PM566 Final Project Report

Echo Tang

2022-12-06

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, particularly Black and Native American women, 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). 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.

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. 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 (diagnosis year) 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. Implausible negative values for years survived were subsequently removed from the dataset. 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 1062 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`.

Table 1: Table 1: 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

There are 109 empty 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.

Table 2: Table 2: Distribution of White vs. Non-White Patients Among Breast Cancer Patients

Racial Group	Number of Individuals
Non-White	264
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", "non-white," or "not reported."

Summary statistics for variables of interest were subsequently generated. Summary statistics for cancer stage show that stages I, II, and III contain different substages on top of their overarching numerical stage; these were then standardized by numerical stage. Summary statistics for age in years shows that the mean and median ages are very similar both at around 58.5 years of age; the minimum age was 26 years of age, and the oldest patient was 90 years old at the time of diagnosis. Summary statistics for diagnosis year shows that most patients in the late 2000s with the last patient being diagnosed in 2013. Finally, summary statistics for years survived after diagnosis showed that the minimum years survived was 9 years after diagnosis, and the maximum was 34 years after diagnosis.

Table 3: Table 3: Distribution of Cancer Stage Among Breast Cancer Patients

Stage of Diagnosis	Number of Patients
Stage II	660
Stage III	268
Stage I	198
Stage X	27

Stage of Diagnosis	Number of Patients
Stage IV	21

Most patients were diagnosed at Stage II with 612 patients being diagnosed at this stage. This is followed by Stage III with 237 patients and Stage I with 184 patients. Stage IV and X diagnoses make up a very small number of diagnoses in this dataset.

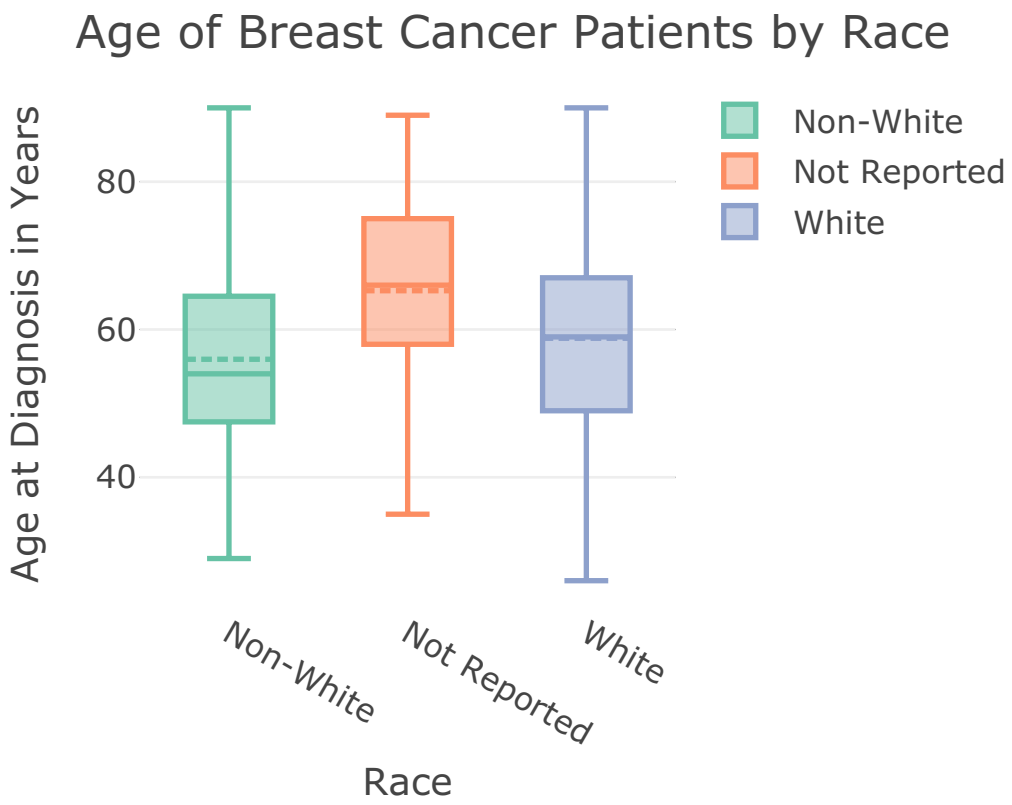
Table 4: Table 4: Vital Status Distribution

