

## An Analysis of Diabetes Risk Factors Using Data Mining Approach

Akkarapol Sa-ngasoongsong and Jongsawas Chongwatpol  
Oklahoma State University, Stillwater, OK 74078, USA

### ABSTRACT

Preventing the disease of diabetes is an ongoing area of interest to the healthcare community. Although many studies employ several data mining techniques to assess the leading causes of diabetes, only small sets of clinical risk factors are considered. Consequently, not only many potentially important variables such as pre-diabetes health conditions are neglected in their analysis, but the results produced by such techniques may not represent relevant risk factors and pattern recognition of diabetes appropriately. In this study, we categorize our analysis into three different focuses based on the patients' healthcare costs. We then examine whether more complex analytical models using several data mining techniques in SAS® Enterprise Miner™ 7.1 can better predict and explain the causes of increasing diabetes in adult patients in each cost category. The preliminary analysis shows that high blood pressure, age, cholesterol, adult BMI, total income, sex, heart attack, marital status, dental checkup, and asthma diagnosis are among the key risk factors.

### 1. INTRODUCTION

Preventing the disease of diabetes is an ongoing area of interest to the healthcare community. Based on the data from the 2011 National Diabetes Fact Sheet, diabetes affects an estimate of 25.8 million people in the US, which is about 8.3% of the population. Additionally, approximately 79 million people have been diagnosed with pre-diabetes [1]. Pre-diabetes refers to a group of people with higher blood glucose levels than normal but not high enough for a diagnosis of diabetes.

Increased awareness and treatment of diabetes should begin with prevention. Much of the focus has been on the impact and importance of preventive measures on disease occurrence and especially cost savings resulted from such measures. Many studies regarding diabetes prediction have been conducted for several years. The main objectives are to predict what variables are the causes, at high risk, for diabetes and to provide a preventive action toward individual at increased risk for the disease. Several variables have been reported in literature as important indicators for diabetes prediction.

Lindstrom and Tuomilehto (2003) develop the diabetes risk score model considering Age, BMI, waist circumference, history of antihypertensive drug treatment, high blood glucose, physical activity, and daily consumption of fruits, berries, or vegetables as categorical variables [2]. Park and Edington (2001) present a sequential neural network model for diabetes prediction. The authors indicate risk factors, in the final model, including blood pressure, cholesterol, back pain, fatty food, weight index or alcohol index [3]. Concaro et al, (2009) present the application of a data mining technique to a sample of diabetic patients. They consider the clinical variables such as BMI, blood pressure, glycaemia, cholesterol, or cardio-vascular risk in the model [4].

Although these studies employ several data mining techniques to assess the leading causes of diabetes, only small sets of clinical risk factors are considered. Consequently, not only many potentially important variables such as pre-diabetes health conditions are neglected in their analysis, but the results produced by such techniques may not represent relevant risk factors and pattern recognition of diabetes appropriately.

This study seeks to fill this gap. Specially, the question arises "What are the most important risk factors to be included in prognostic analysis to prevent prevalence of diabetes?" To answer this research question, we examine whether more complex analytical models using several data mining techniques can better predict and explain the causes of increasing diabetes. In this study, we follow the CRISP-DM Model (Cross Industry Standard Process for Data Mining), which is used as a comprehensive data mining methodology and process model for conducting this data mining study. CRISP-DM breaks down this data mining project in to six phases: business understanding, data understanding, data preparation, modeling, evaluation, and development.

## 2. RESEARCH FRAMEWORK

We provide the in-depth analysis on how data mining approach can be a great help. After understanding the domain of diabetes and developing the objectives of achieving prognostic analysis of diabetes through data mining approach, we begin our analysis by understanding the relevant data source, accessing data quality, and discovering first insights into the data. The next step is toward data preprocessing from the initial raw data to the final dataset, ready for the model development. This preprocessing step takes about 90% of time to clean, transform, construct, and format the relevant data. We then apply analytical data mining techniques to predict and explain factors that increase the prevalence of diabetes in the patient samples. However, we need to evaluate and assess the validity and the utility of our developed predictive models before deploying the data mining results into the domain as stated in the objectives of the study. Figure 1 presents the overall framework of our models to address the research question from the data understanding to model deployment.

### 2.1 Data Description

The data source that is used to perform data mining analysis in this study is provided by SAS in the national 2010 SAS Data Mining Shootout competition. With 50,788 records, the dataset consists of 43 variables in which 35 variables are discrete variables and the other eight variables are continuous variables. This dataset is assumed to be representative of the population and used for analysis as a snapshot of the country and its health care costs at a point in time.

