

Quantitative urine proteomics in pregnant women for early biomarkers of preeclampsia

Supplementary Methods

Regression modeling

Feature selection in terms of which clinical covariates are important for modeling and prediction was done using two methods- relative importance and stepwise regression. R package (relaimpo) was used for both methods.

In this, first, response was linearly fitted against clinical data columns: AGEATENTRY + Height + Weight_at_the_beginning_of_pregnancy+ body mass index (BMI)_at_the_end_of_pregnancy + BMI_in_the_beginning_of_pregnancy. Then, based on relative importance and stepwise regression model selection methods (lowest AIC value), best fitting covariates was chosen.

Generalized linear modeling was used to train the dataset using logistic regression in which first, two types of dataset were prepared: training dataset and test dataset, training dataset was prepared using random sampling of patients of 0s and 1s then merge together into one matrix of 14*29. Second, test dataset was prepared in which random sampling of patients of 0s and 1s was done again then merge together into one matrix of 17*29.

Using glm function and logistic regression link function to prepare the model for training dataset , glm(X0_no_pre_eclampsia__1_Preeclampsia__not_severe__2__severe_preeclampsia ~

BMI_in_the_beginning_of_pregnancy,

data=trainingData, family=binomial(link="logit"))

Prediction on test data set was done using predict function using following below command:

`predict(logitMod, testData, type="response")`. Optimal cutoff threshold value was obtained with help of InformationValue R pkg. Correlation between covariates and protein expression value was obtained by using pearson correlation to correlate “BMI in the beginning of pregnancy (BMI)” with biomarker expression.

Biomarker selection

R Package *ropls* was used for calculating VIP values in Partial Least Square-Discriminant (PLS-DA) analysis modeling for first trimester comparisons. It was done in the following way:

Pre vs Risk

```
PE.plsda1 <- oppls(tdf_wo_last[1:13,], tdf[1:13,362],predI=10 +,scaleC='pareto')
```

13 samples x 361 variables and 1 response

pareto scaling of predictors and response(s)

Pre vs Healthy

```
PE.plsda2 <- oppls(tdf_wo_last[c(1:7,14:17),], tdf[c(1:7,14:17),362], + predI=10,scaleC='pareto')
```

PLS

11 samples x 361 variables and 1 response

pareto scaling of predictors and response(s)

Risk vs Healthy

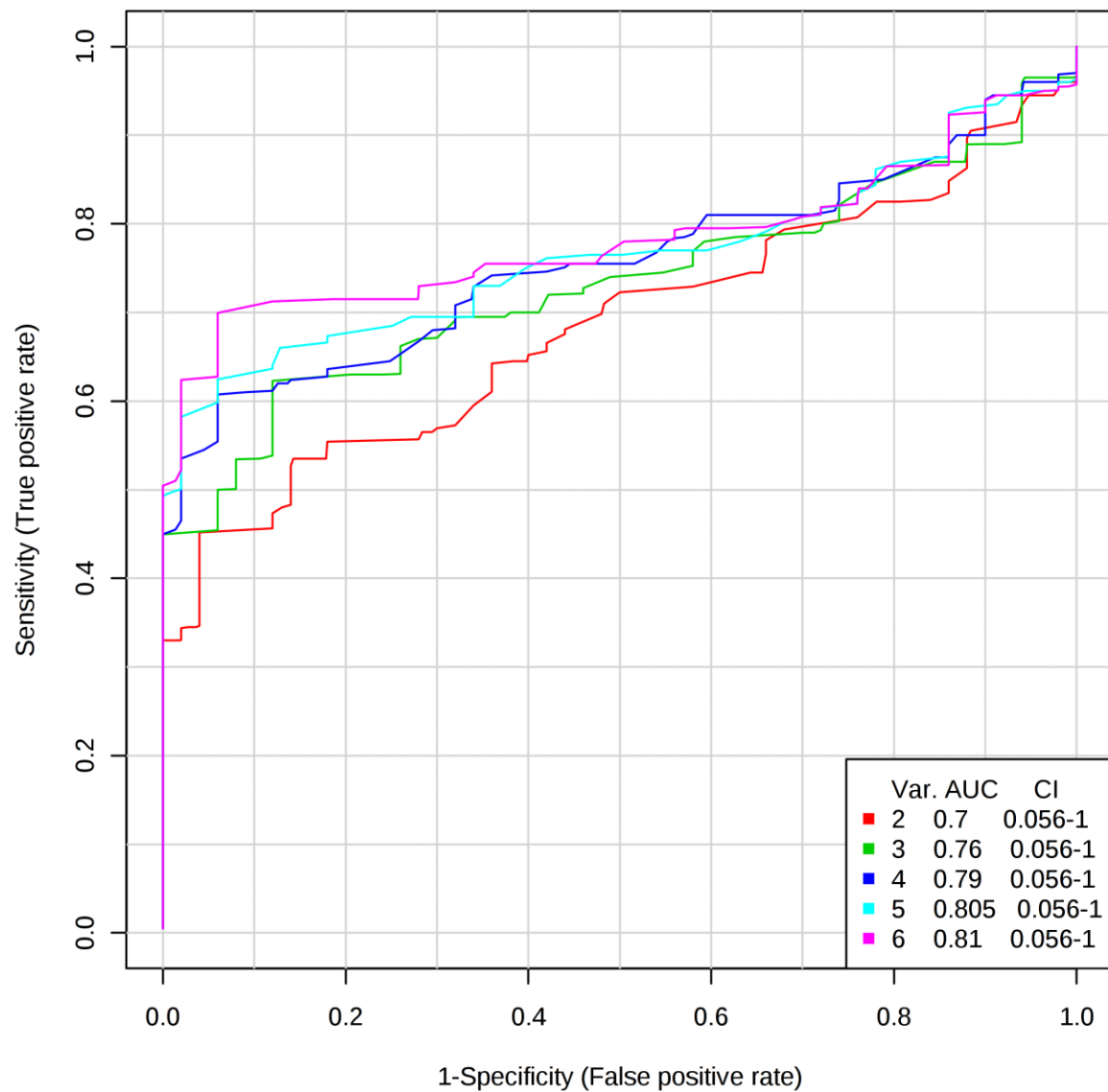
```
PE.plsda3 <- oppls(tdf_wo_last[c(8:17),], tdf[c(8:17),362],predI=10,scaleC='pareto')
```

