A RECEIVER OPERATING CHARACTERISTIC, CURVE-LOGISTIC REGRESSION BASED VARIABLE SELECTION METHOD

by

Rajani Kaimal

A Thesis

Presented to the Division of Biostatistics and the
Oregon Health & Science University
School of Medicine
in partial fulfillment of the requirements for the degree of

Master of Science

June 2015

TABLE OF CONTENTS

Index of Tables

Index of Figures

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Jodi Lapidus, whose knowledge, support and mentoring has guided me throughout this project. I would also like to thank my committee members, Dr. Mara Tableman and Dr. Andrew Combs, who were always available for discussion, advice and guidance. I am grateful to all of the above for sharing this project, which I have immensely enjoyed working on, with me. I would also like to thank Aaron Baraff, his prior work on this project has been a stepping stone in my route to successful completion of this project.

I would like to thank my husband, Sharookh Daruwalla, for being there for me every step of the way, this thesis would not have come to be without him.

I would also like to thank my parents, sisters and friends for their continued patience, encouragement and love.

ABSTRACT

Biomarker discovery is a challenging process. It is rare that a single marker can accurately classify an outcome. Often, classification is improved by finding a combination of markers that can better distinguish between patients with and without a condition of interest. The present study describes a novel method which can potentially function as a diagnostic algorithm to isolate a parsimonious combination of markers that has good classification properties. The method takes advantage of the well-studied properties of receiver operating characteristic (ROC) curves and logistic regression models to select combination of variables that maximize the partial area under ROC (pAUROC), a clinically relevant metric.

Our new procedure proceeds as follows. The partial area under the ROC curve is determined for all potential markers. The model with the maximum pAUROC over the selected false positive fraction (FPF) range is the first variable to step into the model. Next, an adaptation of the jagged-ordered nonparametric algorithm is used to select from the remaining markers based on improvement in the pAUROC. A potential marker is retained in the classification model only if the resultant integrated discrimination index, (IDI) is above a preset threshold. Thus, by using a combination of different classification metrics, the new method hones in on a combination model with good sensitivity and specificity from a moderate-size pool of potential markers.

Contrary to traditional variable selection methods (example, stepwise selection), which are often based on measures of association, the current method is specifically focused on classification metrics. Hence, it eliminates the need for fitting models for every possible combination of candidate markers, and vastly improves the speed of the variable selection process. This is demonstrated by the performance of the method in isolating a combination model for classifying intra-amniotic inflammation in women with preterm labor with intact membranes. The method accurately selected cervicovaginal proteins with optimum classification performance, but contained relatively few proteins, which is desirable from a clinical perspective. Results obtained were comparable to similar parsimonious models, built by using traditional protracted methods of data mining followed by regression, thus supporting the efficiency of current method.

INTRODUCTION

Early diagnosis and management is key to improve mortality and morbidity in several disorders. For many diseases, it is critical to the survival of the patient, and in other cases, early diagnosis can enable clinicians to better manage the disease, enhancing the expected quality of life. Increasingly, studies have focused on early and accurate diagnosis of disease. Current technological advances have enabled accurate and rapid measurement of proteins that could potentially serve as biomarkers (1).However single markers that can accurately classify disease are rare, hence there is often a need to identify combinations of proteins that could be used as a diagnostic tool to identify disease. Current statistical methods are lacking in their ability to select and combine multiple markers to classify disease in a clinically relevant fashion. Data mining methods allow one to narrow down a large number potential classifiers (2); however building and refining classification models is still a time intensive process involving multiple iterations of combining and testing potential models. Methods to study association between an outcome and independent variables are commonly used to build classification models, but again, they are severely limited in being able to focus on clinically relevant metrics, like sensitivity and specificity. In the current investigation, we present a novel method to select variables for entry into a logistic regression model that is based on classification

metrics. The new method is outlined and compared to commonly-used variable selection methods. The method is demonstrated on two datasets from clinical studies aimed at detecting intra-amniotic inflammation from cervicovaginal proteins in women in preterm labor (3,4)

Classification metrics: Sensitivity, Specificity and the Receiver operating curve.

Misclassification of disease can lead to undue stress and financial expenses or worse still, false negative results may worsen prognosis by delaying diagnosis. It is important to specify the sensitivity and specificity of tests to understand the risk and type of misclassification that is more likely to occur with that particular test. It also enables a better understanding of the true classification ability of the test and enables comparison of different tests.

Sensitivity is defined as the proportion of patients, correctly diagnosed as having a condition (or disease/disorder/event) from a pool of patients, all of whom have the condition. It is also called the true positive fraction or TPF. (5)

Specificity is defined as the proportion of patients, correctly diagnosed as not having the condition (or disease/disorder/event) from a pool of patients, all of whom do not have the condition. By definition then, specificity= 1-false positive fraction (FPF).

Consider a binary test (Y) with outcome =1 if test is positive and outcome =0 if test is negative. Let the presence of disease be denoted as D=1 and the absence of disease as D=0 .There are four possible scenarios as shown in Table 1

Table 1: Outcomes of Binary test.

	D=0 (non-disease)	D=1 (disease)
Y=0 (test is negative)	True negatives	False negatives
Y=1 (test is positive)	False positives	True positives

Accordingly,

Sensitivity or TPF is defined as $P[Y = 1|D = 1]$

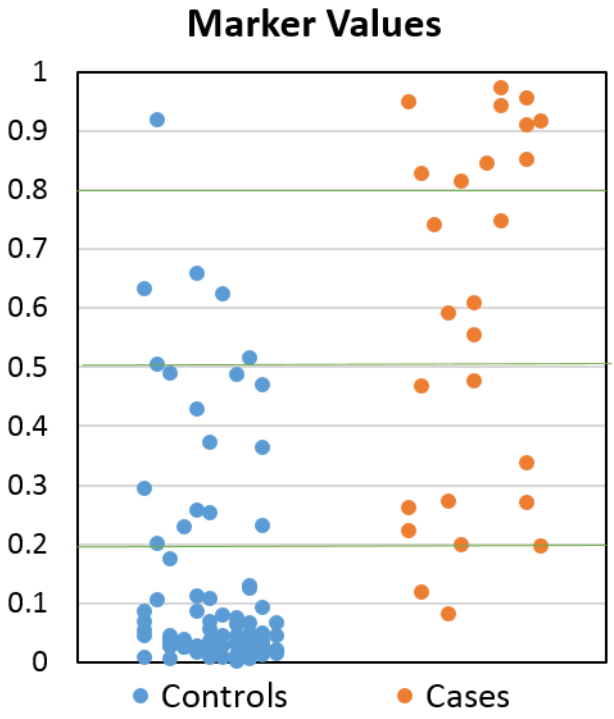
And

1-Specificity or FPF= $P[Y = 1|D = 0]$

Thus sensitivity and specificity are probability of accurate classification in the presence or absence of disease, respectively. Together, TPF and FPF define the performance of the test. A perfect test would have a TPF=1 that is every patient with disease is correctly diagnosed as such, and an FPF =0 that is no disease free patient has a positive test result. However, in practice, a perfect test rarely exists, and usually there is some tradeoff in specificity for better sensitivity and vice versa. Depending on the nature of the disease being tested for, it may be more important for it to be highly sensitive or highly specific. For example, if a condition can be fatal if not diagnosed early, it may be more important to have a high sensitivity so that no positive disease is wrongly classified as non-disease. On the other hand, if early diagnosis is not critical to prognosis, it may be more important to have high specificity in order to avoid unnecessary additional tests. Although both sensitivity and specificity of a test are important characteristics of test performance, the clinical utility of a test also depends on the prevalence of the disease in the intended use population. Additionally, the cost of the test and the availability of other tests also play a role in setting clinically relevant sensitivity and specificity cut points. Being able to evaluate and maximize the performance of a test within restricted sensitivity and specificity is thus critical in developing a clinically relevant test

In a clinical setting few tests are purely binary, but a few do exist, for example, presence or absence of certain bacteria, or relevant gene mutation. A large number of tests measure proteins or other biological markers, on a continuous scale, for example,

protein marker CA125 (U/mL) and hemoglobin (g/dl) levels In such cases it is important to optimize the thresholds over which a test is considered as positive or negative. A continuous marker may have different sensitivity and specificity depending on the threshold set. Hence, optimization under conditions of use are critical to accurate diagnosis. Consider the following scenarios, where we assume that higher levels of marker are indicative of disease; if the threshold is set high we may miss patients who are in the early stages of disease or just constitutively produce less of the marker. In other words, the test will have high specificity but low sensitivity. On the other hand, if we set the threshold too low, while we may catch everyone with potential disease, we may also wrongly diagnose subjects with slightly elevated levels as having disease, i.e. the test will be highly sensitive but not very specific. (Figure1)



Threshold	Sensitivity	Specificity
0.2	0.923	0.78
0.5	0.577	0.927
0.8	0.384	0.988

Figure 1: Sensitivity and Specificity of a marker at different thresholds

As the threshold, above which values of a continuous marker are considered indicative of disease, are increased, the specificity increases and sensitivity decreases.

The sensitivity and specificity then, is highly dependent on the thresholds we set. Hence in case of a continuous tests an additional tool is necessary which illustrates this shifting specificity and sensitivity with threshold. A receiver operating characteristic curve demonstrates the performance of a continuous test by plotting the sensitivity over 1-specificity at each measured value, i.e. each measured value is individually considered as a threshold and FPF and TPF is calculated accordingly.

Consider a value *c* set as the threshold

If $Y \geq c$ the test =positive

If $Y < c$ then the test=negative

Accordingly,

$$TPF(c) = P[Y \geq c|D = 1]$$

And $$FPF(c) = P[Y \geq c|D = 0]$$

The plot of TPF over FPF for all possible threshold values gives the ROC curve (Figure 2).

$$ROC(.) = \{\big(FPF(c), TPF(c)\big), c \in (-\infty, \infty)\}$$

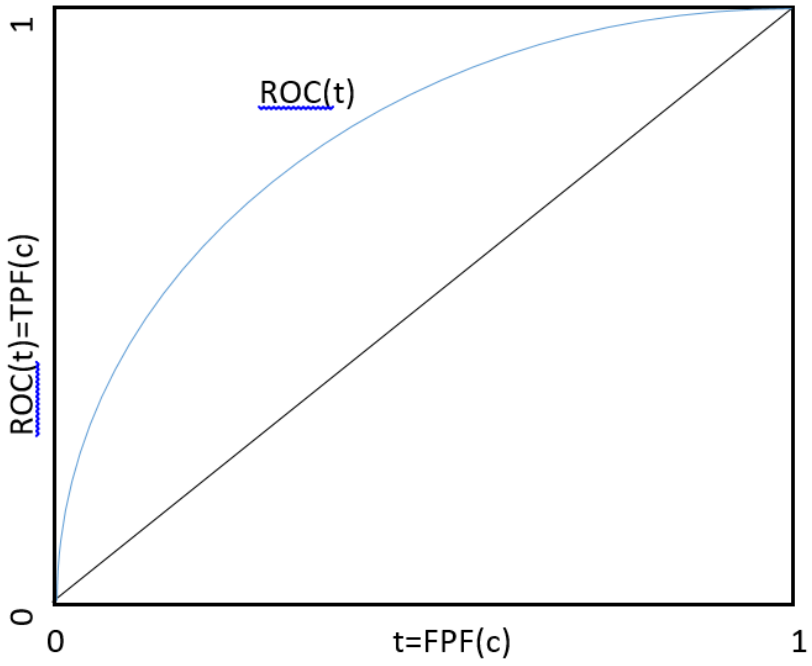


Figure 2: The receiver operating characteristic curve is a plot of sensitivity over 1-specificity at every possible test value

The Receiver Operating characteristic curve is a plot of sensitivity (TPF) over 1-specificity (FPF) at every value a test can take.

As one can imagine, at low values the sensitivity would be very close to 1 and specificity would be close to 0. As the threshold is set at higher measured values the sensitivity decreases and specificity increases. The ROC curve is thus a monotone increasing function and can be written as

$$ROC(.) = \{(t, ROC(t)), t \in (0,1)\}$$

Where the ROC function maps t to$TPF(c)$, and c is the threshold corresponding to $FPF(c) = t$

In order to measure the overall performance of a test over the entire range of values the area under the receiver operating curve (AUROC) is used. If a test has high sensitivity even at high specificity, the ROC will curve further to the left and up, increasing the AUROC, while a test with low sensitivity at high specificity will be curved closer towards the diagonal and have a lower AUROC. Thus the AUROC is a good measure of test performance. It can also be interpreted as the probability that given a randomly selected pair of diseased and non-diseased person, the person with disease will have a higher test value than the one without disease. It is mathematically defined as

$$AUROC = \int_0^1 ROC(t)dt$$

As both sensitivity and specificity are probabilities and have a range of 0 to 1, the maximum possible area under the ROC curve is equal to 1. A test that is 100% sensitive and 100% specific would have an area of 1 while a useless test would have an area close

to 0.5, i.e. a 50% chance of being right just by chance alone. Figure 3 illustrates two ROC curves estimated from fictitious data – one with good and one with poor classification performance.

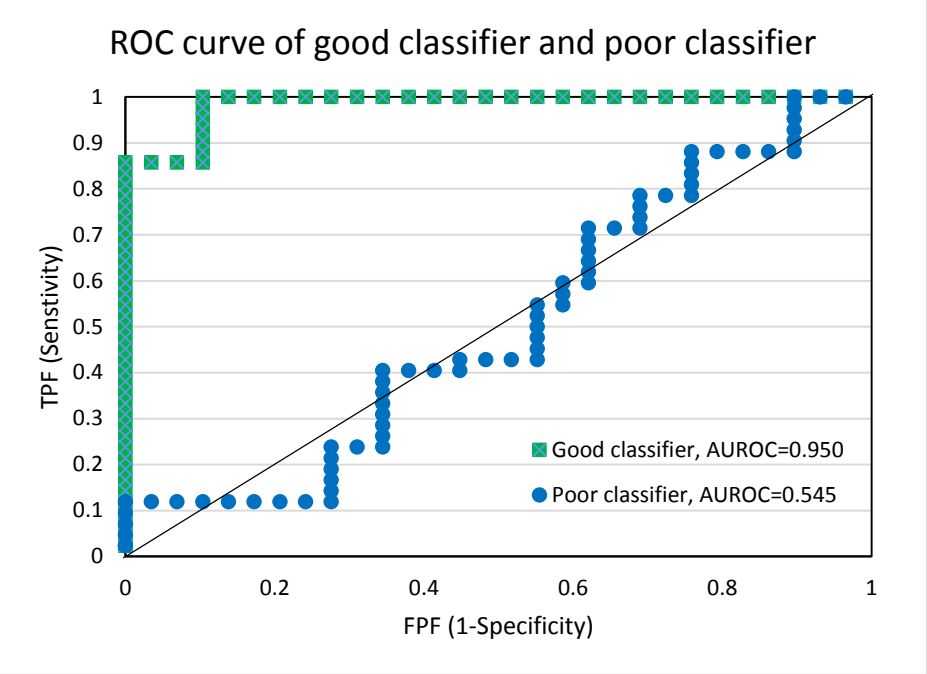


Figure 3: Estimation and comparison of ROC curves.

A good classifier has an ROC curve that curves further up and left compared to an ROC curve of a poor classifier. Above are estimates of ROC curves of a good and poor classifier.

While a perfect test may not be feasible, a test that maximizes the area under ROC, (AUROC) is a test which has high sensitivity even at high specificity and would be considered a good test. Based on Neyman-Pearson Lemma, the optimal criteria for a test Y for classifying a subject as having disease is given by the likelihood ratio > *c*. The Likelihood ratio is a ratio of two probabilities, the probability of a positive test given disease over the probability of a positive test in the absence of disease. A test based on likelihood ratio gives the maximal possible AUROC for any given test and consequently any monotonic transformation of the likelihood ratio also gives the most optimal AUROC. This critical property enables the use of logistic regression in building ROC curves, as will be discussed later.

Another metric that is widely used to illustrate the performance of a test is the partial AUROC. In designing new test, rather than overall performance, the performance of a test in the high specificity region may be more clinically relevant. A threshold which gives a low specificity is rarely of any clinical utility. The partial AUROC is the AUROC over a restricted FPF range (or restricted TPF range).

$$pAUROC(t_0) = \int_0^{t_0} ROC(t)dt$$

The value of pAUROC (t_0) ranges from $\frac{{t_0}^2}{2}$ for a completely uninformative test to t_0 for a perfect test.

The advantages of using the ROC for characterizing a test are manifold. It enables one to set appropriate cut-off or thresholds above which a continuous test would be considered positive. It is also an invaluable tool in comparing the performance of tests. The higher the AUROC of a test the higher the sensitivity of a test for any given specificity.

Logistic regression and ROC

Binomial Logistic regression is a commonly used statistical procedure to determine the probability of one of two outcomes given a set of independent variables. It is an invaluable tool in the clinical field in determining the odds of a person having a disease/ condition given a set of exposures or clinical lab values. The logistic function is defined as

$$\log\left(\frac{\pi(\underline{x})}{1-\pi(\underline{x})}\right) = g(\underline{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

Where $\pi(\underline{x})$ is the probability of case/disease/condition, given $(\underline{x})$

$(\underline{x})$ is the vector of exposures or the independent variables that determines the outcome and, k is the number of such variables. β is the MLE parameter estimates (β_0 is the intercept)

As can be seen the above form is equivalent to the log odds, as it is the ratio of the probability the outcome is a case or has disease/condition over the probability the outcome is not a case (or is a control)(6)

Unlike linear regression, binomial logistic regression does not provide a predicted value for the outcome, that is, it does not assign case or control status, but instead it estimates the probability that the outcome is a case. Hence, the scale of the fitted values has a range of 0 to 1. The closer to unity or higher the predicted probability the more likely the outcome is a case, i.e. has disease. This predicted probability can also be defined by an alternative form of the logistic function.

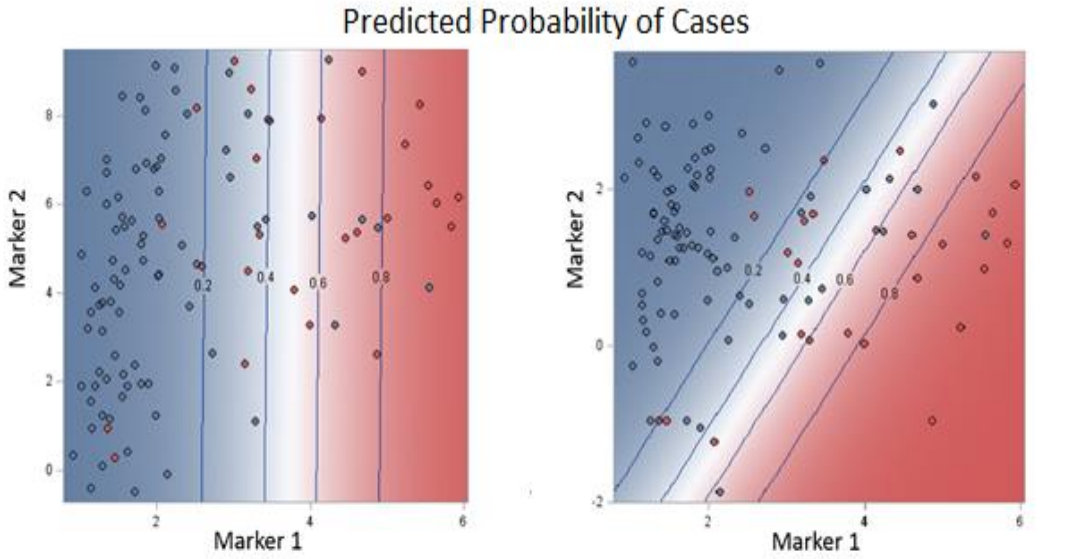
$$\pi(\underline{x}) = \frac{e^{g(\underline{x})}}{1+e^{g(\underline{x})}}$$

Where $g(\underline{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$ is a function of the independent variables

The above probability form makes logistic regression ideal for use in classification. The predicted probabilities have a monotonic relation with the actual test values and can be used to build the ROC curves, just as you would using the actual marker values. Additionally, the predicted probability of a specific outcome (cases) can be calculated for a linear (or nonlinear) combinations of markers using logistic regression, and this can then be used to build an ROC curve, which then gives the classification potential of this combination of markers. The contour plots in Figure 4 illustrate the use of logistic regression to study the classification potential of two linear combinations of markers.

The contour plot is built by plotting the values of Marker 2 against Marker 1. Lines across the plots, either vertical or diagonal represents the predicted probability of cases, at the corresponding marker values.

1. In Model one, Marker 2 is not associated with the outcome and hence does not make any significant contribution. The predicted probability is almost exclusively determined by marker 1 values (the lines are vertical because the estimated regression coefficient for marker 2 equals 0).
2. In Model two, Marker 2 is associated and negatively correlated with outcome. Hence the predicted probability is determined by the values of both the markers, as can be seen by the diagonal logistic contour lines.



