Clinical Trials Versus Health Outcomes Research: SAS/STAT Versus SAS Enterprise Miner

Patricia B. Cerrito, University of Louisville, Louisville, KY

ABSTRACT

Clinical trials typically involve a random selection of subjects to receive an experimental treatment or a control. They often use a minimal sample size, a short term rather than a long term period of study, and clinical trials often use surrogate endpoints. Inclusion/exclusion criteria are used to create a fairly homogeneous set of subjects so that the assumption of normality is relatively valid. Because of the nature of clinical trials, the database is designed to optimize the statistical analysis; most of the techniques used are available in SAS/STAT and there is relatively little preprocessing required prior to the statistical analysis. In contrast, health outcomes typically use data collected in the course of patient care with databases that are not designed for statistical analyses. The data are extremely messy and require considerable preprocessing; They are also observational and require consideration of potential and/or actual confounding factors. Typically, these databases are extremely large, containing thousands to millions of records. The data can be used to investigate real rather than surrogate endpoints in a longitudinal setting as well as to investigate rare occurrences. The subjects are heterogeneous, so the assumption of normality is not reasonable and creates problems when using regression models. In addition, the effect size in any analysis is virtually zero, so the p-value has no real meaning; other measures must be used to determine the adequacy of the model. Many of the statistical tools required to analyze health outcomes data are readily available in SAS Enterprise Miner. We will compare and contrast the differences and similarities between clinical trials and health outcomes research.

INTRODUCTION

Statistical models require an assumption of randomness in order to be valid; observational data violate this assumption. Statistical models have four parameters: type I error, power, sample size, and effect size; large databases used for observational studies generally have an effect size of virtually zero because the sample size is too large. Since the p-value is related to the effect size, an effect size of almost zero will result in p-values that are highly statistically significant. This is particularly true if the sample size is not considered when examining rare occurrences.

In survival analysis, clinical trials tend to be short-term with many censored values. Censored data will always bias the definition of average, and the overall survival rate will be under-estimated. We will demonstrate some of the problems with survival analysis, especially with a heterogeneous population that looks more like a gamma distribution than one that is normal. We will compare statistical models versus predictive modeling when investigating large, observational databases. In addition, we will examine some studies from the medical literature.

BACKGROUND

In this section, we examine some studies from the medical literature and demonstrate some of the problems of using linear models with large samples. Consider a recent study of bypass surgery mortality. (Eifert, Kilian et al. 2010) According to the study,

> " Between 2004 and 2008, 3441 patients (733 women, 2708 men) underwent CABG. 252 women and 854 men were operated using OPCAB, 481 women and 1854 men using extracorporeal circulation (ECC). Medical data was prospectively entered and retrospectively reviewed. 30-days and one year mortality rates were analyzed with Kaplan-Meier estimates and Cox proportional hazards models. Linear and logistic regression models were used to test gender differences."

The p-values in the study are 0.0004 and 0.0008. These p-values suggest an effect size so small that any statistically significant gender difference has no predictive capability at all. In fact, the given effect size is 1.7% versus 2.1% for 30 days. The problem is that the logistic model primarily predicts non-occurrence of mortality. With such a small percentage of deaths, the model will be highly accurate around 98% but with no ability to predict. Logistic regression has never been able to work with such disparate group sizes without some modification to the sample.

Recently, a study that appeared in JAMA used observations from numerous studies to conclude a similar statistical difference with almost identical percentages (2.5% versus 1.7%). (Ranpura, Hapani et al. 2011) A total of 10,217 patients from 16 randomized, controlled trials were included in the analysis for the drug, Avastin. The results were also reported in terms of relative risk, which always tends to make the difference appear larger than it really is. Different medication doses were used across the 16 studies used in the analysis, and different types of cancer, with accompanying different types of chemotherapy, were used in the 16 trials. The inclusion/exclusion criteria were not

taken into consideration across the studies. Usually, however, in clinical trials, only high risk patients are enrolled. Genentech, the manufacturer of the drug stated,

> "...this meta-analysis includes cancer types for which Avastin is not approved by the U.S. Food and Drug Administration (FDA) and should not be used, including advanced squamous* cell non-small cell lung cancer (NSCLC), advanced prostate cancer and advanced pancreatic cancer."

Given this serious flaw in the analysis, it already appears as if it will result in a lower use of the drug in the treatment of cancer:

> " This finding will most likely affect overall usage of Avastin. Because the most significant benefits from Avastin are seen, thus far, in colorectal cancer, it may be that the drug is used with less frequency for treating lung or breast cancers, where the risk may not outweigh the benefits." (Anonymous-Cancer 2011)

There is already considerable pressure as a matter of public policy to reduce the use of Avastin in the treatment of cancer simply as a cost savings measure. (Cerrito and Cerrito 2011) In contrast, a recent study to define a risk index recognized the problems with rare occurrences and used additional information measures as well as the concept of lift to investigate a statin drug. (Raval, Cohen et al. 2010) However, such studies are rare.

METHODOLOGY

Large Samples and Rare Occurrences

A major consideration in the use of logistic is that of rare occurrences. If group A=non-occurrence and group B=rare occurrence, then the logistic model will classify virtually all observations as non-occurrences. In this case, the false negative rate will be almost 100% while the false positive rate will be virtually 0%. While on the surface, the model appears to be a good fit, in practice, it will have almost no predictive ability whatsoever.

Logistic regression requires the two groups to be compared to be almost equal, especially if the sample is large. However, a large sample may be the only way to find a sufficient number of rare occurrences to be able to perform any analyses. The key is to sample a random subset from the non-occurrence group so that the group sizes are close to equal. The result is to make the analysis sample equal to twice the number of rare occurrences. In addition, prior probabilities need to be introduced to choose the optimal model with the prior probabilities giving the true occurrence of both groups. It is also possible to introduce weights to make a false negative more important than a false positive. These changes will make the overall sample of analysis much smaller than expected and to appear to be less accurate, but one that is much more predictive. We will demonstrate through example the impact of using a rare occurrence without compensating through using a subsample of controls.

The Central Limit Theorem and Large Samples

All linear models depend upon the Central Limit Theorem. We want to investigate the consequences of the assumption of normality. Suppose, for example, that we want to determine whether patients with diabetes have a longer length of hospital stay and higher total charges compared to patients without diabetes. We can use different sample sizes from the National Inpatient Sample (NIS) to examine this hypothesis H_0: $\mu_1 = \mu_2$ where group 1=patients with diabetes and group 2=patients without diabetes (available at http://www.ahrq.gov/data/hcup/hcupnis.htm). This database contains a record of all inpatient stays from a stratified sample of 1000 hospitals across 37 states. There are approximately 8 million observations in the database for each year. For purposes of sub-sampling, we can consider this sample to be infinite.

