NAFLD Study Summary

Nan Zhang

August 14th, 2015

1 Introduction

The distribution of steatosis in Nonalcoholic Fatty Liver Disease (NAFLD) can be categorized into Panacinar, Azonal, Zone 1 and Zone 3. Although Focal Zone 1 pathology is rare in adults with NAFLD (<1%), it has been reported in children with NAFLD. We hypothesized that focal Zone 1 steatosis and focal Zone 3 steatosis are two distinct sub-phenotypes of pediatric NAFLD.

1.1 Research Purpose

In order to better understand the potential differences between NAFLD histologic phenotypes of children, we performed multi-center cohort study. Our research aims to determine the relationship between the zonality of steatosis and demographic/clinical/ histologic features of children with NAFLD.

1.2 data

A total number of 813 children (younger than 18 years old) with biopsy-proven NAFLD enrolled in the NASH Clinical Research Network. 165 charactistics were measured at different aspects. Liver histology was reviewed by the Central Pathology Committee based on NASH CRN scoring system. For each biopsy, the predominant location of the fat droplets was recorded at Zone 1 and Zone 3.

1.2.1 Data issue

Over 50 measurements contain missing values exceeding 60% and were excluded from data. After removing 60 irrelevant variables, 42 variables of diesease NAFLD were reserved for the following factor analysis.

2 Methods

Descriptive statistics were generated for the remaining 42 characteristics, which were not illustrated here.

2.1 Factor analysis for reducing matrix dimension

Factor analysis was used to reduce the dimension of predictor matrix.

2.2 Random Forest for identify final variables

Random forest method was used to identify the essential variables for reconstructing the final model.

2.3 Decision tree for constructing preliminary classifier for NAFLD

Decision tree was used to build up the preliminary model for zonality partition.

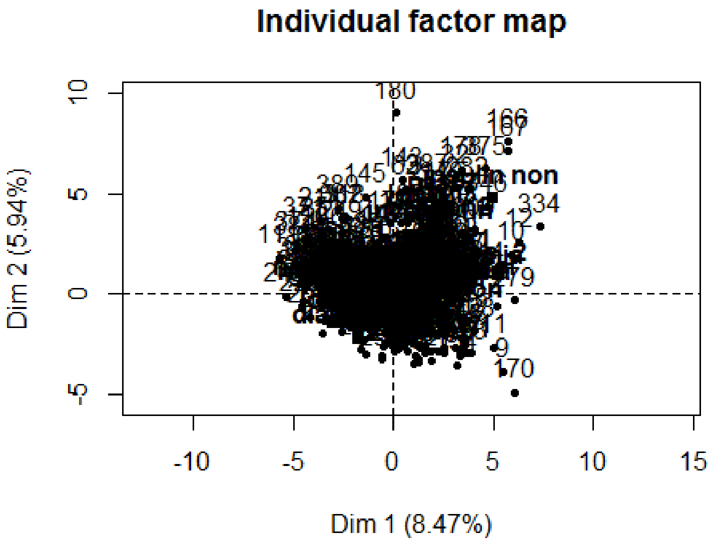
2.4 Logistic regression

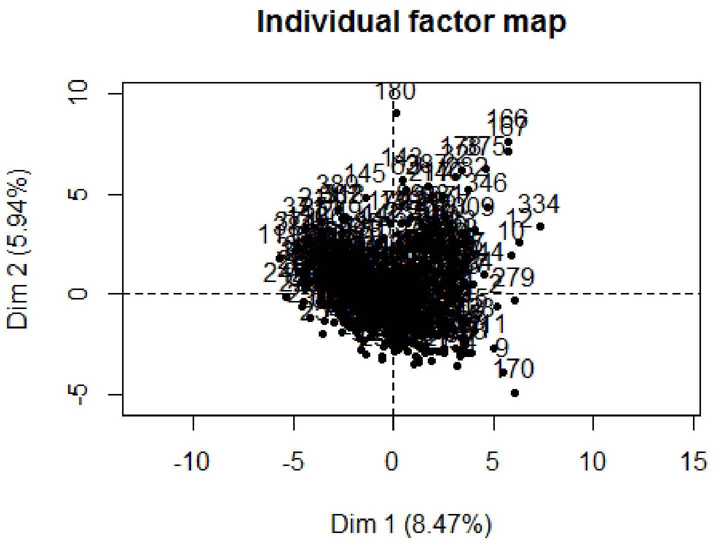
Logistic regression is build up for the final model for zonality classification.

