Paper 148 -2008

Use of Text Mining to Predict Patient Compliance

Christiana Petrou, University of Louisville, Louisville, KY

ABSTRACT

The purpose of this study is to examine standards of care in the Dental School of the University of Louisville. The central theme is to consider issue of compliance on behalf of the patients and how to define it in an unbiased way. We will examine the relationship of visit intervals, treatment needs, and patient compliance. With the use of SAS 9.1.3 software, data mining techniques such as clustering, kernel density, linear models and mixed models estimation will be used to define and analyze compliance. Statistical methods such as text mining were used to examine the severity of patient conditions. Confidence interval estimation and bootstrapping were explored to assist with the allocation of patients to compliance levels. Patients who were within the 95% confidence interval of the median for visit intervals at least 80% of the time were defined as fully compliant, with decreasing levels of compliance as the percentage decreases to 60%, 40%, 20% and 0%.

INTRODUCTION

The purpose of this study is to define standards of care in the dental school setting at the University of Louisville. Patients on a maintenance regimen are seen by a student on a 6 – 12 month interval depending upon their periodontal status. The needs of patients involved in active treatment vary considerably, and the interval between appointments is typically 1-3 weeks. An issue of concern to health care providers is patient non-compliance. To investigate the issue of non-compliance, the timing of patient visits will be an important factor.

This study focuses on the University of Louisville School of Dentistry (ULSD) in Louisville, Kentucky. ULSD is situated in downtown Louisville and is a force of attraction for many people who need immediate attention or do not have any dental insurance but want to have their teeth examined. The study examines the standards of care in the dental school setting to gain a better understanding on the issue of patient compliance. In the field of behavioral and medical sciences, one of the most important variables that needs to be defined is patient compliance. The dental field of orthodontics is an area of health care that is ideal for the study of patient compliance because the rate and degree of treatment success largely depends upon patient cooperation. As noted at the Kentucky Dental Access Summit, as well as claimed by Robertson and Maduxx, data are scarce and studies on the rate of noncompliance are few. In the case of the medical field, compliance is basically adherence to a drug regimen such as taking medications correctly and on time. It captures the patient's active participation in his or her own healthcare, including seeking medical advice, keeping appointments, following recommendations concerning lifestyle, as well as following medical regimens.

In the dental field, compliance is composed of the interplay between several factors or behaviors. For example, with regards to orthodontic treatment, as the number of changes increases in an individual's lifestyle demands because of treatment, the probability of compliance decreases. In dentistry, compliance varies by treatment; that is, whether the patient is treated for a life threatening condition or merely to relieve pain temporarily, the compliance level differs. It is not unusual for patients who begin treatment as emergency patients to discontinue treatment after their pain has been addressed, even if treatment has not been completed. For example, in the case of any dental school setting, emergency patients generally come to the facility because they do not have a dentist. Once their immediate pain is relieved, they most likely choose not to return to complete the treatment. In an educational institution, provider ability also comes into play because students vary in their ability to complete procedures during a given clinic session. Patient education, loyalty to the student, economics and motivation for better oral health are all factors in patient compliance, because dentistry is costly and it evokes fear in many patients.

The measure of compliance will be defined on the data set provided by ULSD. The data set contains information such as demographics, treatment code (ICD-9 code), date of visit, as well as physician and clinic. The database contains well over 30,000 patient visits. With the use of SAS 9.1.3 software, data mining techniques will be used to analyze and define compliance.

The University of Louisville Dental Clinic advertises the following services in addition to offering emergency care: cosmetic bleaching, on site lab, crowns and bridges, periodontal services, dentures and partials,

relines, implant dentistry and repairs. Therefore, ULSD has a rich database containing information on a wide variety of treatments and care options. The data are entered by treatment and identification codes; hence, the data from one visit and treatment must be linked to all other visits and treatments by patient identifier code. The definition for measure of compliance will be defined directly from the data. The patient habits will indicate their level of compliance. There will be no need for any questionnaires to be answered by patients or dentists. The time interval between visits will be computed from visit dates. Median treatment intervals will be estimated from the data (median is used to avoid the influence of outliers). Then patients who are within the 95% confidence interval of the median at least 80% of the time will be defined as fully compliant, with decreasing levels of compliance as the percentage decreases to 60%, 40%, 20% and 0%. The 80% cutoff for compliance has been used in the past, so it is a reasonable cutoff point.

CLUSTERS AND SEVERITY

In most statistics studies, the central theme is to learn new information from the data. Data are generated daily in many fields such as healthcare and manufacturing, and they are in the form of numbers or text that can be useful in statistical analysis projects. What the statistician is interested in are the patterns, associations and relationships that exist within the data in order to make new statements regarding unknown facts and to provide further information to validate original assumptions. Vast amounts of data are available for exploration. The rate at which data are produced and stored is proportional to the technological advancement of today's society.

The naked eye of a statistician might take a while to detect patterns in a data set, and in a world where time is money, it is not effective to spend days looking at data merely for pattern recognition. The answer to pattern detection is data mining. Generally, data mining (also referred to as knowledge-discovery in databases) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is an analytical tool. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of process. Typical data mining applications use structured information that is carefully prepared. The data may be transformed by a "data preparation" process, or the data may be collected by having a predefined design for mining. Data variables are categorized into quantitative (order numerical) and qualitative(categorical). Ordered numerical data have the property that the scale assignment utilized the "greater than," "equal to," or "less than," quantifiers, whereas categorical data can be nominal with no ordering significance.

