AN ANALYSIS OF

BREAST CANCER METASTASIS

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

Jennifer Lee Gildner

December 2011

Advisor: Rebecca Ottesen

Abstract

The main objective of this paper is to evaluate possible socio-economic status, clinical, and treatment associations with the occurrence of distant metastasis in Stage I – III breast cancer patients. After analysis in a logistic regression model, four variables were found to be significant with occurrence of distant metastases. These variables were: education, disease group (Triple-negative, Her2Neu-positive and Luminal A), stage at diagnosis, and concordance to chemotherapy based on the NCCN guidelines. Patients without a college degree were found to be more likely to develop distant metastasis than those with a college degree (OR = 2.46 95% CI 1.44 – 4.23). Triple-negative and Her2Neu-positive patients had higher odds of having distant metastasis than those in with luminal A disease (OR = 3.88 and 3.22 95% CI 2.25 – 6.69 and 1.88 – 5.52, respectively). Stage III patients also had higher odds of having distant metastasis than those with Stage I disease (OR = 5.41 95% CI 2.74 – 10.65). Finally, an unusual result was discovered where patients who were not classified to a chemotherapy guideline were significantly less likely to have distant metastasis than their counterparts who received the recommended chemotherapy (OR = .32 95% CI 0.17 - 0.58).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

Cancer is a disease of growing concern in today's world. New research is done every day in an attempt to discover more about this disease and how it works. There is no doubt that many questions regarding cancer are unanswered. There is a great need for research on this subject to find out better ways to treat and cure cancer patients. This research project focuses on breast cancer, specifically distant metastasis, and attempts to answer questions about the characteristics a patient may have in relation to distant metastasis. A distant metastasis is defined as cancer that has spread from the original (primary) tumor to distant organs or distant lymph nodes. In this case, a distant metastasis would mean the cancer has spread to an area other than the breast (the primary site).

There were two main research questions investigated in this analysis. The first was whether triple-negative breast cancer is associated with developing a distant metastasis. The second was whether a patient's type of treatment is associated with developing a distant metastasis. Typically, triple-negative patients are harder to treat because of their unique status, so it was thought that this group might be associated with higher probability of getting a distant metastasis. In addition, it was thought that patients who received both radiation and chemotherapy, instead of just one or the other, might have a lower probability of developing distant metastasis. Other variables were included in the analysis as controls and/or additional variables of interest. These variables included information about socio-economic status, clinical, and treatment.

The data for this analysis was based on a sample of patients from the City of Hope National Medical Center. The data is stored in the National Comprehensive Cancer Network (NCCN) Outcomes Database. This database consists of many different data tables, each with many different variables within the table. Data is grouped in "raw" data sets that are just pure data as entered into the database and also "derived" data sets that are created from the raw data using predefined programming algorithms. All data is stored in a SQL database and downloaded into SAS data sets that use a hierarchal data structure. Each data set has a corresponding data dictionary which was reviewed to identify variables of interest. The focus for this analysis was a group of patients diagnosed with stage I-III breast cancer and the objective was to choose and test the significance of possible variables that might have an association with whether a breast cancer patient experiences a distant metastasis, using logistic regression.

2. Data Cleaning and the Final Data Set

A major component of this project included programming in SAS to read in data, make necessary exclusions, obtain variables of interest, and create new variables from existing ones for analysis. In order to make the data manageable, the data sets had to be reviewed to identify the cohort and variables of interest. Only certain data sets were read into SAS and then merged to create the final cohort data set.

2.1 Exclusions and Controls

In order to carry out an accurate analysis, exclusions had to be made. Some patients were excluded right when the data was read in to SAS. This included patients who did not have enough follow-up for an analysis that involved radiation therapy (270 days from presentation). Some patients had second episodes of breast cancer, so observations were limited to the first occurrence of cancer and patients with bilateral disease were also excluded. The goal was to investigate treatment and disease group effects on metastases, so any patients with no treatment information were to be deleted. Only patients with treatment defined as adjuvant therapy (treatment given after the primary surgery) and neo-adjuvant therapy (treatment given prior to the primary surgery to shrink the tumor) were included, since these treatments are given to lower the risk of recurrence. Since the other primary goal of the analysis was to identify if disease group is associated with distant metastasis, patients with missing or unknown estrogen or progesterone receptor or Her2Neu status were excluded, because they would not be able to be classified into a disease grouping.

Stage IV patients were excluded because they are a different subset of people. These patients have metastases at diagnosis, where the stage I-III patients do not. Stage IV patients are usually terminal and too far along in the disease to be considered in an analysis looking at predictors of distant recurrence. Since the primary interest was in the characteristics that help predict if a patient gets distant metastasis, it does not make sense to include Stage IV patients that have already metastasized. Stage 0 patients were also excluded because chemotherapy, one of the treatments of interest in this analysis, is not usually given to this group of patients. Stage 0 patients are typically treated as a preventative measure and thus did not fit in to the cohort. Out of the 2,200 total patients in the data set, 964 fit the criteria and were included in the analysis.

2.2 SAS Programming

The first step in the programming process was to identify all variables of interest from the various data dictionaries. Patients were excluded that did not fit the criterion as described in the previous section. A total of 11 different data sets were used in this analysis. These data sets are grouped in a hierarchical database model, meaning that the data is organized into a tree-like structure and linked together by keys. All of the data sets have a common key variable, patient ID (pid), and some share diagnosis ID (dxid) and possibly tumor ID (tumorid) key variables as well. The patients are identified by pid and then each different diagnosis is identified by dxid within the pid key. In addition, the tumorid key identifies different tumors within each diagnosis. Since many of the data sets contain information on patients with recurrence or more than one tumor, there is the possibility for more than one record per patient. In this case, these

records are organized by the dxid and tumorid variables within each pid. However, some of the data sets are already narrowed down to one record per patient, like the demographics data sets. Because of this, data sets had to be sorted and merged in a specific manner in order to get the correct information based on one record per patient. Table 1 summarizes the variables and the corresponding tables they were found in.

Table 1: Data Sets and Variables

Data Set	Variables	Key Variables
Patient Characteristics (derived)	Age at Diagnosis, Income, Race, Follow-Up Radiation Therapy Flag	pid
Clinical Characteristics (derived)	Final Stage	pid, dxid, tumorid
Adjuvant Drug Therapy (derived)	Flags for Adjuvant Treatment Group	pid, dxid
Metastasis Patient Characteristics (derived)	Age at Metastasis	pid, dxid
Surgical Information (derived)	Definitive Surgery Group	pid, dxid, tumorid
Metastatic Sites (raw)	Distant Metastasis variables	pid, dxid
Study Accession (raw)	Education Level, Employment Status, Height, Weight	pid
Insurance (raw)	Insurance Provider	pid
Solid Tumor Stage (raw)	Radiation Flags and Disease Group	pid, dxid, tumorid
Treatment (raw)	Radiation Flags and Neo-adjuvant/Adjuvant Flags	pid, dxid
Concordance (derived)	Concordance Information	pid, dxid, tumorid

Within these data sets, code was used to manipulate and create new variables to be used in the model. In the Adjuvant Drug Therapy data set, a binary variable was created for whether or not a patient had chemotherapy. The Metastasis Patient Characteristics data was used to create a variable for time to metastasis from original diagnosis. As described later, this variable was intended for use in a subsequent analysis. The Metastatic Sites data set included variables that were used to determine where the metastasis occurred and a variable (the response variable) was created for whether or not a patient had distant metastasis or not. The Solid Tumor Stage data set provided useful data for laterality of the patient's breast cancer and for creating a categorical variable for a patient's disease group based on certain clinical characteristics. The raw Treatment data set provided data to classify patients into neo-adjuvant/adjuvant radiation treatment groupings and to ensure patients were only flagged for radiation treatment if the treatment side was equal to the laterality of the breast cancer.

