Name: Huijun An
UID: 904499500
BIOS M215 Final Project

Survival Analysis on San Francisco Breast Cancer Patients from 1988-2016

Introduction

Breast cancer is one of the most common cancers in the world. Even though breast cancer is one the most curable cancers, there were approximately 570,000 breast cancer deaths in 2015 due to the fact that over 1.5 million women worldwide are diagnosed with breast cancer every year.[1,2] Metastasis, which refers to the spread of cancer cells to other areas of the body, accounts for most cases of breast cancer incurability.[3] Risk factors for breast cancer include being female, obesity, older age, family history, lifestyle, etc. Common development of breast cancer starts in cells from the lining of milk ducts and the lobules that supply the ducts with milk, known as ductal carcinomas and lobular carcinomas.[4] A common indicator in a breast cancer patient's pathology report is the hormone receptor status, which affects the cancer therapy. A cancer has positive ER status when it has receptors for estrogen and has positive PR status when it has receptors for progesterone.[6]

Surprisingly, the incidence rate of breast cancer varies greatly in different countries and districts. Previous research has shown that San Francisco Bay area has incidence rates of breast cancer for non-Hispanic white women among the highest in the world.[5] In light of the high incidence rate in San Francisco, this project plans to conduct exploratory survival analysis on breast cancer patients in San Francisco. Specifically, this project aims to find related covariates contributing to patient survival time and fit multiple survival models to find the best model for those patients.

Methods & Results

A. Data Acquisition and Pre-processing

The data for this project was acquired from SEER (The Surveillance, Epidemiology, and End Results) research data record from NIH National Cancer Institute. The data was originally splitted into several txt files based on the type of cancer with description listed in detail in https://seer.cancer.gov/data-software/documentation/seerstat/nov2018/TextData.FileDescription.pdf. The txt file labeled "Breast" was chosen and it originally contains 840,666 records and 142 variables. Since this project's mainly interest was in San Francisco patients, patients only in San Francisco area were selected using the variable "registry ID" with the value "0000001501" based on the information in the description file, resulting in the initial data file of 137,103 records and 142 variables.

Given the 142 variables, extensive data pre-processing was done to obtain a clean dataset for survival analysis. In order to obtain complete cancer stage information from each patient, data prior to 1988 were eliminated. Due to the limited scale of this project, the focus was on malignant breast cancer patients and only on patients with one record in the dataset. Therefore, data labeled as carcinoma in situ were eliminated. Columns that have one value or missing value across all data were eliminated. There were multiple columns sharing same information in different format about race, histology, and age and only one of them in each characteristic was chosen to be used in this project. Not cancer related variables like edition of data manual were eliminated from the dataset. Multiple variables in the dataset were used for determination of cancer stage, which means that their information was included in the Stage variable. Hence, those variables, such as tumor size, lymph nodes containing cancer cell or not, and metastasis, were eliminated from the dataset. Moreover, variables containing more than 70% missing values

were eliminated from the dataset in order to keep the sample size of this project. Dummy variables were created for each categorical variable. Lastly, rows with missing values were eliminated for easiness of model fitting, resulting in the final dataset with 13 covariates (marital status, sex, diagnosis age, number of primary tumors, laterality, grade, primary site surgery, histology, race, ER status, PR status, tumor marker, and stage) and 10,519 records. The outcome variable in this dataset is survival time in months, while the right-censoring variable is called "VSRTSADX" in the dataset standing for patient death or not. Rows with 0 survival month are eliminated for the purpose of analysis of the project. The univariate analyses for covariates and outcome variable are summarized in Table 1.

Table 1. Summary Statistics

Variables	Patient Number (%)
Marital Status	
Single	1,864 (17.7%)
Married	5,947 (56.5%)
Separated / Divorced	1,115 (10.6%)
Widowed	1,593 (15.1%)
Sex	
Male	61 (0.6%)
Female	10,458 (99.4%)
# of Primary Tumors	
One in lifetime	8,840 (84.0%)
More than one	1,679 (16.0%)
Diagnosis Age, mean ± standard deviation	59.5 ± 13.7
Laterality	
Left	5,130 (48.8%)
Right	5,389 (51.2%)
Grade	
1	2,640 (25.1%)
2	4,425 (42.1%)
3	3,188 (30.3%)
4	266 (2.5%)
Surgery at Primary Site	
Yes	10,357 (98.5%)
No	162 (1.5%)
Histology	

