Joshua Chang

Pre-hospital Stroke Diagnosis

Abstract

There is strong interest in pre-hospital diagnosis of stroke in order to enable rapid acquisition in the acute care setting right when patients arrive. The data provided included both clinical variables and pre-transformed Electroencephalogy (EEG) signals. The main goals of this study are to examine the association between the method of prediction (RACE) in the presence of just clinical variables, as well as in the presence of EEG and clinical variables.

In the examination of the effect of the clinical variables on stroke, RACE scores were still found to be a strong predictor of stroke, as each increase in RACE value multiplicatively increased the odds that the patient had a stroke by 2.216. There was also found to be a spline function associated with age of patient. Prior to the age of 67, a one year increase in age multiplicatively increased the odds of having a stroke by 1.383. After the age of 67, a one year increase in age multiplicatively decreased the odds of a female having a stroke by .96.

A prediction model with only clinical variables was compared against a prediction model that had both EEG and clinical variables. The comparison of performance between them was assessed through a 10-fold cross validation averaged value of AUC. The model including just the clinical variables of age and RACE scores ended up being the superior model for prediction. Setting the threshold of Stroke classification at 0.45, the model correctly diagnosed those who had strokes 64.28% of the time.

Introduction

The current method of assessing whether patients had a stroke or not is the Rapid Arterial Occlusion Evaluation (RACE) score. However it has been shown that some additional clinical variables, along with EEG changes immediately after brain ischemia could be helpful in assessing patients with strokes.

Patients admitted at a local hospital's Emergency Department were examined. The clinical variables considered are RACE score, Last Known Well (LKW) time in hours, Gender, and Age. Additionally, 100 EEG signals denoted E1 to E100 were measured. The outcome of interest is whether or not the patient suffered a stroke.

The main questions of interest are as follows:

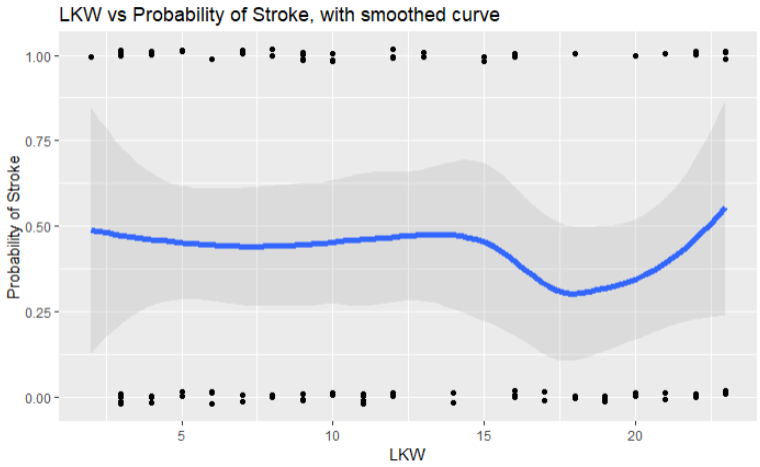
(1) There is interest in seeing how the other recorded clinical variables affects the relationship between RACE and Stroke.
(2) Analyze the hypothesis that the association between RACE and stroke varies by age (in the scope of clinical variables)
(3) Build a model for pre-hospital stroke prediction using both the EEG and clinical variables. Evaluate and compare the performances of the simpler model with just clinical variables and the model that includes EEG variables.