While data mining utilized data in numerical format, there exists a counterpart for text data. Text mining considers data of character format. The reality is that data can also be stored in the form of text, and that text can be very complicated for pattern recognition and correlations to be observed. The purpose of Text Mining is to process unstructured text information, to extract meaningful numeric indices from the text, and to make the information contained in the text accessible to the various data mining, statistical and machine learning algorithms. One of the main themes supporting text mining is the transformation of text into numerical data. In effect, the unstructured data set becomes structured. In text mining, the goal is to discover unknown information, something that no one yet knows.

The power of text mining can be extended to nominal data sets. By definition, a nominal data set has no ranking between members. The reason for using text mining in this project was to determine clusters and classify the patients in those clusters based on the conditions that were recorded from individual visits to the dental school. The data set consisted of the ICD-9 codes that are nominal data. For example '87363' refers to 'tooth broken due to trauma'. Each patient has a collection of theses codes that identify their dental condition at unique visits. The data need to be preprocessed in order to perform text mining. The original data set consisted of ICD-9 codes in different rows, which implies that there were several codes listed for the same patient in subsequent rows. In order to put all the ICD-9 codes into one observation per patient, the data set was transposed.

Proc Transpose in SAS code creates one observation per patient by creating a string of variables. The total number of variables created is equal to the maximum number of ICD-9 codes listed for any one patient. Due to the fact that a text sting needs to be created, the next step is to concatenate all of the ICD-9 codes obtained when transposing the original data set. Text mining is conducted using the SAS 5.2® Enterprise Miner (SAS Institute; Cary, NC). What SAS does is to arrange the text data into an unstructured string of words. The different ICD-9 codes are separated by spaces in the text string. Using the Text Miner node, the clustering of the text string of ICD-9 codes is performed through the expectation maximization algorithm. As an output, the clusters of descriptive words are obtained. The preprocessing code used for the above procedure is given below.

```
Proc transpose data=sasuser.charges1 out=sasuser.tran       (drop=_name_
_label_) prefix=icd9_;
var chdiag;
 by pat;
Run;
Data sasuser.concat( keep= pat chdiag );  //concatenate all icd-9
length chdiag $395;                         codes per patient//
set work.tran;
array chconcat (*) icd9_:;
chdiag = left( trim( icd9_1 ));
do i = 2 to dim( chconcat );
chdiag = left(trim(chdiag)) || ' ' || left(trim( chconcat[i] ));
end;
Run;
```

Note that the length of the text string is unknown, but needed in the above code. It is not obvious how many distinct ICD-9 codes need to be concatenated. There are two ways of addressing this issue, either use the maximum possible length defined as 32767 or use Proc SQL to find the minimum possible value of the text string:

```
Proc sql;
select max( length( chdiag )) into :chdiag_len from       sasuser.concat;
quit;
%put chdiag_len=&chdiag_len;
data sasuser.icd9textstrings;
length chdiag $ &chdiag_len;
set sasuser.concat;
Run;
```

When data are analyzed, the following table (Table 1) is given as an output consisting of five distinct clusters. The most important ICD-9 codes identify the clusters.

Table 1. Clusters of ICD-9 codes

#	Descriptive Terms	Freq	Percentage	RMS Std.
1	4200, 52750, 52720	534	11%	0.0084063278
2	87350, 87343, 87353, 87362, 52510, 52890, 80240, 87340, 87342, 87363	2122	44%	0.1390112514
3	14530, 87130, 14520, 16020, 14550, 17000, 16000, 14490, 17010, 92290	215	4%	0.1366941644
4	80225, 80235, 80221, 80238, 80228, 99667, 47822, 68200, 73382, 52463	888	19%	0.1353886064
5	52200, 52100, 52250, 31810, 14600, 14190, 14100, 16190, 14200, 34380	1036	22%	0.1165227904

The five clusters are translated in Table 2.

Table 2. Defined Text Clusters of ICD-9 codes

Cluster Number	Diagnosis
1	Heart infection associated with some other disease, stone in salivary gland, infection of the salivary gland
2	Cuts on face , open wounds on lips, cut on the tissue surrounding the teeth, toothache, sores of the mouth, inflammation of the mouth, fracture of cheekbones and upper jaw bones, cuts on forehead, broken tooth.
3	Malignant cancer of soft and hard part of mouth, malignant cancer of sinus, roof of the mouth, bones of skull, face, lower jaw bone and area of jaw closest to teeth, malignant cancer of nasal cavities, under tongue, avulsion of the eye, contusion of body.
4	Closed fracture of angle jaw, lower jaw by the ear and upper jaw, open fracture of lower jaw and upper jaw, infection due to implant, infection of the throat, infection of the face, nonunion of fracture, joint disorder, articular disc disorder.
5	Tooth ache, necrosis of the tooth, inflammation of the tooth, malignant cancer of throat, tongue, base of the tongue, salivary glands and larynx, carcinomas, cerebral palsy, mental retardation.

