Estimating the Impact of Lost to Follow-up on Breast Cancer Patients' Five-Year Disease-free Survival

A Senior Project
Presented to
the Faculty of the Statistics Department
California Polytechnic State University, San Luis Obispo

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Science

By
Megan Whitworth
June 2016

Abstract

Background Nationally, the 5-year survival rate for patients with breast cancer is relatively higher than patients diagnosed with other types of cancer. In addition to the higher survival rates, breast cancer patients also tend to have increased rates of lost to follow-up as compared to other cancers. When a patient becomes lost, the occurrence of distant metastasis cannot be reliably ascertained, unless the patient had a breast cancer-specific death. As a consequence of the missing information from lost patients, results from statistical analyses that contain lost patients may not adequately reflect the actual recurrence and disease-free survival rates. The impact of lost patients on the unadjusted and adjusted disease-free survival (DFS) was explored in breast cancer patients seen at the City of Hope National Medical Center in Los Angeles from 1997 to 2012.

Methods Female breast cancer patients with stage 0, I, II, or III at diagnosis were included in the analyses (N = 2,358). Of these patients, 1,937 were deemed non-lost and 421 were lost. Kaplan-Meier estimates for DFS were stratified by lost status. Cox proportional hazard models were built to adjust for multiple predictors such as age group at diagnosis, race, comorbidity score, stage at diagnosis, health insurance type, employment status, and lymphovascular invasion (LVI). Patients were separated into 20 groups based on propensity scores from a logistic model using the variables categorical distance between the patient's residence and the City of Hope, age at diagnosis, stage at diagnosis, insurance type, hormone receptor status, and her2/neu status to predict the probability of being lost. Lost patients were then removed from their assigned propensity score group and replaced with simulated lost patients from the corresponding propensity score group. Simulated lost patients were sampled with replacement from the non-lost patients within each group and then information from different assessment periods were removed from those patients. The new 5-year DFS rate and hazard ratios were calculated. The process of simulating lost patients and recalculating the 5-year DFS and hazard ratio was bootstrapped 1,000 times

Results The 5-year DFS was 95.1% for lost patients and 84.6% for non-lost patients. Adjusting for age, race, comorbidity score, stage, insurance, employment, and LVI, the risk of death or recurrence is 61.5% lower for lost patients compared to non-lost patients (HR = .385, $P<0.01$). After simulating losing assessment periods, there was a higher than average amount of assessments to be lost to capture the DFS rates of the actual lost patients. The actual cohort of patients went lost after an average of three assessments, while in the simulated cohort it took between seven to ten assessments lost before the unadjusted and adjusted 5-year disease free survival rates reflected the 5-year disease free survival rates of the actual lost cohort.

Conclusion A higher than average number of assessments needed to be lost to capture the disease-free survival rates of the actual lost patients. This indicates that the differences in disease-free survival rates between non-lost and actual lost patients is not only due to missing information, but also that lost patients may actually be healthier than their non-lost counterparts— which could be a reason that the patients stopped following-up at City of Hope.

Table of Contents

List of Equations

List of Figures

List of Tables

Introduction

City of Hope National Medical Center is leading research and treatment center for cancer, diabetes, and other life-threatening diseases. City of Hope is one of only forty-five comprehensive cancer centers and is a leader in research of cancer treatment. City of Hope offers a Women's Cancer program that aims to better understand the cancers that affect women, including breast cancer. Studies include focusing on "engaging every stage of the immune response to defeat breast cancer." With these new breakthrough treatments, one of the goals is to increase the survival rates among breast cancer patients.

Nationally, the 5-year survival rate for patients with breast cancer is relatively higher than patients diagnosed with other types of cancer.* In addition to the higher survival rates, breast cancer patients also tend to have increased rates of lost to follow-up as compared to other cancers. No official conclusions have been made about the cause of higher rates of lost to follow-up in breast cancer patients as compared to other cancers. However, this may be due to patients seeking follow-up care outside the clinic they received treatment from. Patients may be more likely to follow-up with providers other than their oncologist due to breast cancer being relatively easier to treat than other cancers and more information being known about the disease.

Patients were deemed as lost if they had not followed-up for two or more years. If a patient does not follow-up then we do not have current, and therefore accurate information about if their breast cancer has recurred. Breast cancer can recur at the original site (called local recurrence), but can also return and spread to other parts of the body (called metastasis or distance recurrence). When determining the disease-free survival (DFS) rates we are not only considering if the patient died, but also if the patients had a distant metastatic site. As a consequence of the missing recurrence information from lost patients, results from statistical analyses that contain lost patients may not adequately reflect the actual recurrence and survival rates. This issue was previously explored by a former student from California Polytechnic State University, San Luis Obispo, Debbie Yan Qun Huang (2013). This project is a continuation of Huang's senior project and contains work from her project in addition to expanding on her previous findings. The impact of lost patients on the unadjusted and adjusted disease-free survival was explored in breast cancer patients seen at the City of Hope National Medical Center in Los Angeles from 1997 to 2012.

*The national average for 5-year survival rates for breast cancer patients is about 89.7% while the average for 5 year survival rates for all types of cancer is about 66.9% according to data from a National Cancer Institute Surveillance, Epidemiology, and End Results Program (SEER) 2006-2012 study.

Methods

Description of Cohort

Female breast cancer patients seen at the City of Hope National Medical Center in Los Angeles from 1997 to 2012, with stage 0, I, II, or III at diagnosis were included in the study (N=2,358). Stage IV breast cancer patients were omitted due to metastases at diagnosis— the cancer had spread beyond the breasts and local lymph nodes, and thus, there was no measurable recurrence. Of the 2,358 breast cancer patients, 1,937 were considered non-lost and 421 were lost.

Description of Variables

Patients were defined as lost if they had not been to a follow-up assessment for two or more years. A recurrence is considered to be a metastasis to a distant site. A breast specific death occurs when a patient's cause of death was due to breast cancer.

Lymphovascular invasion (LVI) occurs when cancer cells are present in blood vessels or lymphatic vessels. The presence of LVI indicates that treatment should most likely include chemotherapy or hormone therapy. Tumor grade is an assessment of the growth patterns and features of cell differentiation. Well-differentiated cells have a low tumor grade, meaning that the growth and spread of the cancer tends to be slower than undifferentiated cells (high grade). Hormone receptor status and her2/neu status are related to the likelihood of the patient responding to certain drug treatments.

Stage at diagnosis was categorized as 0, I, II, or III. Age group was categorized as "pre-menopausal" for patients younger than 50, "post-menopausal" for patients 50 to 70, and "elderly" for patients older than 70. Race was categorized into "White", "Black", "Hispanic", "Asian", and "Other". Comorbidity score was categorized as "low" for a score of 0, "medium" for 1 to 2, and "high" for 3 to 6.

Distance was calculated using the patients' zip codes. Data from SAS® Maps Online was used to match the patients' zip codes to corresponding latitudes and longitudes. Conversions of degrees to radians (Equation 1) and the Great Circle Distance Formula (Equation 2) were used to calculate the shortest distance in miles between the patients' residences and the City of Hope (COH). Distance was categorized as "close" if the patient lived 0 to 50 miles away from the City of Hope, "medium" for 51 to 100 miles, and "far" for 100 or more miles, and "foreign" for patients residing in foreign countries.

Equation 1. Converting Degrees to Radians

$$Radians = \frac{arctan(1)}{45} \times Degrees$$

Equation 2. Great Circle Distance Formula

$$Distance = 3949.99 \times arcos\big(sin(Patient's\ Latitude)\big) \times sin(COH's\ Latitude) + \cos(Patient's\ Latitude) \times \cos(COH's\ Latitude) \times \cos(Patient's\ Longitude - COH's\ Longitude)$$

Statistical Methods

Kaplan-Meier estimates for disease-free survival probabilities were stratified by lost status. Cox proportional hazards models were built to adjust for multiple predictors such as age at diagnosis, race/ethnicity, comorbidity score, stage at diagnosis, health insurance type, and LVI status. Control variables that did not meet the proportional hazards assumption by the supremum test were included as strata variables. These included tumor grade, hormone receptor status, and her2/neu status.

To model the impact of missing distant recurrences on disease-free survival, a logistic regression model was built to calculate propensity scores that determine which covariates predict becoming lost. Patients with similar propensity sores will have similar characteristics that might explain why they became lost.

Propensity scores were then used to create 20 groups of similar size, such that patients in the same group have similar propensity scores, and thus, similar characteristics in terms of becoming lost. The proportion of lost patients per group was calculated and lost patients were removed. Sampling with replacement was then used to simulate the lost patients based on the patients with complete information. These simulated patients were assigned a status of lost to see how this would impact the DFS rates. During each iteration of the simulation, we lose one assessment period at a time. When deleting assessment periods, a new censoring date was determined for these simulated lost patients and any recurrence that occurred in the previous assessment periods became unknown. Figure 1 shows a timeline of a simulated lost patient and the process of losing assessment periods.

