Paper 11681-2016

Predicting occurrence rate of Systemic Lupus Erythematosus (SLE) and Rheumatoid Arthritis (RA) in pregnant women

Ravikrishna Vijaya Kumar, Shrieraam Sathyanarayanan, Dr. William D. Paiva and Dr. Goutam Chakraborty, Oklahoma State University

ABSTRACT

Years ago, doctors advised women with autoimmune diseases such as systemic lupus erythematosus (SLE) and rheumatoid arthritis (RA) not to become pregnant for fear of maternal health. Now, it is known that healthy pregnancy is possible for women with lupus but at the expense of higher pregnancy complication rate. The main objective of this research is to identify key factors contributing to these diseases and to predict the occurrence rate of SLE and RA in pregnant women. Based on the approach used in this study, prediction of adverse pregnancy outcomes for women with SLE, RA and other diseases such as DM (Diabetes Mellitus) and APS (Anti-Phospholipid Antibody Syndrome) can be carried out. These results will help pregnant women to undergo healthy pregnancy by proper medication at an earlier stage. The data set was obtained from Cerner Health Facts data warehouse. The raw data set contains 883,473 records and 85 variables such as diagnosis code, age, race, procedure code, admission date, discharge date, total charges etc. Analyses were carried out with two different datasets, one for SLE patients and the other for RA patients. The final datasets had 397,898 and 398,742 records each for modeling RA and SLE patients respectively. To provide an honest assessment of models, the data was split into training and validation using data partition node. Variable selection techniques such as LASSO, LARS, Stepwise Regression, and Forward Regression were used. Using decision tree, prominent factors that determines the SLE and RA occurrence rate were identified separately. Of all the predictive models run using SAS® Enterprise Miner™ 12.3, the model comparison node identified Decision tree (Gini) as the best model with the least misclassification rate of 0.308 to predict the SLE patients and 0.288 to predict the RA patients.

INRODUCTION

It is estimated that prevalence of SLE in the US is 1 to 4 per 1,000 women [2]. Studies have shown that women comprise 90 % of lupus patients. Similarly it is estimated that among women of ages 16-44 years the prevalence of RA (Rheumatoid Arthritis) to be 1 to 2 cases per 1,000 women in the UK [3]. It is important to study the pregnancy outcomes in women with autoimmune diseases in order to take preventive measures to result in a healthy pregnancy. There are increased rates of cesarean deliveries in patients with SLE [4]. Factors like length of stay, age, ethnicity affects the pregnancy outcomes of patients with SLE and RA [5]. Women with SLE and RA has significantly increased rates of hypertensive disorders, longer hospital stays, higher risk of cesarean delivery, and are older than the general population [5].

Researches are being done to reduce the risks involved in pregnancy while having autoimmune diseases. With careful management of medication prior and during pregnancy these risks can be minimized. As a result of this analysis, strategies can be developed to improve pregnancy outcomes, especially in women with autoimmune diseases. The objective of this paper is to predict the occurrence of SLE and RA in pregnant women. SAS® Enterprise Miner™ 12.3 is used in this paper to identify patients among various pregnancy hospitalizations who display a higher likelihood of having SLE and RA. The combination of patients with both SLE and RA is neglected for this analysis. Future development of the project will involve adding variables that captures adverse pregnancy outcomes and use them as inputs to predict the SLE and RA patients.

LITERATURE REVIEW

Health care organizations are trying to develop, innovate and implement new and adaptive health care system models and products that focuses on care, treatment and health efficiency. Research by Dr. Eliza Chakravarty et al in 2006 using 2002 Nationwide Inpatient Sample (NIS), was based on obstetric hospitalizations in the United States for women with SLE and RA. This was the first study to examine pregnancy outcomes in national data on women with common rheumatic diseases. The software used to perform the analyses was Stata version 8.0. Models like logistic regression and linear regression were built to determine coefficient of length of stay and age as the covariates. Yasmeen et al 2001 [4] studied the pregnancy outcomes in women with SLE using the California Health Information. The study suggested that there are increased rates of cesarean deliveries reported for SLE patients.

Skomsvoll JF et al (2007) [6] studied the Medical Birth Registry of Norway during the years 1967–95 in women. The results showed that women with RA had significantly higher rates of cesarean section. Another study by Nossent HC et al (1990) [7] was done on influence of systemic lupus erythematosus (SLE) on pregnancy. Gimovsky ML et al (1984) [8] studied about pregnancy outcome in women with SLE, and mentioned relationships between the women affected by SLE with and without renal manifestation. Another study by Symmons D et al (2002) [7] used Norfolk Arthritis Register (NOAR) to estimate the prevalence of rheumatoid arthritis in the United Kingdom and estimated that about 1 to 2 cases per 1,000 women were diagnosed with RA. To our knowledge none of the authors have used SAS® to model. Most of the authors did basic descriptive analysis to compare means with control groups. Eliza's results and methodology were easy to interpret.

DATA

This study involves data obtained from the Cerner Health Facts database. Data is real-world, HIPAA-complaint, de-identified, sequenced and time-stamped with its source coming from over 480 hospitals. Cerner Health Facts is the largest relational database on health care. It is the industry's only data warehouse that includes pharmacy, laboratory, billing, clinical events and admission data of the patients. Cerner Health Facts database consists of over 58 million total unique patients with more than 2.4 billion laboratory results. It has more than 14 years of detailed laboratory, pharmacy, registration and billing data.

Years ago, doctors advised women with SLE not to become pregnant for fear of maternal health. Now, it is seen that healthy pregnancy is possible for women with lupus but at the expense of higher pregnancy complication rate.

For this study, we extracted dataset that had 85 variables and 883,473 records. These records include information about various complications related to women during pregnancy. Following table represents the sample of the dataset with variables used for modeling.