It seem as if cluster 1 describes conditions that involve infection of some sort, while cluster 2 describes conditions that deal with sores and cuts that may or may not penetrate the skin (open or closed). Cluster 3 is dominated by conditions that involve cancer of the maxillofacial region (excluding tongue and throat), cluster 4 describes fractures and infection of oral and maxillofacial region, and cluster 5 deals with cancer of the tongue, throat and salivary glands along with tooth complications.

It makes sense that a person who had an accident and broke a tooth or noticed any abnormalities involving the mouth would want to consult a dentist. But why would a person that suffers from cancer of the sinus or infection of the throat see a dentist? It seems that, for example, a person suffering from nasal or sinus cancer will have as one of the very early symptoms, pain that can be confused as tooth ache. Other patients are advised by their physicians to get a dental checkup prior to treatment for cancer. This is important because some cancer treatments, especially ones involving the maxillofacial part of the human body, can make dental problems worse and increase the risk of mouth infection.

The next step is to place a ranking on the clusters. The way the ranking was conducted was to have a private dentist review the cluster definitions and help identify the ordering of the severity. The table of the severity ranking of clusters is given in Table 3.

Table 3. Ranking of Clusters

Cluster ID	Severity Level
1	5
2	4
3	1
4	3
5	2

Severity ranking was assigned based on how complicated a dental procedure is with respect to the disease, whether treatment for a condition is lengthy and has the potential for long term effects. Cluster 1 is ranked as the least severe of all clusters because infections can be treated by antibiotics. Cluster 3 is ranked as the second least severe cluster because cuts on the face or on the head can be addressed by stitching; fractures of cheek bones or jaw bones require a maxillofacial surgeon to operate and fix the fracture while broken teeth can be fixed with implants. Cluster 3 is ranked as the most severe cluster since it describes conditions dealing with potentially deadly cancer diseases. The cancer conditions described involve areas that are hard to detect and require complicated means of treatment. Cluster 4 is ranked as the 3rd most severe cluster because fractures can be complicated to address in terms of dental work and become even more difficult to address if a fracture has not healed properly. But no condition described in this cluster is deadly, unlike clusters 3 and 5. Cluster 5 is ranked as the second most severe cluster. The main difference between the two most severe clusters in terms of the cancer conditions included in each is that for cluster 5,

the treatments for the cancer diseases might not affect the teeth directly, whereas for cluster 3, for example, a surgery to remove part of the jaw requires dental work to reconstruct the jaw, whereas removing the tongue in the extreme case for tongue cancer requires a dentist in the sense of taking care of dental conditions prior to surgery.

CONFIDENCE INTERVALS

Confidence interval theory will be the central component in the definitions for patient compliance. One of the main themes in statistics is to make general statements about a population based on information that is extracted from a sample. We want to make inferences about parameters. Parameters are characteristics of the population. They are constants that appear in the probability function of a distribution. For example, the parameters of a normal distribution are μ, the population mean and σ, the population variance.
The measure of compliance will be defined on the data set provided by ULSD. The data set contains information such as demographics, treatment codes (ICD-9 codes), date of visit, as well as physician and clinic. The database contains well over 30,000 patient visits. With the use of SAS 9.1.3 software, data mining techniques will be used to analyze and define compliance.

The data are entered by treatment and identification codes; hence, the data from one visit and treatment must be linked to all other visits and treatments by patient. The definition for measure of compliance will be directly from the data. The patient habits and performance will indicate their level of compliance. There will be no need for any questionnaires to be answered by patients or dentists. The time interval between visits will be computed from visit dates. Median treatment intervals will be estimated from the data (median is used to avoid the influence of outliers). Then patients who are within the 95% confidence interval of the median at least 80% of the time will be defined as fully compliant, with decreasing levels of compliance as the percentage decreases to 60%, 40%, 20% and 0%. The 80% cutoff for compliance has been used in the past, so it is a reasonable cutoff point.

The dental data contain the visit intervals as well as the median interval visit for each patient. The median was employed to avoid the undue influence of any outliers. An assumption had to be made in regards to visit intervals. Typically, patients on a maintenance regimen go to the dentist every 6 to 12 months. A patient, however, in active treatment can have a visit interval of 1 to 3 weeks. So, a patient exceeding a 12 month visit interval can be considered as an outlier. By employing the use of the median statistic, we eliminate the influence of outliers primarily because the median is a robust statistic. After ordering the median interval visits of individual patients, we note that the 50^{th} percentile for the data set of visit intervals is 14; what we are interested in is the interval for the median in which the parameter exists. When the interval is obtained, then the visit intervals for the patients will be classified in levels of compliance based on how many visit intervals fall within the parameter interval.

The interval estimator is a rule that determines a method for using the sample data to calculate two numbers that result in the endpoints of the interval. The formal name for the interval estimator is confidence interval. The endpoints of the confidence interval will vary depending on the sample taken from the population since they are functions of the selected samples. When computing a confidence interval, certain criteria need to be satisfied. First and foremost, the interval should contain the parameter of interest and secondly, the interval should be as narrow as possible. The aim is to find an interval that is narrow enough but has a high probability of enclosing the parameter. More formally, a 100(1-α) % confidence interval for an unknown population parameter, θ, is an interval between two numbers with an associated probability (1-α) that is generated from a random sample of an underlying population such that if the sampling were repeated numerous times, and the confidence interval recalculated from each sample according to the same method, a proportion (1-α) of the confidence intervals would contain the population parameter in question. There are several methods for constructing a confidence interval; some examples are Pivotal Quantity, Maximum Likelihood Estimator and Bootstrapping.