Statistical Methods

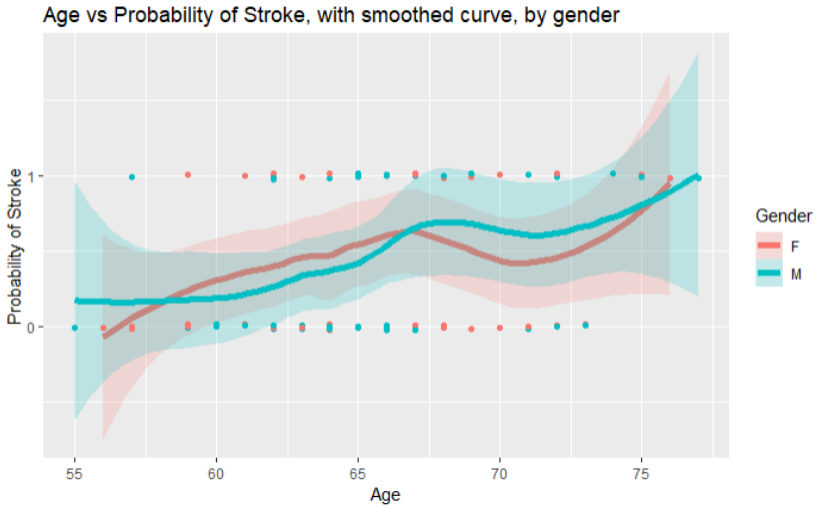
Among the data associated with clinical variables, six individuals out of the one hundred had some missing data. With the assumption that the data is missing at random, the six observations were omitted from model building in this situation.

However, among the data associated with the clinical and EEG variables, 65 people had some sort of missing data, thus some method of imputation was required. The MICE library was used to perform multiple imputation on the data, again with the assumption that data are missing at random. The assumption seemed valid, since none of the variables had more than 4% of their observations missing. Five different imputations were pooled together for the purpose of analysis on the EEG variables.

Insights from Data Exploration



The effect of LKW on probability of having a stroke seems to be non existent from values between 3 and 13, however there is indication of a possible quadratic or cubic effect during the years after. A possible spline component will be looked at.



The lowess smoothed curves on the plot of age vs probability of stroke suggests

that there is a linear trend between age and the probability of having a stroke from the ages of 55 to 67. When separating the data by gender, it seems that males are at higher risk of having a stroke after the age of 67, while females are at higher risk of having a stroke before the age of 67. A possible age related spline effect and interaction between gender and age will be taken into account when model building.

Model Building and Selection

A simple backwards selection algorithm using AIC as performance metric found the simple model with just RACE and Age to be the best. While investigating more complex models that incorporated splines, I made a note of the models that had for the most part significant coefficients while also having a comparable AIC score. The three models that I was left with in the end are as follows:

(1) Stroke ~ Age + RACE
- The model that includes just age and RACE.

(2) Stroke ~ Age + AgeS + RACE + Gender + Gender*AgeS + LKW + LKWS + LKWS^2
- The model additionally has a spline function on the interaction between age and gender after the age of 67, as well as a quadratic spline function for LKW values higher than 14.

Model	AIC	BIC	LRT p.	LOO CV Acc.	Signif. Variables (alpha = .05)
Model (1)	110.7	118.3277		0.732	Age, RACE
Model (2)	109.7	133.5793	0.039	0.714	Age, AgeS, RACE, LKW, LKWS, LKWSQ, AgeS*Gender

Table of AIC, BIC, LRT p-values, and Leave-one-out CV accuracy

Model (2) performed well despite its complexity in comparison to Model (1). The likelihood ratio test comparing Model (1) to Model (2) concluded that Model (2) provided significantly information to predicting stroke than Model (1). The AIC for Model (2) was smaller than Model (1)'s, although the BIC was much larger in (2) due to the large amount of extra variables added.

The results of the AIC, LRT, and significant coefficients indicate that (2) is the better model, but it all comes at the risk of overfitting to the data due to the low number of observations we have (100), and six extra variables that (3) takes into account. To assess the degree to which (2) overfit the data, I performed LOO CV, with the default threshold of $p > .5$ to detect stroke, which showed that the full model was not too far off from the reduced model in terms of variance. The VIF values were checked for the variables in Model (2), and none of them seemed to indicate troublesome multicollinearity.