1 = epithelial neoplasms, NOS	33 (0.3%)
2 = squamous cell neoplasm	14 (0.13%)
5 = adenomas and adenocarcinomas	213 (2.0%)
6 = adnexal and skin appendage neoplasm	21 (0.2%)
8 = cystic, mucinous, and serous neoplasm	203 (1.9%)
9 = ductal and lobular neoplasm	10,021 (95.3%)
11 = complex epithelial neoplasm	13 (0.12%)
22 = fibroepithelial neoplasms	1 (0.01%)
Race	
White	8,010 (76.2%)
Black	861 (8.2%)
Other	1,648 (15.7%)
ER Status	
Positive	8,477 (80.6%)
Negative	2,042 (19.4%)
PR Status	
Positive	7,324 (69.6%)
Negative	3,195 (30.4%)
Stage	
I	5,227 (49.7%)
IIA	2,549 (24.2%)
IIB	1,070 (10.2%)
III	27 (0.26%)
IIIA	770 (7.3%)
IIIB	183 (1.7%)
IIIC	409 (3.9%)
IV	284 (2.7%)
Tumor marker	
Positive	8,642 (82.1%)
Negative	1,877 (17.9%)
Survival time in months, mean ± standard deviation	145.1 ± 63.9
Patient Vital status	
Alive	8600 (81.8%)
Dead	1919 (18.2%)

Note: Data are in format of number (%) of inputs, unless otherwise indicated.

B. Kaplan Meier Estimator