Our first task in this study is to get a sense of the dataset for any inconsistencies, errors, or extreme values in the data. Frequency distribution, descriptive statistics, and cross-tab analysis are used in this section. Table 1 presents the summary of inapplicable data for nominal variables. Below are the key findings, which are very important for data preparation in the next section:

**Table 1: Summary of inapplicable data for nominal variables**

Variable Name	Level	% of inapplicable data	Frequency
CENSUS_REGION	Nominal	0.01%	5
SEX	Nominal	0.00%	0
MARITAL_STATUS	Nominal	0.00%	0
YEARS_EDUC	Nominal	10.50%	5334
HIGHEST_DEGREE	Nominal	0.41%	209
SERVED_ARMED_FORCES	Nominal	0.90%	459
FOODSTAMPS_PURCHASE	Nominal	0.95%	483
MORE_THAN_ONE_JOB	Nominal	54.38%	27619
WEARS_EYEGASSES	Nominal	1.25%	634
PERSON_BLIND	Nominal	95.58%	48544
WEAR_HEARING_AID	Nominal	1.24%	628
IS_DEAF	Nominal	95.09%	48296
DENTAL_CHECK_UP	Nominal	0.63%	318
CHOLEST_LST_CHK	Nominal	33.38%	16951
LAST_CHECKUP	Nominal	31.21%	15849
LAST_FLUSHOT	Nominal	30.91%	15697
LOST_ALL_TEETH	Nominal	29.66%	15064
LAST_PSA	Nominal	82.95%	42130
LAST_PAP_SMEAR	Nominal	63.68%	32340
LAST_BREAST_EXAM	Nominal	63.59%	32294
LAST_MAMMOGRAM	Nominal	71.63%	36380
BLD_STOOL_TST	Nominal	31.30%	15795
SIGMOIDOSCOPY	Nominal	30.69%	15588
WEAR_SEAT_BELT	Nominal	26.65%	13537
ASTHMA_DIAGNOSIS	Nominal	0.04%	22
HIGH_BLOOD_PRESSURE_DIAG	Nominal	26.69%	15078
HEART_DISEASE_DIAG	Nominal	29.73%	15097
ANGINA_DIAGNOSIS	Nominal	29.70%	15086
HEART_ATTACK_DIAG	Nominal	29.66%	15062
OTHER_HEART_DISEASE	Nominal	29.73%	15097
STROKE_DIAGNOSIS	Nominal	29.62%	15043
EMPHYSEMA_DIAGNOSIS	Nominal	29.61%	15039
JOINT_PAIN	Nominal	29.73%	15101
CURRENTLY_SMOKE	Nominal	38.81%	19710
DIABETES_DIAG_BINARY	BINARY	0.00%	0

