

## **ANALYSIS OF BREAST CANCER AND SURGERY AS TREATMENT OPTIONS**

Beatrice Ugiliweneza, University of Louisville, Louisville KY

### **ABSTRACT**

In this paper, we analyze breast cancer cases from the MEPS (Medical Expenditure Panel Survey), the NIS (National Inpatient Sample) and the Thomson Medstat Market Scan® data using SAS and Enterprise Guide 4. First, we find breast cancer cases using ICD9 codes. We are interested in the age distribution, in the total charges of the entire treatment sequence and in the length of stay at the hospital during treatment. Then, we study two major surgery treatments: Mastectomy and Lumpectomy. For each one of them, we analyze the total charges and the length of stay. Then, we compare these two treatments in terms of total charges, length of stay and the association of the choice of treatment with age. Finally, we analyze other treatment options. The objective is to understand the methods used to obtain some useful information about breast cancer and also to explore how SAS and Enterprise Guide 4 can be used to examine specific healthcare problems.

### **INTRODUCTION**

Breast cancer can be defined as a malignant tumor that develops in the breast. There are many types of treatment options including surgery, chemotherapy, radiation therapy, and hormonal therapy. The medical team and the patient analyze the options together and decide the best treatment. However, surgery is so far defined as the best therapeutic option for breast cancer. There are mainly two types of surgical options: mastectomy and lumpectomy. Lumpectomy is a breast conserving procedure that consists of removing the tumor and a margin of tissue surrounding it to make sure all breast cancer has been removed. Mastectomy is the surgery that removes the whole breast. Many studies have shown that there is no statistical difference between the outcomes of these two treatments. In this paper, we compare them in terms of length of stay and total charges. Then, we look at the association of the choice of treatment with age. Also, we analyze other procedures operated on breast cancer treatment who received mastectomy or lumpectomy. The data used are the MedStat and a complete MEPS dataset by the NIS dataset. We analyze the data by summary statistics, frequencies, graphs, kernel density estimation and tests of association.

### **METHOD**

For the analysis in this paper, we use SAS software and the SAS Enterprise Guide 4.1. Enterprise Guide 4 is a point and click interface which uses SAS in-built functions. It helps the user to explore the power of SAS without writing the codes. We divide our study in three parts: Analysis of breast cancer, analysis of mastectomy and lumpectomy as breast cancer treatments and other frequent treatments used for breast cancer.

### **MEPS AND NIS**

In this project, we use two sets of data. The first set is a composition of two datasets incomplete in themselves: the National Inpatient Survey (NIS) and the Medical Expenditure Panel Survey (MEPS). The NIS data is used to complete the MEPS data using predictive modeling. Data Mining can be defined as the process to extract the implicit, previously unknown, and potentially useful information from data. Predictive modeling is the process by which a statistical model is created or chosen to find the best predictor of an outcome.

First, mastectomy and lumpectomy are studied using a suitable model in the NIS dataset, and then the result is scored to the MEPS data. The resulting dataset is then analyzed with more traditional statistical tools. Once the MEPS data are scored, we can examine differences in follow up treatments when comparing the two procedures. We first filter both datasets MEPS and NIS for breast cancer cases only. The MEPS data are downloaded from the MEPS website: [www.meps.ahrq.org](http://www.meps.ahrq.org). For these data, we choose four sets of files: inpatient, outpatient, physician visits, and medications. To execute the entire task of preparing this set of data for this research, we used SAS Enterprise Guide 4.1. The NIS data are very detailed on the kind of procedure, but are not complete for the follow up because they do

not contain any links to patients across observations. The MEPS data are not precise on the procedure because of the HIPAA de-identification of the information. We use these two incomplete datasets to get one complete dataset that would satisfy our requirements for this study and further research while still respecting the privacy policy.

Here, first, we work with the NIS data. We extract the surgical cases among others using the procedure codes. The code for lumpectomy is 85.21. The procedure codes for the different types of mastectomies are 85.41, 85.43, 85.44, 85.34, 85.33, 85.36, 85.47, 85.48, 85.23, 85.45, and 85.46. Then, we create a code variable with 1= mastectomy and 2= lumpectomy. These two are merged into one sorted table. In this new dataset, we consider both the total charges reported and the length of Stay (LOS) as the variables to predict procedure. In order to have an idea of the distribution of these two variables, we use kernel density estimation, Proc KDE in SAS. The SAS code that we use for our two variables, Length Of Stay (LOS) and Total Charges is

```
data meps3.kde_mastectomy_lumpectomy;
set meps3.mastectomy_lumpectomy;
proc kde data=meps2.mastectomy_lumpectomy gridl=0 gridu=10 method=SNR
out=kdeLOS;
    var LOS;
    by codation;
run;
proc kde data=meps3.mastectomy_lumpectomy gridl=0 gridu=131172 method=SNR
out=kdeTotal_charges;
    var Total_charges;
    by codation;
run;
```

Kernel Density Estimation is a way of estimating the probability function of a random variable. After using the kernel density estimation on our data, we use predictive modeling with a logistic regression model. What we obtain is scored to the MEPS data in order to complete the observations.

With each set of files in the MEPS data, we first merge all the files into one table. Then, we extract the cases of breast cancer using the ICD9-CM diagnosis code, 174, a three digit code for breast cancer. ICD9-CM stands for “the International Classification of Disease, 9<sup>th</sup> division, Clinical Modification”. There are two types of ICD9 codes: diagnosis and procedure codes. These codes and their translations are available online at <http://icd9cm.chrisendres.com/>. Among these cases, we extract those with an ICD9 procedure code of surgery, 85. Then we use information from the NIS to score the surgical procedures and examine the distributions of the resulting datasets.