Analysis of Maximum Likelihood Estimates		
Parameter	Estimate (β)	Pr> Chisq
Marker 1	1.1937	<.0001
Marker 2	-0.00797	0.9481

Analysis of Maximum Likelihood Estimates		
Parameter	Estimate (β)	Pr> Chisq
Marker 1	1.4249	<.0001
Marker 2	-0.9113	0.0028

Figure 4: Contour Plot of 2-marker logistic regression models for classification

Two different models showing the contribution of two markers to the predicted probability of cases. Marker 2 in the left hand side model makes no significant contribution, as is reflected in the perpendicular predicted probability contours, while the marker 2 in the right hand side model makes a significant contribution to predicted probability and this is reflected in the diagonal contour lines.

The above plots show how the logistic procedure can be used for classification. The cases in the above plots are shown in red, while controls are denoted by blue dots. It can be seen that in both combination models, at a predicted probability of greater than 0.8, all the subjects are cases, except for 1. Amongst subjects with predicted probability less than 0.2, there are only 5 cases in combination 1, and 2 cases in combination 2, while all the rest are controls. Thus, this demonstrates that predicted probabilities from logistic regression models can be used to classify cases vs. controls, and accuracy of doing so will depend on the threshold used. Based on threshold of predicted probability=0.2, the sensitivity and specificity of the two models in discriminating cases from controls is compared in Table 2.

Table 2: Comparison of Sensitivity and Specificity of two logistic regression models at identical thresholds.

	# of Controls (<0.2)	Specificity	# of Cases (≥0.2)	Sensitivity
Actual	82		26	
Model 1	67	0.817	21	0.808
Model 2	64	0.780	24	0.923

Similarly, sensitivity and specificity can be determined at every such threshold and plotted to build ROC curves.

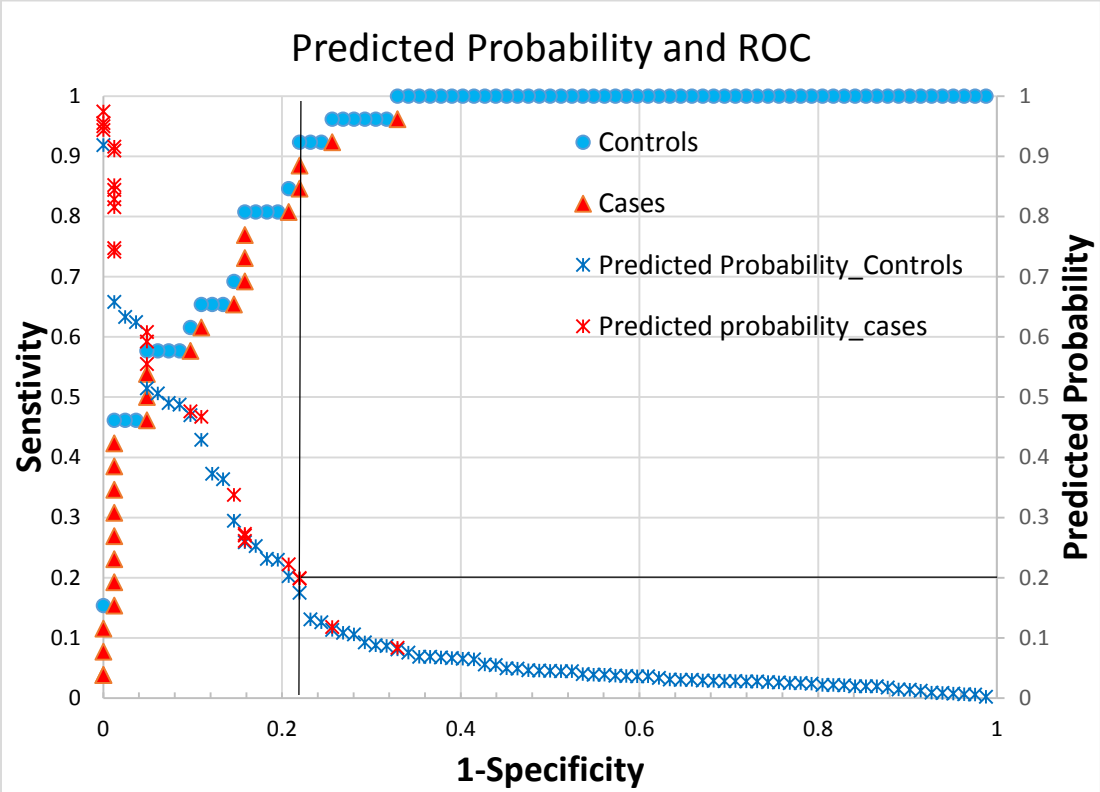


Figure 5: ROC curve and predicted probability of logistic regression

The predicted probability of marker combinations can be used as cut offs for classification of outcome. The sensitivity and specificity at each value of predicted probability is plotted to give the ROC curve

When a logistic model has a good fit, it is the ideal tool for classification as the likelihood ratio-based decision rule is the uniformly most powerful test, which means that a test based on likelihood ratio achieves the highest statistical power among all test with the same type-1 error rate.(8).

When Y is a potential marker, (or combination of markers,X) for the presence $(D=1)$ or absence of disease $(D=0)$, then the likelihood ratio is given by

$$LR(Y) = \frac{P(Y|D=1)}{P(Y|D=0)}$$

Based on an application of the Neyman Pearson Lemma, the optimal classification rule is to predict $Y=1$ (or a positive test result) for a sample, if $LR(Y)>c$, where c is a predetermined cut-point for a set sensitivity and specificity. If two subjects are tested for the marker, the person with disease is more likely to have a higher test value than a person with no disease. The logit function is a monotone increasing function of the likelihood ratio and hence it shares this property of optimality.(7,8)

$$\log\left(\frac{P(D=1|Y)}{1-p(D=1|Y)}\right) \rightarrow \mathrm{LR(Y)} = \left(\frac{P(Y|D=1)}{p(Y|D=0)}\right)$$

It should be noted, that since the Logistic regression is a parametric procedure, the model should meet all the assumptions and be checked for goodness of fit, before any conclusions can be drawn. The Akaike information criterion, (AIC) is a tool to compare the fit of different models, which may or may not be nested. While the AIC value itself has no intrinsic meaning, the lower the AIC value the better the fit when comparing different

models. When choosing a logistic model for classification purposes, a lower AIC value may be used to choose amongst the multiple optimum models. Other metrics based on likelihood are also good indicators of model performance. The Score statistic is one such metric based on likelihood, it is the slope of the log-likelihood with respect to P. In other words, it indicates how sensitively the likelihood function $L(P;Y)$ depends on its parameter, P. (For logistic regression)

It can also be defined as

$$Score(P,Y)=\frac{\delta}{\delta P}\log L(P;Y) = \frac{1}{L(P;Y)}\frac{\delta L(P;Y)}{\delta P}$$

The score statistic is one of the many methods for variable selection for association models using logistic regression. The score statistic tests for improvement in fit when a variable is added to the null model, i.e. model with no explanatory or independent variables. It is also called the global score statistic and has a Chi-square distribution with j degrees of freedom, where j is the number of explanatory variables in the model. The best subset selection procedure uses the score statistic in evaluating association models. The process involves building multiple models with varying number of variables and determining the score statistic of each model. The subset of models at each parameter level that have the higher score statistic are the better association models. This method of model selection may be used as a surrogate for building classification models as it is based on the likelihood ratio. The best subset selection thus provides one with multiple models and each model should be individually checked for classification potential. This

can be tedious when there are a large number of potential markers giving many potential models. Hence this method is optimal only for variable selection from smaller sets of potential markers. When a moderate to large number of candidates need to be tested for inclusion, it may be better to use data mining methods to first narrow down the number of potentials, before using the best subset selection method.

Data Mining for Classification: Decision Trees, Random Forests

Data mining is a useful technique for selecting relevant data, in the presence of noise. In the context of classification, decision trees or recursive partitioning are widely used data mining methods for variable selection. They are especially useful when the number of potential markers are equal to or even larger than the number of available samples. In the current study, random forests, a recursive partitioning method is used for narrowing down the list of potential markers for further analysis using logistic regression based method.

Decision trees or recursive partitioning builds models in the form of a tree structure or flowchart by breaking down data into smaller and smaller subsets based on the number of variables (Figure 6). When used for classification, each possible outcome is called a class and each variable forms a node. The variables could be binary, categorical, ordinal or continuous. The tree is built top down from a single node, which corresponds to the best predictor node and is called the root node. The data is consecutively subsetted over each variable or node until no further split can be achieved, i.e. the terminal node is homogenous. A 'greedy' algorithm that makes a series of locally optimum decisions is

most commonly applied to choose which variable will be used at each node for partitioning the data. While decision trees are very effective in classification, it is biased towards variables with more categories as they are more likely to give homogenous splits. It also over-fits the data, as each node makes a decision based on given data set (2). This over-fitting gives overly optimistic measures of performance such as sensitivity, specificity and AUROC, which is not reproducible in different datasets. To avoid this issue of over-fitting, Breiman introduced the Random forest method(9).

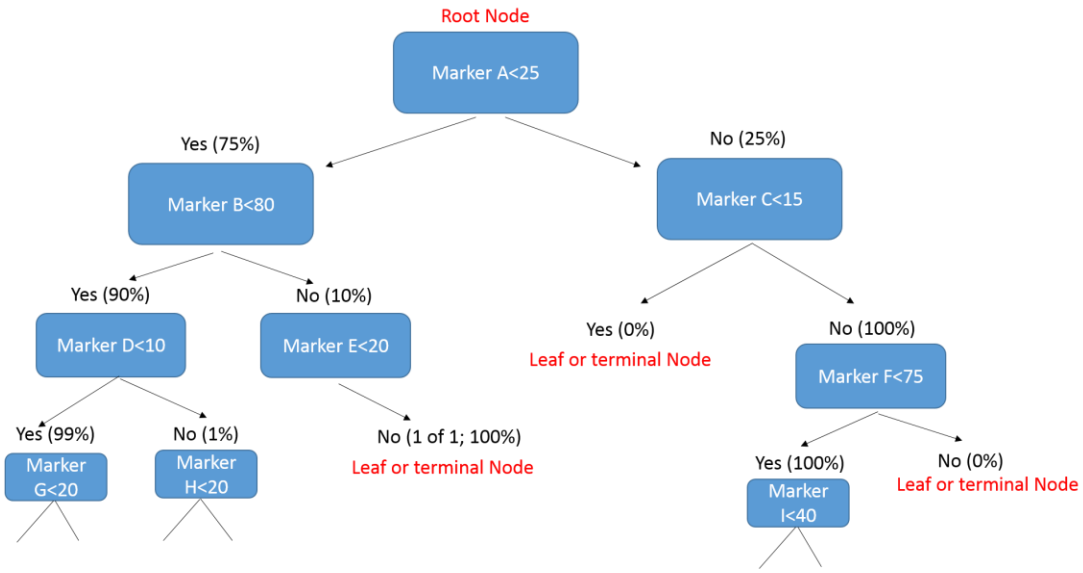


Figure 6: Example of a Decision Tree

Decision trees, also called recursive partitioning are tools to predict an outcome or pick out good classifiers.

Random forest is a recursive partitioning method that is widely used for both regression and classification. It overcomes the issue of over fitting data by utilizing two tools, randomization and multiplicity. It introduces randomness at two different steps that ensures that diverse trees are built at each run. The first step of randomization occurs when it chooses a fraction of the data usually about 2/3rd of the data as training set, this is called bagging and is usually done by bootstrapping with replacement, but can also be done by subsampling a predetermined fraction. The remaining 1/3rd of data, also called OOB or 'out of bag' is used as a test set to validate the model at each step. The next randomization occurs in the set of variables available at each node to split over. This ensures that variables which would not have been considered early on, also get a chance to be included in model building and introduces additional variation in the trees built. The number of random variables to choose from at each node (mtry), can be user-defined, and is typically set to the square root of the total number of variables(11). Each candidate from the randomly chosen set is checked for its association with the outcome, the predictor/classifier which yields the smallest P-value is selected and the next split is chosen within this variable. This process of randomizing variables at each node is continued until a tree is built, to either a preset depth or until it reaches a homogeneous end with no pruning. A large number of such random and diverse trees are thus built. While each tree on its own may not be stable, the aggregate of predictions/classifications has been shown to have very good accuracy and low variance. Each tree over fits, however in different ways, and aggregating a large number of trees smooth out the decision boundaries created by each individual tree. (11)

Random forest also generates a parameter called the variable importance factor (VIF), based on the OOB error rate before and after permuting the values of the variable. If a variable is a good classifier or is significantly associated with outcome, permuting this variable will break this association. When this now permuted variable along with the remaining non permuted variables are used to predict the outcome in the OOB or test set the prediction accuracy will be decreased, thus giving a large OOB error rate. Stronger the association of the variable with outcome larger will be the OOB error rate after permutation, while the prediction accuracy would not change much if the variable was not associated with outcome at all. The difference in prediction accuracy before and after permutation averaged over all trees gives a measure of the how critical the variable is for accurate classification or prediction. Based on this error rate, variables are assigned a variable importance factor (VIF). Higher the VIF the more important the variable.

$$VI_j^{(ER)} = \frac{1}{ntree} \sum_{t=1}^{ntree} (ER_{t\tilde{j}} - ER_{tj})$$

Where ntree denotes the number of trees

ER_{tj} denotes the mean error rate over all OOB observations in tree t before permuting predictor j

$ER_{t\tilde{j}}$ denotes the mean error rate over all OOB observations in tree t after randomly permuting predictor j.

While random forest works well in most settings, it tends to assign higher VIF's to correlated variables, irrespective of the actual association of the variable to outcome(12). That is to say, it is not capable of distinguishing spurious associations due to confounding factors that may be independently associated with a variable but not with the outcome of interest. For example, consider two correlated variables, age, and shoe size. If we test their association with reading skills, shoes size would have a high VIF, even though the true predictor is age and on its own shoe size would be irrelevant to reading skills. In a biomedical setting, variables often show high correlation with each other and it is important to account for this conditional effect when classifying an outcome. To overcome the issue of correlated variables Strobl, introduced the conditional variable importance factor (12,13). This takes into account not just the marginal influence of a variable but also its influence conditional on another variable. Here the assumption is that a variable, X is not truly independent from the some or all of the rest of the variables, Z and hence it is easy to see that by permuting only variable X the correlation structure between X and the other remaining variables is destroyed giving spurious results. To account for this correlation, Strobl suggest a conditional permutation scheme, where X_j is permuted only within groups of observations with $Z=z$, i.e. the null hypothesis is that X is associated with Y conditional on Z

$$P\left(Y, X_j \middle| Z\right) = P(Y|Z).P(X_j|Z)$$

And accordingly the permutation scheme is now shown in Figure 7.

Y	X_j	Z
y_1	$x_{\pi_j(1),j}$	z_1
$\vdots$	$\vdots$	$\vdots$
y_i	$x_{\pi_j(i),j}$	z_i
$\vdots$	$\vdots$	$\vdots$
y_n	$x_{\pi_j(n),j}$	z_n

Y	X_j	Z
y_1	$x_{\pi_{j\|Z=a}(1),j}$	$z_1 = a$
y_3	$x_{\pi_{j\|Z=a}(3),j}$	$z_3 = a$
y_{27}	$x_{\pi_{j\|Z=a}(27),j}$	$z_{27} = a$
y_6	$x_{\pi_{j\|Z=b}(6),j}$	$z_6 = b$
y_{14}	$x_{\pi_{j\|Z=b}(14),j}$	$z_{14} = b$
y_{21}	$x_{\pi_{j\|Z=b}(21),j}$	$z_{21} = b$
$\vdots$	$\vdots$	$\vdots$

Figure 7: Permutation scheme of unconditional and conditional Random Forest procedures

*Permutation scheme for the original marginal (left) does not take correlation of the permuted variable with other variables into consideration. The newly suggested conditional (right) permutation permutes a variable within groups of observation that it is conditional on.**

**Picture courtesy: Conditional variable importance for random forests. Strobl, C., Boulesteix, AL., Kneibi, T., Augustin, T., Zeileis, A.,* BMC Bioinformatics *2008*

X may be associated with all or some of the remaining variables. To determine the variables, Z to be conditioned on, the empirical correlation of variables with X is determined by conditional inference tests. If this correlation coefficient or P-value exceeds a certain preset threshold then $X$is conditioned on that variable. It must be noted that the conditioning variables may not necessarily appear all together with the variable of interest, X. However since a large number of trees are built, different combinations of the conditioning variables along with X may be seen in different trees. The conditioning grid for each tree is determined by the partition of that particular tree only. The permutation importance for each individual tree is then averaged over all trees. While there is still some selection bias, when choosing correlated variables over variables that are very poorly associated with outcomes, it is greatly attenuated by using the conditional approach. This bias and resultant prediction accuracy is also dependent on the value of mtry, or number of variables available at each node to partition on and can be optimized by changing mtry.

Both conditional and unconditional Random forest methods make use of OOB error rate before and after permutation to determine VIF. This is ideal for association prediction and classification when there are equal or similar number of samples in each outcome class. However when the dataset is unbalanced the RF classifier is more likely to allocate new observations to the majority class at the same time, permutation may not change the class affiliation of the majority hence blunting the overall error rate.

The error rate regards all misclassifications equally important, and does not weight the misclassification/ class by frequency. Even if the entire minority class is misclassified if the minority is only 20% the maximum error rate would be 40%.

To make the Random Forest method more relevant to unbalanced datasets, Janitza introduced the use of AUC instead of error rates(14). The AUC weights each class equally and hence is not affected by imbalance in the number of samples per class. The AUC for a tree corresponds to the proportion of concordant events, i.e. A larger class probability for $Y=1$ is assigned to an observation that truly belongs to class $Y=1$ compared to an observation that belongs to class $Y=0$.

$$VI_j^{(AUC)} = \frac{1}{ntree} \sum_{t=1}^{ntree} (AUC_{tj} - AUC_{t\tilde{j}})$$

Where ntree denotes the number of trees

AUC_{tj} denotes the area under the curve computed from the OOB observations in tree t before permuting predictor j

$ER_{t\tilde{j}}$ denotes the area under the curve computed from the OOB observations in tree t after randomly permuting predictor j.

By using the AUC, permutations that change class affiliations of either the majority or minority class give a high error rate and thus it is more robust to unbalanced datasets. For

the current study the party package in R is used, which allows for conditional partitioning and use of change in AUC instead of error rate for assigning VIF (15).

The ROC-Logistic Variable selection method:

Background:

The ROC curve developed by Hanley (16) is a powerful tool in determining the classification ability of single biomarkers. By plotting the sensitivity and specificity of the marker at different cut points, it is possible to visually assess the classification potential of the marker. However, often single markers are not very accurate classifiers but patterns in movement of multiple markers may better classify outcome. In such cases multiple marker models can be built using logistic regression. The predicted probabilities of these combination models can then be used to build ROC curves to classify outcome(17). However when a large number of potential markers are present, it can be challenging to determine an optimum model for accurate classification. It is also not geared towards variable selection for classification purposes. Data mining methods on the other hand, can be used to narrow down potential candidates, but additional analysis is still required to build an optimum model. Additionally, both data mining methods and logistic regression models test for classification performance or fit over the whole range of test values. In a clinical setting, often the performance of a test in the high sensitivity or high specificity region are more critical to its utility rather than overall performance. Neither of the above methods are capable of honing in on a classification optimal model

without extensive trial and error. The proposed new method attempts to address this issue by choosing variables based on classification metric alone, i.e maximizing the partial AUROC.

It takes advantage of the optimality of risk scores and uses a combination of empirical, and parametric methods for optimum variable selection for modelling classification.

In order to understand how the ROC-Logistic method works, it is important to understand a few existing methods in use to determine and evaluate a classification model.

The Baker method (Jagged Algorithm):

Baker(18) suggested a non-parametric method for combining multiple markers. The method takes advantage of the optimality of likelihood ratio to extend the idea of ROC cut points to multivariate positivity region.

In order to accomplish this Baker suggested the following procedure:-

Consider a two marker model with continuous markers, $\mathrm{X1}$ and $\mathrm{X2}$. Each marker is divided into n quantiles and are cross tabulated to give $n*n$ cells. The TPF for each cell is then given by $\mathrm{TPF_{ij}} = \Pr(X_1 = i, X_2 = j/Y = 1)$ and similarly $\mathrm{FPF_{ij}} = \Pr(X_1 = i, X_2 = j/Y = 0)$. Next the ratio of $\mathrm{TPF_{ij}}/\mathrm{FPF_{ij}}$ for each cell is determined. Within the limitations of a predetermined pattern, the cells are ranked in descending order of the $\mathrm{TPF_{ij}}/\mathrm{FPF_{ij}}$ ratio. The corresponding TPF and FPF of each cell is then sequentially added and a curve of cumulative TPF over cumulative FPF is plotted. This gives the ROC curve for the combination of the two markers.