NIS data come only in ascii format, and the SAS and SPSS coding needed to translate the datasets into standard SAS or SPSS format is available on the NIS web site. There is exactly one observation per patient, with no longitudinal linkage between observations. There are three main outcomes available for each patient observation: mortality in the hospital, total charges, and length of stay. There is information concerning patient age, race, and gender. There are fifteen columns to record diagnosis codes and another fifteen columns to record procedure codes. There are thousands of possible diagnoses and procedures. In the NIS, we can resample as needed. More information concerning the data is located at http://www.hcup-us.ahrq.gov/nisoverview.jsp.

With a sample of size 50 for an unpaired t-test, we get the result that the difference is not statistically significant. The confidence interval for the difference in the length of stay is (-1.819, 2.2194) and for total charges is equal to (-16,438, 6299.50). These are quite large and include the value of zero, the null hypothesis. To compute a sample that is stratified proportionally to the occurrence of diabetes to get a sample of size 50, we use the following SAS code:

```
PROC SURVEYSELECT DATA=nis.nis_with_diabetescode;
       OUT=NIS.DIABETESSAMPLE50
       METHOD=SRS
       N=50
       NOPRINT
       ;
       STRATA diabetes;
       ID LOS TOTCHG diabetes; RUN;
```

We can modify the above SAS code for different sample sizes. We use this code to generate a sample of size 200. The t-test remains not significant, but the confidence intervals are considerably smaller at (-1.298, 0.5078) and (-9344, 2581) for length of stay and total charges respectively. At n=1000, the confidence width shrinks even more to (-0.618,0.018) and (-3542,55.64). When n increases to 10,000, the p-values now become highly statistically significant with intervals (-0.579, -0.400) and (-4402, -3453). In other words, the effect size for length of stay is less than 0.15 of a day; the effect size for cost is approximately $500. If the sample size is increased any more, the effect size will be smaller still. It is already so small, that while it has statistical significance, it has no real practical importance. In fact, if we used the complete data sample, the confidence intervals shrink to (-0.443, -0.429) and (-3783, -3702) for a statistically significant difference of $80.

Regression requires the assumption that the residuals are normally distributed. However, most healthcare data are exponential or gamma because of the presence of extreme outliers. The mean of a distribution is highly susceptible to the existence of outliers. Usually, it is better to truncate outliers, to use nonparametric tests based upon the median, or to use a model that accepts a skewed distribution. The assumption of the Central Limit Theorem does not eliminate the problem of outliers.

Linear regression requires moderately large samples to be effective. Power analysis tends to assume that the population distribution is sufficiently homogeneous to be normally distributed. As healthcare outcomes tend to be exponential or gamma distributions because the populations in outcomes research are heterogeneous, we must consider just how large n has to be before the Central Limit Theorem is realistic. (Battioui 2007) To examine the issue, we take samples of different sizes to compute the distribution of the sample mean. The following code will compute 100 mean values from sample sizes starting with 5 and increasing to 10,000. We will graph the distribution of the means and compare it to the distribution of the actual population.

```
PROC SURVEYSELECT DATA=nis.nis_205 OUT=work.samples METHOD=SRS N=5 rep=100
noprint;
RUN;
proc means data=work.samples noprint;
  by replicate;
  var los;
  output out=out mean=mean;
run;
```

Once we have computed the means, we can graph them using kernel density estimation (a smoothed histogram). We show the difference between the distribution of the population, and the distribution of the sample mean for the differing sample sizes. Figures 1-4 show the distribution of the sample mean compared to the distribution of the population for differing sample sizes. To compute the distribution of the sample mean, we collect 100 different samples using the above code. We compute the mean for the patient length of stay using the National Inpatient Sample. The graphs show that the use of the average can be highly misleading when the data are as skewed as represented in a heterogeneous population of patients. Moreover, as the sample size increases, the problem gets worse rather than better.