Finally, to create the most complex variables, the derived Concordance data set was used to manipulate the concordance variables for each patient. Since it is possible for patients to be put on several guidelines, some patients had multiple observations to account for concordance on different guidelines. The concordance data set contained variables for concordance (Yes/No/Not evaluated), guideline, version, and reason if the patient was not concordant. The data was transposed to get a row for each corresponding concordance, guideline, version, and reason variable for each patient. The data was then merged with itself to create one row per patient and four sets of each of the four concordance variables (since any patient is on at most four

guidelines). The following merge creates a four columns, one for each guideline a patient may be on. This was repeated for each of the three other variables.

```
**Merge data with itself to get one record per patient;
data Mconcord;
merge trconcord (where=(_NAME_='guideline') keep=pid dxid tumorid
_NAME_ col1-col4 rename=(COL1=Guideline1) rename=(COL2=Guideline2)
rename=(COL3=Guideline3)rename=(COL4=Guideline4))
...
by pid dxid tumorid;
run;
```

The last step in the initial programming was to combine and merge all of the relevant data sets into one final data set with the variables of interest. Care had to be taken to combine data sets with the set of common keys first starting at the tumor level (pid, dxid, tumorid). Then this result could be combined with those data sets at the next level (pid and dxid), and finally combined with those data sets that just use pid as a key. Several sets of merges with trackers were used to accomplish this. In addition, binary variables were created at the appropriate parts of the program to indicate if a patient had radiation treatment and/or a distant metastasis. For example, if the patient had a missing value for metastatic site, then the patient was recorded as having no metastasis. At the end, the final data set was reviewed to ensure there was only one record per patient.

2.3 Predictors

Once the final patient level data set was created, an analysis data set was made. In this analysis data set, variables were created and manipulated so that they could be used in the logistic regression models. A body mass index (BMI) variable was created from the height and weight data, as well as a categorical variable based on BMI. Many of the predictors, such as age, were collapsed into interesting groups that were clinically meaningful. Breakpoints of <50 (pre-menopause), 50-<70 (midrange) and 70+ (older) were chosen due to their relationship with breast cancer care at a biological and clinical level. Continuous variables were tested for linearity in the logit and categorical variables were included if the variable met the appropriate assumptions for the analysis. Groups for categorical variables were collapsed, defined by what seemed to make sense and would be easy to understand and interpret. Collapsing these categorical variables prevented convergence issues from having too few observations in any one group. The analytic data set is also where final exclusions were made, such as patients who were stage IV at diagnosis. Instead of deleting excluded patients, they were output to a second data set to double check that patients were excluded for correct reasons.

In order to be able to use logistic regression, the response variable (distant metastasis) was collapsed into two groups. One group was patients who experienced distant metastasis, and the other was patients who experienced local metastasis or no metastasis. For ease of discussion of the predictor variables, they were grouped into three categories/types: personal and socio-economic status characteristics, clinical characteristics, and treatment characteristics.

The personal and socio-economic status variables included: age at diagnosis, body mass index (BMI), BMI group, race, insurance, education level, employment status, and income. Age at diagnosis was tested for linearity in the logit by using a Box-Tidwell transformation, and failed. BMI was tested for linearity in the logit using the same method, and passed. However, as stated earlier, BMI was not significant as a continuous predictor, so it was grouped by clinically meaningful cutoffs for underweight, normal, overweight, and obese, and then tested as a categorical variable as well. Race was collapsed into the three largest groups: Caucasian, Hispanic, and African American. A fourth group, "Other," was added for those who did not fall into any of those categories. Insurance was broken up into four groups: Managed Care, Medicare, Medicaid, and other. Education level was collapsed into three categories: college degree (college, some college/AA and graduate school), no college degree, and unknown/other. Employment status was broken up into employed (including employed students), student, and other. Table 2 provides a summary of these variables.

Table 2: Personal and Socio-Economic Status Characteristics[1]

Variable	Group	Local or No Metastasis (N=863) n (%)	Distant Metastasis (N=101) n (%)	Total (N=964) n (%)
Age at Diagnosis	--	Mean: 53.5 Median: 52.8	Mean: 53.1 Median: 51.3	Mean: 53.5 Median: 52.6
	Below 50	352 (40.8)	43 (42.6)	395 (41.0)
	50-70	439 (50.9)	42 (41.6)	481 (50.0)
	Above 70	72 (8.3)	16 (15.8)	88 (9.1)
BMI	--	Mean: 27.4 Median: 26.5	Mean: 28.7 Median: 27.6	Mean: 27.5 Median: 26.6
	Underweight	18 (2.09)	1 (0.99)	19 (1.97)
	Normal	320 (37.08)	27 (26.73)	347 (36.00)
	Overweight	281 (32.56)	36 (35.64)	317 (32.88)
	Obese	244 (28.27)	37 (36.63)	281 (29.15)
Race	Caucasian	452 (52.38)	54 (53.47)	506 (52.49)
	Hispanic	230 (26.65)	29 (28.71)	259 (26.87)
	African American	46 (5.33)	6 (5.94)	52 (5.39)
	Other	135 (15.64)	12 (11.88)	147 (15.25)
Insurance	Managed	480 (55.62)	42 (41.58)	522 (54.15)
	Medicare	139 (16.11)	23 (22.77)	162 (16.80)
	Medicaid	234 (27.11)	35 (34.65)	269 (27.90)
	Other	10 (1.16)	1 (0.99)	11 (1.14)

[1] Continued on next page

Table 2 Continued: Personal and Socio-Economic Status Characteristics

Variable	Group	Local or No Metastasis (N=863) n (%)	Distant Metastasis (N=101) n (%)	Total (N=964) n (%)
Education Level[2]	College Degree	411 (47.62)	23 (22.77)	434 (45.02)
	No College	339 (39.28)	50 (49.50)	389 (40.35)
	Unknown	113 (13.09)	28 (27.72)	141 (14.62)
Employment Status	Student	8 (0.93)	2 (1.98)	10 (1.04)
	Employed	391 (45.31)	37 (36.63)	428 (44.40)
	Other	464 (53.77)	62 (61.39)	526 (54.56)
Income	--	Mean: $52,251.31 Median: $50,047.00	Mean: $48,066.07 Median: $42,108.00	Mean: $51,820.80 Median: $49,818.00

[2] Italicized variables in Tables 2-4 were significant in the final model

The clinical variables were stage at diagnosis and disease group. The stage at diagnosis variable was collapsed into stage I, II or III. Of these groups, stage I patients are the least advanced and stage III patients are the most advanced. A disease group variable was created based on estrogen receptor (ER) status, progesterone receptor (PR) status, and Her2Neu status. ER, PR and Her2 are markers on cancer cells that identify how a patient will respond to different therapies. A patient was classified as triple-negative if ER is negative, PR is negative, and Her2Neu status is negative or low positive. A patient was classified as Her2Neu-positive if Her2Neu status is high positive or positive NOS, regardless of ER and PR status. A patient was classified as luminal A if they did not fit in to any of the previously mentioned groups. Table 3 shows a summary of these variables.

Table 3: Clinical Characteristics

Variable	Group	Local or No Metastasis (N=863) n (%)	Distant Metastasis (N=101) n (%)	Total (N=964) n (%)
Stage at Diagnosis	I	262 (30.36)	15 (14.85)	277 (28.73)
	II	433 (50.17)	42 (41.58)	475 (49.27)
	III	168 (19.47)	44 (43.45)	212 (21.99)
Disease Group	Triple-negative	134 (15.53)	33 (32.67)	167 (17.32)
	Her2Neu-positive	170 (19.70)	34 (33.60)	204 (21.12)
	Luminal A	559 (64.77)	34 (33.66)	593 (61.51)

The treatment variables included: definitive surgery group, radiation and/or chemotherapy treatment group, and concordance to NCCN guidelines. Definitive surgery group was defined as: breast conserving surgery, mastectomy, or no cancer directed surgery. The treatment group variable was created to indicate whether a patient had only chemotherapy, only radiation, or both. The concordance variable was created to indicate whether the patient received the

treatment they were recommended to receive by the NCCN guidelines. Each patient is placed on as many guidelines as they are eligible for. The guidelines are different pathways that a patient can be classified on based on their clinical characteristics. These guidelines can recommend different modalities of treatment including chemotherapy, radiation therapy, and/or other treatment. For this analysis, patients were could only be on a radiation and/or chemotherapy guideline or no guideline because of the exclusions that were made. The concordance variable used in this analysis indicates whether or not the patient followed the care of the given guideline that they were on. Table 4 contains summaries for these clinical variables.

