Survival analysis of patients diagnosed with breast cancer in Greater California Area using Cox's model

Bryan Lin, Lumi Huang, Landi Luo, Joowon Lee

Background/Objective:

About 1 in 8 U.S. women (about 12.4%) will develop invasive breast cancer over the course of her lifetime (Breast Cancer statistics, 2018). In 2018, an estimated 266,120 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., along with 63,960 new cases of non-invasive (in situ) breast cancer. Our project aims to examine the effects of several demographic and patient characteristics on breast cancer survival. In particular, we wish to focus on the effects of sex, race, and stage of cancer on patient survival.

Methods:

Sample:

We used the breast cancer data from the Surveillance, Epidemiology, and End Results Program (SEER) provided by the National Cancer Institute, which contains 590,646 cases and 133 variables. We limited our analyses to the Greater California area, which included the following California Cancer Registry regions: Central California, Sacramento, Tri-County, Desert Sierra, Northern California, San Diego/Imperial, and Orange County. The data were collected from years 2000 - 2015. There were 253,771 subjects after subsetting the data to include only the Greater California region. We selected 13 covariates of interest and dropped all observations with missing or unknown values for each covariate. In addition, we dropped all observations that had a survival time of 0. Our final sample included 55,333 subjects. Sample characteristics are summarized in Table 1.

Variables:

The variables used as predictors in the analysis are defined as follows: cancer stage (stage 0 to stage 4), sex (0 = male, 1 = female), race (1 = white, 2 = black, 3 = American Indian, Alaska Native, 4 = Asian or Pacific Islander), marital status at diagnosis (1 = single, 2 = married, 3 = separated, 4 = divorced, 5 = widowed, 6 = unmarried or domestic partner), breast subtype (1 = Her2+/HR+, 2 = Her2+/HR-, 3 = Her2-/HR+, 4 = Triple Negative), the site in which the primary tumor originated (0 = Nipple, 1 = Central portion of breast, 2 = Upper-inner quadrant of breast, 3 = Lower-inner quadrant of breast, 4 = Upper-outer quadrant of breast, 5 = Lower-outer quadrant of breast, 6 = Axillary tail of breast, 7 = Overlapping lesion of breast, 8 = Breast, NOS), Estrogen Receptor status (0 = positive, 1 = negative), Progesterone Receptor status (0 = positive, 1 = negative), and insurance status (0 = Uninsured, 1 = Any Medicaid, 2 = Insured, 3 = Insured/No specifics). The continuous variables include age, age at diagnosis, total number of in situ/malignant tumors for patient, and total number of benign/borderline tumors for patient.

The SEER breast cancer data included a variable for competing risks analysis; however, for our project we did not consider competing risks. Our censoring indicator was created by dropping subjects who died of other cause-specific classifications before the end of the study period: delta (0 = alive, 1 = dead due to breast cancer). Our event time indicator was survival time in months.

With our chosen 13 predictor variables, we used the LASSO, SCAD and MCP variable selection methods to select the best predictors. We then produced Kaplan-Meier curves, fit the Cox regression model, checked the proportional hazards assumption and ran model diagnostics.

Statistical Analysis:

55,333 patients from years 2000 to 2015 were included in the sample and were evaluated for the associations between gender and survival months after controlling for a number of other predictors. Of those, 94.5% subjects are right-censored and the mean survival time is 30 months. Characteristics of the study sample are shown in Table 1. The stages of cancer are labelled into five categories: 2.2% of patients are in stage 0 (no cancer), 48% of patients are in stage 1, 33.7% patients are in stage 2, 11.5% are in stage 3, and 4.7% patients are in stage 4. Most patients (99.4%) in our sample are female, while only 0.6% of patients are male. In term of race, 85.% of the subjects are white, 4.5% are black, 0.7% are American Indian or Alaska Native, and 9.3% are Asian or Pacific Islander. In terms of marital status, 15.0% of patients are single, 60% patients are married, 1.4 % are seperated, 11.4% are divorced, 11.9% are widowed, and 0.4% are unmarried or have a domestic partner. For breast cancer subtype, 11% are Her2+/HR+, 4.8% are Her2+/HR, 73.2% are Her2-/HR+ and 10.9% are triple negative. The mean age of patients at diagnosis is 60 years with IQR 51-69. The mean number of in situ/malignant tumors is 1.1 and the mean number of benign/borderline tumors is 0.005. 83.3% of patients have a positive estrogen receptor status and 73% have a positive progesterone receptor status.

In terms of model selection, we constructed penalized regressions that regressed survival time and censoring indicator on the 13 covariates of interest. The penalized regression techniques we used included LASSO, SCAD, and MCP. Table 2 shows the coefficients of the penalized regressions. After sorting the absolute value of the coefficients in descending order, we derived the best 8 predictors for each penalized regression. For example, for lasso regression, the best 8 predictors were stage, ER status, breast subtype, PR status, insurance, marital status at diagnosis, total number of in situ/malignant tumors for patient, and race. Combining the results from Lasso, SCAD, and MCP, we decided to fit a Cox model with 10 predictors: sex, stage, ER status, PR status, total number of in situ/malignant tumors for patient, race, the site in which the primary tumor originated, breast subtype, marital status, and insurance.

We plotted the product-limit (Kaplan-Meier) estimates and the cumulative hazard functions for breast cancer survival (Figure 1-2). In addition, we fitted Kaplan-Meier survival curves for breast cancer survival by sex, race, and cancer stage (see Figures 3-5). Using the Breslow method of handling ties, we fit a Cox's proportional hazards model to the data including the 10 covariates listed above. We assumed noninformative right-censoring (the censoring time was independent of the survival time, given our covariates). We constructed an Analysis of Variance table to summarize estimates of the risk coefficients and the results of the one degree-of-freedom tests for each covariate in the model. The coefficients, standard errors, relative risks, and 95% confidence intervals of the relative risk for each covariate were also calculated. The results are shown in Table 3.

Cox-Snell residual plots are used for assessing the overall fit of a Cox's model based on the Cox's proportional hazard model. For simplicity, we assumed that all covariates are fixed-time covariates. If the model is correct and the estimated regression coefficients are close to the true regression coefficients, then the Cox-Snell residuals follow a unit exponential distribution. Looking at Figure 6, the model is not adequate because the plot does not follow a 45 degree line. However, we also show stratification on fixed time covariates may be more appropriate.