Figure 1. Sample Mean With Sample=5

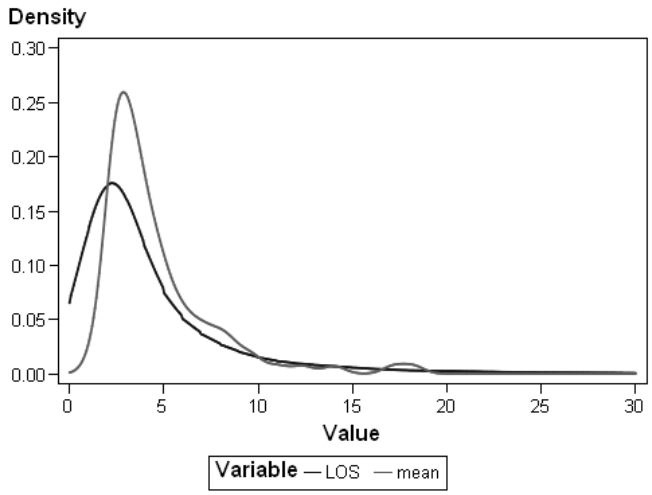


Figure 2. Sample Mean With Sample=30

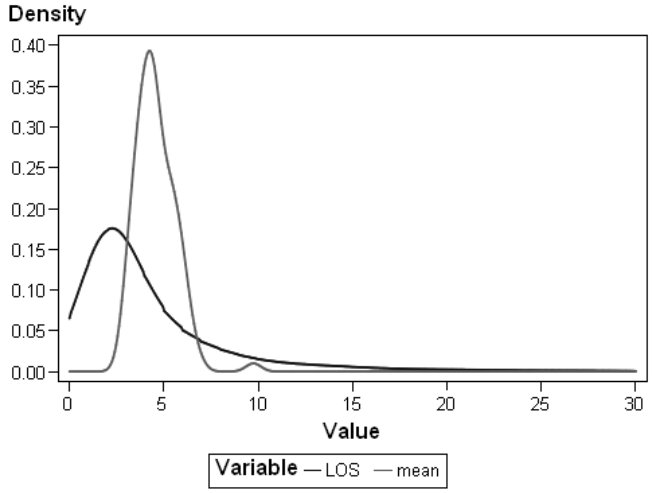


Figure 3. Sample Mean With Sample=100

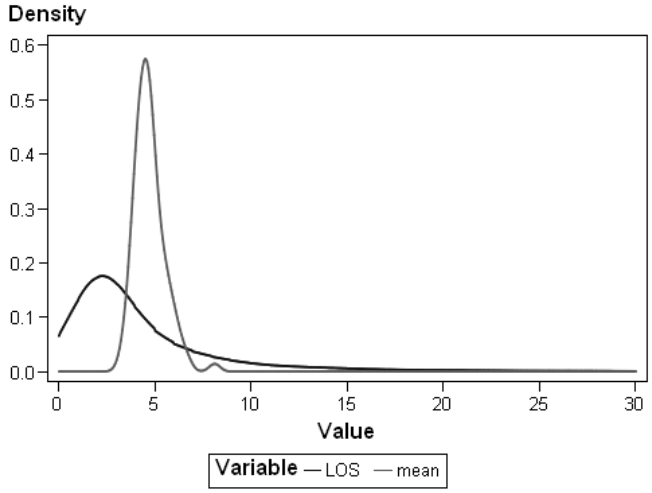
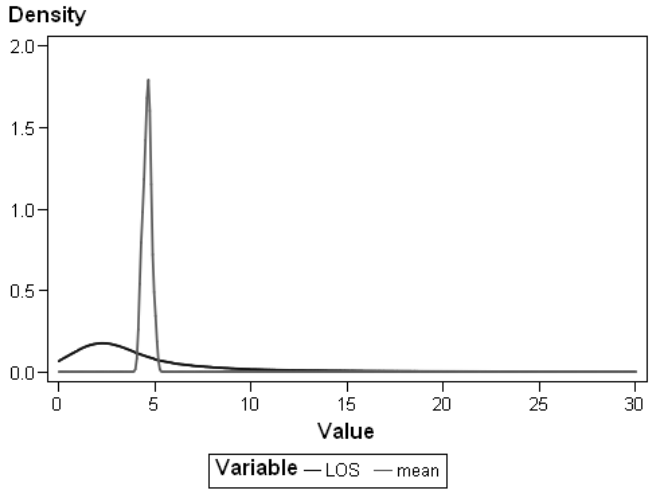


Figure 4. Sample Mean With Sample=1000



In Figure 2, the sample mean peaks slightly to the right of the peak of the population distribution; this peak is much more exaggerated in Figure 3. The reason for this shift in the peak is because the sample mean is susceptible to the influence of outliers, and the population is very skewed. Because it is so skewed, the distribution of the sample mean is not entirely normal. As the sample increases to 100 and then to 1000, this shift from the population peak to the sample peak becomes much more exaggerated. We use the same sample sizes for 1000 replicates (Figures 5-8).

Figure 5. Sample Mean for Sample Size=5 and 1000 Replicates

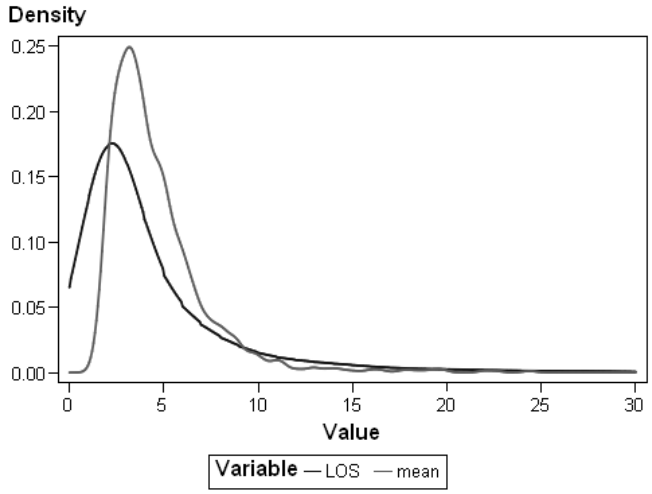


Figure 6. Sample Mean for Sample Size=30 and 1000 Replicates

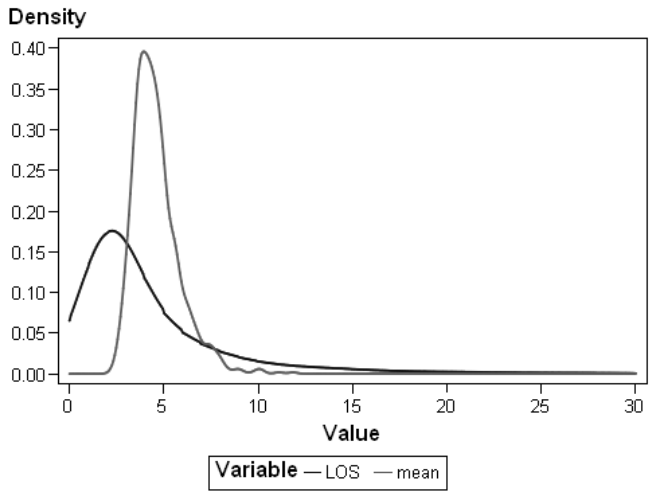


Figure 7. Sample Mean for Sample Size=100 and 1000 Replicates

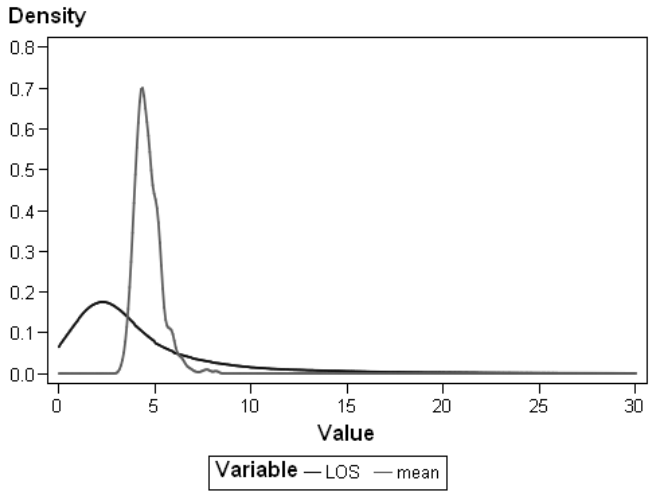
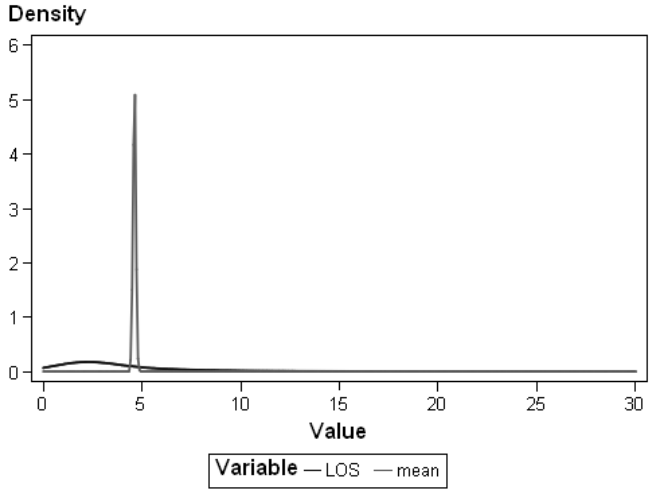