## THOMSON MEDSTAT

A second set of data used for analysis is the Thomson MedStat Market Scan data. The Market Scan data are healthcare data. They are complete and detailed for analysis. The Market Scan data contain all claims for all individuals enrolled by 100 different insurance providers. For our study, we use the inpatient data files. For the analysis, we use SAS software and the SAS Enterprise Guide 4.

We first extract breast cancer cases among all other inpatient events; we look at the age distribution and then we analyze the stratified data with respect to age. Finally, we study the total charges of treatment and the length of stay at the hospital during treatment.

To extract breast cancer cases, we use also the ICD9-CM diagnosis codes. The diagnosis codes we use are:

- 174.0: Nipple and Areola
- 174.1: Central portion
- 174.2: Upper Inner quadrant
- 174.3: Lower Inner quadrant
- 174.4: Upper Outer quadrant
- 174.5: Lower Outer quadrant
- 174.6: Axillary tail
- 174.8: Other specified sites of female breast

174.9: Breast unspecified  
199.0: Disseminated (cancer unspecified site)  
199.1: Other (cancer unspecified site)  
233.0: Breast

In these data, these codes are written without the decimal point; they are four digit numbers. They are recorded in fifteen variable columns: dx1 thru dx15. To extract breast cancer cases, we use the following code.

#### Code1: Extract breast cancer cases from the dataset

```
data sasuser.inpatientfiles1;
set sasuser.inpatient_files;
diagnoses=catx(' ', dx1, dx2, dx3, dx3, dx4, dx5, dx6, dx7, dx8,dx9, dx10, dx11,
dx12, dx13, dx14, dx15;
if ((rxmatch('1740',diagnoses)>0) or (rxmatch('1741',diagnoses)>0) or
(rxmatch('1742',diagnoses)>0)
    or (rxmatch('1743',diagnoses)>0) or
(rxmatch('1744',diagnoses)>0) or (rxmatch('1745',diagnoses)>0)
    or (rxmathc('1746',diagnoses)>0) or
(rxmatch('1747',diagnoses)>0) or (rxmatch('1748',diagnoses)>0)
    or (rxmatch('1749',diagnoses)>0) or
(rxmatch('1990',diagnoses)>0) or (rxmatch('1991',diagnoses)>0)
    or (rxmatch('2330',diagnoses)>0)) then code=1;
else code=0;
data sasuser.inpatient_files_breastcancer;
set sasuser.inpatientfiles1;
where code=1;
run;
```

This code results in a dataset containing only breast cancer cases.

Breast cancer is diagnosed in women of different ages, but we want to know if, according to these data, there is a particular age for diagnosis. For this, we take the breast cancer data set and we try to estimate the distribution of the variable age using the kernel density estimation (kde). The kernel density estimation is a non-parametric way of estimating the probability density function of a variable that is usually continuous. The kernel density estimation can be used in SAS though the proc kde procedure. We use proc kde on our data, then we plot the density to see what it looks like.

#### Code2: Estimate the density function

```
proc kde data=M2008.breastcancer out=book.bcagedensity;
var age;
run;
```

#### Code3: Plot the density function

```
proc gplot data=book.bcagedensity;
title1 'The density distribution of age';
title2 'in breast cancer data';
plot density*age;
symbol color=green i=spline w=3 v=none;
run;
```

Another way to study breast cancer with respect to age is to analyze the stratified data. We next stratify the age into 6 groups:

Group1: From 0 to 30 years old  
Group2: From 31 to 40 years old  
Group3: From 41 to 50 years old  
Group4: From 51 to 60 years old  
Group5: From 61 to 70 years old  
Group6: From 71 to 80 years old

In the MedStat data, the age is recorded in a variable called age. To stratify, we use the following code.

#### Code4: Stratify the age

```
data book.stratifiedbreastcancer;
set book.mastlump;
if (age lt 30) then group=1;
else if ( 31 lt age lt 40) then group=2;
else if ( 41 lt age lt 50) then group=3;
else if ( 51 lt age lt 60) then group=4;
else if ( 61 lt age lt 70) then group=5;
else if ( 71 lt age lt 80) then group=6;
run;
```

With these stratified data, we produce a summary table that gives us the exact number of patients in each group, a bar chart that shows us the frequency of each group in the data and a pie chart that gives us the percentage of each group in the data. The summary table, the bar chart and the pie chart are produced with the in-built commands in Enterprise Guide 4. It is important to note that the data we have available cover only one year. So, the results we will get here may not represent the total charges for the whole treatment sequence, but the treatment for just one year.

We look at the data regardless of the specific treatment. Later in the paper, we will analyze the specific treatments. In our data, the total charges are registered in the variable, TOTPAY, which stands for total pay and represents the cost of treatment in the whole year. We use the summary statistics built into Enterprise Guide 4 to get the description of the TOTPAY variable. In the data, the Length of Stay is recorded in a variable named DAYS. Again here, we use the summary statistics from Enterprise Guide4.

