

Paper 247-2008

## ANALYSIS OF BREAST CANCER USING SAS

Beatrice Ugiliweneza, University of Louisville, Louisville, KY

### ABSTRACT

This research project focused on breast cancer. Two treatment processes were examined using statistical methods, namely, two surgical procedures, lumpectomy and mastectomy. Data were used from the Medical Expenditure Panel Survey (MEPS) and also from the National Inpatient Sample (NIS). These data are incomplete in themselves. The MEPS data are incomplete on procedures and it is not possible to determine whether the procedure was lumpectomy or mastectomy because of the HIPAA required de-identification of information. The NIS data have no information as to follow up treatment after the surgical procedure, but have complete information on procedures. The initial research was to apply statistical methods to these two datasets in order to obtain a complete dataset that can be used for analysis. To examine differences in procedures, therefore, we first gather elements from the NIS data concerning type of surgical procedure and cost, and use predictive modeling on the surgical procedure. Then we use that information to score the procedures in the MEPS dataset. Finally, we analyze the MEPS data statistically using SAS by merging the inpatient dataset to the outpatient and physicians visit datasets. As a result of the study, we show that mastectomy is more likely to be performed in the treatment of breast cancer by surgery.

### INTRODUCTION

This project examined the breast cancer cost and treatment of mastectomy versus lumpectomy. Mastectomy is the surgical procedure in which the entire breast is removed and lumpectomy, also called wide local excision, is a conservative surgery in which only the cancer, along with a border of healthy tissue around it is removed. We used two datasets incomplete in themselves: the National Inpatient Survey (NIS) and the Medical Expenditure Panel Survey (MEPS). The NIS data are 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. It includes the general technique of predictive modeling. 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 were studied using a suitable model, and then the result was scored to the MEPS data that were then analyzed with more traditional statistical tools. Mastectomy is higher in cost and hospital length of stay, and is more likely to be used as a treatment than lumpectomy. Once the MEPS data are scored, we can examine differences in follow up treatments when comparing the two procedures.

### METHOD

The data used were from NIS and MEPS. These data were first filtered for breast cancer cases only. The MEPS data were downloaded from the MEPS website: [www.meps.ahrq.org](http://www.meps.ahrq.org). For these data, four sets of files were chosen: inpatient, outpatient, physician visits, and the medication dataset. To execute the entire task for this research, we used SAS Enterprise Guide 4.1. SAS Enterprise Guide is a point and click interface that provides access to the power of SAS. We wanted to study the treatment procedures of breast cancer, the follow up and different complications.

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 used these two incomplete datasets to get one complete dataset that would satisfy our requirements for further research while still respecting the privacy policy.

First, we worked with the NIS data. We extracted 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 created a code variable with 1= mastectomy and 2= lumpectomy. These two were merged into one sorted table. In this new dataset, we considered 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 used kernel density estimation, Proc KDE in SAS. The SAS code that we used 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. If  $x_1, x_2, \dots, x_N$  are independent and identically distributed random variables, then the kernel density approximation of their probability density function is

$$f_h(x) = (1/Nh) \sum_{i=1, N} K((x-X_i)/h),$$

where  $K$  is some probability density function and  $h$  is the bandwidth (smoothing parameter). Quite often,  $K$  is taken to be a standard normal distribution function with mean zero and variance one. PROC KDE uses

$$K(x) = (1/\sqrt{2}) \exp((-1/2)x^2).$$

After using the kernel density estimation on our data, we used predictive modeling with a logistic regression model. What we obtain was scored to the MEPS data in order to complete the observations.

With each set of files in the MEPS data, we first merged all the files into one table. Then, we extracted the cases of breast cancer using the ICD9 diagnosis code, 174, a three digit code for breast cancer. Code translations are available online at <http://icd9cm.chrisendres.com/>. Among these cases, we extracted those with an ICD9 procedure code of surgery, 85. Then we used information from the NIS to score the surgical procedures and examine the distributions of the resulting datasets.

## RESULTS

The NIS data contain various surgical treatment procedures for breast cancer. After filtering the cases of mastectomy and lumpectomy, the number of observations was 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.

