlas (TCGA) was used, a cancer genomics program that has collected genomics, epigenomics, clinical, transcriptomic, and proteomic data of over 20,000 primary cancer samples across 33 different cancer types (TCGA, n.d.). For this study, only the clinical data for breast cancer patients were accessed to answer the research question. By analyzing the characteristics of breast cancer across multiple different demographics, we hope to gain a more holistic perspective of the disease.

# Methods

Breast cancer clinical data was accessed from TCGA using the R package TCGAbiolinks with the accession code “BRCA.” For data wrangling and cleaning, the clinical data was converted to a data table using the data.table package.

To prepare the data, missing racial data was imputed based on the most common value. Then, missing numerical data was imputed based on mean by sex. Reported substages of cancer diagnosis were standardized to the five stages I through X. Years survived after diagnosis was used as a metric for survivability and was calculated by taking the difference of year of death calculated from the variable days to death (or the current year 2022 for patients who are still alive) and year of cancer diagnosis. After data cleaning and wrangling, the resulting dataset contained the imputed categorical and numerical variables, standardized stage, and years survived after diagnosis for the remaining 1174 breast cancer patients.

Descriptive and summary statistics for variables of interest were generated and tabulated using the R package dplyr and standardized using the kable function from knitr. Data visualization and exploratory data analysis were done through the R package ggplot.

For logistic regression analysis, univariate analyses were first conducted for each variable of interest. Linearity assumptions per univariate model were tested using grouped smooth, LOESS, and fractional polynomial methods to determine variable coding. Continuous variables were chosen to be coded as categorical based on visual inspection of the grouped smooth method or transformed into polynomials based on fractional polynomials. To build a multivariate main effects model, confounding effects were found by adding potential confounding covariates into the model and calculating the percent change of parameter estimates; variables leading to a 10-20% chance in a parameter estimate were kept in the models as confounders. Interaction effects were also tested between each combination of variables, and interaction effects with a p value less than 0.05 were kept in the model to make a preliminary main effects model. Goodness of fit statistics (Hosmer-Lemeshow test and pseudo R-squared) and model diagnostics were run to check for model fit and influential points. After assessing goodness of fit and model diagnostics, the final model was generated keeping covariates and interactions that were significant with a p value less than or equal to 0.05.

There are 109 empty race values, which makes up around 9% of the dataset. These values were changed to “not reported” to examine if the relationship between survival and cancer diagnosis stage differ among those with missing race information and the other racial groups in this study.

Distribution of Race Among Breast Cancer Patients

| Race | Number of Individuals |
| --- | --- |
| AMERICAN INDIAN OR ALASKA NATIVE | 1 |
| ASIAN | 62 |
| BLACK OR AFRICAN AMERICAN | 201 |
| NOT REPORTED | 109 |
| WHITE | 801 |

White patients make up the vast majority of this dataset. Due to the distribution of this dataset, the patients were categorized into “white,” “not reported,” or “non-white.”

*Table 1: Distribution of White vs. Non-White Patients Among Breast Cancer Patients*

| Racial Group | Number of Individuals | Percentage |
| --- | --- | --- |
| Non-White | 264 | 22.49 |
| Not Reported | 109 | 9.28 |
| White | 801 | 68.23 |

After this grouping, there are 264 non-white patients, 109 missing race patients, and 801 white patients.

There are 111 empty values under diagnosis method, which make up 9% of the dataset. Because this is still under 10% of the data, it was imputed by the most common diagnosis method by gender.

Table 2: Distribution of Diagnosis Method Among Breast Cancer Patients

| Diagnosis Method | Number of Individuals | Percentage |
| --- | --- | --- |
| Core needle biopsy | 764 | 65.08 |
| Tumor resection | 171 | 14.57 |
| Fine needle aspiration biopsy | 103 | 8.77 |
| Other method, specify: | 68 | 5.79 |
| Excisional Biopsy | 29 | 2.47 |
| Cytology (e.g. Peritoneal or pleural fluid) | 22 | 1.87 |
| Incisional Biopsy | 17 | 1.45 |

Core needle biopsy remains the most common diagnosis method, followed by tumor resection and fine needle aspiration biopsy.