PLS

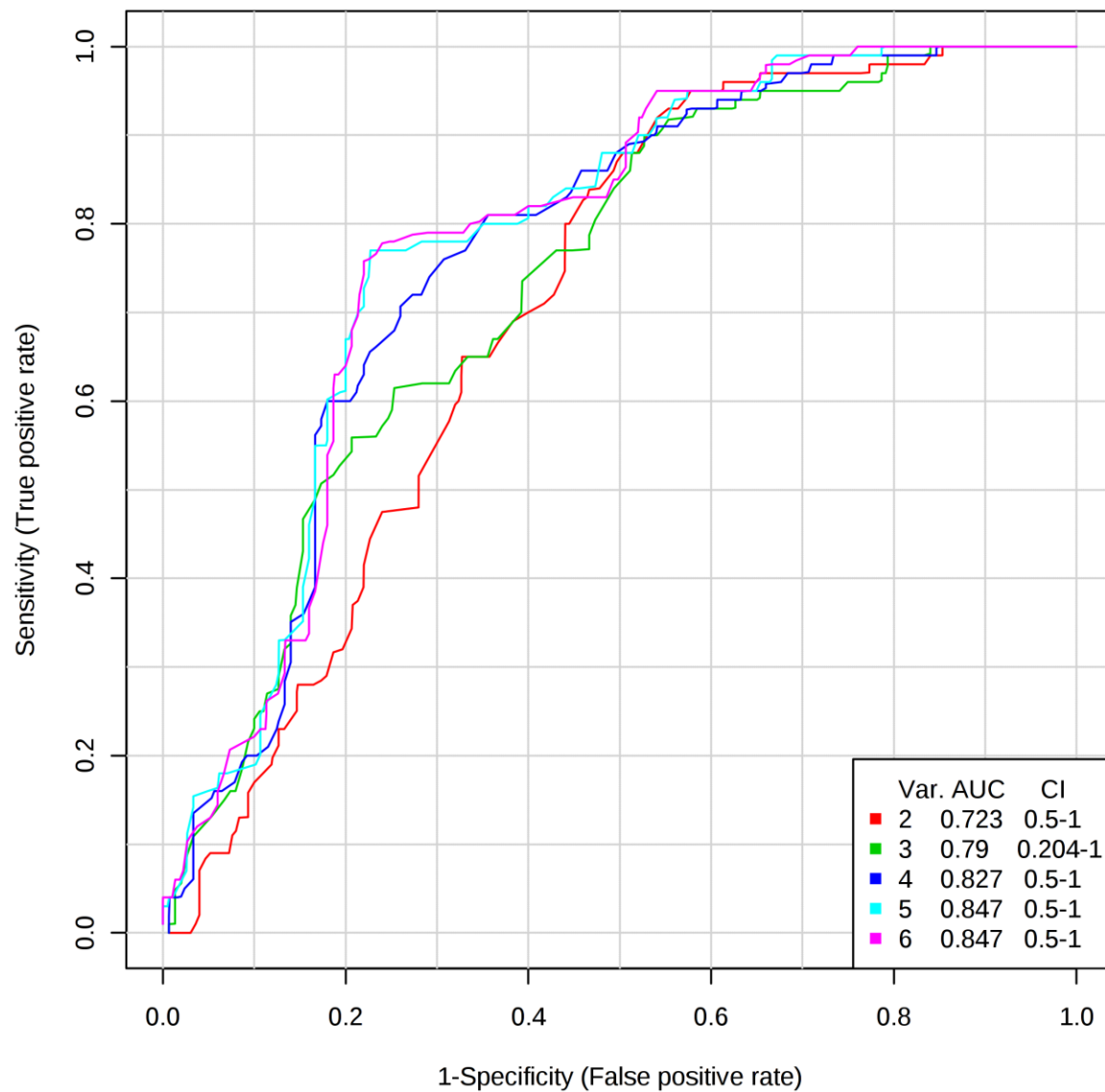
10 samples x 359 variables and 1 response

pareto scaling of predictors and response(s)

2 excluded variables (near zero variance)



Supplementary Figure 2: Receiver operating characteristics (ROC) Curve Analysis. ROC curve was used to create a biomarker panel of 6 proteins selected by 3 independent methods of variable selection to separate Preeclampsia vs Healthy controls. Combined area under the curve values of various combinations of these are shown in the figure with confidence intervals.



Supplementary Figure 3: Receiver operating characteristics (ROC) Curve Analysis. ROC curve was used to create a biomarker panel of 6 proteins selected by 3 independent methods of variable selection to separate Preeclampsia vs Risk controls. Combined area under the curve values of various combinations of these are shown in the figure with confidence intervals.