Estimating the Impact of Lost to Follow-up on Breast Cancer Patients' 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
Debbie Yan Qun Huang
May 2013

© 2013 Debbie Yan Qun Huang

Abstract

Background The 5-year survival rate for patients with breast cancer is much higher than patients with other types of cancer. Due to this longer survival period, breast cancer patients also tend to have increased rates of lost to follow-up, when 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 (BC) death. The impact of lost patients on recurrence rates and disease-free survival (DFS) was explored in breast cancer patients seen at the City of Hope from 1997 to 2012.

Methods Female breast cancer patients with a stage of 0, I, II, or III at diagnosis were included in these 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 hazards models were built to adjust for multiple predictors such as age group at diagnosis, race, comorbidity score, final cancer stage at diagnosis, and lymphovascular invasion (LVI) status. BC death rates were compared between non-lost and lost patients using a chi-square test. Missed recurrences were estimated and recurrence rates were calculated. Patients were separated into 20 groups based on propensity scores from a logistic regression model using categorical distance between the patient's residence and the City of Hope, age group at diagnosis, final cancer stage at diagnosis, hormone receptor status, and her2/neu status to predict the probability of becoming lost. Lost patients were removed and replaced with simulated lost patients. Simulated lost patients were sampled with replacement from the non-lost patients within each group and then one year of information was removed from those patients. The new 5-year DFS rate was calculated. This process of simulating lost patients and recalculating the 5-year DFS was bootstrapped 1,000 times.

Results The 5-year DFS rate was 84.6% for non-lost patients and 95.1% for lost patients. Adjusting for age, race, comorbidity score, stage, and LVI, the risk of death or recurrence is 61.0% lower for lost patients compared to non-lost patients. The BC death rate was 8.2% for non-lost patients and 2.4% for lost patients. This difference in BC death rates may be due to delays in death information for lost patients. There were 66 observed missed recurrence and 42 estimated unobserved missed recurrences. The observed recurrence rate was 7.1% and the estimated recurrence rate was 11.7%. The mean 5-year DFS rate for simulated lost patients was 86.2%.

Conclusion There are missing recurrences for both non-lost and lost patients, yielding a lower observed recurrence rate than estimated and inflated DFS rates. Lost patients lead to even more missing recurrence information, yielding larger differences in the observed rates and estimated rates. Researchers could mention the lost to follow-up rate and the possible effects on DFS to avoid misleading rates.

Table of Contents

List of Figures

List of Equations

List of Tables

Introduction

The 5-year survival rate for patients with breast cancer is much higher than patients with other types of cancer. Due to this longer survival period, breast cancer patients also tend to have increased rates of lost to follow-up, when 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. The impact of lost patients on recurrence rates and disease-free survival was explored in breast cancer patients seen at the City of Hope from 1997 to 2012.

Methods

Description of Cohort

Female breast cancer patients seen at the City of Hope from 1997 to 2012, with a stage of 0, I, II, or III at diagnosis were included in these analyses (N=2,358). Stage IV breast cancer patients were excluded due to metastases at diagnosis. This means that the cancer of stage IV patients had spread beyond the breasts and local lymph nodes; therefore there was no recurrence to be measured. Of the 2,358 breast cancer patients, 1,937 were deemed 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 metastases 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 to doctors 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 was categorized as 0, I, II, or III for the final cancer stage at diagnosis. 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, "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

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

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, race, comorbidity score, stage, and LVI status. Controls 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.

Breast cancer-specific death rates were compared between non-lost and lost patients using a chi-square test. The number of breast cancer deaths was also compared to distant recurrence to ascertain the completeness of recurrence information. A “missed” distant recurrence is considered when a patient has a breast cancer-specific death, but there is no distant recurrence information in the patient chart.

The number of missed distant recurrences was estimated; then the observed distant recurrence rate was compared to the estimated distant recurrence rate. The number of missed recurrences for patients that experience a breast cancer-specific death can be computed (B from Table 1), but this number is unknown for patients that did not experience a breast cancer-specific death. To estimate the number of missed recurrences for patients that without a breast cancer-specific death, an assumption that the ratio of missed recurrences is proportional to the ratio of observed recurrences was made. With this assumption, the number of missed recurrences for patients without a breast cancer-specific death was estimated, shown by Equation 3.

The observed recurrence rate is the total number of observed recurrences divided by the total number of patients, shown in Equation 4. The estimated recurrence rate is the sum of the total number of observed recurrences, the number of observed missed recurrences, and the number of estimated missed recurrences, divided by the total number of patients, show in Equation 5. The difference between these two recurrence rates is the estimated percentage of missed recurrences.

Table 1. Distant Recurrence by Breast Cancer-Specific Death (Without Data)

	No BC Death	BC Death	Total
No Distant Recurrence	A	B	A+B
Distant Recurrence	C	D	C+D
Total	A+C	B+D	N

Equation 3. Estimating Missed Recurrences for Non Breast Cancer-Specific Death Patients

$$Estimate = B \times \frac{C}{D}$$

Equation 4. Calculating Observed Recurrence Rate

$$Observed\ Recurrence\ Rate = \frac{C + D}{N}$$

Equation 5. Calculating Estimated Recurrence rate

$$Estimated\ Recurrence\ Rate = \frac{C + D + B + (B \times \frac{C}{D})}{N}$$

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 scores 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, 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 patients with complete information. These simulated patients were assigned a status of lost to see how this would impact the DFS rates. For actual lost patients, the median time from the date when a patient was first lost to the date of the last assessment is 12.025 months, which is approximately one assessment period. Therefore to simulate lost patients from non-lost patients, one assessment period of known information was deleted from the last assessment period. By deleting one assessment period, there was a new censoring date for these simulated lost patients and any recurrence that occurred in that last assessment period became unknown.