I decided to continue the analysis using Model (2), mainly because I wanted to investigate how the other clinical variables affected the relationship between RACE and Stroke. Picking Model (2) instead of Model (1) will risk overfitting to the limited data that we have, which is something that will have to be kept in mind

Model Equation

$$logit(P(y=1)) = \beta_0 + \beta_1 RACE + \beta_2 Male + \beta_3 Age + \beta_4 (Age-67)_+ + \beta_5 (Age-67)_+ Male + \beta_6 LKW + \beta_7 (LKW-13)_+ + \beta_8 (LKW-13)_+^2$$

- Where y = 1 is incidence of stroke.
- $(LKW-13)_+ = 0$ if $LKW - 13 \leq 0$, $LKW - 13$ else

Model Analysis:

Variables	Exp. Coeff	Lower 95% CI	Upper 95% CI	p-value
RACE	2.216	1.486	3.5122	.0002
GenderM	0.303	0.077	1.057	.0703
Age	1.383	1.125	1.765	.0042
AgeS	0.958	0.922	0.99	.0165
AgeS*GenderM	1.034	1.002	1.073	.0556
LKW	0.897	0.720	1.107	.3197
LKWS	0.680	0.473	0.942	.0264
LKWSQ	1.022	1.004	1.04	.0190

Association between RACE and STROKE given other clinical variables

One of the main study goals was to examine the association between RACE and Stroke given the other clinical variables. There are no interactions between RACE and the other variables, so interpretation of RACE on the probability of stroke will be done assuming all other factors are held fixed.

We are 95% confident that a one unit increase in RACE corresponds to the odds of having a stroke increasing multiplicatively within the range of (1.486, 3.5122), assuming all other factors are fixed. Thus, higher values of RACE is a strong indicator of whether or not a patient has suffered a stroke.

RACE ranges in values from 0 to 4, so in the most extreme scenario we would say that the odds of having a heart attack for a patient with RACE score of 4 is 17.47 times higher than the odds for a patient with a score of 0, assuming the other variables are held fixed.

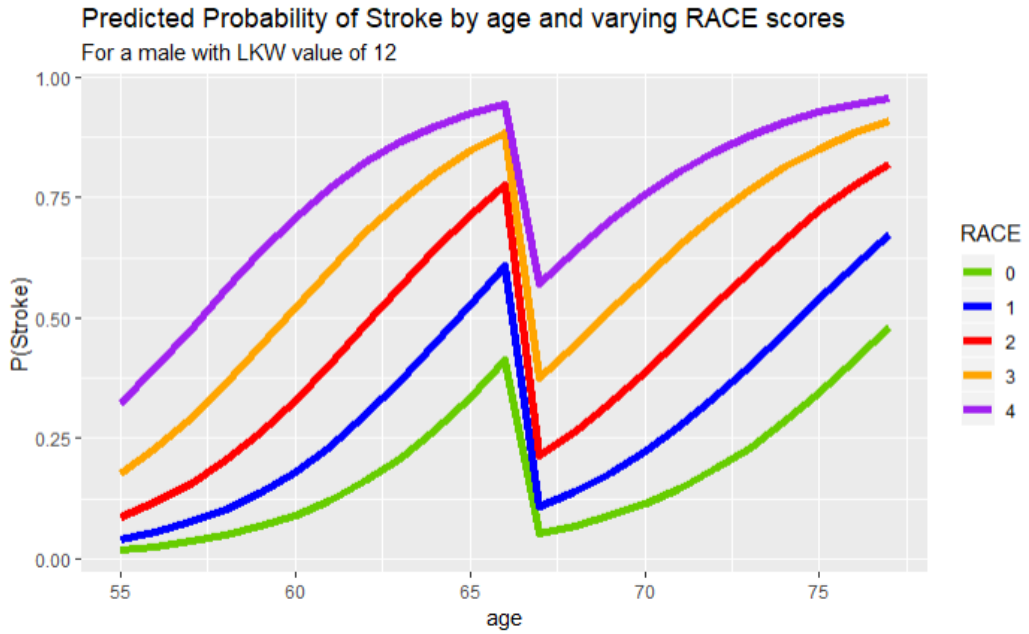
Association between RACE and stroke by age

It was hypothesized that association between RACE and stroke varies by age, and the analysis of our model seems to confirm that belief. The significance of the spline component suggests that increasing values of age prior to hitting 67 increases the odds that the patient suffered a stroke, but after the age of 67 the effects of increasing age starts to dissipate.