The Kaplan Meier estimator (KME) is a non-parametric statistic used to estimate the survival function for this data. The equation for this estimator is: $\hat{S}(t) = \prod_{i:\, t_i \le t}(1 - \frac{d_i}{r_i})$, where t_i is the time when at least one event happened, d_i is the number of events (i.e. deaths) happened at t_i,

and r_i is the number of individuals in the risk set. The KME for the whole data set is shown in Figure 1. The stratified KME curve by different categories are shown in Figure 2.1-12. It can be clearly seen using eyeballs from the stratified KME curve that the survival probability for patients with different grades, having primary site surgery or not, different histology, and different stages are not the same between groups. In order to statistically test if the differences in survival probability are significant, the next part will be log rank test.

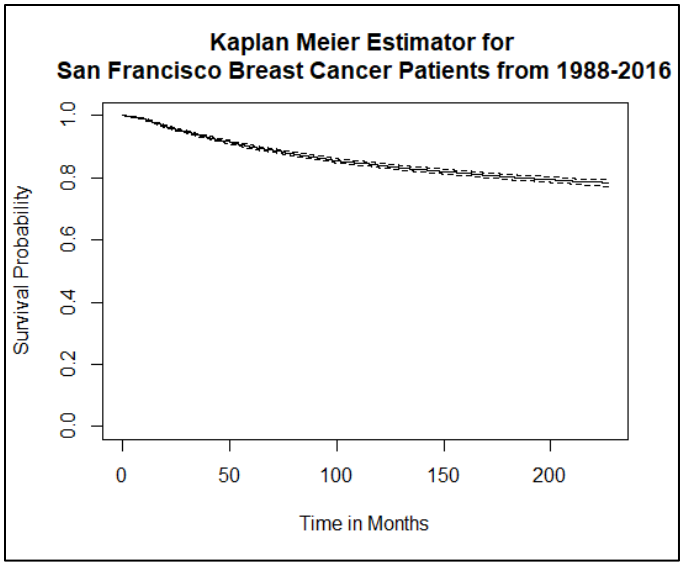
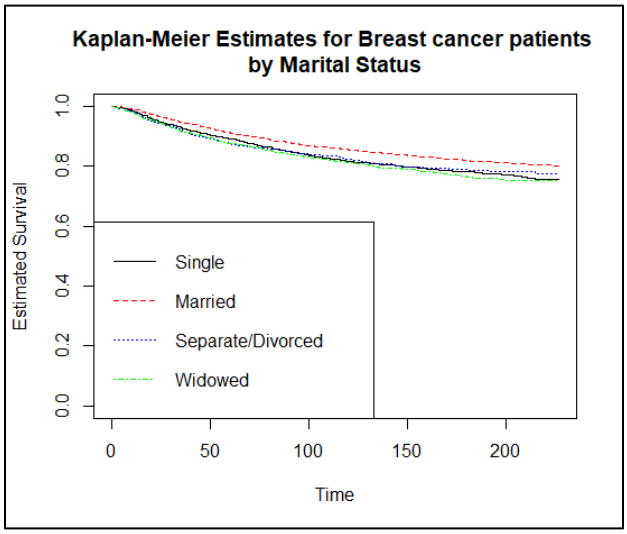
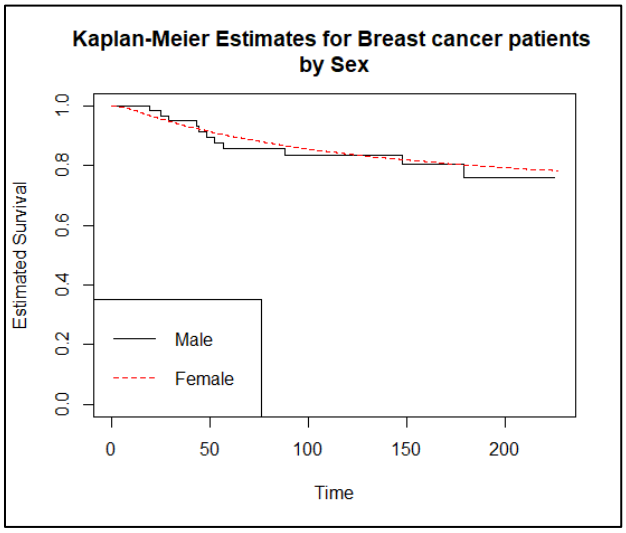
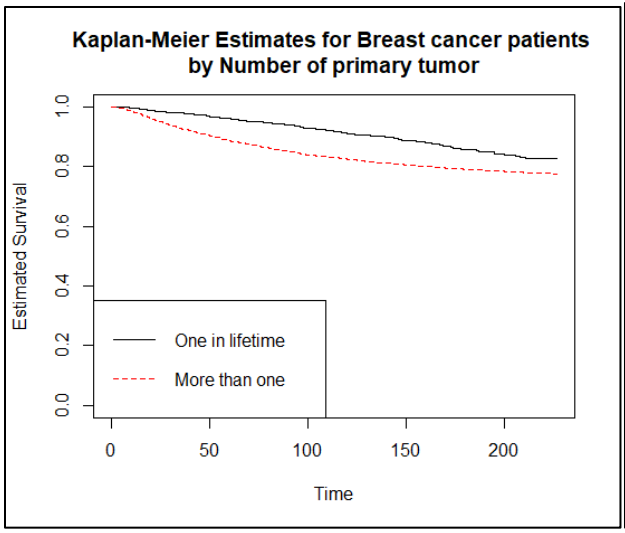
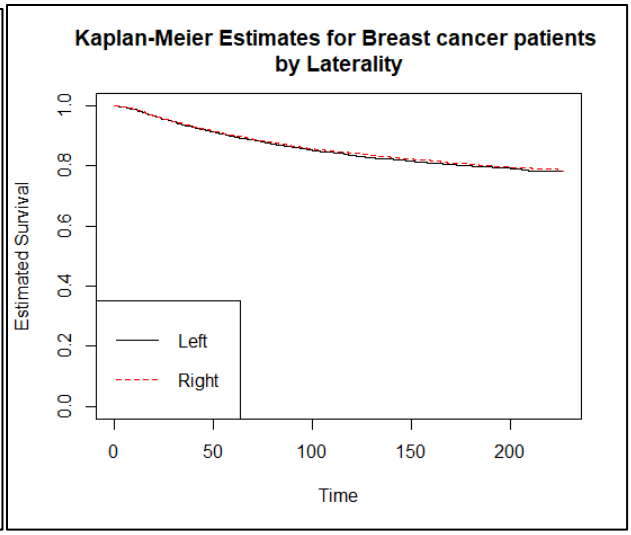