Originally, one concordance variable was created that considered both radiation and chemotherapy together. The variable was first broken up into many specific categories for combinations of chemotherapy and radiation concordance. However, during analysis, SAS gave an error due to quasi-complete separation of the data points. The concordance variable was then collapsed into groups for concordant chemotherapy and radiation, concordant radiation (not considering whether or not the patient had a discordant chemotherapy or not on a chemotherapy guideline), concordant chemotherapy (again, not considering discordant radiation or not on a radiation guideline), and not concordant at all. Again, the concordance variable gave some unusual results which were difficult to interpret. So the variable was then split into two variables, one based on chemotherapy and one based on radiation therapy, to investigate where the odd behavior was coming from. A concordance variable was defined for patients on a radiation therapy guideline (yes, no, n/a for concordance) and another variable for patients on a chemotherapy guideline (same categories). From these variables, the analysis was able to pinpoint which therapy was significant, if any.

Table 4: Treatment Characteristics

Variable	Group	Local or No Metastasis (N=863) n (%)	Distant Metastasis (N=101) n (%)	Total (N=964) n (%)
Definitive Surgery Group	Breast Conserving Surgery (BCS)	484 (56.08)	46 (45.54)	530 (54.98)
	Mastectomy	367 (42.53)	53 (52.48)	420 (43.57)
	No Definitive Surgery	12 (1.39)	2 (1.98)	14 (1.45)
Treatment Group	Chemotherapy Alone	213 (24.68)	26 (25.74)	239 (24.79)
	Radiation Alone	174 (20.16)	8 (7.92)	182 (18.88)
	Chemotherapy and Radiation	476 (55.06)	67 (66.34)	543 (56.33)
Chemotherapy Concordance	Yes	465 (53.88)	61 (60.40)	526 (54.56)
	No	157 (18.19)	18 (17.82)	175 (18.15)
	N/A	241 (27.93)	22 (21.78)	263 (27.28)
Radiation Concordance	Yes	488 (56.55)	45 (44.55)	533 (55.29)
	No	45 (5.21)	9 (8.91)	54 (5.60)
	N/A	330 (38.24)	47 (46.53)	377 (39.11)

The concordance variable required creative programming in order to create the variable since the sets of guideline and concordance variables had to be examined and combined to create one variable. As previously described, concordance and guideline each had a set of four variables for each patient. To create the chemotherapy and radiation concordance variables, arrays were set-up for the guideline and concordance variables. In a do loop, each guideline was tested for type of therapy recommended and the corresponding concordance was extracted. The concordance variables were increased by one if the patient was concordant. For example:

```
do i=1 to 4;
  if conc(i)='Yes' and (substr(guide(i),1,4)='invx' or
     guide(i)='invtx1' or guide(i)='invtx1b' or guide(i)='invtx1c')
   then RTConcord=rtconcord+1;
  if conc(i)='Yes' and (substr(guide(i),1,4)='inva' or
     guide(i)='invtx1' or guide(i)='invtx1b' or guide(i)='invtx1c' or
     guide(i)='invtx1a' or guide(i)='invtx2')
   then ChemoConcord=chemoconcord+1;
end;
```

After deciding when a patient was concordant, there had to be a way to break up not concordant patients and the "not applicable" patients. At this point, patients either had a value of 1 if they were concordant or 0 if they were not concordant or not applicable. These "N/A" patients included those that were not evaluated for concordance or were put on a different guideline not relevant to the given modality. This occurs when a patient does not have the clinical characteristics specific to a guideline. In order to break up the groups, another do loop was created to check the two sets of four variables. The loop would only start if the patient did not have a "Yes" for either concordance variable. The following code was used to break up these groups:

```
if (chemconcord=0 or rtconcord=0) and guide(1)~='Not evaluated for
  concordance' then do i=1 to 4;
 if conc(i)='No' and (substr(guide(i),1,4)='invx' or guide(i)='invtx1'
     or guidei)='invtx1b'
     or guide(i)='invtx1c')
  then rtconcord=2;
 if conc(i)='No' and (substr(guide(i),1,4)='inva' or guide(i)='invtx1'
     or guide(i)='invtx1b'
     or guide(i)='invtx1c' or guide(i)='invtx1a' or guide(i)='invtx2')
  then chemconcord=2;
end;
```

After the do loop and testing for non-concordance, the concordance variables ended up with a value of 1 if the patient was concordant, a value of 2 if the patient was not concordant, and a value of 0 of the patient was not on a guideline.

3. Analysis

Once the final analytic data set was created, `PROC FREQ` was run on categorical predictors and `PROC MEANS` was run on continuous predictors to get a sense of the data and indicate any missing values. These results are summarized in the three tables shown in the previous section. In SAS, `PROC LOGISTIC` was used to run logistic regression models and investigate relationship between the predictors of interest and the distant metastasis response variable. The outcome modeled was distant metastases equal to yes, so the odds ratio interpretations are thus predicting the odds of a patient getting distant metastasis. Since most of the predictors were categorical, reference groups had to be assigned for each variable. These reference groups were defined based on what clinically would be least likely to result in distant metastasis for ease of interpretation of the results. This means that the groups of greatest interest (more likely to get distant metastasis) would have a predictor coefficient in the model. For example, Stage "I" was chosen as the reference group for the stage variable since Stage I patients are less advanced and thus were predicted to be less likely to get a distant metastasis. In addition, "Luminal A" was chosen as the reference for disease group, "College degree" was chosen as the reference for education, and "Yes" was chosen as the reference group for the Concordance variables.

In addition to testing all the covariates, possible interactions were also considered. Instead of relying on a stepwise regression to come up with interactions that might not make sense clinically, meaningful interactions were established a priori and then tested in the model. First, the clinically meaningful interactions were tested, such as age with treatment group, treatment group and disease group, and various other combinations. However, all of the interactions tested turned out to be insignificant. In the end, stepwise selection was used as a back-up in case any significant interactions were missed, but still, none came up as significant and it was decided that the effect of any of the predictors did not depend on other predictors.

Three models were considered: a saturated model with all covariates of interest, a significant borderline model with all covariates that had a p-value <0.10, and a significant terms only model. The saturated model included many insignificant variables and did not turn out to be a useful model. From the saturated model, least significant variables were removed one at a time until all variables fit within the significance level desired. The final significant model included one socio-economic variable for education level, two clinical variables for disease group and stage, and one treatment variable for chemotherapy concordance. The borderline model also included the categorical age group variable and continuous BMI variable in addition to the significant model predictors, however these variables did not add much more information to the model. The significant model was used as the final model as reported in this paper.

3.1 Assumptions

One of the biggest advantages of logistic regression is that there are not as many restrictive assumptions as in linear regression. The first assumption is that there is linearity between the logits and explanatory variables. However, since all of the significant predictors in the final model are categorical, this was not an assumption of concern in this analysis. Independent observations are also needed, but since this data is from de-identified retrospective data collected, this was not something that could be controlled for during this analysis. However, it is

obvious that one patient getting a distant metastasis has nothing to do with another patient's experience with distant metastasis and there also is no reason to think that the patients have any genetic relationships. Large sample size is ideal for logistic regression and this was not a problem (with a final sample size of 964 patients) for this data set. Another assumption is no multicollinearity between the explanatory variables, and no outliers. These two assumptions will be addressed in the Diagnostics section of the paper. Otherwise, no assumptions were violated.

3.2 Results

The final model included the variables for disease group, chemotherapy concordance, stage at diagnosis, and education level. Table 5 provides a summary of odds ratio estimates and confidence intervals for each significant variable in the model.

Table 5: Significant Model Estimates

Variable	Overall P-Value	Group	Odds Ratio Estimate	95% Confidence Interval
Disease Group	<0.0001	Luminal A	Reference	
		Her2Neu-positive	3.223	(1.884, 5.516)
		Triple-negative	3.881	(2.250, 6.694)
Chemotherapy Concordance	0.0008	Yes	Reference	
		No	0.911	(0.492, 1.686)
		N/A	0.317	(0.173, 0.583)
Stage at Diagnosis	<0.0001	I	Reference	
		II	1.234	(0.650, 2.341)
		III	5.406	(2.744, 10.650)
Education Level	<0.0001	College Degree	Reference	
		No College	2.462	(1.435, 4.225)
		Unknown	4.810	(2.571, 8.999)

As seen in Table 5, the significance of the disease group variable provided an explanation to the first research question of interest. Patients with triple-negative disease were estimated to be 3.88 times more likely to develop a distant metastasis than those with luminal A (95% CI 2.25 – 6.69). Similarly, patients in the Her2Neu-positive disease group were also associated with higher odds of having distant metastasis than patients in the luminal A group (OR = 3.22 95% CI 1.88 – 5.52).