We are 95% confident that a one year increase in age prior to the age of 67 multiplicatively increases the odds of having a stroke between the range of (1.125, 1.765), assuming all other variables are held constant.

We are 95% confident that a one year increase in age after the age of 67 multiplicatively decreases the odds of having a stroke between the range of (0.922, 0.99) for females, assuming all other variables held constant.

We are 95% confident that a male older than 67 had odds of having a stroke being between (1.002,1.073) times higher than the odds of stroke for a female at the same age and with the same other variable values. (should be interpreted with caution, as not quite significant).



The above is a plot of predicted probabilities of a patient having had a stroke, by increasing age and varying race scores, for a given male with an LKW value of 12. It can be seen that the predicted probabilities increase linearly from ages 55 to 67, and then drop sharply after the age of 67 due to the spline coefficient. The probabilities then start to increase over time again. One possible reason for the drop in probability after the age of 67 could be due to other medical incidents that more commonly happen to patients after that age.

The varying influence of RACE scores on predicted probabilities can also be seen. For instance, at the age of 65, an individual with a RACE score of 0 is assessed a 35% probability of having had a stroke, while an individual at that age with a RACE score of 0 is assessed roughly a 92% probability of having had a stroke.

EEG Model

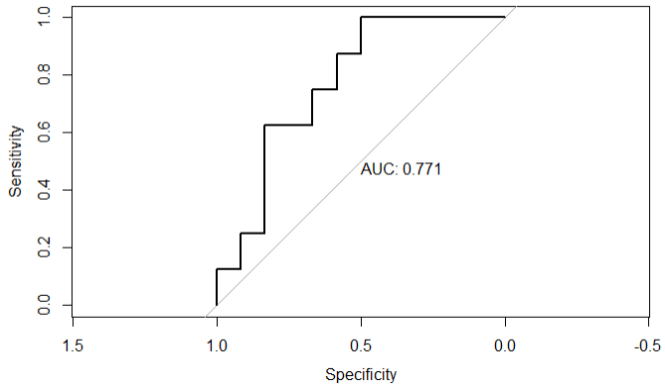
When taking into consideration the EEG variables, there are now more variables (106), than there are observations (100), which poses a problem for many model selection techniques. LASSO regression is one way to deal with this problem, as it induces shrinkage into the majority of the variables, leaving only a few parameters. I will be fitting the data including the EEG variables subject to the LASSO regression's constraints, using the glmnet and caret library.

Prior towards model building, the data was first randomly divided such that 80 observations were put in the training set, and 20 observations were put in a testing set. Cross validation was then used to find the optimal value of the LASSO's regularization parameter, which turned out to be 0.0653. The model was then fit using the parameter and the results are as follows.

Significant Variables on training set

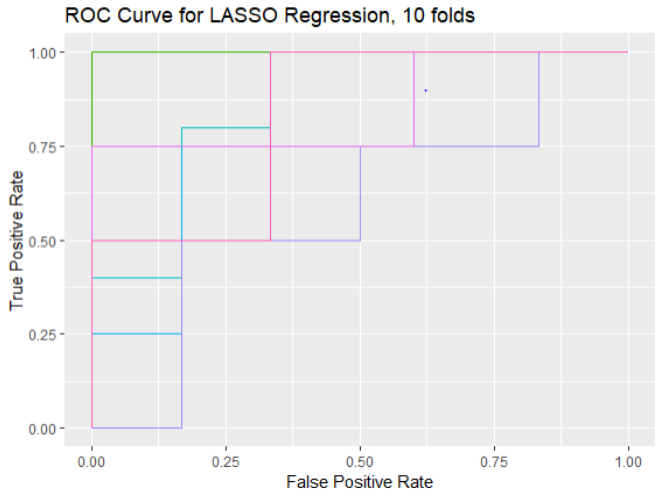
Variable	Coefficient
RACE	0.2029
E6	-0.2328
E25	0.2996
E53	-.0205
E54	-.0183