3 Results

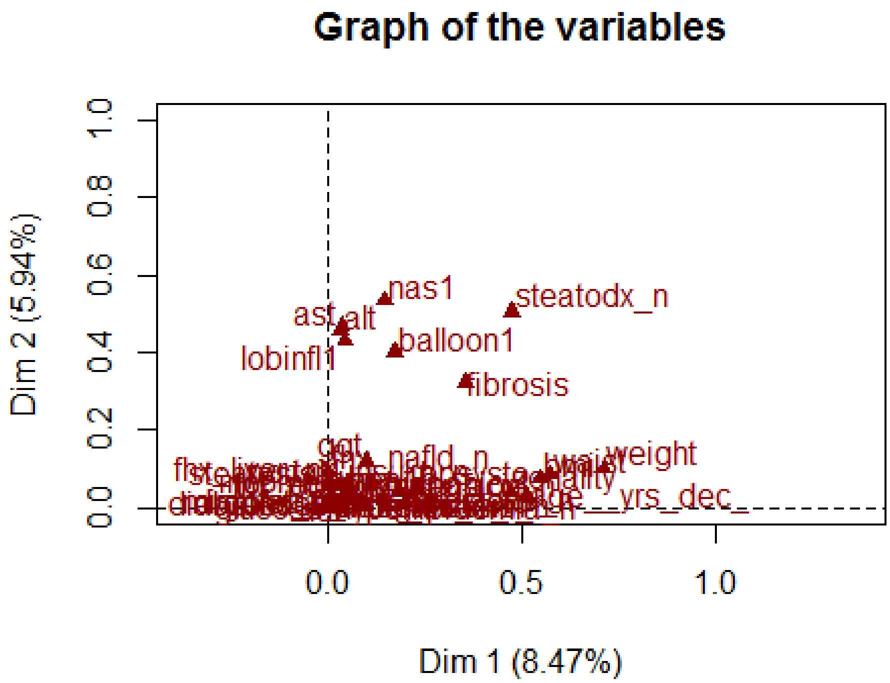
3.1 Factor analysis for reducing matrix dimension