Baker details three different pattern for choosing cells, Unordered, Jagged ordered and Rectangular ordered. Each method has its advantages and disadvantages (17). For the current study, we focused on the jagged ordered algorithm for choosing sequence of cells. The jagged algorithm ranks the cells based on LR and also the assumption that movement in either one of the marker is indicative of disease, i.e both markers need not move similarly. In the context of two marker combination, where higher values are indicative of disease, starting from the cell for which both markers have high values, it ranks cells such that one or the other marker is maximized. Cells can be chosen sequentially in either direction, horizontal or vertical or both, maximizing either one or both markers. However a cell can be ranked only if cells with higher values have already been accounted for, i.e. only if cells to the right (higher values of Marker B) and cells below (Higher values of Marker A) are already taken. A cell (i,j) can be ranked only if both cells $(i+1,j)$ and $(j+1,i)$ have already been ranked.

Within the above constraints, cells with higher TPF/FPF ratio are preferentially ranked over lower ratios until all the cells have been covered. The TPF and FPF of the cells are then sequentially added based on their ranking to give the ROC curve.

Consider the combination of two markers A and B split into 5 quantiles each and cross tabulated. The Jagged algorithm is used to rank the $\mathrm{TPF/FPF}$ ratio table of these two markers.

TPF/FPF

Marker A/Marker B	1	2	3	4	5
1		0.141		0	0
2		0			
3		0.318		0	0
4		0.891		0	3.714
5		8.023		4.457	5.349

The algorithm starts at cell (5,5) under the assumption that higher values are indicative of disease. It then moves to the left to cell (5,4) as the $\mathrm{TPF/FPF}$ ratio is larger than that of cell(4,5). Next it ranks cell (4,5) as amongst the adjacent cells it has the higher value. From here it next ranks cell (5,3) and then (5,2), as (5,2) has a high ratio.

The rankings of cells based on jagged ordering are given below.

Marker A/Marker B	**1**	**2**	**3**	**4**	**5**
1	25	23	21	19	17
2	24	22	20	18	13
3	16	12	11	10	9
4	15	8	7	6	3
5	14	5	4	2	1

It should be noted that cells with $\mathrm{FPF}=0 \text{ and } \mathrm{TPF}>0$ give an undefined ratio. These cells are given preference as this means an increase in sensitivity with no loss in specificity.

Cells where TPF=0, instead of the ratio the FPF value is used. Based on jagged ordered ranking the FPF and TPF of the cells are sequentially added and ROC curve plotted as shown in Figure 8.

Below are the cross tabulation of the TPF and FPF of the two marker combination.

TPF:

Marker A/Marker B	**1**	**2**	**3**	**4**	**5**
1	0	0.0571	0	0	0
2	0	0	0	0	0
3	0	0.0571	0	0	0
4	0	0.114	0	0	0.143
5	0	0.257	0	0.029	0.343

FPF:

Marker A/Marker B	**1**	**2**	**3**	**4**	**5**
1	0	0.404	0	0.006	0.006
2	0	0.058	0	0	0
3	0	0.179	0	0.032	0.026
4	0	0.128	0	0.019	0.038
5	0	0.032	0	0.006	0.064

Cumulative FPF and TPF based on Jagged Algorithm

FPF	TPF	Cumulative FPF	Cumulative TPF

0.064103	0.342857	0.064103	0.342857
0.00641	0.028571	0.070513	0.371429
0.038462	0.142857	0.108974	0.514286
0	0	0.108974	0.514286
0.032051	0.257143	0.141026	0.771429
0.019231	0	0.160256	0.771429
0	0	0.160256	0.771429
0.128205	0.114286	0.288462	0.885714
0.025641	0	0.314103	0.885714
0.032051	0	0.346154	0.885714
0	0	0.346154	0.885714
0.179487	0.057143	0.525641	0.942857
0	0	0.525641	0.942857
0	0	0.525641	0.942857
0	0	0.525641	0.942857
0	0	0.525641	0.942857
0.00641	0	0.532051	0.942857
0	0	0.532051	0.942857
0.00641	0	0.538462	0.942857
0	0	0.538462	0.942857
0	0	0.538462	0.942857
0.057692	0	0.596154	0.942857
0.403846	0.057143	1	1

0	0	1	1
0	0	1	1

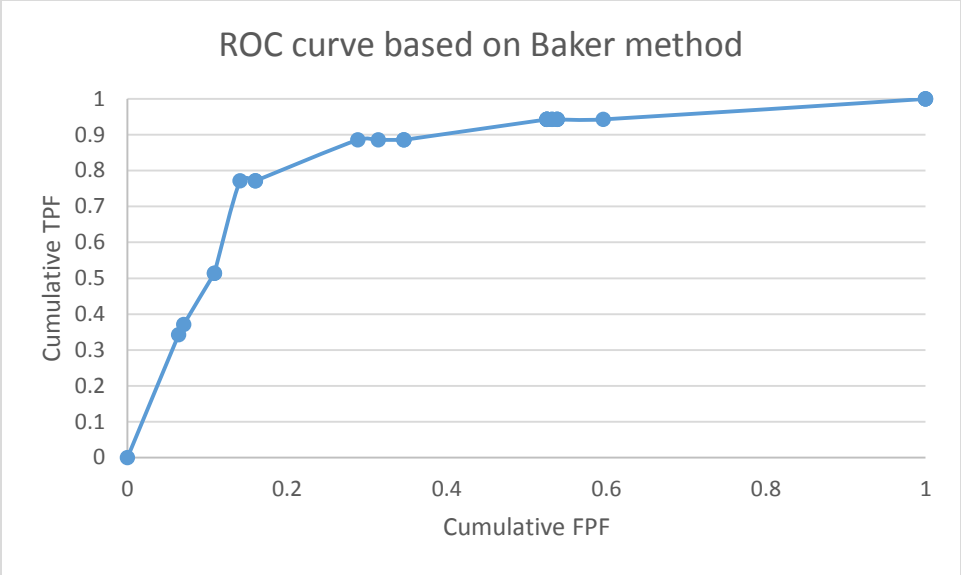


Figure 8: ROC curve based on Baker Jagged ordered Algorithm

The ROC curve is built by plotting cumulative TPF over cumulative FPF. The order of TPF and FPF is determined by ranking the TPF/FPF ratio obtained by cross-tabulating quintiles of two markers. This ranking is based on the Jagged ordered algorithm suggested by Baker. AUROC=0.846

For more than two markers Baker recommended higher dimensional tables. i.e. three dimensional cross-tabulation for 3 markers and so on. The number of groups for tabulation was also suggested by number of markers to be combined. However combining markers in this multi-dimensional manner is not feasible for large number of potential markers(8), hence the current method recommends a modified method using predicted probability for multiple markers.

IDI:

Integrated Discrimination Improvement (IDI) index is a tool to evaluate the benefit of adding a single or multiple variables to an existing classification model(19). It is complementary to the AUC in principal and is defined as

$$IDI = (IS_{new} - IS_{old}) - (IP_{new} - IP_{old})$$

- Where IS is the integral of sensitivity and IP is the integral of specificity.
- The subscript old is used for the original model (nested) and the subscript new is used for the model with additional variables.

Alternatively, the IDI can be easily determined by the below equation.

$$\widehat{IDI} = \left(\overline{\hat{p}_{new,cases}} - \overline{\hat{p}_{old,cases}}\right) - \left(\overline{\hat{p}_{new,ctrls}} - \overline{\hat{p}_{old,ctrls}}\right)$$

Where $\hat{p}$ is the predicted probability

The IDI is intuitive to understand. If adding a variable improves the classification ability of the model then the predicted probability of the cases as determined by the new model should be higher than that determined by the original model. The better the movement in cases the greater will be the difference of predicted probability in the first part of above equation. Alternatively the new model may improve classification by lowering the predicted probability of controls as compared to the original model. The better the new model in this regard the more negative the second term in the equation. Thus improving the overall IDI.

In the current method IDI is used as an exclusion criteria. If adding a new variable does not improve IDI, the variable is not included in the model. It should be noted that IDI is dependent on initial model as well as case-control ratio and hence the IDI threshold to be used for each dataset should be optimized.(20)

The Z value for IDI is given by

$$z_{IDI} = \frac{\widehat{IDI}}{\sqrt{(\widehat{SE}_{cases})^2 + (\widehat{SE}_{ctrls})^2}}$$

However it has been shown that for IDI approaching zero the distribution may not be normal and the z-score may not be ideal (21). In the current method we use an IDI cut off, and do not evaluate if it is significantly different from zero.

The ROC-Logistic variable selection process:

The ROC-Logistic method is an amalgamation of classification methods that hones in on a model that maximizes partial AUROC. While clinical development of any diagnostic test is

focused on enhancing sensitivity and specificity over specific ranges, none of the existing variable selection method chooses based on partial AUROC. The suggested variable selection method combines different existing methods to optimize the variable selection process, as detailed in Figure 9

The first step is to determine the partial AUROC of each marker at a predefined clinically relevant FPF range. The potential marker that has the highest pAUROC, (corrected for negative correlation by multiplying by -1) is the first variable into the model (denoted as Marker A). If two potential markers have identical pAUROC's then the model which has better classification potential over the entire range of values is selected. This is accomplished by using the Mann-Whitney U test statistic. The Mann-Whitney U is a nonparametric rank-based test used to evaluate the separation of two distributions, in the current instance, the separation of cases from controls. The Mann-Whitney U statistic is equivalent to the empirical ROC curve, thus it represents the performance of a variable as a classifier. However in the current method, the pAUROC tie-breaking selection was done based on P value rather than actual statistic so as to not miss good classifiers that are negatively correlated to outcome and hence would have an AUROC lower than 0.5 and subsequently higher Mann-Whitney U statistic .

Next the Baker method, with jagged ordered multivariate optimization algorithm is utilized to build the partial ROC of all binary models with the initially chosen variable, Marker A. For the current method, the number of groups to divide each variable into was set to 5. We found a smaller number of groups gave a very irregular stepped curve, as the

ROC was plotted with fewer points. Also the resultant AUROC was not comparable to that arrived at by using predicted probabilities from logistic regression. Larger number of groups made the cross-tabulation sparse, introducing subjectivity in ranking cells and minor inconsistencies. It is possible that larger groups would work better for a larger sample sizes, or in a dataset with less skewed variables. The slope of line from origin to TPF corresponding to a specific FPF is determined for every binary model. This FPF is user-defined, based on the clinical relevant AUROC region to maximize. The variables are then reordered based on decreasing slope. This, ensures that variables which have a higher sensitivity at relevant specificity when combined with the first variable get tested for inclusion first.

The variable are then sequentially combined with the initial variable one by one to check for improvement in partial AUROC. If the pAUROC is improved, the IDI is computed. If the IDI is above a preset threshold then the variable (say, Marker B) is added to the model. The IDI threshold is also user-defined, and may need optimization for different sized datasets. The predicted probability of this new combination model is determined using logistic regression.

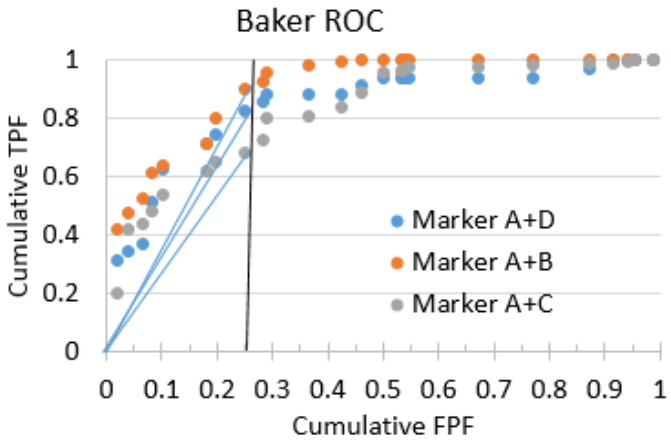
$$\hat{p}_{A+B} = \frac{e^{\beta_0+\beta_A+\beta_B}}{1 + e^{\beta_0+\beta_A+\beta_B}}$$

The next variable in the decreasing slope order will now be combined with this predicted probability using Bakers jagged ordered algorithm and the resultant new model will be checked for improvement in partial AUROC over the previous combination model. Instead

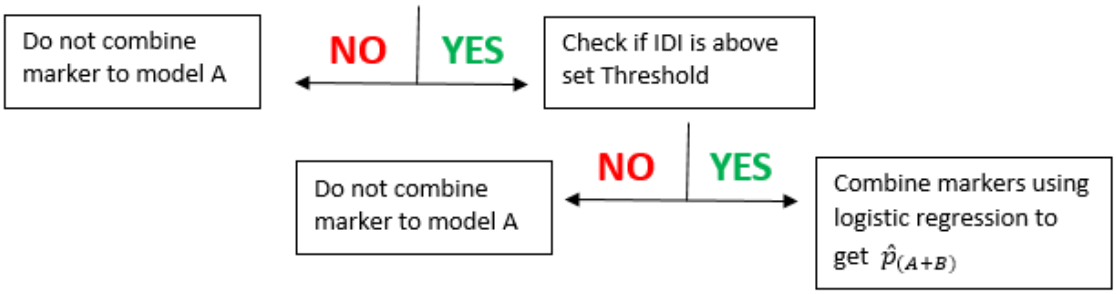
of using the actual values of the markers in the model, the predicted probability are quintiled and cross tabulated with next variable to be checked. Thus solving the multi-dimension issue of using the Baker method for multiple variable modelling. If the next variable in sequence does not improve partial AUROC, then it is ignored and the process checks the variable after that for inclusion. In such a manner, the process is continued until all variables are checked for improvement in partial AUROC and IDI. This classification targeted approach is quick and effective in choosing variable combination that are parsimonious and have good classification properties.

1. Choose the marker with the highest pAUROC, (Marker A) as the first marker

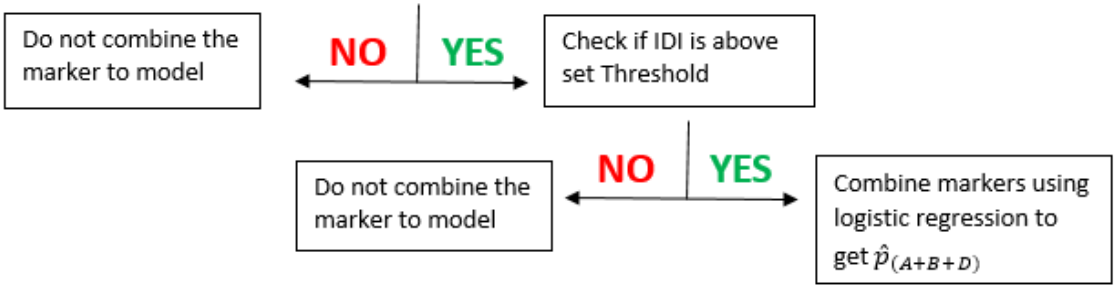
2. Test Baker binary combination of each marker with marker A and order the markers wrt to slope



3. Check if the Baker pAUROC of the best binary combination [A+B] is better than that of model A alone



4. Check if the Baker pAUROC of the next marker[say D] + predicted probability of the model so far, is improved



5. Repeat step 4 until all markers have been checked.

Figure 9: The ROC- Logistic variable selection procedure

CASE STUDY:

Preterm birth, defined as births at less than 37 completed weeks of gestation, is the leading cause for neonatal mortality and morbidity both globally and in the US (22). Women may be at risk for preterm birth due to a number of factors, including pregnancy complications such as preeclampsia and fetal birth defects.(23). However in two thirds of preterm births there is a spontaneous onset of preterm labor. Numerous factors have been shown to be associated with spontaneous preterm labor, including vascular disorders, cervical disease, stress, intra-amniotic inflammation or infection (24). Of the above, intra-amniotic inflammation is a known risk factor for spontaneous preterm labor(25). Intra-amniotic inflammation, represented by high concentration of inflammatory proteins in the amniotic fluid, could be due to a microbial invasion of the amniotic cavity (MIAC) or, in certain cases, inflammation in the absence of infectious agents. The latter has been termed as sterile intra-amniotic inflammation. In the examples presented in the current investigation, no differentiation was made between MIAC and sterile intra-amniotic inflammation, as in either case the adverse neonatal outcomes were comparable (3,26). For the current study, women with amniotic interleukin-6 (IL-6) levels greater than or equal to 11.3ng/ml were considered positive for intra-amniotic inflammation (3).

When a woman presents with spontaneous preterm labor, the course of treatment can be critical to the subsequent mortality and/or morbidity of the neonate.

The decision is a complicated one for obstetricians, because often the etiology of presented preterm labor is unknown. If there is an infection or inflammation, treatment with antibiotics or anti-inflammatory steroids is recommended,(27) however indiscriminate use of antibiotics may be harmful to the neonate(28). This makes accurate diagnosis critical to course of therapy. Unfortunately, quick and safe procedures to check the status of the mother are limited. Analyzing the amniotic fluid by amniocentesis for IL-6 is the most widely used method to check for inflammation; however, amniotic fluid must be sampled and tested, which is an invasive procedure. In the absence of a reliable, safe and non-invasive diagnostic procedure, studies to determine the best course of action are limited, and there is no consensus on the right treatment course to follow for women in spontaneous preterm labor(29). A rapid, safe and accurate diagnostic tool to aid in the detection of intra-amniotic inflammation is thus critical to improved understanding of preterm labor etiology, and consequently the course of action that best alleviates the mortality and morbidity associated with preterm labor. In recent years, short cervix, gestational age and maternal serum white blood cell count, have all been considered as non-invasive diagnostic tools for intra-amniotic inflammation (30,31,32). Another option being considered are examination of protein levels the cervicovaginal fluid (CVF). The CVF is obtained by swabbing the posterior vaginal fornix for 60secs, a non-invasive procedure and, thus presents a relatively safer, quicker and easier procedure to determine the inflammation status of the amniotic sac. The CVF consists of numerous proteins, many of which could be indicative of intra-amniotic inflammation. A recent study has shown the possible benefit of using CVF IL-6 or interleukin-8 (IL-8)

concentrations as potential markers for intra-amniotic inflammation in a limited dataset(33). A different study examined CVF fetal fibronectin as a marker for upper genitourinary tract infection and a predictor, along with intra-amniotic inflammation, as predictors of imminent birth(34). A combination of multiple CVF protein measures may serve as a multi-marker panel to detect intra-amniotic inflammation, improving upon the sensitivity and specificity of individual CVF proteins, and result in a clinically-applicable diagnostic tool for detecting intra-amniotic inflammation. Analyzing combination of markers could be key to developing a diagnostic test based on CVF.

METHODS:

Sample collection:

To develop and validate the presented methodology, we use data from a multicenter prospective study conducted in multiple phases. For the current study, data was restricted to a subset of the first two phases of the study. A large number of candidate CVF proteins were screened in Phase 1 of the original study to determine potential intra-amniotic inflammation markers. Phase 2 was restricted to proteins that passed initial screening criteria, and were further refined and their classification ability estimated. In this examination, we first explore the proposed new methodology in Phase 2 due to the smaller dimension (number of candidate proteins), and then examine Phase 1 where 81 CVF protein measures were studied. Therefore, we note that nine CVF proteins (i.e. candidate markers) were advanced from Phase 1 into Phase 2 based on AUROC, robustness, biologic plausibility, and simplicity of performing the assays. This restricted data set of 9 CVF proteins is further referred to as set 1. Following this, the Phase 1 data which had a larger number of potential markers was used to study the method performance in the presence of noise, further referred to as set 2.

The data was collected at 16 tertiary-care perinatal centers across the United States. Women above 18yrs of age, with singleton pregnancies and in spontaneous preterm labor with intact fetal membrane, were eligible for the study. Preterm labor is

defined as regular uterine contractions with either a cervical dilation of ≥ 2cm or a cervical length, as measured by transvaginal sonography, of ≤ 30mm or a positive cervicovaginal fetal fibronectin test. Cases with ruptured fetal membrane, major fetal anomaly, fetal aneuploidy or a medical indication for preterm birth were excluded from the study. Consented and enrolled eligible women underwent amniocentesis as part of their routine care, with an extra aliquot collected for the study. Amniotic fluid IL6 levels were measured and an IL6 concentration of 11.3ng/ml or higher was considered positive for intra-amniotic inflammation and served as cases. Additionally CVF was collected to measure proteins that could potentially serve as a marker for intra-amniotic inflammation. Further management of preterm labor and other clinical decisions were independent of the study and was left to the caregiver's discretion.

Statistical Analysis:

All CVF protein measures had a highly skewed non-normal distributions in all women (regardless of inflammation status), hence the values were converted to the natural) log form for all further analysis. Additionally, a number of CVF proteins in set 1 were also below the lower limit of detection.

The distribution of cases (women with amniotic fluid IL-6 levels ≥11.3 ng/ml) and controls

(<11.3 ng/ml) were compared using the Mann- Whitney U test, a nonparametric rank-based test to determine differences in the distribution of two groups. The overall AUROC and the pAUROC over the FPF region 0 to 0.25 for each CVF protein measure was determined using the ***comproc*** procedure in Stata (35) and ***roccontrast*** option in ***proc***

logistic in SAS/STAT® software (36) . Additionally both procedures were also used to compare model performances for AUROC and pAUROC. The ***comproc*** procedure in Stata calculates a confidence interval for the difference between ROC summary indices. The observed difference is divided by its standard error and is compared to the standard normal distribution. Confidence intervals and standard errors are derived from the bootstrap distribution of the estimators.