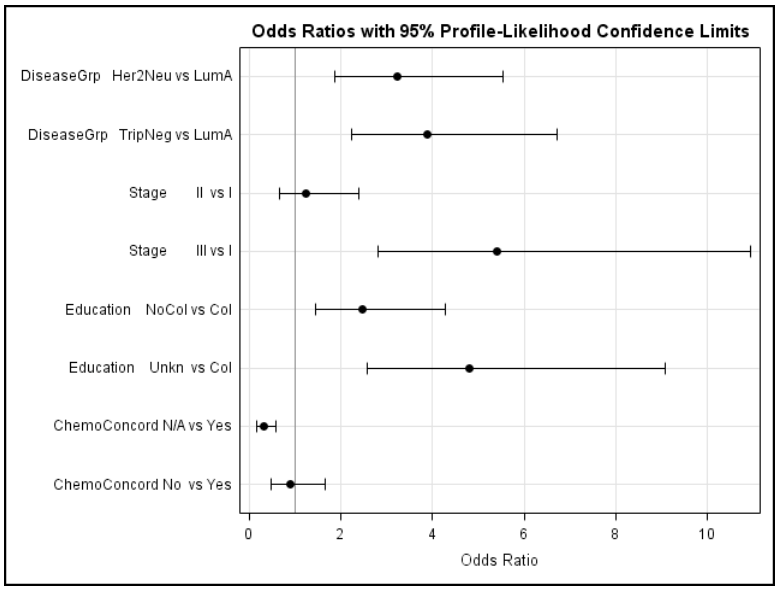
As previously mentioned, the concordance variable gave some unusual results. But, the significance of the chemotherapy concordance provided some explanation for the second research question relating to a patient's treatment. Concordance was originally one variable that considered both radiation and chemotherapy. Since this variable behaved unusually (with the N/A patients less likely to get distant metastasis), the variable was broken down into two variables, one for radiation concordance and one for chemotherapy concordance, to target where this problem was coming from. The chemotherapy concordance turned out to be the culprit. The radiation variable turned out not to be a predictor of distant recurrence and was not included in the final model. As seen in Table 5, the "N/A" patients (those who were not evaluated for concordance or not on a guideline) were associated with being significantly less likely to have

distant metastasis than patients who were concordant for chemotherapy (OR = 0.32 95% CI 0.17 – 0.58). Patients who were not concordant were not significantly associated with developing distant metastasis than those who were concordant.

Stage at diagnosis had one significantly different group, which were the Stage III patients. Stage III patients were associated with 5.41 higher odds of having distant metastasis than Stage I patients (95% CI 2.74 – 10.65), after considering the other variables in the model. Stage II patients were not significantly more or less likely to have distant metastasis than Stage I patients.

Finally, education level was the only significant non-treatment or non-clinical variable. Patients with no college education were 2.46 times more likely to develop distant metastasis than those with a college education (95% CI 1.44 – 4.23), considering the other covariates in the model. Similarly, patients with unknown education level also were estimated to have higher odds of getting distant metastasis than patients with a college degree (OR = 4.81 95% CI 2.57 – 9.00). Figure 1 provides a visual representation of the odds ratios for each level of the covariates. The width of the band corresponds to the width of the 95% confidence interval.

Figure 1: Odds Ratios Plot



3.3 Diagnostics

In order to ensure the validity of the model and to solidify any conclusions, it is essential to check model fit and other diagnostics. SAS was used to run a Hosmer-Lemeshow Goodness-of-Fit Test. The result was a Chi-Square test statistic of 2.4116 with 8 degrees of freedom and a p-value of 0.9657, which means that there is no evidence that this model is not a good fit to the data. In addition, Deviance and Pearson Goodness-of-Fit Statistics for the logistic regression residuals were tested. Deviance gave a value of 70.4679 with 68 degrees of freedom, which gave a ratio (value/df) of 1.0363 and a p-value of 0.3951. Also, Pearson gave a value of 60.9382 with 68 degrees of freedom, which is a ratio of 0.8961 and p-value of 0.7157. Since the ratios were close to 1, there is no indication of severe under- or over-dispersion. As there were no dispersion issues, there does not appear to be evidence that the binomial assumption is not valid

for this model. SAS also gives an R^2 value of 0.2057. Although this value is not very impressive, this is sometimes the case with real patient based data and is not something to be very concerned with in this analysis.

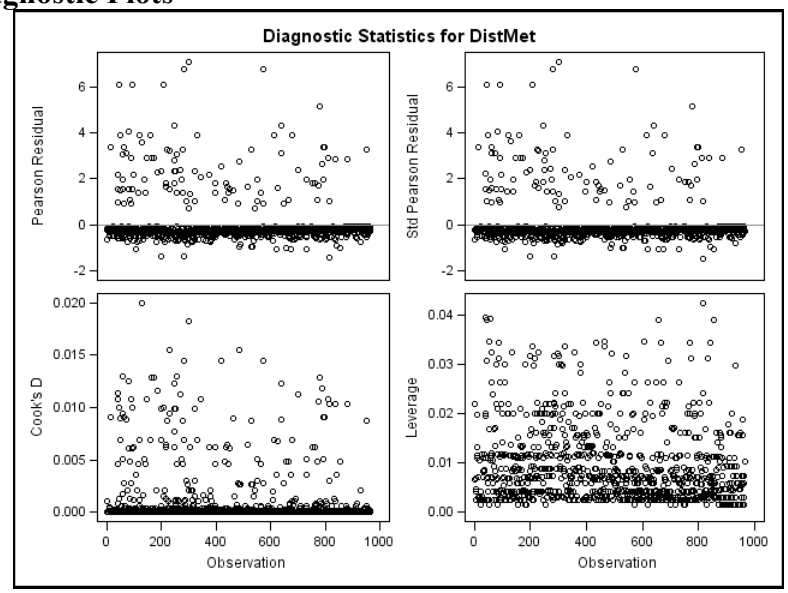
Multicollinearity among the predictors was tested using `PROC CORR` with the Spearman correlation coefficients, as the predictors were categorical variables. Table 6 shows a summary of these results. Some of the covariates have significant correlations, however this is usually the case with large sample sizes. Since none of the correlations appear to be very large and standard errors of the estimates are low, multicollinearity does not seem to be an issue in this model.

Table 6: Spearman Correlation Coefficients

$\mathbf{r_s}$ *Prob>|r_s|*	Disease Group	Stage at Diagnosis	Education Level	Chemotherapy Concordance
Disease Group	**1.000**	**-0.114** *0.0004*	**-0.035** *0.2768*	**0.126** *<0.0001*
Stage at Diagnosis	**-0.114** *0.0004*	**1.000**	**0.087** *0.0069*	**-0.282** *<0.0001*
Education Level	**-0.035** *0.2768*	**0.087** *0.0069*	**1.000**	**0.024** *0.4482*
Chemotherapy Concordance	**0.126** *<0.0001*	**-0.282** *<0.0001*	**0.024** *0.4482*	**1.000**

Finally, unusual observations and case-influence diagnostics were investigated. Figure 2 shows diagnostic plots for the model. The top two plots are residual plots. Note that the distinct separation of the residuals is due to the different response groups. Distant metastasis patients have the positive residuals, while the no metastasis patients have the negative residuals. If anything, the residual plots show a distinct difference between the no distant metastasis patients and the distant metastasis patients. These plots also show a random scatter of points and there is no distinct curvature or pattern. In this case, Pearson residuals are equivalent to the standardized Pearson residuals because of the large sample size. The Pearson and standardized Pearson residuals were output to a new data set and compared to ensure that their values were equal. In most cases, the values differed only after the first decimal place. It appears that many of the residual values are high (>3.3), however this is not unusual for real life data. There were actually only 21 observations out of the total 964 with a standardized Pearson residual > 3.3 out of 964 observations. These 21 observations were printed with their corresponding leverage values and Cook's distance to investigate possible high leverage or influence. None of these observations had high leverage or influence, so these observations were not considered a problem. The leverage and influence plots in Figure 2 also show that none of the observations had very high leverage or influence. These measure potential effects of the observations on the model results.