VARIABLE NAME	MEASUREMENT LEVEL	POTENTIAL VALUES
ADMISSION_SOURCE_CODE	NOMINAL	1-9, A, B,C, N, O, P, Q, R, -1, 88888, 99999
ADMISSION_SOURCE_CODE_DESC	NOMINAL	Examples: Physician Referral, Clinic Referral, Emergency Room, Transfer from a hospital, Not Available
admission_source_id	INTERVAL	1 to 26
admission_type_id	INTERVAL	1 to 8
admitted_dt_tm	INTERVAL	ddmmmyyyy:hh:mm:ss

admitting_physician_id	INTERVAL	Min -3995844 thru + 44500000, -1 (Physician NULL), -9 (Physician Not Found)
age_in_years	INTERVAL	0 to 90
BED_SIZE_RANGE	NOMINAL	<6, 6-99, 100-199, 200-299, 300-499, 500+
CARESETTING_DESC	NOMINAL	Examples: Ambulatory Unit, Cardiology, Family Practice Clinic, Genetics, Medical/Surgical, Obstetrics & Gynecology, Oncology, Intensive Care Unit, Intensive Care Unit - Neonatal
CARESETTING_ID	INTERVAL	1 to 178
CENSUS_REGION	NOMINAL	Northeast, Midwest, South, West
discharged_dt_tm	INTERVAL	ddmmmyyyy:hh:mm:ss
DISCHG_DISP_CODE	NOMINAL	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 41, 42, 43, 50, 51, 61, 62, 63, 64, 70, 71,72,, 100-109, -1
DISCHG_DISP_CODE_DESC	NOMINAL	Examples: Discharged to home, Expired, Discharged/transferred to a SNF
DISCHG_DISP_ID	INTERVAL	1 to 31
ENCOUNTER_ID	INTERVAL	20 digit number
gender	UNARY	Female, Male, Null, Unknown/Invalid, Null
marital_status	NOMINAL	Divorced, Legally Separated, Married, Single, Unknown, Widowed, Null
patient_sk	INTERVAL	
PATIENT_TYPE_DESC	NOMINAL	Inpatient, Emergency, Outpatient, Pre-Admit, Observation, Recurring, Short Stay, Outpatient Surgery, Clinic, Billing, Dental, Hospice, Non-patient
patient_type_id	INTERVAL	75 to 145
PAYER_CODE	NOMINAL	Examples: BC, CH, HM, MC, SP
PAYER_CODE_DESC	NOMINAL	Examples: Blue Cross/Blue Shield, CHAMPUS (Military dependents), HMO/Managed Care (undesignated), Medicare, Self-Pay
payer_id	INTERVAL	1 to 23

race	NOMINAL	Caucasian, African American, Asian, Native American, Unknown, Hispanic, Other, Not Mapped
TEACHING_FACILITY_IND	BINARY	1 (Teaching), 0 (Nonteaching), -1 (NULL)
total_charges	INTERVAL	
URBAN_RURAL_STATUS	BINARY	U (Urban), R (Rural)
Classification	NOMINAL	Control, SLE, RA
Sle_yes_no	BINARY	1 (Has SLE) ,0 (Does not have SLE)
RA_yes_no	BINARY	1 (Has RA) ,0 (Does not have RA
LOS (Length of Stay)	NOMINAL	0.01 to 500

Table 1. Variables in the analysis data set

World Health Organization (WHO) maintains the data set for the use of International Statistical Classification of Diseases known as ICD consists of information and records for patients with different health conditions. These ICD-9-CM diagnosis and procedure codes for the pregnancy related complications were identified and the dataset was extracted by matching these codes. Datasets corresponding to SLE and RA diseases were extracted separately which contained 17,385 and 41,599 total patients respectively.

Diagnosis and procedure codes related to Normal delivery, Premature and Distress deliveries are ICD-9-CM 650, V22, V23.41, and 669. Those related to Cesarean section, Early or Threatened Labor are ICD-9-CM 74, 644, 654.2, and 642.

DATA CLEANSING AND PREPARATION

The original dataset had 85 variables and 883,473 observations. To prepare the data for modeling, data was subjected to intensive cleansing procedures. The final datasets had 397,898 and 398,742 records with 72 variables each for modeling RA and SLE patients respectively. Variables like total charges, length of stay, age in years, admission time, and discharge time had missing values. In order to avoid modeling bias and imputing the missing values, the observations with missing values were removed. PROC SQL queries and DATA steps were used to remove the missing values. Certain variables like race, patient type, and admission data had ambiguous values like ‘Null’, and intentionally entered values like ‘Not Mapped’, and so on. For few records the admission date of the patient was future dated compared to the discharge date because of which we had negative values when calculating the length of hospital stay for all these patients. So they had to be cleaned and recoded.

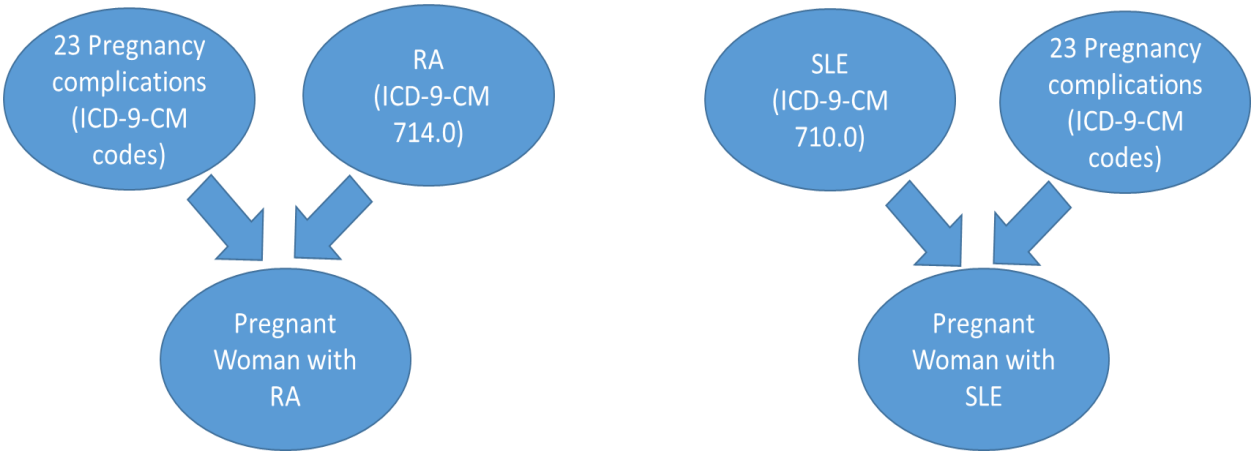

Figure 1. Data consolidation schematic view

Patient ID is associated with the encounter_id of the patient. Although this patient_id is considered as the primary key, it varies for a single patient whenever the patient is entered newly in the database. So, we had duplicate records for a single patient with varying patient ids. For this reason, patient_sk which is a unique identifier for each patient was considered when merging datasets or to pull records for a diagnosis.

New variables LOS (Length of Stay), Sle_yes_no and Ra_yes_no were created using DATA steps in SAS® Enterprise Guide. As this is a manually entered data, there were numerous duplicate records found. We obtained the data in xlsx format and while importing them into SAS® Enterprise Guide we had variable format issues. The datasets for each of the pregnancy types were extracted separately and merged together. Issues while merging them were taken care. Then we matched the pregnancy data with SLE and RA data which were extracted and cleaned the same way, to obtain two different final data sets to model. As this is a manually entered data, even pregnancy instances were recorded for male patients, which were removed later.