In this part, we analyze lumpectomy and mastectomy individually in terms of total charges and length of stay. Then, we compare these two treatments in terms of total charges and length of stay. Finally, we test if there is any association between the choice of treatment and age. In the MarketScan data, the procedures are recorded in the variables PROC1 thru PROC15. We use these variables to extract Mastectomy and Lumpectomy datasets. We use the following ICD9 procedure codes:

#### Lumpectomy

85.21: Local excision of lesion of breast (Lumpectomy)

#### Mastectomy

85.23: subtotal mastectomy  
85.33: Unilateral subcutaneous mammectomy with synchronous implant  
85.34: Other unilateral subcutaneous mammectomy  
85.36: Other bilateral subcutaneous mammectomy  
85.41: Unilateral simple mastectomy  
85.43: Unilateral extended simple mastectomy  
85.44: Bilateral extended simple mastectomy  
85.45: Unilateral radical mastectomy  
85.46: Bilateral radical mastectomy  
85.47: Unilateral extended radical mastectomy  
85.48: Bilateral extended radical mastectomy

To extract the two datasets Mastectomy and Lumpectomy, we use the following SAS code

#### Code5: Extract Mastectomy and Lumpectomy datasets among breast cancer cases

```
libname Book "F:\Book";
libname M2008 "F:\M2008";
/*Extract the mastectomy and the lumpectomy cases*/
data book.mastlump;
set M2008.BREASTCANCER;
procedures= catx(' ', proc1, proc2, proc3, proc4, proc5, proc6, proc7,proc8,
proc9, proc10, proc11, proc12, proc13, proc14, proc15);
```

```

if ((rxmatch('8541', procedures)>0) or (rxmatch('8543', procedures)>0) or
(rxmatch('8544', procedures)>0)
    or (rxmatch('8534', procedures)>0) or (rxmatch('8533', procedures)>0) or
(rxmatch('8536', procedures)>0)
    or (rxmatch('8547', procedures)>0) or (rxmatch('8548', procedures)>0) or
(rxmatch('8545', procedures)>0)
    or (rxmatch('8546', procedures)>0) or (rxmatch('8523', procedures)>0) or
(rxmatch ('8521', procedures)>0))
    then code=1;
else if (rxmatch ('8521', procedures)>0) then code=2;
else code=0;
run;
data Book.mastectomy;
set book.mastlump;
where code=1;
run;
data Book.lumpectomy;
set book.mastlump;
where code=2;
run;

```

With the datasets obtained, we are interested in the total charges and length of stay for each patient. After the individual analysis, we compare mastectomy and lumpectomy use. First, we compare the two treatments with respect to total pay. For this, we first estimate the density of this variable in both datasets using the kernel density estimation in SAS. We use the following proc kde procedure code.

#### Code6: Compute the densities of total pay in both datasets

```

/*densities of total pay*/
data mast1;
set book.mastectomy;
keep totpay;
run;
proc kde data=mast1 out=book.mast_totdensity;
var totpay;
run;
data lump1;
set book.lumpectomy;
keep totpay;
run;
proc kde data=lump1 out=book.lump_totdensity;
var totpay;
run;

```

After computing the two densities, we merge the two resulting tables and we plot them on the same graph.

#### Code7: Merge and plot the densities for total pay

```

/*Combine the two tables with the two densities*/
data book.mastlump_densities;
set book.mast_totdensity (rename=(totpay=mast_totpay density=mast_totdensity
count=mast_totcount));
set book.lump_totdensity (rename=(totpay=lump_totpay density=lump_totdensity
count=lump_totcount));
merge book.mast_totdensity book.lump_totdensity;
run;
/*Plot the two total pay densities in the same coordinate system*/
title1 'Comparison of the total pay';
title2 'in Mastectomy and Lumpectomy';
proc gplot data=book.mastlump_densities;
symbol1 color=red i=spline w=3 v=none;
symbol2 color=blue i=spline w=3 v=none;
/*red=mastectomy blue=lumpectomy*/

```

```

plot mast_totdensity*mast_totpay lump_totdensity*lump_totpay/overlay haxis=0 to
30000 by 5000;;
label mast_density='density'
      mast_totpay='total pay'
      lump_density='density'
      lump_totpay='total pay';
run;

```

Second, we compare mastectomy and lumpectomy in terms of days spent at the hospital. We use the same procedures as in the comparison of the charges.

**Code8: Compute the densities of days in both datasets**

```

/*densities of days of stay*/
data mast2;
set book.mastectomy;
keep days;
run;
proc kde data=mast2 out=book.mast_daysdensity;
var days;
run;
data lump2;
set book.lumpectomy;
keep days;
run;
proc kde data=lump2 out=book.lump_daysdensity;
var days;
run;

```

**Code9: Merge and plot the densities for length of stay**

```

/*Combine the two tables with the two densities*/
data book.mastlump_densities;
set book.mast_daysdensity (rename=(days=mast_days density=mast_daysdensity
count=mast_dayscount));
set book.lump_daysdensity (rename=(days=lump_days density=lump_daysdensity
count=lump_dayscount));
merge book.mast_daysdensity book.lump_daysdensity;
run;
/*Plot the two days of stay densities in the same coordinate system*/
title1 'Comparison of the days of stay';
title2 'in Mastectomy and Lumpectomy';
proc gplot data=book.mastlump_densities;
symbol1 color=red i=spline w=3 v=none;
symbol2 color=blue i=spline w=3 v=none;
/*red=mastectomy blue=lumpectomy*/
plot mast_daysdensity*mast_days lump_daysdensity*lump_days/overlay
haxis=0 to 10 by 2;

label mast_density='density'
      mast_totpay='total pay'
      lump_density='density'
      lump_totpay='total pay';
run;

```

Finally, we want to know if there is any association between the choice of treatment and the patient's age. For this, we first produce a table with R rows and C columns with groups of age as row entries and procedures as column entries using the FREQ procedure in Enterprise Guide 4. Then, we use the R\*C contingency table procedure.