Diagnosis year and lymph node count contained missing values as well; this was imputed based on the average within the gender variable.

There are 112 patients who have passed away, making up about 9.5% of the dataset.

Table 3: Distribution of Vital Status

| Vital Status | Number of Patients | Percentage |
| --- | --- | --- |
| Alive | 1062 | 90.46 |
| Dead | 112 | 9.54 |

After data wrangling and cleaning, the resulting dataset consists of 1174 observations for 116 variables, of which include newly imputed categorical variables and the years survived after diagnosis variable.

Table 3: Distribution of Cancer Stage Among Breast Cancer Patients

| Stage of Diagnosis | Number of Patients | Percentage |
| --- | --- | --- |
| Stage II | 660 | 56.22 |
| Stage III | 268 | 22.83 |
| Stage I | 198 | 16.87 |
| Stage X | 27 | 2.30 |
| Stage IV | 21 | 1.79 |

Most patients were diagnosed at Stage II with 660 patients being diagnosed at this stage. This is followed by Stage III with 268 patients and Stage I with 198 patients. Stage IV and X diagnoses make up a very small number of diagnoses in this dataset.

Table 4: Summary Statistics for Diagnosis Age and Lymph Node Count

|  | Mean | Standard Deviation | Median |
| --- | --- | --- | --- |
| Age at Initial Pathologic Diagnosis (years) | 58.77087 | 13.257994 | 59 |
| Lymph node count | 10.39608 | 8.144937 | 10 |

# The mean age of diagnosis was 58.77 years old with a standard deviation of 13.26; the median was very similar at 59 years of age. The mean lymph node count was 10.40 lymph nodes with a standard deviation of 8.14; the median was also very similar at 10 lymph nodes.

# Exploratory Data Analysis

Chart, box and whisker chart

Description automatically generated

There does not seem to be much difference in the distribution of diagnosis age among non-white and white patients. White patients seem to have a bigger range in diagnosis ages with a lower third quartile, and higher mean and median diagnosis age. Among non-white patients, the mean diagnosis age is a little higher than the median diagnosis age. However, the patients with missing racial data have higher means and medians than both non-white and white patients.

Chart, box and whisker chart

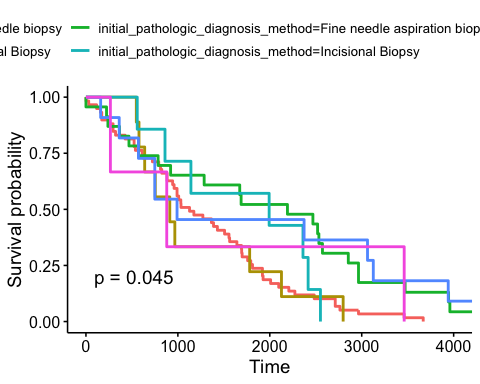
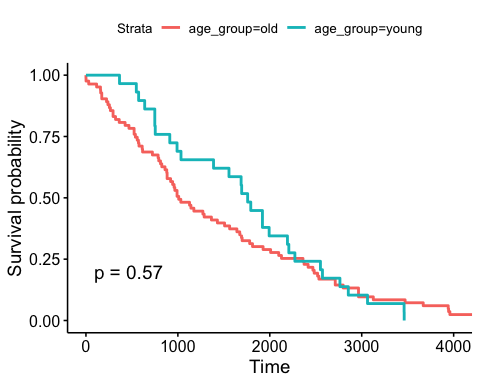
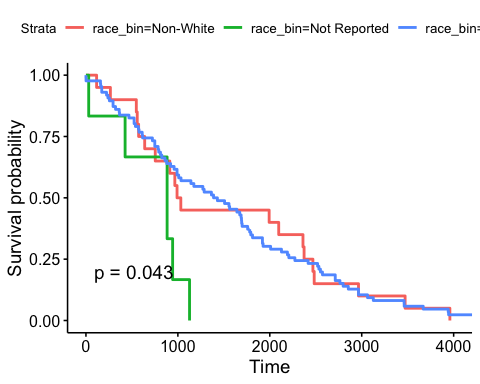
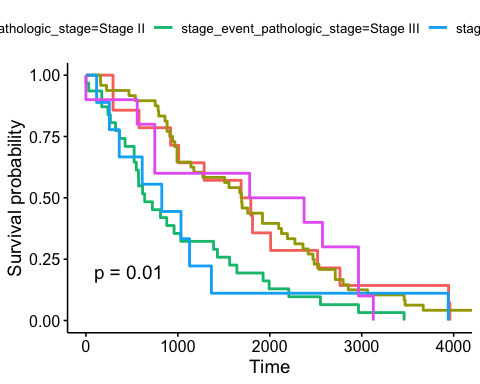
Description automatically generated