Vital Status	Number of Patients
Alive	1062
Dead	112

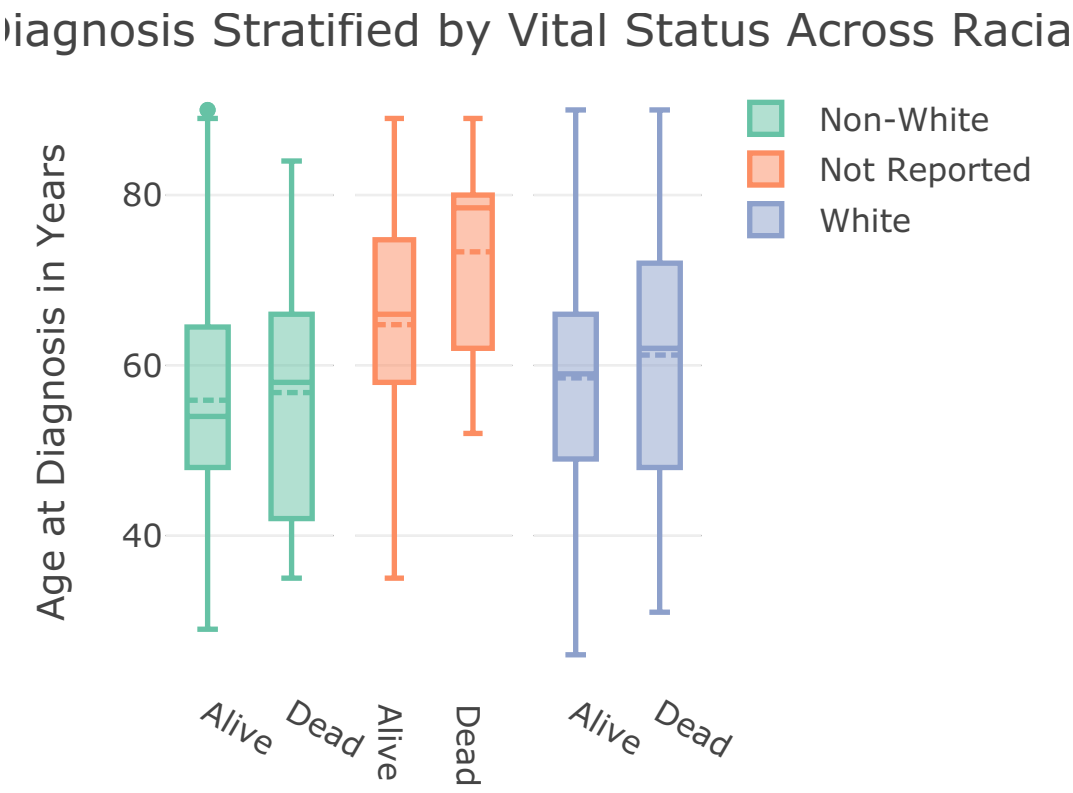
There are 112 patients who have passed away, making up about 9.5% of the dataset.

After data wrangling and cleaning, the resulting dataset consists of 1062 observations for 116 variables, of which include newly imputed categorical variables and the years survived after diagnosis variable.