Figure 8. Sample Mean for Sample Size=1000 and 1000 Replicates



It is again noticeable that the sample mean is shifted away from the peak value of the population distribution because of the skewed distribution. However, the distribution of the mean is not normally distributed. In other words, the sample converges on a value that is much greater than the actual population peak.

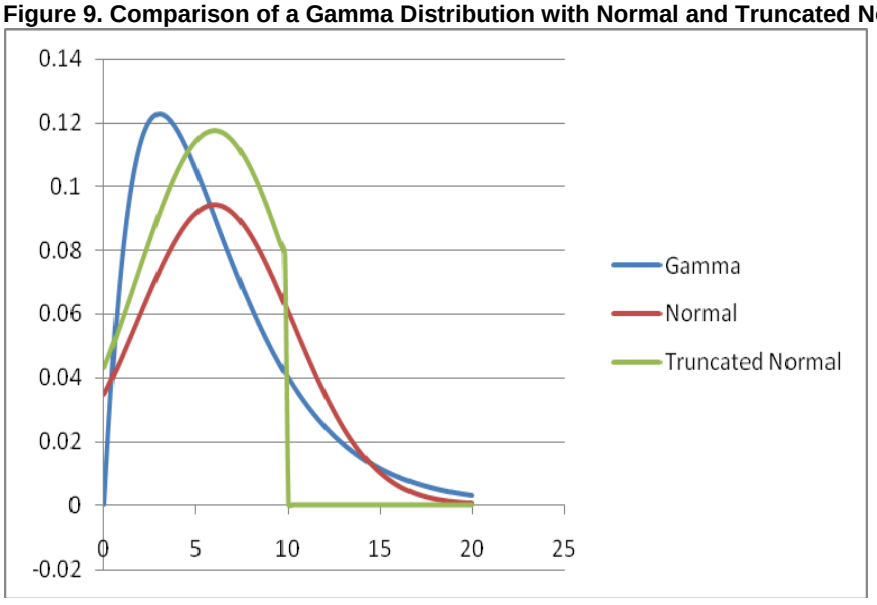
Large Samples and Survival Analysis
Consider a clinical trial involving a cancer drug. It is usually closed after a fixed period of time. There will be patients in the trial who are still alive and who are still without recurrence. Because of the closing of the trial, the survival time of those still alive is truncated. Suppose, for example, that there are 1000 patients enrolled in a trial of drug A. Suppose further that the study closes after 2 years and that 500 of the patients are still alive. Then the 500 patients will be given an average survival of 2 years. The remaining patients will have an average that is less than 2 years, giving an overall average that is also less than 2 years.

Suppose the 500 patients who are not censored have an average survival of 1 year. Then the overall average will be 1.5 years. Without making some assumptions concerning the nature of survival of the censored patients, it is impossible to provide a better estimate. However, then the estimate depends upon the validity of the assumptions used to construct the model. In addition, adjustments because of censoring can be so complex that they are rarely done. (Zhao, Lee et al. 2009) Instead, the average time is not necessarily reported in favor of comparing different groups that have censoring. (Yao, Barlow et al. 2010) Another way is to specify a fixed time point and the proportion still alive in each group at that time point. (Kalager, Haldorsen et al. 2009) Another approach is to consider surrogate endpoints in the place of survival. However, there must be a way to validate the relationship between the surrogate endpoint and actual survival before the surrogate can be used. (Ray, Bae et al. 2009)

Now suppose the average of the 500 censored patients have an average survival of 4 years. Then the actual average survival would be 2.5 years but is computed at 1.5 years. Therefore, the actual survival will be under-estimated. The use of an electronic medical record and the use of longitudinal data will make survival estimates much more accurate since patients can be followed for a much longer time than clinical trials will allow.

Because of this bias in censored data, an alternative approach would be better to use the median, provided that no more than half the subjects are censored; more than half still hav the problem of bias. (Vieitez, Carrasco et al. 2003) Figure 9 shows the problem of truncating outcomes compared to the true population distribution.

Figure 9. Comparison of a Gamma Distribution with Normal and Truncated Normal



Both the normal and the truncated normal have the mean shifted to the right compared to the true, gamma distribution. Moreover, the truncated normal will mis-represent the individuals on the right side of the curve; the overall result will be to under-estimate the potential for survival in an experimental treatment.

Linear Models Versus Predictive Modeling
in contrast, predictive modeling has built-in measures to improve upon the actual predictability of the model. It makes frequent use of the sampling node (Figure 10) to ensure that both false positives and false negatives are considered along with the prior probability and the potential for decision weights.
The sampling node uses all of the observations with the rare occurrence, and then takes a random sample of the remaining data. While the sampling node can use any proportional split, we recommend a 50:50 split. Figure 11 shows how the defaults are modified in the sampling node of SAS Enterprise Miner to make predictions.

Rule induction is a special case of a decision tree model. Figure 10 also shows three different neural network models and two regression models. The second regression model automatically categorizes all interval independent variables. There is one remaining model in Figure 10; the MBR or memory-based reasoning model. It represents nearest neighbor discriminant analysis. We first discuss the use of the sampling node in the process of predictive modeling. We start with the defaults for sampling node as modified in Figure 11. Because there are so many observations in a large sample, it is possible to use multiple models. The data are routinely partitioned into training, validation, and testing sets; the testing set represents a holdout sample that can be used to validate the final results.

The first arrow indicates that the sampling is stratified, and the criterion is level based. The rarest level (in this case, mortality) is sampled so that it will consist of half (50% sample proportion) of the sample to be used in the predictive model.