**Table1: Summary of NIS data.**

	Mastectomy		Lumpectomy	
	Length Of Stay	Total Charges	Length Of Stay	Total Charges
Number Of Observations	289	277	26	24

	Mastectomy		Lumpectomy	
	Length Of Stay	Total Charges	Length Of Stay	Total Charges
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**

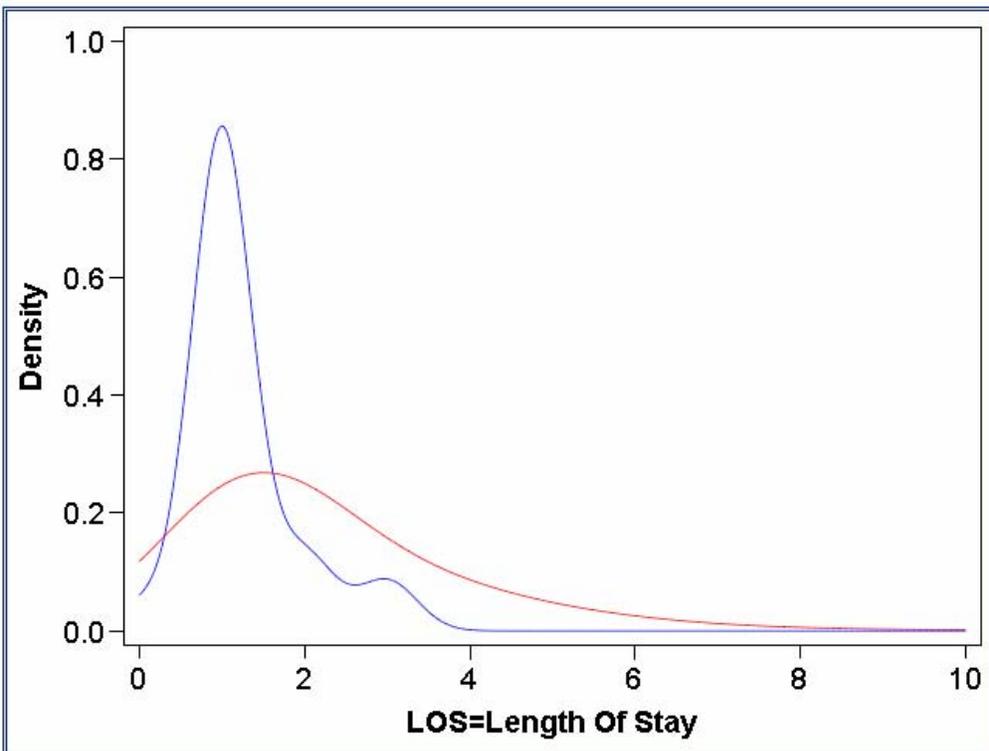
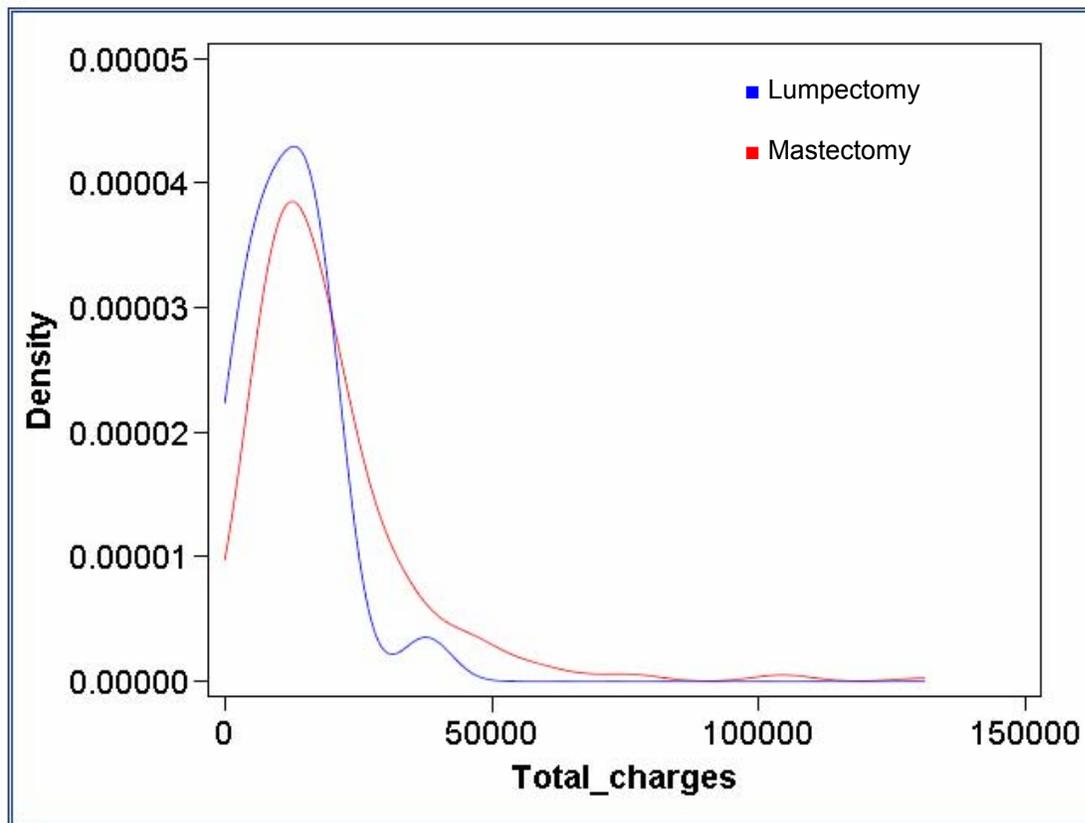


Figure 1 shows that the LOS is normally distributed for both mastectomy and lumpectomy. This is very important because many statistical tests require data to be normally distributed. This graph shows that the patients having a mastectomy stay longer than those having a lumpectomy. Figure 2 gives the 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, 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 by the Enterprise Guide 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.