Figure 1: Timeline of Losing Assessment Periods for Simulated Lost Patient

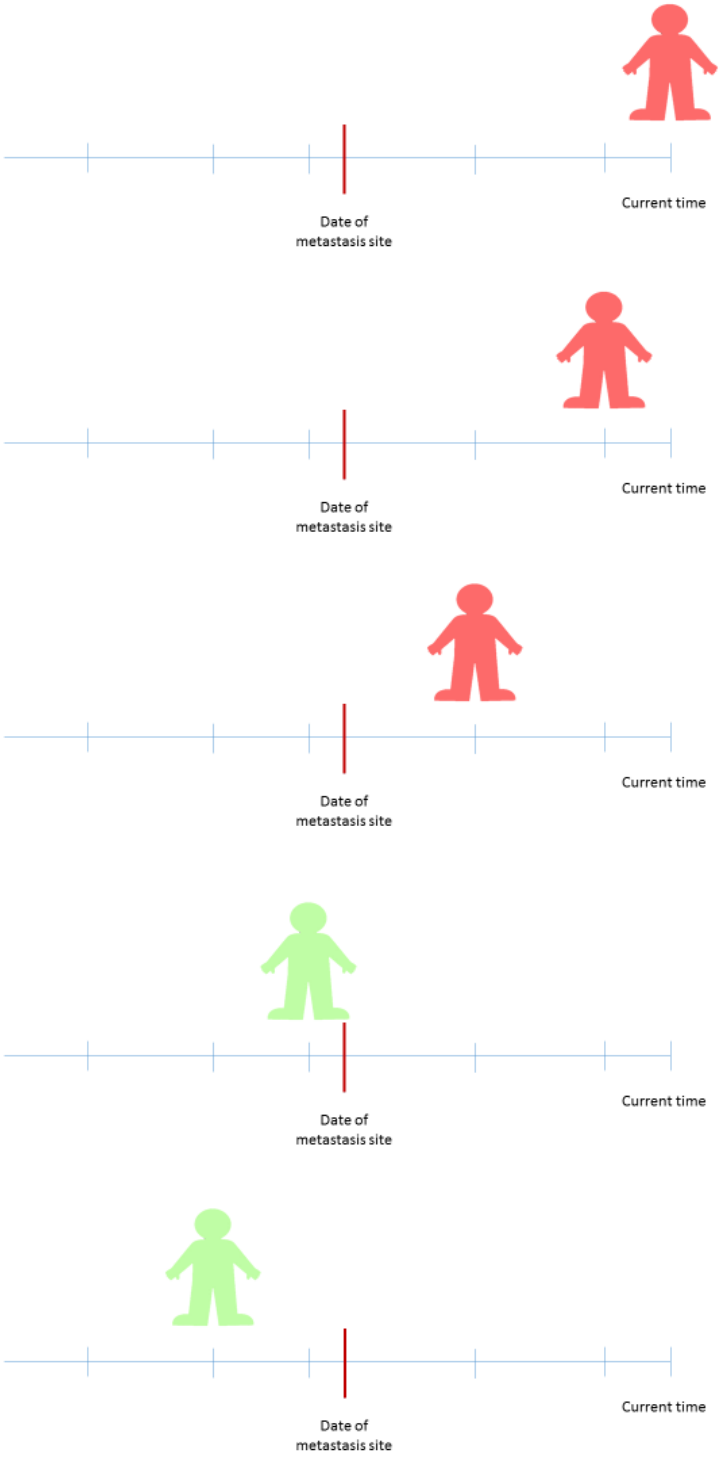


Figure 1: A generic timeline for simulation losing assessment periods where each minor tick mark represents an assessment periods. During each iteration of the simulation, we lose one assessment period at a time. If we lose information before the date of recurrence for a metastatic site then the patient appears to be healthy (highlighted in green) because we no longer have the information about the recurrence. Timelines will differ depending on special characteristics of the patients, such as if the patient died.

Two methods were used to select the sample of simulated lost patients from the non-lost cohort. The first method involved selecting patients from the non-lost cohort to essentially 'replace' those that were lost. For every patient that was flagged lost, a patient within the same propensity score group was randomly selected (with replacement) to replace the lost patient. The second method involved initially taking a simple random sample from the non-lost cohort of similar size to the actual cohort (sampled 1,800 out of 1,937 non-lost patients). In the new sample of non-lost patients, we take a further sample that is equivalent to the percentage of lost patients in each propensity group, which will become our simulated lost patients.

For each simulation, Kaplan-Meier estimates were recalculated for the unadjusted 5-year disease-free survival probabilities and a Cox regression model was used to recalculate the adjusted hazard ratio for the risk of recurrence/death for the lost patients. The 5-year disease free survival probability for non-lost patients remained the same, as no changes were made to non-lost cohort; however the 5-year disease-free survival probability for simulated lost patients was expected to change. Similarly as to the Kaplan-Meier estimates, the hazard ratio for the non-lost remained the same while the hazard ratio for simulated lost patients was expected to change. To stabilize results, the process of resampling from the non-lost cohort, simulating lost patients, and recalculating the Kaplan-Meier estimate for 5-year disease-free survival probability and the hazard ratio for simulated lost patients was bootstrapped 1,000 times for each of the two methods. The estimated 95^{th} percentile confidence intervals for the 5-year disease-free survival probabilities and hazard ratios for these 1,000 replications was calculated for the simulated lost patients and compared to the original 5-year disease-free survival probability and hazard ratio for the actual lost patients.

All tests were two-sided and evaluated at the 0.05 significance level. All data management and analyses were performed in SAS® 9.3. Due to confidentiality, the breast cancer data from the City of Hope cannot be made publicly available.

Programming Methods

In the previous project, the datasets provided by the City of Hope were primarily managed in SAS using DATA steps. To have the code run more efficiently, PROC SQL was used to read-in and merge the data. By implementing PROC SQL we avoided resorting each newly merged dataset and remerge the datasets multiple times (due to the merge statement being able to only merge two datasets at a time).

Another method that was altered to have the code run more efficiently was the bootstrapping process. The bootstrapping algorithm requires taking 1,000 resamples from the non-lost cohort and calculating the unadjusted and adjusted survival rates of the simulated lost patients. Originally, the bootstrapping algorithm was created in a MACRO and ran for 1,000 iterations. This method produced repetition in the code that slowed down the processing time. The new method omits the MACRO and instead takes 1,000 random samples from our non-lost cohort in a single PROC step and then runs losing the assessment periods and survival analyses concurrently on the 1,000 samples using BY group processing.

Results

Preliminary Analysis

Table 1 shows the race/ethnicity proportions for the cohort in for our study. A majority of the patients seen at the City of Hope National Medical Center in Los Angeles were Caucasian (54.5%), with the next highest demographic being Hispanic (24.7%). From previous research, there is an association between the patient's ethnicity and transferring care— non Spanish/Hispanics were more likely to transfer care as compared to those of different race/ethnicities.

Table 1: Frequency of Cohorts' Race/Ethnicity

Race/Ethnicity	Frequency	Percent in Cohort	Percentage Lost in Cohort
White	1285	54.5%	20.3%
Hispanic	583	24.7%	12.9%
Asian	323	13.7%	20.1%
Black	116	4.9%	18.1%
Other	51	2.2%	13.7%

Table 1: Percent in Cohort is percentages of each race/ethnicity in the total cohort (N=2,380). Percentage Lost in Cohort is calculated by the number of patients lost for each race/ethnicity (not in table) divided by the number of patients in the respective race/ethnicity.

Table 2 shows descriptive statistics for the time (in months) when a patient stops following-up. The average time that we are missing information on a lost patient is about 36 months, which is approximately 3 assessment periods. Due to patients being defined as lost if they had not been to a follow-up assessment in two or more years, the lowest estimates for the time that we are missing information on a lost patient can only be 24 months (2 years).

Table 2. Descriptive Statistics of Months of Missing Information for Lost Patients

Months of Missing Information for Lost Patients					
N	Mean	Standard Deviation	Lower Quartile	Median	Upper Quartile
421	35.96	23.62	24.00	24.00	48.00

From Table 3 we can see discrepancies the disease-free survival between non-lost and lost patients in the actual cohort before sampling. Lost patients have a higher median time of survival (6.59 months) as compared to the non-lost patients (3.85 months). The lost patients appear to be healthier with having 91.7% of the patients did not have a distant recurrence

or death as compared to the non-lost patients where 85.7% of patients did not have a distant recurrence or death.

Table 3. Frequencies of Death and Recurrence Statuses by Lost Status

	Non-Lost (N=1937)				Lost (N=421)			
	Alive		Dead		Alive		Dead	
	No Metastasis Site	Metastasis Site	No Metastasis Site	Metastasis Site	No Metastasis Site	Metastasis Site	No Metastasis Site	Metastasis Site
N	1660	61	116	100	386	10	20	5
%	85.7%	3.1%	6.0%	5.2%	91.7%	2.4%	4.8%	1.2%
Median	3.85	2.49	3.13	1.79	6.59	0.85	6.64	2.33

Initially, the entire cohort (N=2,385) was analyzed for the disease-free survival rates stratified by lost status. Figure 2 indicates that lost patients tend to have significantly higher disease-free survival rates than non-lost patients. We estimate that 95.1% of the lost population will survive without recurrence or death for the first 5-years, while 84.2% of the non-lost population will survive without recurrence or death for the first 5-years.