Figure 2: Diagnostic Plots



Those without distant metastasis tend to have smaller residuals and leverage, which indicates these patients probably have more traits in common. On the other hand, the distant metastasis patients have many more outliers, higher leverage, and residuals, which might indicate that these patients are more unique and have many different characteristics. This observation just reinforces the unpredictability of cancer and why it is so difficult to make predictions about the disease. In addition, these disparities might have come about because of the sample size differences between the groups. The local or no metastasis group had 863 observations and the distant metastasis group had 101 observations. Thus it makes sense that since the distant metastasis patients are so different from their counterparts, they would have higher leverage values. The sample size disparity also makes sense because distant metastasis is much less common for the average breast cancer patient. A study from the Journal of Clinical Oncology reported that a small percentage (31 out of 226, 13.7%) of breast cancer patients with pathologic complete response after chemotherapy experience distant metastasis. This is similar to the results in this study, with 101 out of 964 (10.5%) patients experiencing distant metastasis.

4. Conclusions

The last part of the analysis for this project includes interpretation and explanations for the model results. From the maximum likelihood estimates, a function for the log-odds of developing distant metastasis can be calculated:
$\log(odds) = -3.8191 + 1.1704x_1 + 1.3560x_2 + 0.2101x_3 + 1.6875x_4 + 0.9010x_5 + 1.5708x_6 - 1.1475x_7 - 0.0929x_8$
where each variable is 1 if the patient is in that group or 0 if the patient is not. The variables are:

x_1: Disease group "Her2Neu-positive"
x_2: Disease group "Triple-negative"
x_3: Stage "II"
x_4: Stage "III"
x_5: Education level "No college"
x_6: Education level "Unknown"
x_7: Chemotherapy concordance "N/A"
x_8: Chemotherapy concordance "No"

The log-odds of event occurrence is difficult to interpret, so the log-odds function can be converted into an odds function using the formula: $odds = e^{\log(odds)}$. Finally, the probability function is then given by $\hat{p} = \frac{odds}{1 + odds}$. We end up with the final probability function for this analysis: $\hat{p} = \frac{e^{-3.8191+1.1704x_1+1.3560x_2+0.2101x_3+1.6875x_4+0.9010x_5+1.5708x_6-1.1475x_7-0.0929x_8}}{1+e^{-3.8191+1.1704x_1+1.3560x_2+0.2101x_3+1.6875x_4+0.9010x_5+1.5708x_6-1.1475x_7-0.0929x_8}}$. This function models the probability of a patient developing a distant metastasis, given the characteristic values of the patient.

As displayed in Figure 3, triple-negative patients have a consistently higher probability of developing distant metastasis, while luminal A patients are much less likely to have distant metastasis. Triple-negative patients have negative estrogen receptors, progesterone receptors, and Her2Neu, which means the growth of the cancer is not supported by the hormones estrogen and progesterone, nor by the presence of too many Her2Neu receptors. As a result, triple-negative breast cancer does not respond to hormonal therapy or other targeted therapies. Since triple-negative patients require unusual treatment courses, it makes sense that these patients would be more likely to develop a distant metastasis because they are so different from other patients. On the other hand, luminal A patients are much more common and easier to treat. Patients in the luminal A group have markers on their cells to indicate that they will respond well the hormone therapies that are common for breast cancer treatment. These patients generally respond better to treatment and are therefore less likely to develop distant metastasis.

Figure 3: Probability Plot by Disease Group

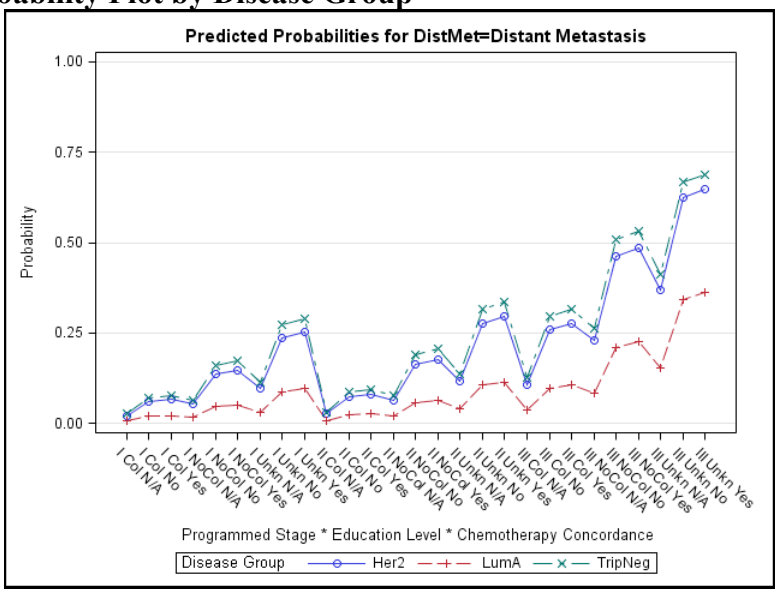


Figure 4 illustrates the probability of having distant metastasis for patients in the different chemotherapy concordance groups. Patients in the N/A group were consistently less likely to have distant metastasis. This result is unusual and unexpected, which makes it somewhat difficult to interpret. Patients in this group cannot be placed on a guideline due to some unknown clinical characteristics so this may just be a chance observation on a group of patients with mixed clinical characteristics. The rates of concordance groupings were evaluated to compare the no metastasis patients with the distant metastasis patients across all the variable groupings in an attempt to discover why these patients received such an unusual result. No apparent differences were discovered to explain any special characteristics that the N/A patients might have. However, this was expected since there were also no significant interactions. The N/A patients cannot be placed on a guideline due to some missing clinical characteristic or some other unknown reason, so it is difficult to explain any reasons these patients might be less likely to have a distant metastasis. Further research is necessary to understand why this result was obtained and to investigate whether these results were observed only for this subset of patients.

Figure 4: Probability Plot by Chemotherapy Concordance

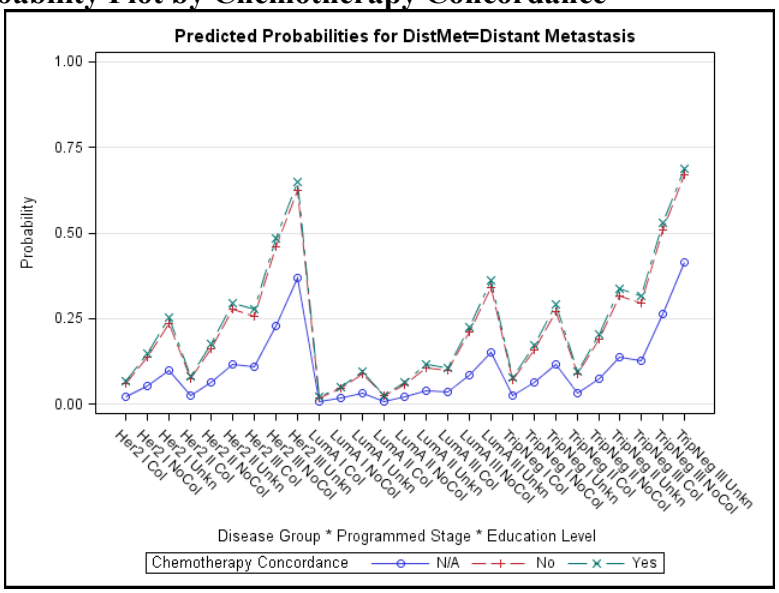
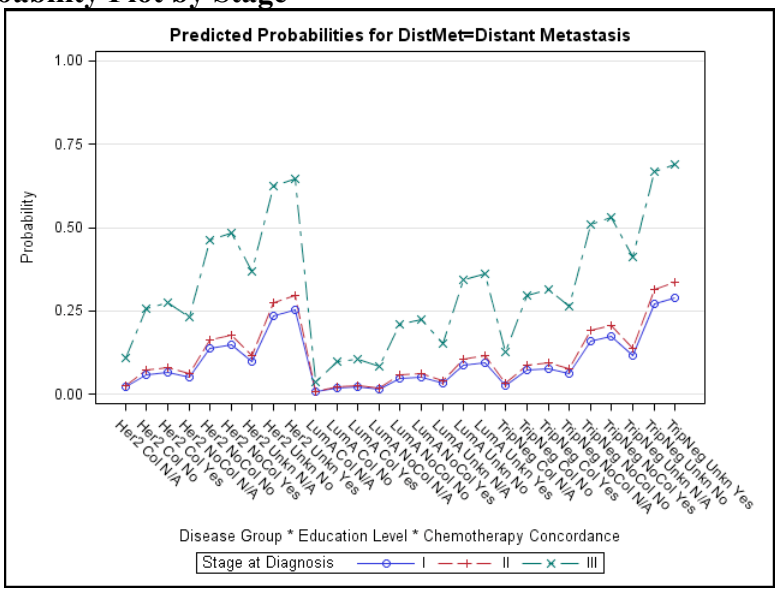