There are more non-white patients who have passed away with younger diagnosis ages than white and non-white patients; however, the mean and median age of diagnosis among white and non-white patients who have passed away are still quite similar. Regardless of vital status, patients with missing race information have higher ages of diagnosis with the mean age of diagnosis being much higher than the median for those who have passed away.

## 

## Kaplan Meier Plots

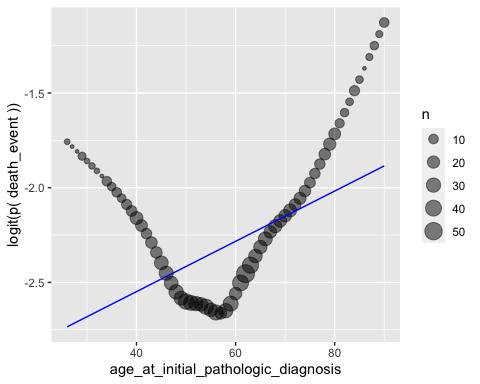
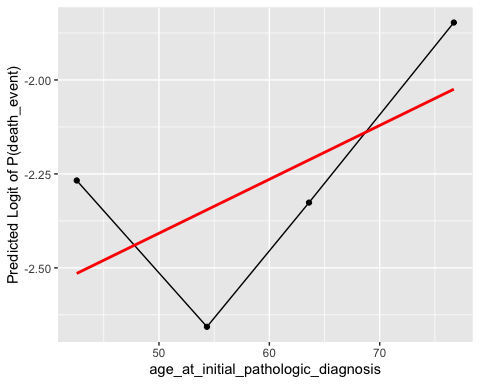
Kaplan-Meier curves were graphed to see if a Cox model was appropriate for this analysis.



None of the variables of interest were seen to have proportional hazards. Therefore, a logistic regression model was chosen, as the outcome is binary.

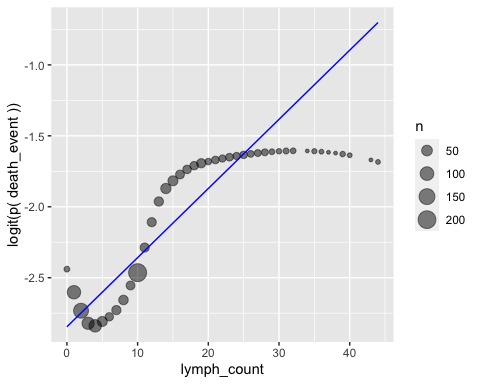
# Logistic Regression Univariate Analysis

Missing race patients have exp(-0.3415) = 0.711 times the odds of death than non-white patients; however, this value was not significant. White patients have exp(0.3835) = 1.47 times the odds of death than non-white patients; however, this value was also not significant.



Fractional polynomials suggests that the best transformation for age is a cubic transformation; this was validated by visual inspection of the grouped smooth and LOESS methods, which suggested that a continuous coding of age violated linearity. After transformation, the coefficient estimate for age cubed is 1.602e-6 and statistically significant (p = 0.00774).

Patients diagnosed at stage II have exp(0.03035) = 1.03 times the odds of death of patients diagnosed at stage I; however, this value was not significant. Patients diagnosed at stage III have exp(0.54181) = 1.719 times the odds of death of patients diagnosed at stage I; however, this value was not significant. Patients diagnosed at stage IV have exp(2.2882) = 9.857 times the odds of death of patients diagnosed at stage I (p = 1.12e-5). Patients diagnosed at stage IV have exp(2.04525) = 7.731 times the odds of death of patients diagnosed at stage I (p = 2.52e-5).