In the new cohort, 1,937 were non-lost and 421 were simulated lost. Kaplan-Meier estimates were recalculated for the 5-year disease-free survival probabilities. The 5-year disease-free survival probability for non-lost patients remained the same, as no changes were made to non-lost patients; however the 5-year disease-free survival probability for simulated lost patients was not expected to be the same. To stabilize results, this process of resampling from the non-lost cohort, simulating lost patients, and recalculating the Kaplan-Meier estimate for 5-year disease-free survival probability for simulated lost patients was bootstrapped 1,000 times. Each iteration of this process used the seed corresponding to that iteration for sampling (e.g. seed=1 for 1st iteration, seed=2 for 2nd iteration, etc.). The mean 5-year disease-free survival probability for these 1,000 replications was calculated for the simulated lost patients and compared to the original 5-year disease-free survival probability for 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.

Results

Disease-free Survival

Initially the entire cohort of 2, 358 patients was analyzed for disease-free survival stratified by lost status. Figure 1 indicates that non-lost patients tend to have significantly lower disease-free survival rates than lost patients (Log-rank P<0.01). The disease-free survival rates from Table 2 also concurred with the notion that non-lost patients have lower disease-free survival rates (5-year survival 84.6% for non-lost vs. 95.1% for lost). The log-rank chi-square test statistic of 24.75 with 1 degree of freedom yields a p-value <0.01, which provides evidence that the survival functions for non-lost and lost patients are significantly different.

Figure 1. Kaplan-Meier Curves for Disease-free Survival

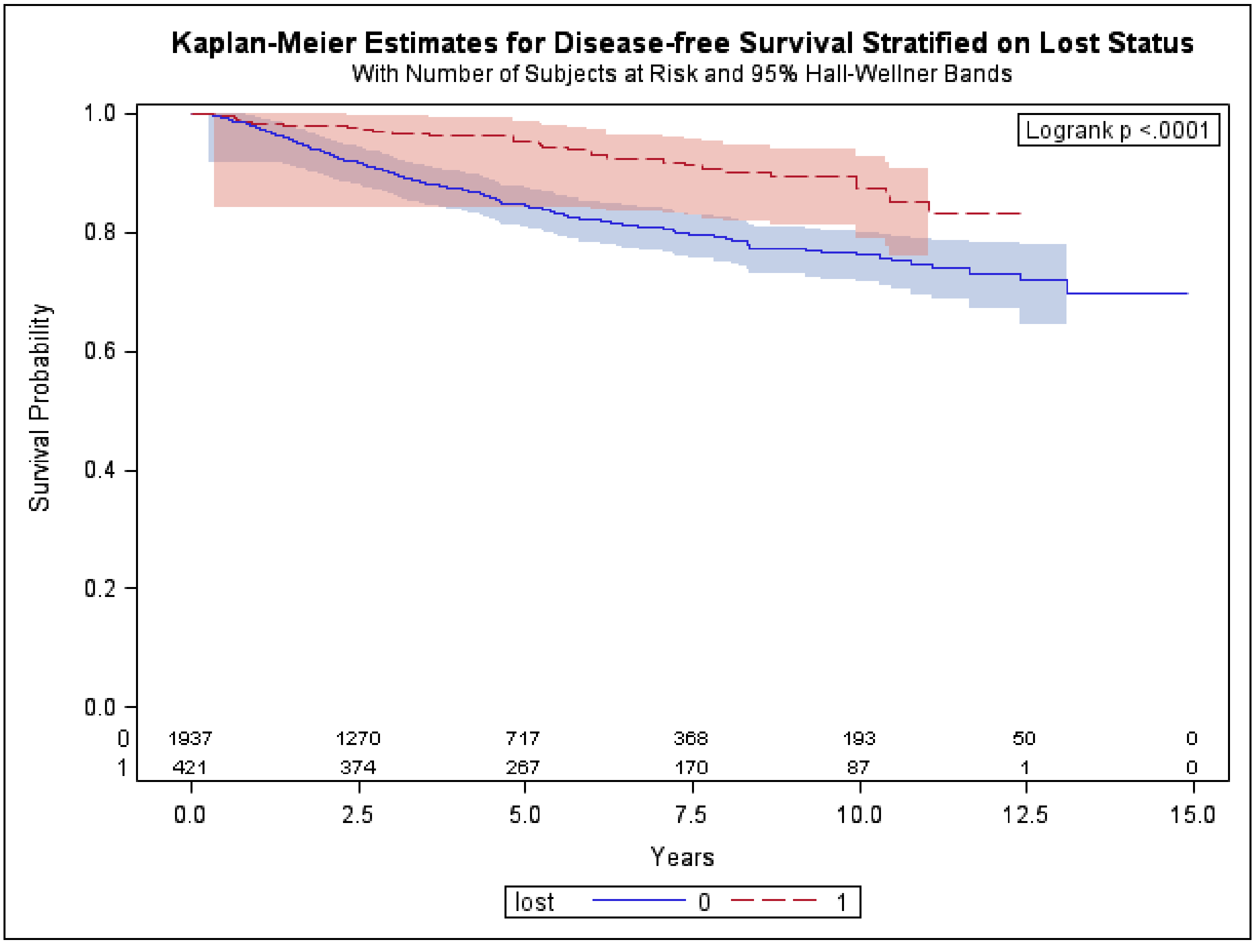

Table 2. Kaplan-Meier Estimates for Disease-free Survival

	Non-lost	Lost
1-Year	97.6%	98.1%
3-Year	90.0%	96.8%
5-Year	**84.6%**	**95.1%**
10-Year	75.7%	86.3%
Log-Rank Chi-Square=24.75	DF=1	P-value<0.01