The ***roccontrast*** statement was specified under ***Proc logistic*** to compare different ROC models using the SAS/STAT software. The reference was specified in each case and used with the nofit option. This produces a contrast matrix of differences between each ROC curve and the specified reference curve following a logistic regression. The ***Proc logistic*** process was also used to determine AIC of each model to compare the goodness of fit of all models to each other as AIC allows comparison of non-nested models too.

For the first study, using set 1 data, the ROC-logistic method result was compared to the models built with best subset variable selection technique which is based on the score statistic, a measure of the slope of the likelihood function. Models with high score statistic based on best subsets were fit by logistic regression and the AUROC and pAUROC estimated. This dataset also had a high concentration of early onset preterm labor (defined here as ≤ 34weeks). Analysis of the model performance in this early onset subset was also done. Additionally, gestational age was added as a covariate to the model to check for improvement in classification performance.

Finally, as there is no definitive method to classification modeling and the best subset selection does not always select the best model in terms of classification, the model was compared to all possible combination models to evaluate classification performance.

For the next step of method study, the ROC-Logistic method was used to determine optimum classification model in a noisier dataset with more potential candidates. To compare the performance of the model, a combination of traditional methods of data mining and best subset selection was used. As this dataset had 81 measures, it was not feasible to use Best subset selection directly and hence a data mining procedure, random forest was used as an initial step to select good classifiers. This was followed by best subset selection on this selected group of potential classifiers. The final models were compared to the ROC method model for overall AUROC and partial AUROC. The exact parameters used and a brief description of methods follow:

ROC–Logistic method applied to set 1 and set 2 of intra-amniotic inflammation data set: The partial AUROC was determined for all individual protein markers over the FPF range of 0 to 0.25. Variables were ordered in decreasing order of partial AUROC followed by increasing p-value for the Mann Whitney U-statistic. The variable with the largest partial AUROC was set as the first marker into the classification model. Each of the remaining variables were individually combined with the first (best classifying) variable using the jagged ordered multivariate optimization algorithm developed by Baker (18). The slope of line drawn from the origin to the FPF=0.25 point of the resultant ROC was determined. The variables were reordered in descending order of this slope. As the Baker method does not involve fitting the entire curve for the determining the TPF at a set FPF, this is a very

efficient and time saving step that ensures that variables that combine better with initial variable get the opportunity to be added sooner rather than later. Each marker was then combined, individually, in order of slope, with the initially selected variable to check for improvement in the partial AUROC via Baker method. The variable was retained in the model if the resultant partial AUROC of combination was higher than the first variable alone and, the IDI of this new combination model was above a preset IDI threshold of 0.05. If both criteria were not met the variable was not included in the model and the next variable in the sequence was checked. Variables that, on combination improved pAUROC and crossed the IDI threshold were combined using the logistic regression to give the predicted probability. The next marker to be checked will now be combined with this predicted probability using Baker method. By using the logistic regression instead of Baker for combining selected variables the multidimensionality issue of Baker is avoided. In this way each variable was sequentially checked for inclusion into the model built by improvement in partial AUROC. This forward selection ensures that there is no markers get added into the model if the pAUROC is not improved, this ensures that the final model is parsimonious, while also have good classification properties in the relevant FPF range. The code for the entire process was written in R and is attached.

Best Subset Selection:

As there are no gold standard methods to select variables based on classification ability, we used best subset selection as a surrogate method for choosing classification models. Best subset selection is a variable selection procedure that selects models with highest likelihood score (chi-square) statistic from all possible models with 1 or more variables

(parameters) included. That is, the procedure selects "best" 2-marker, 3-marker, 4-marker, etc. models. The number of models per parameter level was restricted to 5 for this investigation. As both set 1 and set 2 had less than 200 samples with 18% to 24% cases, analysis was restricted to models with a maximum of 4 parameters (4 proteins) to avoid over-fitting and to ensure reliability in classification. The ***Proc logistic*** procedure in SAS v 9.3 was used to build the best subset models using the score statistic selection criteria. The AUROC, pAUROC and AIC values were than determined for the selected models to allow comparison with the model selected by the ROC-logistic regression based variable selection method.

Random Forest:

The 'party' package in R was used to determine optimum classifiers of intra-amniotic inflammation based on 81 CVF proteins measured in Phase 1 (set 2). The advantage of using this package was twofold. First, it allows us to account for correlation between measured CVF proteins, which is critical in any biological data set. Second, it allows for determining the prediction accuracy based on change in 'Out of Bag' (OOB) AUROC instead of using the traditional, OOB Error rate. OOB error rate has suboptimal accuracy in determining the variable importance factor (VIF) when used in unbalanced data sets (14). The set 2 dataset has a case to control ratio of 0.317 and hence determining VIF based on AUROC was suitable.

Intra-amniotic Inflammation status was coded as factor and controls for ***cforest*** was set as unbiased. The following default parameters were used: The teststat used to determine which variable to split on was set to quadratic and testtype, which specifies the

computation of distribution of test statistic was set to univariate. Both the above settings ensure unbiased split. The number of observations to draw per sampling was set to 0.632 of the data, without replacement.(10, 12, 37)

The number of trees, ntree was set to 2000 and mtry was set to 9, i.e $\sqrt{81}$. The parameters mincriterion, minsplit and minbucket were set to zero to grow conditional inference trees to maximal depth with no pruning. varimpAUC was determined under conditional=TRUE, with a threshold set to 0.2, i.e Only, if the 1 - p-value of the association between the variable of interest and a covariate is less than 0.2, will they be considered as independent of each other. The number of permutations was set to 1. OOB was set to true, so that the VIF would be calculated based on OOB data. The process was repeated 7 times with different seed to ensure diverse trees were built. The variables that had the highest VIF were consistent through the 7 runs.

RESULTS:

Set 1 results

Cervicovaginal fluid collected from 191 patients, was analyzed and 9 different proteins concentrations known to play an active role in intra-amniotic inflammation were measured (Table 3a). Additionally the gestational age at onset of preterm labor was included in the dataset. This cohort of patients with measures on limited number of CVF proteins, set 1, was an ideal starting point to test the efficiency of the ROC-Logistic method. Of the 191 patients, 35 had elevated IL-6 (≥11.3 ng/mL) in the amniotic fluid, indicative of severe intra-amniotic inflammation. The remaining 156 patients served as controls. A high proportion of controls and cases had measures below the lower limits of detection (LOD) for each of the CVF proteins, and most of the distributions were heavily skewed to the right. Hence, all measures of proteins were converted into log scale to correct some of the skewness. The distribution of cases and controls differed significantly from each other in 8 of the 9 CVF proteins as determined by the Mann Whitney U test (Table 3b). The mean gestational age at onset of preterm labor was also significantly different between the two groups with cases having a lower mean of 27.86 weeks, as compared to controls at 31.33 weeks. The area under ROC curve (AUROC) for each of the CVF proteins was computed. Additionally the AUROC of gestational age was calculated as 0.7332. The partial AUROC was calculated over the FPF range of 0 to 0.25. CVF IL6 had the

highest AUROC at 0.8612 (0.7902-0.9322) while within FPF 0.25, the α-fetoprotein, AFP had the larger partial AUROC at 0.1540. Note: When using STATA the partial AUROC of AFP was slightly lower at 0.147 (0.108-0.186), this could be because of how the different software, STATA against R, handle ties in their estimation of AUROC.

Table: 3a Geometric means and lower detection limits of 9 CVF proteins and arithmetic mean of gestational age at onset of preterm labor.

			Inflammation Absent (n=156)		**Inflammation Present (n=35)**	
Variables	Protein abbrv.	LOD	Geometric mean(std dev)	% freq at LOD	Geometric mean(std dev)	% freq at LOD
Gestational Age (wk)			31.33 (3.67)*		27.86 (3.96)*	
α-1-Acid Glycoprotein	A1AG	312.50	719.81 (2.63)	41.67%	1596.29 (3)	11.43%
α-fetoprotein	AFP	15625	16466 (1.31)	94.23%	43537 (3.09)	37.14%
β2 microglobulin	B2MG	6250	18465 (2.89)	38.46%	54462 (3.14)	14.29%
Chemokine Ligand 1	GROα	312.50	670.1 (2.35)	26.28%	548.93 (1.8)	17.14%
Insulin-like growth factor-binding protein 1	IGFBP1	313.00	429.07 (1.88)	66.03%	873.52 (3.15)	42.86%
Interleukin 6	IL6	62.60	149.26 (2.63)	41.67%	713.77 (2.67)	5.71%
lipopolysaccharide binding protein	LBP	39.00	160.01 (4.09)	28.85%	349.57 (5.34)	11.43%
Chemokine Ligand 2	MCP1	31.25	38.9 (1.73)	80.13%	68.65 (2.5)	48.57%
Chemokine Ligand 4	MIP-1β	312.50	354.27 (1.38)	76.92%	540.53 (2.17)	51.43%

***Arithmetic mean (std dev)**

Table: 3b Difference in distribution of Cases and Controls for each of the CVF proteins

Variables	Mann Whitney U (p value)	AUROC *(95% CI)	pAUROC(0.25)* (95% CI)
A1AG	1554 (<0.0001)	0.715 (0.625-0.806)	0.075 (0.038-0.111)
AFP	1129.5 (<0.0001)	0.793 (0.709-0.878)	0.147 (0.108-0.186)
B2MG	1375 (<0.0001)	0.748 (0.652-0.844)	0.106 (0.069-0.146)
GROα	2874 (0.6253)	0.526 (0.431-0.622)	0.012 (-0.003-0.027)
IGFBP1	1805 (0.0005)	0.669 (0.566-0.773)	0.100 (0.062-0.138)
IL6	758 (<0.0001)	0.861 (0.790-0.932)	0.151(0.114-0.188)
LBP	1990 (0.0124)	0.636 (0.530-0.741)	0.065 (0.032-0.099)
MCP1	1768.5(<0.0001)	0.676 (0.581-0.771)	0.096 (0.060-0.132)
MIP-1β	1819 (<0.0001)	0.667 (0.572-0.762)	0.092 (0.058-0.127)

* From STATA –using **comproc**.

While IL6 had good classification performance with an AUROC of 0.861 (0.790-0.932), a combination of CVF proteins may perform better. Two different methods were compared for their ability to determine an optimum combination of the proteins that would accurately classify intra-amniotic inflammation.

The ROC-Logistic method outlined previously was used to build a classifying model by maximizing the partial AUROC in the FPF 0 to 0.25 region. For a test to be clinically applicable it has to fit certain sensitivity and specificity criterion, which often lie within this region of the ROC curve and hence maximizing this partial area may enhance the clinical utility of the test. Using a combination of the Baker algorithm and logistic regression, the method chose the combination of AFP and IL6 that maximizes partial AUROC, while also meeting the IDI cut off set at 0.05.

The AUROC of the final model, including IL6 and AFP, was 0.8804 via logistic regression, and 0.8866 when using the Baker jagged-algorithm. The pAUROC over the FPF range 0-0.25 was equal to 0.1711 and 0.1784 using the logistic and Baker approach respectively. The model characteristics are presented in Table 5.

Early onset preterm labor, defined in this study as labor on or before 34 weeks, had a higher incidence of inflammation in the current cohort as is the case in general(25). It is probable that the etiology of early and late onset preterm labor are different and consequently may present different markers. We checked the above model in the early onset subset to determine if the combination of AFP+IL6 performs equally well in this

subset. In the current data set we have 150 patients who had early onset preterm labor, 31 of whom had inflammation, detailed in Table 4.The model performed equally well with an AUROC of 0.8786 and 0.8806 using Logistic and Baker method respectively. The pAUROC were comparable as well.

Table 4: Distribution of cases and controls in early and late onset preterm labor

	Early Onset preterm labor	Late onset preterm labor	Total
No inflammation	119	37	156
Inflammation	31	4	35
Total	150	41	191

Table 5: ROC-Logistic based method model 1

	AFP+IL6	
Model	**Entire data set (n=191)**	**Early onset PTL (n=150)**
AUROC_logistic	0.880	0.879
AUROC_Baker	0.887	0.881
pAUROC_logistic	0.171	0.170
pAUROC_Baker	0.178	0.175

The Baker method is a simple procedure of cross-tabulation of markers and ranking the TPF/FPF ratio of each cell. The cumulative TPF and FPF based on the ranking is used to build the ROC curve and determine the pAUROC and AUROC. The logistic method involves fitting the marker combinations to determine the predicted probability of outcome based on marker combination. These predicted probabilities are then used to build the ROC curve and determine pAUROC and AUROC.

Gestational age is associated with inflammation, hence, we checked if including gestational age into the AFP+IL6 model improves the classification accuracy. When added, the overall AUROC improved from 0.8804 to 0.9086. The partial AUROC was improved from 0.1711 to 0.186 (Table 6a). The IDI on addition of gestational age improved by 0.056. The new model, however, was not statistically significantly different when compared to the initial model containing only the CVF proteins IL6 and AFP alone (Table 6b). Adding gestational age as covariate selectively increased the predicted probability of majority of early onset cases as well as controls. Thus separating early and later-onset preterm labor rather than the presence of intra-amniotic inflammation (Figure 10).

Table 6a: ROC-Logistic based method model 2

	AFP+IL6+gestational Age	
Model	**Entire data set (n=191)**	**Early onset PTL (n=150)**
AUROC_logistic	0.909	0.898
AUROC_Baker	0.920	0.898
pAUROC_logistic	0.186	0.186
pAUROC_Baker	0.185	0.189

Table 6b: Comparison of model with and without gestational age as an independent variable

	Entire data set (n=191)		Early onset PTL (n=150)	
Model	**AUROC**	**pvalue**	**AUROC**	**pvalue**
IL6+AFP	0.880 (0.810-0.951)	0.079	0.899 (0.832-0.965)	0.798
IL6+AFP+GA	0.909 (0.853-0.964)		0.897 (0.830-0.947)	

In the current data set, gestational age was significantly associated with inflammation only in early onset preterm labor subset (beta=-0.256; pvalue <0.0001). However we did not see significant improvement in classification on adding gestational age to this subset either.

We noted that CVF AFP and gestational age are highly correlated (-0.3763 ;pvalue<0.0001) in the current data set (including early onset subset (-0.3671; pvalue<0.0001)) the utility of gestational age in classifying inflammation becomes redundant

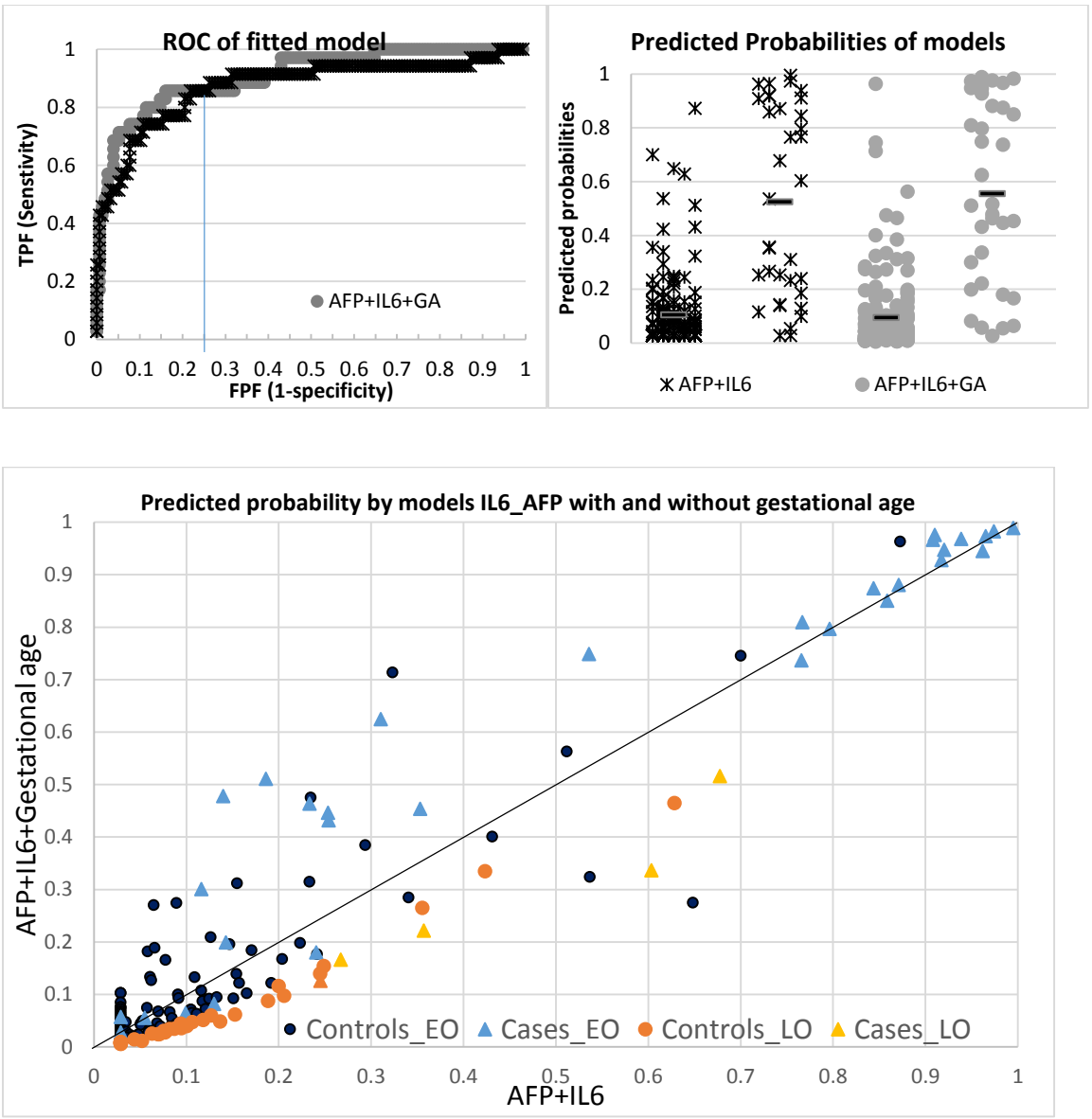


Figure 10: Comparison of predicted probability of intra-amniotic inflammation

The predicted probability of CVF IL6 and AFP combination models with and without gestational age are plotted to show how adding gestational age to the model increases the predicted probability of both early onset cases and controls.