Figure 10. Addition of Sampling Node

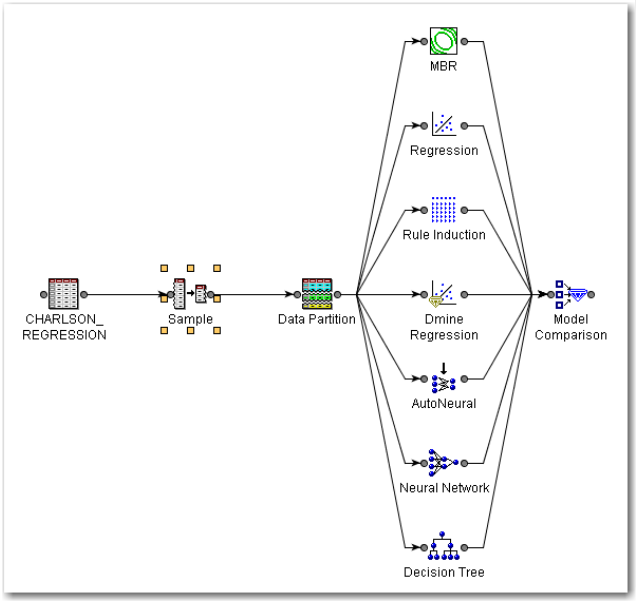
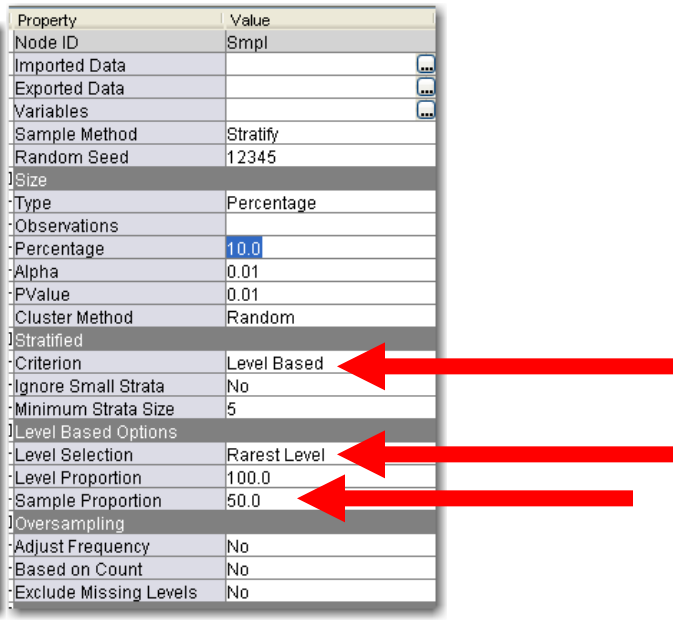


Figure 11. Change to Defaults in Sampling Node



In the following examples, we use a 50/50 split in the data. We use just three patient diagnoses of pneumonia, septicemia, and immune disorder to predict mortality. We use all of the models depicted in Figure 10. According to the model comparison, the rule induction provides the best fit, using the misclassification criterion as the measure of "best". We first look at the regression model, comparing the results to those when a 50/50 split was not performed. The overall misclassification rate is 28%, with the divisions as shown in Table 1. Table 2 shows a 25/75 split in the data. The contrast in the misclassification rate is clear. Table 3 gives the results for a 10/90 split.

Table 1. Misclassification in Regression Model with 50/50 Split

Target	Outcome	Target Percentage	Outcome Percentage	Count	Total Percentage
Training Data					
0	0	67.8	80.1	54008	40.4
1	0	32.2	38.3	25622	19.2
0	1	23.8	19.2	12852	9.6
1	1	76.3	61.7	41237	30.8
Validation Data					
0	0	67.7	80.8	40498	40.4
1	0	32.3	38.5	19315	19.2
0	1	23.8	19.2	9646	9.6
1	1	76.2	61.5	30830	30.7

Table 2. Misclassification with a 25/75 Split in the Data

Target	Outcome	Target Percentage	Outcome Percentage	Count	Total Percentage
Training Data					
0	0	80.4	96.6	10070	72.5
1	0	19.6	70.9	2462	17.7
0	1	25.6	3.3	348	2.5
1	1	74.4	29.1	1010	7.3
Validation Data					
0	0	80.2	97.1	7584	72.8
1	0	19.8	71.7	1870	17.9
0	1	23.7	2.9	229	2.2
1	1	76.2	28.2	735	7.0

Table 3. Misclassification Rate for a 10% Sample

Target	Outcome	Target Percentage	Outcome Percentage	Count	Total Percentage
Training Data					
0	0	91.5	99.3	31030	89.4
1	0	8.5	83.5	2899	8.3
0	1	27.3	0.7	216	0.6
1	1	72.6	16.5	574	1.6
Validation Data					
0	0	91.5	99.2	23265	89.3
1	0	8.4	82.4	2148	8.2
0	1	27.8	0.7	176	0.7
1	1	72.2	17.5	457	1.7

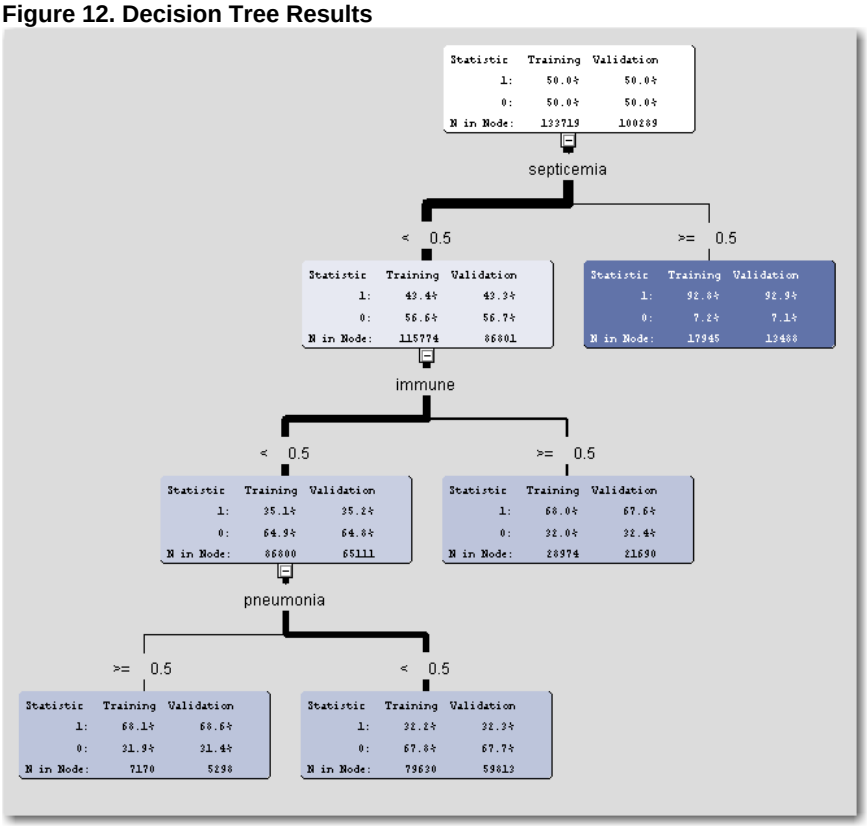
The misclassification becomes more balanced between false positives and false negatives with a 50/50 split in the data. The model gives heavier weight to false positives than it does to false negatives. Table 4 shows the contrast without a 50/50 split using all of the available data. The false negative rate is extremely high even while the overall accuracy is 97%.