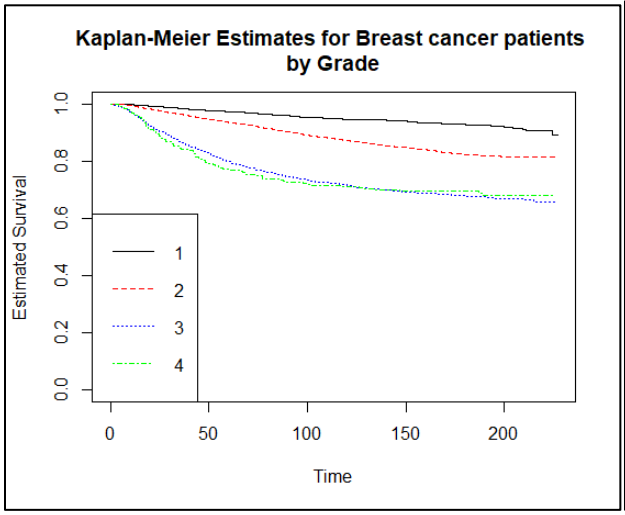
Figure 1. Kaplan Meier Estimator for San Francisco Breast Cancer Patients from 1988-2016

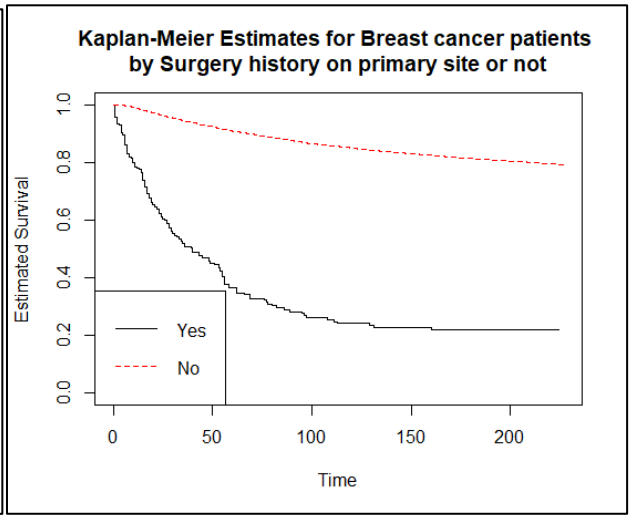


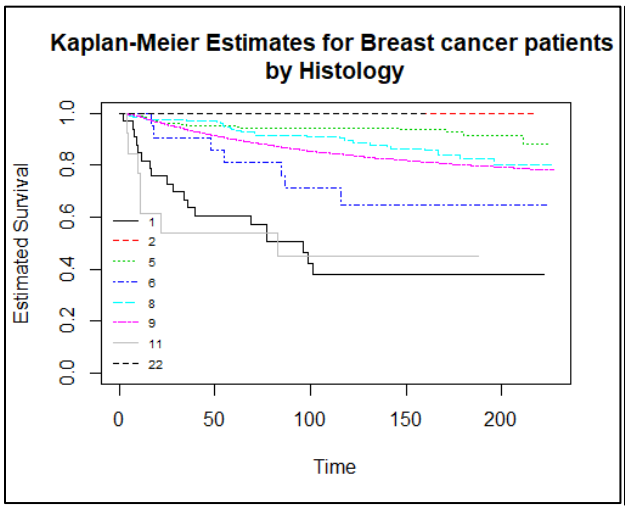


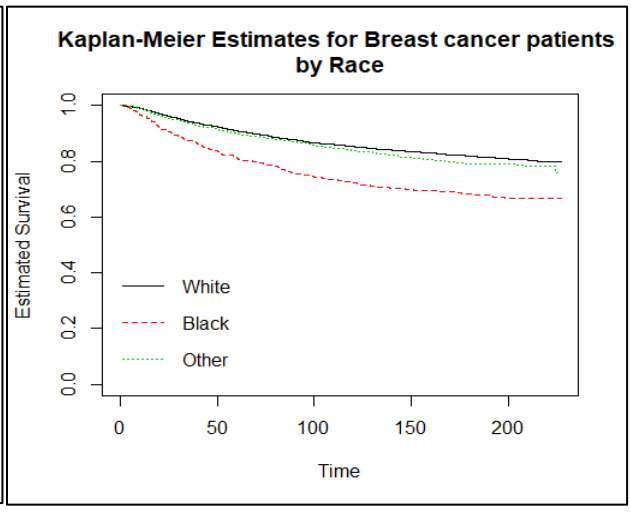

Kaplan-Meier Estimates for Breast cancer patients by Number of primary tumor
Estimated Survival
Time
One in lifetime
More than one


Kaplan-Meier Estimates for Breast cancer patients by Laterality
Estimated Survival
Time
Left
Right


Kaplan-Meier Estimates for Breast cancer patients by Grade
Estimated Survival
Time
1
2
3
4


Kaplan-Meier Estimates for Breast cancer patients by Surgery history on primary site or not
Estimated Survival
Time
Yes
No


Kaplan-Meier Estimates for Breast cancer patients by Histology
Estimated Survival
Time
1
2
5
6
8
9
11
22


Kaplan-Meier Estimates for Breast cancer patients by Race
Estimated Survival
Time
White
Black
Other