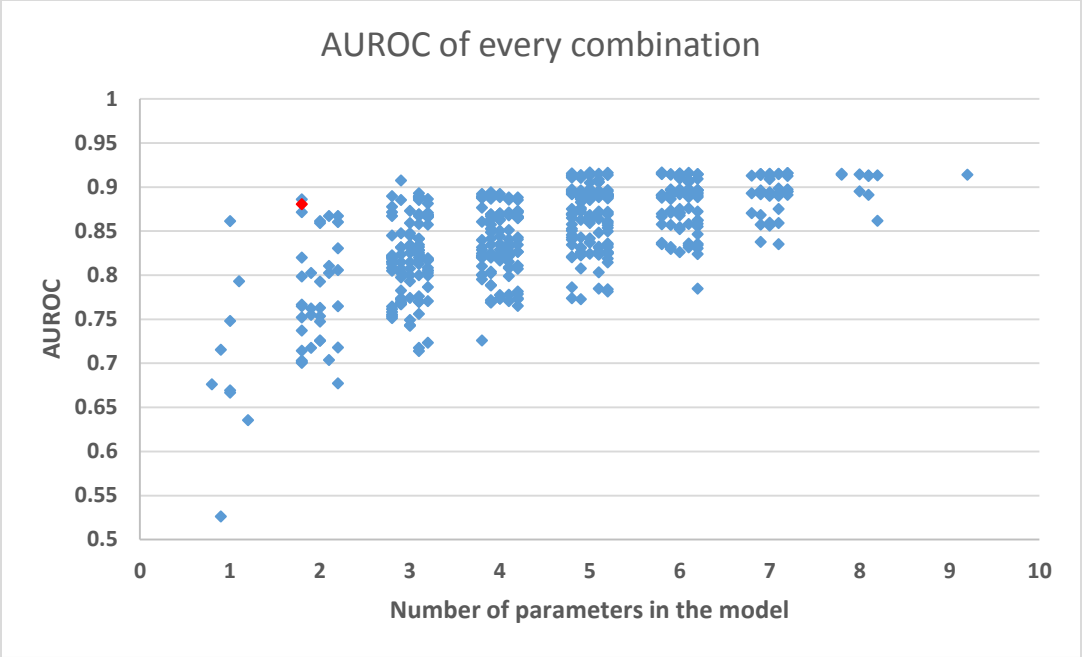
To validate the results from ROC-logistic method we compared the model built to models built by best subset selection. This method is traditionally used for small number of variables as in the current case. Best subset selection generates multiple models at increasing number of parameters that maximizes the chi square score statistic. For the current analysis, the number of models at each parameter level was restricted to 5 (all models in Supplementary Table 1 – Appendices) so that we examined a total of 41 models for 9 CVF proteins. Of all the combination models built by best subset selection, two models comparable to the IL6+AFP in terms of AUROC were AFP+GROα+IL6 and AFP+ B2MG+GROα+ IL6, both models were however, not statistically significantly different. Table 7 shows the top model chosen by best subset selection method based on the Chi-square score statistic

Table 7: Best subset selection models for set 1

No. of Vars	Models	Chi-Square Score	AUROC (95% CI)	Pvalue	AIC	pAUROC (95% CI)	Pvalue
1	MIP1b	23.42	0.667 (0.572-0.762)	<.0001	167.36	0.092 (0.0567-0.128)	<0.0001
1	IGFBP1	22.87	0.669 (0.566-0.773)	<.0001	167.04	0.100 (0.064-0.136)	0.00015
1	B2MG	25.31	0.748 (0.652-0.844)	<.0001	159.65	0.106 (0.068-0.144)	<0.0001
1	AFP	63.27	0.793 (0.709-0.878)	0.0141	133.70	0.147 (0.108-0.186)	0.078
1	IL6	53.94	0.861 (0.790-0.932)	0.0515	131.40	0.151 (0.113-0.189)	0.041
2	AFP+MIP1b	65.20	0.793 (0.698-0.887)	0.0178	134.30	0.153 (0.114-0.192)	0.16
2	AFP+B2MG	65.36	0.803 (0.705-0.901)	0.0079	133.76	0.154 (0.116-0.193)	0.15
2	AFP+ IGFBP1	66.99	0.810 (0.717-0.904)	0.0509	133.28	0.160 (0.123-0.196)	0.29
2	AFP+ MCP	66.35	0.830 (0.747-0.914)	0.0872	133.13	0.162 (0.125-0.199)	0.42
2	**AFP+ IL6**	**77.75**	**0.880 (0.810-0.951)**	**Ref**	**118.59**	**0.171 (0.137-0.206)**	**Ref**
3	AFP+IL6+ MCP	77.81	0.878 (0.807-0.949)	0.5646	120.02	0.168 (0.133-0.203)	0.47
3	AFP+IL6+LBP	80.41	0.889 (0.829-0.950)	0.4586	117.15	0.172 (0.138-0.206)	0.87
3	A1AG+ AFP+ IL6	80.42	0.890 (0.828-0.951)	0.3697	115.67	0.170 (0.134-0.206)	0.86
3	AFP+ B2MG+ IL6	79.03	0.893 (0.833-0.953)	0.2216	118.02	0.172 (0.139-0.206)	0.83
3	AFP+ GROα+ IL6	82.05	0.907 (0.855-0.960)	0.0490	112.91	0.178 (0.146-0.211)	0.39
4	AFP+ GROα+ IL6+ MIP1b	82.14	0.905 (0.852-0.959)	0.0713	114.63	0.178 (0.145-0.211)	0.45
4	AFP+ GROα+ IL6+ MCP	82.10	0.906 (0.853-0.959)	0.0620	114.59	0.177 (0.144-0.210)	0.47
4	AFP+ B2MG+ GROα+ IL6	82.29	0.912 (0.863-0.961)	0.0491	114.15	0.178 (0.144-0.212)	0.40

4	AFP+ GROα+ IL6+ LBP	83.63	0.914 (0.868-0.961)	0.0472	112.93	0.178 (0.144-0.212)	0.41
4	A1AG+AFP+ GROα+IL6	84.15	0.915 (0.867-0.963)	0.0493	111.07	0.150 (0.146-0.213)	0.39

Both methods gave comparable results, however the chi square score statistic used by the Best subset selection method does not always track with classification improvement, hence it is possible to miss good classifier models. In the absence of a gold standard method and because we had a limited number of CVF proteins in set 1, the current model of AFP+IL6 was compared with every possible combination of the 10 potential markers to check if there were better models than that chosen by presented method or best subset method. A total of 511 models were built based on every possible combination of the 9 CVF proteins and were analyzed for classification performance. We compared the AUROC of all the models individually with IL6+AFP model. Only 7 out of 511 models were significantly better at alpha=0.05, however when restricted to a maximum 4 variables/ model, i.e. 255 models, no additional combination models were significantly better than IL6+AFP. Thus the ROC-logistic model was efficient in terms of both pAUROC and parsimony as shown in Figure 11. Of all possible models, the only model which was comparable to the one chosen by ROC-logistic method in terms of AUROC and partial AUROC, while also being parsimonious was the model AFP+IL6+GROα. However as mentioned before it was not statistically significantly different in partial AUROC. The AFP+IL6 model chosen by ROC-logistic method also had a good fit based on AIC. (Figure 12).



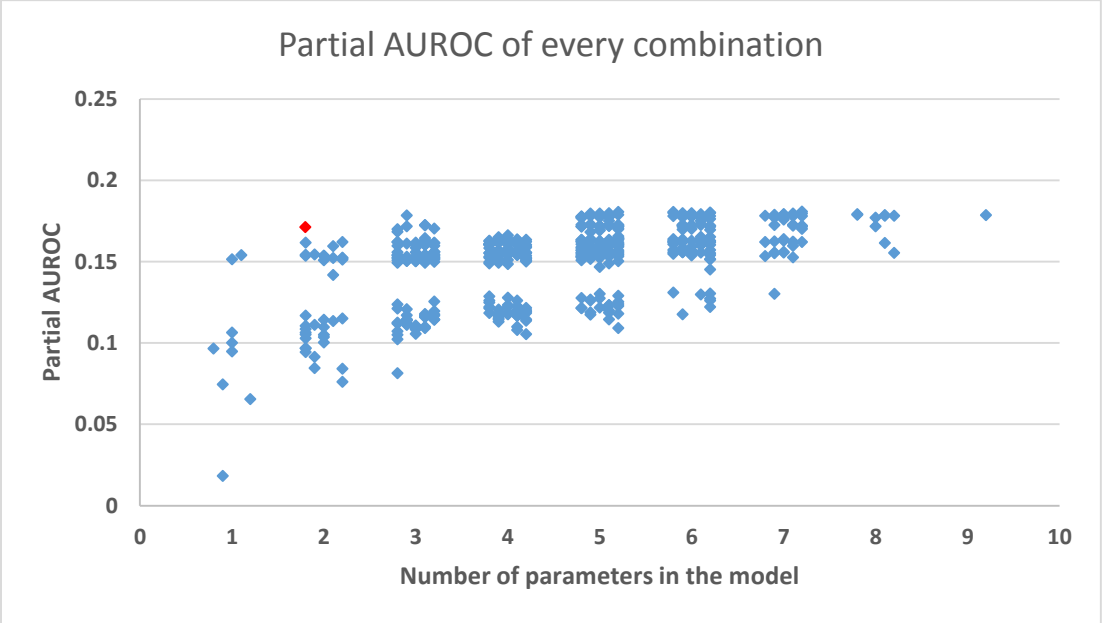


Figure 11: AUROC and pAUROC of models of all possible combinations of nine CVF proteins

The point in red corresponds to model AFP+IL6, chosen by ROC-Logistic variable selection method. The partial AUROC of the model is comparable to that of higher models, while it itself has only two variables.

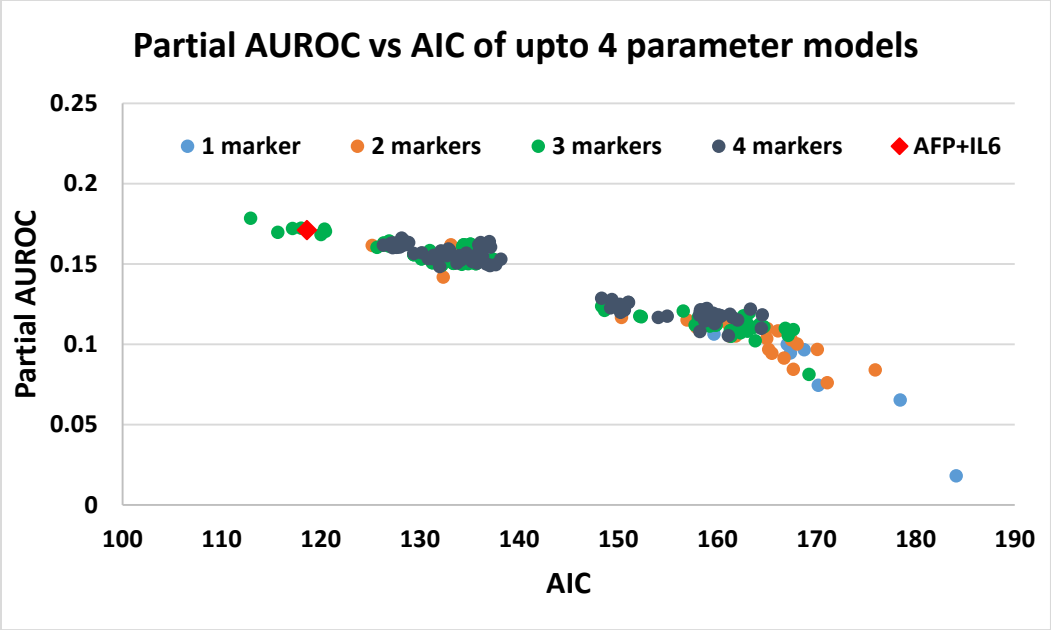


Figure 12: pAUROC and AIC of all models of all possible combinations of nine CVF proteins

The model chosen by the ROC-Logistic method is optimal in AUROC, pAUROC and AIC.

Table 8: Selected models with high AUROC and partial AUROC is compared to model chosen by the ROC-Logistic method.

No. of vars	**Models**	**AUROC (95% CI)**	pvalue	AIC	pAUROC **(95% CI)**	Pval
2	**AFP+IL6**	0.880 (0.810-0.951)	.Ref	**118.59**	0.171 (0.135-0.206	.Ref
2	GROα+ IL6	0.886 (0.828-0.944)	0.7233	125.18	0.162 (0.125-0.198)	0.46
3	AFP+ IL6+GROα	0.907 (0.855-0.960)	0.049	112.91	0.178 (0.146-0.211)	0.39
4	IL6+ MIP1b GROα+ LBP	0.894 (0.841-0.947)	0.4347	126.32	0.166 (0.131-0.201)	0.43
4	IL6+ MIP1b GROα+ IGFBP1	0.892 (0.835-0.950)	0.4397	128.16	0.167 (0.124-0.199)	0.65

Set 2 results

Next we checked the performance of the presented method when choosing a combination of proteins from a larger pool of potential classifiers. Seventy four unique CVF proteins were assessed in set 2 for their ability to classify inflammation. Of these 6 proteins were measured using two different methods to give a total of 81 measures. Gestational age in weeks at onset of preterm labor was also collected. Set 2 had fewer number of women at 108, including fewer women (7 women) with late onset preterm labor. The majority of women (94%) were of gestational age less than 34 weeks. Table 9a details the characteristics of 21 of the measures that were found to be important for classification, and mean gestational age of patients with and without inflammation. Three CVF proteins had measures at lower limit of detection greater than 10%. Myoglobin had 58.5% of controls and 19.23% of cases at lower limits of detection. The CVF protein Sortilinn had 13% of controls and 23% of cases and Human Neutrophile Peptides 1-3 had 14.63% of controls and 7.69% of cases at lower limits of detection. Table 9b shows the difference in distribution of the cases from the controls.

Table 9a: Geometric means of 21 CVF proteins and arithmetic mean of gestational age.

		Intra-amniotic Inflammation Absent (n=82)	**Intra-amniotic Inflammation present (n=26)**
Variables	Protein abbrv.	Geometric mean (std dev)	Geometric mean (std dev)
Gestational Age	GA	29.97 (3.57)*	26.05 (3.1)*
α-fetoprotein	AFP	7.4 (2.61)	50.08 (3.78)
Interleukin-1 beta	IL-1β	322.76 (4.89)	1287.03 (5.32)
interleukin 6	IL6 (poc)	463.76 (2.15)	2441.79 (4.08)
interleukin 6	IL6 (rbm)	109.79 (5.83)	1160.73 (6.64)
interleukin 8	IL8	8379.4 (3.34)	30579.03 (3.3)
Chemokine Ligand 4	MIP-1β	372.69 (3.61)	1381.61 (4.39)
Plasminogen Activator Inhibitor 1	PAI1	3414.75 (7.08)	35820.88 (4.59)
β2 microglobulin	B2MG	28.03 (3.84)	86.43 (4.57)
Carcinoembryonic Antigen	CEAG	321.83 (3.15)	118.97 (2.81)
Insulin-like growth factor-binding protein 1	IGFBP1	268.39 (6.02)	1465.53 (5.74)
Hepatocyte Growth Factor	HGF	2.7 (2.44)	5.93 (2.72)
Chemokine Ligand 2	MCP1	231.2 (3.07)	665.15 (3.36)
Human Neutrophile Peptides 1-3	HNP-1-3	1279638.15 (3.59)	3223728.61 (3.51)
Myoglobin	Myog	27.96 (4.04)	101.07 (4.25)
Prolactin	Prolactin	1297.64 (2.67)	3498.35 (4.1)
Pregnancy-Associated Plasma Protein	PAPP	41.28 (3.47)	17.79 (2.2)
Interleukin-1 alpha	IL1α	321.64 (2.96)	696.65 (2.52)

Sortilinn	Sortilinn	1.11 (3.21)	0.69 (2.81)
Cancer Antigen 125	CA125	1838.46 (3.56)	1233.43 (4.3)
Chemokine Ligand 1	GROα (rbmbr)	4.2 (2.9)	2.49 (3)
Pregnancy-Associated Plasma Protein A	PAPPA	76.34 (5.62)	96.51 (5.24)

*Arithmetic mean (std dev)

Table 9b: Difference in the distribution of cases and controls, as determined by Mann Whitney U, AUROC and partial AUROC, of selected 21 CVF proteins

	PROTEIN ABBRV.	Mann Whitney U (p-value)	AUROC (95% CI)	pAUROC (0.25) (95% CI)
Gestational Age	GA		0.677 (0.562-0.792)	0.071 (0.027-0.116)
α-fetoprotein	AFP	267.5 (<0.0001)	0.875 (0.793-0.956)	0.162 (0.116-0.207)
Interleukin-1 beta	IL-1β	510.5 (<0.0001)	0.760 (0.645-0.875)	0.104 (0.56-0.152)
Interleukin 6	IL6 (poc)	340 (<0.0001)	0.841 (0.745-0.936)	0.158 (0.115-0.202)
Interleukin 6	IL6	328 (<0.0001)	0.846 (0.744-0.948)	0.153 (0.108-0.199)
Interleukin 8	IL8	380 (<0.0001)	0.822 (0.713-0.931)	0.151 (0.107-0.195)
Chemokine Ligand 4	MIP-1β	490 (<0.0001)	0.769 (0.649-0.889)	0.126 (0.082-0.170)
Plasminogen Activator Inhibitor 1	PAI1	375 (<0.0001)	0.824 (0.729-0.919)	0.125 (0.078-0.173)
β2 microglobulin	B2MG	525.5 (0.0001)	0.753 (0.630-0.877)	0.120 (0.072-0.168)
Carcinoembryonic Antigen	CEAG	1596.5 (0.0001)	0.748 (0.641-0.854)	0.097 (0.049-0.144)
Insulin-like growth factor-binding protein 1	IGFBP1	536 (0.0001)	0.749 (0.641-0.856)	0.094 (0.048-0.140)
Hepatocyte Growth Factor	HGF	543 (0.0002)	0.745 (0.628-0.862)	0.100 (0.051-0.148)
Chemokine Ligand 2	MCP1	551 (0.0002)	0.741 (0.630-0.853)	0.092 (0.045-0.139)
Human Neutrophile Peptides 1-3	HNP-1-3	576 (0.0004)	0.724 (0.594-0.855)	0.106 (0.059-0.152)
Myoglobin	Myog	603 (0.0004)	0.660 (0.519-0.80)	0.064 (0.019-0.109)
Prolactin	Prolactin	581.5 (0.0005)	0.727 (0.605-0.848)	0.103 (0.057-0.149)
Pregnancy-Associated Plasma Protein	PAPP	1512 (0.0014)	0.708 (0.591-0.825)	0.080 (0.032-0.127)
Interleukin-1 alpha	IL1α	626.5 (0.0016)	0.706 (0.597-0.815)	0.076 (0.032-0.120)

Sortilinn	Sortilinn	1311 (0.0783)	0.596 (0.482-0.710)	0.050 (0.136-0.087)
Cancer Antigen 125	CA125	1206.5 (0.3144)	0.566 (0.433-0.698)	0.044 (0.07-0.081)
Chemokine Ligand 1	GROα	1338.5 (0.0506)	0.627 (0.502-0.752)	0.063 (0.026-0.101)
Pregnancy-Associated Plasma Protein A	PAPPA	975.5 (0.5177)	0.537 (0.411-0.663)	0.036 (0.005-0.067)

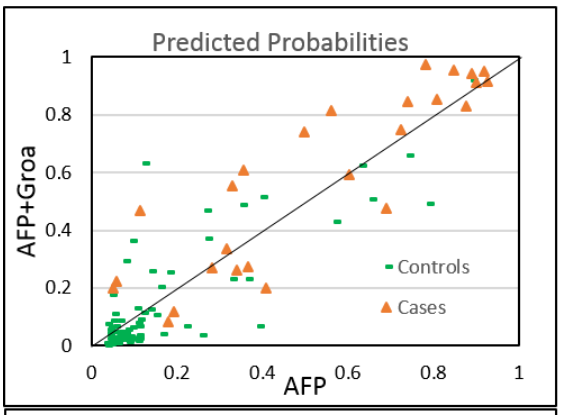
The ROC method was applied on set 2 data. The best combination of proteins maximizing partial AUROC in FPF region 0 to 0.25, was found to be AFTP+GROα+IL6+PAPPA. Table 10 details the proteins and the resultant AUROC by baker and logistic method as well as the partial AUROC and IDI as each protein is individually added into the model.

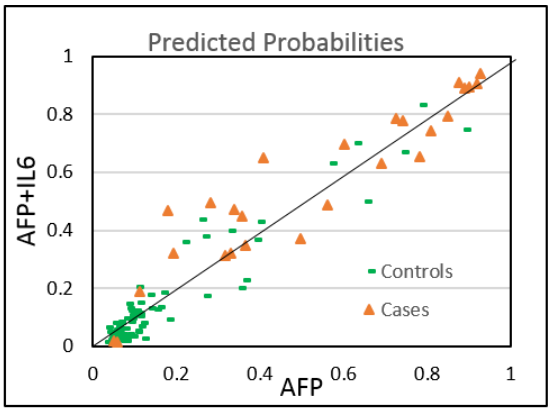
Table 10: The model built by the ROC-Logistic method

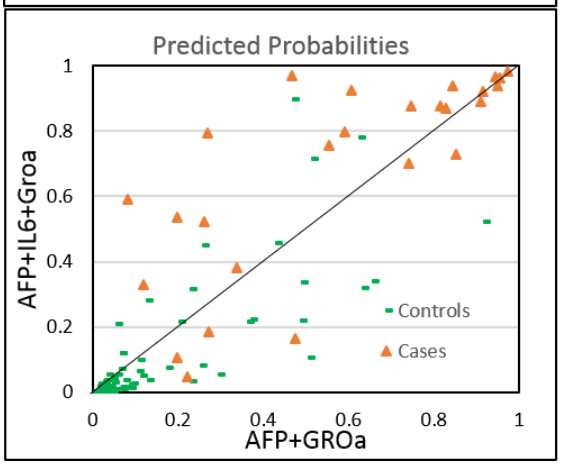
CVF Protein	AFP	GROα	IL6	PAPPA
Order of addition into model	1	2	3	4
Baker_pAUROC	0.161585	0.16909	0.19606	0.206614
Baker_AUROC	NA	0.91909	0.94606	0.954503
Logistic_AUROC	0.874531	0.911351	0.94606	0.954034
IDI	NA	0.069351	0.133552	0.081022

Partial AUROC of the model=0.2165 using logistic regression.

The above model was arrived at by using an IDI cut off of 0.05, we wanted to check if a parsimonious model could be equally efficient. By increasing the IDI cut off, fewer variables will meet criteria. The analysis was redone with a higher IDI cut off of 0.1. Interestingly, no additional markers were included in the model except the initial AFP. On further examination, we saw that, if GROα is not retained in the model, then none of the other variables reach the IDI cut off. Only when GROα is in the model does the IDI increase above 0.05 irrespective of improvement in partial AUROC. This may be because there is more perturbation on addition of GROα, increasing the IDI on its addition and also when other variables are added after it (Figure 13).







