Statistical Studies for SNP association in Acute Coronary Syndrome ex vivo Use of Agonists and Nanoparticles

Puspita Das Roy1, Cynthia Basu2, Sonali Das3 and Anjan Das Gupta1
1 Department of Biochemistry, University of Calcutta
2 Indian Statistical Institute, Kolkata
3 Council for Scientific and Industrial Research, Pretoria

Background information

Acute Coronary Syndrome (ACS), nick named as heart attack, is becoming one of the more frequent causes of death in today's fast paced stressed out life, 4.8% of total death around the globe is caused by ACS making it the 5th largest leading cause of death[1]

ACS, medically referred to as myocardial infarction, results from pathological thrombus formation and vascular occlusion in the coronary artery[2]. In layman's term, this means that the blood contains platelets which play their role by aggregating within themselves. Platelets are specialized disk-shaped cells in the blood stream that are involved in the formation of blood clots. Usually this is a good role they play especially when it occurs in order to prevent further blood loss during bleeding. However, sometime this platelet aggregation within the blood vessels cause the blocking of blood flow to the heart and hence cause what we know as heart attack, strokes, and peripheral vascular disease. In cardiovascular disease, abnormal clotting occurs that can result in heart attacks[3].

One of the important events in this process is the positive feedback cycle, initiated when platelets release several secondary mediators, such as Adenosine-di-phosphate (ADP), serotonin. These agonists lead to activation and further aggregation of platelets. Activated platelets release more ADP. Abnormalities in ADP receptors, which are mediators in the above cycle, can predispose individuals to the formation of abnormal thrombus and hence ACS[4] (give some Ref).

It is believed that ACS is therefore caused by two types of factors, namely environmental factors such as stress and pollution and genetic factors).

---

[1] . World Health Organisation (2008)

[2] M. J. Davies, T. Thomas, J. McMichael and P. D. Richardson; The Pathological Basis and Microanatomy of Occlusive Thrombus Formation in Human Coronary Arteries [and Discussion] ; Phil. Trans. R. Soc. Lond. B 1981 vol. 294 no. 1072

[3] David Gregg, Pascal J. Goldschmidt-Clermont; 2003; Platelets and Cardiovascular Disease; Circulation; 108: e88-e90;pg 2

[4] José Luis Ferreiro and Dominick J. Angiolillo; 2011; Diabetes and Antiplatelet Therapy in Acute Coronary Syndrome; Circulation; 123:798-813

Blood vessels injured by smoking, cholesterol, or high blood pressure develop cholesterol-rich plaques that line the blood vessel[5] ; these plaques can rupture and present sites for unwanted platelet binding. Following this, events occur which lead to formation of abnormal thrombus which blocks an intact blood vessel. Again, inherited abnormalities in components of the feedback cycle predispose individuals to development of disease.

Problem

In this investigation, we have data from both ACS patients and Non-ACS individuals. Since we know that environmental and genetic factors are responsible for triggering ACS, our primary focus question is "Is there any feature I can look at and predict whether a person is more or less prone to ACS?"

We thus l look for factors that help determine if a person is more or less prone to the condition of ACS from the given factors. We will then investigate if there are any interactions between these factors, i.e. a factor on its own may cause, say, a person to be more vulnerable to the disease but when this very feature appears in a person with another feature, the person becomes less prone to the disease. We also want to identify the variables that are more informative than the others, so that in future if these results are put to the simple practical purpose of predicting for a person, we can do so by noting down less information and hence saving on time and data collection. The other futuristic use of this variable filtering would be for the next generation DNA sequencing where we will have a very large amount of data and one good way to tackle that would be to extract and deal with only the bit that is important. In other words, we are looking for parsimony in our result[67]. (give ref on next gen dna).

Since we know genetic factors may cause ACS we will look at the Single Nucleotide Polymorphism (SNP) combinations.

Method of Experiment

Blood samples were collected from a hospital in India. A total of 177 blood samples were collected from both patients with ACS (N=91) and non-ACS (N=86) individuals depending upon their matching ethnicity and particular age bar was maintained during blood collection. As ACS patients were brought in their blood samples were collected and this group of individuals formed the case group.

---

[5] David Gregg and Pascal J. Goldschmidt-Clermont, Platelets and Cardiovascular Disease, Circulation 2003, 108:e88-e90.