3.1.1 Individual factor map (in dim1 and dim2)

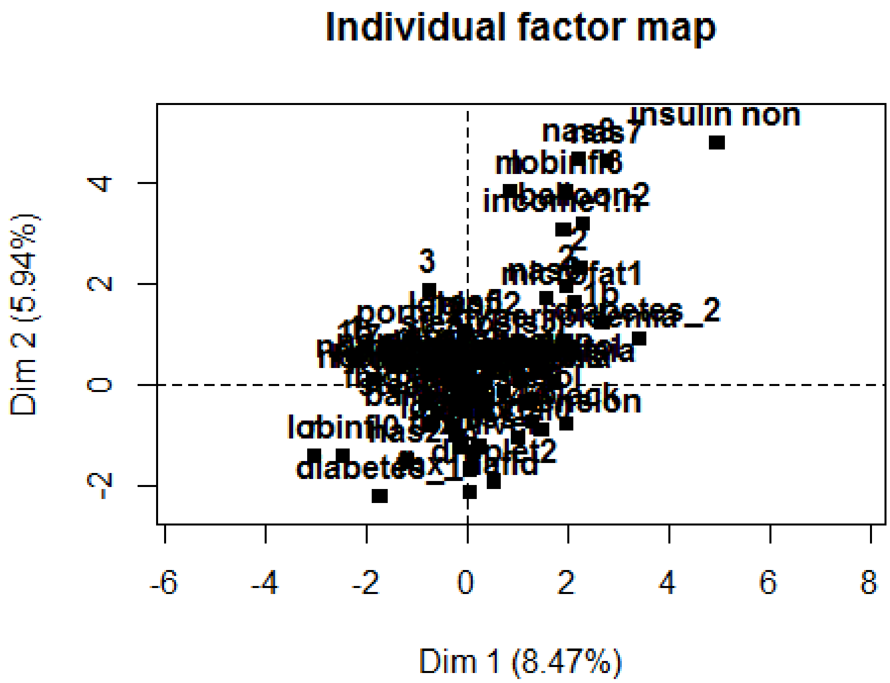




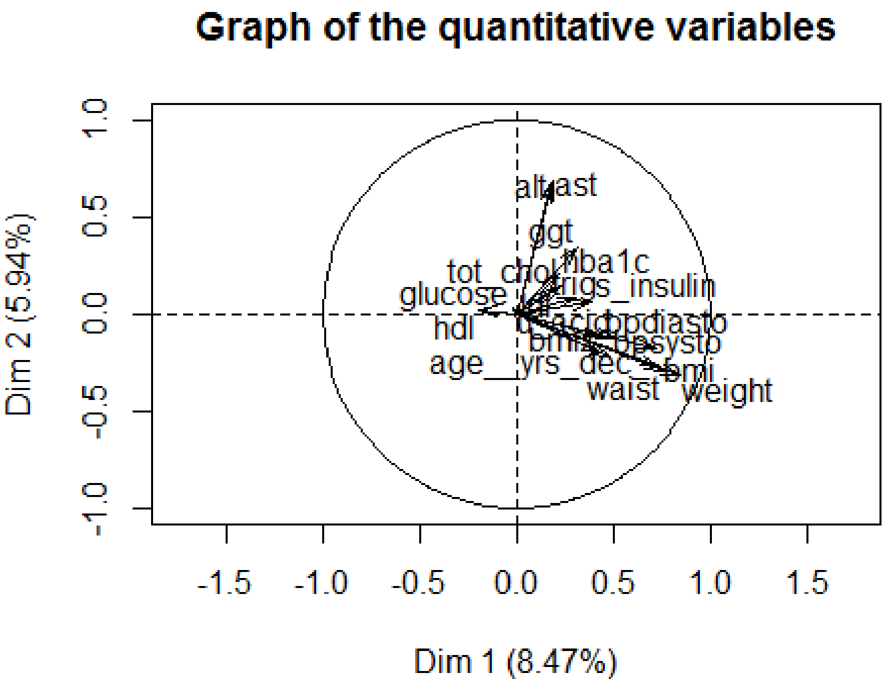
3.1.2 All variables projection (in dim1 and dim2)



3.1.3 Categorical variables (in dim1 and dim2)

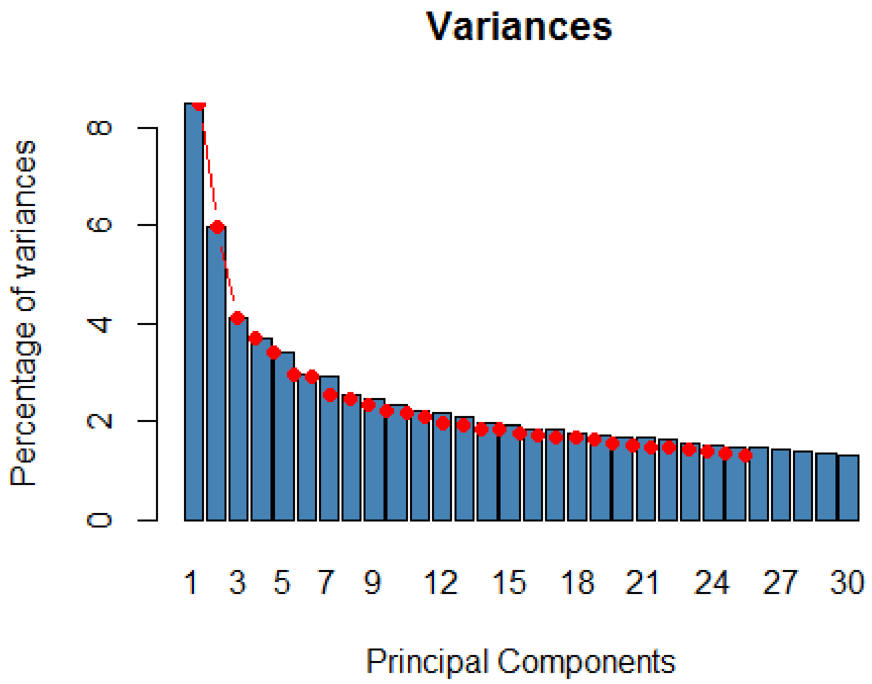


3.1.4 Numeric variables (in dim1 and dim2)



3.1.5 Bar plot

Bar plot shows that ~92% of the information (variances) contained in the data are retained by the first two principal components.