ROC Curve and AUC on the testing set:



LASSO Regression on the training set decided to keep the clinical variable RACE and the EEG variables E6, E25, E53, and E54. Area under the ROC curve was found to be 0.771, so the model has decent capability in correctly diagnosing patients.

Since we only have 100 observations, a train-test split is not enough to obtain a reliable performance metric. 10-fold cross validation will be conducted in order to obtain an averaged AUC metric which we will use to compare this model against the clinical variable model. The plot below illustrates the 10 ROC curves fitted on each of the cross validation folds, as well as averages the computed AUC for each fold. In this case, the average AUC was found to be 0.71

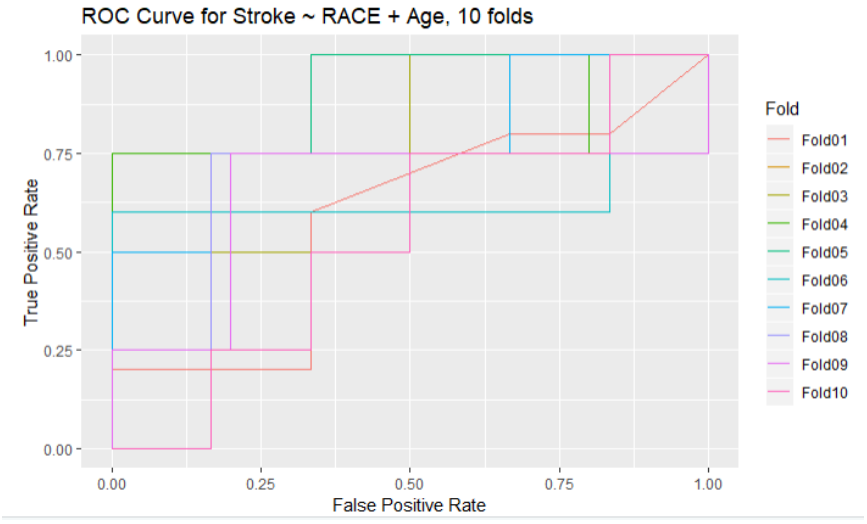


Resample <chr>	auc <chr>
Fold01.Rep1	0.6363636
Fold02.Rep1	0.9
Fold03.Rep1	0.7
Fold04.Rep1	0.6666666
Fold05.Rep1	0.8
Fold06.Rep1	0.6363636
Fold07.Rep1	0.7
Fold08.Rep1	0.6
Fold09.Rep1	0.7777777
Fold10.Rep1	0.7
average	0.7117171

Clinical Variables model:

The more complicated model that I investigated in the earlier parts would most likely be too complex to refit and perform prediction with across the folds, so I will be analyzing the performance of the simple model with just Age and RACE as variables, more specifically the average AUC values.

The output below indicates that the clinical model with just RACE and Age has a noticeably higher average averaged AUC in comparison to the one obtained by the LASSO model. I will be continuing the analysis by looking more in depth at the predictive capabilities of this clinical model.



Resample <chr>	auc <chr>
Fold01	0.6363636
Fold02	0.8
Fold03	0.8
Fold04	0.7777777
Fold05	0.8
Fold06	0.7272727
Fold07	0.7
Fold08	0.8
Fold09	0.7777777
Fold10	0.6
average	0.7419191

Predicted Probabilities

To further examine the performance of the clinical model, a pooled confusion matrix was generated through 10-fold cross validation. The predicted vs actual stroke diagnosis was added up from each of the testing folds and summarized in these tables. The threshold for predicting stroke was set to 0.5.