Variable Summary

Role	Measurement Level	Frequency Count
ID	INTERVAL	4
INPUT	BINARY	2
INPUT	INTERVAL	3
INPUT	NOMINAL	3
REJECTED	BINARY	7
REJECTED	INTERVAL	24
REJECTED	NOMINAL	19
REJECTED	UNARY	9
TARGET	BINARY	1

Table 2. Summary of variables

The key issue with the final datasets is that the ratio of pregnant women with SLE against pregnant women without SLE was too low (844/398,742), i.e. around 0.21 %. Similarly for RA patients the percentage

incident was 0.16% (660/398,742). Predictive accuracy to evaluate performance of the classifier might not be appropriate when the data is imbalanced [9]. We used a sample of 30 % target and 70% non-target variable [10] [11]. So, in this case since we had 844 records in SLE data, a random sampling from the data for pregnant women without SLE diagnosis (Control group) was done with (844*1.7) 1,435 records. Similarly we carried out this process for the RA dataset and took 1,122 records (660*1.7) as a random sample from the data for pregnant women without RA diagnosis. Modeling was carried out with two different final datasets – one for SLE and the other for RA.

DESCRIPTIVE ANALYSIS

In order to represent the results analogous to the authors referred, we performed basic descriptive statistics on the data. Most of the findings were in accordance with our analysis. For example, the study conducted by Eliza Chakravarty et al (eliza) showed that pregnant women with RA and SLA have significantly higher rates of hypertensive disorders compared with general population (14.50%, 22.28 % and 10.70% respectively). They also have longer hospital stays (3.07, 3.40, and 2.87 days respectively) and older than the general population (34.3, 30.62, and 27.98 years respectively). In addition to that we also found that on an average RA and SLE patients spend more than the general population ($13,308, $16,207, and $11,215 respectively). The effects of adverse pregnancy outcomes in those women with SLE and RA are yet to be studied.

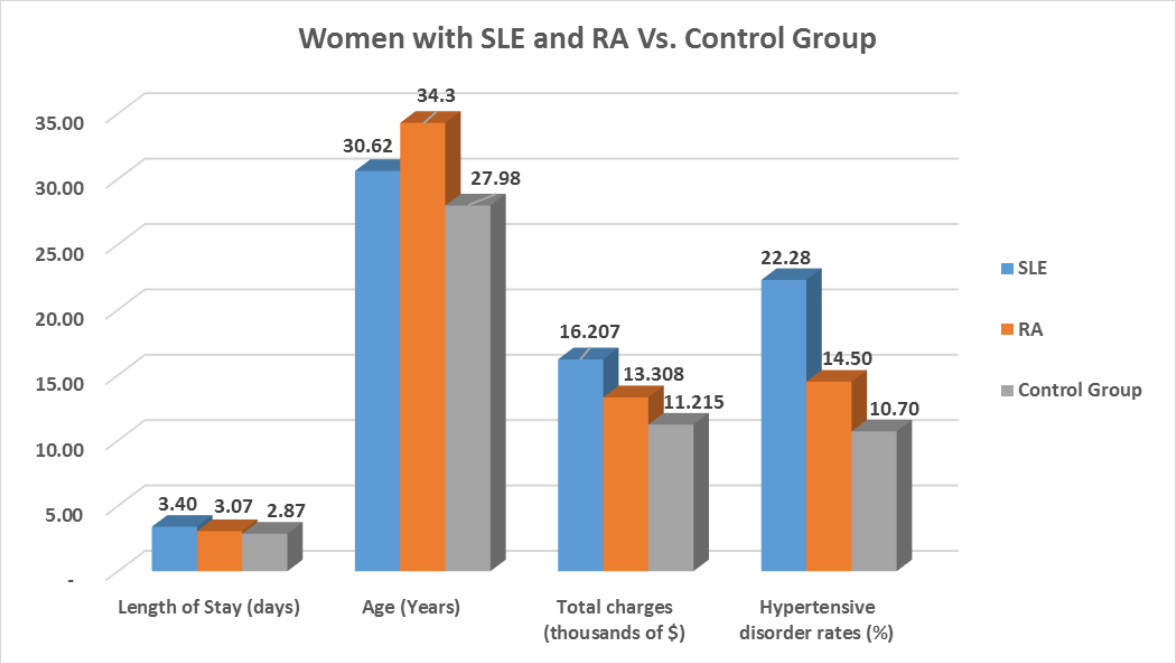

Figure 2. Descriptive statistics

As far as the variables' distribution, after removing missing values, the statistics for the interval variables seemed satisfactory. Similarly distribution and statistics of nominal variables in the SLE data also seemed satisfactory.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
DAY_NUMBER_IN_MONTH	INPUT	15.69241	8.856761	2279	0	1	16	31	0.006125	-1.19795
ENCOUNTER_ID_0001	INPUT	1.0136E8	68488260	2279	0	1491843	84436959	2.8181E8	0.476102	-0.7084
ENCOUNTER_ID_0002	INPUT	1.0136E8	68488260	2279	0	1491843	84436959	2.8181E8	0.476102	-0.7084
HOSPITAL_ID_0001	INPUT	111.4493	126.7567	2279	0	11	67	669	2.625637	7.222316
LOS	INPUT	3.064923	2.598312	2279	0	0.017361	2.666667	40.12361	6.680988	67.01892
WEEK_NUMBER_IN_YEAR	INPUT	26.38833	14.85067	2279	0	1	27	53	0.008907	-1.17089
age_in_years	INPUT	28.88328	7.906304	2279	0	3	28	90	2.029518	10.10838
patient_id_0001	INPUT	62952052	46286528	2279	0	524565	53443261	1.6638E8	0.522566	-0.81052
patient_sk	INPUT	1.992E12	2.692E12	2279	0	1.8962E8	7.19E11	1.566E13	1.615816	1.545439
total_charges	INPUT	12841.53	18752.48	2279	0	0.34	9394.91	500341.9	12.86564	257.6068
weight	INPUT	22.53326	46.77884	2279	0	0	0	274	2.224559	4.937752

Table 3. SLE Interval Summary Statistics

After removing missing values from variables like age, total charges, and length of stay, almost all of the other variables did not have missing values. Even after removing the records with missing values, we still had duplicate values which we had to remove. Regarding the distribution of the variables, most of the selected variables are normally distributed and had less kurtosis values.