Table 2 shows the unadjusted 1, 3, 5, and 10-year disease-free survival rates by lost status. To adjust for covariates in the presence of lost status, Cox proportional hazards regression was used. Accounting for age at diagnosis, race, comorbidity score, stage, and LVI, the risk of death or distant recurrence is 61.0% lower for lost patients than non-lost patients (HR=0.39, P<0.01). Although race was an insignificant predictor, it remained in the model as a control, due to breast cancer incidence rates being different between races. From the Kaplan-Meier estimates (Table 2) and hazard ratio for disease-free survival (Table 3), it appears that lost patients have a significantly higher disease-free survival rate than non-lost patients, but this may likely be a result of missing recurrences. Distant recurrences are difficult to ascertain once a patient becomes lost, so the disease-free survival rate may be inflated, because these missing recurrences are not accounted for.

Table 3. Adjusted Hazard Ratios for Disease-free Survival

Variable (Reference)	**Level**	**Adjusted Hazard Ratio**	**95% CI**	**Overall P-value**
Lost Status (Non-lost)	**Lost**	**0.39**	**(0.27, 0.57)**	<0.01
Age (Post-Menopausal)	**Elderly**	**2.73**	**(1.99, 3.74)**	<0.01
	Pre-Menopausal	1.29	(0.98, 1.69)	
Race (White)	Asian	0.83	(0.57, 1.20)	0.80
	Black	1.13	(0.68, 1.90)	
	Hispanic	1.03	(0.78, 1.36)	
	Other	0.79	(0.25, 2.51)	
Comorbidity Score (Low)	**High**	**3.43**	**(2.10, 5.62)**	<0.01
	Medium	**1.38**	**(1.03, 1.86)**	
Stage (0)	**I**	**8.64**	**(2.79, 26.73)**	<0.01
	II	**12.90**	**(4.10, 40.65)**	
	III	**31.33**	**(9.87, 99.51)**	
*LVI Status (No)	**Yes**	**1.52**	**(1.16, 2.01)**	<0.01

* 71 patients with unknown LVI status
NOTE: Model is stratified on tumor grade, hormone receptor status, and her2/neu status.

Breast Cancer-Specific Death

To account for missing recurrences, breast cancer-specific death rates were compared by lost status. Table 4 shows that, overall 7.1% of the cohort experienced a breast cancer-specific death and 92.9% did not. A chi-square test statistic of 17.47 with 1 degree of freedom yields a p-value <0.01, meaning that there appears to be a significant association between breast cancer-specific death and lost status. From Table 4, 8.2% of non-lost patients and 2.4% of lost patients experienced a breast cancer-specific death. The non-lost patients have a 5.8% higher breast cancer-specific death rate than lost patients. This may be due to the delay in death information for lost patients to the City of Hope. Death information for lost patients who died could be obtained through sources such as the National Death Index, but due to limited resources, it is not possible to search for the lost patients daily.

Table 4. Breast Cancer-Specific Death by Lost Status

Frequency *Column Percent*	Non-Lost	Lost	Total
No BC Death	1779 91.8%	411 97.6%	2190 92.9%
BC Death	158 8.2%	10 2.4%	168 7.1%
Total	1937	421	2358
Chi-Square=17.47	DF=1	P-value<0.01	

Missed Distant Recurrences and Recurrence Rates

Patients with a breast cancer-specific death must have had a distant metastatic site, so theoretically the number of patients that had a breast cancer-specific death, without distant recurrence should be 0. From Table 5, the number of patients that had a breast cancer-specific death, but did not have a distant recurrence is 66. This means that there are 66 missed recurrences in the cohort for patients who experienced a breast cancer-specific death. Of these 66 observed missed recurrences, 60 occurred for non-lost patients and 6 for lost patients.

Table 5. Distant Recurrence by Breast Cancer-Specific Death

	No BC Death	BC Death	Total
No Distant Recurrence	2125	**66**	2191
Distant Recurrence	65	102	167
Total	2190	168	2358

Equation 6 shows the calculation for the estimated number of missed recurrences for patients that did not experience a breast cancer-specific death, which was approximately 42 patients. Equation 7 shows the calculation for the observed recurrence rate, which in this cohort was 7.1%. Equation 8 shows the calculation for the estimated recurrence rate, which was 11.7%. From these calculations, there was an estimated 108 missed recurrences, which is approximately 4.6% of the entire cohort.