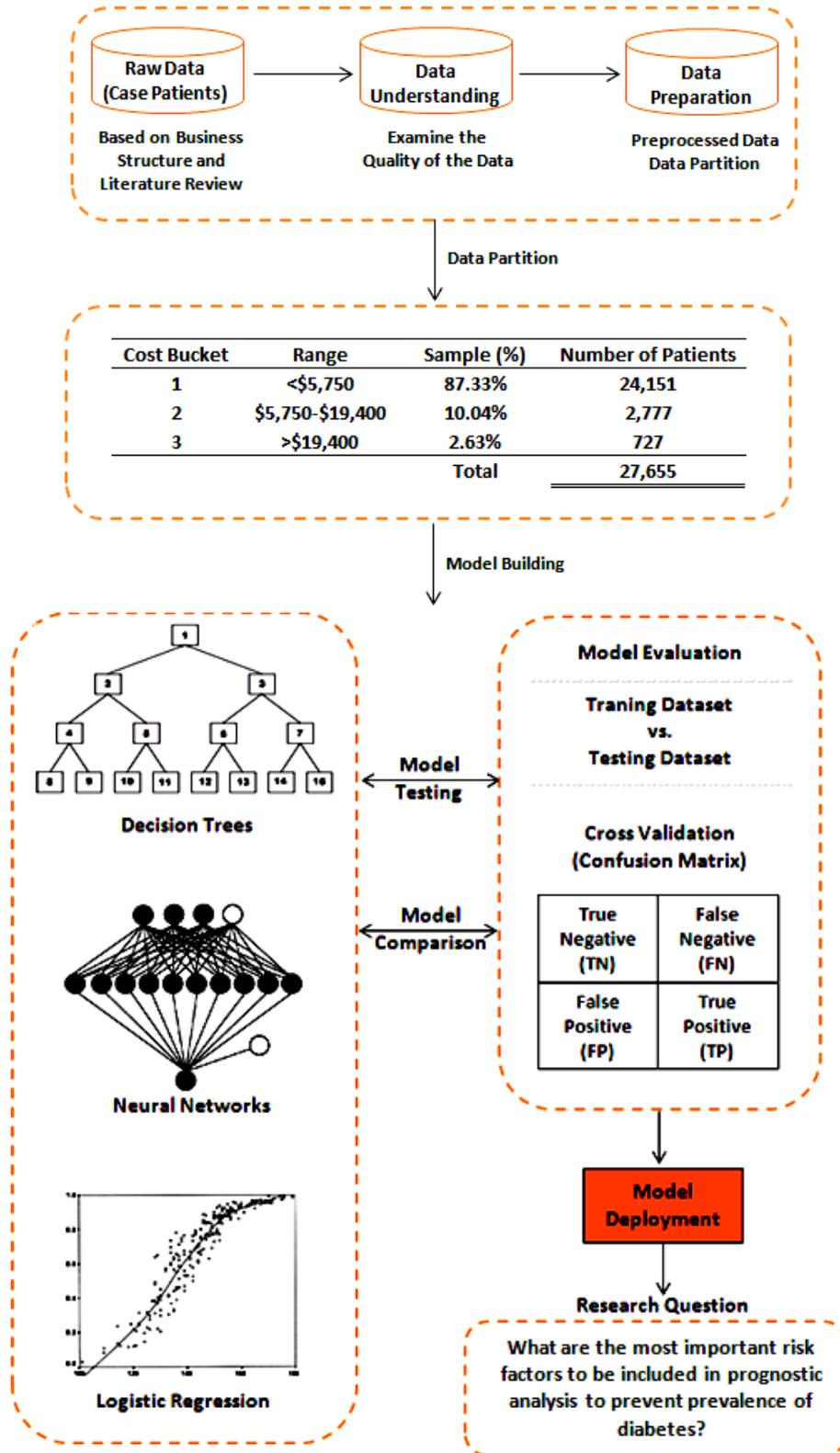


Figure 1: Research Framework

Below are the key findings, which are very important for data preparation in the next section:

- LAST\_PSA, LAST\_PAP\_SMEAR, LAST\_BREAST\_EXAM, LAST MAMMOGRAM are only taken by one type of gender (Deleted).
- More than 95% of the data in PERSON\_BLIND and IS\_DEAF are inapplicable.
- 34.96% of patients with diabetes have no educational degree, compared to 17.81% of patients without diabetes
- 81.11% of patients with diabetes wear eyeglasses, compared to 44.58% of patients without diabetes
- 84.24 % of patients with diabetes checked their cholesterol within last year, compared to 31.79 % of patients without diabetes
- 83.20% of patients with diabetes checked their health within last year, compared to 38.92% of patients without diabetes
- 52.07% of patients with diabetes had flu-shot within last year, compared to 16.55% of patients without diabetes
- 32.48% of patients with diabetes had sigmoidoscopy or colonoscopy test, compared to 10.74% of patients without diabetes
- 63.23% of patients with diabetes have been diagnosed with high blood pressure, compared to 13.86 % of patients without diabetes
- 13.40% of patients with diabetes have been diagnosed with coronary heart disease diagnosis, compared to 1.57% of patients without diabetes

Based on these findings, we can observe that most diabetes patients have been diagnosed with other concurrent diseases such as high blood pressure, heart disease, and cholesterols, etc.

## 2.2 Data Preparation

The next step in this analysis is to examine the quality of the data. We would like to figure out whether or not the data is complete or has missing values and what variables to be included in the model. We make an assumption that the model excludes variables with over 50% information missing; see similar methodology from Park and Edington (2001), [3]. Thus, the following seven variables are excluded from our analysis: (1) MORE\_THAN\_ONE\_JOB, (2) PERSON\_BIND, (3) IS\_DEAF, (4) LAST\_PSA, (5) LAST\_PAP\_SMEAR, (6) LAST\_BREAST\_EXAM, and (7) LAST\_MAMMOGRAM.