Chart, line chart

Description automatically generated

Assessing linearity of lymph node count with fractional polynomials, the ideal transformation given was a complex polynomial combination. The LOESS plot also revealed that a continuous coding of the variable may violate linearity. Therefore, the grouped smooth method was also used to assess linearity and see if a categorical coding would be better at modeling lymph node count. In the grouped smooth method, it looks like a categorical coding of lymph node count is more appropriate; therefore, a model was refit with lymph nodes as categorical quartiles. Those with a lymph count of 3 to 10 have exp(0.4338) = 1.543 times the odds of death compared to those with a lymph count of 0 to 3; however, this value was not significant. Those with a lymph count of 10 to 15 have exp(1.0275) = 2.794 times the odds of death compared to those with a lymph count of 0 to 3 (p = 0.002362). Those with a lymph count of 15 to 44 have exp(1.0972) = 2.996 times the odds of death compared to those with a lymph count of 0 to 3 (p = 0.000444).

Excisional biopsies, fine needle aspiration biopsies, incisional biopsies, “other” methods, and tumor resections were the diagnosis methods that yielded statistically significantly different odds than core needle biopsies in univariate analyses. Excisional biopsies, fine needle aspiration biopsies, incisional biopsies, and “other” methods led to higher odds of death, whereas tumor resections led to lower odds of death in this univariate analysis. Those who were diagnosed with excisional biopsies have exp(1.6822) = 5.377 times the odds of death as those who were diagnosed with core needle biopsies (p = 7.17e-5). Those who were diagnosed with fine needle aspiration biopsies have exp(1.2341) = 3.435 times the odds of death as those who were diagnosed with core needle biopsies (p = 6.01e-6). Those who were diagnosed with incisional biopsies have 8.365 times the odds of death as those who were diagnosed with core needle biopsies (p = 3.24e-5). Those who were diagnosed with “other” methods have exp(0.8355) = 2.306 times the odds of death as those who were diagnosed with core needle biopsies (p = 0.019). Those who were diagnosed with tumor resections have exp(-1.5477) = 0.213 times the odds of death as those who were diagnosed with core needle biopsies (p = 0.0098).

## Summary Results of Univariate Analyses

 Table 5: Summary Results of Univariate Analyses

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Predictors | Odds Ratio | CI | p | Odds Ratio | CI | p | Odds Ratio | CI | p | Odds Ratio | CI | p | Odds Ratio | CI | p |
| Intercept | 0.08 | 0.05-0.13 | **<0.001** | 0.07 | 0.05-0.10 | **<0.001** | 0.08 | 0.04-0.13 | **<0.001** | 0.06 | 0.03-0.09 | **<0.001** | 0.08 | 0.06-0.11 | **<0.001** |
| Not Reported Race | 0.71 | 0.25-1.72 | 0.477 |  |  |  |  |  |  |  |  |  |  |  |  |
| White Race | 1.47 | 0.90-2.50 | 0.139 |  |  |  |  |  |  |  |  |  |  |  |  |
| Age at initial pathologic diagnosis ^3 |  |  |  | 1.00 | 1.00-1.00 | **0.007** |  |  |  |  |  |  |  |  |  |
| Stage II |  |  |  |  |  |  | 1.03 | 0.57-1.98 | 0.923 |  |  |  |  |  |  |
| Stage III |  |  |  |  |  |  | 1.72 | 0.90-3.42 | 0.108 |  |  |  |  |  |  |
| Stage IV |  |  |  |  |  |  | 9.86 | 3.51-27.60 | **<0.001** |  |  |  |  |  |  |
| Stage X |  |  |  |  |  |  | 7.73 | 2.95-20.10 | **<0.001** |  |  |  |  |  |  |
| Lymph count [4-10] |  |  |  |  |  |  |  |  |  | 1.54 | 0.85-2.91 | 0.164 |  |  |  |
| Lymph count [11-15] |  |  |  |  |  |  |  |  |  | 2.79 | 1.45-5.51 | **0.002** |  |  |  |
| Lymph count [16-44] |  |  |  |  |  |  |  |  |  | 3.00 | 1.65-5.67 | **<0.001** |  |  |  |
| Cytology |  |  |  |  |  |  |  |  |  |  |  |  | 0.00 | NA-2003559 | 0.978 |
| Excisional biopsy |  |  |  |  |  |  |  |  |  |  |  |  | 5.38 | 2.24-12.03 | **<0.001** |
| Fine needle aspiration biopsy |  |  |  |  |  |  |  |  |  |  |  |  | 3.44 | 1.98-5.80 | **<0.001** |
| Incisional biopsy |  |  |  |  |  |  |  |  |  |  |  |  | 8.36 | 2.94-22.59 | **<0.001** |
| Other method |  |  |  |  |  |  |  |  |  |  |  |  | 2.31 | 1.10-4.48 | **0.019** |
| Tumor resection |  |  |  |  |  |  |  |  |  |  |  |  | 0.21 | 0.05-0.59 | **0.010** |