Table 4. Classification Table for Logistic Regression With Pneumonia and Septicemia

Classification Table									
Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.580	782E4	0	167E3	0	97.9	100.0	0.0	2.1	.
0.600	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.620	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.640	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.660	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.680	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.700	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.720	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.740	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.760	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.780	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.800	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.820	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.840	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.860	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.880	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.900	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.920	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.940	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.960	731E4	63391	104E3	517E3	92.2	93.4	37.9	1.4	89.1
0.980	731E4	63391	104E3	517E3	92.2	93.4	37.9	1.4	89.1
1.000	0	167E3	0	782E4	2.1	0.0	100.0	.	97.9

We now examine additional predictive models to contrast them with the logistic model, again using the 50/50 split. We first want to examine the decision tree model. While it is not the most accurate model, it is one that clearly describes the rationale behind the predictions. This tree is given in Figure 12. The tree shows that the first split occurs on the variable, Septicemia. Patients with Septicemia are more likely to suffer mortality compared to patients without Septicemia. The Immune Disorder has the next highest level of mortality, followed by Pneumonia.

Figure 12. Decision Tree Results



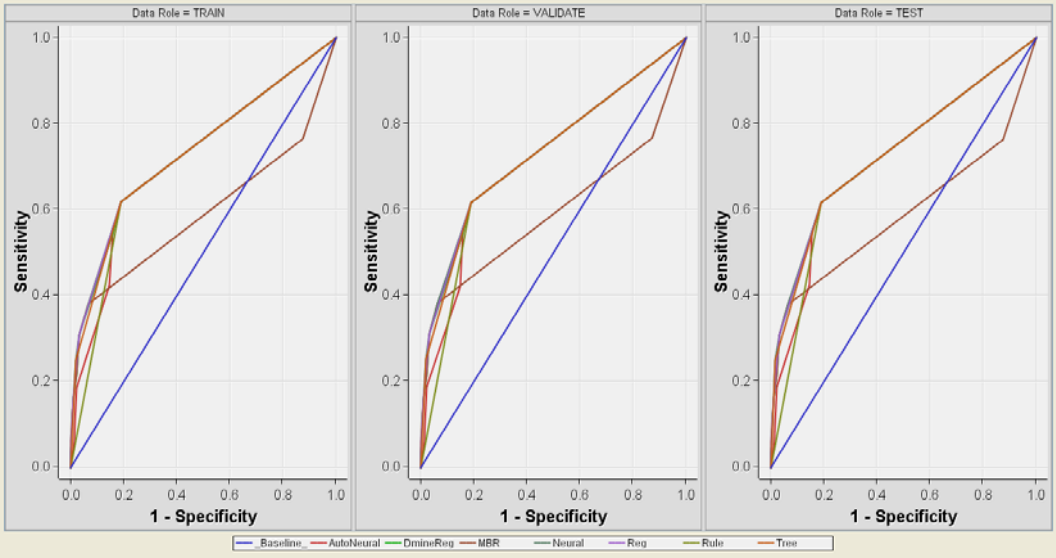
Since rule induction is identified as the best model, we examine that one next. The misclassification rate is only slightly smaller compared to the regression model. Table 5 gives the classification table.

Table 5. Misclassification in Rule Induction Model

Target	Outcome	Target Percentage	Outcome Percentage	Count	Total Percentage
Training Data					
0	0	67.8	80.8	54008	40.4
1	0	32.2	38.3	25622	19.2
0	1	23.8	19.2	12852	9.6
1	1	76.3	61.7	41237	30.8
Validation Data					
0	0	67.7	80.8	40498	40.4
1	0	32.3	38.5	19315	19.2
0	1	23.8	19.2	9646	9.6
1	1	76.2	61.5	30830	30.7

The results look virtually identical to those in Table 1. For this reason, the regression model, although not defined as the best, can be used to predict outcomes when only these three variables are used. The similarities in the models can also be visualized in the ROC (receiver-operating curve) that graphs the sensitivity versus one minus the specificity (Figure 13). The curves for rule induction and regression are virtually the same.