Below are two examples on how we analyze those excluded variables. Note that, besides looking at the descriptive statistics, we have actually tested whether or not these seven excluded variables are significant in the models (overall decision tree model or even the model by age and gender). Due to the high-missing-values issue, the results show that these seven variables are not included in any models tested.

- For instance, a variable “PERSON\_BLIND” has 48,544 missing observations (approximately 95% missing values) and out of its applicable data, there are only 37 observations indicating that those patients have been diagnosed with diabetes and blindness. Even though there has been claimed that roughly 40% of patients diagnosed with diabetes in the United States have some form of diabetic retinopathy [5], we still exclude that variable from our analysis due to the relatively-low relevant amount of data in this dataset.
- Another critical variable to be excluded from our analysis is “MORE\_THAN\_ONE\_JOB.” There have been several studies on how unhealthy eating or sleeping habits (sedentary life styles) for those people that work more than one job increases the risk factor for diabetes [6]. Only approximately less than 0.3% (68 out of 23,169 observations excluding all non-applicable data) indicates the group of patients who have been diagnosed with diabetes and worked more than one job (see Figure 2). Thus, we decide to exclude this variable from our analysis. Similar reasons for excluding such variable can be seen for the other five variables.

		PERSON IS BLIND			Total	
		-1	1	2		
DIABETES DIAG BINARY	0	Frequency	46349	113	1737	48199
	Row Pct	96.16	0.23	3.60		
	Col Pct	95.48	75.33	82.95		
1	Frequency	2195	37	357	2589	
	Row Pct	84.78	1.43	13.79		
	Col Pct	4.52	24.67	17.05		
Total	Frequency	48544	150	2094	50788	

		HAS MORE THAN ONE JOB			Total	
		-1	1	2		
DIABETES DIAG BINARY	0	Frequency	25990	1823	20386	48199
	Row Pct	53.92	3.78	42.30		
	Col Pct	94.10	96.40	95.81		
1	Frequency	1629	68	892	2589	
	Row Pct	62.92	2.63	34.45		
	Col Pct	5.90	3.60	4.19		
Total	Frequency	27619	1891	21278	50788	

Figure 2: Cross-Tab Analysis (Diabetes by PERSON\_BIND and Diabetes by MORE\_THAN\_ONE\_JOB)

Additionally, including such variables with high-missing values in the model or even applying missing value imputation method can lower the quality of our findings. In fact, we believe that the dataset is still quite large enough (over 30,000 observations) for the analysis after excluding these variables and any other missing values.

### 2.3 Missing Values, Possible Outliers, and Data Transformation

After excluding suspected variables, we decide to impute missing values for better model prediction and to increase the total number of observations with diabetes. Although, imputation can bias the model prediction, omitting these missing values in the sample dataset may also produce an extremely biased prediction when applied to the new dataset. Another issue that should be addressed is the possible outliers in the data set. Some of the patients in the data set have income lower than zero. Also some of the most expensive patients have the total health care cost, Medicaid, or Medicare within very extreme percentile. In our model, we decide to filter the patients with these outliers out of the data set. Consequently, a total of 27,655 observations are ready for modeling analysis.

For interval input variables, the missing values are replaced with the mean of the non-missing values. In contrast, the missing values for nominal variables are replaced with the most frequently occurring class variable value. Initially, we decide to transform the variables, which look right skewed, in order to improve the fit of a model to the data. We apply “Maximize Normality” method to the following six variables: AMOUNT\_PAID\_MEDICAID, AMOUNT\_PAID\_MEDICARE, NUMBER\_VISITS, ADULT\_BMI, TOTALEXP, and TOTAL\_INCOME. However, it seems that only TOTAL\_INCOME and NUMB\_VISITS show significant improvement after the transformation, while the distribution of other variables remain closely the same. Thus, only TOTAL\_INCOME and NUMB\_VISITS variables are transformed for further analysis (see Figure 3).

Source	Method	Variable Name	Formula	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
Input	Original	NUMB_VISITS		2969	0	218	7.625126	12.90659	5.541453	51.77834
Input	Original	TOTAL_INCOME		2969	0	217950	23587.71	25328.7	2.714735	11.09394
Output	Computed	LOG_NUMB_VISITS	$\log(\text{SCALEVAR}_+ + 1)$	2969	0	0.693147	0.032993	0.050703	4.2769	29.59589
Output	Computed	SQRT_TOTAL_INCOME	$\text{Sqrt}(\text{SCALEVAR}_)$	2969	0	1	0.286132	0.162366	0.562733	1.051097

Figure 3: Variable Transformation Statistics

## 3. PREDICTION MODEL

In this study, three popular data mining techniques including logistic regression, decision tree, and artificial neural network are applied and compared to each other based on their predictive accuracy on the hold-out sample.