Since our primary three covariates of interest were sex, cancer stage, and patient race, we assessed the proportional hazards assumption for these three variables to check whether a Cox model would be appropriate to use for our analyses. In order to test whether or not our Cox's model satisfied the proportional hazard assumption for sex (a fixed-time covariate), we created a time-dependent covariate, $Z2(t)$, defined as $Z2(t) = Z1 \times g(t)$, where $g(t)$ is a known function of the time t. In our case, we used $g(t) = \ln(t)$. A test of the null hypothesis that $\beta 2 = 0$ was performed. Using $g(t) = \ln(t)$, the Wald p-value for the test of $H0 : \beta 2 = 0$ was 0.0044, which is significant at $\alpha = 0.05$. Thus, there is evidence that the sex covariate has nonproportional hazards. We then fit a stratified proportional hazards model stratified on sex. A key assumption in using this stratified proportional hazards model is that all covariates are performing similarly on the baseline hazard function in each stratum. That is, we assume common β's in each stratum. In order to test whether or not this assumption is satisfied, we obtained the log partial likelihood using only data from each stratum and performed a likelihood ratio test. The likelihood ratio chi square for the test that the β's are the same in each stratum is 174.0883, which has a large-sample, chi-square distribution with 9 degree of freedom under the null hypothesis. The p-value is $<<0.0001$, so the assumption of using a stratified model is not met; the covariate effects are not the same between the two strata. We concluded that a stratified model is not appropriate.

We also used graphical checks of the proportional hazards assumption for gender, race and cancer stage. Specifically, we used the Andersen plot to check for proportional hazards (Figures 7-9) and expect the Andersen curves to be straight lines through the origin if the proportional hazards assumption is met. For these covariates, proportional hazards does not seem to be met. The Andersen plot for the cumulative hazard for Black vs. White is much higher among the other races. Likewise, in Figure 9, the cumulative hazard for stage 4 vs. stage 1 cancer is much higher than the other stages.

Using the Cox model, we tested whether there was an effect of each variable on survival using the Likelihood Ratio Test, Wald test, and Score test (see Tables 4-6). For example, for the cancer stage covariate, we wish to test the null hypothesis that there is no effect of cancer stage on survival. We can do this by testing the null hypothesis that all coefficients for the cancer stage covariate are equal to 0 using 4 degrees of freedom, since we have 4 dummy variables for the 5-level factor cancer stage (stages 0 - 4). Similarly, we tested the null hypothesis that there is no effect of race, sex, breast subtype, marital status, malignant tumor count, primary tumor origin site, estrogen receptor status, progesterone receptor status, and insurance status on survival using all three methods at a significance level = 0.05.

We also performed log-rank tests to compare whether the hazard rates were the same for the following variables: sex, stage, race, breast subtype, marital status, ER status, PR status, and insurance status. We used both the log-rank χ^2 test statistic (W(t) = 1) and weighted log-rank test using Gehan's test, which uses a weight function of $W(t_i) = n_i$, where n_i is the number of people at risk. The results are summarized in Table 7.

Results:

Table 3 shows the results from fitting the Cox proportional hazards model. In particular, we focused on the results of covariates cancer stage, race, and sex. The estimated relative risk of death for females as compared to males is 0.705, controlling for the other covariates; thus the risk of death is 29.5% lower for female as compared to male breast cancer subjects. The 95% confidence interval around the relative risk can be calculated as $\exp(-0.35 \pm 1.96 * 0.214) =$ (0.464, 1.072). Since the 95% confidence interval includes 1, we conclude that the relative risk of death is not significantly different for females versus males at an $\alpha = 0.05$ significance level. This result corresponds with the 1 degree of freedom p-value given in the table (p-value = 0.10).

The relative risk of death for subjects increases significantly as cancer stage increases from stage 1 - 4 compared to stage 0, controlling for the other covariates. The relative risk of death for subjects who are in stage 1 cancer as compared to stage 0 cancer is 6.40; the relative risk of death for subjects who are in stage 2 cancer as compared to stage 0 is 21.68; the relative risk for stage 3 cancer as compared to stage 0 is 65.63 and for stage 4 vs. stage 0 it is 341. All of

the relative risks comparing cancer stages are highly significant, suggesting that there is a significantly higher risk of death as cancer stage increases.

The relative risk of death is 21% higher for black subjects as compared to white subjects, controlling for the other covariates, and it is significantly higher for black subjects (p-value = 0.004). The relative risk of death for American Indians/Alaskan Natives as compared to white subjects is 1.07, and the risk of death is not significantly different between these two groups. The relative risk of death for Asians/Pacific Islanders compared to white subjects is 0.73, and it is significant suggesting that subjects of Asian/Pacific Islander race have significantly lower risk of death as compared to white subjects.

Among the other covariates, we noted that subjects with breast cancer subtype Her2+/HR- had a significantly lower risk of death as compared to Her2+/HR+ subjects, controlling for the other covariates. On the other hand, Her2-/HR+ subjects had a significantly higher relative risk as compared to Her2+/HR+ subjects. Married patients had a significantly lower risk of death (27.1% lower), and widowed patients had a significantly higher risk of death compared to single patients. The number of malignant tumors, ER status, and PR status were significantly associated with survival whereas primary site of tumor origin did not appear to be significantly associated with the survival, controlling for the other covariates.

The results of the likelihood ratio test, Score test, and Wald test for each covariate effect in the Cox model are summarized in Tables 4-6. Each table provides the χ^2 test statistics, degrees of freedom, and p-values. We can see that for all three tests, every covariate is highly significant at $\alpha = 0.05$, with the exception of sex. In particular, we are interested in the effects of the race, cancer stage, and sex covariates. Evidence suggests that there is significant difference in the hazard rates between subjects in different cancer stages and among subjects with different races. There is not a significant difference in the hazard rates between male and female breast cancer subjects, however.

From the results of the log-rank tests (Table 7), we have sufficient evidence to reject the null hypothesis that hazard rates of the groups divided by all variables are the same at a significance level of 0.05, except the sex variable. At significance level 0.05, we do not have enough evidence to reject the null hypothesis that the hazard rates between males versus females are the same (log-rank p-value = 0.24 and Gehan's p-value = 0.28).

We calculated 5 year survival rates for our covariates of interest. The survival rate of breast cancer patients is highly affected by cancer stage. Specifically, we compared the 5-year survival rate with a 95% confidence interval (Table 8). We found that only about 1/3 of the patients with stage IV breast cancer can survive more than 5 years. Since the maximum observed

death time of stage 0 in our data is 26 months, we did not include an estimate of the 5-year survival rate for stage 0 breast cancer. These results are similar with the official report from American Cancer Society.

The survival rates between males and females are not significantly different at a significance level of 0.05 (Table 9). However, since sex is a basic demographic variable, we decided to include covariate sex in our final model. Next, we looked into two types of hormone receptor-positive breast cancer, ER status and PR status. Cancer cells grow in response to the hormone estrogen and progesterone, respectively, so breast cancer has one or both of these hormone receptors. Even though hormone therapy is effective in treating breast cancer, it does not work on tumors that are hormone receptor negative. This is why patients with hormone receptor positive have higher 5-year survival rates (Table 10).