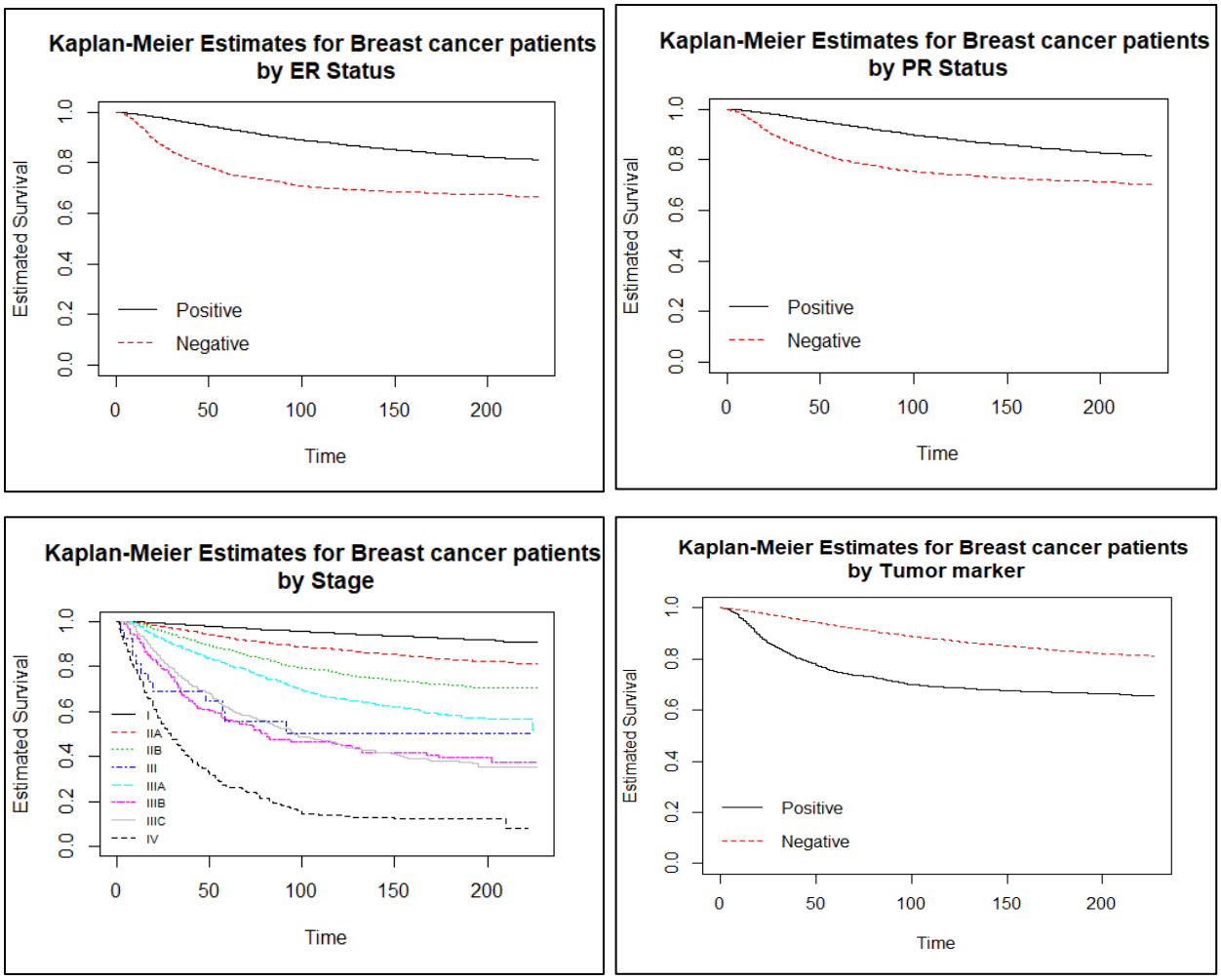


Figure 2.1-2.12. Kaplan Meier Estimator for San Francisco Breast Cancer Patients from 1988-2016 by different covariate categories

C. Log-rank test

The log-rank test is a hypothesis test to compare the estimates of the hazard functions between two independent right-censored samples. The null hypothesis is that the hazard rates in the two groups are the same at all time points. Under null hypothesis, the test statistics $z = \frac{z_1}{\sqrt{\sigma_{11}}} \sim N(0, 1)$ where $z_j = \sum_{i=1}^{D} w(t_i)[d_{ij} - r_{ij}\frac{d_i}{r_i}]$, $\widehat{\sigma_{jj}} = \sum_{i=1}^{D} w^2(t_i) r_{ij} \frac{d_i}{r_i}(1 - \frac{d_i}{r_i})(\frac{r_i - r_{ij}}{r_i - 1})$. d_{ij} equals to the number of uncensored events in the j^{th} sample at time t_i; r_{ij} equals to the number of individuals in the j^{th} sample that are at risk prior to time t_i; d_i equals to the number of uncensored

events in the pooled sample at time t_i; r_i equals to the number of individuals at risk in the pooled sample prior to time t_i. Without specific interest from physician in the weight of the test, this project uses weight equals 1, the standard log-rank test. The log-rank test results for each covariate are summarized in Table 2. We have sufficient evidence to reject the hypothesis that the hazard rates are the same in different groups in all covariates except Sex and Laterality.

Table 2. Summarized result of log-rank test on each covariate

Covariate	p-value	Significant at $\alpha = 0.05$
Marital Status	8e-07	Yes
Sex	0.7	No
Number of primary tumors	2e-11	Yes
Laterality	0.5	No
Grade	<2e-16	Yes
Surgery on primary site	<2e-16	Yes
Histology	<2e-16	Yes
Race	<2e-16	Yes
ER status	<2e-16	Yes
PR status	<2e-16	Yes
Stage	<2e-16	Yes
Tumor marker	<2e-16	Yes

D. Fitting survival model

In order to better explore the relationship between survival of breast cancer patients and covariates we discussed above, different models including Cox proportional hazards model and Accelerated failure time (AFT) model will be used to fit this dataset.

i. Cox Model

The Cox model is expressed by the hazard function (denoted as h(t)). The equation is shown as: $h(t|z(t)) = h_0(t) \times \exp(\beta^T z(t))$ where t represents the survival time, $h_0(t)$ represents the baseline hazard, z represents the vector of covariates, and β represents the impact of covariates. The quantities $\exp(\beta)$ are called hazard ratios (HR). If $\exp(\beta)$ is greater than 1, it means that when the

value of that specific covariate increases, the event hazard increases. The fitted cox model using all covariates are summarized in Table 3.