- Logistic regression is often used to predict an outcome variable that is binary or multi-class dependent variables. It allows the prediction of discrete variables (dependent variables) by a mix of continuous and discrete predictors as the relationship between dependent variables and independent variables is non-linear. It builds the model to predict the odds of its occurrence instead of point estimate event in the traditional linear regression model.
- Decision tree is another data classification and prediction method commonly used due to its intuitive explainability characteristics. Decision tree divides the dataset into multiple groups by evaluating individual data record, which can be described by its attributes. It is also simple and easy to visualize the process of classification where the predicates return discrete values and can be explained by a series of nested if-then-else statements.
- Artificial Neural Network (ANN) is a mathematical and computational model for pattern recognition and data classification through a learning process. It is a biologically inspired analytical technique, simulating biological systems, where learning algorithm indicates how learning takes place and involves adjustments to the synaptic connections between neurons. Data input can be discrete or real valued; meanwhile the output is in a form of vector of values and can be discrete or real valued as well.

For a technical summary including both algorithm and its applications of each method in medical and health care areas see [7-11].

### 3.1 Cost Bucketing and Model Classification

In this study, we categorize our analysis into three different focuses based on the patients' healthcare costs. We partition our patients into these three groups (see Table 2) in such a way that the sum of all patients' healthcare costs in each group is approximately the same. The range of healthcare costs is varied significantly. According to the population's cumulative healthcare expenses, approximately 87% of the overall expense of the population originates from only 3% of the most expensive patients.

- The cost bucket #1 can be interpreted as candidates for the low-risk group of patients diagnosed with diabetes.
- The cost bucket #2 can be interpreted as the pre-diabetes group.
- The cost bucket #3 can be interpreted as the high-risk group of patients diagnosed with diabetes.

**Table 2: Cost bucket information**

Cost Bucket	Range	Sample (%)	Number of Patients
1	<\$5,750	87.33%	24,151
2	\$5,750-\$19,400	10.04%	2,777
3	>\$19,400	2.63%	727
Total			27,655

We decide to stratify our samples to construct a model set with approximately equal numbers of each target value (DIABETES\_DIAG\_BINARY) in each cost bucket. With a 50% adjusting for oversampling, we believe that the contrast between the two values is minimized, which make it easier and more reliable to recognize patterns in this dataset. The most important goals of this study are to predict the outcome probability of people with diabetes. In order to provide more accurate diabetes prediction, we apply and compare ANN, Logistic Regression, and Decision tree models in each cost bucket as well.

### 3.2 Performance Measures

The complicity of the model is controlled by fit statistics calculated on the testing dataset. We use three different criteria to select the best model on the testing dataset. These criteria include false negative, prediction accuracy, and misclassification rate. False negative (Target = 1 and Outcome = 0) represents the case of an error in the model prediction where model results indicate that diabetes occurrence is not present, when in reality, there is an incident. The false negative value should be as low as possible. The proportion of cases misclassified is very common in the predictive modeling. However, the observed misclassification rate should be also relatively low for model justification. Lastly, prediction accuracy is evaluated among the three models on the testing dataset. The higher the prediction accuracy rate, the better the model to be selected.

## 4. RESULTS AND DISCUSSIONS

After excluding variables with outliers and high missing values, we first develop predictive models on the original sample dataset without categorizing our sample into three cost-bucketing groups as presented in Table 2. The results of the overall model are compared to those from each cost-bucketing group to determine whether important risk factors of diabetes are different. The dataset is allocated to the training (70%) and testing (30%) partitions. The binary variable of patients with diabetes (Target = 1 for patients with diabetes and Outcome = 0 for patients without diabetes) is the output variable of the prediction models. After recoding all categorical input variables, the selected variables are tested whether the association between the input variables and the logit of binary target variable satisfy the linearity assumption. The problematic variables are then transformed to satisfy such assumption. Different models are constructed and compared in order to predict patients with diabetes.

Training dataset is only used to extract models by the data mining algorithms. Then, those models derived in the training data set are then applied on the testing dataset for the correct discovery of intrusions. In other words, this testing dataset is used to prune the models generated by the data mining process in the training dataset to avoid overfitting and instabilities in the classification accuracy. Statistical analyses are performed using SAS® Enterprise Guide 4.3 for data preparation and SAS® Enterprise Miner™ 7.1 for model development and comparison.