Figure 2: Kaplan-Meier Curves for Disease-free Survival

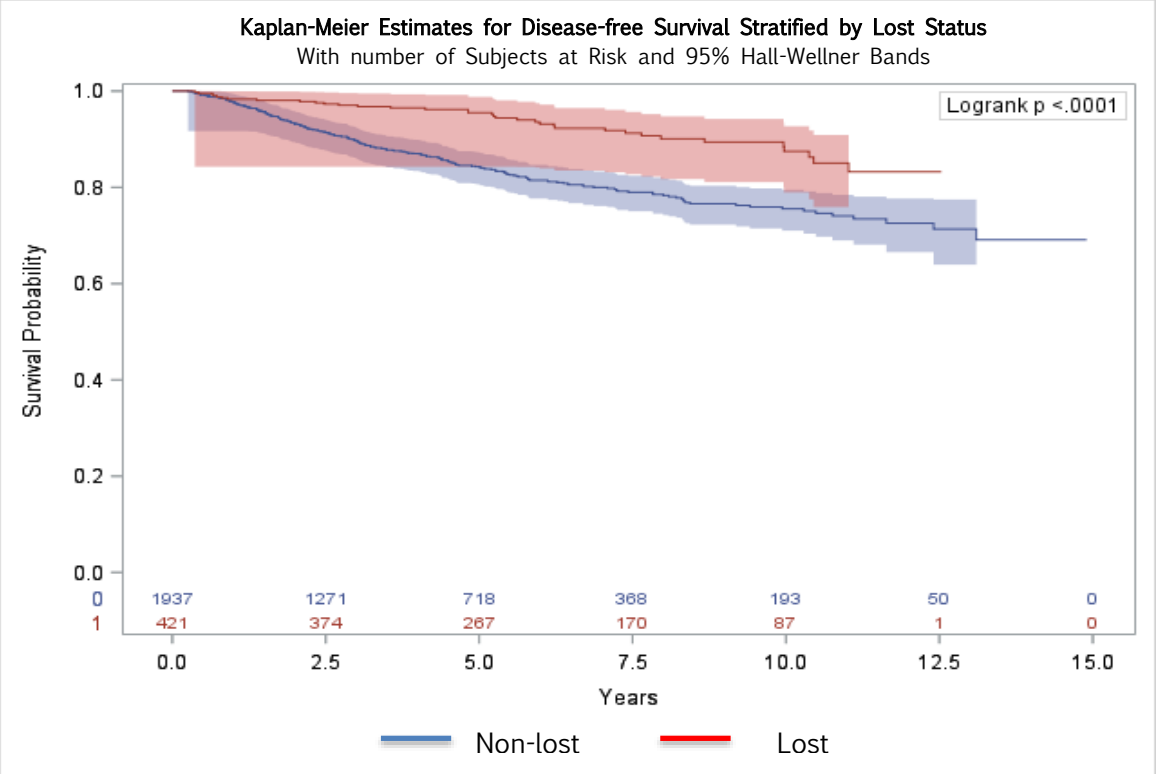


Table 4 concur with the notion that lost patients have a lower hazard of recurrence or death. To adjust for covariates in the presence of lost status, Cox proportional hazards regression was used. After adjusting for age at diagnosis, race/ethnicity, comorbidity score, stage at diagnosis, health insurance type, and LVI, the risk of death or distant recurrence is 61.5% lower for lost patients than non-lost patients (HR = .385, P<0.01). Although race/ethnicity was an insignificant predictor, it remained in the model as a control, due to breast cancer incidence rates being different between races and the difference in follow-up between race/ethnicity.

Table 4: Adjusted Hazard Ratio for Disease-free Survival

Variable (Reference)	**Level**	**Adjusted Hazard Ratio**	**95% CI**	**Overall P-value**
Lost Status (Non-lost)	**Lost**	**0.39**	**(0.27, 0.56)**	<0.01
Age at Diagnosis (Post-Menopausal)	**Elderly**	**1.84**	**(1.26, 2.68)**	<0.01
	Pre-Menopausal	**1.51**	**(1.14, 2.00)**	
Race/Ethnicity (White)	Asian	0.85	(0.58, 1.24)	0.85
	Black	1.09	(0.66, 1.81)	
	Hispanic	0.91	(0.69, 1.21)	
	Other	0.78	(0.24, 2.48)	
Comorbidity Score (Low)	**High**	**3.11**	**(1.90, 5.11)**	<0.01
	Medium	1.29	(0.96, 1.74)	
Stage at Diagnosis (0)	**I**	**6.97**	**(2.30, 21.11)**	<0.01
	II	**10.64**	**(3.47, 32.65)**	
	III	**24.78**	**(8.00, 76.72)**	
*LVI Status (No)	**Yes**	**1.53**	**(1.61, 2.01)**	<0.01
Health Insurance (Managed)	**Medicaid**	**1.69**	**(1.25, 2.27)**	<0.01
	Medicare	**2.14**	**(1.47, 3.11)**	

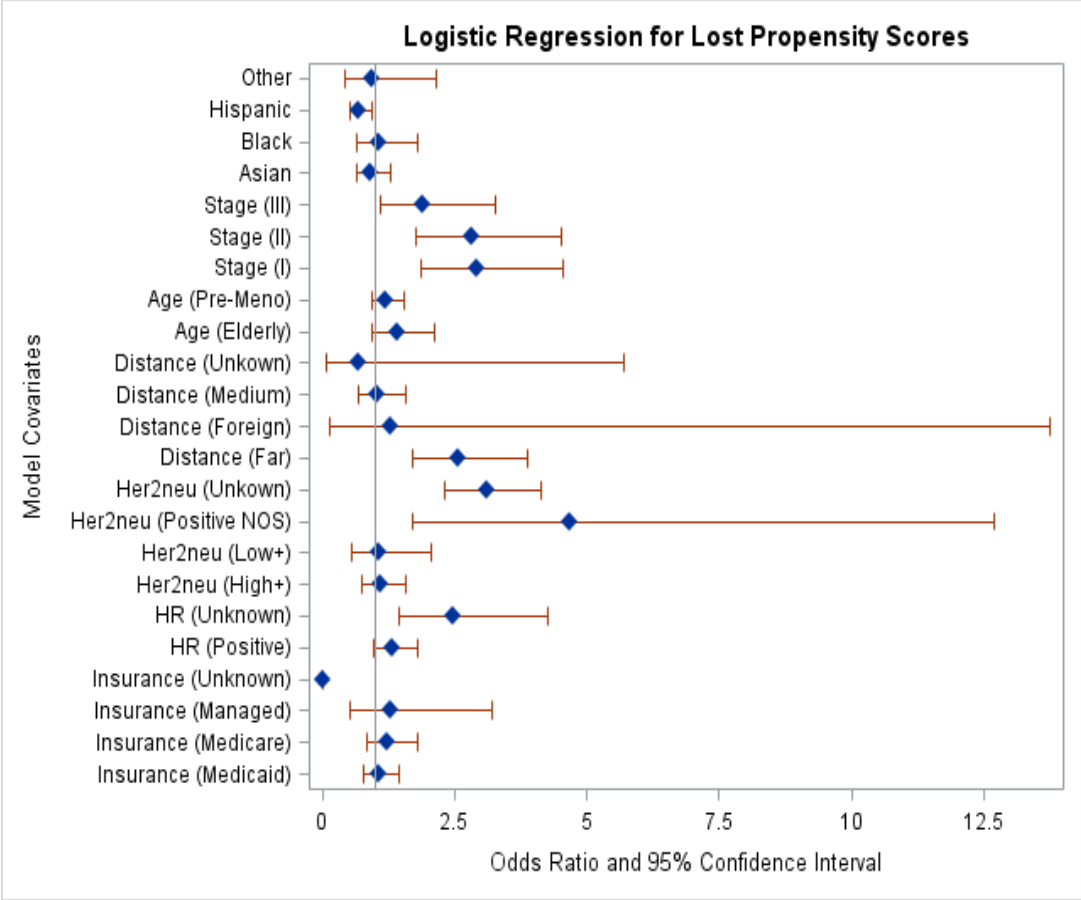
From the Kaplan-Meier estimates and hazard ratio for disease-free survival, it appears that lost patients have a significantly lower hazard of recurrence or death than non-lost patients. This may be a result of missing recurrences— distant recurrences are difficult to ascertain once a patient becomes lost. Therefore, the hazard of recurrence or death may not be accurate due to missing recurrences not being accounted for. Another reason for the differences in the survival rates could be that lost patients actually tend to be healthier, which may be a reason they stopped following-up with City of Hope.

Disease-free Survival for Simulate Lost Patients

To explore if there are underlying reasons that a patient goes lost, we will be simulating "losing patients" based on a sample from our non-lost cohort and comparing the results to our preliminary analysis of the actual lost patients. If those that become lost are due to random chance, then we would expect our simulated lost patients to mirror the survival rates from our actual lost population after losing three assessments.

A logistic regression model was constructed using distance from the medical center, age at diagnosis, stage at diagnosis, race/ethnicity, health insurance type, hormone receptor status, and her2/neu status to predict the probability of being lost. Although health insurance type and age at diagnosis were insignificant predictors, they remained in the logistic regression model due to being significant predictors in the disease-free survival model. Figure 3 and Table 5 shows the variables used in the logistic regression model and the corresponding adjusted odds of becoming lost as compared to the reference group.

Figure 3. Forest Plot of the Adjusted Odds Ratio of Becoming Lost



The reference group for each covariate is as follows:
Race/Ethnicity (White), Distance (Close), Age at Diagnosis (Post-Menopausal), Stage at Diagnosis(0), Hormone Receptor Status (Negative, Her2neu (Negative), Health Insurance (Managed)

Table 5: Associations for Becoming Lost Using Logistic Regression

Variable (Reference)	Level	Adjusted Odds Ratio	95% CI	Overall P-value
Distance (Close)	Medium	1.085	(0.72, 1.64)	<.001
	Far	**2.669**	**(1.78, 4.01)**	
	Foreign	1.289	(0.13,13.34)	
Age at Diagnosis (Post-Menopausal)	Elderly	1.462	(0.98, 2.18)	0.108
	Pre-Menopausal	1.17	(0.91, 1.51)	
Stage (0)	**I**	**2.915**	**(1.87, 4.55)**	<.001
	II	**2.784**	**(1.74, 4.45)**	
	III	**1.873**	**(1.09, 3.24)**	
*Hormone Receptor Status (Negative)	Positive	1.343	(0.994, 1.81)	0.004
**Her2/neu Status (Negative)	High+	1.093	(0.75, 1.59)	<.001
	Low+	1.081	(0.57, 2.07)	
	Positive NOS	**4.861**	**(1.80, 13.17)**	
Race/Ethnicity (White)	Asian	0.909	(0.65, 1.28)	.154
	Black	1.070	(0.64, 1.79)	
	Hispanic	**0.683**	**(0.51, 0.92)**	
	Other	0.941	(0.42, 2.15)	
Health Insurance (Managed)	Medicaid	1.060	(0.77, 1.45)	.149
	Medicare	1.234	(0.85, 1.79)	
	Other	1.296	(0.53, 3.20)	

* 184 patients with unknown hormone receptor status
** 678 patients with unknown her2/neu status

The probabilities calculated from the logistic regression model were used as the propensity scores for grouping patients with similar characteristics in terms of becoming lost. The patients were sorted and separated into 20 similar sized groups based on the similar propensity scores. Within each group, the frequencies of lost patients were calculated and used for sampling the patients that will become simulated lost. Table 6 shows the distribution of lost and non-lost patients within each group.