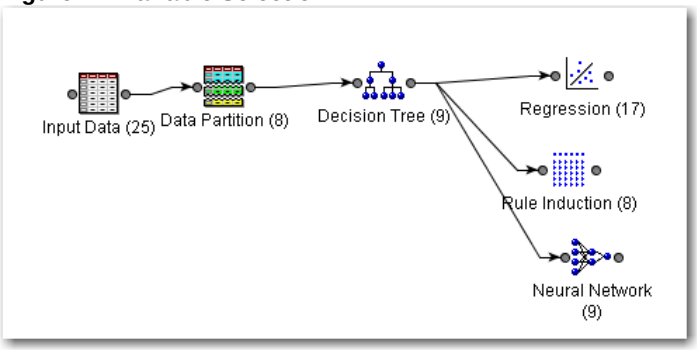
Figure 13. Comparison of ROC Curves



Many Variables in Large Samples

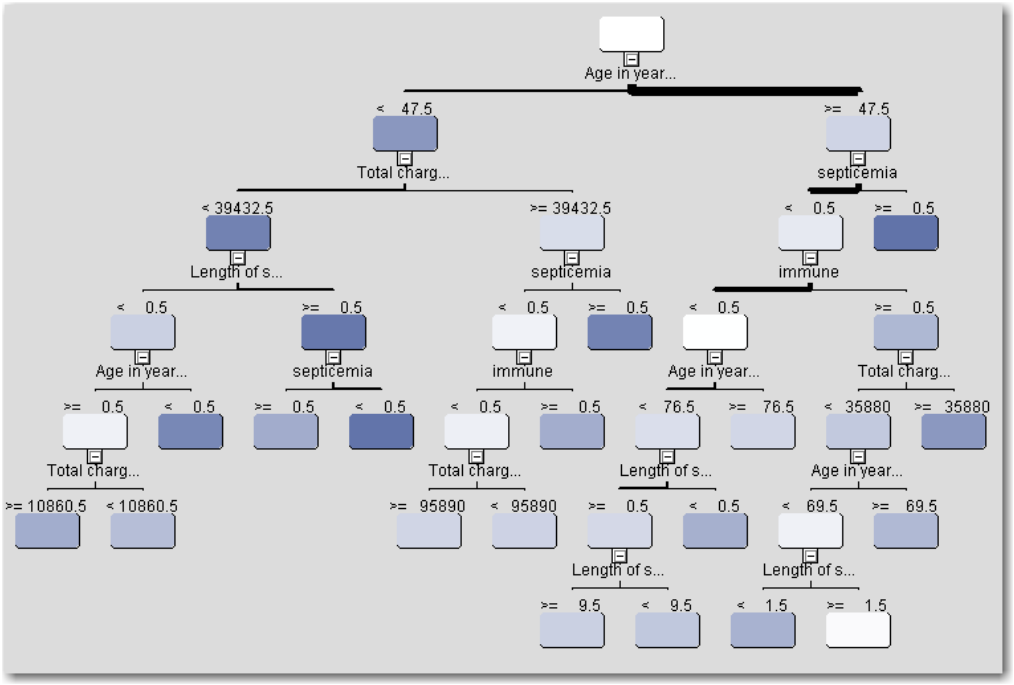
There can be hundreds if not thousands of variables collected for each patient. There can be far too many to include in any predictive model. We want to include all those variables that are crucial to the analysis, including potential confounders, but the use of too many variables can cause the model to over-fit the results, inflating the outcomes. Therefore, there needs to be some type of variable reduction method. In the past, factor analysis has been used to reduce the set of variables prior to modeling the data. However, there is now a more novel method available (Figure 14). In our example, there are many additional variables that can be considered in this analysis. Therefore, we use the variable selection technique to choose the most relevant. We first use the decision tree followed by regression, and then regression followed by the decision tree. The first model gives the variable selection; the second model gives the prediction.

Figure 14. Variable Selection



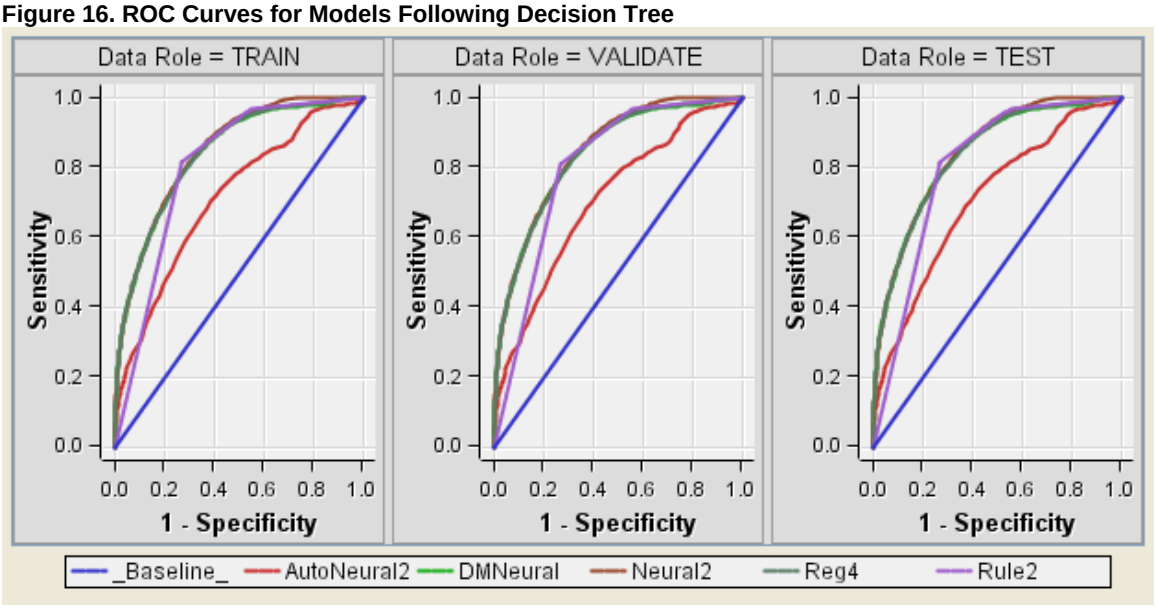
Using the decision tree to define the variables, Figure 15 shows the ones that remain for the modeling. Note that age, charges, and length of stay are at the beginning of the tree.

Figure 15. Decision Tree Variables



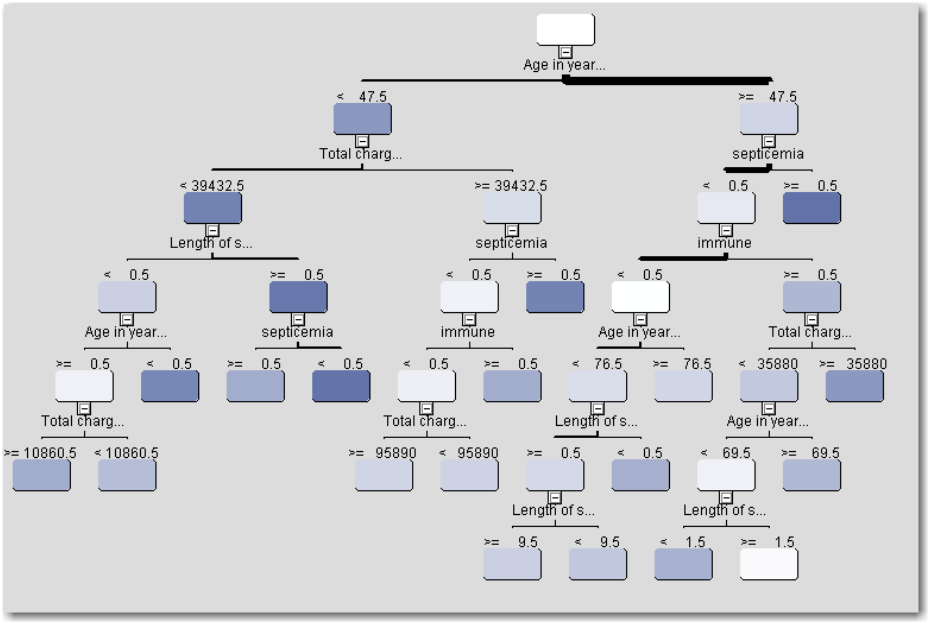
This tree shows that age, length of stay, having septicemia, immune disorder, length of stay and total charges are related to mortality. The remaining variables have been rejected from the model. The rule induction is the best model, and the misclassification rate decreases to 22% with the added variables. The ROC curve looks considerably improved (Figure 16).