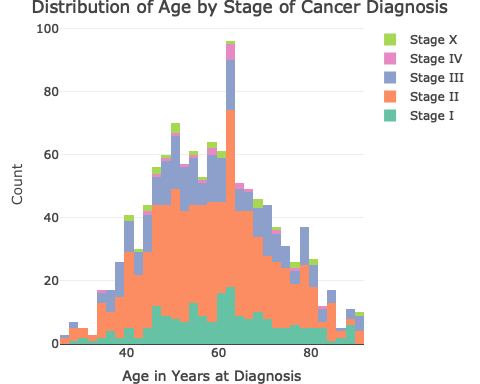
# Visualization of Age, Vital Status, and Race

Chart, box and whisker chart

Description automatically generated

There are more non-white patients who have passed away with younger diagnosis ages than white and non-white patients; however, the mean and median age of diagnosis among white and non-white patients who have passed away are still quite similar. Regardless of vital status, patients with issing race information have higher ages of diagnosis with the mean age of diagnosis being much higher than the median for those who have passed away.

The distributions for age by cancer stage was then examined.



The distributions of age look relatively normal for every cancer stage; from this stacked histogram, the distribution of age doesn’t seem to differ between stage at diagnosis.

# Checking for Confounding Effects

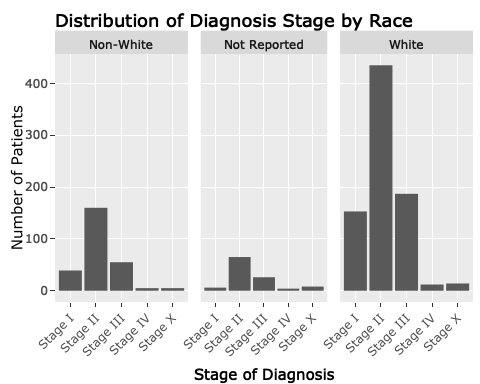
In a model containing both race and age, those with missing race information have 0.588 times the odds of death compared to non-white patients adjusting for age; however, this value was not significant. White patients are associated with 1.397 times the odds of death compared to non-white patients adjusting for age; however, this value was not significant. The parameter estimate for age cubed was 1.728e-6 (p = 0.00438).

Stage was then added into the model as a potential confounder. Stage changed the parameter estimate of age cubed by 7.9% and the parameter estimate of the not reported race patients by more than 10% (19.6%), so it was kept in the model as a confounder.

Diagnosis method was also examined as a confounder to see if the relationship between stage, race, and death event was confounded by diagnosis method.

Diagnosis method significantly changed the parameter estimate for Stage III (11%), Stage X (18%), as well as both parameter estimates for missing race information patients (46%) and white patients (56%), so it was kept in the model as a confounder.

# Examining Distributions of Stage by Race



From these bar graphs, it looks like the distribution of diagnosis stage also does not differ by race. However, as diagnosis stage is a confounder, there is evidence that it changes vital status based on racial group despite no significant differences in diagnosis stage’s distribution by race.