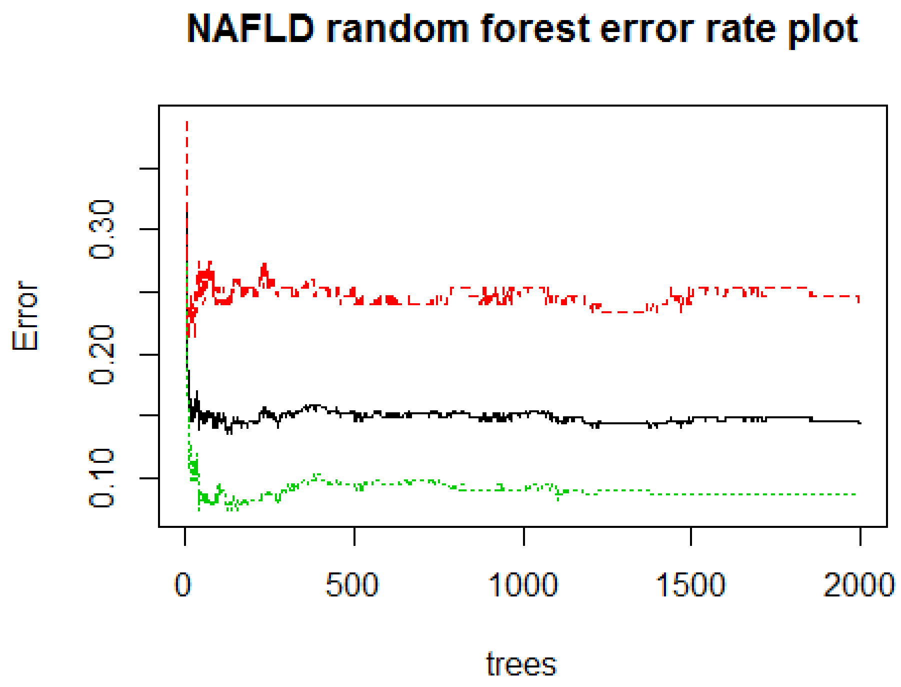
Numeric variables and category variables selected from the factor analysis are listed as followings. All the selected variables have p-values less than 0.01. Among them, 28 variables are reserved in the next analysis. Numeric variables (first 16 selected): weight, bpsysto, bpdiasto, uacid, hba1c, ast, alt, tot_chol, ldl, glucose, bmiz, age, ggt, waist, trig, and hdl. Categories variables (first 14 selected): gender, nas, race, hispanic, hba1c, steatosis, fibrosis, lobinfl, portal, balloon, sinsulin and nash.

```
## $quanti
##               correlation       p.value
## weight          0.8442461 3.822462e-107
## waist           0.7583611  4.091437e-74
## bmi             0.7415995  2.917777e-69
## age__yrs_dec_   0.7169356  9.093304e-63
## u_acid          0.5130195  1.449202e-27
## bpsysto         0.4810663  5.553538e-24
## bpdiasto        0.4468289  1.543188e-20
## bmiz            0.4300402  5.499137e-19
## s_insulin       0.3872886  2.094763e-15
## ggt             0.3153869  1.868946e-10
## trig            0.3093852  4.262230e-10
## tot_chol        0.2343967  2.876385e-06
## hba1c           0.2268703  6.037083e-06
## ldl             0.2036934  5.072266e-05
## ast             0.1890050  1.737137e-04
## alt             0.1787241  3.899607e-04
## hdl            -0.1991580  7.489567e-05
```