Table 6: Distribution of Lost and Non-lost Within the Propensity Score Groups

Group	Lost	Non-Lost	% Lost	Total
1	8	110	6.8%	118
2	9	109	7.6%	118
3	12	106	10.2%	118
4	13	105	11.0%	118
5	12	106	10.2%	118
6	17	101	14.4%	118
7	14	104	11.9%	118
8	11	107	9.3%	118
9	16	102	13.6%	118
10	22	96	18.6%	118
11	13	105	11.0%	118
12	18	100	15.3%	118
13	17	101	14.4%	118
14	26	92	22.0%	118
15	18	100	15.3%	118
16	25	93	21.2%	118
17	32	86	27.1%	118
18	46	72	39.0%	118
19	41	77	34.7%	118
20	51	65	44.0%	116
Total	421	1937	17.9%	2358

Table 7: 95th Percentile Confidence Intervals of Disease-free Survival Rates Based on Simulated Number of Assessment Lost

	Bootstrapping Method 1		Bootstrapping Method 2	
Number of Assessments Lost	**95 Percentile CI of Hazard Ratios of Death/Recurrence**	**95 Percentile CI of 5-Year Disease-free Survival**	**95 Percentile CI of Hazard Ratios of Death/Recurrence**	**95 Percentile CI of 5-Year Disease-free Survival**
0 (non-lost)	(0.85, 1.32)	(0.79, 0.87)	(0.85, 1.35)	(0.78, 0.88)
1 (non-lost)	(0.88, 1.39)	(0.78, 0.87)	(0.88, 1.39)	(0.77, 0.87)
2	(0.90, 1.41)	(0.77, 0.86)	(0.90, 1.43)	(0.75, 0.87)
3	(0.88, 1.39)	(0.76, 0.86)	(0.87, 1.41)	(0.75, 0.87)
4	(0.87, 1.38)	(0.77, 0.87)	(0.85, 1.39)	(0.75, 0.88)
5	(0.78, 1.31)	(0.79, 0.90)	(0.75, 1.30)	(0.78, 0.91)
6	(0.72, 1.23)	(0.82, 0.93)	(0.69, 1.26)	(0.81, 0.93)
7	(0.59, 1.09)	(0.91, 0.98)	(0.57, 1.11)	(0.91, 0.99)
8	(0.55, 1.05)	(0.94, 0.99)	(0.53, 1.10)	(0.94, 1.00)
9	(0.44, 1.01)	(0.93, 1.00)	(0.42, 1.05)	(0.92, 1.00)
10	(0.39, 0.98)	(0.90, 1.00)	(0.38, 1.10)	(0.89, 1.00)

For the bootstrapping methods, we want to simulate losing patients from our non-lost cohort. Before we begin the simulations, the lost patients were removed, leaving only non-lost patients in the sampling cohort.

The bootstrapping algorithms consisted of simulating lost patients by sampling non-lost patients with replacement from each propensity score group. By removing assessment periods from those sampled patients, they become simulated lost patients and any distant recurrences that may have taken place during those years became unknown. Kaplan-Meier estimates for 5-year disease-free survival and hazard ratios were recalculated based on the simulated lost patients. Table 7 shows the recalculated 95th percentile confidence intervals for the hazard ratio and 5-year disease-free survival rates based on the number of assessment periods lost for both bootstrapping methods.

Both sampling methods for the bootstrapping process produced similar results for the 95th percentile confidence of the hazard ratios and 5-year disease free survival. Recall that for actual lost patients, the average number of missing assessments was three assessments and the hazard ratio and 5-year disease free survival rate were 0.385 and 0.951 respectively. After simulating losing three assessments, the 95th percentile confidence

intervals do not capture the disease-free survival rates for lost patients— the simulated lost patients have a higher hazard of death/recurrence and a lower 5-year disease-free survival rate than the actual lost patients. The only 95th percentile interval that nearly captures the actual cohort hazard ratio is the simulation that loses 10 assessment periods. However, the 5-year disease-free survival rate for lost patients was captured in the 95th percentile confidence interval after losing at least 7 assessment periods. This indicates that a higher than average number of assessments are needed to be lost to capture the disease-free survival rates of the actual lost patients. Therefore, the differences in disease-free survival rates between non-lost and actual lost patients seem to be not only due to missing information on distant recurrence.

Conclusion

The Kaplan-Meier estimates for disease-free survival and adjusted hazard ratios that were run on the actual cohort of 2,358 patients suggest that lost patients have lower hazard of recurrence or death than non-lost patients. However, we wanted to examine if the differences in the disease-free survival rates between the two cohorts were actually due to missing recurrence information or another underlying factor, such as lost patients tending to be healthier than non-lost patients. If the difference was due to simply missing information on the lost patients, then we expected 5-year disease-free survival rates and hazard ratio for simulated lost patients to concur with the results from the actual lost patients. Through bootstrapping simulation, we found that a higher than average number of assessments needed to be lost to capture the disease-free survival rates of the actual lost patients. This indicates that the differences in disease-free survival rates between non-lost and actual lost patients is not due to missing recurrence information alone. It is likely that lost patients may actually be healthier than their non-lost counterparts— which could be a reason that the patients stopped following-up with City of Hope.

References

Huang, Debbie Y. (2013) Estimating the Impact of Lost to Follow-up on Breast Cancer Patient's Disease-free Survival. (Senior project, California Polytechnic State University San Luis Obispo). Retrieved from http://digitalcommons.calpoly.edu/statsp/34/

"About Us." *City of Hope*. City of Hope, n.d. Web. <http://www.cityofhope.org/about-city-of-hope>.

"NCI Dictionary of Cancer Terms." *National Cancer Institute*. N.p., n.d. Web. 10 March 2016.

SAS Maps Online.(2016). Zipcode (V8 and V9) [Data file]. Retrieved from <http://support.sas.com/rnd/datavisualization/mapsonline/html/misc.html>.

"Surveillance Research Program." *Cancer Survival Statistics* -. N.p., n.d. Web. 10 May 2016. < http://surveillance.cancer.gov/statistics/types/survival.html>.

“Usage Note 5325: Calculating the distance between ZIP codes.” SAS. SAS, 04 June 2009. Web. 05 March. 2016. <http://support.sas.com/kb/5/325.html>.