# Examining Years Survived After Diagnosis and Age of Diagnosis by Vital Status and Race

Company name

Description automatically generated

For all stages, those who have passed away have less years survived after diagnosis across all ages compared to those who are still alive. However, some relationships between years survived after diagnosis and age at diagnosis differ across stage; although there seems to be a slight negative relationship between years survived after diagnosis and diagnosis age for stages I, II, III, and X across vital status, there is a slightly positive relationship among those diagnosed at Stage IV across vital status. However, due to Stage IV diagnoses only making up a small percentage of all diagnoses in this dataset, this relationship is hard to confirm.

Chart

Description automatically generated with medium confidence

There seems to be very slightly negative relationship between years survived after diagnosis and age in stages I, II, and III across all racial groups. There is, however, a slight positive relationship between years survived after diagnosis and age for white and missing race patients in Stage IV and non-white patients in Stage X. This suggests that the relationship between years survived and age does differ to an extent by stage.

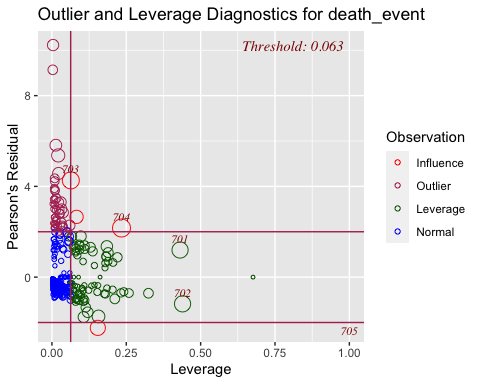
These visual differences suggest that race of a patient and stage of diagnosis may have interaction effects.

# Checking for Interaction Effects

Lastly, interaction effects were tested between stage and lymph node count, as well as between race and stage. The interaction between stage and lymph node count was significant among patients diagnosed at Stage X for the third and fourth lymph count quartiles; this suggests that the relationship between odds of death and stage differ by lymph node count.

Race was removed from the model because its effects did not remain significant.

## Model Diagnostics and Influential Points



There seems to be a few points that may be influential. One point has very high dBhat and dChisq values; however, the probability for death event for this covariate pattern was 6.39e-8 and the only person with this covariate pattern survived. This suggests that the actual outcome for this covariate pattern is not far off from the predicted. On the other hand, the covariate pattern with the next highest dBhat value shows that the probability of death event for this individual was 0.218, but the only individual in this covariate pattern passed away. Similarly, the two covariate patterns with the next highest dbHat values predicted that death event were not as likely as survival, but the only patients in those covariate patterns also passed away.

The patient who passed away was a 44 year old Stage I patient with a low lymph count and diagnosed with fine needle aspiration biopsy; the patient passed away with a low stage of diagnosis and lymph count, than expected. On the other hand, the patient who survived was a 64 year old Stage I patient with the lowest lymph count and diagnosed with core needle biopsy. This corresponds to the patient with the highest dbHat value; however, the probability of death event was very small for this patient and this patient has survived, indicating that the actual and predicted values are not that far off.

## Goodness of Fit Statistics

There is no evidence for a lack of fit for the final model (p = 0.4425). The McFadden pseudo-R2 is 0.1867, suggesting 18.67% of the variance in vital status is explained by the model.