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	ACUTE_STATUS	INPUT	3	1	Acute	99.25	Non-Acute	0.70
TRAIN	ADMISSION_SOURCE_CODE	INPUT	11	0	1	81.75	2	6.49
TRAIN	ADMISSION_SOURCE_CODE_DESC	INPUT	11	0	Physician Referral	81.75	Clinic Referral	6.49
TRAIN	ADMISSION_SOURCE_ID_0001	INPUT	11	0	1	81.75	2	6.49
TRAIN	BED_SIZE_RANGE	INPUT	8	0	300-499	31.90	200-299	20.58
TRAIN	CATH_LAB_DIAGNOSTIC_IND	INPUT	2	0	1	72.09	0	27.91
TRAIN	CATH_LAB_FULL_IND	INPUT	2	0	1	75.34	0	24.66
TRAIN	CENSUS_DIVISION	INPUT	9	0	2	32.34	6	24.79
TRAIN	CENSUS_REGION	INPUT	4	0	Northeast	42.26	South	30.72
TRAIN	DAY_NUMBER_OF_WEEK	INPUT	7	0	3	17.07	2	16.45
TRAIN	DAY_OF_WEEK	INPUT	7	0	TUESDAY	17.07	MONDAY	16.45
TRAIN	DISCHG_DISP_CODE	INPUT	11	0	1	91.93	6	3.77
TRAIN	HOLIDAY_IND	INPUT	2	0	0	98.82	1	1.18
TRAIN	MONTH	INPUT	12	0	9	9.39	3	8.91
TRAIN	MONTH_NAME	INPUT	12	0	SEP	9.39	MAR	8.91
TRAIN	PATIENT_TYPE_DESC	INPUT	6	0	Inpatient	91.75	Emergency	4.04
TRAIN	PATIENT_TYPE_ID_0001	INPUT	6	0	87	91.75	84	4.04
TRAIN	PAYER_CODE	INPUT	19	0	-1	25.49	MD	16.24
TRAIN	PAYER_CODE_DESC	INPUT	19	0	NULL	25.49	Medicaid	16.24
TRAIN	PAYER_ID_0001	INPUT	19	0	22	25.49	10	16.24
TRAIN	QUARTER	INPUT	4	0	3	26.68	1	25.10
TRAIN	TEACHING_FACILITY_IND	INPUT	2	0	1	70.21	0	29.79
TRAIN	URBAN_RURAL_STATUS	INPUT	2	0	Urban	99.78	Rural	0.22
TRAIN	WEEKDAY_IND	INPUT	2	0	1	82.19	0	17.81
TRAIN	YEAR	INPUT	14	0	2010	16.67	2009	14.30
TRAIN	discharged_tm_valid_ind	INPUT	2	0	1	97.54	0	2.46
TRAIN	marital_status	INPUT	9	0	Married	38.88	Single	30.19
TRAIN	race	INPUT	12	0	Caucasian	65.20	African American	18.43
TRAIN	Sle_yes_no	TARGET	2	0	0	62.97	1	37.03

Table 4. SLE Nominal Statistics

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
DAY_NUMBER_IN_MONTH	INPUT	15.77385	8.736901	1782	0	1	16	31	-0.01777	-1.16902
ENCOUNTER_ID_0001	INPUT	1.0363E8	67336661	1782	0	1452872	95465352	2.8181E8	0.439306	-0.67863
ENCOUNTER_ID_0002	INPUT	1.0363E8	67336661	1782	0	1452872	95465352	2.8181E8	0.439306	-0.67863
HOSPITAL_ID_0001	INPUT	108.2357	113.0116	1782	0	11	67	669	2.549927	7.664929
LOS	INPUT	2.987076	2.759735	1782	0	0.024306	2.649306	59.76806	11.12916	188.9781
WEEK_NUMBER_IN_YEAR	INPUT	26.78171	15.15375	1782	0	1	27	53	-0.00922	-1.23826
age_in_years	INPUT	30.13917	10.81874	1782	0	3	29	90	2.612171	10.0037
patient_id_0001	INPUT	65127384	46113046	1782	0	528355	54394499	1.6653E8	0.472928	-0.83848
patient_sk	INPUT	2.003E12	2.753E12	1782	0	2.6933E9	7.075E11	1.582E13	1.668201	1.745696
total_charges	INPUT	12253.13	14572.05	1782	0	0.19	9212.84	275405.9	8.000746	110.5936
weight	INPUT	22.92222	48.09969	1782	0	0	0	305	2.284105	5.17464

Table 5. RA Interval Summary Statistics

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	ACUTE_STATUS	INPUT	2	0	Acute	99.27	Non-Acute	0.73
TRAIN	ADMISSION_SOURCE_CODE	INPUT	12	0	1	81.31	2	6.06
TRAIN	ADMISSION_SOURCE_CODE_DESC	INPUT	12	0	Physician Referral	81.31	Clinic Referral	6.06
TRAIN	ADMISSION_SOURCE_ID_0001	INPUT	12	0	1	81.31	2	6.06
TRAIN	BED_SIZE_RANGE	INPUT	8	0	300-499	32.27	200-299	19.36
TRAIN	CATH_LAB_DIAGNOSTIC_IND	INPUT	2	0	1	73.29	0	26.71
TRAIN	CATH_LAB_FULL_IND	INPUT	2	0	1	76.43	0	23.57
TRAIN	CENSUS_DIVISION	INPUT	9	0	2	29.46	6	25.98
TRAIN	CENSUS_REGION	INPUT	4	0	Northeast	38.83	South	31.82
TRAIN	DAY_NUMBER_OF_WEEK	INPUT	7	0	2	18.46	4	16.84
TRAIN	DAY_OF_WEEK	INPUT	7	0	MONDAY	18.46	WEDNESDAY	16.84
TRAIN	DISCHG_DISP_CODE	INPUT	10	0	1	89.90	6	4.32
TRAIN	HOLIDAY_IND	INPUT	2	0	0	98.04	1	1.96
TRAIN	MONTH	INPUT	12	0	3	9.88	10	9.88
TRAIN	MONTH_NAME	INPUT	12	0	MAR	9.88	OCT	9.88
TRAIN	PATIENT_TYPE_DESC	INPUT	6	0	Inpatient	91.92	Emergency	5.50
TRAIN	PATIENT_TYPE_ID_0001	INPUT	6	0	87	91.92	84	5.50
TRAIN	PAYER_CODE	INPUT	18	0	-1	25.93	MD	15.04
TRAIN	PAYER_CODE_DESC	INPUT	18	0	NULL	25.93	Medicaid	15.04
TRAIN	PAYER_ID_0001	INPUT	18	0	22	25.93	10	15.04
TRAIN	QUARTER	INPUT	4	0	4	26.32	1	25.70
TRAIN	TEACHING_FACILITY_IND	INPUT	2	0	1	66.55	0	33.45
TRAIN	WEEKDAY_IND	INPUT	2	0	1	80.13	0	19.87
TRAIN	YEAR	INPUT	14	0	2010	16.44	2009	15.77
TRAIN	discharged_tm_valid_ind	INPUT	2	0	1	98.04	0	1.96
TRAIN	marital_status	INPUT	8	0	Married	39.96	Single	29.18
TRAIN	race	INPUT	11	0	Caucasian	68.74	African American	13.24
TRAIN	ra_yes_no	TARGET	2	0	0	62.96	1	37.04