Figure 16. ROC Curves for Models Following Decision Tree



The ROC curve is much higher compared to that in Figure 13. If we use regression to perform the variable selection, the results remain the same. In addition, a decision tree is virtually the same when it follows the regression compared to when it precedes regression (Figure 17).

Figure 17. Decision Tree Following Regression



The above example only used three possible diagnosis codes. We want to expand upon the number of diagnosis codes, and also to use a number of procedure codes. However, to a certain extent, the problem becomes just how many diagnosis and procedure codes should be used to investigate the data. There are approximately 17,000 possible diagnosis codes and a similar number of procedure codes. Just how many codes should be used to be able to define a model? If as many as 50 are used, there is still the problem that mortality is a rare occurrence in all of them, and the ability of the model to predict mortality will remain poor as long as each is a rare occurrence. For more information on predictive modeling and patient severity measures, we refer the interested reader to Cerrito (2009).

DISCUSSION

Great care must be taken with large samples to ensure that the results are not misleading because of the highly significant p-values and the high level of accuracy in the model. This is especially true when investigating rare occurrences since the group sizes in the regression model will be extremely disparate. If this is not taken into consideration, the model will be highly accurate but will not have any predictive ability.

It is possible to focus attention on those patients most at high risk for a rare occurrence by using the concept of lift, which is very commonly used in data mining. Using lift, true positive patients with highest confidence come first, followed by positive patients with lower confidence. True negative cases with lowest confidence come next, followed by negative cases with highest confidence. Based on that ordering, the observations are partitioned into deciles, and the following statistics are calculated:

- The *Target density* of a decile is the number of actually positive instances in that decile divided by the total number of instances in the decile.
- The *Cumulative target density* is the target density computed over the first *n* deciles.
- The *lift* for a given decile is the ratio of the target density for the decile to the target density over all the test data.
- The *Cumulative lift* for a given decile is the ratio of the cumulative target density to the target density over all the test data.

Lift, then can find the 20% of patients most at risk, given a model that can focus on these 20%. Standard logistic regression cannot distinguish between patients at high risk and those at moderate risk. For more information, we again refer the interested reader to Cerrito (2009).

REFERENCES

1. Anonymous-Cancer (2011) Avastin in combination with chemo associated with increased fatality. Cancer Network **February 9, 2011**,
2. Battioui, C. (2007). Cost Models with Prominent Outliers, Digital Dissertations.
3. Cerrito, P. (2009). Text mining techniques for healthcare provider quality determination: methods for rank comparisons. Hershey, PA, IGI Publishing.
4. Cerrito, P. and J. Cerrito (2011). Assumptions and Consequences of Comparative Effectiveness Analysis Using Data Mining. PharmaSug, Nashville, TN, PharmaSug.
5. Eifert, S., E. Kilian, et al. (2010). "Early and mid term mortality after coronary artery bypass grafting in women depends on the surgical protocol: retrospective analysis of 3441 on- and off-pump coronary artery bypass grafting procedures." Journal of Cardiothoracic Surgery **5**(90).
6. Kalager, M., T. Haldorsen, et al. (2009). "Improved breast cancer survival following introduction of an organized mammography screening program among both screened and unscreened women: a population-based cohort study." Breast Cancer Research & Treatment **11**: R44.
7. Ranpura, V., S. Hapani, et al. (2011). "Treatment-related mortality with bevacizumab in cancer patients." JAMA **305**(5): 487-494.
8. Raval, M. V., M. E. Cohen, et al. (2010). "Improving American college of surgeons national quality improvement program risk adjustment: incorporation of a novel procedure risk score." Journal of the American College of surgeons **211**(6): 715-723.
9. Ray, M. E., K. Bae, et al. (2009). "Potential surrogate endpoints for prostate cancer survival: analysis of a phase III randomized trial." Journal of the National Cancer institute **101**: 228-236.
10. Vieitez, J. M., J. Carrasco, et al. (2003). "Irinotecan in the treatment of advanced colorectal cancer in patients pretreated with fluorouracil-based chemotherapy." American Journal of Clinical Oncology **26**(2): 107-111.
11. Yao, S., W. E. Barlow, et al. (2010). "Manganese superoxide dismutase polymorphism, treatment-related toxicity and disease-free survival in SWOG 8897 clinical trial for breast cancer." Breast Cancer Research & Treatment **124**: 433-439.
12. Zhao, Y., A. H. Lee, et al. (2009). "A score test for assessing the cured proportion in the long-term survivor mixture model." Statistics in Medicine **28**: 3454-3466.

ABOUT THE AUTHOR

Patricia Cerrito, PhD, Professor of Mathematics at the University of Louisville, has spent over 20 years investigating health outcomes using data mining tools. Dr. Cerrito has been recognized as one of America's Elite Educators. She has published a number of books on the general topic including

- Cases on Health Outcomes and Clinical Data Mining: Studies and Frameworks
- Text Mining Techniques for Healthcare Provider Quality Determination: Methods for Rank Comparisons
- Clinical Data Mining for Physician Decision Making and Investigating Health Outcomes: Methods for Prediction and Analysis

CONTACT INFORMATION

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

Name	Patricia Cerrito
Enterprise	University of Louisville
Address	Department of Mathematics
City, State ZIP	40292
Phone:	502-852-6826 502-742-0889
Fax:	502-852-7132
E-mail:	pcerrito@gmail.com

Web site: drpatriciacerrito.com