Results

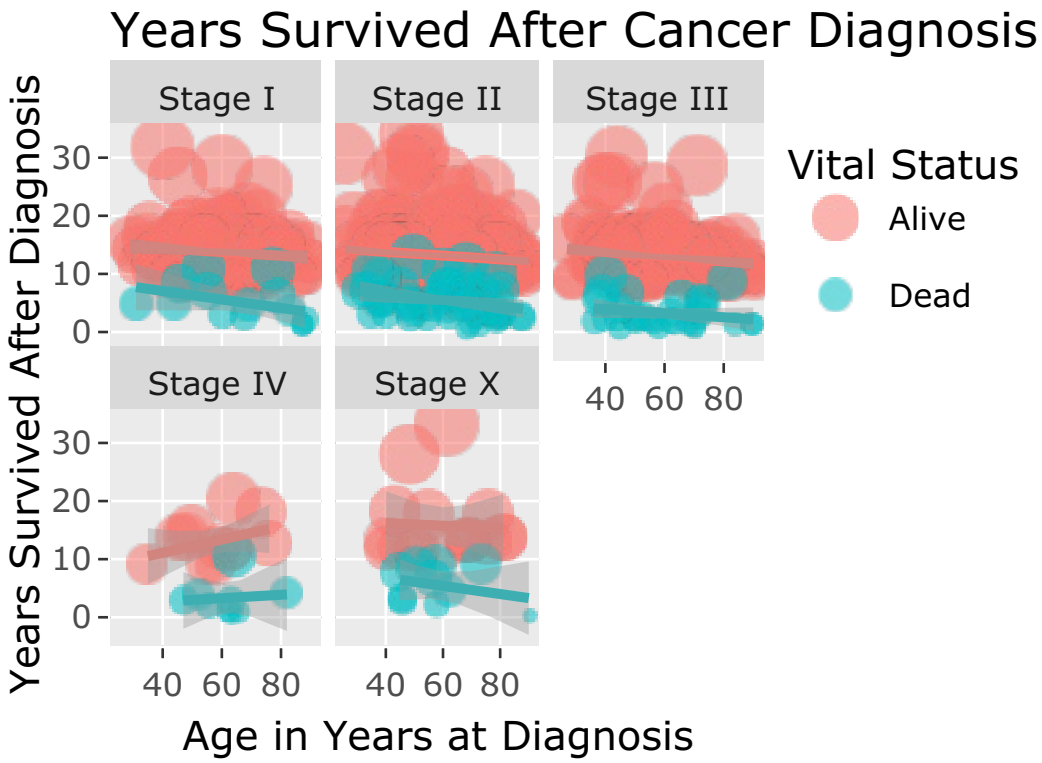


There does not seem to be much difference in the distribution of diagnosis age among non-white and white patients. White patients seem to have a bigger range in diagnosis ages with a lower third quartile, and higher mean and median diagnosis age. Among non-white patients, the mean diagnosis age is a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