Figure 5 shows probability differences for the stage groups. It is clear that Stage III patients are much more likely to develop distant metastasis. Stage III patients are more advanced in the disease at diagnosis than Stage I patients. Stage III patients are more advanced because they have larger tumors and cancer in more lymph nodes (which allows cancer to spread), but the cancer has not actually metastasized yet. Thus, because Stage III patients are further along they are usually harder to treat and have a higher probability of developing distant metastasis.

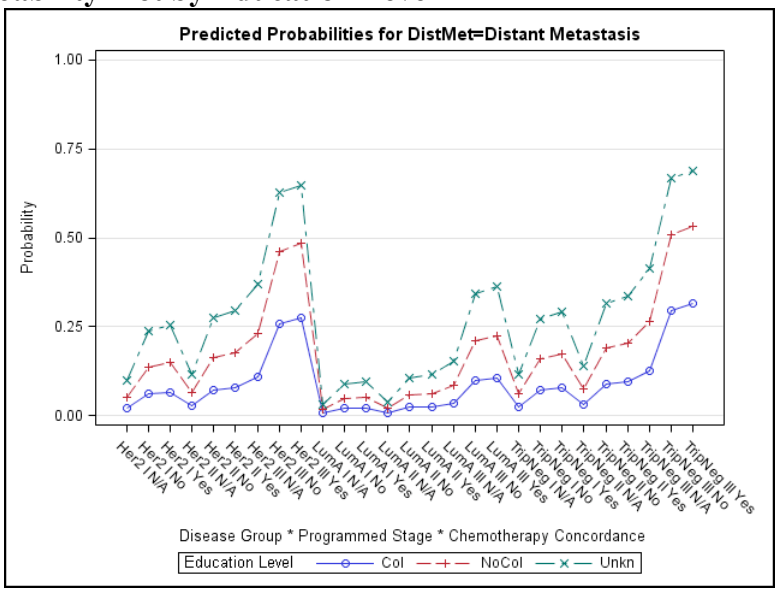
Figure 5: Probability Plot by Stage



As discussed previously, it was found that patients with a college degree were less likely to develop a distant metastasis. This result could be due to patients with a higher education level being more informed and more likely to get the treatments they need. Patients with less education may go untreated for longer or not receive the treatment they need at all. It is also

possible that patients with more education have better jobs and might have better access to healthcare. There are many explanations for this result, but further research could be used to pinpoint the exact reasons. Figure 6 provides a visual representation of these probability differences for education.

Figure 6: Probability Plot by Education Level



5. Limitations

This analysis had several limitations. The first limitation was that recurrence data is difficult to come by in breast cancer patients. These patients have to be followed and observed for long periods of time in order to find any recurrence of the disease. These patients survive long after their initial diagnosis so there is a chance patients will drop out of the study for one reason or another, or even expire due to another non-breast cancer related cause and before experiencing any recurrence. For this reason, records on distant metastasis can be incomplete and difficult to obtain.

In addition, these patients are from a limited sample from the City of Hope. If there were a way to access all patient records, including care at outside clinics, to be sure distant metastasis had or had not occurred, this would provide an ideal analysis. However, this is certainly unrealistic and a possible limitation of this study.

Another limitation of this study is lack of knowledge of clinical meaning and in depth knowledge of the disease. It would be ideal to work with a physician to help identify possible predictors, because of their knowledge about what contributes to the disease. Additionally, a physician could provide clinical opinions on the analysis and results. This would allow for better interpretation and would significantly strengthen the results.

6. Suggestions for further research

Originally, the analysis included an additional research question about whether treatment group or disease group is associated with "quicker" time to metastasis. As proposed, this analysis would have included only patients that actually experienced a distant metastasis and investigated a regression model, which would have included an age at metastasis variable and a time to metastasis variable. Since the data set for this analysis only included 101 patients with distant metastasis, the sample size was not sufficient to conduct such an analysis. This resulted in a more in-depth analysis for the first research question. However, it would be interesting to conduct analysis on this secondary question if provided a larger sample size. This could be done using a disease-free survival analysis to look at time to recurrence. However, the limited follow-up information as noted in the limitations section makes this analysis far beyond the scope of this research paper.

Another interesting question is what factors influence where the breast cancer spreads? The National Cancer Institute states that the main sites of breast cancer metastasis are the lungs, liver, and bones. It makes sense that cancer would spread to the closest areas to the breast, but perhaps there are other reasons for recurrence patterns. An analysis could be done on patient characteristics and clinical variables and with different metastatic sites to see if there is any relation between certain these characteristics and their primary metastatic sites. A larger data set would be necessary for this analysis as well.

In addition, much research has been done on gene expression and genetics, and it would be interesting to do further research in this area. The National Foundation for Cancer Research has done research on which genes suppress metastasis and how their metastasis-suppressing function is regulated, but a more complete understanding has yet to be discovered. This provides another question to be analyzed: whether risk for getting distant metastasis is related to genetics in any way. However, the genetic testing data in this database was very limited.

Research is key to developing new treatments and understandings of this disease. Continued investigation of these research ideas, in addition to other important questions, is essential to understanding cancer and working towards preventing and treating it effectively.

7. References

- Agresti, Alan. An Introduction to Categorical Data Analysis. New York: John Wiley & Sons, Inc. 1996.
- Doi, Jimmy. "Categorical Data Analysis, Class Notes." Cal Poly, San Luis Obispo. Lecture.
- Hosmer, David W, and Stanley Lemeshow. Applied Logistic Regression. New York: John Wiley & Sons, Inc. 2000.
- National Cancer Institute. "Dictionary of Cancer Terms." 27 Oct 2011. <http://www.cancer.gov/dictionary>.
- Ottesen, Rebecca. "In Class Lecture Notes." Cal Poly, San Luis Obispo. Lecture.
- Pagano, Marcello, and Kimberlee Gauvreau. Principles of Biostatistics. Belmont: Duxbury Press. 1993.
- Rosner, Bernard. Fundamentals of Biostatistics. Pacific Grove: Brooks/Cole, 2000.
- SAS Customer Support Knowledge Base. "SAS 9.2 Product Documentation." 20 Sep 2011. <http://support.sas.com/documentation/92/index.html>.
- UCLA Academic Technology Services. "Resources to help you learn and use SAS." 10 Oct 2011. <http://www.ats.ucla.edu/stat/sas/>.