Table 3. Cox Model using all covariates

Covariates	**Coefficient**	**exp(Coefficient)**	**p-value**
Diagnosis Age	1.50E-02	1.02	5.4e-14 *
Sex	-4.43E-02	0.957	0.88
Laterality	-1.80E-02	0.982	0.69
ER status	-0.25	0.782	0.17
PR status	-0.20	0.815	0.004 *
Tumor marker	-3.68E-02	0.964	0.85
Marital: Married	-0.19	0.831	0.0024 *
Marital: Separated/Divorced	4.07E-02	1.04	0.63
Marital: Widowed	4.49E-02	1.05	0.6
One primary tumor in lifetime or not	0.30	1.35	6.4e-05 *
Grade2	0.47	1.60	4.7e-08 *
Grade3	0.87	2.39	< 2e-16 *
Grade4	1.03	2.81	6.4e-14 *
Surg	-1.03	0.359	< 2e-16 *
Histology: squamous cell neoplasm	-13.2	1.78E-06	0.98
Histology: adenomas and adenocarcinomas	-0.66	0.515	0.052
Histology: adnexal and skin appendage neoplasm	8.43E-03	1.01	0.99
Histology: cystic, mucinous, and serous neoplasm	0.37	1.45	0.22
Histology: ductal and lobular neoplasm	-8.94E-02	0.915	0.71
Histology: complex epithelial neoplasm	0.65	1.91	0.15
Histology: fibroepithelial neoplasms	-14.9	3.40E-07	0.99
White	-7.95E-02	0.924	0.21
Black	0.23	1.25	0.0091 *
Stage: IIA	0.71	2.04	< 2e-16 *
Stage: IIB	1.23	3.41	< 2e-16 *
Stage: III	1.65	5.21	3.9e-08 *
Stage: IIIA	1.75	5.73	< 2e-16 *
Stage: IIIB	2.24	9.37	< 2e-16 *
Stage: IIIC	2.36	10.6	< 2e-16 *
Stage: IV	3.08	21.7	< 2e-16 *

In order to keep the model parsimonious and pick out the only important features, variable selection needs to be done. In this project, we use stepwise AIC (Akaike Information

Criterion) method to do variable selection. AIC = -2log(-likelihood) + 2p where p is the number of parameters. We want to choose a model with the smallest AIC. The fitted cox model after variable selection are summarized in Table 4. Note that only three variables, Sex, Laterality, and Tumor marker were dropped from the model.

Table 4. Cox Model after variable selection using stepwise AIC

Covariates	**Coefficient**	**exp(Coefficient)**	**p-value**
Diagnosis Age	0.02	1.02	4.9E-14 *
ER status	-0.28	0.76	0.00012 *
PR status	-0.21	0.81	0.00147 *
Marital: Married	-0.18	0.83	0.00239 *
Marital: Separated/Divorced	0.04	1.04	0.62
Marital: Widowed	0.04	1.05	0.60
One primary tumor in lifetime or not	0.30	1.35	6E-05 *
Grade2	0.47	1.60	4.7E-08 *
Grade3	0.87	2.39	< 2e-16 *
Grade4	1.03	2.81	5.6E-14 *
Surg	-1.03	0.36	< 2e-16 *
Histology: squamous cell neoplasm	-13.22	1.82E-06	0.98
Histology: adenomas and adenocarcinomas	-0.67	0.51	0.051
Histology: adnexal and skin appendage neoplasm	0.01	1.01	0.98
Histology: cystic, mucinous, and serous neoplasm	0.37	1.45	0.22
Histology: ductal and lobular neoplasm	-0.09	0.91	0.71
Histology: complex epithelial neoplasm	0.65	1.92	0.15
Histology: fibroepithelial neoplasms	-14.88	3.433E-07	0.99
White	-0.08	0.92	0.21
Black	0.23	1.25	0.01 *
Stage: IIA	0.71	2.04	< 2e-16 *
Stage: IIB	1.23	3.41	< 2e-16 *
Stage: III	1.65	5.19	4.1E-08 *
Stage: IIIA	3.08	21.66	< 2e-16 *
Stage: IIIB	2.24	9.38	< 2e-16 *
Stage: IIIC	2.36	10.63	< 2e-16 *
Stage: IV	3.08	21.66	< 2e-16 *

ii. AFT Model

The AFT model provides an alternative to the Cox model, where it assumes that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant. It can be written as: $\log(x_0) = \log(x) + \beta^T z + \varepsilon$, where x is survival time, z is vector of covariates, and ε is the error term. When the error term has a parametric distribution, we call the AFT model as parametric AFT models. Common distributions include exponential, Weibull, and lognormal, which are the three different AFT models we used below to fit the data. The AFT model results are summarized in Table 5. Note that the three AFT models produced very similar results. Same logic of variable selection was also executed on AFT models, yet they all produced the same model as before.