Equation 6. Estimated Number of Missed Recurrences for Non Breast Cancer-Specific Death Patients

$$Estimate = 66 \times \frac{65}{102} = 42.06$$

Equation 7. Observed Recurrence Rate

$$Observed\ Recurrence\ Rate = \frac{167}{2358} = 0.0708 = 7.1\%$$

Equation 8. Estimated Recurrence rate

$$Estimated\ Recurrence\ Rate = \frac{167 + 66 + 42.06}{2358} = 0.1166 = 11.7\%$$

Disease-free Survival for Simulated Lost Patients

A logistic regression model was built using distance, age, stage, hormone receptor status, and her2/neu status to predict the probability of being lost. As seen in Table 6, the predictors previously listed were all significantly associated with a patient's lost status. There was a noticeable pattern for distance; patients residing further away from the City of Hope have higher odds of becoming lost, with the exception of foreign patients, but this may be due to the low number of foreign patients (n=4). There was a noticeable pattern for stage as well; patients with a higher stage of breast cancer have lower odds of becoming lost.

Table 6. Associations for Becoming Lost Using Logistic Regression

Variable (Reference)	**Level**	**Adjusted Odds Ratio**	**95% CI**	**Overall P-value**
Distance (Close)	Medium	1.08	(0.72, 1.64)	<0.01
	Far	**2.60**	**(1.74, 3.90)**	
	Foreign	1.26	(0.12, 13.04)	
Age (Post-Menopausal)	**Elderly**	**1.53**	**(1.12, 2.09)**	0.03
	Pre-Menopausal	1.16	(0.91, 1.48)	
Stage (0)	**I**	**2.82**	**(1.81, 4.39)**	<0.01
	II	**2.65**	**(1.66, 4.22)**	
	III	**1.79**	**(1.04, 3.09)**	
*Hormone Receptor Status (Negative)	**Positive**	**1.36**	**(1.01, 1.83)**	<0.01
**Her2/neu Status (Negative)	High+	1.08	(0.74, 1.58)	<0.01
	Low+	1.12	(0.59, 2.13)	
	Positive NOS	**5.09**	**(1.88, 13.76)**	

* 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 groups, such that patients in the same group have similar propensity scores. Within each group, the frequencies of lost patients were saved and used for sampling. Table 7 shows the distribution of lost and non-lost patients within each group.

Table 7. Distribution of Lost and Non-lost Within the 20 Propensity Score Groups

Group	Lost	Non-lost	Total
1	13	105	118
2	9	109	118
3	11	107	118
4	16	102	118
5	15	103	118
6	16	102	118
7	19	99	118
8	8	110	118
9	24	94	118
10	15	103	118
11	9	109	118
12	13	105	118
13	14	104	118
14	18	100	118
15	20	98	118
16	27	91	118
17	42	76	118
18	40	78	118
19	41	77	118
20	51	65	116
Total	421	1937	2358

The lost patients were then removed, leaving only non-lost patients in the sampling cohort. Within each group, non-lost patients were sampled with replacement to simulate lost patients. By removing one assessment period (roughly one year) from those patients, they become simulated lost patients and the distant recurrences that took place in that year became unknown. Kaplan-Meier estimates were recalculated for 5-year disease-free survival based on the new cohort. The process of resampling to simulate lost patients and estimating the 5-year disease-free survival rate was bootstrapped to stabilize results. From the 1,000 repetitions, the average 5-year disease-free survival rate for simulated lost patients was 86.2%. Recall that the 5-year survival rate for the actual lost patients was 95.1%. This showed an 8.9% difference between the actual and simulated 5-year disease-free survival rates. Therefore the disease-free survival rates for lost patients appear to be inflated due to missing recurrence information.

Conclusion

The initial Kaplan-Meier estimates for disease-free survival and adjusted hazard ratios suggest that lost patients have higher disease-free survival rates than non-lost patients, which may actually be due to the fact that they are lost, and their recurrence information is missing. From the known data, observed missed recurrences were calculated and used to estimate the unobserved missed recurrences. It seems reasonable to assume that due to the missed recurrences for both non-lost and lost patients, disease-free survival rates would be misleading. However, the number of missing recurrences for lost patients appears to be much higher, leading to even more distorted recurrence rates than in the non-lost population. The impact of losing patients was modeled and the 5-year disease-free survival rates also concurred with the notion that the disease-free survival rates for lost patients are inflated.

The missing recurrences due to patients becoming lost to follow-up must be accounted for to accurately estimate disease-free survival. One solution would be for hospitals to track down lost patients and collect the appropriate information, but this is not very plausible due to limited resources. A more realistic solution would be for researchers to mention this caveat when presenting disease-free survival analyses. The researchers could provide descriptive statistics about the lost to follow-up rate in the cohort and mention the possible effects of lost to follow-up on disease-free survival.

Limitations and Future Studies

There were limitations in this study due to time constraints. The proportional hazard assumption was not met for lost status, tumor grade, hormone receptor status, and her2/neu status for the Cox proportional hazards model. The control variables tumor grade, hormone receptor status and her2/neu status were accounted for by stratifying on those variables in the model. However, lost status was the main variable of interest. In order to obtain a hazard ratio for comparison, lost status could not be included as a strata variable. One method for accounting for lost status being time dependent would be to include it as an interaction term with time, or more commonly the natural log of time. Another solution would be to include lost as a time dependent covariate in the model.

There were a few more analyses that could also be explored in the future. The propensity score analysis computed the 5-year disease-free survival probabilities after resampling. In order to account for other factors, a Cox proportional hazards model could also be bootstrapped in a similar way, yielding hazard ratios for exploration.