Appendix

SAS Code Data Management

```
options rightmargin=1in     leftmargin=1in   topmargin=1in
bottommargin=1in   nodate nonumber ls=85 center     FORMDLIM="-";

libname raw "C:\Users\Megan\Desktop\Senior Project\Senior Project
Information\Data\Raw";
run;

libname derived "C:\Users\Megan\Desktop\Senior Project\Senior Project
Information\Data\Derived";
run;

libname zip "C:\Users\Megan\Desktop\Senior Project\Zip Codes";
run;

libname sp "C:\Users\Megan\Desktop\Senior Project\SP";
run;

libname spm1 "C:\Users\Megan\Desktop\Senior Project\SP\SPM1";
run;

libname spm1CI "C:\Users\Megan\Desktop\Senior Project\SP\SPM1\CI";
run;

libname spm1done"C:\Users\Megan\Desktop\Senior
Project\SP\SPM1\FinalSets";
run;


libname spm1CI2 "C:\Users\Megan\Desktop\Senior Project\SP\SPM1v2\CI2";
run;

libname newfmt "C:\Users\Megan\Desktop\Senior Project\Senior Project
Information\Data\Formats";
run;

Options fmtsearch=(raw.formats derived.ddformats);
run;


proc sql;
create table sp.bcaTEST2 as
select *
from derived.clinical_characteristics as d
         LEFT JOIN raw.diagnosis as r
            ON r.pid = d.pid AND r.dxid = d.dxid


             LEFT JOIN derived.patient_characteristics as dpc
                ON d.pid = dpc.pid AND d.dxid = dpc.dxid
```

```
              LEFT JOIN raw.metastatic_sites as rms
                   ON d.pid = rms.pid AND d.dxid = rms.dxid

               LEFT JOIN raw.demographics as rd
                   ON d.pid = rd.pid

               LEFT JOIN raw.study_accession as rsa
                   ON d.pid = rsa.pid

               LEFT JOIN raw.breast_diagnosis as rbd
                   ON d.pid = rbd.pid AND d.dxid = rbd.dxid AND
d.tumorid = rbd.tumorid

               LEFT JOIN raw.solid_tumor_stage as rst
                   ON d.pid = rst.pid AND d.dxid = rst.dxid AND
d.tumorid = rst.tumorid

  order by pid
  ;
 quit;


 data sp.bcaTEST3;
   set sp.bcaTest2;
    by pid;

*The variable initial is binary where 0=no and 1=yes and is formatted
with yes/no;
      *Is causing a lot of patients to be dropped;
      if initial = 0 or initial = .;

            if lostflag = 1 then lost =1;
        else lost = 0;

years = yrdif(dxdt, osdt,  'act/act');

length insur $10. Comorbid $10. Race $15. HR_grp $10. Her2neu_grp $15.
age_grp $10.;

   /*create age groups pre-meno, post-meno, & elderly*/
      if agedx < 50 then age_grp="Pre-Meno";
      else if 50 <= agedx <= 70 then age_grp="Post-Meno";
      else if agedx >70 then age_grp="Elderly";

      if stage_final=22 then stage=0;
      else if stage_final=23 | stage_final=23.1 | stage_final=23.2 then
stage=1;
      else if stage_final=23.5 | stage_final=24 | stage_final=25 then
stage=2;
      else if stage_final=25.5 | stage_final=26 | stage_final=27 |
stage_final=27.5 then stage=3;
      if stage=. then delete;

      if race_eth=1 then race="White";
      else if race_eth=3 then race="Black";
      else if race_eth=5 then race="Asian";
```

```
      else if race_eth=7 | race_eth=9 | race_eth=10 then race="Other";
      else race="Hispanic";

      if comorbidity=0 then comorbid="Low";
      else if comorbidity=1 | comorbidity=2 then comorbid="Med";
      else comorbid="High";

      if insurance=1 then insur="Managed";
      else if insurance=5 | insurance=5.5 | insurance=5.75 then
insur="Medicare";
      else if insurance=4 then insur="Medicaid";
      else if insurance=0 | insurance=2 | insurance=6 then
insur="Other";
      else if insurance=-1 | insurance=. then insur="Unknown";

      if HR=0 then HR_grp="Negative";
      else if HR=1 then HR_grp="Positive";
      else HR_grp="Unknown";

      if her2neu=1 then Her2neu_grp="Negative";
      else if her2neu=2 then Her2neu_grp="Low +";
      else if her2neu=3 then Her2neu_grp="High +";
      else if her2neu=4 then Her2neu_grp="Positive NOS";
      else Her2neu_grp="Unknown";

 /* distant met: intra-ab 5, bone 6, lung 7, pleural effusion 8,
pericardial effusion 9,liver 10, bone marrow 11, brain/cns 12, LN other
distant 15,  LN other distant visceral 17, LN other distant non-
visceral 18, skin 20, contralateral breast 14, Contralateral
supraclavicular nodes 19, Ipsilateral supraclavicular nodes 16 ONLY IF
DIAGNOSED AFTER JAN 01, 2003  Meninges 13 but deactivated & changed to
brain/cns 12
                    */
      if site=5 | site=6 | site=7 | site=8 | site=9 | site=10 | site=11
| site=12 | site=15 | site=17 | site=18 | site = 19 | site=20 | site=14
then met=1;
      else if site=16 & sitedt >= '01jan2003'd then met=1;
      else met=0;

*** evt = new event variable for death/recurrence;
      if met=1 | event=1 then evt=1;
      else evt=0;

***** censordt = date of first met/event/censor;
      if sitedt=. then cendt=osdt;
      else if osdt <= sitedt then cendt = osdt;
      else if sitedt < osdt then cendt= sitedt;
  format cendt DATE9.;

  censordt= yrdif(dxdt, cendt, 'act/act');


      if deathicd=" " then dead=0;
      else dead=1;

   /* ICD codes for breast specific death 174, C50, 233 */
```

```
      if substr(deathicd, 1, 3)= "174" | substr(deathicd, 1, 3)= "C50"
|     substr(deathicd, 1, 3)= "233" then breastdeath=1;
      else breastdeath=0;

  ******* education status ******;
      if edustat=0 | edustat=1 | edustat=2 then edu="Other/Less than HS
Grad";
      else if edustat=3 then edu="HS Grad";
      else if edustat=4 | edustat=5 then edu="AA/Tech";
      else if edustat=6 | edustat=7 then edu="College Grad";
      else edu="Unknown";

  ******* Employment status at diagnosis;
      if empstatdx=1 | empstatdx=2 | empstatdx=3 | empstatdx=4 |
empstatdx=5  then employ="Employed/Student";
      else if empstatdx=9 then employ="Unemployed";
      else employ="Other";
  **** Other: homemaker, medical leave, retired, disabled;

  *** BMI = weight(in KG)/(height*height(in M));   *** height in CM &
weight in KG;
      if heightpres=-1 | weightpres=-1 then BMI="Unknown";
      else BMI =  weightpres / (0.01*heightpres*0.01*heightpres);

 length BMI_grp $15.;
      if BMI="Unknown" then BMI_grp="Unknown";
      else if 0 <= BMI < 18.5 then BMI_grp="Underweight";
      else if 18.5 <= BMI < 25 then BMI_grp="Normal";
      else if 25 <= BMI < 30 then BMI_grp="Overweight";
      else if BMI >= 30 then BMI_grp="Obese";

length grade_grp $15. LVI_grp $10;
      if stage=0 then grade=put(dcishistogrd, 8.);
      else if stage=1 | stage=2 | stage=3 then grade=put(invcahistogrd,
8.);


      if grade=1 then grade_grp="Low";
      else if grade=2 then grade_grp="Intermediate";
      else if grade=3 then grade_grp="High";
      else grade_grp="Unknown";

 LVI=lymphvascinv;
      if LVI=0 then LVI_grp="No";
      else if LVI=1 then LVI_grp="Yes";
      else LVI_grp="Unknown";

zipcode=zip; ****** rename zip for merging later;

drop createdby modifiedby cid studyid;

run;

proc sort data=sp.bcatest3;
      by pid evt descending sitedt;
run;
```

```
data sp.bcaTest4;
      set sp.bcatest3;
      by pid;
      if last.pid;
run;



/** zip codes from SAS Maps Online to create distance variable **/
  ***** Imports and outputs zipcode SAS datasets ****;
proc cimport infile="C:\Users\Megan\Desktop\Senior Project\Zip
Codes\zipcode_Jan13_v9.cpt"
      lib=zip;
run;



data sp.zip.Zip1;
      set zip.zipcode_13q1_unique;
run;


  ****** Final zip code data set = zip.zipcode *************;
proc sort data=zip.Zip1 out=zip.zipcode;
      by zip zip_class;
run;

/*Producing writen output*/
  ******** recreate the index ******;
proc datasets lib=zip;
      modify zipcode;
      index create zip;
run;



  **** X = longitude in degrees & Y = latitude in degrees;   *** COH
NCCN zip = 91010, lat=34.1357  long=-117.9655;
data sp.zipp;
      set zip.zipcode;  zipcode = put(zip, 5.);
*** convert zipcode to character to match zipcode in BCA data set;
            drop zip;
            rename zipcode=zip;
run;

proc sort data=sp/bcatest4;
      by zip;
run;

data sp.final;
 merge sp.bcatest4(IN=clinical) sp.zipp;
 by zip;
 if clinical=1;

      *** Convert long & lat from degrees to radians;
      COH_long = atan(1)/45 * -117.9655;
      COH_lat = atan(1)/45 * 34.1357;
```

```
      long = atan(1)/45 * X;
      lat = atan(1)/45 * Y;

   *** dist in miles;

      *** Great Circle Distance Formula;
      dist = 3949.99 * arcos(sin(lat) * sin(COH_lat) + cos(lat) *
cos(COH_lat) * cos(long - COH_long));

      Length dist_grp $8.;

       if 0 < dist <= 50 then dist_grp="Close";
       else if 50 < dist <= 100 then dist_grp="Medium";
       else if 100 < dist then dist_grp="Far";
       else if zip=0 | zip=1 | zip=2 then dist_grp="Foreign";
       else dist_grp="Unknown";
 run;
```