#### 4.1 Overall Model

Logistic regression produces the best results with overall misclassification rate of 22.89%. Although Artificial Neural Networks (ANN) has the lowest false negative rate of 20.55%, we still select the logistic regression model as our final model to predict patients with diabetes. After performing the stepwise logistic regression, which is a process of building a model where the choice of predictive variables is carried out based on the t-statistics of their estimated coefficients, the final model is presented in Figure 4. There are total 15 risk factors used to predict the prevalence of diabetes.

Classification Table					
Data Role=TRAIN Target Variable=DIABETES_DIAG_BINARY					
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	78.5303	75.7470	1090	37.8735
1	0	21.4697	20.7088	298	10.3544
0	1	23.4228	24.2530	349	12.1265
1	1	76.5772	79.2912	1141	39.6456

Data Role=VALIDATE Target Variable=DIABETES_DIAG_BINARY					
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	78.4380	74.7573	462	37.3786
1	0	21.5620	20.5502	127	10.2751
0	1	24.1113	25.2427	156	12.6214
1	1	75.8887	79.4498	491	39.7249

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	HIGH_BLOOD_PRESSURE_DIAG	1	1	599.6806		<.0001
2	AGE	1	2	250.6042		<.0001
3	CHOLEST_LST_CHK	5	3	136.1832		<.0001
4	ADULT_BMI	1	4	95.8065		<.0001
5	NUMB_VISITS	1	5	36.0891		<.0001
6	HIGHEST_DEGREE	6	6	44.3947		<.0001
7	LAST_FLUSHOT	5	7	25.4388		0.0001
8	HEART_ATTACK	1	8	12.2377		0.0005
9	DENTAL_CHECKUP	3	9	13.7331		0.0033
10	WEARS_EYEGLASSES	1	10	9.4814		0.0021

Figure 4: Classification Table and Summary of Stepwise Selection (Overall Model)

#### 4.2 Cost Bucket #1 (<\$5,750)

For the low-risk group (Cost Bucket #1), Decision tree produces the best results with overall misclassification rate of 24.56%, followed by the Logistic regression and artificial neural networks with misclassification rates of 25.43% and 25.68%, respectively. Figure 5 presents the final variable selection and classification table for Decision Tree model. Seven variables including age, cholesterol, adult BMI, high blood pressure, total income, sex, and asthma diagnosis are important factors to predict diabetes in this cost bucket group. Compared to the overall model, Only age, high blood pressure, and cholesterol are the common risk factors.

Classification Table					
Data Role=TRAIN Target Variable=DIABETES_DIAG_BINARY					
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	81.7204	72.9211	684	36.4606
1	0	18.2796	16.3113	153	8.1557
0	1	24.4466	27.0789	254	13.5394
1	1	75.5534	83.6887	785	41.8443

Data Role=VALIDATE Target Variable=DIABETES_DIAG_BINARY					
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	79.3696	68.7345	277	34.3672
1	0	20.6304	17.8660	72	8.9330
0	1	27.5711	31.2655	126	15.6328
1	1	72.4289	82.1340	331	41.0670

Obs	NAME	NRULES	IMPORTANCE	VIMPORTANCE	RATIO
1	AGE	1	1.00000	1.00000	1.00000
2	CHOLEST_LST_CHK	2	0.57768	0.79885	1.38285
3	ADULT_BMI	2	0.45938	0.29888	0.65061
4	HIGH_BLOOD_PRESSURE_DIAG	2	0.45176	0.57993	1.28371
5	TOTAL_INCOME	1	0.21563	0.23269	1.07913
6	SEX	1	0.13927	0.00000	0.00000
7	ASTHMA_DIAGNOSIS	1	0.12192	0.23284	1.90981

**Figure 5: Classification Table and Variable Importance (Cost Bucket #1)**

### 4.3 Cost Bucket #2 (\$5,750 - \$19,400)

For the pre-diabetes group, Logistic Regression produces the best results with overall misclassification rate of 29.74%, followed by artificial neural networks and decision tree, respectively. Figure 6 presents the final variable selection based on the stepwise regression approach and classification table for Logistic regression model. Eight variables including cholesterol, adult BMI, high blood pressure, highest degree, last flu shot, heart attack, wears eyeglasses, and marital status are important factors to predict diabetes in this cost bucket group. Compared to the overall model, adult BMI, high blood pressure, last flu shot, heart attack, wears eyeglasses, and marital status are the common risk factors. On the other hand, only adult BMI and high blood pressure are the common risk factors between this group and the cost bucket #1.