	Actual Stroke Diagnosis	
Predicted Stroke	No	Yes
No	45	20
Yes	13	22

Table of Pooled Confusion Matrix

	Percentage
No Stroke Prediction Accuracy	0.6923
Stroke Prediction Accuracy	0.6286
Percent of No Stroke Correctly Diagnosed	0.7759
Percent of Stroke Correctly Diagnosed	0.5238

Table of Pooled Model Performance

The model correctly assessed that the patient had a stroke 62.86% of the time, and correctly assessed that the patient did not have a stroke 69.23% of the time. Out of the people that actually had strokes, only 52.38% of them were correctly diagnosed, and out of the people that did not have strokes, 77.59% of them were correctly diagnosed.

The default threshold at p = 0.5 is probably not ideal for this situation. We would like to try to maximize the percentage of people with strokes that are correctly diagnosed, in order to more accurately assess those that truly need emergency care. However, we also do not want to incorrectly diagnose those who did not actually suffer a stroke, as it could take away valuable time and resources away towards helping the true stroke victims.

I looked at varying values of threshold and believe that p = .45 was able to maximize the percent of stroke correctly diagnosed, while also minimizing incorrect diagnosis on those who didn't have strokes.

	Percentage
No Stroke Prediction Accuracy	0.732
Stroke Prediction Accuracy	0.6136
Percent of No Stroke Correctly Diagnosed	0.707
Percent of Stroke Correctly Diagnosed	0.6428

Performance at threshold p = .45

Changing the threshold improved the predictive capability on patients who actually suffered strokes by roughly 12%, while also minimizing and even improving some of the other performance metrics. Lower threshold values were also examined, but they did not provide much improvement towards correct stroke diagnosis in comparison to the loss in performance of predictive abilities.

Discussion

With the other clinical variables introduced to the model, increasing RACE scores was still a strong indication of having had a stroke. Each unit increase in RACE value multiplicatively increased the odds that the patient had a stroke by 2.216. There was also an indication that spline functions on both age and LKW values helped to improve model significance. LKW values before 13 did not seem to have an impact on the model, but the values after 13 appeared to quadratically decrease the odds of having a stroke. Prior to the age of 67, a one year increase in age multiplicatively increased the odds of having a stroke by 1.383, assuming other variables were held constant. After the age of 67, a one year increase in age multiplicatively decreased the odds of a female having a stroke by .96, assuming other variables were held constant..

A prediction model with only clinical variables was compared against a prediction model that had both EEG and clinical variables. The comparison of performance between them was assessed through a 10-fold cross validation averaged value of AUC. The model including just the clinical variables of age and RACE scores ended up being the superior model for prediction in comparison to a LASSO model that kept the variables RACE, E6, E25, E53, E54.

10-fold cross validation was used to generate a pooled confusion matrix, over varying threshold values of the predicted probabilities. The best threshold value was found to be at $p = 0.45$, as it led to correctly assessing patients who had strokes 64% of the time, while not sacrificing the model's performance on those who did not suffer a stroke.

Limitations

Only 100 observations were available for analysis, so the spine model examined in the early part of the report is likely to overfit the available data. It would have been better to use cross validation to fit a model numerous times to the data and perhaps use the averaged AIC as a performance metric.

Missing data was not an issue in the clinical setting because only 6 individual had missing observations, and they could be safely removed assuming the data was missing at random without too much loss of data. However, when including the EEG variables multiple imputation was needed in order to preserve all of the observations. This introduces some possible bias to the dataset, but there was an attempt to minimize the bias by pooling over 5 separate imputations.

I had expected the LASSO model to perform better than the clinical model, but perhaps the LASSO model was too overfit to the training data.

There are many metrics of performance of predictive capabilities for logistic regression models other than AUC, so it would have potentially been better to have also examined other metrics like deviance or correctly predicted outcomes. Between the clinical only model and the model including EEG variables.