Table 5. Summarized result of three parametric AFT models using all covariates

	Exponential		Lognormal		Weibull	
Covariates	**Coef**	**p-value**	**Coef**	**p-value**	**Coef**	**p-value**
(Intercept)	7.7	< 2e-16 *	7.4	< 2e-16 *	7.7	< 2e-16 *
Diagnosis Age	-0.016	< 2e-16 *	-0.017	< 2e-16 *	-0.017	3.6E-16 *
ER status	0.077	0.80	-0.009	0.98	0.077	0.80
PR status	0.018	0.70	0.027	0.56	0.018	0.70
Marital: Married	0.20	0.26	0.39	0.027 *	0.21	0.26
Marital: Separated/Divorced	0.21	0.003 *	0.28	8.3E-05 *	0.21	0.0032 *
Marital: Widowed	0.059	0.76	-0.035	0.85	0.059	0.76
One primary tumor in lifetime or not	0.19	0.002 *	0.16	0.013 *	0.19	0.0021 *
Grade2	-0.044	0.61	-0.045	0.61	-0.044	0.61
Grade3	-0.056	0.51	-0.099	0.26	-0.057	0.51
Grade4	-0.29	8.2E-05 *	-0.22	0.0013 *	-0.30	8.9E-05 *
Surg	-0.48	2.7E-08 *	-0.42	1.3E-08 *	-0.48	3.3E-08 *
Histology: squamous cell neoplasm	-0.88	< 2e-16 *	-0.84	< 2e-16 *	-0.89	< 2e-16 *
Histology: adenomas and adenocarcinomas	-1.04	4.2E-14 *	-0.99	3.1E-12 *	-1.05	7.6E-14 *
Histology: adnexal and skin appendage neoplasm	1.02	< 2e-16 *	0.99	2.2E-10 *	1.03	< 2e-16 *
Histology: cystic, mucinous, and serous neoplasm	14.05	0.98	8.02	0.98	14.36	0.98

Histology: ductal and lobular neoplasm	0.72	0.034 *	0.53	0.18	0.730	0.035 *
Histology: complex epithelial neoplasm	-0.025	0.96	-0.016	0.98	-0.024	0.96
Histology: fibroepithelial neoplasms	-0.36	0.29	-0.29	0.43	-0.33	0.29
White	0.14	0.57	0.094	0.77	0.14	0.57
Black	-0.57	0.21	-0.72	0.21	-0.57	0.21
Stage: IIA	15.74	< 2e-16 *	8.75	< 2e-16 *	14.92	< 2e-16 *
Stage: IIB	0.075	0.24	0.048	0.46	0.076	0.24
Stage: III	-0.23	0.0078 *	-0.27	0.0032 *	-0.23	0.008 *
Stage: IIIA	-0.72	< 2e-16 *	-0.62	< 2e-16 *	-0.73	< 2e-16 *
Stage: IIIB	-1.25	< 2e-16 *	-1.15	< 2e-16 *	-1.26	< 2e-16 *
Stage: IIIC	-1.66	3.2E-08 *	-1.97	2.7E-08 *	-1.68	3.8E-08 *
Stage: IV	-1.78	< 2e-16 *	-1.69	< 2e-16 *	-1.80	< 2e-16 *
log(Scale)	-	-	-2.33	< 2e-16 *	-2.30	< 2e-16 *

iii. Prediction Accuracy Measures and Model Diagnostics

In order to evaluate how good the model fits the data, two types of prediction accuracy measures are used. The first measure, nonparametric R^2 measure, defined as the proportion of explained variance, quantifies the potential predictive power of the nonlinear prediction function. The second measure, L2, defined as the proportion of explained prediction error, gauges the closeness of the prediction function to its corrected version.[7] The two measures for each model fitted above are summarized in Table 6. Since R^2 is the most important measure to determine the prediction accuracy, we select the original Cox model as the best model fitting this survival data as it has the highest R^2. The goodness of fit of the model is analyzed using Cox-Snell residual plot as shown in Figure 3.