	IDI	AUROC	pAUROC
AFP		0.876	0.162
AFP+GROa	0.069	0.911	0.169
AFP+IL6	0.039	0.89	0.177
AFP+GROa+IL6	0.134*	0.946	0.196

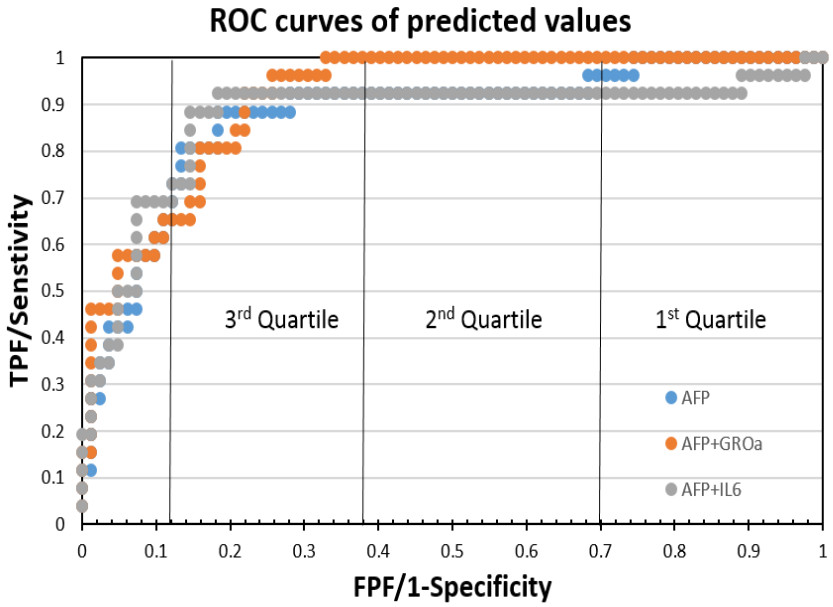
* Compared to AFP+IL6

Figure 13: Change in predicted probability of cases and controls on addition of new marker

The graphs shows the change in predicted probability on adding GROα to AFP, on adding IL6 to AFP and adding IL6 to AFP after addition of GROα. Addition of GROα to AFP causes greater change in the predicted probability of both cases and controls compared too addition of IL6 to AFP. This is reflected in the IDI.

The IDI of AFP+GROα is higher than AFP+IL6 in the first, second and third quartile, while it is slightly lower than AFP+IL6 in the fourth quartile. Hence even though the partial AUROC of AFP+IL6 is higher than AFP+GROα, the IDI is lower as AFP+GROα does better overall. Also when IL6 is added to AFP+GROα the IDI improves mainly due to positive movement in the third quartile (Figure 14).

The above example illustrates how using IDI may not be straightforward when only interested in partial AUROC. An alternative is to quartile predicted probabilities and use IDI specific for the quartile that needs to be maximized. In the current case IDI improvement in the third and fourth quartile would be equivalent to improvement in partial AUROC. Indeed the IDI for the two quartiles is 0.006 for AFP+GROα and 0.085 for AFP+IL6 reflecting the better partial AUROC of AFP+IL6. However using a smaller bin of IDI may make it more susceptible to outliers for example in case of AFP+GROα the IDI improvement in the first and second quartile is due to a single point (1 case). Additional studies are required to fine tune the use of IDI as variable retention criteria in the ROC-logistic procedure. Smoothing the ROC curve or binning the IDI may be a solution. For the current study, the overall IDI is used.



AFP and AFP+GROα	IDI	1st Quartile	2nd Quartile	3rd Quartile	4th Quartile
Cases	0.053	0.147 (1)	0.164 (1)	0.048 (7)	0.042 (17)
Controls	-0.017	-0.006 (26)	-0.014 (26)	0 (20)	-0.085 (10)
Sum	0.069	0.141	0.15	0.048	-0.042
AFP and AFP+IL6	**IDI**	**1st Quartile**	**2nd Quartile**	**3rd Quartile**	**4th Quartile**
Cases	0.03	-0.033 (1)	-0.044 (1)	0.117 (7)	0.002 (17)
Controls	-0.009	-0.01 (26)	-0.011 (26)	0.013 (20)	-0.046 (10)
Sum	0.039	-0.043	-0.055	0.13	-0.045
AFP+GROα and AFP+GROα+IL6	**IDI**	**1st Quartile**	**2nd Quartile**	**3rd Quartile**	**4th Quartile**
Cases	0.101	0 (0)	0 (0)	0.171 (9)	0.065 (17)
Controls	-0.032	-0.01 (27)	-0.011 (27)	-0.058 (18)	-0.104 (10)
Sum	0.134	-0.01	-0.011	0.113	-0.039

Figure 14: IDI for each quartile of the ROC curve for three different marker combinations.

The IDI for different sections of the AUROC curve is generated to compare movement of predicted probability corresponding to different parts of the ROC curve.

** AFP+GROα and AFP+GROα+IL6 not shown as ROC curve.*

Using the model AFP+IL6+GROα+PAPPA, the model was checked for improvement in classification ability on adding gestational age as a covariate. No improvement was seen. This is expected as the model had an AUROC of 0.954, with very less room for improvement, (Table 11)

Table 11: Comparison of models with and without Gestational age as covariate

Model	AUROC (95% CI)	pvalue	AIC	pAUROC (95% CI)	pvalue
IL6+AFP+GROα +PAPPA	0.954 (0.908-1.000)	1	55.484	0.216 (0.188-0.245)	
IL6+AFP+GROα +PAPPA+GA	(0.954 (0.909-1.000)	0.56	57.073	0.217 (0.188-0.246)	0.52

Next, the presented method was compared with traditional methods of model building. As the number of variables in set 2 was 81 it was not feasible to check every single model, nor is it practical to use best subset selection with such a large number of variables. Hence data mining methods were used to narrow down good classifiers and then the best subset selection method was utilized to identify optimum combinations of these selected classifiers.

Random forest is ideal for the set 2 data set as it has only 108 patient samples and 81 variables. Many of the CVF proteins are highly correlated with each other. The dataset also had values for the same protein that were measured using different methods (for example IL6 was measured using two different methods coded as POC and RBM) and hence would be expected to have high correlation. To account for this a random forest application which takes correlation into account was used. The R package 'party' is one such package that mines variables and builds classification trees taking the correlation of variables with each other into account. Figure 15 shows the average variable importance factor of 7 runs for the top 20 classifiers derived using the party package. Variable importance factor was based on change in AUROC before and after permutation as the current dataset was unbalanced at 26 cases and 82 controls.

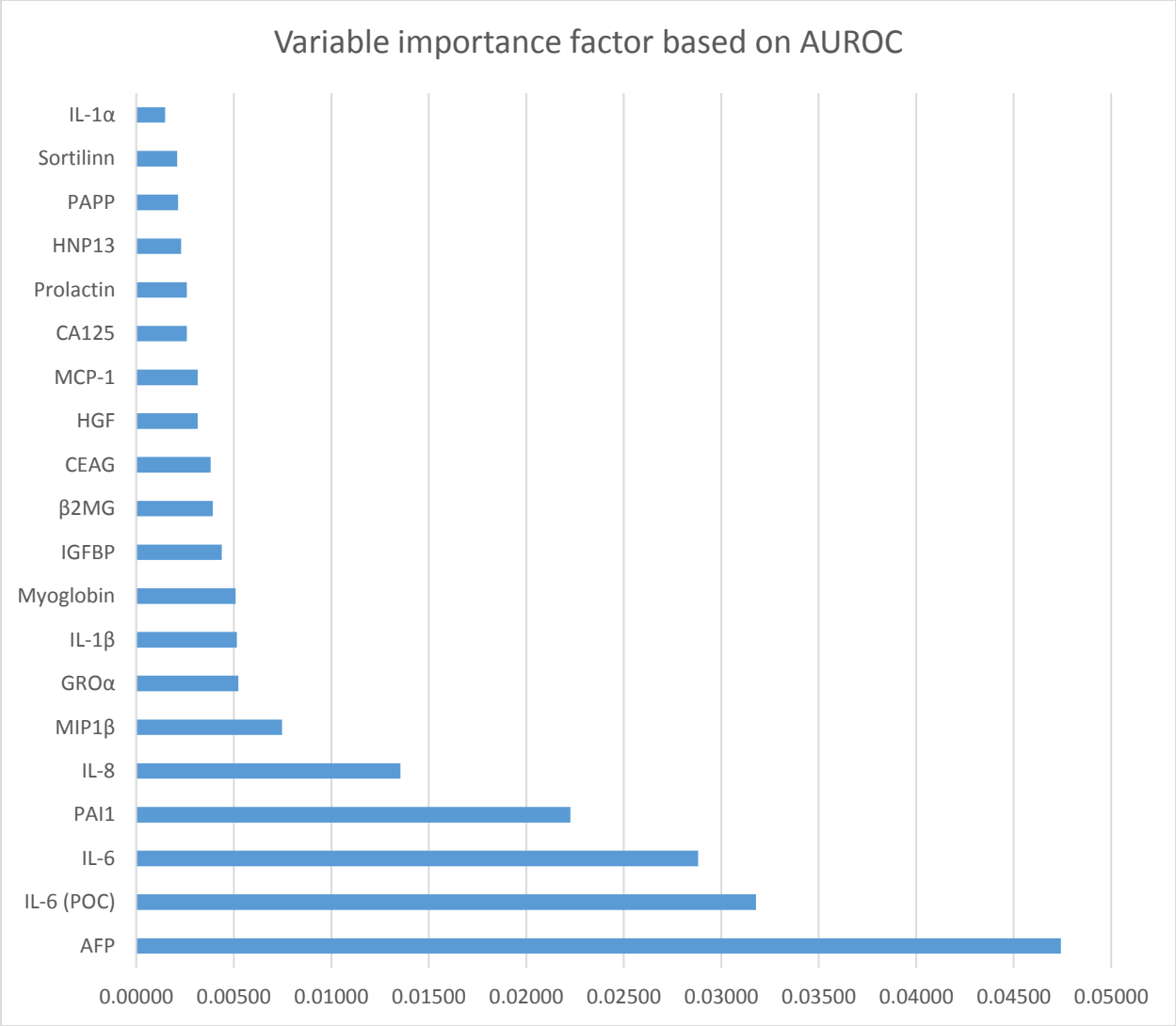


Figure 15: Variable Importance Factor of top 20 CVF proteins

The VIF was generated using OOB AUC instead of error rate in conditional Random forest

Best subset selection was carried out with these 20 variables to determine combinations of variables that would optimize classification of inflammation. The majority of models with greater than 14 variables did not converge, presented in Figure 16 are models with a maximum of 5 variables. Note many of the models are at AUROC 0.98 at 5 parameter level and hence models restricted to 5 parameters. Models with greater number of variables may be over fitting the data, especially as we have only 108 samples, and hence may not be reproducible.

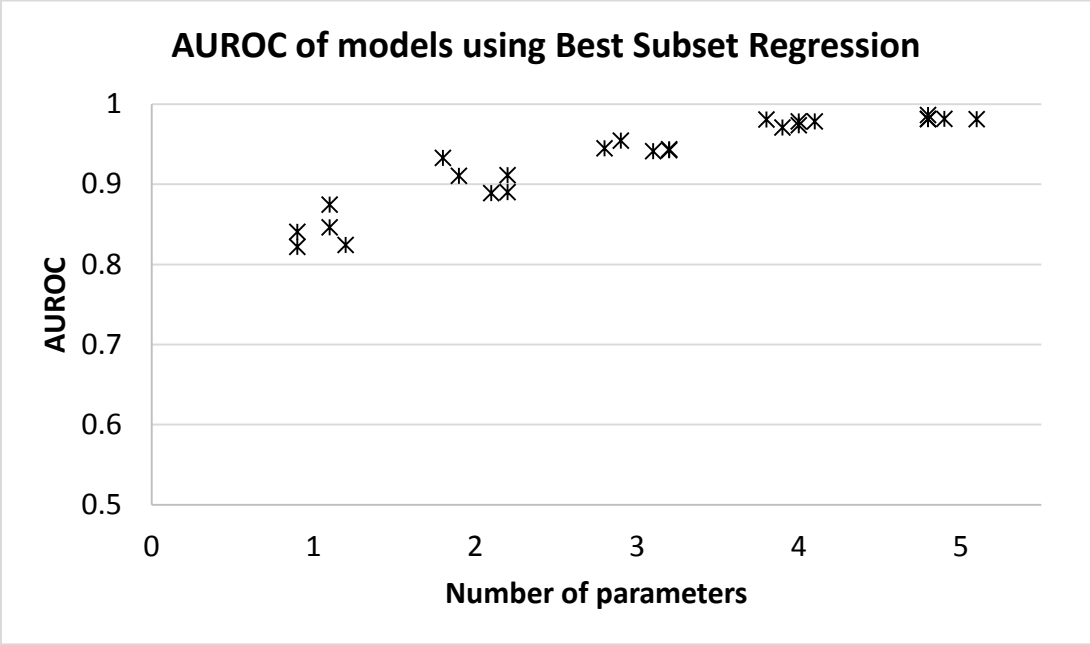


Figure 16: AUROC of models with up to 5 parameters

The AUROC of models with more than 4 parameters were very close to 1 (the maximum)

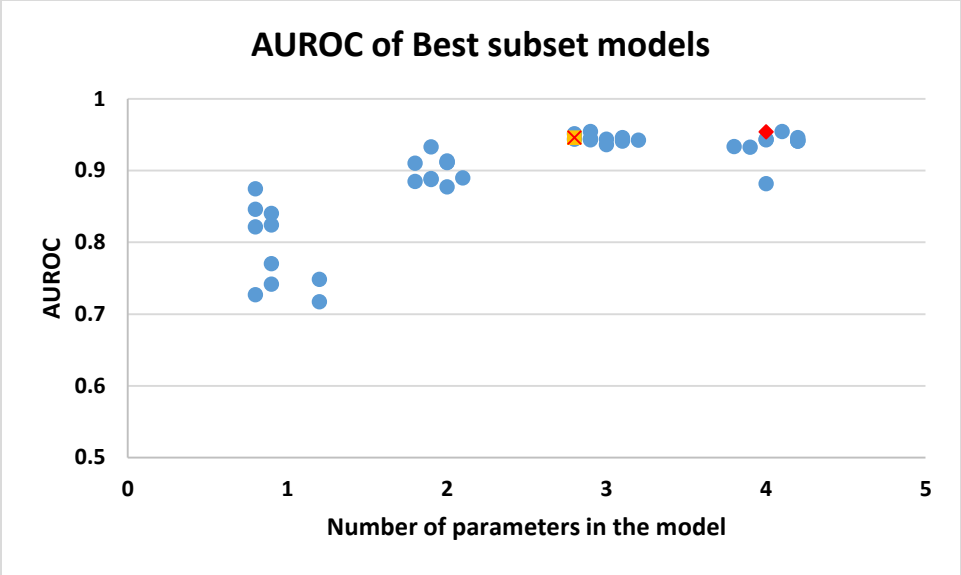
The traditional methods of Data mining and best subset selection did not arrive at the model built using the ROC-Logistic method. This is because the best subset selection was restricted to variables chosen by random forest and Pappa-1 was not one of the top 20 variables as ranked by VIF. Pappa1 had a very low VIF of 0.00007. Table 12 compares the AUROC and pAUROC of all the models built using best subset to the model chosen by ROC-logistic method. None of the models were significantly better than the model chosen by ROC-logistic method.

Table 12: Comparing Best subset regression models to that chosen by ROC-logistic method.

No. of Var	Variables in Model	Score Chi-Square	AUROC (95% CI)	pvalue	AIC	pAUROC (0.25) (95% CI)	pvalue
4	IL6+AFP+GROα+PAPPA	.	0.954 (0.908-1)	1	55.48	0.216 (0.187-0.246)	Ref
1	AFP	40.90	0.875 (0.793-0.956)	0.012	83.20	0.162 (0.116-0.207)	0.01
1	IL6 (POC)	38.81	0.841 (0.745-0.936)	0.0056	87.02	0.158 (0.115-0.202)	0.0035
1	IL6	26.29	0.846 (0.744-0.948)	0.0056	92.04	0.153 (0.108-0.199)	0.0032
1	PAI1	24.68	0.824 (0.729-0.919)	0.0032	94.85	0.125 (0.078-0.173)	0.00024
1	IL8	19.15	0.822 (0.713-0.931)	0.0061	98.71	0.151 (0.107-0.195)	0.0071
2	IL6+GROα	49.20	0.933 (0.881-0.985)	0.1212	65.46	0.195 (0.157-0.233)	0.033
2	IL6+CA125	48.81	0.91 (0.846-0.975)	0.098	68.79	0.182 (0.141-0.223)	0.097
2	IL6(POC)+GROα	48.49	0.889 (0.805-0.973)	0.1284	73.95	0.184 (0.143-0.224)	0.077
2	AFP+IL6	47.59	0.89 (0.809-0.971)	0.0317	77.51	0.173 (0.128-0.218)	0.028
2	AFP+GROα	47.07	0.911 (0.858-0.964)	0.1232	74.72	0.165 (0.118-0.211)	0.011
3	IL6+GROα+CA125	59.04	0.955 (0.916-0.993)	0.9796	54.38	0.207 (0.174-0.239)	0.4
3	IL6(POC)+IL8+GROα	57.10	0.941 (0.895-0.988)	0.6262	62.45	0.198 (0.163-0.234)	0.23
3	AFP+IL6(POC)+GROα	56.12	0.942 (0.9-0.985)	0.6209	64.36	0.196 (0.160-0.231)	0.14
3	AFP+IL6(POC)+CA125	55.78	0.944 (0.901-0.987)	0.6957	63.94	0.199 (0.163-0.235)	0.34
3	AFP+IL6+CA125	55.69	0.945 (0.903-0.986)	0.6358	63.13	0.199 (0.163-0.235)	0.32
4	IL6+GROα+CA125+IL1a	65.79	0.981 (0.961-1)	0.2069	39.74	0.231 (0.211-0.250)	0.26

4	IL6+IL8+GROα+CA125	65.29	0.978 (0.958-0.999)	0.224	45.62	0.228 (0.209-0.248)	0.31
4	IL6+GROα+HGF+CA125	63.93	0.974 (0.949-0.999)	0.3126	44.27	0.224 (0.197-0.250)	0.55
4	IL6+GROα+IL1b+CA125	63.58	0.978 (0.958-0.999)	0.2102	44.69	0.228 (0.207-0.250)	0.3
4	IL6+GROα+IL1a+Sortilinn	62.82	0.971 (0.943-0.999)	0.2547	47.58	0.221 (0.194-0.248)	0.59

The ROC-Logistic was also performed on the random forest chosen subset of variables to compare with best subset selection models. The model chosen was AFP+GROα+IL6. This model was not amongst the top 5 based on score statistic but was amongst the top 10. It had comparable performance in terms of AUROC and pAUROC to the other models selected based on score statistic. Figure 17 shows the top 10 models at 4 parameter levels. Additionally PAPPA was added to the random forest selected variables, to check if it would be picked up in best subset modelling. AFP+GROα+IL6+PAPPA was amongst the top 10 but not amongst the top 5 models selected based on score statistic.



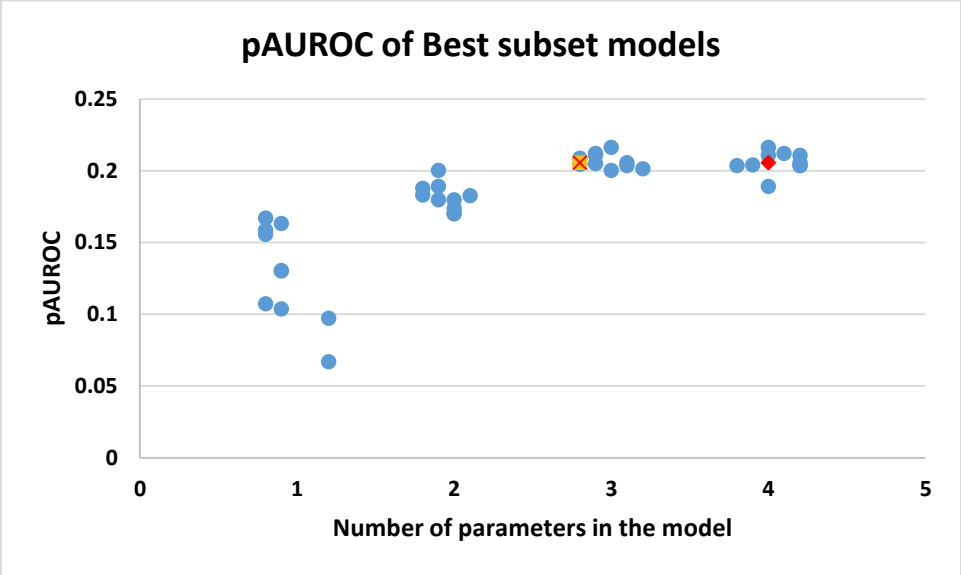


Figure 17: AUROC and pAUROC of 10 models at upto 4 parameter level chosen by the best subset selection or ROC-logistic method.

The "x" marks the model AFP+IL6+GROα and the red diamond marks the model, AFP+GROα+IL6+PAPPA chosen by the ROC-Logistic regression variable selection method on the random forest selected 20 variables and the entire set of 81 measures. The models are comparable to similar parsimony models chosen by best subset selection method from the 20 variables preselected by randomforest.