**Code10: Test of association between procedure and age**

```

data book.count;
input code group count;

```

```

cards;
1 1 29
1 2 318
1 3 1383
1 4 2424
1 5 942
1 6 2
2 1 3
2 2 23
2 3 152
2 4 120
2 5 47
2 6 0
;
run;
proc freq data=book.count;
table code*group/chisq expected nopercnt fisher;
weight count;
run;

```

First, we look at the 5 most frequent treatments per procedure. For this, we use the summary tables in Enterprise Guide 4 to produce the list of procedure codes ordered in descending order for each of the 15 procedure variables (PROC1-PROC15). Then, we gather the first 5 appearing procedure codes for each procedure in descending order. From the table obtained; we record all the procedure codes present in this table and we order them in descending order of their frequency.

## RESULTS

### MEPS AND NIS DATA

The NIS data contain various surgical treatment procedures for breast cancer. After filtering the cases of mastectomy and lumpectomy, the number of observations is considerably reduced. The analysis was performed on 315 observations for the variable, LOS (Length Of Stay) and 301 observations for the Total Charges. Table 1 gives the summary statistics.

**Table 1: Summary of NIS data.**

	Mastectomy		Lumpectomy	
	Length Of Stay	Total Charges	Length Of Stay	Total Charges
Number Of Observations	289	277	26	24
Mean	2.45	19,564	1.23	11,912
Variance	7.89	2.57E8	0.42	7.04E7
Standard deviation	2.81	16038	0.65	8391

The Kernel Density Estimation helps visualize the density function and test for normality. PROC KDE for Length of Stay is a way of examining the procedures in detail.

**Figure 1: Kernel Density Estimation for LOS for Mastectomy and Lumpectomy in the NIS data**

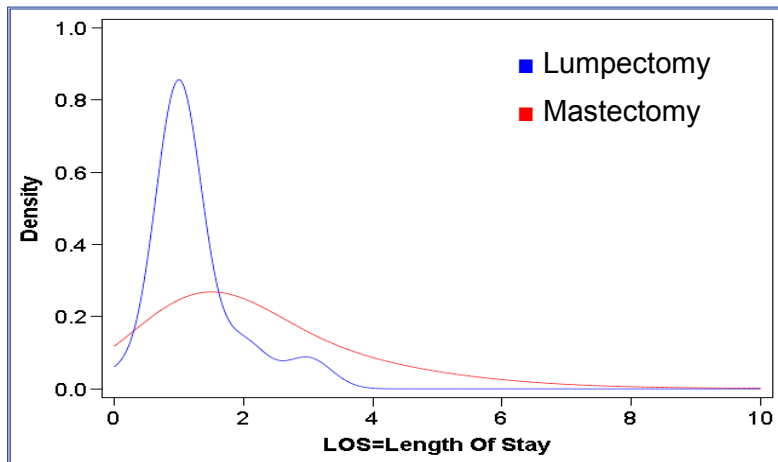
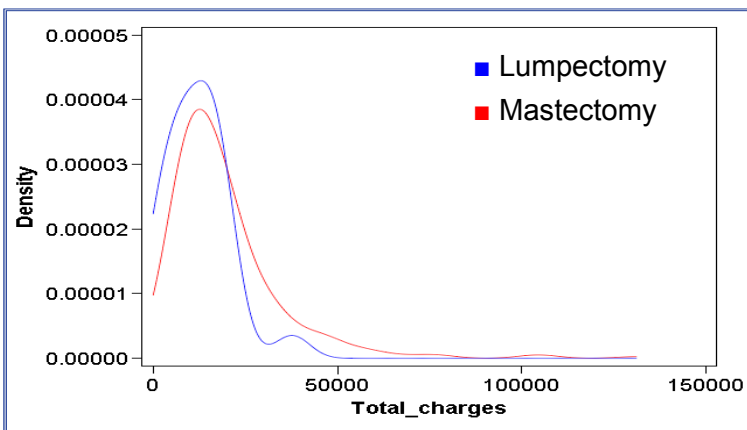


Figure 1 shows that the LOS is normally distributed for both mastectomy and lumpectomy. This is very important because many statistical tests require data to be normally distributed. This graph shows that the patients having a mastectomy have a higher probability of staying longer than those having a lumpectomy. Figure 2 gives the kernel density for Total Charges.

**Figure 2: Kernel Density Estimation for Total Charges for Mastectomy and Lumpectomy in the NIS data**



The total charges variable is also normally distributed for both mastectomy and lumpectomy. This facilitates the research because all statistical tests can be performed on these data. This graph points out that the total cost of mastectomy has a higher probability of a higher cost compared to the cost of lumpectomy.

