**Supplementary File on Statistics**

This supplementary file provides a brief description of the penalized logistic regression. For a fuller description of the method, we refer to the paper by Park and Hastie [12](#_ENREF_9). We begin by discussing linear regression in order to understand the issue of collinearity. This will be followed by logistic regression and penalized logistic regression.

*Linear Regression*

In linear regression, the aim is to find the least squares solution *β* that minimise the error term of the equation.



The least squares solution for *β* refers to the fitting of the line between the intercept and the predictor variables X such that the Euclidian distance between the observed variables *Y* and predicted variables  are as small as possible. The solution for the regression coefficient is possible if the correlation matrix (between the predictor variables and its transpose) or  can be inverted. If the inverse solution does not exist and the determinant of the correlation matrix is zero, then collinearity exists.

*Collinearity*

Collinearity of the predictor variables implies redundancy in the data because of linear dependence between each of the predictor variables [6,11](#_ENREF_22). This issue can lead to overfitting of the data and unstable regression coefficients. In our study, collinearity exists because the ASPECTS regions have related function and belong to the same arterial territory6,11. Problems relating to collinearity can be seen when one examines the relationship between the variance of the regression coefficient *V(βi)* and the multiple correlation coefficientbetween an ASPECTS region with each of the other nine regions.



As the strength of the linear relationship between the predictor variables (ASPECTS regions) increases,  approaches 1 and *V()* becomes very large and tends towards infinity. Further, small changes in the dependent variables lead to fluctuations in the regression solution. Diagnostic tests for collinearity are based on this relationship between the variance of the regression coefficient *V(βi)* and the multiple correlation coefficient.

*Logistic regression*

In this study, we used binary outcome measures as the dependent variable (infarct versus no infarct). The regression model now requires a non-linear link function *g* such as logit link function.

.

This requires transformation into logit variables, the natural logarithm of the odds of the outcome (probability of the outcome divided by one minus the probability of the outcome). The term *L* here refers to the likelihood function and *l* refers to the log function.



For binary dependent variable where the error variances are not equal, the assumptions of least squares fitting for the errors to be normally distributed are not met. Consequently, the regression coefficients are better estimated by the maximum likelihood estimation (MLE). Maximum likelihood estimation (MLE) is the process of finding the parameter that maximises the likelihood for the observed data. Unlike linear least squares regression, there is no close solution for the MLE and an iterative trial and error method is used to find the MLE (iterative reweighted least squares).

*Penalized regression*

Penalized/Ridge regression is a method used to overcome collinearity in the columns of the predictor variables. From the discussion of collinearity given above, the inverse solution can be found by introducing a bias term to the correlation matrix. The effect of this bias term is that it leads to restriction in the size of the variance for the parameter estimate. The tuning parameteris added to the diagonal elements of the matrix to be inverted  to encourage non-singularity. The mean squared error of parameter estimates decreases as value increases for a certain extent beyond zero.

The cost-complexity statistic is defined in this case as the deviance plus the product of the complexity parameter and the degree of freedom12. Deviance is defined as the residual sum of squares or errors between the observed values and the model.

.

The  value, which results in the regression model with the minimal Bayesian Information Criterion (*BIC),* is chosen as the desired value. refers to the maximum likelihood.

We therefore introduced robust estimation of maximum likelihood estimators by penalizing the  norm or ridge estimator [12](#_ENREF_9). The  norm refers to the quadratic penalization of the parameter estimate of the maximum likelihood. The ridge estimator behaves like a curved estimator and as the penalty is increased, all parameters are reduced while still remaining non-zero. The resultant estimates are smaller than those obtained from logistic regression model. This method of penalisation is different from the penalisation the norm of the coefficient or *LASSO* (least absolute shrinkage and selection operator). The LASSO estimator behave in a more linear fashion in which increasing the penalty will result in more and more of the parameters to be driven to zero.



This PLR method employed a forward and backward stepwise penalized logistic regression and used all the ASPECTS regions in the analysis. It called on the penalized function in R programming environment). This program also assessed the interaction of factors on the regression model in the following manner. The choice of factors to be added to the stepwise regression is based on the cost complexity statistics. The asymmetric hierarchy principle[15](#_ENREF_13) was used to determine the choice of interaction of factors. In this case any factor retained in the model can form interaction with those not yet in the model. In this analysis, we have specified a maximum of five terms to be added to the forward selection procedure.

The advantage of this PLR method is that it can fit interactions between a large number of categorical variables and handle cells with zero values. The disadvantage of this method is that the quadratic penalization term may result in breaking up the large main effect coefficient into a sum of smaller interaction pieces [12](#_ENREF_9). This effect is made less likely by minimizing the tuning parameter  and consequently the penalty term.