DISCUSSION:

The presented ROC-Logistic method is variable selection strategy that makes use of three different procedures, the Baker algorithm, logistic regression and IDI. Each of these methods were developed either to classify, to study the association or to evaluate classification improvement. In the current methodology, by combining certain desirable properties of each of the methods, the ROC-logistic method can be used for variable selection and classification model building. However there are some finer points of the methods used that should be kept in mind when evaluating models built using the presented method.

Baker suggests three different algorithms for model building, unordered, jagged and rectangular. The unordered, as the name suggest, does not follow any pattern, which means intermediate or lower protein levels may be considered before higher protein levels. The rectangular ranks TPF/FPF in an ordered fashion such that it alternates ranking between the two markers. This means that both markers should increase or decrease for the combination to be a good classifier. This works when potential markers have high sensitivity but individually have low specificity. The Jagged order, on the other hand, allows for heterogeneous movement, the classification is improved if either marker shows movement with disease condition (17). The presented method uses the jagged ordering, as the etiology of intra-amniotic inflammation may be heterogeneous, that is, it's possible that a marker maybe more responsive to disease than another. Jagged ordering allows for this unequal contribution of markers to classification. As the number of markers to be

added to the model increases the Baker method becomes multidimensional and impractical. The current method avoids this issue by using logistic regression to combine markers. Since both Baker and Logistic regression us the underlying principle of optimality of likelihood ratio, it is possible to seamlessly combine both methods. Instead of increasing the dimensionality of multi marker combinations, the markers were added sequentially. The markers already in the model are combined using logistic regression and the new marker is added to the predicted probability of this combination using Baker algorithm. Thus the jagged algorithm always works with a two dimensional model, which is much less complicated. The current method resolves the multidimensionality issue, however, it also makes the order of addition of variables critical. Depending on which marker is already in the model a new marker may or may not improve the model despite being a good classifier. To overcome this issue, the initial step of the method reorders variables based on slope from origin to *ROC(desired FPF)*. It is possible, however that good classifiers that come later in the ordering may not be added.

Similarly depending on the partial AUROC to be maximized, the final model may be different as the first model to step in determines the rest of the model. Hence it is important to fully understand the biological and clinical relevance of the FPF range over which the AUROC is to be maximized.

Partial AUROC is a clinically relevant metric in designing classification models, however the IDI used as an exclusion criteria is not restricted to the partial FPF range. IDI, is a relatively new metric for evaluating improvement in classification. It looks at overall improvement in predicted probability of cases and controls on addition of a new marker.

The advantage of the IDI is that it looks at not only the direction but also the magnitude of change. Greater movement get higher values and slight movement in predicted probabilities have lower values. This means that predicted probabilities that have more room to move are likely to contribute more to IDI, as extreme points may not move much (if not misclassified) . This also makes IDI susceptible to outliers. If a single sample has a large movement in predicted probability, on adding a new marker to the model, it may overshadow any other small movement and significantly increase or decrease the overall IDI.

It has been shown that the test statistic for IDI does not have a standard normal distribution under the null assumption that IDI=0 and hence the test based on Z_{IDI} is not valid for determining significant change (21) hence in the current method the choice was made to use a cut off of actual IDI values as against significant change in IDI. The cut off for IDI should be optimized for each sample set as it is susceptible to case-control ratio, basal AUROC, change in AUROC and outliers. The IDI does not have a one to one correlation with AUROC or partial AUROC hence optimizing the IDI to maximize partial AUROC, while also building a parsimonious model may require some trial and error. Cross-validation studies may be required to optimize model especially if the model is not parsimonious. In the current case study, the ROC-Logistic method chose four proteins including PAPPA. This may be too many for this limited data set. On comparing the ROC curves for the model with PAPPA and without (Figure 18), it is observed that the difference between the two models may not be significant and the curves may be very

similar on smoothing. Hence in the interest of parsimony the model combination of AFP+GROα+IL6 may be better.

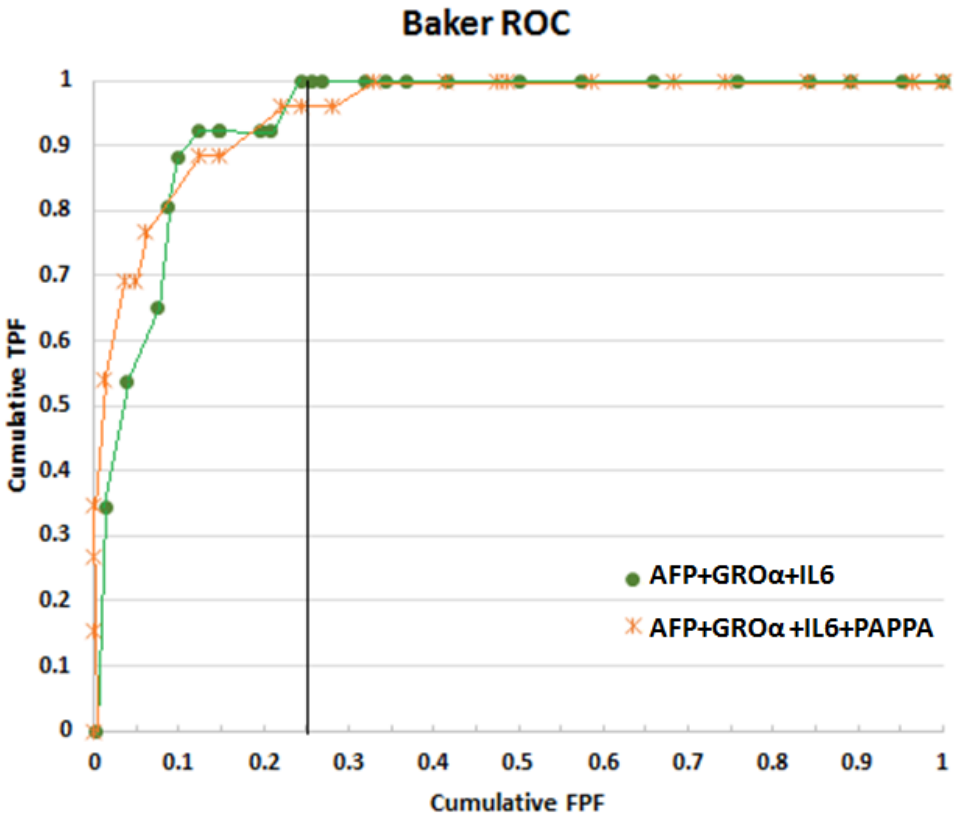


Figure 18: Comparison of ROC curve of models with and without PAPPA

The two models chosen by the ROC-logistic method, using all 81 CVF protein measures and using randomforest preselected 20 CVF measures, have very similar ROC curves.

One of the advantages of the proposed method is its adaptability. The method allows one to isolate models by setting clinically relevant FPF specific to the study. It also allows to retain a known biologically relevant marker and choose the rest of markers conditional on this marker. This makes the method clinically appealing as it allows combining existing knowledge about markers, while finding new ones. In order to further evaluate the performance of the method, validation studies would need to be conducted. This involves testing the performance of the chosen model on a subset of data that was not used for building the model Due to the limited sample size it was not feasible to conduct validation studies in the current dataset. Set 1 and set 2 datasets had values of same proteins measured using different methods and hence they could not be used for validation of method and models were built separately for both datasets. It is, however, worth noting that AFP and IL6 were chosen in both sets.

CONCLUSION:-

For a dataset of moderate size and number of classifiers, the ROC-Logistic method is a quick and efficient method in choosing combination of variables for accurate classification. The method allows one to hone in on combination of classifiers that maximize partial AUROC, the metric of interest, when developing a clinically relevant test. To our knowledge, no current method exist which specifically target improvement in partial AUROC. The method also chooses the most parsimonious model, which is likely to give more reproducible classification results. It allows one to bypass multiple steps of data mining and then checking for efficient models amongst the chosen classifiers. In conclusion, the ROC-Logistic method efficiently isolates optimum combination of classifiers based on clinically relevant metrics.

REFERENCES:

1. *Rifai N., Gillette MA., Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nat Biotechnol. 2006;24:971–983.*
2. *Tan PS, M.; Kumar, V. Introduction to Data Mining. 1 ed: Addison-Wesley [ISBN 978-0321321367].*
3. *Combs CA, Gravett M, Garite TJ, Hickok DE, Lapidus J, Porreco R, et al. Amniotic fluid infection, inflammation, and colonization in preterm labor with intact membranes. Am J Obstet Gynecol 2014;210(2):125 e1-125 e15.*
4. *Combs CA, Garite TJ, Lapidus JA, Lapointe JP, Gravett M, Rael J, et al. Detection of microbial invasion of the amniotic cavity by analysis of cervicovaginal proteins in women with preterm labor and intact membranes. Am J Obstet Gynecol. 2015 Apr;212(4):482.e1-482.e12*
5. *Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford Statistical Science Series 31 [ISBN 978-0198565826]*
6. *Hosmer, DW., Lemeshow, S., Sturdivant, R. Applied Logistic Regression. Wiley Series in Probability and Statistics. [ISBN 0470582472]*
7. *McIntosh MW, PEPE MS. Combining several screening tests: Optimality of the risk score. Biometrics 2002;58(3):8.*
8. *Feng ZY, Y. Statistical considerations in combining biomarkers for disease classification. Disease Markers 2004;20:7.*
9. *Breiman L. Random forests. Machine Learning 2001;45(1):37.*

10. *Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 2007;8:25.*
11. *Liaw AW, M. Classification and Regression by randomforest. R News 2002;2(3):5.*
12. *Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics 2008;9:307.*
13. *Hornik KZ, A.; Hothorn, T. Unbiased Recursive Partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics 2006;15(3):23.*
14. *Janitza S, Strobl C, Boulesteix AL. An AUC-based permutation variable importance measure for random forests. BMC Bioinformatics 2013;14:119.*
15. *Hothorn TH, K.; Zeileis, A. party: A laboratory for recursive part(y)itioning.*
16. *Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29-36.*
17. *Feng Z. Classification versus association models: should the same methods apply? Scand J Clin Lab Invest Suppl 2010;242:53-8.*
18. *Baker SG. Identifying Combinations of Cancer Markers for further study as triggers of early intervention. Biometrics 2000;56(4).*
19. *Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27(2):157-72; discussion 207-12.*
20. *Pencina MJ, D'Agostino RB, Sr., Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. Stat Med 2012;31(2):101-13.*

21. *Kerr KF, McClelland RL, Brown ER, Lumley T. Evaluating the incremental value of new biomarkers with integrated discrimination improvement. Am J Epidemiol 2011;174(3):364-74.*
22. *Liu L, Johnson HL, Cousens S, Perin J, Scott S, Lawn JE, et al. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. Lancet 2012;379(9832):2151-61.*
23. *Romero R, Miranda J, Chaiworapongsa T, Korzeniewski SJ, Chaemsaithong P, Gotsch F, et al. Prevalence and clinical significance of sterile intra-amniotic inflammation in patients with preterm labor and intact membranes. Am J Reprod Immunol 2014;72(5):458-74.*
24. *Romero R, Espinoza J, Kusanovic JP, Gotsch F, Hassan S, Erez O, et al. The preterm parturition syndrome. BJOG 2006;113 Suppl 3:17-42.*
25. *Romero R, Gomez R, Chaiworapongsa T, Conoscenti G, Kim JC, Kim YM. The role of infection in preterm labour and delivery. Paediatr Perinat Epidemiol 2001;15 Suppl 2:41-56.*
26. *Yoon BH, Romero R, Moon JB, Shim SS, Kim M, Kim G, et al. Clinical significance of intra-amniotic inflammation in patients with preterm labor and intact membranes. Am J Obstet Gynecol 2001;185(5):1130-6.*
27. *Been JV, Degraeuwe PL, Kramer BW, Zimmermann LJ. Antenatal steroids and neonatal outcome after chorioamnionitis: a meta-analysis. BJOG 2011;118(2):113-22.*

28. *Kenyon S, Pike K, Jones DR, Brocklehurst P, Marlow N, Salt A, et al. Childhood outcomes after prescription of antibiotics to pregnant women with spontaneous preterm labour: 7-year follow-up of the ORACLE II trial. Lancet 2008;372(9646):1319-27.*
29. *Tsatsaris V, Carbonne B, Cabrol D. Place of amniocentesis in the assessment of preterm labour. Eur J Obstet Gynecol Reprod Biol 2000;93(1):19-25.*
30. *Palacio M, Cobo T, Bosch J, Filella X, Navarro-Sastre A, Ribes A, et al. Cervical length and gestational age at admission as predictors of intra-amniotic inflammation in preterm labor with intact membranes. Ultrasound Obstet Gynecol 2009;34(4):441-7.*
31. *Hassan S, Romero R, Hendler I, Gomez R, Khalek N, Espinoza J, et al. A sonographic short cervix as the only clinical manifestation of intra-amniotic infection. J Perinat Med 2006;34(1):13-9.*
32. *Jung HJ, Park KH, Kim SN, Hong JS, Oh KJ, Kim G, et al. Non-invasive prediction of intra-amniotic inflammation in women with preterm labor. Ultrasound Obstet Gynecol 2011;37(1):82-7.*
33. *Park JW, Park KH, Lee SY. Noninvasive prediction of intra-amniotic infection and/or inflammation in women with preterm labor: various cytokines in cervicovaginal fluid. Reprod Sci 2013;20(3):262-8.*
34. *Yoon BH, Romero R, Moon JB, Oh SY, Han SY, Kim JC, et al. The frequency and clinical significance of intra-amniotic inflammation in patients with a positive cervical fetal fibronectin. Am J Obstet Gynecol 2001;185(5):1137-42.*
35. *Pepe MS, Longton G, Janes H. Estimation and Comparison of Receiver Operating Characteristic Curves. Stata J. 2009 March 1; 9(1): 1*

36. *SAS Institute Inc. 2008. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc*

37. *Hothorn TB, P.; Dudoit, S.; Molinaro, A.; Van Der Laan, M. Survival Ensembles. Biostatistics 2006;7(3):18.*

APPENDICES:

Supplementary Tables:

Supp. Table 1: Best Subset Models for set1:

Models	Variables	Chi-Square	Variables Included in Model	AUROC(95% CI)	P value when compared with AFP+IL6 model	AIC
1	1	63.27	AFP	0.793 (0.709-0.878)	0.0141	133.70
2	1	53.94	IL6	0.861 (0.79-0.932)	0.0515	131.40
3	1	25.31	β2MG	0.748 (0.652-0.844)	<.0001	159.65
4	1	23.42	MIP-1β	0.667 (0.572-0.762)	<.0001	167.36
5	1	22.87	IGFBP	0.669 (0.566-0.773)	<.0001	167.04
6	2	77.75	AFP IL6;	0.88 (0.81-0.951)	Ref	118.59
7	2	66.99	AFP IGFBP	0.81 (0.717-0.904)	0.0509	133.28
8	2	66.35	AFP MCP_log	0.83 (0.747-0.914)	0.0872	133.13
9	2	65.36	AFP β2MG	0.803 (0.705-0.901)	0.0079	133.76
10	2	65.20	AFP MIP-1β	0.793 (-0.698-0.887)	0.0178	134.30
11	3	82.05	AFP GROα IL6	0.907 (0.855-0.96)	0.049	112.91
12	3	80.42	α1AG AFP IL6	0.89 (0.828-0.951)	0.3697	115.67
13	3	80.41	AFP IL6 LBP	0.889 (0.829-0.95)	0.4586	117.15
14	3	79.03	AFP β2MG IL6	0.893 (0.833-0.953)	0.2216	118.02
15	3	77.81	AFP IL6 MCP_log	0.878 (0.807-0.949)	0.5646	120.02
16	4	84.15	α1AG AFP GROα IL6	0.915 (0.867-0.963)	0.0493	11.07
17	4	83.63	AFP GROα IL6 LBP	0.914 (0.868-0.961)	0.0472	112.93
18	4	82.29	AFP β2MG GROα IL6	0.912 (0.863-0.961)	0.0491	114.15
19	4	82.14	AFP GROα IL6 MIP-1β	0.905 (0.852-0.959)	0.0713	114.63

20	4	82.10	AFP GROα IL6 MCP_log	0.906 (-0.853-0.959)	0.062	114.59
21	5	84.29	α1AG AFP GROα IL6 MIP-1β	0.916 (-0.869-0.963)	0.0447	113.04
22	5	84.24	α1AG AFP GROα IL6 LBP	0.914 (0.865-0.962)	0.0535	113.06
23	5	84.18	α1AG AFP β2MG GROα IL6	0.916 (0.87-0.963)	0.0508	113.05
24	5	84.17	α1AG AFP GROα IL6 MCP_log	0.914 (0.865-0.962)	0.0568	112.92
25	5	84.17	α1AG AFP GROα IGFBP IL6	0.915 (0.868-0.963)	0.0514	112.91
26	6	84.41	α1AG AFP GROα IL6 LBP MIP-1β	0.914 (0.866-0.962)	0.0553	115.02
27	6	84.38	α1AG AFP GROα IL6 MCP_log MIP-1β	0.915 (0.866-0.963)	0.0489	114.91
28	6	84.30	α1AG AFP β2MG GROα IL6 MIP-1β	0.917 (0.87-0.963)	0.0489	115.02
29	6	84.30	α1AG AFP GROα IGFBP IL6 MIP-1β	0.916 (0.868-0.963)	0.0485	114.87
30	6	84.27	α1AG AFP GROα IGFBP IL6 LBP	0.914 (-0.865-0.963)	0.0563	114.86
31	7	84.46	α1AG AFP GROα IL6 LBP MCP_log MIP-1β	0.913 (0.863-0.962)	0.0581	116.84
32	7	84.43	α1AG AFP GROα IGFBP IL6	0.915 (-0.867-0.962)	0.0533	116.85

			MCP_log MIP-1β			
33	7	84.42	α1AG AFP GROα IGFBP IL6 LBP MIP-1β	0.913 (0.864-0.962)	0.0616	116.81
34	7	84.41	α1AG AFP β2MG GROα IL6 LBP MIP-1β	0.916 (0.869-0.963)	0.0495	117.00
35	7	84.39	α1AG AFP β2MG GROα IL6 MCP_log MIP-1β	0.915 (0.868-0.962)	0.0575	116.87
36	8	84.51	α1AG AFP GROα IGFBP IL6 LBP MCP_log MIP-1β	0.912 (0.863-0.962)	0.0657	118.76
37	8	84.46	α1AG AFP β2MG GROα IL6 LBP MCP_log MIP-1β	0.913 (0.865-0.962)	0.0621	118.81
38	8	84.43	α1AG AFP β2MG GROα IGFBP IL6 MCP_log MIP-1β	0.915 (0.868-0.962)	0.0566	118.82
39	8	84.42	α1AG AFP β2MG GROα IGFBP IL6 LBP MIP-1β	0.914 (0.866-0.962)	0.06	118.80
40	8	84.30	α1AG AFP β2MG GROα IGFBP IL6 LBP MCP_log	0.913 (-0.865-0.962)	0.0626	118.76

41	9	84.51	α1AG AFP β2MG GROα IGFBP IL6 LBP MCP_log MIP-1β	0.914 (0.866-0.962)	0.0604	120.75