SAS Code Preliminary Analysis

```
/****Risk of Death and Breast Cancer Recurrence Non-lost vs. Lost****/

proc freq data = sp.final;
      tables lost
              age_grp
              race
              comorbid
              stage
              grade_grp
              HR_grp
              Her2neu_grp
              LVI_grp
              insur
              employ;
run;

/*Median number of missing assessment periods for lost patients*/
data numMissingAssess;
      set sp.final;
            if missassess ^= . and firstlost ^=.;
                  NumMissingAssess = missassess - firstlost;
run;

proc means data = numMissingAssess n median mean std clm q1 q3;
      var NumMissingAssess;
run;


/*Median number of years before a patients becomes lost*/
data lostpopulation;
      set sp.final;
    if lost = 1;
run;
proc means data = lostpopulation n mean median std;
      var censordt;
run;


/*Finding the median number of months for when patients first stop
following-up*/
proc means data = sp.final n mean median std;
  var firstlost;
run;

proc sort data = sp.final out=sort;
   by PID;
run;

proc phreg data = sp.final outtest=testing;
      title "Cox Regression for Risk of Death and Breast Cancer
Recurrence";
      class lost(ref="0")
```

```
                age_grp(ref="Post-Meno")
                race(ref="White")
                comorbid(ref="Low")
                stage(ref="0")
                grade_grp(ref="Low")
                HR_grp(ref="Negative")
                Her2neu_grp(ref="Negative")
                LVI_grp(ref="No")
                insur (ref="Managed")
        model censordt*evt(0) = lost age_grp race comorbid stage LVI_grp
              insur;
        strata grade_grp HR_grp Her2neu_grp;
        assess ph / resample;

        hazardratio lost / CL=WALD DIFF=REF;
        hazardratio age_grp / CL=WALD DIFF=REF;
        hazardratio race / CL=WALD DIFF=REF;
        hazardratio comorbid / CL=WALD DIFF=REF;
        hazardratio stage / CL=WALD DIFF=REF;
        hazardratio LVI_grp / CL=WALD DIFF=REF;
        hazardratio insur / CL=WALD DIFF=REF;
        hazardratio employ / CL=WALD DIFF=REF;
run;



proc lifetest data = sp.final plots=survival(atrisk=0 to 15 by 2.5
CB=HW test nocensor) Method=KM outsurv = sp.Surv ;
        time censordt*evt(0);
        strata lost / test=logrank;
        label censordt = "Years";
        title "KM Estimates for Disease-free Survival Stratified Lost
Status";
run;


proc sort data = sp.Surv;
        by lost censordt;
run;

data sp.FiveYearSurv;
        set sp.Surv;
        by lost censordt;
                if censordt >= 5 and survival ^=. then output;
run;


proc sort data=sp.FiveYearSurv;
        by Replicate;
run;


data testing2;
        set testing;
                HRLost = exp(lost1);
run;
```

```
proc univariate data = testing2;
      histogram HRLost;
      Title "Histogram of Estimated Hazard Ratios";
run;

data sp.freqtable;
      set sp.final;
      if deathdt ^= . then dead = 1;
      else dead = 0;
      if sitedt=. then metsite=0;
      else metsite=1;
run;

proc freq data = sp.freqtable;
      tables lost*dead;
run;

proc means data=sp.freqtable median mean;
var censordt;
class lost dead met;
run;



/******** Propensity Scores for LTFU on DFS **********/;

/*Need command to make Forest Plot*/
ods output "Odds Ratios" = ORci;

/*Logistic Model(same variables as phreg model)*/
proc logistic data=sp.final descending;
      title "Logistic Regression for Lost Propensity Scores";
      class dist_grp(ref="Close")
              age_grp(ref="Post-Meno")
              race(ref="White")
              edu(ref="College Grad")
              employ(ref="Unemployed")
              insur(ref="Managed")
              comorbid(ref="Low")
              stage(ref="0")
              grade_grp(ref="Low")
              HR_grp(ref="Negative")
              Her2neu_grp(ref="Negative")
              LVI_grp(ref="No")/ param=ref;
      model lost = dist_grp age_grp stage HR_grp Her2neu_grp insur race
/ lackfit;
   score out=sp.scores;
run;
ods output off;
/*Note: insur is not significant in the logistic model, but still
include it because they are used in the Cox Regression models later
(the variable is significant when accounting for hazard of recurrence).
Because those variable is of interest in latter models we are keeping
them in to help sort the patients into their respective propensity
score buckets*/
```

```
/**FOREST PLOT**/
/*Dataset used for making the forest plots. Contains the odds ratio
est, lower CL, and upper CL for each variable in the logistic
regression model*/
data orci;
      set orci;
      effect = UPCASE(effect);

/*Renaming the variables for aethetics reasons for the forest plot*/
      if effect = "DIST_GRP    FAR     VS CLOSE" then effect =
"Distance (Far)";
      else if effect = "DIST_GRP    FOREIGN VS CLOSE" then effect =
"Distance (Foreign)";
      else if effect = "DIST_GRP    MEDIUM  VS CLOSE" then effect =
"Distance (Medium)";
      else if effect = "DIST_GRP    UNKNOWN VS CLOSE" then effect =
"Distance (Unkown)";

      else if effect = "AGE_GRP     ELDERLY  VS POST-MENO" then effect
= "Age (Elderly)";
      else if effect = "AGE_GRP     PRE-MENO VS POST-MENO" then effect=
"Age (Pre-Meno)";

      else if effect = "STAGE       1 VS 0" then effect = "Stage (I)";
      else if effect = "STAGE       2 VS 0" then effect = "Stage (II)";
      else if effect = "STAGE       3 VS 0" then effect = "Stage
(III)";

      else if effect = "HR_GRP      POSITIVE VS NEGATIVE"  then effect
= "HR (Positive)";
      else if effect = "HR_GRP      UNKNOWN  VS NEGATIVE" then effect =
"HR (Unknown)";

      else if effect = "HER2NEU_GRP HIGH +       VS NEGATIVE" then
effect = "Her2neu (High+)";
      else if effect = "HER2NEU_GRP LOW +        VS NEGATIVE" then
effect = "Her2neu (Low+)";
      else if effect = "HER2NEU_GRP POSITIVE NOS VS NEGATIVE" then
effect = "Her2neu (Positive NOS)";
      else if effect = "HER2NEU_GRP UNKNOWN      VS NEGATIVE" then
effect = "Her2neu (Unkown)";

      else if effect = "INSUR       MEDICAID VS MANAGED" then effect =
"Insurance (Medicaid)";
      else if effect = "INSUR       MEDICARE VS MANAGED" then effect =
"Insurance (Medicare)";
      else if effect = "INSUR       OTHER    VS MANAGED" then effect =
"Insurance (Managed)";
      else if effect = "INSUR       UNKNOWN  VS MANAGED" then effect =
"Insurance (Unknown)";

run;

/*The Forest Plot*/
proc sgplot data = orci;
```

```
      scatter x = oddsratioest y=effect / xerrorlower = lowercl
                                                                  xerrorupper
= uppercl
                                                                  markerattrs
= or

      (symbol=DiamondFilled size = 8);
      refline 1 / axis = x;

      xaxis label = "Odds Ratio and 95% Confidence Interval" min=0;
      yaxis label = "Model Covariates";
run;


proc sort data = sp.final out =sp.final_sortedlost;;
      by lost;
run;

proc freq data = sp.final_sortedlost;
      tables race;
      by lost;
run;


/***** Simulating Lost Patients: Setting Up Propensity Buckets ****/
   *** group by propensity scores;
proc sort data=sp.scores;
      by P_1;
run;

data sp.bucket;
set sp.scores;
 total = 2358;  /*total number of rows in sp.scores*/
 size = round(total/20);  /*Assiging patients into equal size
propensity score groups called "buckets"*/
      if 1 <= _N_ <= size then bucket=1;
        else do i= 2 to 20;
            if size*(i-1) < _N_ <= i*size then bucket=i;
        end;
run;

  *** save frequencies/proportions of lost;
proc freq data=sp.bucket;
      table lost*bucket / nopercent norow nocol out=dist;
run;

data sp.dist;
  set dist;
      if lost=1;
run;

proc sort data = sp.bucket;
      by bucket;
run;

  *** non-lost cohort;
data sp.nonlost;    *** includes all variables;
```

```
      merge sp.dist sp.bucket;
      by bucket;
        if lost=0;      *** non-lost only (n=1684);
    _NSIZE_=count;   *** _NSIZE_ needed for proc surverselect;
run;

proc sort data=sp.nonlost;
      by pid;
run;

proc sort data=raw.Continuous_Status;
      by pid descending assessid;
run;

data sp.constat;
      merge sp.nonlost(IN=non) raw.Continuous_Status;
      by pid;
       if non=1;    *** get assessid info for nonlost patients;
run;

proc sort data=sp.nonlost_population;
      by bucket pid;
run;

proc sort data =sp.final out = sp.finalSorted;
      by lost;
run;
proc freq data = sp.finalSorted;
      tables race;
      by lost;
run;
```