### 4.4 Cost Bucket #3 (>\$19,400)

For the high-risk group, Decision Tree produces the best results with overall misclassification rate of 41.80%, followed by artificial neural networks and Logistic Regression, respectively. Figure 7 presents the final variable selection and classification table for Decision Tree model. The final model indicates that adult BMI, age, wear eyeglasses, and marital status are among the key risk factors in this cost bucket group. Compared to the overall model, adult BMI, age, and, wears eyeglasses are the common risk factors. On the other hand, only adult BMI and age are the common risk factors between this group and the cost bucket #1; meanwhile adult BMI, wears eyeglasses, and marital status are the common risk factors compared to the cost bucket #2

Data Role=TRAIN Target Variable=DIABETES_DIAG_BINARY					
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	73.9011	73.4973	269	36.7486
1	0	26.0989	25.9563	95	12.9781
0	1	26.3587	26.5027	97	13.2514
1	1	73.6413	74.0437	271	37.0219

Data Role=VALIDATE Target Variable=DIABETES_DIAG_BINARY					
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	70.0000	70.8861	112	35.4430
1	0	30.0000	30.3797	48	15.1899
0	1	29.4872	29.1139	46	14.5570
1	1	70.5128	69.6203	110	34.8101

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	HIGH_BLOOD_PRESSURE_DIAG	1	1	92.8821		<.0001
2	ADULT_BMI	1	2	57.4458		<.0001
3	CHOLEST_LST_CHK	5	3	30.9138		<.0001
4	HIGHEST_DEGREE	6	4	30.7163		<.0001
5	LAST_FLUSHOT	5	5	16.7130		0.0051
6	HEART_ATTACK	1	6	5.9002		0.0151
7	WEARS_EYEGASSES	1	7	4.4965		0.0340
8	MARITAL_STATUS	4	8	11.3634		0.0228

Figure 6: Classification Table and Summary of Stepwise Selection (Cost Bucket #2)

## 5. CONCLUSION

This study demonstrates that data mining based approaches can be used to assess predictor variables influencing the risk of diabetes in adult patients. As opposed to the traditional descriptive statistical analysis methods or the approaches adopting only expert-selected variables, the employment of logistic regression, decision trees, or neural network models provide an interesting list of risk factors that some have been included in the existing studies; meanwhile some others have been absent from the related literatures. We categorize our analysis into three different focuses based on the patients' healthcare costs. Figure 8 presents the summary of the key findings of this study. Only adult BMI is the key risk factor among these four groups. Age is the most important factor for the overall model, cost buckets #1, and # 3; meanwhile, high blood pressure is the key indicator for the overall model, cost buckets #1 and #2. The results presented in Figure 8 confirm the previous literature to some extent. The following key risk factors: High blood pressure, adult BMI, cholesterol, and heart attack derived from our analysis are consistent with the studies from Lindstrom and Tuomilehto (2003), Park and Edington (2001), and Concaro et al. (2009). On the other hand, several interesting factors such as total income, asthma diagnosis, wears eyeglasses, and marital status reported to be critical in this study have not been included in the existing studies. A prospective study is needed to provide better understanding of the relationship between these undisclosed factors and the increased risk of diabetes.

Data Role=TRAIN Target Variable=DIABETES_DIAG_BINARY					
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	62.0321	87.2180	116	43.4457
1	0	37.9679	52.9851	71	26.5918
0	1	21.2500	12.7820	17	6.3670
1	1	78.7500	47.0149	63	23.5955

Data Role=VALIDATE Target Variable=DIABETES_DIAG_BINARY					
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	56.9444	69.4915	41	35.0427
1	0	43.0556	53.4483	31	26.4957
0	1	40.0000	30.5085	18	15.3846
1	1	60.0000	46.5517	27	23.0769

Variable Importance						
Obs	NAME	LABEL	NRULES	IMPORTANCE	VIMPORTANCE	RATIO
1	ADULT_BMI	ADULT BMI	1	1.00000	0.00000	0.00000
2	AGE		1	0.95316	0.90958	0.95428
3	WEARS_EYEGASSES	WEARS EYEGASSES	2	0.67493	0.00000	0.00000
4	MARITAL_STATUS	MARITAL STATUS	1	0.50110	1.00000	1.99560