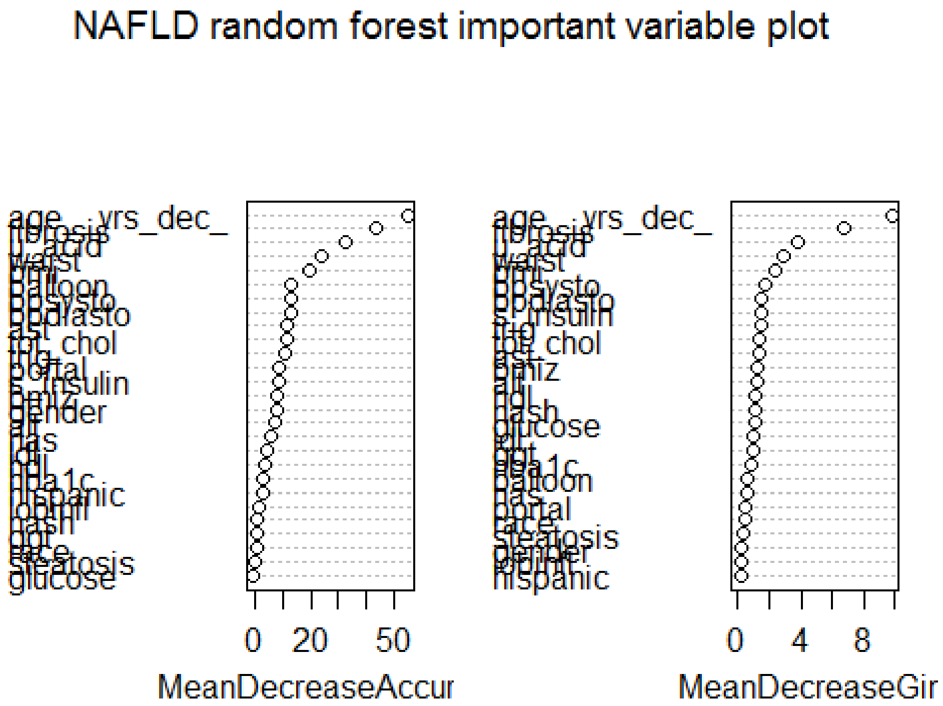
```
##
## $quali
##                          R2      p.value
## zonality       0.47334031 5.562401e-56
## steatodx_n     0.47553270 8.740758e-54
## fibrosis       0.35464630 8.998332e-34
## balloon1       0.17448912 7.688870e-17
## hispanic       0.14437496 7.885964e-14
## nas1           0.14726840 2.415324e-11
## diabetes_2_n   0.09823338 2.452995e-10
## race           0.08639149 4.744673e-06
## portal1        0.05252270 2.924595e-05
## insulin_n      0.04186270 4.685685e-05
## hypertension_n 0.03584691 1.691719e-04
## lobinfl1       0.04566724 4.262764e-04
## steatosis1     0.02670442 5.313243e-03
## microfat_n     0.01771760 8.489750e-03
## income1        0.04342366 8.952898e-03
```

3.2 Random Forest for identify final variables

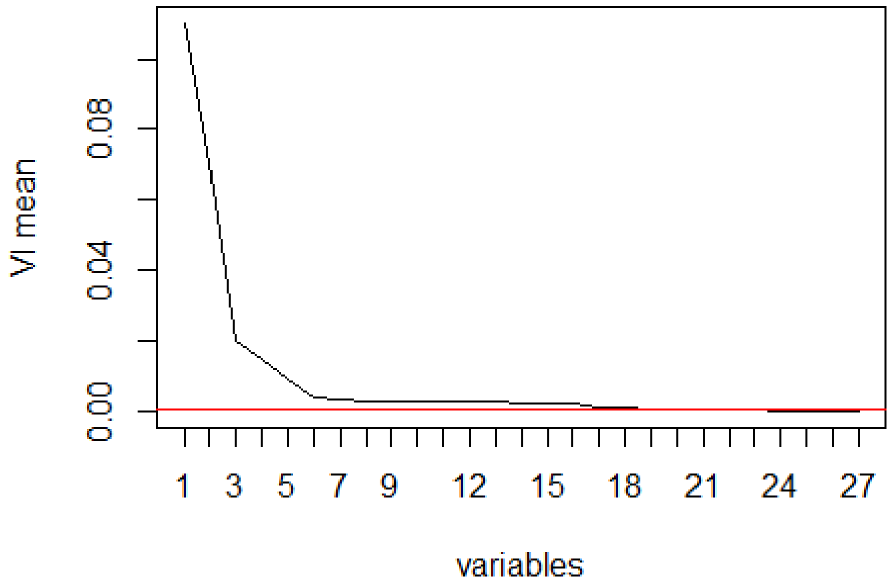
The following plot indicates that the average error rate is 0.15, while the difference between the error rate of zone1 and zone3 is 0.2. The difference between error rates is likely to reduce if an increasing amount of data was used to train the model.



The below plots show the rank of different variables contributing to MeanDecreaseAccurary and MeanDecreaseGini, respectively. The results also indicate that the top four variables have clear impact accuracy and Gini score compared with other variables.