The definition of lost in this study was a patient not going to a follow-up visit for two or more years, so originally two assessment periods were supposed to be deleted. In this analysis, only one assessment period was deleted, due to the median time between the date when a patient was first lost and the date of their last assessment period being approximately one assessment period. Going forward the removal of two assessment periods could also be explored.

References

"About BMI for Adults." *Centers for Disease Control and Prevention*. CDC, 13 Sep. 2011. Web. 26 Feb. 2013. <http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html>.

"Breast Cancer Rates by Race and Ethnicity." *Centers for Disease Control and Prevention*. CDC, 19 Dec. 2012. Web. 29 Jan. 2013. <http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html>.

"Pathology Report." *Dr. Susan Love Research Foundation.* Dr. Susan Love Research Foundation. Web. 28 May 2013. <http://www.dslrf.org/breastcancer/content.asp?CATID=28&L2=1&L3=6&PID=&sid=132&cid=1104>

SAS Maps Online. (2013). Zipcode (V8 and V9) [Data file]. Retrieved from http://support.sas.com/rnd/datavisualization/mapsonline/html/misc.html

"Tumor Grade." *National Cancer Institute*. NIH, 03 May 2013. Web. 10 May 2013. <http://www.cancer.gov/cancertopics/factsheet/detection/tumor-grade>.

"Usage Note 5325: Calculating the distance between ZIP codes." *SAS*. SAS, 04 June 2009. Web. 05 Feb. 2013. <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 "D:\Senior Project\Data\Raw";
libname derived "D:\Senior Project\Data\Derived";
libname zip "D:\Senior Project\Zip";
libname sassy "D:\Senior Project\Sassy";
libname newfmt "D:\Senior Project\Data\New Formats";


*** Laptop SAS9.2 is 32bit;
Options fmtsearch=(raw.formats derived.formats);

*** Campus CPU SAS9.3 is 64bit;
*Options fmtsearch=(newfmt.formats);




data bca1;

      merge raw.diagnosis derived.clinical_characteristics(IN=clinical);
      by pid dxid;

      if clinical=1;

run;




data bca2;

      merge bca1(IN=clinical) derived.patient_characteristics;
      by pid dxid;

      if clinical=1;

            *** only use patients in clinical_characteristics data set;
            **** recode LTFU, missing is non-lost and 1&2 are lost;

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



            *** years between diagnosis and event/censor date;

      years = (osdt-dxdt)/365.25;




      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";



run;



data bca3;

      merge bca2(IN=clinical) raw.metastatic_sites;
      by pid dxid;


      if clinical=1;

      *** only use patients in clinical_characteristics data set;

      if first.pid=1;
```

```
/* 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=20 | site=14 then
            met=1;
             else if site=16 & sitedt >= '01jan2003'd then met=1;
             else if site=19 & 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=(cendt-dxdt)/365.25;

run






data bca4;

      merge raw.demographics bca3(IN=clinical);
      by pid;


      if clinical=1;


      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;





run;






data bca5;

      merge bca4(IN=clinical) raw.study_accession;
      by pid;

      if clinical=1;

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



            ******* 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";



run;







data bca6;

      merge bca5(IN=clinical) raw.breast_diagnosis;
      by pid dxid tumorid;

      if clinical=1;


run;




data bca7;

      merge bca6(IN=clinical) raw.solid_tumor_stage;
      by pid dxid tumorid;

      if clinical=1;

      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";


run;




proc sort data=bca7;
      by zipcode;
run;





******** zip codes from SAS Maps Online to create distance variable ********;

            ***** Imports and outputs zipcode SAS datasets ****;

proc cimport infile="C:\Users\Debbie\Documents\Senior
      Project\Zip\zipcode_Jan13_v9.cpt"
      lib=zip;
run;


            *** Put 2 zipcode data sets together;

proc append base=zip.Zip1 data=zip.Zip2;
run;

            ****** Final zip code data set = zip.zipcode *************;

proc sort data=zip.Zip1 out=zip.zipcode;
      by zip zip_class;
run;

            ******** 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 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;




data sassy.final;
      merge bca7(IN=clinical) 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;
```

```
******** transporting formats from 32bit SAS to 64bit SAS **************;


      *** In 32bit ***;

filename rawfmt 'D:\Senior Project\Data\New Formats\rawformats.cpt';  /*
transport file you are creating */
filename derfmt 'D:\Senior Project\Data\New Formats\derivedformats.cpt';

proc cport lib=raw file=rawfmt memtype=catalog;
   select formats;
run;

proc cport lib=derived file=derfmt memtype=catalog;
   select formats;
run;