Breast cancer subtype is determined by whether some cancer cells have hormones receptors as well as an overexpression of the human epidermal growth factor receptor 2 (HER2) gene. Triple negative is defined as a cancer which is not receptive to ER and PR hormones and does not have overexpression of HER2. It is known that triple negative breast cancer consist of 15 to 20 percent of all breast cancer. Our data have 10.9 percent of triple negative patients. Triple negative breast cancer cells do not respond well to hormone therapy or medications that block HER2 receptors. A previous study found that 77 percent of women with triple negative breast cancer survived at least five years, which is similar to our results (Table 11).

We also examined the survival rates according to race variable. Asian or Pacific Islander has the best 5-year survival rate, on the other hand, black patients have the worst 5-year survival rate. A previous study found that the highest mortality rates are black ethnic group, followed by whites and then American Indian/Alaska Native. The lowest mortality rate is for Asian/Pacific Islanders, which is similar to our results (Table 12).

We further examined the survival rates by the site in which the primary tumor originated. We found that there is significant different between the survival rates depending on the tumor origins (Table 13). Invasive ductal carcinoma, not otherwise specified (NOS) has the lowest 5-year survival rate, whereas patients with tumors from upper-inner quadrant have the highest 5-year survival rate.

Among marital status at diagnosis, married patients have the best 5-year survival rate, whereas widowed patients have the worst 5-year rate. This result (Table 14) suggests the importance of social relationships and how social relationships support in the treatment and affect the survival of breast cancer.

We found that survival rates can also be affected by the total number of in situ/malignant tumors in a patient (Table 15). Patients whose in situ/malignant tumors are less than 1 have the better 5-year survival rate. Finally, we looked into insurance information. As can be seen in Table 16, patients without insurance have the lowest 5-year survival rate, except category Other. Patients who get Medicaid service have the best 5-year survival rate.

Conclusion:

Our three main covariates of interest on predicting breast cancer survival are sex, race and stages of cancer. We found sex is not a significant predictor in our model, which may be because there are very few male breast cancer patients (0.6%) in our dataset. However, the Kaplan Meier curves (Figure 3) show that the estimated survival of breast cancer patients are different. Specifically, male breast cancer patients have had worse survival outcomes than female patients. Men are less likely to get breast cancer, as breast cancer is usually detected as a hard lump underneath the nipple and areola. They are less likely to assume a lump under their nipple is breast cancer, so awareness of breast cancer among men is low. Therefore, this can cause a delay in seeking medical care.

We found that race is a significant predictor of survival in our model. The Kaplan Meier curves (Figure 4) for race show that Black people have the lowest survival. This may be because of racial inequities in healthcare, as well as inherent health disparities among races. However, both Cox's proportional hazards assumption and stratified Cox's model assumption were not met for the race covariate based on our graphical checks. Similarly, we found stages of cancer is a significant predictor of survival, with stage IV patients having significantly lower survival. This makes sense, the cells in stage IV cancer patients have metastasized and the chances of survival are low. However, Cox's proportional hazards assumption is not met. Therefore, we need to be cautious in interpreting the results from our model.

Limitations:

There are some limitations to our analyses. First, since we performed a complete-case analysis using only 13 of the original 133 variables, our final sample only included 21.8% of the Greater California breast cancer data. Thus, our sample may not be representative of the entire Greater California breast cancer population and the results may not be generalizable. These issues may be addressed in future analyses by better handling of the missing data. Second, the Cox's proportional hazards assumptions were not met for our three primary covariates of interest: cancer stage, race, and sex. Therefore, a Cox model may not be appropriate for our analyses. In addition, 94.5% of our data were right-censored and only 0.6% of our sample were male breast cancer subjects. Furthermore, Cox's proportional hazards assumption and the stratified Cox's model assumption is not met for sex, one of our primary covariates of interest.

We can consider fitting a stratified Cox's model with an "uncommon" β, which might provide a better solution to our research questions. Another option to consider is to fit an accelerated failure time (AFT) model with an exponential distribution because our cumulative hazard function looks linear from Figure 2. Furthermore, in future analyses, we may consider analyzing the data including competing risks.

REFERENCES

1. U.S. Breast Cancer Statistics Web-based Report. Ardmore, PA. Available at: http://www.cdc.gov/uscs.
2. https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html#references
3. National Breast Cancer Foundation (https://www.nationalbreastcancer.org/)
4. Laura, M. (2017). Types of Breast Cancer, WebMD Medical Reference
5. Pritchard, K. (2013). Adjuvant endocrine therapy for non-metastatic, hormone receptor-positive breast cancer.
6. Bauer, K. R., Brown, M., Cress, R. D., Parise, C. A., & Caggiano, V. (2007). Descriptive analysis of estrogen receptor (ER)‐negative, progesterone receptor (PR)‐negative, and HER2‐negative invasive breast cancer, the so‐called triple‐negative phenotype: a population‐based study from the California cancer Registry. *Cancer*, *109*(9), 1721-1728.
7. Sinn, H. P., & Kreipe, H. (2013). A brief overview of the WHO classification of breast tumors. *Breast Care*, *8*(2), 149-154.
8. Daly, B., & Olopade, O. I. (2015). Race, ethnicity, and the diagnosis of breast cancer. *JAMA*, *313*(2), 141-142.

TABLES

Table 1. Sample Characteristics of Breast Cancer Patients in Greater California Area (N = 55,333)

Characteristic	Variable Type	Variable Description
stage	Categorical	Cancer Stage 0 Stage 0: 1200 patients (2.2%) 1 Stage 1: 26555 patients (48.0%) 2 Stage 2: 18643 patients (33.7%) 3 Stage 3: 6348 patients (11.5%) 4 Stage 4: 2587 patients (4.7%)
SEX	Categorical	Sex 0 Male: 343 patients (0.6%) 1 Female: 54990 patients (99.4%)
RAC_RECY	Categorical	Race 1 white: 47283 patients (85.5%) 2 black: 2474 patients (4.5%) 3 American Indian, Alaska Native: 403 patients (0.7%) 4 Asian or Pacific Islander: 5173 patients (9.3%)
MAR_STAT	Categorical	Marital status at diagnosis 1 Single: 8276 patients (15.0%) 2 Married: 33172 patients (60.0%) 3 Separated: 762 patients (1.4%) 4 Divorced: 6302 patients (11.4%) 5 Widowed: 6610 patients (11.9%) 6 Unmarried or domestic partner: 211 patients (0.4%)
BRST_SUB	Categorical	Breast Subtype 1 Her2+/HR+: 6095 patients (11.0%) 2 Her2+/HR-: 2673 patients (4.8%) 3 Her2-/HR+: 40529 patients (73.2%) 4 Triple Negative: 6036 patients (10.9%)
AGE_DX	Continuous	Age at diagnosis (IQR) 60 (51, 69) year
MALIGCOUNT	Continuous	Total number of in situ/malignant tumors for patient (Mean = 1.069)
BENBORDCOUNT	Continuous	Total number of benign/borderline tumors for patient (Mean = 0.004536)