# Final Model

*Table 6: Summary Results of Final Model*

|  |  |  |  |
| --- | --- | --- | --- |
| Predictors | Log Odds | CI | p |
| Intercept | -3.77 | -4.89—-2.85 | **<0.001** |
| Stage II | -0.32 | -1.56-0.97 | 0.613 |
| Stage III | 0.66 | -2.38-2.69 | 0.577 |
| Stage IV | -12.27 | NA-160.48 | 0.993 |
| Stage X | 5.05 | 2.59-8.38 | **<0.001** |
| Age at Initial Pathologic Diagnosis ^3 | 0.00 | 0.00-0.00 | **0.001** |
| Lymph count [4-10] | 0.28 | -1.18-1.68 | 0.692 |
| Lymph count [11-15] | 1.69 | -0.42-3.47 | 0.076 |
| Lymph count [16-44] | 2.67 | 0.71-4.67 | **0.006** |
| Cytology | -13.70 | -156.63-12.46 | 0.978 |
| Excisional biopsy | 1.58 | 0.61-2.48 | **0.001** |
| Fine needle aspiration biopsy | 1.13 | 0.54-1.70 | **<0.001** |
| Incisional biopsy | 2.23 | 1.14-3.29 | **<0.001** |
| Other method | 0.30 | -0.62-1.11 | 0.492 |
| Tumor resection | -1.73 | -3.25—-0.64 | **0.007** |
| Stage II \* lymph count [4-10] | 0.32 | -1.37-2.09 | 0.716 |
| Stage III \* lymph count [4-10] | -0.24 | -2.66-3.02 | 0.862 |
| Stage IV \* lymph count [4-10] | 14.83 | -155.65-NA | 0.991 |
| Stage X \* lymph count [4-10] | -3.44 | -7.09—-0.45 | **0.034** |
| Stage II \* lymph count [11-15] | -0.38 | -2.45-1.95 | 0.725 |
| Stage III \* lymph count [11-15] | -1.48 | -4.22-2.02 | 0.329 |
| Stage IV \* lymph count [11-15] | 14.64 | -153.45-NA | 0.991 |
| Stage X \* lymph count [11-15] | -5.40 | -9.76—-1.73 | **0.006** |
| Stage II \* lymph count [16-44] | -1.18 | -3.39-1.01 | 0.283 |
| Stage III \* lymph count [16-44] | -2.12 | -4.87-1.30 | 0.153 |
| Stage IV \* lymph count [16-44] | -3.63 | -40.51-46.42 | 0.999 |
| Stage X \* lymph count [16-44] | -6.41 | -10.96—-2.54 | **0.002** |

The final model included cubic age at diagnosis, diagnosis method, diagnosis stage, and lymph node count as covariates (with stage and lymph node count having interaction effects). The odds ratio for the cubic age term was exp(2.191e-6) = 1.00 (p = 0.000927).

Diagnosis methods leading to significant differences in odds of death were excisional biopsy, fine needle aspiration biopsy, incisional biopsy, and tumor resection (adjusting for stage, age, and lymph node count). Adjusting for all other covariates, patients diagnosed with excisional biopsies had exp(1.582) = 4.86 times the odds of death compared to those diagnosed with core needle biopsies (p = 0.000775); those diagnosed with fine needle aspiration biopsies had exp(1.134) = 3.11 times the odds of death compared to those diagnosed with core needle biopsies (p = 0.000119); those diagnosed with incisional biopsies had exp(2.235) = 9.35 times the odds of death compared to those diagnosed with core needle biopsies (p = 3.46e-5); and those diagnosed with tumor resection had exp(-1.735) = 0.634 times the odds of death compared to those diagnosed with core needle biopsies (p = 0.007012).

Stage X patients were the only stage that had significantly different odds of death compared to Stage I patients adjusting for all other covariates; Stage X patients had exp(5.05) = 156.02 times the odds of death compared to Stage I patients (p = 0.000288). Furthermore, the highest lymph node count quartile was the only quartile with significantly different odds of death compared to the lowest quartile adjusting for all over covariates; patients with lymph node counts between 15 and 44 had exp(2.672) = 14.47 times the odds of death compared to the lowest quartile (0.006455).

However, the relationship between lymph node count changed depending on stage. In particular, Stage X patients for all lymph node count quartiles showed significantly different odds of death compared to Stage I lowest quartile patients adjusting for age and diagnosis method. Stage X patients who had lymph node counts between 3 and 10 had exp(-3.442) = 0.032 times the odds of death compared to Stage I patients with the lowest lymph node counts adjusting for age and diagnosis method. Stage X patients with lymph node counts between 10 and 15 had exp(-5.401) = 0.0045 times the odds of death compared to Stage I patients with the lowest lymph node counts adjusting for age and diagnosis method. Stage X patients with lymph node counts between 15 and 44 had exp(-6.410) = 0.0016 times the odds of death compared to Stage I patients with the lowest lymph node counts adjusting for age and diagnosis method.