Bootstrapping is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample. One of the first statisticians to consider bootstrapping was Bradley Efron. The bootstrap was introduced in 1979 as a computer based method for estimating the standard error of θ, where θ is an estimate of a parameter of interest. The main theme of the bootstrap method is to use the data collected for a single experiment to stimulate what the results might be if the experiment was repeated over and over with a new sample. The new experiments are called bootstrap samples and they are obtained from the original dataset through the procedure of resampling. When Efron initiated the bootstrap procedure, the underling goal was to randomly draw a large number of observations, or “re-samples” from the population. Each re-sample will have the same number of elements, and due to the replacement that takes place, some elements may be included one time, more than once, or not at all. Hence, each one of these

samples will vary in the observations included in it. For each re-sample, the statistic $\hat{\theta}_i$ will be calculated, where i refers to the number of the sample. Due to the slight differences from sample to sample, each $\hat{\theta}_i$ will have a slightly different value. There are three widely known bootstrap methods for confidence interval estimation: the percentile method, bootstrapped t-method and finally the BCa method.
The BCa method is the best method for constructing confidence intervals using bootstrap samples compared to the percentile method and the bootstrapped t-method. BCa stands for Bias-corrected and accelerated. Some knowledge regarding bias and jackknife is essential in understanding the method. Bias is said to measure the average accuracy of an estimator. Formally, Bias ($\hat{\theta}$) = E ($\hat{\theta}$)-θ. Ideally, the bias should be as small as possible because the smaller the magnitude of the bias, the better the estimator tends to be. The confidence interval is of the form ($\hat{\theta}^{*(\alpha 1)}$, $\hat{\theta}^{*(\alpha 2)}$), where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1-\hat{\alpha}(\hat{z}_0 + z^{(\alpha)})}\right)$$

and

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1-\hat{\alpha}(\hat{z}_0 + z^{(1-\alpha)})}\right) .$$

Note that Φ is the standard normal distribution function and $z^{(\alpha)}$ is the 100xα^{th} percentile point of the standard normal distribution. To calculate the bias- correction, we first append the bootstrap samples to the original data set and find the proportion of samples that have a sample statistic less than the statistic calculated from the original data set.

For example, in the case of the dental data, the median was 14; we are looking for the samples that have a median less than 14. Hence, $\hat{\theta}$ is the original data statistic, $\hat{\theta}_i^*$ is the statistic calculated from the i^{th} bootstrap sample, and B is the number of bootstrap samples with $\hat{z}_0 = \Phi^{-1}(\#\{\hat{\theta}_i^* \le \hat{\theta}\}/B)$, where $\Phi^{-1}(x)$ indicates the inverse function of a standard normal distribution. This procedure can be done very quickly using SAS in utilizing the following code:

```
data bootsamp;
do sampnum =1 to 10000;
do i = 1 to nobs;
x = round(ranuni(0)*nobs);
set dataset nobs=nobs point = x;
output;
end;
end;
stop;
run;
proc means data=bootsamp noprint nway;
class sampnum;
var median;
output out=bootmedian median=median;
run;
data bootdata;
set dataset (in=a) bootsamp;
if a then sampnum=0;
run;
data proportionless;
set bootmedian;
retain origmedian;
if sampnum=0 then origmedian=median;
if median lt origmedian then lessthan=1;
else lessthan=0;
retain nless 0;
if sampnum gt 0 then nless=nless+lessthan;
```

```
if sampnum ne 0 then output proportionless;
run;
data bias;
set proportionless;
by sampnum;
if last.sampnum then do;
propless=nless/sampnum;
bias=probit(propless);
output bias;
end;
run;
```

The acceleration statistic $\hat{\alpha}$ represents the rate of change of the standard error of the estimate of the sample statistic with respect to the true value of the statistic.

Efron and Tibshirani suggest calculating the acceleration using the jackknife procedure. The jackknife method bears a great deal of similarity with the bootstrap procedure in the notion of re-sampling and is actually what initiated the study on bootstrapping. The goal of the procedure is to estimate the bias and standard error of an estimate by focusing on the samples that leave out one observation at a time. So ultimately, we are gathering information on how much each individual record influences the estimate. The statistic of a jackknife sample is denoted by $\hat{\theta}_{(i),}$ and the mean of n jackknife samples is denoted by $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \frac{\hat{\theta}_{(i)}}{n}$. The acceleration is calculated by the following equation:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6\{(\sum_{i=1}^{n} (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2\}^{\frac{3}{2}}}$$