Age	Continuous	Age at the end of the study
PRIMSITE	Categorical	The site in which the primary tumor originated 0 Nipple: 235 patients (0.4%) 1 Central portion of breast: 2587 patients (4.7%) 2 Upper-inner quadrant of breast: 6695 patients (12.1%) 3 Lower-inner quadrant of breast: 2928 patients (5.3%) 4 Upper-outer quadrant of breast: 18806 patients (34%) 5 Lower-outer quadrant of breast: 4192 patients (7.6%) 6 Axillary tail of breast: 245 patients (0.4%) 7 Overlapping lesion of breast: 13401 patients (24.2%) 8 Breast, NOS: 6247 patients (11.3%)
ERSTATUS	Categorical	Estrogen Receptor status 0 Positive: 46093 patients (83.3%) 1 Negative: 9240 patients (16.7%)
PRSTATUS	Categorical	Progesterone Receptor status 0 Positive: 40410 patients (73.0%) 1 Negative: 14923 patients (27%)
INSREC_PUB	Categorical	INSURANCE RECODE 0 Uninsured: 470 patients (0.8%) 1 Any Medicaid: 7819 patients (14.1 %) 2 Insured: 39783 patients (71.9%) 3 Insured/No specifics: 7261 patients (13.1%)
SRV_TIME_MON	Continuous Outcome	Survival months survival time (IQR) 30 (14, 48) months
delta	Binary Outcome	Death Indicator 0 Censored 52300 patients (94.5%) 1 Uncensored 3033 patients (5.4%)

Table 2. Coefficients of Penalized Regression using Lasso, SCAD and MCP Variable Selection Methods

	lasso.est	scad.est	mcp.est
stage1	0.0000	0.0000	0.0000
stage2	1.2482	1.3133	1.3119
stage3	2.3907	2.4524	2.4507
stage4	4.0271	4.0854	4.0869
RAC_RECY2	0.2236	0.2294	0.2368
RAC_RECY3	0.0190	0.0000	0.0000
RAC_RECY4	-0.2371	-0.2623	-0.1627
SEX	-0.1887	-0.1980	0.0000
BRST_SUB2	-0.8124	-0.8340	-0.8386
BRST_SUB3	0.1451	0.1944	0.1982
BRST_SUB4	0.0000	0.0000	0.0000
AGE_DX	0.0245	0.0252	0.0257
Age	0.0000	0.0000	0.0000
MAR_STAT2	-0.3598	-0.3705	-0.3942
MAR_STAT3	-0.0823	-0.0684	0.0000
MAR_STAT4	-0.1268	-0.1523	-0.1342
MAR_STAT5	0.0895	0.0816	0.0000
MAR_STAT6	0.0000	0.0000	0.0000
MALIGCOUNT	0.2404	0.2493	0.2513
BENBORDCOUNT	0.0277	0.0000	0.0000
PRIMSITE1	-0.0674	0.0000	0.0000
PRIMSITE2	0.0000	0.0832	0.0000
PRIMSITE3	0.0885	0.2284	0.0620
PRIMSITE4	-0.1129	0.0000	0.0000
PRIMSITE5	-0.1289	0.0000	0.0000
PRIMSITE6	0.0000	0.0000	0.0000
PRIMSITE7	0.0497	0.1731	0.1044
PRIMSITE8	0.1799	0.2967	0.2576
ERSTATUS	0.9324	0.9724	0.9796
PRSTATUS	0.6222	0.6380	0.6352
INSREC_PUB1	0.0000	0.0000	0.1879
INSREC_PUB2	-0.3647	-0.3850	-0.1742
INSREC_PUB3	-0.2232	-0.2553	0.0000

Table 3. Cox PH Model Summary

Covariate	Coefficient	Std. Error	Relative Risk	Lower 95% CI	Upper 95% CI	P-Value
factor:(SEX)1	-0.3498104	0.2137943	0.7048217	0.4635506	1.0716706	0.1017976
factor:(stage)1	1.8567879	0.7097021	6.4031362	1.5932840	25.7331098	0.0088891
factor:(stage)2	3.0762746	0.7081560	21.6774947	5.4103535	86.8545430	0.0000140
factor:(stage)3	4.1840138	0.7081035	65.6287428	16.3815621	262.9255904	0.0000000
factor:(stage)4	5.8319125	0.7079154	341.0102520	85.1508065	1365.6710583	0.0000000
factor:(RAC_RECY)2	0.1910369	0.0669042	1.2105041	1.0617379	1.3801148	0.0042985
factor:(RAC_RECY)3	0.0673499	0.1975770	1.0696697	0.7262261	1.5755330	0.7331945
factor:(RAC_RECY)4	-0.3123954	0.0772355	0.7316922	0.6289055	0.8512780	0.0000524
factor:(BRST_SUB)2	-0.9653283	0.1523746	0.3808581	0.2825282	0.5134105	0.0000000
factor:(BRST_SUB)3	0.2512626	0.0639033	1.2856476	1.1342986	1.4571912	0.0000843
factor:(BRST_SUB)4	-0.1036935	0.1423817	0.9015015	0.6819789	1.1916864	0.4664430
factor:(MAR_STAT)2	-0.3157273	0.0494999	0.7292583	0.6618307	0.8035553	0.0000000
factor:(MAR_STAT)3	-0.0995482	0.1469079	0.9052464	0.6787636	1.2072995	0.4980108
factor:(MAR_STAT)4	-0.0089949	0.0643074	0.9910455	0.8736853	1.1241704	0.8887604
factor:(MAR_STAT)5	0.5158527	0.0579821	1.6750663	1.4951250	1.8766639	0.0000000
factor:(MAR_STAT)6	-0.1266944	0.3361255	0.8810029	0.4558971	1.7025030	0.7062288
MALIGCOUNT	0.2880071	0.0548764	1.3337668	1.1977578	1.4852201	0.0000002
factor:(PRIMSITE)1	-0.2458512	0.2558266	0.7820386	0.4736617	1.2911840	0.3365485
factor:(PRIMSITE)2	-0.2446495	0.2545332	0.7829789	0.4754348	1.2894638	0.3364671
factor:(PRIMSITE)3	-0.1206211	0.2595897	0.8863697	0.5329074	1.4742734	0.6421744
factor:(PRIMSITE)4	-0.3639285	0.2484321	0.6949409	0.4270533	1.1308725	0.1429479
factor:(PRIMSITE)5	-0.3709508	0.2569803	0.6900779	0.4170192	1.1419317	0.1488799
factor:(PRIMSITE)6	-0.1753562	0.3367633	0.8391580	0.4337009	1.6236679	0.6025684
factor:(PRIMSITE)7	-0.1856234	0.2486804	0.8305863	0.5101615	1.3522652	0.4554054
factor:(PRIMSITE)8	-0.0446436	0.2484893	0.9563382	0.5876206	1.5564172	0.8574194
factor:(ERSTATUS)1	1.0333016	0.1286672	2.8103290	2.1839137	3.6164198	0.0000000
factor:(PRSTATUS)1	0.6786592	0.0530830	1.9712329	1.7764526	2.1873701	0.0000000
factor:(INSREC_PUB)1	-0.1079585	0.1284547	0.8976648	0.6978682	1.1546624	0.4006614
factor:(INSREC_PUB)2	-0.3572593	0.1261293	0.6995910	0.5463649	0.8957889	0.0046188
factor:(INSREC_PUB)3	-0.1487134	0.1319291	0.8618161	0.6654514	1.1161251	0.2596487