Table 6. RA Nominal Statistics

PREDICTIVE MODELING

The data was split into training (70 %) and validation (30 %) before modeling, to provide an honest assessment of the model. Before using the data to build models, important variables were identified using standard variable selection methods such as LARS (Least Angle Regression), LASSO, Adaptive LASSO, Stepwise regression, forward regression, and decision tree. The variables selected by decision tree were more contributory in reducing the misclassification rate. Variables such as age in years, length of stay, total charges and census region had more potential in predicting the classifier for the RA data.

VARIABLE SELECTION

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
age_in_years		2	1.0000	1.0000	1.0000
LOS		2	0.7400	0.5914	0.7992
total_charges		1	0.2600	0.2499	0.9610
CENSUS_REGION		1	0.1884	0.2445	1.2980

Table 7. Variable Selection

Similarly for SLE, variables such as age in years, length of stay, total charges, patient type, and census region were selected as the most important predictors of the target.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
age_in_years		6	1.0000	0.9052	0.9052
LOS		5	0.9889	0.7526	0.7611
total_charges		6	0.8860	1.0000	1.1287
PATIENT_TYPE_ID_0001		1	0.5492	0.6410	1.1671
CENSUS_REGION		1	0.3162	0.4441	1.4046

Table 8. Variable Selection

PREDICTING FOR SLE PATIENTS

After importing the data into SAS® Enterprise Miner™ 12.3, we used various models like decision tree (gini, entropy, and default) as the nominal target criterion, linear regression, gradient boosting (default settings), SVM (Support Vector Machine), MBR (Memory Based Reasoning), and rule induction (binary model as tree and cleanup models as neural) with the variables selected using the decision tree as inputs to predict the binary target Sle_yes_no (whether a patient has SLE or not: 0 for no and 1 for yes). Then we used the model comparison algorithm in SAS® Enterprise Miner™ to compare the models according to the validation misclassification rate as the target variable is binary.

Decision tree (Gini) as the nominal target criterion turned out be the champion model with a validation misclassification rate of 0.31140.

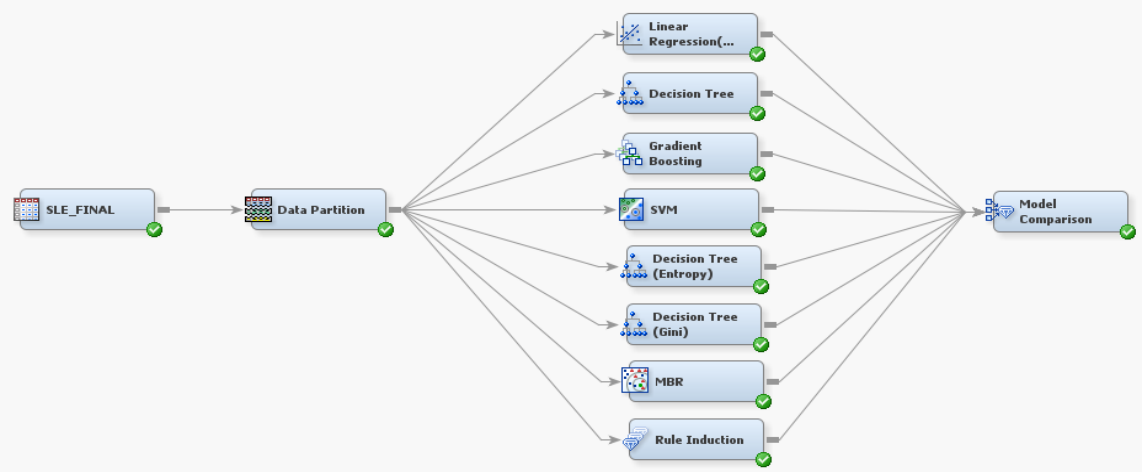

Figure 3. Model Comparison

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree3	Decision Tree (Gini)	0.31140	0.19661	0.29906	0.21267
	Tree2	Decision Tree (Entropy)	0.31287	0.19538	0.29781	0.21490
	Tree	Decision Tree	0.32164	0.21289	0.32038	0.21458
	Reg2	Linear Regression(Stepwise)	0.32310	0.20848	0.31661	0.21129
	SVM	SVM	0.32895	0.21167	0.33041	0.21140
	Rule	Rule Induction	0.32895	0.21282	0.32163	0.22597
	Boost	Gradient Boosting	0.33480	0.21007	0.32163	0.21455
	MBR	MBR	0.33918	0.20591	0.32100	0.22647