SAS Code Bootstrapping Method 1 and DFS Analysis

```
/**********************Bootstrapping Method 1**********************/

%let numreps = 1000;
%let k =2; /* change k to go back different assessment periods, k=2
goes back 1 assessment periods, k= 3 goes back 2 assessment periods,
etc.. do for k = 1*/


/* sampling from each bucket the number of lost patients we removed to
get back orginal sample size but with only non-lost patients */
/*use reps= to specify the number of sample replications*/
/*the variable: Replicate  indicates the sample replicate number during
the surveyselect (is alread sorted in descending order)*/
proc surveyselect data=sp.nonlost_population
                          method=URS
                          seed=1
                          sampsize=sp.nonlost_population
                          reps = &numreps
                          outhits
                          out = spm1.select;

      strata bucket;

      title "Sampling With Replacement From Non-lost Patients:
Testing";
run;


data spm1.sample;   /* Simulated lost patients */
      set spm1.select;
      lost=1;

      /*Creating a new ID variable that will be unique for each
      patient.
      For lost patients it will be the original PID with L (for lost)
concatenated onto it eg: ##L*/
      Newid = cats(PID, "_", "L");

run;

/*Need to sort sample by new ID to set up the dataset for the counter
variable in the next block of code*/
proc sort data = spm1.sample;
      by pid newid descending assessid;
run;


data spm1.ID;
set spm1.sample;
      by pid;
```

```
      /*Recoding new id to make them unique for each observation
       These new ids will have a counter concated to the new id eg:
      50L1,50L2, 50L3 */
      if first.pid then do;
            counter = 1;
            newid_unique=cats(newid,"_",counter);
      end;

      else do;
            counter+1;
            newid_unique = cats(newID,"_",counter);
       end;
run;

/*For getting the assessment periods*/
data spm1.cont_status;
      set raw.Continuous_status;
run;
proc sort data = spm1.cont_status;
      by pid descending assessid;
run;


/*Setting data up so we can lose assesment periods*/
proc transpose data=spm1.cont_status out=spm1.cs;
  by pid;
  var assessid;
run;

data GetSimLastAssess;
      set spm1.cs;

      Simlastassess=largest(&k, OF COL1-COL17); /*k=2 to go back 1
assessment period, change k for going back more periods*/

      drop COL1-COL17;
run;



data GetSimLastAssess2;
      merge GetSimLastAssess spm1.cont_status;
      by pid;
      if SimLastAssess = assessid;

      lostdt = datacollectdtdm;
      vitalstatdt2 = vitalstatdt;
      vitalstat2 = vitalstat;
      nccncarestatdt2 = nccncarestatdt;
      dzstatdt2 = dzstatdt;

      if vitalstatdt2 > lostdt then delete;
      /*this should not happen, would be caused by error of inputting
information*/
      /*one person (pid 256622) in this dataset has a lostdt <
vitalstatdt2*/
```

```
      keep pid lostdt vitalstatdt2 vitalstat2 nccncarestatdt2
dzstatdt2;
      format lostdt DATE9. vitalstatdt2 DATE9. nccncarestatdt2 DATE9.
dzstatdt2 DATE9.;
run;



data spm1.id2;
  merge GetSimLastAssess2 spm1.id (in=a);
  by pid;
  if a;
run;

/*For finding the sitedt -- First date of diagnosis of site of
disease.*/
data spm1.met_site;
      set raw.metastatic_sites;
      keep pid sitedt;
run;
proc sort data = spm1.met_site;
      by pid;
run;

/*preparing data for the merge*/
proc sort data = spm1.id2;
      by pid;
run;


/*Creating a new dataset that has the lost data and the date of the met
site*/
data spm1.id3;
  merge spm1.met_site spm1.id2 (in=b);
  by pid;
  if b;
run;

proc sort data=spm1.id3;
 by newid_Unique;
run;


data spm1.DFS;
      set spm1.id3;
      by newid_unique;

if lostdt >= dxdt then do;
  if deathdt = . then do;
        if lostdt >= sitedt and sitedt ^= . then do;  /*lostdt is the
date of the most recent information we have on them and sitedt first
date of diagnosis of site of disease.*/
            newmet = 1; /*recurrance has happened*/
             newcendt = sitedt;
            newcensordt = yrdif(dxdt, newcendt, 'act/act'); /*dxdt is
the date of diagnosis*/
        end;
```

```
        else if lostdt < sitedt and sitedt ^= . and vitalstat2 = 1 then
do;
            newmet = 0;
            newcendt = max(vitalstatdt2, NDidt);
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
       end;
       else if lostdt < sitedt and sitedt ^= . and vitalstat2 = . then
do;
            newmet = 0;
            newcendt = max(dzstatdt2, NDidt);
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
       end;
       else if sitedt = . and vitalstat2 = 1 then do;
            newmet = 0;
            newcendt = max(vitalstatdt2, NDidt);
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
       end;
       else if sitedt = . and vitalstat2 = . then do;
            newmet = 0;
            newcendt = max(dzstatdt2, NDidt);
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
       end;
   end;



 if deathdt ^= . and sitedt ^= .  then do;
  event = 1;
      if lostdt >= sitedt then do;
            newcendt = sitedt;
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
      end;
      else if lostdt < sitedt then do;
          newcendt = deathdt;
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
      end;
  end;

  else if deathdt ^= . and sitedt = . then do;
            event =1;
            newcendt = deathdt;
            *newcendt = osdt;
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
  end;
end;



else if lostdt < dxdt then delete;


   if newmet=1 | event=1 then newevt=1;
   else newevt=0;  /* newevt = new event variable for death/recurrence
*/

       keep newid_unique pid  lost newevt newcensordt replicate bucket
                  /*Variables important to the Cox Model*/
```

```
                age_grp race comorbid stage grade_grp HR_grp Her2neu_grp
LVI_grp insur employ years
                   vitalstatdt lostdt censordt datacollectdtdm

                   newcendt deathdt osdt sitedt;
run;


/*Creating 1000 identical datasets to sp.nonlost_population that will
be all in the same dataset
      (nonlost_population_reps) and creating a new variable, Replicate,
that keeps track of the replication number
      for each patient and will be used with the variable Replicate
from the proc survyselect*/
data spm1.nonlost_population_reps;
      set sp.nonlost_population;
            do Replicate=1 to &numreps;
                  ** make newevt and newcensordt here, same as original
data;
                  newevt = evt;
                  newcensordt = censordt;
            output;
      end;
run;


/*Creating a dataset will all patients (both nonlost and simulated
lost)*/
data spm1done.simulated_all_&k;
      set spm1.nonlost_population_reps spm1.DFS;
run;

proc sort data = spm1done.simulated_all_&k;
      by Replicate pid;
run;


/*Finding Kaplan-Meier Estimates of disease free survival on both
nonlost and simulated lost patients*/
ods graphics off;
proc lifetest data = spm1done.simulated_all_&k outsurv=
spm1done.Surv_&k noprint;
      by Replicate;
      time newcensordt*newevt(0);
      strata lost / test=logrank;
      label newcensordt = "Years";
      title "KM Estimates of DFS by Lost Status: SIMULATED Lost
Patients";
run;
ods graphics on;

proc sort data = spm1done.surv_&k;
      by lost newcensordt;
run;

data spm1done.Simulated_Survival5yr_&k;
      set spm1done.surv_&k;
```

```
      by lost newcensordt;
            if newcensordt >= 5 and survival ^=. then output;
run;


proc sort data=spm1done.Simulated_Survival5yr_&k;
      by Replicate;
run;


data spm1.Simulated_Survival5yr_Lost
spm1.Simulated_Survival5yr_NonLost;
      set spm1done.Simulated_Survival5yr_&k;
      by Replicate;

      if lost=1 then output spm1.Simulated_Survival5yr_Lost;
      else if lost = 0 then output spm1.Simulated_Survival5yr_NonLost;
run;

data spm1.Simulated_Survival5yr_Lost;
   set spm1.Simulated_Survival5yr_Lost;
   by Replicate;

   if first.Replicate;

   keep lost Replicate newcensordt survival;
run;

data spm1.Simulated_Survival5yr_NonLost;
   set spm1.Simulated_Survival5yr_NonLost;
   by Replicate;

   if first.Replicate;

   keep lost Replicate newcensordt survival;
run;


proc sort data = spm1.Simulated_Survival5yr_Lost out =
spm1.SurvLostCI_&k;
      by SURVIVAL;
run;

data spm1.SurvLostCI_&k;
      set spm1.SurvLostCI_&k;
      if _N_ = 25 or _N_ = 975 or _N_=500;
      keep SURVIVAL;
run;




proc sgplot data = spm1.Simulated_Survival5yr_Lost;
      histogram SURVIVAL;
      xaxis values = (.7, .75, .8,  .85 , .9, .95, 1);
      title "Histogram of Estimated 5 Year Disease Survival Rates for
Simulated Lost Patients";
```

```
run;



ods graphics off;
proc phreg data = spm1done.simulated_all_&k noprint
outest=spm1done.Hrtable_&k;
      by Replicate;
      title "Cox Regression for Risk of Death and Breast Cancer
Recurrence";
      class lost(ref="0")
              age_grp(ref="Post-Meno")
              race(ref="White")
              comorbid(ref="Low")
              stage(ref="0")
              grade_grp(ref="Low")
              HR_grp(ref="Negative")
              Her2neu_grp(ref="Negative")
              LVI_grp(ref="No")
              insur (ref="Managed";
      model newcensordt*newevt(0) = lost age_grp race comorbid stage
LVI_grp insur ;
      strata grade_grp HR_grp Her2neu_grp;


      hazardratio lost / CL=WALD DIFF=REF;
        /*  Displays model coefficients, tests of significance,
              and exponentiated coefficient as hazard ratio*/
run;
ods graphics on;


data spm1done.HRtable_final_&k;
      set spm1done.HRtable_&k;
             HRLost = exp(lost1);
run;


proc sort data = spm1done.Hrtable_final_&k out = spm1CI2.HrTableCI_&k;
      by HRlost;
run;

data spm1CI2.HRtableCI_&k;
      set spm1CI2.HrtableCI_&k;
      if _N_ = 25 or _N_ = 975;
      keep HRlost;
run;

proc sgplot data = spm1done.HRtable_final_&k;
      histogram HRLost;
      xaxis values = (.2, .4, .6,  .8 , 1, 1.2, 1.4, 1.6);
      title "Histogram of Estimated Hazard Ratios for Nonlost vs.
Simulated Lost Patients";
run;
```