Table 4. Wald Tests for Cox model

	test	df	pvalue
SEX	1.36	1	0.243
stage	5861.42	4	0.000
RAC_RECY	143.10	3	0.000
BRST_SUB	1049.04	3	0.000
MAR_STAT	77.28	1	0.000
MALIGCOUNT	41.38	1	0.000
PRIMSITE	565.35	8	0.000
ERSTATUS	1013.79	1	0.000
PRSTATUS	1014.57	1	0.000
INSREC_PUB	508.75	3	0.000

Table 5. Score Tests for Cox model

	test	df	pvalue
SEX	1.370	1	0.242
stage	13294.571	4	0.000
RAC_RECY	149.035	3	0.000
BRST_SUB	1193.626	3	0.000
MAR_STAT	77.816	1	0.000
MALIGCOUNT	41.743	1	0.000
PRIMSITE	611.374	8	0.000
ERSTATUS	1138.773	1	0.000
PRSTATUS	1133.384	1	0.000
INSREC_PUB	544.119	3	0.000

Table 6. Likelihood Ratio Tests for Cox model

	test	df	pvalue
SEX	1.260	1	0.262
stage	5794.524	4	0.000
RAC_RECY	128.648	3	0.000
BRST_SUB	895.337	3	0.000
MAR_STAT	73.849	1	0.000
MALIGCOUNT	36.586	1	0.000
PRIMSITE	495.260	8	0.000
ERSTATUS	885.310	1	0.000
PRSTATUS	985.679	1	0.000
INSREC_PUB	452.691	3	0.000

Table 7. Log Rank Tests and Weight for Cox model.

variable	Log.rank_p	Gehan_p
stage	0.0000	0.0000
RAC_RECY	0.0000	0.0000
SEX	0.2414	0.2807
BRST_SUB	0.0000	0.0000
PRIMSITE	0.0000	0.0000
ERSTATUS	0.0000	0.0000
PRSTATUS	0.0000	0.0000
INSREC_PUB	0.0000	0.0000

Table 8. Comparison of 5 year Survival Rate Between Our Data and Official Report from American Cancer Society

	Our data		Official report from American Cancer Society
	5-year survival rate	95% CI	5-year survival rate
Stage 0	-	-	close to 100%
Stage I	97.8%	(97.5, 98.1)	close to 100%
Stage II	91.6%	(90.9, 92.3)	93%
Stage III	76.6%	(74.9, 78.3)	72%
Stage IV	32.8%	(29.8, 35.7)	22%

Table 9. 5-year Survival Rates for Breast Cancer by Gender.

	Our data	
	5-year survival rate	95% CI
Male	84.7%	(77.8, 91.5)
Female	90.2%	(89.9, 90.6)

Table 10. 5-year Survival Rate Based on ER and PR Status

ER status	Our data	
	5-year survival rate	95% CI
ER+	92.3%	(91.9, 92.7)
ER-	79.8%	(78.6, 81.0)

PR status	Our data	
	5-year survival rate	95% CI
PR+	93.1%	(92.7, 93.5)
PR-	82.4%	(81.5, 83.4)

ER/PR status	Our data	
	5-year survival rate	95% CI
ER+ / PR+	93.3%	(92.9, 93.7)
ER+ / PR-	86.3%	(85.0, 87.5)
ER- / PR+	81.0%	(75.6, 86.4)
ER- / PR-	79.7%	(78.4, 81.0)

Table 11. 5-year Survival Rate for Breast Cancer Based on Breast Cancer Subtype

	Our data	
	5-year survival rate	95% CI
Her2+/HR+	89.9%	(88.6, 91.2)
Her2+/HR-	84.3%	(82.1, 86.5)
Her2-/HR+	92.5%	(92.1, 92.9)
Triple Negative	77.8%	(76.2, 79.3)

Table 12. 5-year Survival Rate for Breast Cancer Based on Race

	Our data	
	5-year survival rate	95% CI
White	90.3%	(89.9, 90.7)
Black	81.6%	(79.1, 84.1)
American Indian, Alaska Native	89.3%	(85.0, 93.5)
Asian or Pacific Islander	93.5%	(92.4, 94.6)

Table 13. 5 year Survival Rate Based on Tumor Origin

	Our data	
	5-year survival rate	95% CI
Nipple	83.7%	(74.6, 92.9)
Central portion	87.6%	(85.5, 89.6)
Upper-inner quadrant	93.4%	(92.5, 94.3)
Lower-inner quadrant	90.3%	(88.5, 92.0)
Upper-outer quadrant	92.1%	(91.5, 92.7)
Lower-outer quadrant	91.8%	(90.5, 93.2)
Axillary tail of breast	85.0%	(76.7, 93.4)
Overlapping lesion	90.3%	(89.5, 91.2)
Breast, NOS	81.5%	(80.0, 82.9)

Table 14. 5-year Survival Rate Based on Marital Status

	Our data	
	5-year survival rate	95% CI
Single	85.6%	(84.4, 86.8)
Married	92.6%	(92.2, 93.1)
Separated	88.1%	(84.6, 91.6)
Divorced	88.8%	(87.5, 90.1)
Widowed	84.9%	(83.6, 86.3)
Unmarried or domestic partner	91.4%	(85.7, 97.1)