The MEPS data are not precise on different treatments, especially on surgical treatments of breast cancer. In order to get a complete data set, the previous results were scored to this data set. The different data sets (inpatient, outpatient and physician visit) obtained after conversion to time series were merged together and then attached to the data set of mastectomy and lumpectomy from NIS. In order to do this, the variables, Total charges, Patient ID, LOS (Length of Stay), and Procedures were extracted from both datasets with the procedure value left blank for the MEPS data. Before merging, we created a new variable in each table called number. To define the variable, number, we let 1=mastectomy\_lumpectomy, 2=inpatient, 3=outpatient, 4=physician visit. We merged the tables with respect to this variable number. Logistic regression, as a predictive modeling procedure, is applied to the result. The basic logistic regression model is performed by the PROC GENMOD. We apply the logistic regression to the result using SAS Enterprise Guide 4. The code used is:



```

Input Data: SASUSER.APPEND_TABLE_0011
Server: Local
----- */
PROC SQL;
%_SASTASK_DROPDS(SASUSER.PREDLogRegPredictionsAPPEND_TABL);
%_SASTASK_DROPDS(WORK.SORTTempTableSorted);
%_SASTASK_DROPDS(WORK.TMP1TempTableForPlots);
QUIT;
/* -----
Data set SASUSER.APPEND_TABLE_0011 does not need to be sorted.
----- */
PROC SQL;
CREATE VIEW WORK.SORTTempTableSorted
AS SELECT * FROM SASUSER.APPEND_TABLE_0011;
QUIT;
TITLE;
TITLE1 "Logistic Regression Results";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL) on
%SYSFUNC(DATE(), EURDFDE9.) at %SYSFUNC(TIME(), TIMEAMP8.)";
PROC LOGISTIC DATA=WORK.SORTTempTableSorted
;
MODEL procedures1= /
SELECTION=NONE
LINK=LOGIT
;
OUTPUT OUT=SASUSER.PREDLogRegPredictionsAPPEND_TABL(LABEL="Logistic regression
predictions and statistics for SASUSER.APPEND_TABLE_0011")
PREDPROBS=INDIVIDUAL;
RUN;
QUIT;
TITLE;
TITLE1 "Regression Analysis Predictions";
PROC PRINT NOOBS DATA=SASUSER.PREDLogRegPredictionsAPPEND_TABL
;
RUN;
/* -----
End of task code.
----- */
RUN; QUIT;
PROC SQL;
%_SASTASK_DROPDS(WORK.SORTTempTableSorted);
%_SASTASK_DROPDS(WORK.TMP1TempTableForPlots);
QUIT;

```

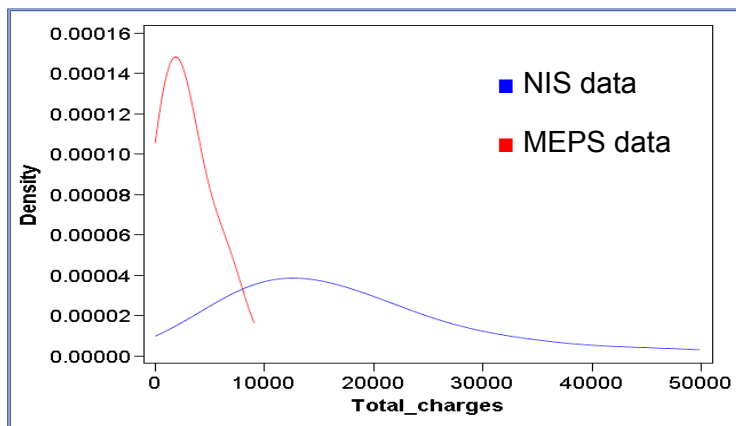
By doing this, we use the model of NIS procedures to score the MEPS procedures. After this step, we separated the MEPS data from the NIS data. This is one of the first steps to preprocess the MEPS data for further analysis. The summary statistics of the MEPS data are given in Table 2. **Table 2. Summary of MEPS Data**

	Total Charges	
	Inpatient	Physician visit
Number of observations	5	185
Mean	2773	271
Variance	5.54E6	822,601

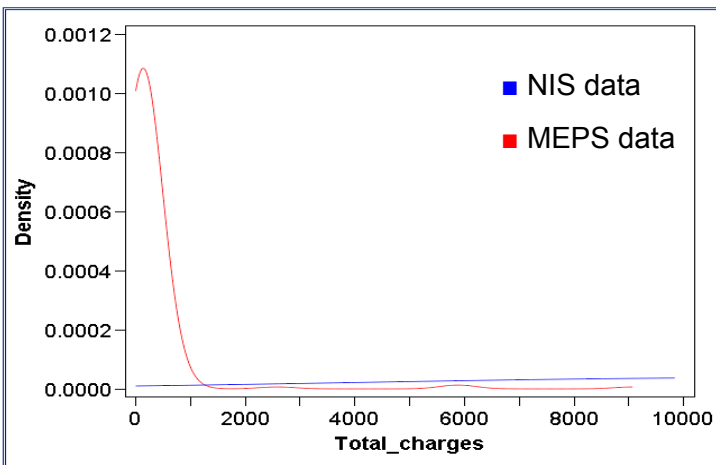
	Total Charges	
	Inpatient	Physician visit
Standard deviation	2353	907

The outpatient number of observations is too small to give a significant output. The LOS has an average of one day for both inpatient and physician visits. We applied Kernel Density Estimation to the total charges of each data set, inpatient and physician visits. Figure 3 compares the MEPS to NIS for total charges in the inpatient data set; Figure 4 compares it in the physician visit data set.

**Figure 3: Kernel Density Estimation for Total Charges for Mastectomy in MEPS inpatient data set compared to NIS dataset**



**Figure 4: Kernel Density Estimation for Total Charges for Mastectomy in MEPS physician visits data set compared to NIS dataset**



Figures 3 and 4 show that the resulting Total Charges for Mastectomy in the MEPS data are skewed and normally distributed compared to the Mastectomy in NIS, which is fairly normally distributed. For this reason, after merging the physician visit data and the inpatient data, minor changes are needed for this variable before proceeding in the analysis.