Table 9. Model selection for SLE

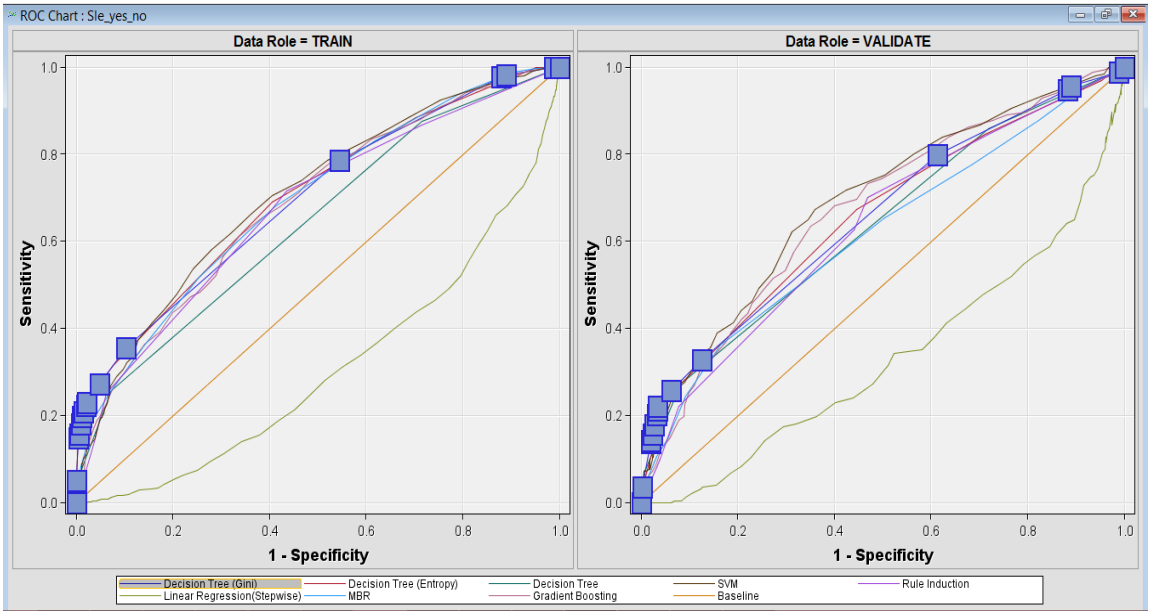

Figure 4. ROC Chart

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
age_in_years		6	1.0000	0.9052	0.9052
LOS		5	0.9889	0.7526	0.7611
total_charges		6	0.8860	1.0000	1.1287
PATIENT_TYPE_ID_0001		1	0.5492	0.6410	1.1671
CENSUS_REGION		1	0.3162	0.4441	1.4046

Table 10. Variable Importance

Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation
SIe_yes_no	_NOBS_	Sum of Frequencies	1595	684
SIe_yes_no	_MISC_	Misclassification Rate	0.29906	0.311404
SIe_yes_no	_MAX_	Maximum Absolute Error	0.923077	1
SIe_yes_no	_SSE_	Sum of Squared Errors	627.1902	290.9277
SIe_yes_no	_ASE_	Average Squared Error	0.196611	0.212666
SIe_yes_no	_RASE_	Root Average Squared Error	0.443409	0.461158
SIe_yes_no	_DIV_	Divisor for ASE	3190	1368
SIe_yes_no	_DFT_	Total Degrees of Freedom	1595	.

Table 11. Fit Statistics

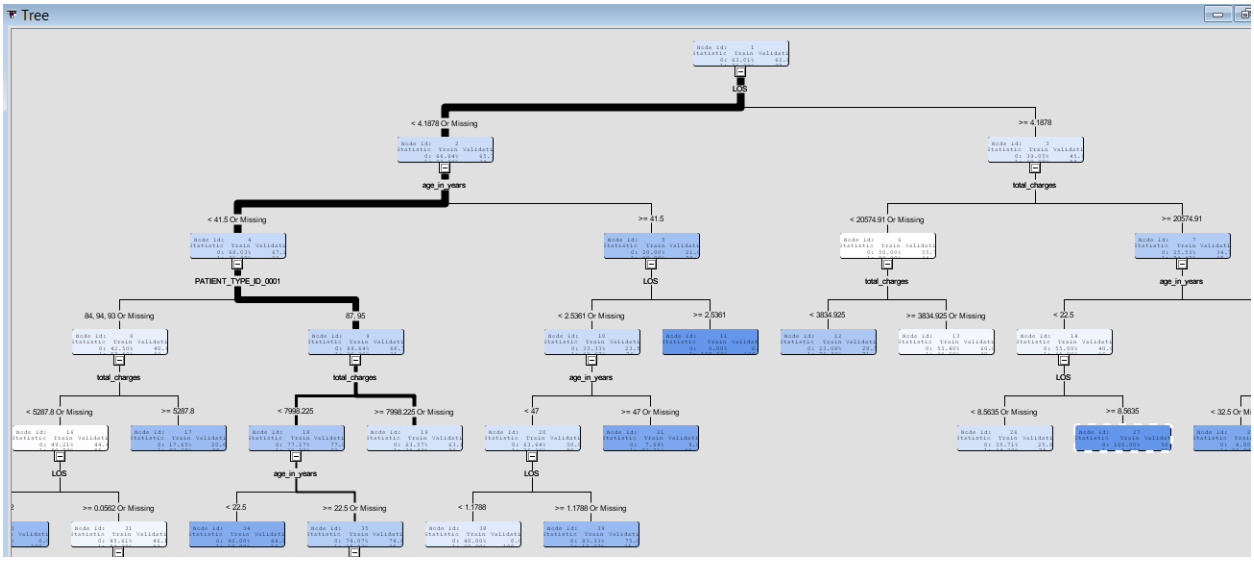

Figure 5. Decision Tree

```
if total_charges >= 20574.9
AND age_in_years < 37.5 AND age_in_years >= 32.5 or MISSING
AND LOS >= 4.18785
AND CENSUS_REGION IS ONE OF: SOUTH or MISSING
then
 Tree Node Identifier   = 73
 Number of Observations = 10
 Predicted: Sle_yes_no=1 = 0.70
 Predicted: Sle_yes_no=0 = 0.30
```

Figure 6. Rules

```
if total_charges >= 5287.8
AND age_in_years < 41.5 or MISSING
AND PATIENT_TYPE_ID_0001 IS ONE OF: 84, 94, 93 or MISSING
AND LOS < 4.18785 or MISSING
then
 Tree Node Identifier   = 17
 Number of Observations = 17
 Predicted: Sle_yes_no=1 = 0.82
 Predicted: Sle_yes_no=0 = 0.18
```

```
if total_charges < 7.765
AND age_in_years < 41.5 AND age_in_years >= 22.5 or MISSING
AND PATIENT_TYPE_ID_0001 IS ONE OF: 87, 95
AND LOS < 4.18785 or MISSING
then
 Tree Node Identifier   = 56
 Number of Observations = 6
 Predicted: Sle_yes_no=1 = 0.83
 Predicted: Sle_yes_no=0 = 0.17
```

Figure 7. Rules

PREDICTING FOR RA PATIENTS

Similarly data set for RA patients was imported into SAS® Enterprise Miner™ 12.3. Then we used various models like decision tree (gini, entropy, and default) as the nominal target criterion, linear regression, gradient boosting (default settings), SVM (Support Vector Machine), MBR (Memory Based Reasoning), and rule induction (binary model as tree and cleanup models as neural) with the variables selected using the decision tree as inputs to predict the binary target Ra_yes_no (whether a patient has RA or not: 0 for no and 1 for yes). Likewise we used the model comparison algorithm in SAS® Enterprise Miner™ to compare the models according to the validation misclassification rate.

Even for predicting RA patients, Decision tree (Gini) as the nominal target criterion turned out be the champion model with a validation misclassification rate of 0.29423. The English rules we analyzed to get a clear insight of the model.