Once again, the code for SAS programming is as follows:

```
data origjack;
set dataset end=eof;
obsnum=_n_;
if eof then call symput('nobs',put(obsnum,2.));
run;
%macro jackdata;
data jackdata;
set %do i=1 %to 307;
sasuser.origjack (in=in&i where=(obsnum ne &i)) %end;;
%do i=1 %to 307;
if in&i then repeat=&i;
%end;
run;
%mend;
%jackdata;
proc means data=work.jackdata noprint nway;
class repeat obsnum;
var median;
output out=sasuser.jacksum median=median;
run;
proc sql noprint;
select median(median)
into :medianjack from sasuser.jacksum;
quit;
data jacksum1;
set sasuser.jacksum;
```

```
cubed=(&medianjack - median)**3;
squared=(&medianjack - median)**2;
run;
proc means data=jacksum1 noprint;
output out=jacksum2 sum(cubed)=sumcube sum(squared)=sumsquar;
run;
data accel;
set jacksum2;
accel=sumcube/(6*(sumsquar**1.5));
keep accel;
run;
```

Having both the acceleration and the bias correction estimates, we can find the values for α_1 and α_2. The final step is to order the random variable of interest and compute the value $N_1 = \alpha_1 * B$ as well as $N_2 = \alpha_2 * B$. The N_1^{st} order value of the random variable will be the lower endpoint of the confidence interval while the N_2^{nd} order value will be the upper endpoint of the confidence interval. The SAS code to use is

```
data ciends;
part1=(2.17 +probit(0.125))/(1-(-0.000071647*(2.17+probit(0.125))));
part2=(2.17 +probit(0.875))/(1-(-0.000071647*(2.17+probit(0.875))));
alpha1=probnorm(2.17+part1);
alpha2=probnorm(2.17+part2);
n1=alpha1*1000;
n2=alpha2*1000;
call symput ('n1',put(floor(n1),5.));
call symput ('n2',put(floor(n2),5.));
run;
proc sort data=sasuser.bootmed;
by median;
run;
data ci_bca;
set sasuser.bootmed end=eof;
retain conf_lo conf_hi;
if _n_=&n1 then conf_lo=median;
if _n_=&n2 then conf_hi=median;
if eof then output;
keep conf_lo conf_hi;
run;
```

When this procedure was performed for the Dental data, the confidence interval derived was (14, 14). The reason for this outcome is primarily due to the statistic of interest within the dental data. The median is not a “smooth” statistic nor is it a “robust” statistic. That means that small changes in the data can cause considerable change in the statistic. The lack of smoothness causes the jackknife estimate of standard error to be inconsistent for the median.

There is a way to fix this inconsistency. Instead of leaving one observation at a time, we can leave out two or three, or even more observations. The procedure is known as the deleted-d jackknife. The rule of thumb is to leave out d observations at a time, where $d = \sqrt{n}$ and n is the total number of observations in a data set. This rule is applied merely to achieve consistency for the jackknife estimate of standard error. With all the computer capabilities that we have, this can still be quite a challenging task, since the more data values deleted, the more the resulting table of appended jackknife samples grows exponentially. For example, the dental data set consisted of 307 observations. Deleting one observation at a time resulted in a jackknife sample of 306, repeated 307 times, and the resulting merged tables for the jackknife data set had 93,942 records. In the case of deleting two observations, each sample had 305 records, and there were 46,971 possible combinations for deleting 2 values at a time. The result was a data set with 4,398,129,585 records! Once that was done, the confidence interval was still (14, 14). It gets worse with deleting three observations, and it was too much for the computer to handle. A more traditional approach was implemented to derive the confidence interval for the median visit interval: pivotal quantities.

The pivotal quantity method is a very well known approach towards constructing confidence intervals. The main goal is to define a function of the sample and the unknown parameter, where the unknown parameter

is the only one in the function, and the probability distribution function does not depend on the unknown parameter. The logic behind this method is that for a random variable X, the probability $P(a \le X \le b)$ is unaffected by a change of scale or translation on X. Therefore, if we know the probability distribution of a pivotal quantity, we may be able to use operations such as scaling and translation to create an interval estimator.

To construct the confidence interval, we proceed as follows:

1. Let $X_1, X_2, \ldots, X_n$ be a random sample from the population X with probability density function (pdf) $f(x;\theta)$, where θ is an unknown parameter.
2. Consider the pdf to see if a pivot exists and determine the pivotal quantity $Q(X, \theta)$.
3. For a specific value of α, we can find numbers a and b that do not depend on θ, to satisfy $P(a \le Q(X, \theta) \le b) \ge 1- \alpha$.
4. Algebraically manipulate the inequalities to isolate θ in the middle, $P(L \le \theta \le U) = 1- \alpha$ where $L=L(X_1, X_2, \ldots, X_n)$ and $U=(U X_1, X_2, \ldots, X_n)$.