      **** In 64bit ***;

filename rawfmt 'F:\Senior Project\Data\New Formats\rawformats.cpt';   /*
same as in Step 1 above */
filename derfmt 'F:\Senior Project\Data\New Formats\derivedformats.cpt';

proc cimport infile=rawfmt lib=newfmt;
run;

proc cimport infile=derfmt lib=newfmt;
run;
```

SAS Code for Analyses

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




libname raw "F:\Senior Project\Data\Raw";
libname derived "F:\Senior Project\Data\Derived";
libname zip "F:\Senior Project\Zip";
libname sassy "F:\Senior Project\Sassy";
libname newfmt "F:\Senior Project\Data\New Formats";



*** Laptop SAS9.2 is 32bit;
*Options fmtsearch=(raw.formats derived.formats);

*** Campus CPU SAS9.3 is 64bit;
Options fmtsearch=(newfmt.formats);

ods html close;
ods listing;


ods graphics on;
ods listing sge=on;




****************************************************************************;


      ********* Survival from death: Non-lost vs Lost ************;

            **** p-value<0.0001;

proc lifetest data=sassy.final plot=survival METHOD=KM;

      *** product-limit is the same as Kaplan-Meier;

      time years*event(0);
      strata lost / test=logrank;
      title "KM Survival Estimates for Death by Lost Status";
run;
```

```
      *** lost p-value<0.0001;


proc phreg data=sassy.final;
      class lost(ref="0");
      model years*event(0) = lost ;
      title "Cox Regression for Risk of Death";
run;




***************** stratify on GRADE, HR, Her2neu  ************************;

            ****** Lost p-value<0.0001;



proc phreg data=sassy.final;

      title "Cox Regression for Risk of Death";

      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");

      model years*event(0) = lost age_grp race comorbid stage LVI_grp;

      strata grade_grp HR_grp Her2neu_grp;

      assess ph / resample;


run;





********************************* DFS *************************************;



*** Survival from death and breast cancer recurrence: Non-lost vs Lost ****;



            *** p-value<0.0001;


proc lifetest data=sassy.final plot=survival (atrisk=0 to 15 by 2.5 CB=HW
      test nocensor) METHOD=KM;

            *** product-limit is the same as Kaplan-Meier;

      time censordt*evt(0);
      strata lost / test=logrank;
      label censordt="Years";
      title "KM Estimates for DFS by Lost Status";
run;        *** p-value<0.0001;

proc phreg data=sassy.final;
      class lost(ref="0");
      model censordt*evt(0) = lost;
      title "Cox Regression for Risk of Death and Breast Cancer Recurrence";
run;



********************** stratafy on GRADE, HR, Her2neu  ********************;

      ** lost p-value<0.0001;

proc phreg data=sassy.final;

      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");

      strata grade_grp HR_grp Her2neu_grp;

      model censordt*evt(0) = lost age_grp race comorbid stage LVI_grp;


      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;

      assess ph / resample;

run;



****************************************************************************;
```

```
************ Method 1 - Breast Specific Survival: Lost vs Non-lost *********;


proc freq data=sassy.final;
      table breastdeath*lost / chisq;
run;

      /* chisq=17.4705 , p-value<0.0001 */





********** Method 2 - Estimate Number of Missed Recurrences *************;


            /* 66 missed recurrences */

proc freq data=sassy.final;
      title "Two-way Table: Dist Met x Breast Death";
      tables met*breastdeath;
run;



proc freq data=sassy.final;
      title "66 Missed Recurrences - 6 lost & 60 non-lost";
      where met=0 and breastdeath=1;
      table lost / chisq;
run;



proc freq data=sassy.final;
      title "Death Status for 66 non-BCA Deaths with Dist Mets";
      where breastdeath=0 & met=1;
      table dead;
run;


proc freq data=sassy.final;
      title "Two-way Table: Lost by Dist Met";
      table met*lost / chisq;
run;
```

```
*************** Method 3 - Propensity Scores for LTFU on DFS **************;



proc logistic data=sassy.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 / lackfit;

            *** lack fit = Hosmer Lemeshow test;


      score out=sassy.scores;
run;





*********** Simulating Lost Patients FOR METHOD 3 *******************;



                  *** group by propensity scores;

proc sort data=sassy.scores;
      by P_1;
run;



data sassy.bucket;
      set sassy.scores;

      total = 2358;

      size = round(total/20);

      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=sassy.bucket;
      table lost*bucket / nopercent norow nocol out=dist;
run;



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





            *** non-lost cohort;


data sassy.nonlost;                       *** includes alllllll variables;
      merge sassy.dist sassy.bucket;
      by bucket;

      if lost=0;                          *** non-lost only n=1684;

      _NSIZE_=count;                *** _NSIZE_ needed for proc surverselect;

run;






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


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






data sassy.constat;
      merge sassy.nonlost(IN=non) raw.Continuous_Status;
      by pid;

      if non=1;               *** get assessid info for nonlost patients;

run;
```

```
*** to get all the assessid's for deleting follow-ups;


data sassy.total_constat;
      set raw.Continuous_status;
      pidnum=1;
run;


%macro repeats(mydset, mypidnum);

      data &mydset;
            set raw.Continuous_status;
            pidnum=&mypidnum;
      run;


      proc append base=sassy.total_constat data=&mydset; run;

%mend;


%repeats(temp_constat2, 2);
%repeats(temp_constat3, 3);
%repeats(temp_constat4, 4);
%repeats(temp_constat5, 5);
%repeats(temp_constat6, 6);
%repeats(temp_constat7, 7);
%repeats(temp_constat8, 8);
%repeats(temp_constat9, 9);
%repeats(temp_constat10, 10);



proc sort data=sassy.total_constat;
      by pid pidnum descending assessid;
run;





data sassy.nonlost_population;
      set sassy.constat;
      by pid descending assessid;

      if first.pid=1 & first.assessid=1;