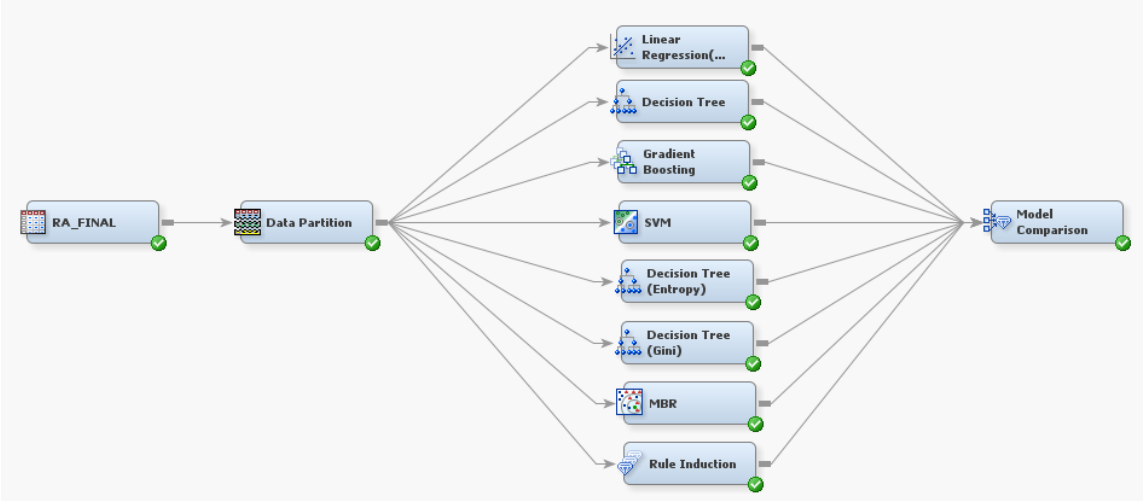

Figure 8. Model Comparison

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree3	Decision Tree (Gini)	0.29423	0.20255	0.29157	0.20386
	Tree	Decision Tree	0.29609	0.20519	0.29398	0.20710
	Rule	Rule Induction	0.29609	0.20830	0.29398	0.20932
	Reg2	Linear Regression(Stepwise)	0.29981	0.19759	0.29960	0.19955
	Tree2	Decision Tree (Entropy)	0.30168	0.19426	0.29076	0.20650
	Boost	Gradient Boosting	0.31099	0.20243	0.31968	0.20797
	SVM	SVM	0.31099	0.21155	0.30683	0.21241
	MBR	MBR	0.34637	0.20755	0.31807	0.23084

Table 12. Model Selection for RA

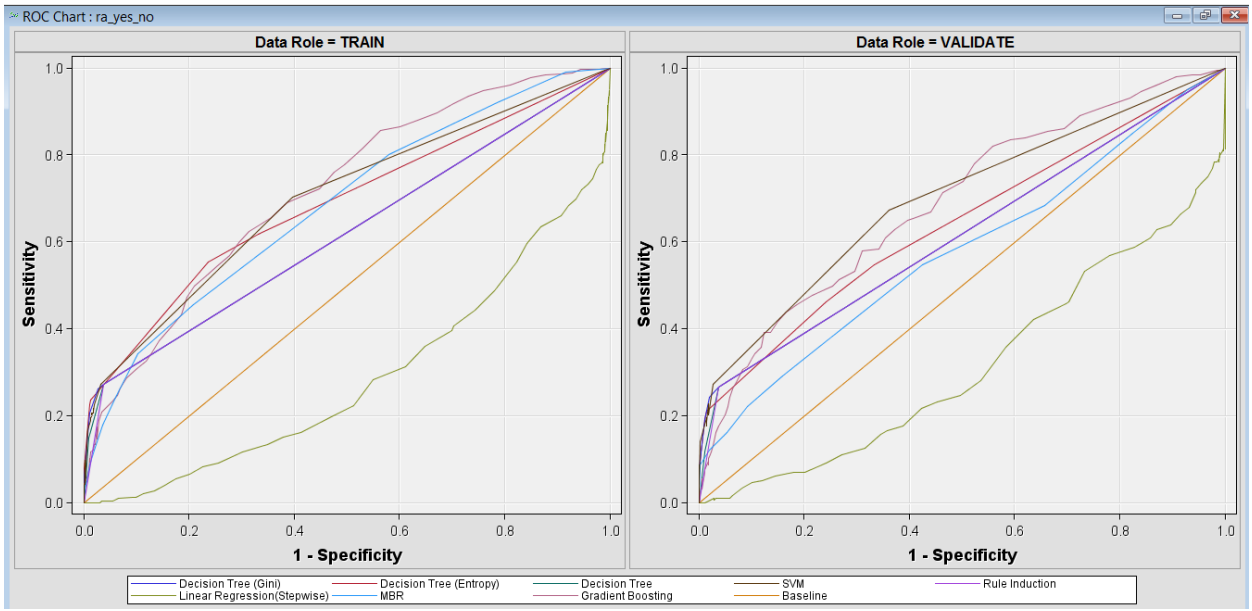

Figure 9. ROC Chart

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
age_in_years		2	1.0000	1.0000	1.0000
LOS		2	0.7400	0.5914	0.7992
total_charges		1	0.2600	0.2499	0.9610
CENSUS_REGION		1	0.1884	0.2445	1.2980

Table 13. Variable Importance

Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation
ra_yes_no	_NOBS_	Sum of Frequencies	1245	537
ra_yes_no	_MISC_	Misclassification Rate	0.292369	0.288641
ra_yes_no	_MAX_	Maximum Absolute Error	0.912281	0.818182
ra_yes_no	_SSE_	Sum of Squared Errors	507.5975	215.6781
ra_yes_no	_ASE_	Average Squared Error	0.203854	0.200818
ra_yes_no	_RASE_	Root Average Squared Error	0.451502	0.448127
ra_yes_no	_DIV_	Divisor for ASE	2490	1074
ra_yes_no	_DFT_	Total Degrees of Freedom	1245	.