8. Appendix

The following SAS code was used to carry out this analysis:
(Please note that the data for this analysis is confidential and is on file with Rebecca Ottesen)

```
**create libraries for derived and raw data;
      libname ddata 'G:\Senior Project\sasdata\derived';
      libname rawdata 'G:\Senior Project\sasdata\raw';
**tell SAS where to look for format catalogs;
      option fmtsearch=(ddata.ddformats rawdata.formats);
**Read in Patient Characteristics derived data set;
      data patchar; set ddata.patient_characteristics (drop=cid)
      **delete all patients without Follow-up for Radiation Therapy Analyses;
            if FUrt=1;
      run;
**Read in Clinical Characteristics derived data set;
      data clnchar; set ddata.clinical_characteristics (drop=cid);
      **Keep only stage I-III patients;
            if finalstg>=23 and finalstg<=27.5;
      run;
**Read in Adjuvant Drug Therapy derived data set: Flags;
      data adjtreat; set ddata.adjuvant_drug_therapy (drop=cid);
      **Create flags indicating if a patient received chemo
            if adjtxgroup=-99 then delete;
                  else if adjtxgroup in(1,2,6,7) then chemflag=1
                  else chemflag=0;
      run;
**Read in Metastasis Patient Characteristics derived data set;
      data mets; set ddata.mets_patient_characteristics (drop=cid);
      run;
**Read in Surgical Information derived data set;
      data surginf; set ddata.surgical_information (drop=cid);
      run;
**Read in Concordance derived data set;
      data concord; set ddata.concordance_status (drop=cid);
      run;
**Transpose concordance data to get concordance,guideline,version,nreason
**rows for each patient;
      proc transpose data=concord out=trconcord;
            var concordance guideline version nreason;
            by pid dxid tumorid;
      run;
**Merge data with itself to get one record per patient;
      data Mconcord;
            merge trconcord (where=(_NAME_='concordance')
                    keep=pid dxid tumorid _NAME_ COL1-COL4
                    rename=(COL1=Concordance1) rename=(COL2=Concordance2)
                    rename=(COL3=Concordance3) rename=(COL4=Concordance4))
                  trconcord (where=(_NAME_='guideline')
                    keep=pid dxid tumorid _NAME_ col1-col4
                    rename=(COL1=Guideline1) rename=(COL2=Guideline2)
                    rename=(COL3=Guideline3) rename=(COL4=Guideline4))
                  trconcord (where=(_NAME_='version')
                    keep=pid dxid tumorid _NAME_ col1-col4
                    rename=(COL1=Version1) rename=(COL2=Version2)
                    rename=(COL3=Version3) rename=(COL4=Version4))
                  trconcord (where=(_NAME_='Nreason')
```

```
                    keep=pid dxid tumorid _NAME_ col1-col4
                    rename=(COL1=Reason1) rename=(COL2=Reason2)
                    rename=(COL3=Reason3) rename=(COL4=Reason4));
            by pid dxid tumorid;
      run;
**Read in Metsites raw data set;
      data metsite; set rawdata.metastatic_sites (drop=cid studyid)
            if initial=0 and ((site in(5,6,7,8,9,10,11,12,13,15,17,18,19,20))
                  or (site=16 and sitedt<mdy(1,1,2003)))
                  then DistMet=3; **Distant Met;
            else if site~=. then DistMet=2; **Local Met;
      run;
**Read in Study Accession raw data set;
      data stdyacc; set rawdata.study_accession (drop=cid studyid);
      run;
**Create format for new insurance variable, store in ddformats catalog;
      proc format library=ddata.ddformats;
            value newins
                  .='Missing'
                  -1='Unknown'
                  0='Other'
                  1='Managed'
                  2='Indemnity'
                  4='Medicaid/Indigent'
                  5='Medicare alone'
                  5.5='Medicare + Supp'
                  5.75='Medicare + Managed'
                  6='Self-Pay';
      run;
**Read in Insurance raw data set;
      data insurance; set rawdata.insurance (drop=cid studyid);
            if assessid=0; **keep only first record of insurance;
            *** reprogram insurance;  * per MH grid 043003;
            if ins1type=2 or ins2type=2 then do;
                  if ins1type~=5 and ins2type~=5 then newinsur=2;
                  else if ins1type=5 or ins2type=5 then newinsur=5.5;
            end;
            else if ins1type=1 or ins2type=1 then do;
                  if ins1type not in(4,5) and ins2type not in(4,5) then
                        newinsur=1;
                  else if ins1type=5 or ins2type=5 then newinsur=5.75;
                  * medicare+managed;
                  else if ins1type=4 or ins2type=4 then newinsur=4;
            end;
            else if ins1type in(4,7) or ins2type in(4,7) then do;
                  if ins1type~=5 and ins2type~=5 then newinsur=4;
                  else if ins1type=5 or ins2type=5 then newinsur=5.5;
            end;
            else if ins1type=6 or ins2type=6 then do;
                  if ins1type~=5 and ins2type~=5 then newinsur=6;
                  else if ins1type=5 or ins2type=5 then newinsur=5.5;
            end;
            else if ins1type in(0,3,8,9,10,11) or ins2type in(0,8,3,9,10,11)
                  then do;
                  if ins1type~=5 and ins2type~=5 then newinsur=0;
                  else if ins1type=5 or ins2type=5 then newinsur=5.5;
            end;
            else if (ins1type in(-1) and ins2type in(-1,.)) or
```

```
                  (ins1type in(-1,.) and ins2type in(-1)) then newinsur=-1;
            else if ins1type=5 and ins2type in(0,5) then newinsur=5.5;
            else if ins1type=5 and ins2type in(-1,.) then newinsur=5;
            else if ins1type in(-1,.) and ins2type=5 then newinsur=5;
            else if ins1type=. and ins2type=. then newinsur=.;
            format newinsur newins.;
      run;
**Read in Solid Tumor Stage raw data set;
      data ststge; set rawdata.solid_tumor_stage (drop=cid studyid);
      run;
**Read in raw Treatment data set where txcat=10(radiation) and indication
**in(1,2,3,4);
      data treat; set rawdata.treatment (drop=cid studyid);
            if txcat=10;
            where indication in(1,2,4);
      run;
**Merge data by pid dxid and tumorid;
      data merge1;
            merge clnchar (in=a) ststge surginf Mconcord;
            by pid dxid tumorid;
            if a=1 then output;
            **only keeps narrowed down records from derived data (one record
            **per patient);
      run;
**Merge data by pid and dxid;
      data merge2;
            merge merge1 (in=a) treat adjtreat mets metsite;
            by pid dxid;
            if site=. then DistMet=1;  **define no met;
            if a=1 and side=laterality then do;
                  **flag patients as having radiation if side=laterality
                  rtflag=1;
                  output;
            end;
            else if a=1 then do;
                  **otherwise keep narrowed records and flag no radiation
                  rtflag=0;
                  output;
            end;
      run;
**sort to pick distant sites trumping other sites;
      proc sort data=merge2;
            by pid dxid distmet;
      run;
**overwrite merge 2 to narrow down one row per patient by keeping last
**diagnosis record;
      data merge2; set merge2;
            by pid dxid distmet;
            if last.dxid=1;
      run;
**Merge data by pid;
      data mergelast;
            merge patchar (in=a) merge2 (in=b) stdyacc insurance;
            by pid;
            if a=1 and b=1 then output;
            **only keeps narrowed down records from derived data;
      run;
**Check to see if there is more than one of the same value for a variable;
```

```
      proc freq data=mergelast noprint;
            table pid / out=count;
      run;
**print pid's with repeats;
      proc print;
            where count>1;
      run;
**Define formats, save to derived data format catalog;
      proc format library=ddata.ddformats;
            **Disease Groups format;
            value tripneg
                  3='TripNeg'
                  2='Her2'
                  1='LumA';
            **Treatment Groups format;
            value treatgrp
                  1='RT alone'
                  2='RT+Chemo'
                  3='Chemo alone';
            **Distant Met Definition format;
            value dist
                  1='Distant Metastasis'
                  0='Local or no met';
            **BMI group format;
            value bmi
                  1='Underweight'
                  2='Normal'
                  3='Overweight'
                  4='Obese';
            **RT/Chemo Concordance variable format;
            value rtchconc
                  3='N/A'
                  1='Yes'
                  2='No';
            **Age group format;
            value age
                  1='Below 50'
                  2='50-70'
                  3='Above 70';
            **Stage variable format;
            value stage
                  1='I'