            *** to get last assessment period;

      lastassess=assessid;
run;
```

```
proc sort data=sassy.nonlost_population;
      by bucket pid;
run;


/* sampling from each bucket the number of lost patients we removed to get
back original sample size but with only non-lost patients*/



proc surveyselect data=sassy.nonlost_population
      method=URS seed=1 sampsize=sassy.nonlost_population outhits out=sample;

      strata bucket;

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

run;


/* URS=unrestricted random sampling = equal prob with replacement

sampsize= give dataset with _NSIZE_ variable, which is the number
we want to sample for each bucket

outhits = outputs all obs, so replicates all get outputted            */







data sample;      *** Simulated lost patients;
      set sample;
      lost=1;
run;



proc append base=sassy.nonlost_population data=sample force;
run;
            *** now constat2 is data set with all non-lost patients (2358);


proc sort data=sassy.nonlost_population;
      by pid;
run;




data sassy.simulated;
      set sassy.nonlost_population;
      by pid;

                  *** to handle repeat pid for resampled patients;
      if pid=lag(pid) then pidnum+1;
            else if pid^=lag(pid) then pidnum=1;

run;





proc freq data=derived.patient_characteristics;
table race_eth;run;




data sassy.ID;
      merge sassy.total_constat sassy.simulated;
      by pid pidnum descending assessid;

      retain prevlost prevpid prevpidnum;



      if _N_=1 then do;
            prevlost=lost;
            prevpid=pid;
            prevpidnum=pidnum;
            newid=1;
      end;
                                                *** to fix "missing" lost;



      if pid=prevpid & pidnum=prevpidnum & lost^=. then prevlost=lost;

            else if pid=prevpid & pidnum=prevpidnum & lost=. then
                  lost=prevlost;

            else if pid=prevpid & pidnum^=prevpidnum & lost^=. then do;
                        prevlost=lost;
                        prevpid=pid;
                        prevpidnum=pidnum;
                        newid+1;
                  end;

            else if pid=prevpid & pidnum^=prevpidnum & lost=. then delete;

            else if pid^=prevpid then do;
                        if lost=. then delete;
                              else do;
                                    prevlost=lost;
                                    prevpid=pid;
                                    prevpidnum=pidnum;
                                    newid+1;
                              end;
                  end;


run;






proc sort data=sassy.ID;
      by newid assessid;
run;






proc transpose data=sassy.ID out=sassy.trans;
      by newid;
      var assessid;
run;








data sassy.DFS;
      merge sassy.ID sassy.trans;
      by newid;

      lastassess=largest(2, OF COL1-COL17);

*** k=2 to go back 1 assessment period, change k for going back more periods;



      if last.newid=1;
*** new date of first met/event/censor for simulated patients;

*** last assessment date is censordt, if 1st met beyond last assess date then
no met;


      newmet=met;
      newcensordt=censordt;


      if lost=1 then do;


            if vitalstatdt^=. then lostdt=(vitalstatdt-dxdt)/365.25;
                  *** simulated lost date;

                  else if vitalstatdt=. then lostdt=(quality_fudt-
                        dxdt)/365.25;
                  if lostdt<0 then lostdt=(quality_fudt-dxdt)/365.25;
            *** for pid 256570 error bc vitalstatdt is before dxdt;

            if met=1 & lostdt <= censordt then do;
                  newmet=0;
                  *** Met is unknown after lost date;

                  newcensordt=lostdt;
            end;

            if event=0 & lostdt <= osdt then newcensordt=lostdt;
            *** For no death, lost date is new censor date;
      end;



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


      keep newid pid pidnum lost newevt newcensordt;
run;

proc lifetest data=sassy.DFS noprint outsurv=surv;
      *** product-limit is the same as Kaplan-Meier;
      time newcensordt*newevt(0);
      strata lost / test=logrank;
      label newcensordt="Years";
      title "KM Estimates of DFS by Lost Status: SIMULATED Lost Patients";
run;




data sassy.SIMULATED_Survival;
      set surv;
      by lost newcensordt;

      if newcensordt >= 5 & Survival^=. then output;

run;
```

```
data sassy.SIMULATED_Surv5Yr;
      set sassy.SIMULATED_Survival;
      by lost;

      if lost=1;

      if first.lost=1;

      keep lost newcensordt survival;
run;






*************** start macro ;


%macro DFS(n);


      %do i=2 %to &n;


      /* to reset sassy.nonlost_population because of appending */

            data sassy.nonlost_population;
                  set sassy.constat;
                  by pid descending assessid;

                  if first.pid=1 & first.assessid=1;

                   /* to get last assessment period */
                  lastassess=assessid;

            run;




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


/* sampling from each bucket the number of lost patients we removed
   to get back orginal sample size but with only non-lost patients */



            proc surveyselect data=sassy.nonlost_population
```

```
                  method=URS seed=&i sampsize=sassy.nonlost_population
                  outhits out=sample&i;

                  strata bucket;

                  title "Sampling With Replacement From Non-lost
                        Patients: Sample&i";

            run;


/* URS=unrestricted random sampling = equal prob with replacement

sampsize= give dataset with _NSIZE_ variable, which is the number
we want to sample for each bucket

outhits = outputs all obs, so replicates all get outputted          */






            data sample_&i;               /* Simulated lost patients */
                  set sample&i;
                  lost=1;
            run;



            proc append base=sassy.nonlost_population data=sample_&i force;

            run;

      /* now constat2 is data set with all non-lost patients (2358) */