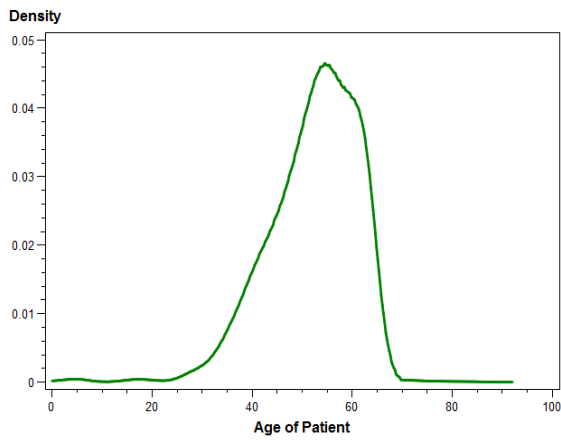
From the two incomplete NIS and MEPS datasets, we are able to construct a complete MEPS dataset. The diagnosis codes in the MEPS are now complete and we can differentiate mastectomy from lumpectomy. The dataset is ready to be used for longitudinal analysis. In the treatment of breast cancer, according to the analysis of these data, the chance of having a mastectomy is significantly higher. The cost of this treatment is high, too, but the length of stay is similar for each procedure.

### THOMSON MEDSTAT DATA

From our breast cancer data, we get the following age distribution:

**Figure 5: The distribution of age in breast cancer data**

*The density distribution of age  
in breast cancer data*



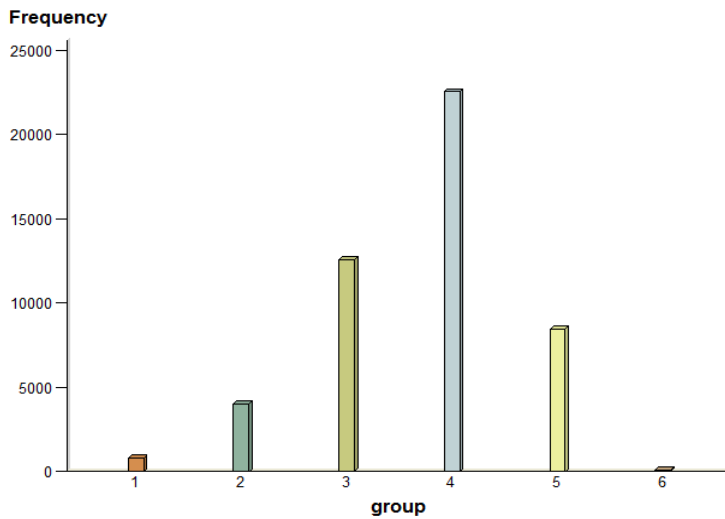
From this graph, we see that the probability of getting breast cancer is close to zero from 0 to almost 30 years old; the probability gets higher as the age increases to a peak at 50-55 years of age. Then, the probability gets closer to zero from 70 years old and older. Looking at this graph, we have a clear idea of which age is more at risk of breast cancer. However, we want to see deeper into age classes. We look at the stratified data.

**Table 3: Summary table of the stratified breast cancer data**

	N
<b>group</b>	
<b>1</b>	785
<b>2</b>	4032
<b>3</b>	12570
<b>4</b>	22536
<b>5</b>	8445
<b>6</b>	87

We can see that the group 4 (51 to 60 years old) has the largest number of patients, followed by group 3 (41 to 50 years old). This alone is enough to confirm the results we obtained with the age distribution. Another way to look at it is to analyze the bar chart.

**Figure 6: Bar chart of the stratified breast cancer data**



Here, group 4 has the highest frequency followed by group 3. This is a visual way to look at what the summary table described. This shows us the association of breast cancer with age. However, we must note that there are many other factors, not considered here, that elevate the probability of getting breast cancer. We next look at the patient outcomes of charges. Table 4 gives the basic summary information.

**Table 4: Summary statistics of total pay and length of stay in breast cancer data**

Analysis variables	Mean	Std dev	Min	Max	N
<b>TOTPAY (Total payments)</b>	16733.30	28085.28	0	626058.00	60394
<b>DAYS (Length of Stay)</b>	5.43	8.65	1	367	6312

The charges can go as high as \$626,058, but the mean is \$16,733. This gives an idea of how much the treatment can cost. Concerning the time spent at the hospital, it can be up to one whole year, but on average, it is about five days and at least 1 day. Here, we looked at the breast cancer data without taking into consideration any particular treatment. In the following part, we are going to study two particular surgery treatments: Mastectomy and Lumpectomy. Below, we summarize the data concerning the total pay and the length of stay for lumpectomy and for mastectomy.

**Table 5: Summary statistics for total pay and length of stay in lumpectomy**

Analysis variables	Mean	Std dev	Min	Max	N
<b>TOTPAY (Total payments)</b>	10219.76	8178.27	0	28275.00	471
<b>DAYS (Length of Stay)</b>	5.43	8.65	1	367	172

The maximum charge is \$28,275 and the mean is \$10,219. The maximum number of days a patient can stay at the hospital is 11 and the average number of days is 1.