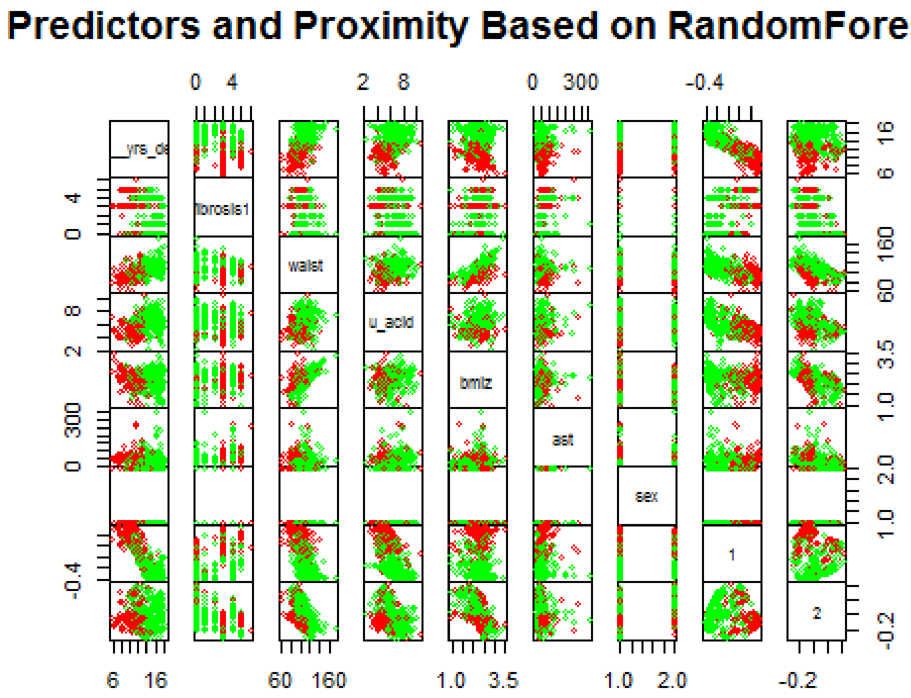


The following analysis automatically outputs sixteen variables, which lie above the threshold (the red line in the plot) and are recognized as important variables. Among them, eight (age, fibrosis, waist, u_acid, bmiz, ast, bmi and bpsysto) are selected as predictors for zonality partition. Since bmiz is the standardization of bmi, bmi can be removed. Random forest also indicates that the top 3 variables have the strongest predicting power.



```
##  [1] 16 27  4 19  1 15  6  2  8 20  3 22 13 14  7 17 24  5  9 21 18
## [1] 16 27  4 19  1 15  6  2
## [1] 16 27  2
```

The plot below illustrates the correlation between each pair of important variables. Moreover, it also suggests whether the selected important variables can distinguish the zonality 1 from zonality 3. The results show that bmiz, ast and u_acid can not distinguish these two zonalities. Four variables (age, fibrosis, sex and waist) will be included in the decision tree analysis.



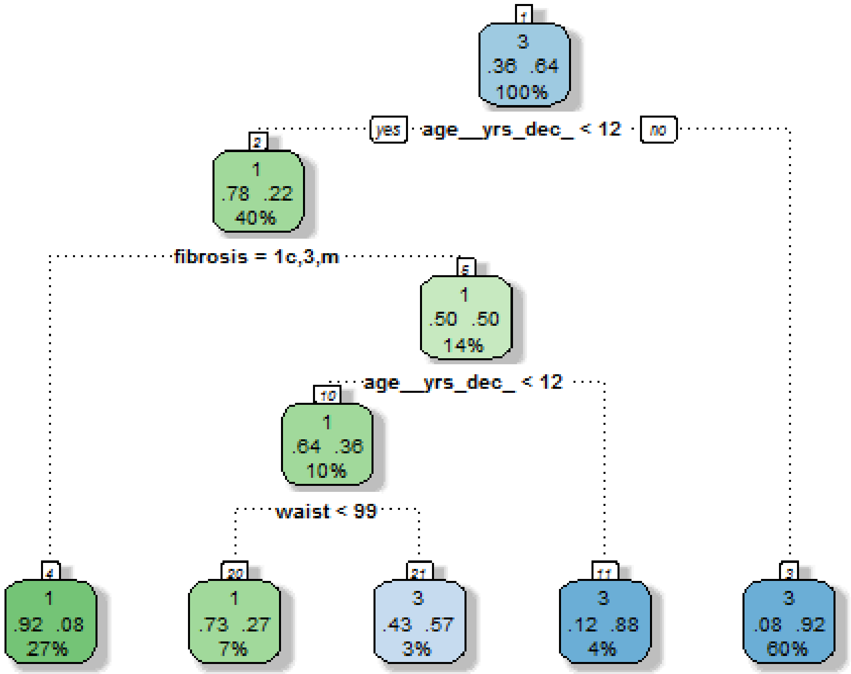
3.3 Construct a preliminary decision tree model

A preliminary decision tree model is build up and will be improved in the following analysis.