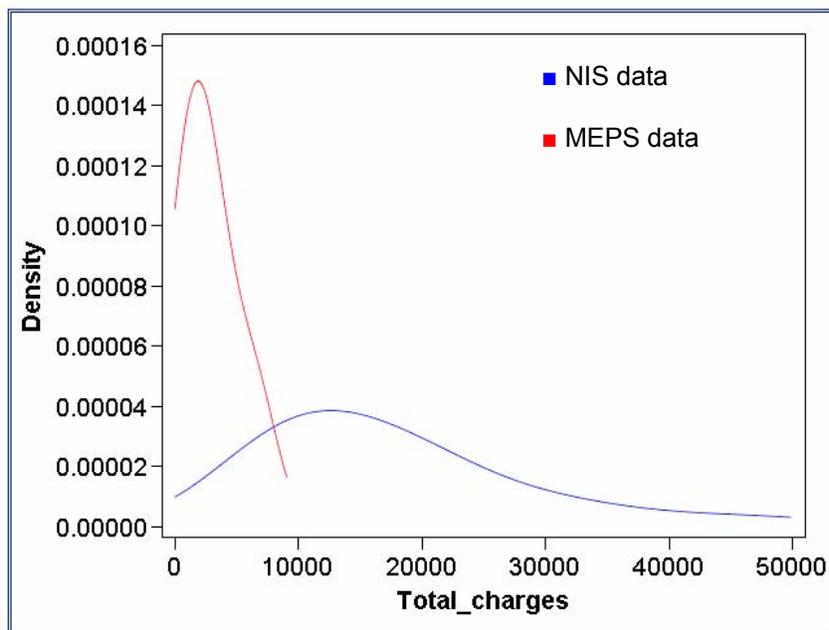
**Table 2. Summary of MEPS Data**

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

**Figure3: Kernel Density Estimation for Total Charges for Mastectomy in MEPS inpatient data set compared to NIS dataset**



**Figure4: Kernel Density Estimation for Total Charges for Mastectomy in MEPS physician visits data set compared to NIS dataset**

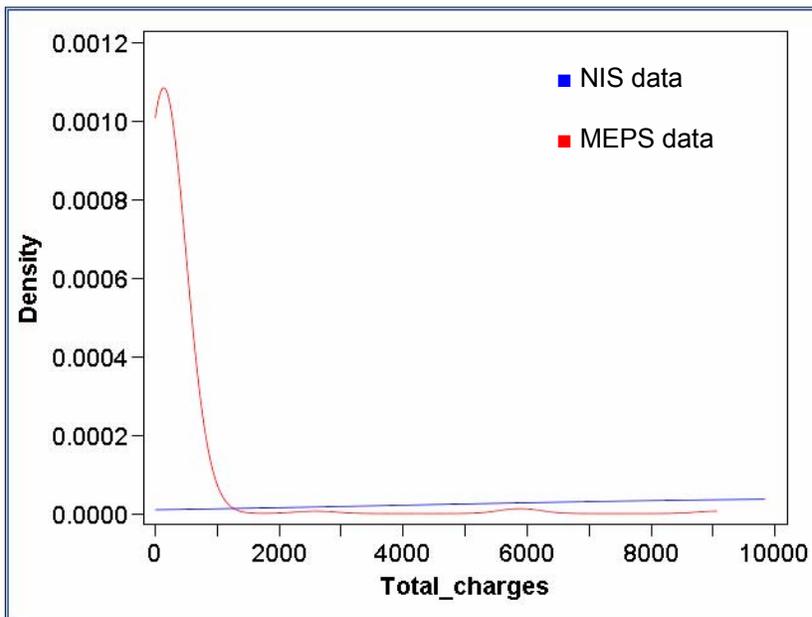


Figure 3 and figure 4 show that the resulting Total Charges for Mastectomy in the MEPS data is skewed and normally distributed compared to the Mastectomy in NIS, which is fairly normally distributed. For this reason, after merging the physician visit data and the inpatient data, minor changes are needed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

## 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. Further analysis will be done with an

ARIMA (Auto regressive Integrated Moving Average) model.

**CONTACT INFORMATION**

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

Beatrice Ugiliweneza  
University of Louisville  
Louisville, KY 40292  
Work Phone: 502 8526022  
E-mail: [b0ugil01@louisville.edu](mailto: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.