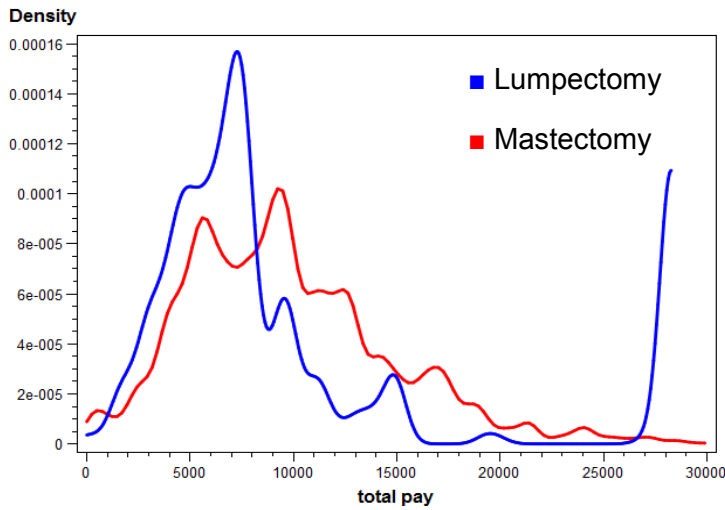
**Table 6: Summary statistics for total pay and length of stay in mastectomy**

Analysis variables	Mean	Std dev	Min	Max	N
<b>TOTPAY (Total payments)</b>	11319.61	11052.25	0	97242.00	6568
<b>DAYS (Length of Stay)</b>	2.2166772	1.6941708	1	25	1583

The maximum charge is \$97,242 and the average is \$11,319. The maximum number of days is 25 and the average is 1. Next, we compare these two surgical treatments in terms of total pay and length of stay. First, we compare their respective cost.

**Figure 7: Compared densities for total pay**

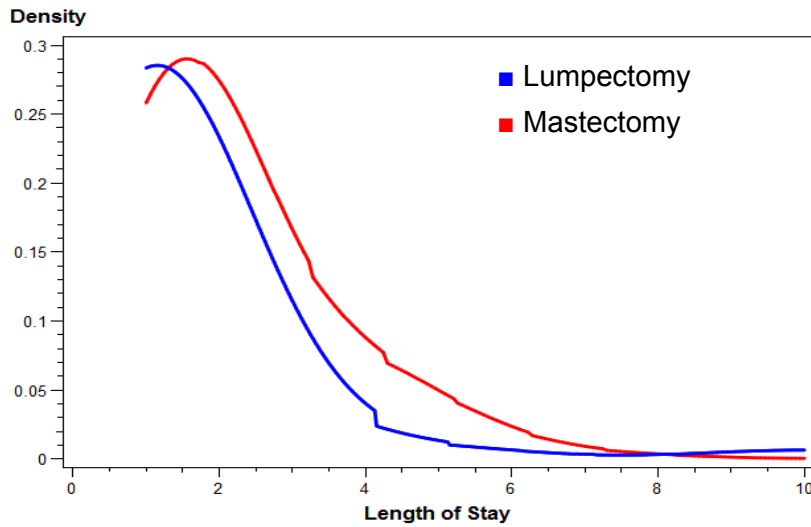
*Comparison of the total pay  
in Mastectomy and Lumpectomy*



This graph shows us that in both treatments, the probability of paying around \$10,000 is high, but it is higher for lumpectomy than mastectomy. Also, we can see that there is higher probability of a higher cost for lumpectomy. Second, we compare their respective length of stay:

Figure 8: Compared densities for days

*Comparison of the days of stay  
in Mastectomy and Lumpectomy*



There is higher probability of a longer stay when treated by mastectomy. These two treatments are the most used for breast cancer. Since breast cancer is associated with age, we test for an association between mastectomy, lumpectomy and age. In other words, we want to know if age is a factor in determining which treatment to use. We use the age distribution by procedure.