Table 15. 5-year Survival Rate Based on the Total Number of In Situ/Malignant Tumors

	Our data	
	5-year survival rate	95% CI
# of malignant tumors >1	86.5%	(84.9, 88.1)
# of malignant tumors ≤ 1	90.5%	(90.1, 90.9)

Table 16. 5-year Survival Rate Based on Insurance Status

	Our data	
	5-year survival rate	95% CI
Uninsured	82.1%	(80.6, 83.5)
Any Medicaid	92.1%	(91.7, 92.5)
Insured	88.5%	(87.3, 89.7)
Other	78.7%	(73.0, 84.4)

FIGURES

Figure 1. Product-Limit (Kaplan-Meier) Estimates for Breast Cancer Survival Probability

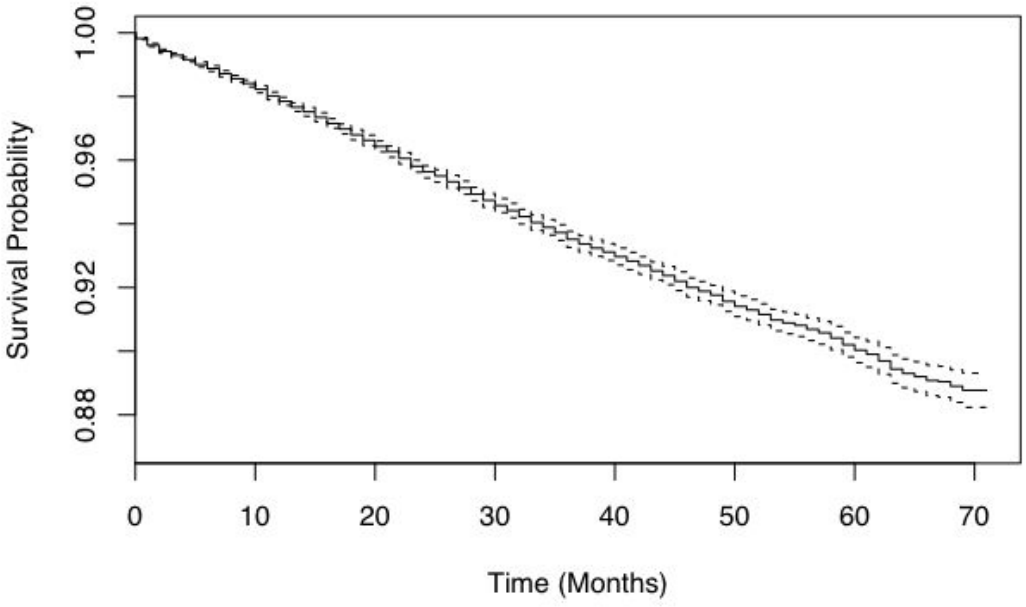

Figure 2. Cumulative Hazard Function for Breast Cancer

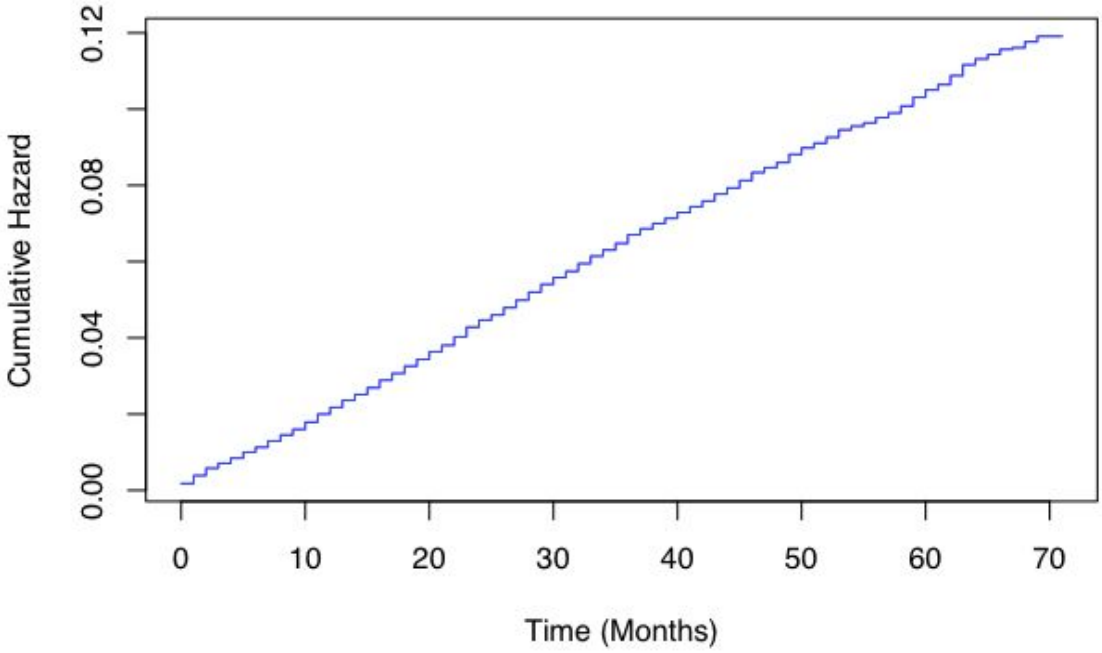

Figure 3. Estimated Survival Functions for Breast Cancer by Gender

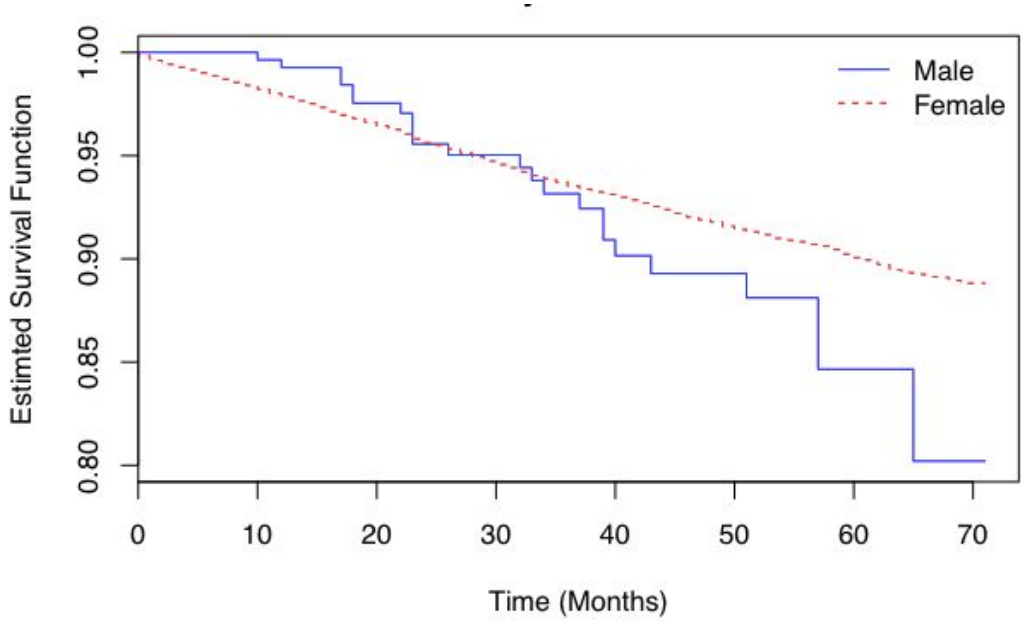

Figure 4. Estimated Survival Functions for Breast Cancer by Race