            proc sort data=sassy.nonlost_population;
                  by pid;
            run;





            data Simulated&i;
                  set sassy.nonlost_population;
                  by pid;

                        /* to handle repeat pid for resampled patients */

                  if pid=lag(pid) then pidnum+1;
                        else if pid^=lag(pid) then pidnum=1;

            run;
```

```
data ID&i;
      merge sassy.total_constat Simulated&i;
      by pid pidnum descending assessid;

      retain prevlost prevpid prevpidnum;



      if _N_=1 then do;
            prevlost=lost;
            prevpid=pid;
            prevpidnum=pidnum;
            newid=1;
      end;

                              /* to fix "missing" lost */



      if pid=prevpid & pidnum=prevpidnum & lost^=. then
            prevlost=lost;

            else if pid=prevpid & pidnum=prevpidnum & lost=. then
            lost=prevlost;

            else if pid=prevpid & pidnum^=prevpidnum & lost^=.
                  then do;

                        prevlost=lost;
                        prevpid=pid;
                        prevpidnum=pidnum;
                        newid+1;
                  end;


            else if pid=prevpid & pidnum^=prevpidnum & lost=.
                  then delete;

            else if pid^=prevpid then do;
                        if lost=. then delete;
                              else do;
                                    prevlost=lost;
                                    prevpid=pid;
                                    prevpidnum=pidnum;
                                    newid+1;
                              end;
                  end;


run;
```

```
            proc sort data=ID&i;
                  by newid assessid;
            run;


            proc transpose data=ID&i out=trans&i;
                  by newid;
                  var assessid;
            run;



            data DFS&i;

             /* all nonlost pateints (Simulated lost patients) */

                  merge ID&i trans&i;
                  by newid;

                  lastassess=largest(2, OF COL1-COL17);

*** k=2 to go back 1 assessment period, change k for going back more periods;


                  if last.newid=1;





/* new date of first met/event/censor for Simulated patients last assessment
date is censordt, if 1st met beyond last assess date then no met */


                  newmet=met;
                  newcensordt=censordt;


                  if lost=1 then do;


                        if vitalstatdt^=. then lostdt=(vitalstatdt-
                              dxdt)/365.25;

                              /* Simulated lost date */

                        else if vitalstatdt=. then lostdt=(quality_fudt-
                              dxdt)/365.25;

                              if lostdt<0 then lostdt=(quality_fudt-
                                    dxdt)/365.25;
```

```
/* for pid 256570 error bc vitalstatdt is before dxdt*/


            if met=1 & lostdt <= censordt then do;
                  newmet=0;

            /* Met is unknown after lost date */


                  newcensordt=lostdt;

            end;


            if event=0 & lostdt <= osdt then newcensordt=lostdt;

      /* No death, lost date is new censor date */

      end;



      if newmet=1 | event=1 then newevt=1;

      /* newevt = new event variable for death/recurrence */
            else newevt=0;


      keep newid pid pidnum lost newevt newcensordt;


run;







proc lifetest data=DFS&i noprint outsurv=surv&i;
/* product-limit is the same as Kaplan-Meier */

      time newcensordt*newevt(0);
      strata lost / test=logrank;
      label newcensordt="Years";
      title "KM Estimates of DFS by Lost Status: SIMULATED Lost
            Patients from Sample&i";
run;




data SIMULATED_Survival&i;
      set surv&i;
      by lost newcensordt;
```

```
                  if newcensordt >= 5 & Survival^=. then output;

            run;


            data SIMULATED_Surv5Yr&i;
                  set SIMULATED_Survival&i;
                  by lost;


                  if lost=1;

      /* only care about lost because nonlost doesn't have any changes */

                  if first.lost=1;

                  keep lost newcensordt survival;
            run;




            proc append base=sassy.SIMULATED_Surv5Yr data=SIMULATED_Surv5Yr&i
force;
            run;


      %end;
%mend;



%DFS(1000);



proc print data=sassy.SIMULATED_Surv5Yr;
      var lost survival newcensordt;
run;
proc means data=sassy.SIMULATED_Surv5Yr;
      title "5-Year Survival for Simulated Lost Patients (Delete 1 Follow-
            up)";
      var survival;
run;


proc means data=sassy.SIMULATED_Surv5Yr median;
      var survival;
run;
```

```
************************************************************************;

**** checking how longer the period is between first lost & last visit for
      LOST patients;



proc sort data=raw.continuous_status;
by pid assessid;
run;

proc sort data=sassy.final;
by pid;
run;

data first;
      merge sassy.final(IN=final) raw.continuous_status
(rename=(assessid=firstlost));
      by pid firstlost;

      if final=1;       ** patients in our cohort only;
      if lost=1;        ** keep lost patients only;

      firstdate=nccncarestatdt;

      keep pid firstlost firstdate;
run;




data last;
      set raw.continuous_status(rename=(assessid=lastassess)
            where=(nccncarestatdt^=.));
      by pid lastassess;

      if last.pid=1 & last.lastassess=1;

      lastdate=nccncarestatdt;

      keep pid lastassess lastdate;

run;



data period;
```

```
      merge first(IN=lostonly) last;
      by pid;

      if lostonly=1;

      period=(lastdate-firstdate)/365.25;
      assess_period=period*12;

run;




proc means data=period;
      var period assess_period;
run;

proc means data=period median;
      var period assess_period;
run;



*** Delete 1 follow-up period instead of 2, median=12months;



****************************************************************************;





ods graphics off;
```