Figure 9: The distribution of age in the lumpectomy data

*The density distribution of age  
in lumpectomy data*

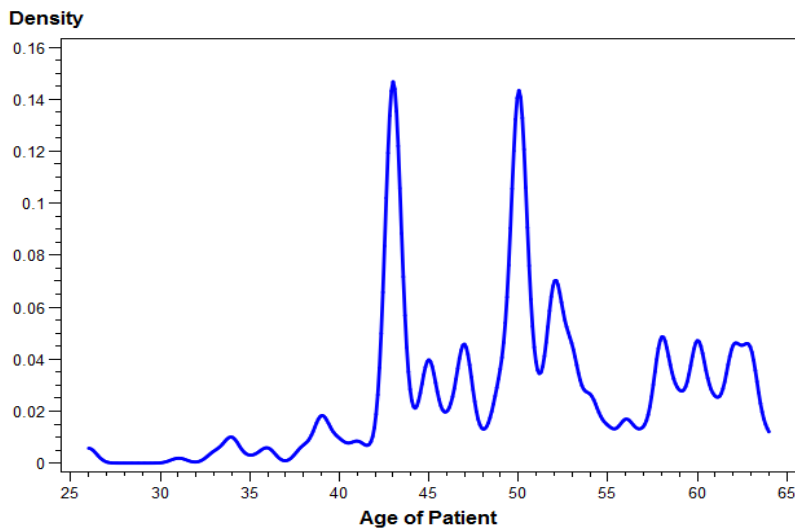


Figure 10: The distribution of in the mastectomy data

*The density distribution of age  
in mastectomy data*

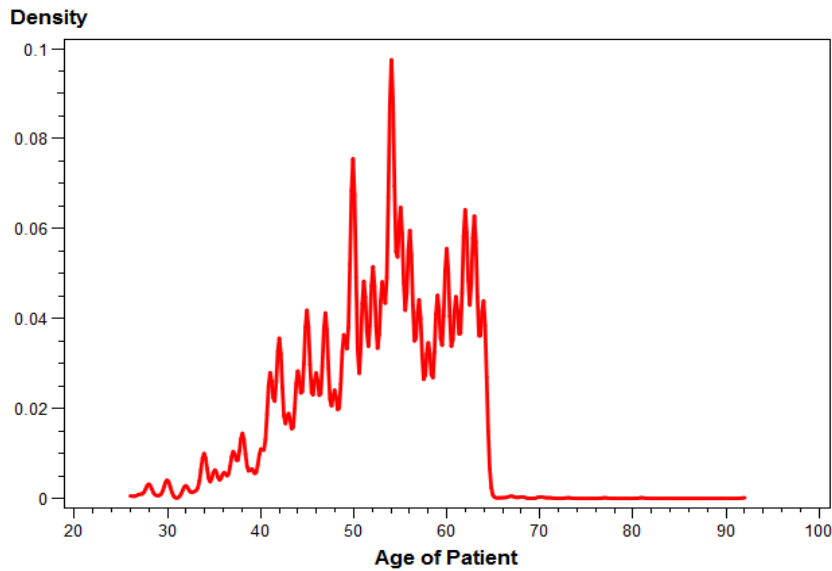


Table 7: Statistics for the test of the association of age and procedures (mastectomy and lumpectomy)

⊕  
*Statistics for Table of code by group* □

Statistic	DF	Value	Prob
Chi-Square	5	49.0250	<.0001
Likelihood Ratio Chi-Square	5	45.9789	<.0001
Mantel-Haenszel Chi-Square	1	25.3967	<.0001
Phi Coefficient		0.0949	
Contingency Coefficient		0.0945	
Cramer's V		0.0949	
<b>WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.</b>			

Fisher's Exact Test	
Table Probability (P)	7.136E-15
Pr <= P	4.480E-09

*Sample Size = 5443*

The results are significant, which means that the choice of treatment does depend on age. However, we must note that there are other stronger factors that affect the choice of treatment. These two surgical treatments (Mastectomy and Lumpectomy) are the most frequent major procedures. Beside these procedures, there are other treatments for breast cancer. We will examine these other treatments.

**Table 8: Most frequent procedure codes in descending order**

99.23	Injection of steroid
99.25	Injection or infusion of chemotherapeutic substance
99.22	Injection of other anti-injective
71	Operations on vulva and perineum
88	Other related diagnosis radiology and related techniques
93	Physical therapy, respiratory therapy, rehabilitation, and related procedures
72	Forceps, vacuum and breech delivery
70	Operations on vagina and cul-de-sac
34.91	Theracentesis
33.27	Closed endoscopic biopsy of lung
01.59	Other excision or destruction of lesion or tissue of tissue of brain

This table shows the most frequent treatments, in descending order, received by breast cancer patients on top of the mastectomy or lumpectomy.

## CONCLUSION

This research shows that data mining can be used to complete one dataset using another one that also has incomplete information. The MEPS dataset, which is incomplete on the procedures because of the HIPAA de-identification, is completed by the NIS dataset using predictive modeling and scoring. We found the variable, Total charges, is normally distributed and the LOS (Length Of Stay) is mostly one day. All this helped us to do the first preparation of the MEPS data.

Breast cancer is a terrible disease for women. According the website, [www.home.thomsonthehealthcare.com](http://www.home.thomsonthehealthcare.com), more than 25% of cancer cases diagnosed are breast cancer. It is important for women to take care of themselves and report any changes in their breasts. But also, it is important to be informed about the realities surrounding breast cancer. We attempted to answer some questions: If I have breast cancer, what are my treatment options? How much is the treatment going to cost me? How long will I be staying at the treatment center? In answering these questions, this paper also shows how to get important information about a particular case from a very large dataset using a statistical tool such as SAS and Enterprise Guide 4. However, the disease of breast cancer is large and many questions and opportunities are not explored in this chapter; for example, there are many other risk factors of breast cancer and there are many things that should be considered to determine the optimal treatment. A treatment sequence can contain more than one treatment... These are interesting challenges to consider in a further analysis.

## REFERENCES

1. Fisher, Bernard; Anderson, Stewart; Redmond, Carol K.; et al (1995). Reanalysis and Results after 12 years of Follow up in a Randomized Clinical Trial Comparing Total Mastectomy with Lumpectomy with or without



Irradiation in the Treatment of Breast Cancer. The New England Journal of Medicine, volume 333(22), pp 1456-1461.

2. Fisher, Bernard; Anderson, Stewart; Bryant, John; et al (2002). Twenty-Year Follow-Up of a Randomized Trial Comparing Total Mastectomy, Lumpectomy, and Lumpectomy plus Irradiation for the Treatment of Invasive Breast Cancer. The New England Journal of Medicine, volume 347(16), pp 1233-1241.
3. Obedian, Edward; Fischer, Diana B.; Haffty, Bruce G. (2000). Second malignancies After Treatment of Early-Stage Breast Cancer: Lumpectomy and Radiation Therapy Versus Mastectomy. Journal of Clinical Oncology, volume 18(12), pp 2406-2412

## **ACKNOWLEDGMENTS**

My acknowledgments go to Dr Patricia Cerrito who helped me and continue to help and support me.

## **CONTACT INFORMATION**

Your comments and questions are valued and encouraged. Contact the author at:

Beatrice Ugiliweneza  
University of Louisville  
328 Natural Science Building  
Louisville KY 40292  
Work Phone: 502 852 6826  
E-mail: b0ugil01@louisville.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.