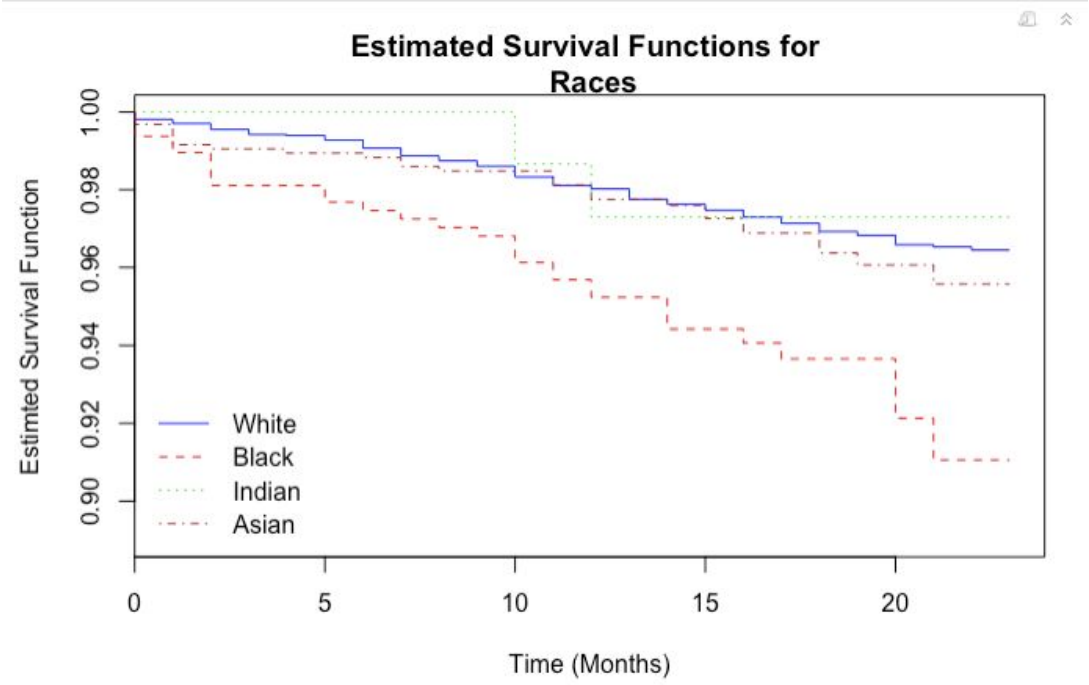

Figure 5. Estimated Survival Functions For Breast Cancer by Cancer Stage

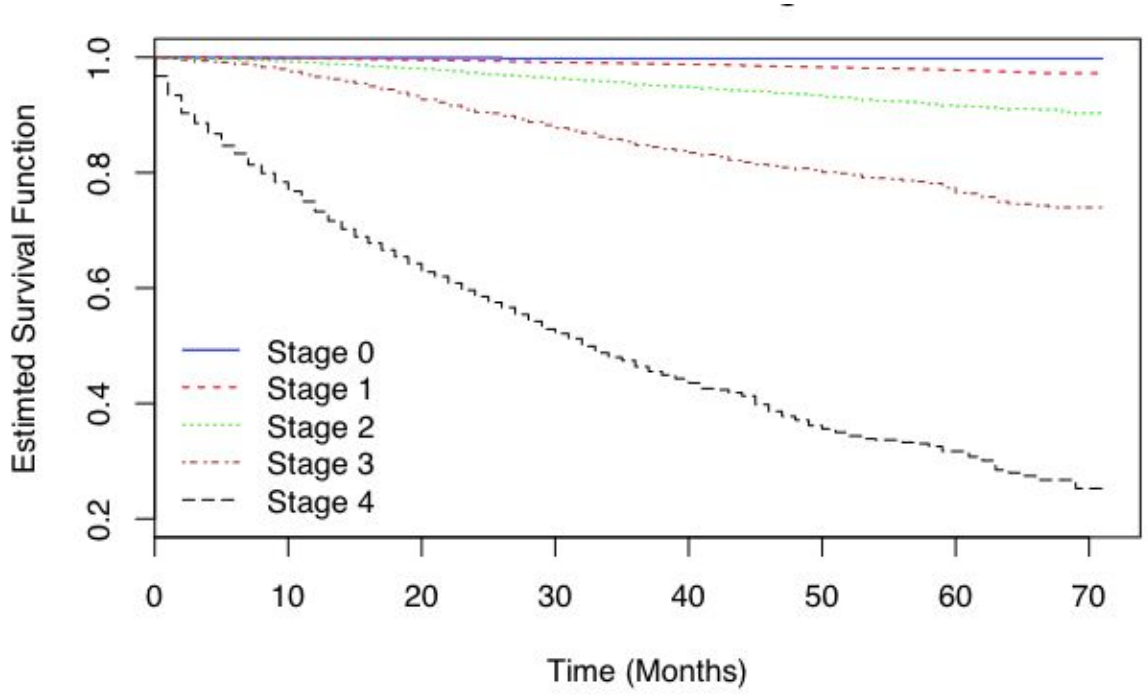

Figure 6. Cox-Snell Residual Plot

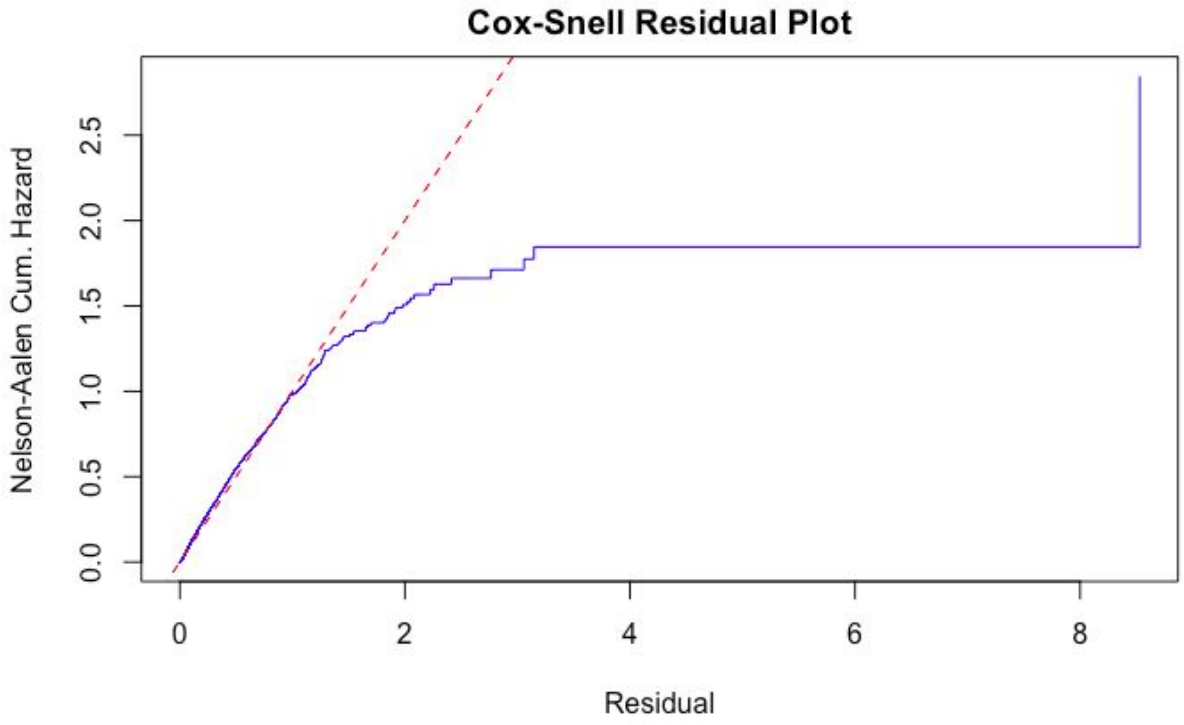

Figure 7. Andersen Plot for Gender

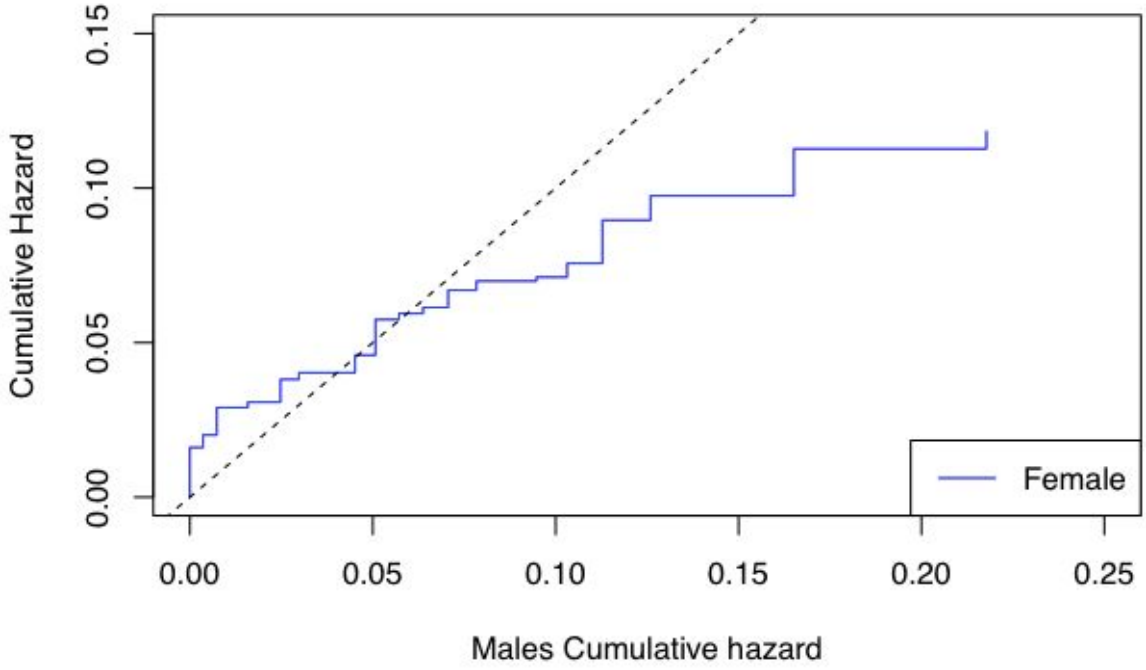

Figure 8. Andersen Plot for Race

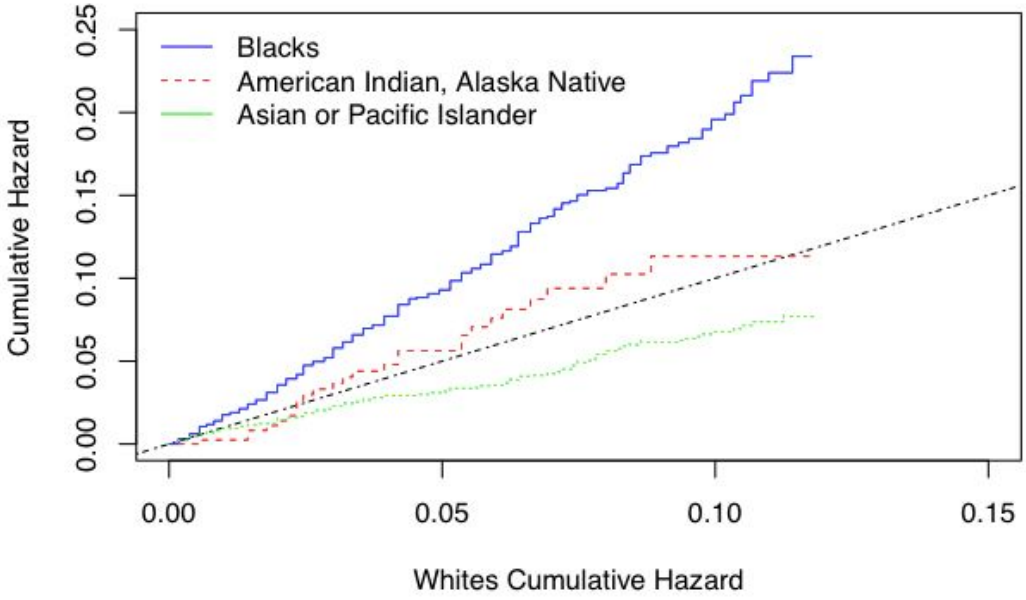

Figure 9. Andersen Plot for Cancer Stage