## Summary Visualization

Chart, box and whisker chart

Description automatically generated

The summary graph suggests that white patients who have survived have on average live longer than non-white patients who have survived; on the other hand, the average years survived after diagnosis for those have passed away do not differ much between non-white and white patients. However, the mean years survived after diagnosis for the missing race patients who have survived is slightly higher than non-white and white-patients, and the mean years survived after diagnosis for those who have passed away are lower than non-white and white patients. Furthermore, the range of years survived after diagnosis for both white and non-white patients are much larger than those who have missing race information. Given the relationships between years survived after diagnosis and age of diagnosis for stage IV and X for non-reported and white patients, it is likely that missing race patients are the most similar to white patients; however, because of similarities in age of diagnosis across vital status and distributions in diagnosis stage for non-white and white patients in this dataset, this cannot be confirmed.

# Conclusion

From this analysis, there were no significant differences found in diagnosis age among racial groups across vital status, nor were there significant differences found in the relationship between years survived after diagnosis and diagnosis age stratified by race for patients diagnosed in Stages I, II, and III.

A logistic regression analysis revealed that race does not have a significant effect on vital status in breast cancer, contrary to previous research; however, age at diagnosis, stage, diagnosis method, and lymph node counts are associated with vital status. Stage X patients had significantly higher odds of death compared to Stage I patients adjusting for age, diagnosis method, and lymph count; and patients with lymph node counts between 15 and 44 also had higher odds of death compared to patients with 0 and 3 lymph nodes, adjusting for age, diagnosis method, and stage. However, interaction effects between diagnosis stage and lymph node counts were also significant with Stage X patients having lower odds of death across lymph node count quartiles, which suggest that being diagnosed at Stage X many have a protective effect for patients with high lymph node counts, adjusting for age and diagnosis method.

However, due to the limited number of patients in both Stage IV and X compared to the other stages, which both had less than 30 patients, future analysis should be done with a larger sample size to determine if these interactions and main effects still hold.

# Addenda

The link to my code can be found here: <https://github.com/eztang12/final_project_pm566/blob/main/final_proj.Rmd>

Code to download data is in the code. Note that the package TCGAbiolinks needs to be installed for data download.

Data needs to be re-queried and loaded for every new R session; however, data only needs to be downloaded once. If this is your first time downloading this data, please uncomment the line GDCDownload.

# References

Azamjah, N., Soltan-Zadeh, Y., & Zayeri, F. (2019). Global trend of breast cancer mortality rate: A 25-year study. Asian Pacific Journal of Cancer Prevention, 20(7), 2015–2020.

Kim, M. K., Park, H. S., Kim, J. Y., Kim, S., Nam, S., Park, S., & Kim, S. I. (2017). The clinical implication of the number of lymph nodes harvested during sentinel lymph node biopsy and its effects on survival outcome in patients with node-negative breast cancer. The American Journal of Surgery, 214(4), 726–732.

McGuire, A., Brown, J., Malone, C., McLaughlin, R., & Kerin, M. (2015). Effects of age on the detection and management of breast cancer. Cancers, 7(2), 908–929.

National Institute of Health. (n.d.). The Cancer Genome Atlas Program. National Cancer Institute. Retrieved October 18, 2022, from https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

Ooi, S. L., Martinez, M. E., & Li, C. I. (2010). Disparities in breast cancer characteristics and outcomes by Race/Ethnicity. Breast Cancer Research and Treatment, 127(3), 729–738.

Wang, L. (2017). Early Diagnosis of Breast Cancer. Biosensors for Cancer Biomarkers, 17(7), 1572.

Yang, J., Long, Q., Li, H., Lv, Q., Tan, Q., & Yang, X. (2017). The value of positive lymph nodes ratio combined with negative lymph node count in prediction of breast cancer survival. Journal of Thoracic Disease, 9(6), 1531–1537. https://doi.org/10.21037/jtd.2017.05.30