- The model is formulated as:

$$Zonality = Age + Fibrosis + Waist + error$$

The following plot shows that how the decision tree model split data into zone1 and zone3. This model was parameterized through the training dataset and was validated by the test dataset.



Rattle 2016-Feb-29 16:24:06 Nancy

```
##    n
## 1 34
```

There are 34 observations between predicted zonality and zonality. The error rate for this model is 0.1588235, which means that the model can achieve around 85% accuracy for predicting zonality compared to the original zonality result in the test dataset. Below are the details of the model:

```
## n= 220
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 220 79 3 (0.35909091 0.64090909)
##    2) age__yrs_dec_< 12.44932 89 20 1 (0.77528090 0.22471910)
##      4) fibrosis=1c,3,m 59  5 1 (0.91525424 0.08474576) *
##      5) fibrosis=0,1a,1b,2 30 15 1 (0.50000000 0.50000000)
##       10) age__yrs_dec_< 11.75068 22  8 1 (0.63636364 0.36363636)
##         20) waist< 98.9935 15  4 1 (0.73333333 0.26666667) *
##         21) waist>=98.9935 7  3 3 (0.42857143 0.57142857) *
##       11) age__yrs_dec_>=11.75068 8  1 3 (0.12500000 0.87500000) *
##    3) age__yrs_dec_>=12.44932 131 10 3 (0.07633588 0.92366412) *
```