                  2='II'
                  3='III';
            **Race variable format;
            value rce
                  1='Caucasian'
                  2='Hispanic'
                  3='African American'
                  4='Other';
            **Education variable format;
            value edu
                  2='Col'
                  1='NoCol'
                  3='Unkn';
            **Employment variable format;
            value emp
                  1='Employed'
```

```
                2='Student'
                3='Other';
          **Insurance variable format;
          value insur
                1='Managed'
                2='Medicare'
                3='Medicaid'
                4='Other';
      run;
**Create analysis data set from final data set;
      data analysis exclude; set mergelast
       (keep=pid dxid tumorid her2neu metage diagage distmet finalstg dsgroup
       racecomp p053001 edustat empstatdx initial ins1type ins2type newinsur
       site furt laterality adjtxgroup indication side heightpres weightpres
       chemflag rtflag tmarker1 tmarker2 concordance: guideline:);
      **create disease groups;
          if tmarker1 in(.,-1) or tmarker2 in(.,-1) or Her2neu in(.,-1)
          then delete;
          else if tmarker1=0 and tmarker2=0 and (her2neu in(1,2)) then
          DiseaseGrp=3;
          else if her2neu in(3,4) then DiseaseGrp=2;
          else DiseaseGrp=1;
      **redifine distmet and combine local/no met for logistic regression;
          if distmet=3 then distmet=1;
                else if distmet=2 or distmet=1 then distmet=0;
      **create BMI variable;
          if heightpres~=-1 then do;
                heightM=heightpres/100;
                BMI=weightpres/(heightM**2);
          end;
      **create BMI groups;
          if BMI<18.5 then BMIGrp=1;
          else if BMI>=18.5 and BMI<25 then BMIGrp=2;
          else if BMI>=25 and BMI<30 then BMIGrp=3;
          else if BMI>=30 then BMIGrp=4;
      **create variables to test for linearity in the logit;
          linage=diagage*log(diagage);
          if p053001>=0 then lininc=p053001*log(p053001);
          if bmi~=. then linbmi=bmi*log(bmi);
      **Define concordance variable;
          array conc (4) concordance:;
          array guide (4) guideline:;
          RTConcord=0;
          ChemoConcord=0;
          do i=1 to 4;
                if conc(i)='Yes' and (substr(guide(i),1,4)='invx' or
                      guide(i)='invtx1' or guide(i)='invtx1b' or
                      guide(i)='invtx1c') then RTConcord=rtconcord+1;
                if conc(i)='Yes' and (substr(guide(i),1,4)='inva' or
                      guide(i)='invtx1' or guide(i)='invtx1b' or
                      guide(i)='invtx1c' or guide(i)='invtx1a' or
                      guide(i)='invtx2') then ChemoConcord=chemoconcord+1;
          end;
      **Break down no guideline/not on inva,invx guideline and not concord;
          if (chemoconcord=0 or rtconcord=0) and guide(1)~='Not evaluated
            for concordance' then do i=1 to 4;
                if conc(i)='No' and (substr(guide(i),1,4)='invx' or
                      guide(i)='invtx1' or guide(i)='invtx1b' or
```

```
                  guide(i)='invtx1c') then rtconcord=2;
            if conc(i)='No' and (substr(guide(i),1,4)='inva' or
                  guide(i)='invtx1' or guide(i)='invtx1b' or
                  guide(i)='invtx1c' or guide(i)='invtx1a' or
                  guide(i)='invtx2') then chemoconcord=2;
      end;
      if chemoconcord=0 then chemoconcord=3;
      if rtconcord=0 then rtconcord=3;
**Create groupings for Age variable;
      if diagage<50 then AgeGroup=1;
      else if diagage>=50 and diagage<=70 then AgeGroup=2;
      else AgeGroup=3;
**Collapse stages into I,II,III only;
      if finalstg>=23 and finalstg<=23.2 then Stage=1;
      else if finalstg>=23.5 and finalstg<=25 then Stage=2;
      else if finalstg>=25.5 and finalstg<=27.5 then Stage=3;
**Collapse race into Caucasian, Hispanic, African American or other;
      if racecomp=1 then Race=1;
      else if racecomp in(2,11,8,6,4) then Race=2;
      else if racecomp=3 then Race=3;
      else Race=4;
**Collapse education variable into College, no college, unknown;
      if edustat in(5,6,7) then Education=2;
      else if edustat in(1,2,3,4) then Education=1;
      else Education=3;
**Collapse employment variable into employed, student, other;
      if empstatdx in(1,2,5) then Employment=1;
      else if empstatdx in(3,4) then Employment=2;
      else Employment=3;
**Collapse insurance variable;
      if newinsur=1 then Insurance=1;
      else if newinsur in(5,5.5,5.75) then Insurance=2;
      else if newinsur=4 then Insurance=3;
      else Insurance=4;
**create treatment groups and output to analysis/exclusions data sets;
      if chemflag=1 and rtflag=1 then do;
            TreatGroup=2;
            output analysis;
      end;
      else if rtflag=1 then do;
            TreatGroup=1;
            output analysis;
      end;
      else if chemflag=1 then do;
            TreatGroup=3;
            output analysis;
      end;
      **create exclusions output to check;
      else output exclude;
**create significant variable labels;
      label DiseaseGrp="Disease Group" distmet="Distant Metastasis"
            chemoconcord="Chemotherapy Concordance"
            Stage="Stage at Diagnosis" Education="Education Level";
**format variables
      format diseasegrp tripneg. distmet dist. bmigrp bmi. Rtconcord
            rtchconc. chemoconcord rtchconc. agegroup age. stage stage.
            race rce. education edu. employment emp. insurance insur.
            treatgroup treatgrp.;
```

```
      **only keep variables of interest for analysis
            keep pid distmet diagage p053001 bmi bmigrp agegroup treatgroup
                  diseasegrp stage dsgroup race education employment
                  insurance rtconcord chemoconcord; **linbmi linage lininc ;
      run;
**Run proc freq on categorical variables;
      proc freq data=analysis;
            tables distmet diseasegrp*distmet stage*distmet education*distmet
                  chemoconcord*distmet agegroup*distmet bmigrp*distmet
                  treatgroup*distmet dsgroup*distmet race*distmet
                  employment*distmet insurance*distmet rtconcord*distmet /
                  norow nopercent;
      run;
**Run logistic regression analysis;
**Q: Are triple negs (TN) or treatment group (TG) associated with having
**distant metastasis?;
**also account/control for stage, age at diagnosis, and other socio-economic
**status variables;
      ods graphics on;
      proc logistic data=analysis plots=all;
            class distmet (ref='Local or no met') diseasegrp (ref='LumA')
                  stage(ref='I') education (ref='Col')
                  chemoconcord (ref='Yes') / param=ref;
            model distmet = diseasegrp stage education chemoconcord
                  / lackfit aggregate scale=n rsquare stb iplots clodds=pl;
      label diseasegrp='Disease Group' stage='Stage at
                        Diagnosis'
                  education='Education Level' chemoconcord='Chemotherapy
                  Concordance';
      run;
      ods graphics off;
**logistic regression for creating effects plots;
      ods graphics on;
      proc logistic data=analysis plots(only)=effect(sliceby=chemoconcord
                  connect);
                  *sliceby=diseasegrp sliceby=stage sliceby=education;
                  *sliceby=chemoconcord ;
            class distmet (ref='Local or no met')  diseasegrp  stage
                  education   chemoconcord/ param=ref;
            model distmet = diseasegrp stage education chemoconcord;
      run;
      ods graphics off;
**Genmod procedure for residuals;
      ods graphics on;
      proc genmod data=analysis plots=(stdreschi reschi leverage dobs);
            class distmet (ref='Local or no met')  diseasegrp (ref='LumA')
                  stage (ref='I') education (ref='Col') chemoconcord
                  (ref='Yes') / param=ref;
            model distmet = diseasegrp stage education chemoconcord /
                  link=logit dist=bin;
            output out=resids reschi=pearson stdreschi=stdpearson
                  leverage=lev cookd=infl;
      run;
      ods graphics off;
**Print observations with high residuals
      proc print data=resids;
            var pid distmet pearson stdpearson lev infl;
            where stdpearson>3.3;
```

```
      run;
**Check for multicollinearity of predictors;
      ods graphics on;
      proc corr data=analysis spearman;
            var diseasegrp stage education chemoconcord;
      run;
      ods graphics off;
**Investigate possible reasons patients not on a NCCN guideline are less
**likely to develop a distant met;
      proc freq data=analysis;
            where distmet=1; *where distmet=0;
            tables chemoconcord diseasegrp*chemoconcord stage*chemoconcord
                  education*chemoconcord;
      run;
```