Table 6. Prediction Accuracy Measures for Cox model and AFT models

Model	**R^2**	**L^2**
Cox	0.2095	0.3055
Cox after variable selection	0.2092	0.3056
AFT - Exponential	2e-04	0.9949
AFT - Lognormal	2e-04	0.9949
AFT - Weibull	2e-04	0.9949

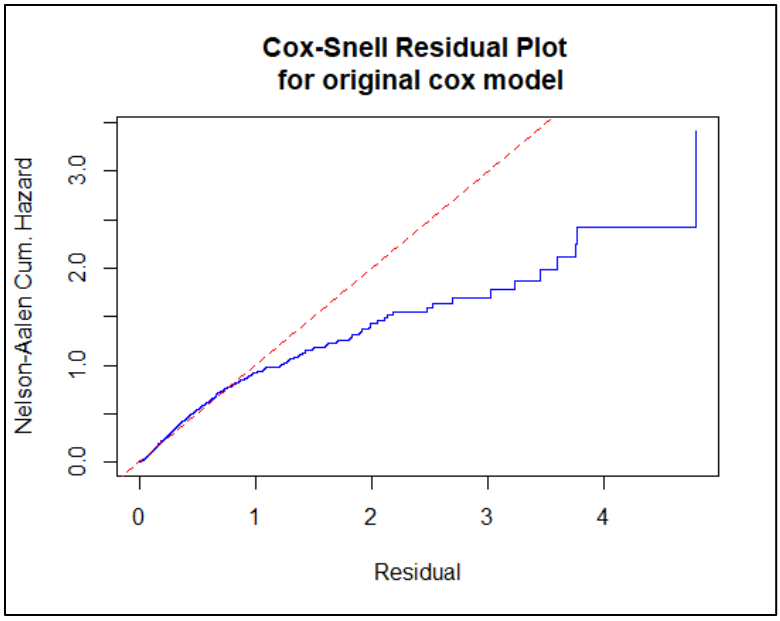


Figure 3. Cox-Snell residual plot for original cox model

Discussion

The primary goal of this project is to identify covariates contributing to breast cancer survival and fit the best model predicting breast cancer survival using SEER data of patients in San Francisco from 1988 too 2016. First, we visualized the survival curve over about 17 years (200 months) for breast cancer patients using Kaplan-Meier Estimation in Figure 1. It can be seen that it has not reached median survival time in the shown figure, indicating that breast cancer patients overall have a high survival rate in San Francisco. With KME curve stratified on different subgroups, it can be seen that patients with higher grade, have surgery history on primary site, histology at complex epithelial neoplasm and fibroepithelial neoplasms, and later cancer stage clearly have lower survival rate. Each covariate's subgroups are being compared using log-rank test to statistically confirm that these covariates have significant effects on breast cancer survival. Other than these covariates, log-rank tests also confirmed the effect of Marital

status, number of primary tumors, race, ER status, PR status, and prognostic tumor marker in breast cancer survival.

In order to fit the best model to predict breast cancer survival, Cox model and AFT model with exponential, lognormal and Weibull distribution are used. After fitting each model with all the covariates, a variable selection method, stepwise AIC, is used to find the best subsets of variables. Yet, AFT models after variable selection give the same sets of variables. Cox model after variable selection excluded sex, laterality, and tumor marker. The prediction accuracy of each of the five models is evaluated using pseudo R^2. However, even for the model given the highest R^2, the original cox model, only gives us the R^2 of 0.2095. There are many possible reasons for the poor fit. First of all, the dataset in SEER only includes very basic demographic information in addition to cancer related covariates. From previous research in introduction, many known risk factors like family history, weight, and lifestyle (i.e. smoking, drinking, etc) are not included in the dataset. Moreover, given the fact that the five-year survival rate for breast cancer patient is over 99%, with limitation in years of the dataset, it is hard to predict patients' survival. This is proven by the alive/dead percentage shown in Table 1 where more than 80% of patients are alive (censored) in the dataset. In general, cancer survival is a completed issue and patient death can be caused by many different reasons. For future studies, a competing risk model which takes into account the cause of death could be more realistic and gives a better prediction of breast cancer patient survival.

Reference

1. Stewart BW, Wild CP. World Cancer Report 2014. Geneva, Switzerland: WHO Press; 2014.

2. WHO: Geneva, Switzerland. Breast cancer. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/
3. Sun YS, et al. Risk factors and preventions of breast Cancer. Int J Biol Sci. 2017;13(11):1387–1397.
4. Breast cancer. (2019, Dec 7). https://en.wikipedia.org/wiki/Breast_cancer.
5. Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB, (Eds)Cancer Incidence in Five Continents, Vol. VIII, Vol. 155. Lyon, France: IARC Scientific Publications; 2002.
6. Hormone Receptor Status. https://www.breastcancer.org/symptoms/diagnosis/hormone_status. 2018.
7. Gang Li & Xiaoyan Wang (2019) Prediction Accuracy Measures for a Nonlinear Model and for Right-Censored Time-to-Event Data, Journal of the American Statistical Association.