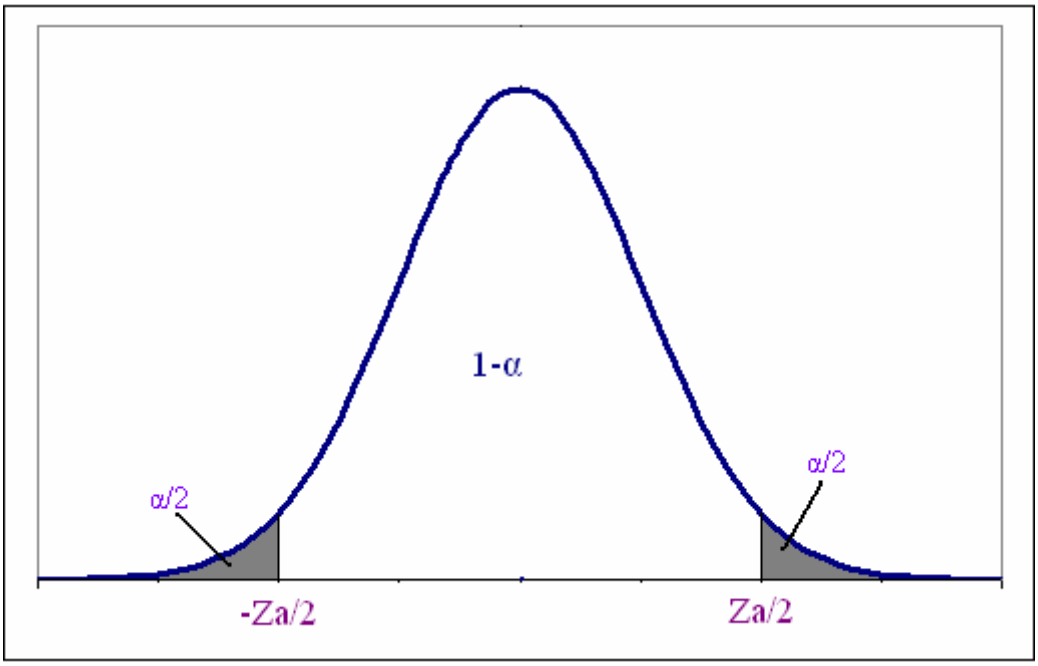
Then the limits of the form (L, U) are a subinterval of the real line $\Re$. Therefore, we conclude that if we take a large number of samples from the underlying population and construct all the corresponding (1- α)% confidence intervals, then approximately (1- α)% of these intervals would include the unknown value of the parameter θ. To illustrate the above theory, let $\hat{\theta}$ be a statistic that is normally distributed with mean θ and standard error $\sigma_{\hat{\theta}}$. Then we want to find a confidence interval for $\hat{\theta}$ that possesses a confidence coefficient equal to 1-α.

The quantity $Z= (\hat{\theta} - \theta)/ \sigma_{\hat{\theta}} \sim N(0, 1)$. We select two values from the tails of the normal distribution, namely $-z_{\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$ such that

$$P(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}) = 1-\alpha.$$

Figure 1 displays the graph of a normal curve, and indicates the proportion of the graph that $-z_{\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$ can be possible estimates.

Figure 1. Two tails with probability α.



Substitute for Z in the probability statement to get

$$P\left(-z_{\frac{\alpha}{2}} \le (\hat{\theta} - \theta)/\sigma_{\hat{\theta}} \le z_{\frac{\alpha}{2}}\right) = 1-\alpha.$$

Multiplying by $\sigma_{\hat{\theta}}$ to get

$$P\left(-z\sigma_{\hat{\theta}}\,{}_{\frac{\alpha}{2}} \le \hat{\theta} - \theta \le z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}}\right) = 1-\alpha.$$

Subtracting $\hat{\theta}$ from both terms of the inequality to obtain

$$P\left(-\hat{\theta} - z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}} \le -\theta \le -\hat{\theta} + z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}}\right) = 1-\alpha.$$

Finally, multiplying each term by -1 and changing the directions of the inequalities, we have

$$P\left(\hat{\theta} - z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}} \le \theta \le \hat{\theta} + z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}}\right) = 1-\alpha.$$

Hence, the confidence endpoints for a (1-α)% confidence interval for θ are

$$L = \hat{\theta} - z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}} \text{ and } U = \hat{\theta} + z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}}.$$

Now this theory can be applied to construct the confidence interval for the median.

For the Dental School data, the population is a total of 320 patients with their corresponding visit intervals and median visit interval. A total of 307 out of 320 patients have a nonzero entry as their median visit. Those

patients were selected as a sample to determine the Maritz-Jarrett estimate of standard error. This procedure was performed using SAS 9.1 and the code used is

```
data sasuser.mj1;
set sasuser.charges7;
%let i=obsnum;
y=cdf('BETA',(&i/307),153,154)-cdf('BETA',((&i-1)/307),153,154);
run;
quit;
data sasuser.mj2;
set sasuser.mj1;
c1=y*median;
c2=y*(median**2);
run;
```

Once the code runs successfully, we get that the standard error (or standard deviation) of the sample median is equal to 1.023878934. Hence, the sampling distribution of the median for this study is approximately normal with mean $\tilde{\mu}$ =14 and variance 1.048328071.