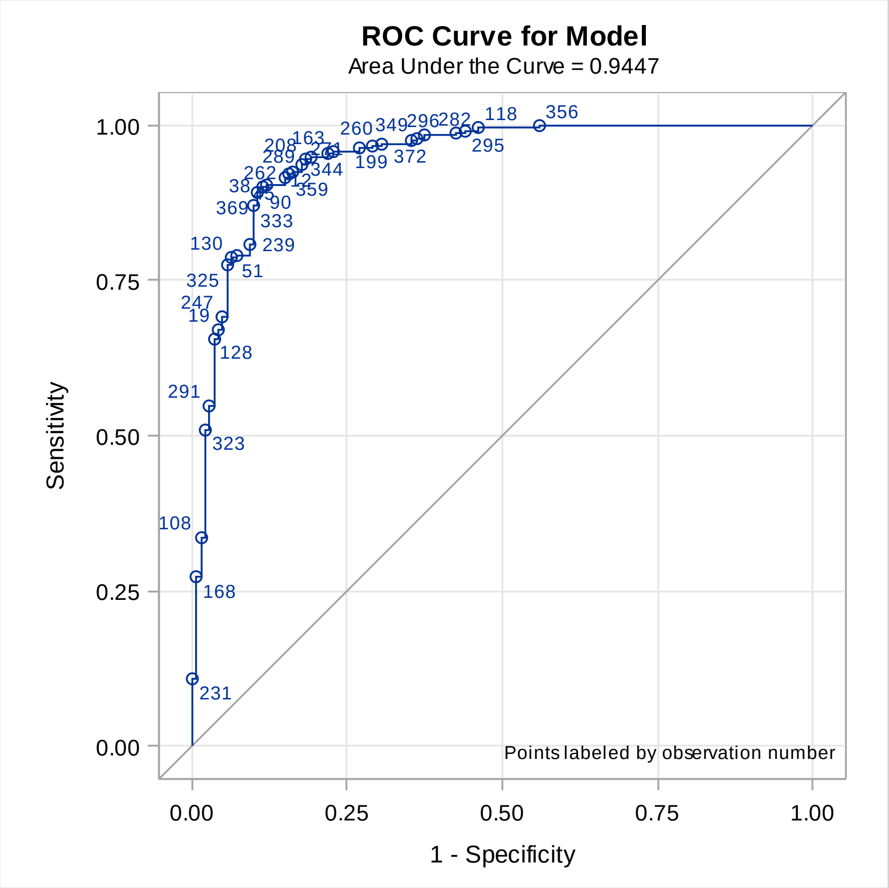
3.4 Logistic regression

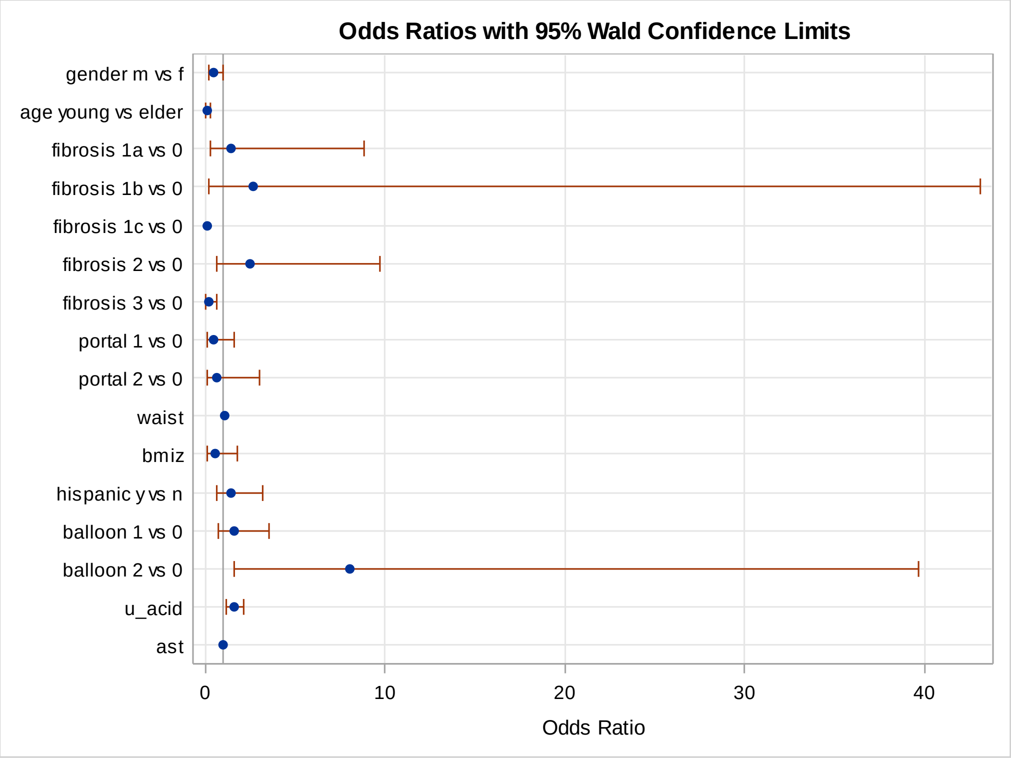
Impact factors for zonality classification were identified using multiple logistic regression models and a candidate set of factors: gender, age, fibrosis, balloon, uric acid ,waist and ast. Goodness of fit of the logistic model was assessed using a Hosmer- Lemeshow chi-square test with P=0.2844>0.05 indicating adequate fit. The likelihood ratio chi-square of 285.2079 with a p-value of 0.0001 and $R^{\wedge 2}$ with value 0.7206 tells us that our model as a whole fits significantly better than an empty model. The Score and Wald tests are asymptotically equivalent tests of the same hypothesis tested by the likelihood ratio test, not surprisingly, these tests also indicate that the model is statistically significant. The area under the ROC curve is estimated by the statistic c in the "Association of Predicted Probabilities and Observed Responses" table. In this example, the area under the ROC curve is 0.9447. The OddRatio plot for each coefficient is shown below. The All analyses assumed nominal, two-sided P values as statistically significant if P<0.05.Analyses were performed using SAS version 9.3 (SAS Institute) and R version 3.2.2

R-Square	0.5279	Max-rescaled R-Square	0.7206

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	285.2079	16	<.0001
Score	226.8007	16	<.0001
Wald	101.4865	16	<.0001

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
9.7308	8	0.2844


ROC Curve for Model
Area Under the Curve = 0.9447
Sensitivity
1 - Specificity
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
Points labeled by observation number
356
118
282
296
349
260
295
372
199
163
208
271
289
344
262
12
359
38
90
369
333
130
239
51
325
247
19
128
291
323
108
168
231



4 Discussion

The above study are limited in two major aspects: small sample size and many missing values. A large dataset can improve the training of classifier. The removal of missing values further reduced the available data for analysis.