**Figure 7: Classification Table and Variable Importance (Cost Bucket #3)**

Overall Model	Cost Bucket #1 (<\$5,750)	Cost Bucket #2 (\$5,750-\$19,400)	Cost Bucket #3 (>\$19,400)
ADULT_BMI	ADULT_BMI	ADULT_BMI	ADULT_BMI
High_Blood_PRESSURE_DIAG	High_Blood_PRESSURE_DIAG	High_Blood_PRESSURE_DIAG	
AGE	AGE		AGE
CHOLEST_LST_CHECK	CHOLEST_LST_CHECK		
NUMB_VISITS			
HIGHEST_DEGREE		HIGHEST_DEGREE	
LAST_FLUSHOT		LAST_FLUSHOT	
HEART_ATTACK		HEART_ATTACK	
DENTAL_CHECKUP			
WEARS_EYEGASSES		WEARS_EYEGASSES	WEARS_EYEGASSES
		MARITAL_STATUS	MARITAL_STATUS
	TOTAL_INCOME		
	SEX		
	ASTHMA_DIAGNOSIS		

**Figure 8: Summary of the Key Findings**

## REFERENCES

- [1] American Diabetes Association, "Diabetes Statistics," June 02, 2012, <<http://www.diabetes.org/diabetes-basics/diabetes-statistics/>>.
- [2] J. Lindstrom and J. Tuomilehto, "The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk," *Diabetes Care*, 26:3 (2003), 725-731.
- [3] J. Park and D. W. Edington, "A Sequential Neural Network Model for Diabetes Prediction," *Artificial Intelligence in Medicine*, 23 (2001), 277-293.
- [4] S. Concaro, L. Sacchi, C. Cerra, and M. Stefanelli, "Temporal Data Mining for the Assessment of the Costs Related to Diabetes Mellitus Pharmacological Treatment," *Proc. AMIA 2009 Symposium Proceedings*, 2009, pp. 119-123.
- [5] American Optometric Association, "Diabetes is the Leading Cause of Blindness Among Most Adults," July 11, 2010, <<http://www.aoa.org/x6814.xml>>.
- [6] Eurekalert, "Insufficient sleep may be linked to increased diabetes risk," July 11, 2010, <<http://lifesciencelog.com/cluster53092620/>>.
- [7] Jackson, J., Data mining: a conceptual overview. *Communications of the Association for Information Systems*, 2002. 8: p. 267-296.
- [8] Turban, E., R. Sharda, and D. Delen, *Decision Support and Business Intelligence Systems*. 2011, Pearson.
- [9] Delen, D., A. Oztekin, and Z.J. Kong, A machine learning-based approach to prognostic analysis of thoracic transplantations. *Artificial Intelligence in Medicine*, 2010. 49(1): p. 33-42.
- [10] Delen, D., A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 2010. 49(4): p. 498-506.
- [11] Delen, D., G. Walker, and A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*, 2005. 34(2): p. 113-127.

## CONTACT INFORMATION

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

Name: Akkarapol Sa-ngasoongsong  
Enterprise: Oklahoma State University  
Address: 245N. University Place#301  
City, State ZIP: Stillwater, Oklahoma, 74075, USA  
E-mail: [akkarap@okstate.edu](mailto:akkarap@okstate.edu)

Name: Jongsawas Chongwatpol  
Enterprise: Oklahoma State University  
Address: Spears School of Business  
City, State ZIP: Stillwater, Oklahoma, 74078, USA  
E-mail: [jongsaw@ostateemail.okstate.edu](mailto:jongsaw@ostateemail.okstate.edu)

Akkarapol Sa-ngasoongsong is a PhD student in School of Industrial Engineering and Management at Oklahoma State University. He has three years of professional experiences as R&D Engineer. He has earned two SAS certifications: SAS Certified Base Programmer for SAS® 9 and Certified Predictive Modeler using SAS® Enterprise Miner™ 6.1. He was 1<sup>st</sup> place award winner of the M2010 conference's Data Mining Shootout, and recipient of SAS ambassador honorable mention award (SGF 2012)

Jongsawas Chongwatpol has earned PhD degree in Management Science and Information Systems from Oklahoma State University. He was 1<sup>st</sup> place award winner of the M2010 conference's Data Mining Shootout. He was also a recipient of the best teaching award from Oklahoma State University, and 2011 best cast study paper award from Decision Sciences Institute (DSI).

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

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