If we standardized $X_{0.5}$, then we get $\frac{X_{\tilde{m}} - \tilde{\mu}}{\sigma_{MJ}}$ ~ N (0, 1). The distribution of the standardized $X_{0.5}$ is independent of $\tilde{\mu}$. The standardized $X_{0.5}$ is the pivotal quantity since it is a function of the population median $\tilde{\mu}$ and its probability distribution is independent of $\tilde{\mu}$. Using the pivotal quantity, we construct the confidence interval the same way we did earlier:

$$1-\alpha = P\left(-z_{\frac{\alpha}{2}} \le \frac{X_{\tilde{m}} - \tilde{\mu}}{\sigma_{MJ}} \le z_{\frac{\alpha}{2}}\right)$$

$$1-\alpha = P\left(X_{\tilde{m}} - \sigma_{MJ} z_{\frac{\alpha}{2}} \le \tilde{\mu} \le X_{\tilde{m}} + \sigma_{MJ} z_{\frac{\alpha}{2}}\right).$$

Hence, the (1-α)% confidence interval for $\tilde{\mu}$ is

$$\left(X_{\tilde{m}} - \sigma_{MJ} z_{\frac{\alpha}{2}},\ X_{\tilde{m}} + \sigma_{MJ} z_{\frac{\alpha}{2}}\right).$$

Therefore, the confidence interval for the sample that consists of all nonzero visit intervals for the dental school turned out to be (11.99, 16.007).

The patients were then classified into the different compliance level groups. Some work was done in Excel in regards to finding out how many visits each patient had that were within the confidence interval endpoints, and from that, the percentage was estimated. Then in SAS 9.1, the following SAS code was used to classify the patients in compliance levels:

```
data sasuser.dependentvar;
set sasuser.compliance;
by compliance;
if compliance lt 20then level=5;
if compliance ge 20 lt 40 then level=4;
if compliance ge 40 lt 60 then level=3;
if compliance ge 60 lt 80 then level=2;
if compliance ge 80 lt 100 then level=1;
run;
```

Hence, a patient that had less than 20% of their visit intervals within the confidence interval was classified in the lowest compliance group. A patient that had more than 80% of their visit intervals in the confidence interval was classified in the compliance level group 1, which means that they are more likely to continue

attending the Dental School and are fully compliant. Table 4 describes the allocation of patients in the different Compliance levels.

Table 4. Frequency Counts for Compliance Levels

Compliance Level	Frequency	Percentage
1	0	0
2	1	0.31
3	14	4.38
4	43	13.44
5	262	81.88

It is evident that the majority of the patients that attend the Dental School are not compliant. There could be several reasons, some of which will be explored in later sections, as to why that is the case.

COMPLIANCE AND PATIENT SEVERITY

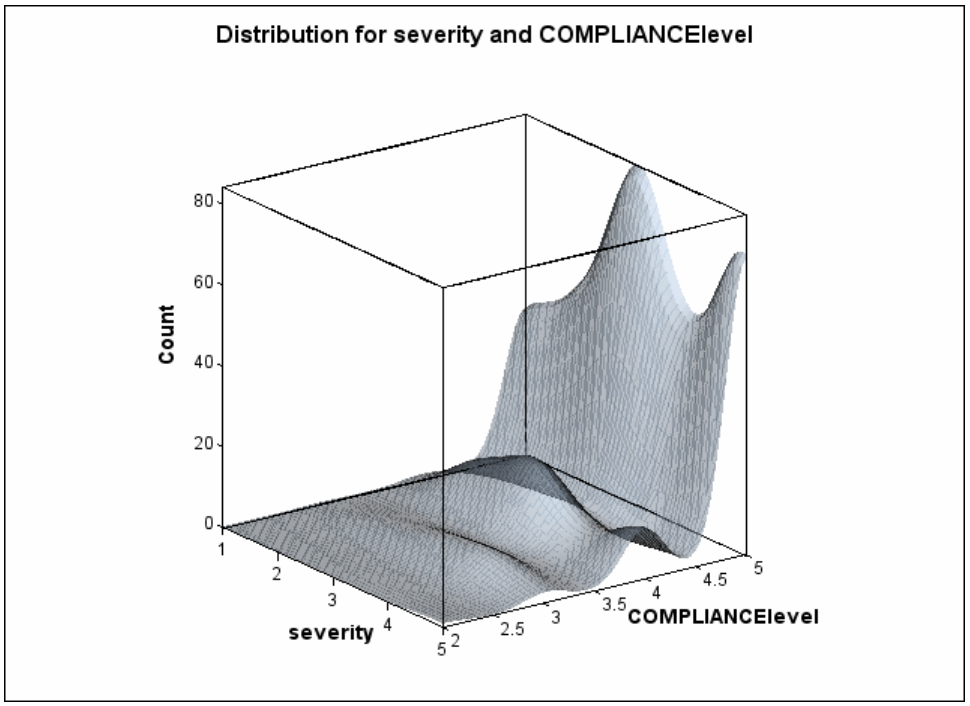
The relationship of compliance to severity will be explored to gain a better understanding of what type of patients attend the Dental school more often. The relationship between the levels of compliance and severity is illustrated in Table 5 through a table analysis.