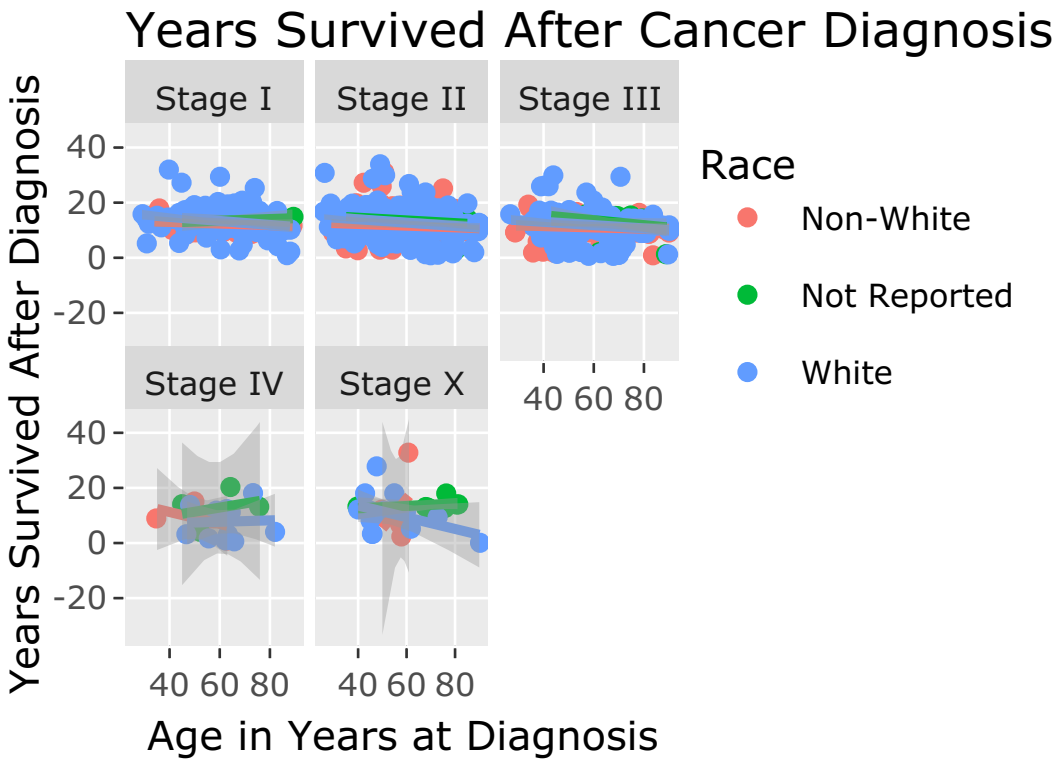


There are more non-white patients who have passed away with younger diagnosis ages than white and non-white patients; however, the mean and median age of diagnosis among white and non-white patients who have passed away are still quite similar. Regardless of vital status, patients with issing race information 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 of age are relatively normal for every cancer stage. The distribution of diagnosis stage also does not seem to differ much by race. See the "Introduction, Methods, and Summary Tables" page on the website for these distributions.

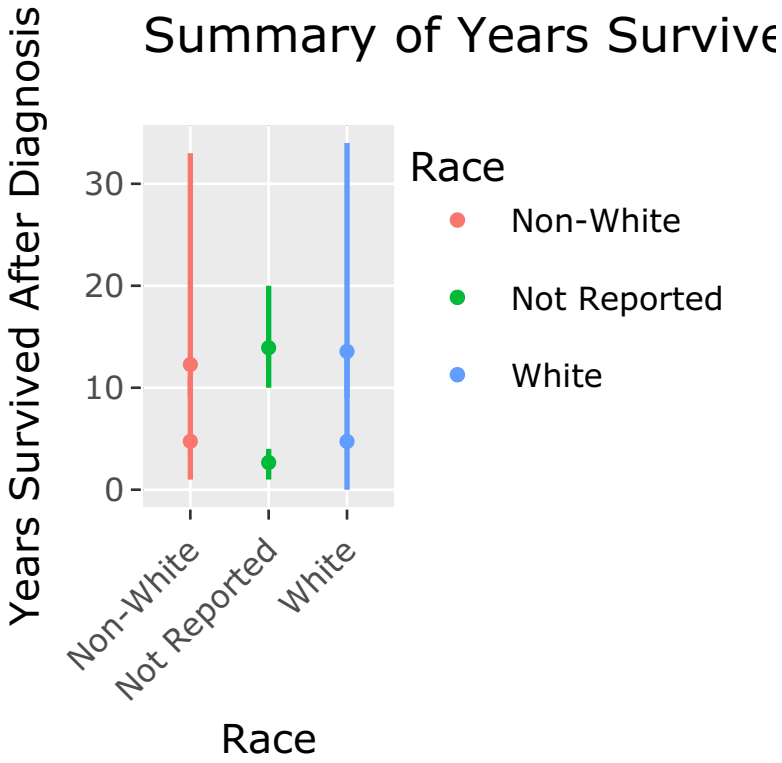


For all stages, those who have passed away have less years survived after diagnosis across all ages compared to those who are still alive. However, some relationships between years survived after diagnosis and age at diagnosis differ across stage; although there seems to be a slight negative relationship 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.



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; however, because Stage IV and Stage X diagnoses only accounted for around 20 patients each out of over 1000 patients in the entire dataset, a bigger sample size may be needed to confirm these relationships.



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, it was revealed that there were no significant differences between the distribution of diagnosis age or stage of diagnosis among white and non-white breast cancer patients. However, it was found that across vital status, the relationship between years survived after diagnosis for those who were diagnosed in Stage IV differs from other stages of diagnosis where there was a slightly positive relationship compared to other stages that have seen slightly negative relationships. Similarly, white and missing race patients saw a slightly positive relationship between years survived after diagnosis and age of diagnosis in Stage IV compared to other stages, where patients of all racial groups saw a slightly negative relationship. However, due to the limited number of patients in both Stage IV compared to the other stages, future analysis should be done with a larger sample size.

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.

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.