CODES:

```
#################
STEP 1 FIRST BRING IN DATA
#DETERMINE WHICH COLUMN IS THE OUTCOME/RESPONSE/STATUS COLUMN, YOU SHOULD HAVE THE STATUS COLUMN BEFORE ALL THE TEST VARIABLE.
#SUBJECT ID AND OTHER NON-TEST VARIABLES MUST GO BEFORE THE STATUS COLUMN CAN BE THE FIRST COLUMN.
#IN THE CURRENT DATA RESPONSE OR OUTCOME IS COLUMN 2 AND THE REST OF THE TEST VARIABLES START FROM COLUMN 3 ONWARDS
#DETERMINE THE TOTALNUMBER OF COLUMNS, YOU WILL USE THIS FOR VARS IN MannU AND MannU_val; VAR IS THE FIRST COLUMN OF TEST VARIABLES=3
STEP 2 DETERMINE AND ORDER COLUMNS BASED ON PAUC AT SPECIFIED CUT POINT FOLLOWED BY MANN WHITNEY U PVALUE (NOT ACTUAL STATISTIC, AS SOME COULD Be NEGATIVELY CORRELATED)
STEP 3 REORDER BASED ON STEEPEST SLOPE AT CUT POINT
STEP 4: SET IDI CUT OFF AND GET YOUR RESULTS. THE FIRST CUT OFF(0.2 IN EG BELOW) IS FPF FOR PARTIAL AUC AND THE SECOND(0.05) IS FOR IDI, START WITH LOWER IDI AND CAN INCREASE IF NEEDED
##ACTUAL CODES##
#to calculate ROC #score is column of variable you want to check and status is status column.
ROC = function(score,status)
{
        n = length(score)
        d = sum(status==1)
        status = status[order(score)]
        score = sort(score)
        FPR = TPR = rep(0,n-sum(duplicated(score))+1)
        AUC = 0
        k = 2
        x = y = 0
        for(i in 1:n)
        {
                x = x + 1 - status[n-i+1]
                y = y + status[n-i+1]
                if(!duplicated(score)[n-i+1])
                {
                        FPR[k] = x/(n-d)
                        TPR[k] = y/d
                        AUC = AUC + (TPR[k]+TPR[k-1])*(FPR[k]-FPR[k-1])/2
                        k = k + 1
                }
        }
        return(list(FPR=FPR,TPR=TPR,AUC=AUC))
}

# to calculate pAUC at certain FPF cutoff,c (c can be 0.1, 0.2,0.25.....and so on)
```

```
pAUC = function(score,status,c) #score is column of variable you want to check and status
is status column. #so for AFP:-ROC(iaf[,6],iaf[,2])
{
        n = length(score)
        d = sum(status==1)
if(cor(score,status)<0){
                                score=-score
                                }
        status = status[order(score)]
        score = sort(score)
        FPR = TPR = rep(0,n-sum(duplicated(score))+1)
        pAUC = 0
        k = 2
        x = y = 0
        for(i in 1:n)
        {
                x = x + 1 - status[n-i+1]
                y = y + status[n-i+1]
                if(!duplicated(score)[n-i+1])
                {
                        FPR[k] = x/(n-d)
                        TPR[k] = y/d
                        pAUC = pAUC + (TPR[k]+TPR[k-1])*(FPR[k]-FPR[k-1])/2
                        if(FPR[k]>c){break}
                        k = k + 1
                }
        }
                closestTPR=TPR[k]
                closestFPR=FPR[k]
                TPF =(((TPR[k]-TPR[k-1])/(FPR[k]-FPR[k-1]))*(c-FPR[k]))+TPR[k]
                pAUCc=pAUC-(TPF*(FPR[k]-c))-(0.5*(TPR[k]-TPF)*(FPR[k]-c))

        return(list(closestFPR=closestFPR,closestTPR=closestTPR,pAUC=pAUC,TPF=TPF,pAUCc=pA
UCc))
}
#pAUC(p1[,9],p1[,2],0.35)
#pAUC(p1ga[,3],p1ga[,2],0.25)

### to generate a table of pAUC, AUC and mann whitney (NOTE: pAUCc IS AT SPECIFIED
CUTPOINT, THE OTHER pAUC IS JUST A BUILT IN CHECK)
MannU_val=function(data,var,vars,status,c){
as.factor(data[,status])
subdata<-data[,var:vars]
mann<- data.frame(statistic= numeric(0), p.value = numeric(0),
AUC=numeric(0),pAUC=numeric(0),pAUCc=numeric(0))
var_name<-NULL
for(i in var:vars){
```

```
A<-wilcox.test(data[,var]~data[,status])
AUC<-ROC(data[,var],data[,status])
pAUC<-pAUC(data[,var],data[,status],c)
mann[nrow(mann)+1,] <- c(A$statistic,A$p.value,AUC$AUC,pAUC=pAUC$pAUC,pAUC$pAUCc)
var_name<-c(var_name,colnames(data[var]))
var=var+1
}
mann<-cbind(var_name,mann)
mann<-mann[order(mann[,3],mann[,2]),]
return(mann=mann)
}

#MannU_val(p1,3,83,2,c)
### to order based on partial pAUC followed by pvalue. i.e picking the best

MannU=function(data,var,vars,status,c){
as.factor(data[,status])
subdata<-data[,var:vars]
mann<- data.frame(statistic= numeric(0), p.value = numeric(0),
AUC=numeric(0),pAUCc=numeric(0))
var_name<-NULL
for(i in var:vars){
A<-wilcox.test(data[,var]~data[,status])
AUC<-ROC(data[,var],data[,status])
AUROC<-pAUC(data[,var],data[,status],c)
mann[nrow(mann)+1,] <- c(A$statistic,A$p.value,AUC$AUC,AUROC$pAUCc)
var_name<-c(var_name,colnames(data[var]))
var=var+1
}
mann<-cbind(var_name,mann)
neworder<-subdata[,order(-mann[,5],mann[,3])]
Mordered<-cbind(data[,1:2],neworder)
return(Mordered=Mordered)
}

## to rearrange data based on baker generated slopes, i,e picking the next best options
bak_slope = function(data,var1,p,s,c)
{
subdata<-data[,(var1+1):(var1+p)]
status=data[,s]
n = length(status)
d = sum(status==1)
bakmat<-data.frame(0)
for(t in 1:p){
        var2=var1+t
        rank1 = rank(data[,var1])
        rank2 = rank(data[,var2])
        quad1 = floor(2*rank1/(n+1))+1
```

```
        quad2 = floor(2*rank2/(n+1))+1
        FPR_table =
table(factor(quad1[status==0],levels=2:1),factor(quad2[status==0],levels=2:1))/(n-d)
        TPR_table =
table(factor(quad1[status==1],levels=2:1),factor(quad2[status==1],levels=2:1))/d
        ratio_table = TPR_table/FPR_table
        max_quad = which.max(ratio_table)# gives the position of max value
        if(max_quad > 1){
        rank1=rank((2*(max_quad%%2)-1)*data[,var1])
        rank2=rank((1-2*((max_quad-1)%/%2))*data[,var2])
        }
        score1 = floor(5*rank1/(n+1))+1
        score2 = floor(5*rank2/(n+1))+1
        FPR_table =
table(factor(score1[status==0],levels=1:5),factor(score2[status==0],levels=1:5))/(n-d)
        TPR_table =
table(factor(score1[status==1],levels=1:5),factor(score2[status==1],levels=1:5))/d
        ratio_table = TPR_table/FPR_table
        ratio_table[is.nan(ratio_table)] = 0
        ratio_table[ratio_table==0] = -FPR_table[ratio_table==0]
        k<-ratio_table
        k[25]=NA
        FPR=FPR_table[25]
        TPR=TPR_table[25]
        cTPR=c(0,TPR)
        cFPR=c(0,FPR)
        pAUC=0
        for (j in 2:n){
                        M<-which((is.na(k[,]))==TRUE)
                        A=c(M-1)
                        B=c(M-5)
                        b<-Reduce(intersect, list(A,B))
                        excl<-which((is.na(k[b]))==TRUE)
                        b<-if (length(excl)>0){b=b[-excl]
                                                }else{
                                                b=b
                                                }
                        LR<-c(k[5,])
                        LRM<-which((is.na(LR))==TRUE)
                        LRP=(5*(min(LRM)-1))
                        LC<-c(k[,5])
                        LCM<-which((is.na(LC))==TRUE)
                        LCP=(20+(min(LCM)-1))
                        lrow<-c(mean(k[5,],na.rm=TRUE),mean(k[,LRP%/%5],na.rm=TRUE),k[LRP])
                        lcolumn<-
c(mean(k[,5],na.rm=TRUE),mean(k[LCP%%5,],na.rm=TRUE),k[LCP])
                        maxes=NULL
                                        for(i in 1:4){
```

```
                                        inter<-c(mean(k[,5-
b[i]%%5],na.rm=TRUE),mean(k[b[i]%%5,],na.rm=TRUE),k[b[i]])
                                        maxes<-cbind(maxes,max(inter,na.rm=TRUE))
                                        i=i+1
                                                }
                        maxes<-cbind(maxes,max(lrow,na.rm=TRUE),max(lcolumn,na.rm=TRUE))
                        if (which.max(maxes)<5){FPR<-
rbind(FPR,FPR_table[b[which.max(maxes)]])
                                                }else{
                                                if(which.max(maxes)==5){FPR<-
rbind(FPR,FPR_table[LRP])
                                                }else{FPR<-rbind(FPR,FPR_table[LCP])
                                                }}
                        if (which.max(maxes)<5){TPR<-
rbind(TPR,TPR_table[b[which.max(maxes)]])
                                                }else{
                                                if(which.max(maxes)==5){TPR<-
rbind(TPR,TPR_table[LRP])
                                                }else{TPR<-rbind(TPR,TPR_table[LCP])
                                                }}
                        if (which.max(maxes)<5){k[b[which.max(maxes)]]=NA
                                                }else{
                                                if(which.max(maxes)==5){k[LRP]=NA
                                                }else{k[LCP]=NA
                                                }}
                cFPR<-c(cFPR,cFPR[j]+FPR[j])
                cTPR<-c(cTPR,TPR[j]+cTPR[j])
                if(cFPR[j+1]>c){break}
                j=j+1}
                ltpr=tail(cTPR, n=1)
                lfpr=tail(cFPR, n=1)
                lg<-length(cFPR)
                TPF =(((cTPR[lg]-cTPR[lg-1])/(cFPR[lg]-cFPR[lg-1]))*(c-cFPR[lg]))+cTPR[lg]
pAUC=pAUC+((FPR[j]*cTPR[j])+(TPR[j]*FPR[j]*0.5))#j is a number so cTPRj is one row above
TPRj
vec<-c(lfpr,ltpr,TPF)
bakmat=cbind(bakmat,vec)
names(bakmat)[ncol(bakmat)] <- paste(colnames(data)[var2])
t=t+1
}
bak<-bakmat[,2:(p+1)]
shbak<-subdata[,order(-bak[3,],-bak[1,])]
byslope<-cbind(data[,1:3],shbak)
return(list(byslope=byslope,bak=bak))
}
```

```
#baker for full curve

mybaker=function(score1,score2,status){
n=length(status)
d=sum(status==1)
rank1=rank(score1)
rank2=rank(score2)
quad1=floor(2*rank1/(n+1))+1
quad2=floor(2*rank2/(n+1))+1
FPR_table=table(factor(quad1[status==0],levels=2:1),factor(quad2[status==0],levels=2:1))/
(n-d)
TPR_table=table(factor(quad1[status==1],levels=2:1),factor(quad2[status==1],levels=2:1))/
d
ratio_table=TPR_table/FPR_table
max_quad=which.max(ratio_table)#givesthepositionofmaxvalue
if(max_quad>1)
return(mybaker((2*(max_quad%%2)-1)*score1,(1-2*((max_quad-1)%/%2))*score2,status))
score1=floor(5*rank1/(n+1))+1
score2=floor(5*rank2/(n+1))+1
FPR_table=table(factor(score1[status==0],levels=1:5),factor(score2[status==0],levels=1:5)
)/(n-d)
TPR_table=table(factor(score1[status==1],levels=1:5),factor(score2[status==1],levels=1:5)
)/d
ratio_table=TPR_table/FPR_table
ratio_table[is.nan(ratio_table)]=0
ratio_table[ratio_table==0]=-FPR_table[ratio_table==0]
k<-ratio_table
k[25]=NA
FPR=FPR_table[25]
TPR=TPR_table[25]
cTPR=c(0,TPR)
cFPR=c(0,FPR)
myAUC=FPR*TPR*0.5
{for(j in 2:25)
{
M<-which((is.na(k[,]))==TRUE)
A=c(M-1)
B=c(M-5)
b<-Reduce(intersect,list(A,B))
excl<-which((is.na(k[b]))==TRUE)
b<-if(length(excl)>0){b=b[-excl]
}else{
b=b
}
LR<-c(k[5,])
LRM<-which((is.na(LR))==TRUE)
LRP=(5*(min(LRM)-1))
LC<-c(k[,5])
```

```
LCM<-which((is.na(LC))==TRUE)
LCP=(20+(min(LCM)-1))
lrow<-c(mean(k[5,],na.rm=TRUE),mean(k[,LRP%/%5],na.rm=TRUE),k[LRP])
lcolumn<-c(mean(k[,5],na.rm=TRUE),mean(k[LCP%%5,],na.rm=TRUE),k[LCP])
maxes=NULL
for(i in 1:4){
inter<-c(mean(k[,5-b[i]%%5],na.rm=TRUE),mean(k[b[i]%%5,],na.rm=TRUE),k[b[i]])
maxes<-cbind(maxes,max(inter,na.rm=TRUE))
i=i+1
}
maxes<-cbind(maxes,max(lrow,na.rm=TRUE),max(lcolumn,na.rm=TRUE))
if(which.max(maxes)<5){FPR<-rbind(FPR,FPR_table[b[which.max(maxes)]])
}else{
if(which.max(maxes)==5){FPR<-rbind(FPR,FPR_table[LRP])
}else{FPR<-rbind(FPR,FPR_table[LCP])
}}
if(which.max(maxes)<5){TPR<-rbind(TPR,TPR_table[b[which.max(maxes)]])
}else{
if(which.max(maxes)==5){TPR<-rbind(TPR,TPR_table[LRP])
}else{TPR<-rbind(TPR,TPR_table[LCP])
}}
if(which.max(maxes)<5){k[b[which.max(maxes)]]=NA
}else{
if(which.max(maxes)==5){k[LRP]=NA
}else{k[LCP]=NA
}}
cFPR<-c(cFPR,cFPR[j]+FPR[j])
cTPR<-c(cTPR,TPR[j]+cTPR[j])
myAUC=myAUC+((FPR[j]*cTPR[j])+(TPR[j]*FPR[j]*0.5))#j is a number so cTPRj is one row
above TPRj
j=j+1}
}
FPR<-rbind(0,FPR)
TPR<-rbind(0,TPR)
result<-cbind(FPR,TPR,cFPR,cTPR)
return(list(result=result,myAUC=myAUC))
}
#mybaker(data[,4],data[,5],data[,2]


# to calulate partial AUC using baker
pAUC_bak<-function(score1,score2,status,c){
n = length(status)
d = sum(status==1)
dataset<-data.frame(score1,score2)
bakmat<-data.frame(colnames(dataset)[1])
        rank1=rank(score1)
        rank2=rank(score2)
```

```
        quad1=floor(2*rank1/(n+1))+1
        quad2=floor(2*rank2/(n+1))+1
        FPR_table =
table(factor(quad1[status==0],levels=2:1),factor(quad2[status==0],levels=2:1))/(n-d)
        TPR_table =
table(factor(quad1[status==1],levels=2:1),factor(quad2[status==1],levels=2:1))/d
        ratio_table = TPR_table/FPR_table
        max_quad = which.max(ratio_table)# gives the position of max value
        if(max_quad > 1){
        rank1=rank((2*(max_quad%%2)-1)*score1)
        rank2=rank((1-2*((max_quad-1)%/%2))*score2)
        }
        score1 = floor(5*rank1/(n+1))+1
        score2 = floor(5*rank2/(n+1))+1
        FPR_table =
table(factor(score1[status==0],levels=1:5),factor(score2[status==0],levels=1:5))/(n-d)
        TPR_table =
table(factor(score1[status==1],levels=1:5),factor(score2[status==1],levels=1:5))/d
        ratio_table = TPR_table/FPR_table
        ratio_table[is.nan(ratio_table)] = 0
        ratio_table[ratio_table==0] = -FPR_table[ratio_table==0]
        k<-ratio_table
        k[25]=NA
        FPR=FPR_table[25]
        TPR=TPR_table[25]
        cTPR=c(0,TPR)
        cFPR=c(0,FPR)
        pAUC=FPR*TPR*0.5
        for (j in 2:n){
                        M<-which((is.na(k[,]))==TRUE)
                        A=c(M-1)
                        B=c(M-5)
                        b<-Reduce(intersect, list(A,B))
                        excl<-which((is.na(k[b]))==TRUE)
                        b<-if (length(excl)>0){b=b[-excl]
                                                }else{
                                                b=b
                                                }
                        LR<-c(k[5,])
                        LRM<-which((is.na(LR))==TRUE)
                        LRP=(5*(min(LRM)-1))
                        LC<-c(k[,5])
                        LCM<-which((is.na(LC))==TRUE)
                        LCP=(20+(min(LCM)-1))
                        lrow<-c(mean(k[5,],na.rm=TRUE),mean(k[,LRP%/%5],na.rm=TRUE),k[LRP])
                        lcolumn<-
c(mean(k[,5],na.rm=TRUE),mean(k[LCP%%5,],na.rm=TRUE),k[LCP])
                        maxes=NULL
```

```
                                        for(i in 1:4){
                                                inter<-c(mean(k[,5-
b[i]%%5],na.rm=TRUE),mean(k[b[i]%%5,],na.rm=TRUE),k[b[i]])
                                                maxes<-cbind(maxes,max(inter,na.rm=TRUE))
                                                i=i+1
                                                        }
                        maxes<-cbind(maxes,max(lrow,na.rm=TRUE),max(lcolumn,na.rm=TRUE))
                        if (which.max(maxes)<5){FPR<-
rbind(FPR,FPR_table[b[which.max(maxes)]])
                                                }else{
                                                if(which.max(maxes)==5){FPR<-
rbind(FPR,FPR_table[LRP])
                                                }else{FPR<-rbind(FPR,FPR_table[LCP])
                                                }}
                        if (which.max(maxes)<5){TPR<-
rbind(TPR,TPR_table[b[which.max(maxes)]])
                                                }else{
                                                if(which.max(maxes)==5){TPR<-
rbind(TPR,TPR_table[LRP])
                                                }else{TPR<-rbind(TPR,TPR_table[LCP])
                                                }}
                        if (which.max(maxes)<5){k[b[which.max(maxes)]]=NA
                                                }else{
                                                if(which.max(maxes)==5){k[LRP]=NA
                                                }else{k[LCP]=NA
                                                }}
                cFPR<-c(cFPR,cFPR[j]+FPR[j])
                cTPR<-c(cTPR,TPR[j]+cTPR[j])
                pAUC=pAUC+((FPR[j]*cTPR[j])+(TPR[j]*FPR[j]*0.5))#j is a number so cTPRj is
one row above TPRj
                if(cFPR[j+1]>c){break}
                j=j+1}
                ltpr=tail(cTPR, n=1)
                lfpr=tail(cFPR, n=1)
                lg<-length(cFPR)
                TPF =(((cTPR[lg]-cTPR[lg-1])/(cFPR[lg]-cFPR[lg-1]))*(c-cFPR[lg]))+cTPR[lg]
pAUCc=pAUC-(0.5*(lfpr-c)*(ltpr+TPF))
score1_2<-c(lfpr,ltpr,TPF,pAUC)
bakmat=cbind(bakmat,score1_2)
return(list(bakmat=bakmat,pAUCc=pAUCc))
}
##pAUC_bak(ch$byslope[,4],ch$byslope[,8],ch$byslope[,2])

# to calculate IDI
IDI = function(old,new,status)
{
        n = length(status)
        d = sum(status==1)
```

```
df <- data.frame(status,old,new,diff = new - old)
p_control = sum(df$diff[df$status==0])/(n-d)
p_case = sum(df$diff[df$status==1])/d
IDI_score = (p_case-p_control)
return(score=IDI_score)
}

### FINAL STEP TO GET OUTPUT I.E BEST VARIABLE COMBO (c=FPF at which partial AUC should
be cut off)
pAUC_IDI<-function(data,var1,varlast,s,c,IDI_cutoff){
iniROC<-ROC(data[,var1],data[,s])
ini<-pAUC(data[,var1],data[,s],c)
parauc=ini$pAUCc
logit<-glm(data[,s]~data[,var1],family=binomial(link="logit"))
res<-matrix(c(var1,ini$pAUCc,NA,iniROC$AUC,NA),5,1)
variable<-c(colnames(data[var1]))
test<-data.frame(data[,var1])
for (i in 1:varlast){
var2=var1+i
pAUC_add<-pAUC_bak(logit$fitted,data[,var2],data[,s],c)
if(pAUC_add$pAUCc>parauc){
                              newtest<-cbind(test,data[,var2])
                              names(newtest)[ncol(newtest)] <-
paste(colnames(data)[var2])
                              baker<-mybaker(logit$fitted,data[,var2],data[,s])
                              newlogit =
glm(data[,s]~.,family=binomial(link=logit),data=newtest)
                              roc<-ROC(newlogit$fitted,data[,s])
                              idi<-IDI(logit$fitted,newlogit$fitted,data[,s])
                              if (idi>IDI_cutoff){
                                                  keep<-
c(var2,pAUC_add$pAUCc,baker$myAUC,roc$AUC,idi)
                                                  res<-cbind(res,keep)
                                                  variable<-
append(variable,colnames(data[var2]))
                                                  logit$fitted=newlogit$fitted
                                                  test=newtest
                                                  parauc=pAUC_add$pAUCc
                                                          }else{
                                                          logit$fitted=logit$fitted
                                                          }
                                          }else{
                                                  logit$fitted=logit$fitted
                                                  }
i=i+1
}
allp<-glm(data[,s]~.,family=binomial(link=logit),data=data[,res[1,]])
roc<-ROC(allp$fitted,data[,s])
```

```
pauc<-pAUC(allp$fitted,data[,s],c)
res=rbind(variable,res)
res=cbind(c("var","var_pos","Baker_pAUCc","Baker_AUC","Logistic_AUC","IDI"),res)
return(list(res=res,roc=roc$AUC,partial_AUC=pauc$pAUCc))
}
#pAUC_IDI(ch$byslope,3,80,2,0.25,0.1)
```