Table 5. Table analysis of Compliance by Severity Indicating Frequency and Column Percentage

Severity(→) Compliance(↓)	1	2	3	4	5	Total
2	0 0.00	1 1.10	0 0.00	0 0.00	0 0.00	1
3	4 3.96	2 2.20	6 13.64	2 4.00	0 0.00	14
4	17 16.83	13 14.29	3 6.82	10 20.00	0 0.00	43
5	80 79.21	75 82.42	35 79.55	38 76.00	34 100.00	262
Total	101	91	44	50	34	320

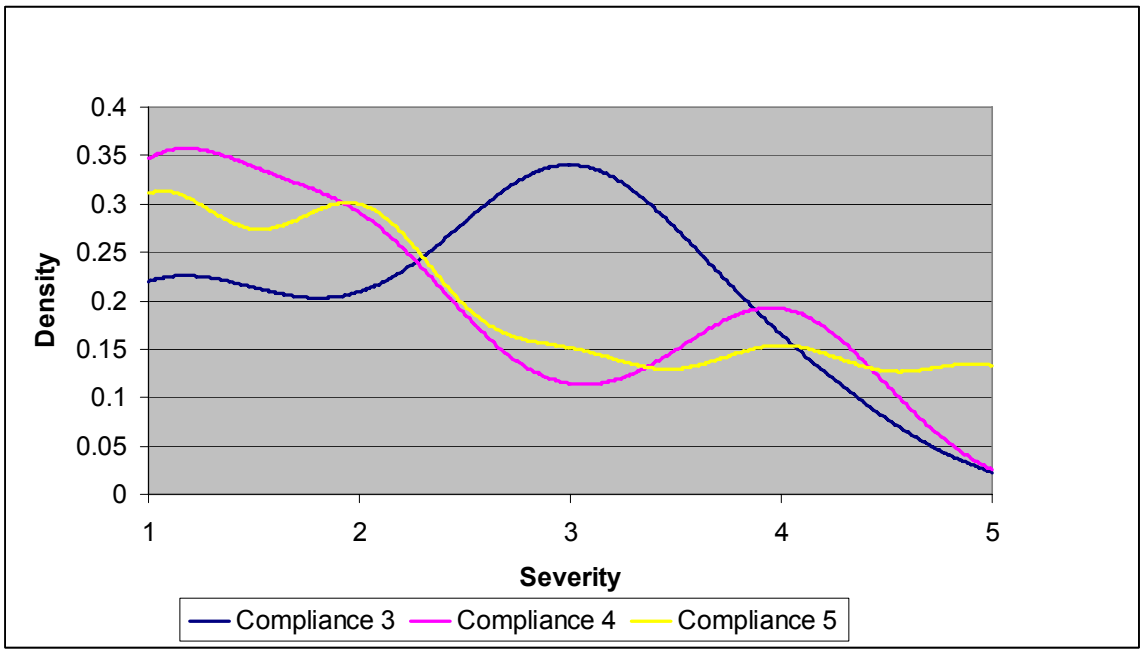
The most obvious relation that emerges from the above table is that all the patients that are listed in the least severe cluster belong in the least compliant level group. In general, there is a pattern of patients being compliant at most 20% of the time in all severity groups. Patients that are classified in the second to last group in terms of severity indicate a higher percentage of compliance than any other severity group; to be specific, they are complaint on an interval at least 20% of the time and at most 40%. Similarly, patients in the 3rd severity group tend to belong in the compliance level group of attending their regular visits at least 40% and at most 60% of the time. For the highest levels of compliance, there is at most 1 patient belonging to both groups, which does not convey enough information. A visual interpretation of Table 5 is given in Figure 2 with the bivariate kernel density estimation of the severity level by compliance level.

Figure 2. Bivariate Kernel Density



Computing the kernel density plot of the compliance levels by the severity levels, Figure 3 is obtained.

Figure 3. Kernel Density Estimation of Severity by Compliance



The different compliance levels are compared based upon the severity ranking of the ICD-9 codes that were recorded at the Dental School. Patients belonging in compliance level 3 have the lowest probability of belonging to any of the two most severe clusters, unlike compliance levels 4 and 5. In fact, patients who are least compliant have the highest probability of suffering from conditions that are the most severe. However, patients in the 3rd level of compliance have the highest probability of belonging to the 3rd cluster in terms of severity. Patients in the 4th compliance level group have the higher probability of being in the second to last severity cluster while patients in compliance level 5 have higher probability of being in the least severe cluster.

CONCLUSION

Bootstrapping is a method that simulates multiple data sets equivalent to the original. Even though in theory it is a nonparametric approach to the confidence interval estimation for the median, in practicality, however it is unattainable unless a good CPU is available. If a CPU is not attainable, then we are handicapped in using a more traditional approach of interval estimation. Once the compliance Confidence Interval was established, it became obvious that approximately 80% of the patients are least compliant. Overall, there is a trend emerging which indicates that more compliant patients are not, in general, suffering from severe conditions, whereas patients that are least compliant, in general, need to undergo treatments for severe conditions. The low rate of compliance on behalf of the patients with severe and complicated treatment needs is an interesting area of exploration on its own.

REFERENCES

1. Bradley Efron; Robert J. Tibshirani,*" An Introduction to the Bootstrap"*, CRC Press LLC, 1998.
2. Scott Robertson; James Maddux, *"Compliance in Pediatric Orthodontic Treatment: Current Research and Issues "*,Children's Health Care, Volume 15 Issue 1, 1986.
3. Nancy Barker, "A Practical Introduction to the Bootstrap Using the SAS System", Oxford Pharmaceutical Science, SAS Conference Proceedings: Phuse 2005.

CONTACT INFORMATION

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

Name: Christiaan Petrou
Enterprise: University of Louisville
Address: 9935 Viuex carre Dr, Louisville, KY, 40223
Work Phone: 502-852-6240
E-mail:cspetr01@louisville.edu