SAS Code Bootstrapping Method 1 and DFS Analysis

```
/******************** Bootstrapping Method 2*************************/

%let sampsize=1800;     /*similar to original sample size*/
%let numreps=1000;
%let k = 2;


/* sampling with replacement from the non-lost population to make a
sub-sample */
/*use reps= to specify the number of sample replications*/
/*the variable, Replicate, indicates the sample replicate number during
the surveyselect (is already sorted in descending order)*/
proc surveyselect data=sp.nonlost_population
                          method=URS
                          seed=1
                          sampsize= &sampsize
                          reps = &numreps
                          outhits
                          out = spm2.SubNonLostPop;

      title "Sampling With Replacement From Non-lost Patients";
run;

/*Need to find the number per bucket from our new sub sample*/
/*Finding percentage of lost for each bucket*/
proc sort data = sp.bucket out = spm2.sortedBucket;
      by lost;
run;

/*Finding percentage of lost patients in each bucket based on the
actual population*/
proc freq data= spm2.sortedBucket noprint;
      tables lost*bucket / out = spm2.BucketCount ;
run;


data spm2.BucketCount2;
      set spm2.BucketCount;
      if lost=1;

/*Determining how many observations should go in each bucket based on
the actual proportion of lost patients in each bucket*/
      BucketCounter = round(&numreps *(&sampsize * (PERCENT*.01)));
run;


/*Finding the percentage lost in our total population*/
proc means data = spm2.BucketCount2 sum noprint;
 var PERCENT;
 output out = FreqLost SUM = PercLost;
run;
```

```
proc sort data = spm2.SubNonLostPop;
 by bucket;
run;

proc sort data = spm2.BucketCount2;
  by bucket;
run;


/*Merging datasets inorder to get the amount of people that should be
lost per bucket for the _NSIZE_ variable*/
data spm2.NonLost;
   merge spm2.BucketCount2 spm2.SubNonLostPop (in=a);
        by bucket;
        if a;
        if lost = 0;
       _NSIZE_=BucketCounter;
run;

proc sort data = spm2.NonLost;
  by Bucket;
run;

proc surveyselect data = spm2.NonLost
                         method= SRS
/*selects units with equal probability and without replacement*/
                         seed= 1
                         sampsize=  spm2.NonLost
                         out = spm2.select;

            strata Bucket;

  title "Sampling Without Replacement From Non-lost Sub-Population";
run;


data spm2.sample;   /* Simulated lost patients */
      set spm2.select;
      lost=1;

      /*Creating a new ID variable that will be unique for each
patient. For lost patients it will be the original PID with L (for
lost) concatenated onto it eg: ##L*/
      Newid = cats(PID, "_", "L");

run;

/*Need to sort sample by new ID to set up the dataset for the counter
variable in the next block of code*/
proc sort data = spm2.sample;
      by pid newid descending assessid;
run;


data spm2.ID;
set spm2.sample;
```

```
      by pid;

/*Recoding new id to make them unique for each observation
These new ids will have a counter concatenated to the new id eg: 50L1,
50L2, 50L3 */
      if first.pid then do;
            counter = 1;
            newid_unique=cats(newid,"_",counter);
      end;

      else do;
            counter+1;
            newid_unique = cats(newID,"_",counter);
       end;
run;

/*For getting the assessment periods*/
data spm2.cont_status;
      set raw.Continuous_status;
run;
proc sort data = spm2.cont_status;
      by pid descending assessid;
run;


/*Setting data up so we can lose assesment periods*/
proc transpose data=spm2.cont_status out=spm2.cs;
  by pid;
  var assessid;
run;

data spm2prep.GetSimLastAssess_&k;
      set spm2.cs;

      Simlastassess=largest(&k, OF COL1-COL17);
/*k=2 to go back 1 assessment period, change k for going back more
periods*/

      drop COL1-COL17;
run;

data spm2prep.GetSimLastAssess2_&k;
      merge spm2prep.GetSimLastAssess_&k spm2.cont_status;
      by pid;
      if SimLastAssess = assessid;

      lostdt = datacollectdtdm;
      vitalstatdt2 = vitalstatdt;
      vitalstat2 = vitalstat;
      nccncarestatdt2 = nccncarestatdt;
      dzstatdt2 = dzstatdt;

      if vitalstatdt2 > lostdt then delete;
/*this should not happen, would be caused by error of inputting
information*/
/*one person (pid 256622) in this dataset has a lostdt < vitalstatdt2*/
```

```
      keep pid lostdt vitalstatdt2 vitalstat2 nccncarestatdt2
dzstatdt2;
      format lostdt DATE9. vitalstatdt2 DATE9. nccncarestatdt2 DATE9.
dzstatdt2 DATE9.;
run;

data spm2prep.id2_&k;
  merge spm2prep.GetSimLastAssess2_&k spm2.id (in=a);
  by pid;
  if a;
run;

/*For finding the sitedt -- First date of diagnosis of site of
disease.*/
data spm2prep.met_site;
      set raw.metastatic_sites;
      keep pid sitedt;
run;
proc sort data = spm2prep.met_site;
      by pid;
run;

/*preparing data for the merge*/
proc sort data = spm2prep.id2_&k;
      by pid;
run;


/*Creating a new dataset that has the lost data and the date of the met
site*/
data spm2prep.id3_&k;
  merge spm2prep.met_site spm2prep.id2_&k (in=b);
  by pid;
  if b;
run;

proc sort data=spm2prep.id3_&k;
 by newid_Unique;
run;


data spm2prep.DFS_&k;
      set spm2prep.id3_&k;
      by newid_unique;


if lostdt >= dxdt then do;
   if deathdt = . then do;
        if lostdt >= sitedt and sitedt ^= . then do;
/*lostdt is the date of the most recent information we have on them and
sitedt first date of diagnosis of site of disease.*/
            newmet = 1; /*recurrance has happened*/
             newcendt = sitedt;
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
/*dxdt is the date of diagnosis*/
        end;
```

```
        else if lostdt < sitedt and sitedt ^= . and vitalstat2 = 1 then
do;
            newmet = 0;
            newcendt = max(vitalstatdt2, NDidt);
            newcensordt = yrdif(dxdt, newcendt, 'act/act');

       end;
       else if lostdt < sitedt and sitedt ^= . and vitalstat2 = . then
do;
            newmet = 0;
            newcendt = max(dzstatdt2, NDidt);
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
       end;
       else if sitedt = . and vitalstat2 = 1 then do;
            newmet = 0;
            newcendt = max(vitalstatdt2, NDidt);
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
       end;
       else if sitedt = . and vitalstat2 = . then do;
            newmet = 0;
            newcendt = max(dzstatdt2, NDidt);
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
       end;
   end;



 if deathdt ^= . and sitedt ^= .  then do;
  event = 1;
      if lostdt >= sitedt then do;
            newcendt = sitedt;
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
      end;

      else if lostdt < sitedt then do;
          newcendt = deathdt;
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
      end;
end;

  else if deathdt ^= . and sitedt = . then do;
            event =1;
            newcendt = deathdt;
            newcensordt = yrdif(dxdt, newcendt, 'act/act');
  end;
end;



else if lostdt < dxdt then delete;


   if newmet=1 | event=1 then newevt=1;
   else newevt=0;  /* newevt = new event variable for death/recurrence
*/

       keep newid_unique pid  lost newevt newcensordt replicate bucket
```

```
                  /*Variables important to the Cox Model*/
               age_grp race comorbid stage grade_grp HR_grp Her2neu_grp
LVI_grp insur employ years
                  vitalstatdt lostdt censordt datacollectdtdm

                  newcendt deathdt osdt sitedt;
run;


/*Creating 1000 identical datasets to sp.nonlost_population that will
be all in the same dataset(nonlost_population_reps) and creating a new
variable, Replicate, that keeps track of the replication number for
each patient and will be used with the variable Replicate from the proc
survyselect*/

data spm2prep.nonlost_population_reps_&k;
      set spm2.SubNonLostPop;
      /* make newevt and newcensordt here, same as original data*/
                  newevt = evt;
                  newcensordt = censordt;
run;


/*Creating a dataset will all patients (both nonlost and simulated
lost)*/
data spm2prep.simulated_all_&k;
      set spm2prep.nonlost_population_reps_&k spm2prep.DFS_&k;
run;

proc sort data = spm2prep.simulated_all_&k;
      by Replicate pid;
run;


/*Finding Kaplan-Meier Estimates of disease free survival on both
nonlost and simulated lost patients*/
ods graphics off;
proc lifetest data = spm2prep.simulated_all_&k noprint outsurv=
spm2done.Surv_&k;
      by Replicate;
      time newcensordt*newevt(0);
      strata lost / test=logrank;
      label newcensordt = "Years";
      title "KM Estimates of DFS by Lost Status: SIMULATED Lost
Patients";
run;
ods graphics on;

proc sort data = spm2done.surv_&k;
      by lost newcensordt;
run;

data spm2done.Simulated_Survival_&k;
      set spm2done.surv_&k;
      by lost newcensordt;
            if newcensordt >= 5 and survival ^=. then output;
run;
```

```
proc sort data=spm2done.simulated_survival_&k;
      by Replicate;
run;


data spm2done.Simulated_Survival5yr_Lost_&k
spm2done.Simulated_Survival5yr_NonLost_&k;
      set spm2done.simulated_survival_&k;
      by Replicate;

      if lost=1 then output spm2done.Simulated_Survival5yr_Lost_&k;
      else if lost = 0 then output
spm2done.Simulated_Survival5yr_NonLost_&k;
run;

data spm2done.Simulated_Survival5yr_Lost_&k;
   set spm2done.Simulated_Survival5yr_Lost_&k;
   by Replicate;

   if first.Replicate;

   keep lost Replicate newcensordt survival;
run;

data spm2done.Simulated_Survival5yr_NonLost_&k;
   set spm2done.Simulated_Survival5yr_NonLost_&k;
   by Replicate;

   if first.Replicate;

   keep lost Replicate newcensordt survival;
run;


proc sort data = spm2done.Simulated_Survival5yr_NonLost_&k out =
spm2CI.SurvNonLostCI_&k;
      by SURVIVAL;
run;

data spm2CI.SurvNonLostCI_&k;
      set spm2CI.SurvNonLostCI &k;
      if _N_ = 25 or _N_ = 975;
      keep SURVIVAL;
run;

proc sgplot data = spm2done.Simulated_Survival5yr_NonLost_&k;
      histogram SURVIVAL;
      xaxis values = (.7, .75, .8,  .85 , .9, .95, 1);
      title "Histogram of Estimated 5 Year Disease Survival Rates for
Sampled Non-Lost Patients";
run;
```

```
proc sort data = spm2done.Simulated_Survival5yr_Lost_&k out =
spm2CI.SurvLostCI_&k;
      by SURVIVAL;
run;

data spm2CI.SurvLostCI_&k;
      set spm2CI.SurvLostCI_&k;
      if _N_ = 25 or _N_ = 975;
      keep SURVIVAL;
run;

proc sgplot data = spm2done.Simulated_Survival5yr_Lost_&k;
      histogram SURVIVAL;
      xaxis values = (.7, .75, .8,  .85 , .9, .95, 1);
      title "Histogram of Estimated 5 Year Disease Survival Rates for
Simulated Lost Patients";
run;



ods graphics off;
proc phreg data = spm2prep.simulated_all_&k noprint
outest=spm2.Hrtable_&k;
      by Replicate;
      title "Cox Regression for Risk of Death and Breast Cancer
Recurrence";
      class lost(ref="0")
              age_grp(ref="Post-Meno")
              race(ref="White")
              comorbid(ref="Low")
              stage(ref="0")
              grade_grp(ref="Low")
              HR_grp(ref="Negative")
              Her2neu_grp(ref="Negative")
              LVI_grp(ref="No")
              insur (ref="Managed");
      model newcensordt*newevt(0) = lost age_grp race comorbid stage
LVI_grp insur;
      strata grade_grp HR_grp Her2neu_grp;

      hazardratio lost / CL=WALD DIFF=REF;
        /*  Displays model coefficients, tests of significance,
              and exponentiated coefficient as hazard ratio*/
run;
ods graphics on;


data spm2done.HRtable_final_&k;
      set spm2.HRtable_&k;
             HRLost = exp(lost1);
run;


proc sort data = spm2done.Hrtable_final_&k out = spm2CI.HrTableCI_&k;
      by HRlost;
run;
```

```
data spm2CI.HRtableCI_&k;
      set spm2CI.HrtableCI_&k;
      if _N_ = 25 or _N_ = 975;
      keep HRlost;
run;


proc sgplot data = spm2done.HRtable_final_&k;
      histogram HRLost;
      xaxis values = (.2, .4, .6,  .8 , 1, 1.2, 1.4);
      title "Histogram of Estimated Hazard Ratios for Nonlost vs.
Simulated Lost Patients";
run;
```