Table 14. Fit Statistics

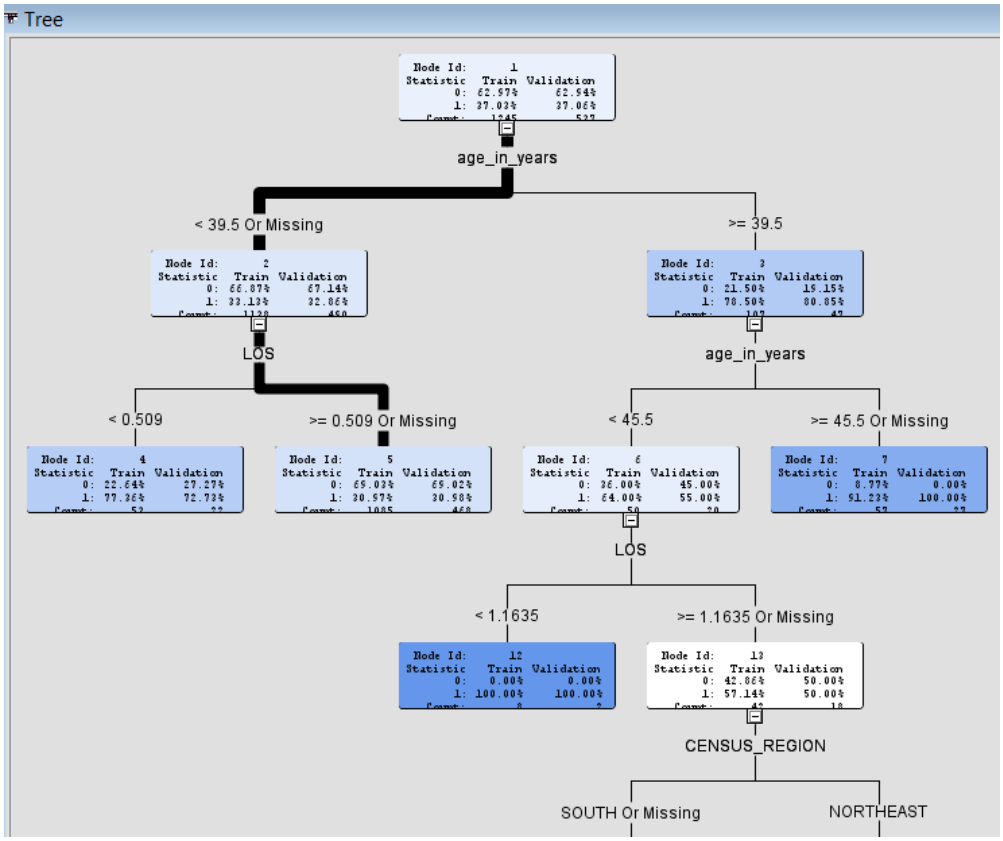

Figure 10. Decision Tree

```
if total_charges >= 10462.5
AND age_in_years < 45.5 AND age_in_years >= 39.5
AND LOS >= 1.16354 or MISSING
AND CENSUS_REGION IS ONE OF: SOUTH or MISSING
then
 Tree Node Identifier   = 39
 Number of Observations = 8
 Predicted: ra_yes_no=1 = 0.75
 Predicted: ra_yes_no=0 = 0.25
```

```
if age_in_years < 45.5 AND age_in_years >= 39.5
AND LOS >= 1.16354 or MISSING
AND CENSUS_REGION IS ONE OF: NORTHEAST
then
 Tree Node Identifier   = 25
 Number of Observations = 23
 Predicted: ra_yes_no=1 = 0.70
 Predicted: ra_yes_no=0 = 0.30
```

Figure 11. Rules

CONCLUSIONS AND FUTURE RESEARCH

For SLE patients, according to the rules of the decision tree, pregnant woman with total charges less than $7,765 and aged less than 41.5 years and be either an inpatient or obstetric patient, and with a length of stay less than 4.18 day have 83% chance of being an SLE patient.

Similarly if a pregnant woman with total charges greater than or equal to $10,462.5 and aged between 39.5 and 45.5 years, and length of stay greater than or equal to 1.16, and residing in south region, have a 75% chance of being a RA patient.

Future extension of this project will involve predicting the pregnancy outcomes in women with SLE and RA. If possible we may also expand the disease range to predict APS (Anti Phospholipid Antibody Syndrome) and DM (Diabetes Mellitus) in pregnant women, and also predict the adverse outcomes of pregnancy in them.

REFERENCES

1. Munnangi, H. Chakraborty, G. 2015. “Predicting Readmission of Diabetic Patients using the high performance Support Vector Machine algorithm of SAS® Enterprise Miner™” 3254.
2. Uramoto, KM. Michet, CJ Jr, Thumboo, J. Sunku, J. O’Fallon, WM. Gabriel, SE. 1999. “Trends in the incidence and mortality of systemic lupus erythematosus, 1950–1992.” *Arthritis Rheum*, 42:46–50.
3. Symmons, D. Turner,G. Webb, R. Asten, P. Barrett, E. Lunt, M. et al. 2002. “The prevalence of rheumatoid arthritis in the United Kingdom: new estimates for a new century.” *Rheumatology (Oxford)*, 41:793–800.
4. Yasmeen, S. Wilkins, EE. Field, NT. Sheikh, RA. Gilbert, WM. 2001. “Pregnancy outcomes in women with systemic lupus erythematosus.” *J Matern Fetal Med*, 10:91–6.
5. Chakravarthy, E. Nelson, L. and Krishnan, E. 2006. “Obstetric Hospitalizations in the United States for Women with Systemic Lupus Erythematosus and Rheumatoid Arthritis.” *Arthritis & Rheumatism*, 54:899–907.
6. Skomsvoll, JF. Ostense, M. Irgens, LM. Baste, V. 2000. “Pregnancy complications and delivery practice in women with connective tissue disease and inflammatory rheumatic disease in Norway.” *Acta Obstet Gynecol Scand*, 79:490–5.
7. Nossent, HC. Swaak, TJ. 1990. “Systemic lupus erythematosus. IV. Analysis of the interrelationship with pregnancy.” *J Rheumatol*, 17: 771–6.
8. Gimovsky, ML. Montoro, M. Paul, RH. 1985. “Pregnancy outcome in women with systemic lupus erythematosus.” *Obstet Gynecol*, 63:686–92.
9. Chawla, VN. et al. 2004. “Data Mining for Imbalanced Data sets: An Overview.” *Data Mining and Knowledge Discovery Handbook*, 40:853-867.
10. Getting Started with SAS® Enterprise Miner™ 7.1. “Create a Gradient Boosting Model of the Data.” Accessed January 18, 2016. http://support.sas.com/documentation/cdl/en/emgsj/64144/HTML/default/viewer.htm#p03iy98sk0c9bvn1r6x7ppx8uj08.htm
11. Deutsch, G. 2010. “Overrepresentation – “SAS”-Oversampling. Accessed January 18, 2016. http://www.data-mining-blog.com/tips-and-tutorials/overrepresentation-oversampling/