[6]Mc Kenna *et al;* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data; Genome Res. 2010 September; 20(9): 1297–1303

[7]Jay Shendure & Hanlee Ji; Next-generation DNA sequencing; Nature Biotechnology, 2008

Patients with ACS were under the medication of anti-platelet drugs for 4-5 days and the non-ACS individuals were without any bleeding disorder and were not taking any anti-platelet drug. Blood samples were taken from some other individuals at the hospital, who could have been there for any other reason but ACS. This group of individuals formed our control group. The study was approved by the Calcutta Medical College Ethics Committee, and all subjects gave written informed consent. The blood samples were then carefully studied and several features were noted down, namely

1. Sex
2. Age
3. Family history of disease
4. Smoking
5. Systole
6. Diastole
7. Pulse rate
8. Medication
9. Percentage aggregation and then disaggregation of blood on addition of Adenosine-di-phosphate (ADP) of quantity
    a. 10mμ
    b. 2.5mμ
10. Percentage aggregation and then disaggregation of blood on addition of Adenosine-di-phosphate (ADP) of quantity
    a. 10mμ plus gold Nano-particles
    b. 2.5mμ plus gold Nano-particles.
11. SNP combinations
    a. SNP1 - rs701265 P2Y1-3 1622 A>G
    b. SNP2 - rs2046934 P2Y12-2 744 T>C
    c. SNP3 - rs6785930 P2Y12 234 G>A
    d. SNP4 - P2Y12 1622 C>T
    e. SNP5- rs6803224 P2Y12 2014C>T

At this point it will be worth a mention that due to the instrument breaking at certain stages of the data collection process, of the 91 individuals in the case group, the values of blood aggregation could be noted for only 45 individuals. For the control group as well, we have data on the blood aggregation of only 44 out of the 85. Hence if we are to discard data on an individual for whom any one variable value is missing, we will be losing a lot of data, affecting result reliability.

Statistical Analysis

The aim in this study is to devise a prediction rule. For this purpose we decided to look at supervised and unsupervised learning, i.e. classification clustering techniques.

Among the classification techniques we have the Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis(QDA) which basically uses the idea that we divide the entire feature space into the class's partitions with a line or a plane or hyper plane for LDA and by a quadratic curve, surface and so on in QDA[8]. Note that both these techniques cannot be used on categorical data and some of our variables, such as SNP combinations are categorical in nature. Hence these techniques don't hold good for our analysis.

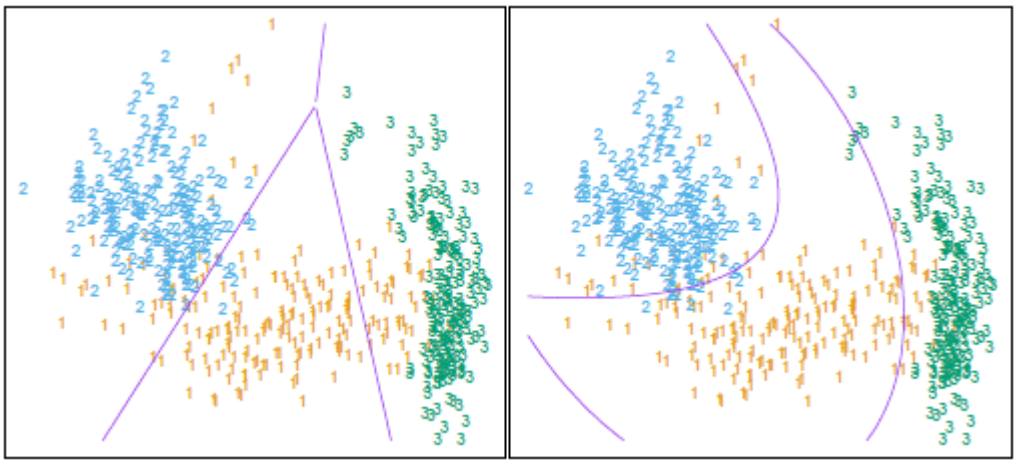

Figure 1: The figure on the left is that of an LDA classification and on the right that of a QDA classification

The k-nearest neighbour technique also requires the variables to be continuous in nature and hence has to be ruled out for this analysis[9] (Ref). Since our dependant variable, or output is binary in nature, ACS or non-ACS, decision tree approach also doesn't hold good.

Other classification techniques include the Naive Bayes. In Naive Bayes we are require to calculate

$P(X|C_i)$

$P(C_i)$

$P(X)$

Then calculate:

$$P(C_i|X) = P(X|C_i)\, P(C_i) / P(X)$$

By the maximum posteriori hypothesis the Naïve Bayes classifier predicts:

[8] Hastie, Friedman, Tibshirani; The Elements of Statistical Learning, Second Edition, Chapter 4; page 102

[9] Hastie, Friedman, Tibshirani; The Elements of Statistical Learning, Second Edition, Chapter 6; page 210

X belongs to Class Ci iff P (Ci|X) > P(Cj|X) for 1<= j <= m , j <> i

However, keeping in mind the method of sample collection, we are dealing with a case-control study and the Naive Bayes cannot be used directly. This is because in this study the individuals are not randomly selected and then tested for ACS, rather we are choosing individuals because they were admitted for ACS.

Now let us consider the logistic regression to demonstrate the modification required.

Let

$$Z = \begin{cases} 1 \; if \; an \; individual \; is \; sampled \\ 0 \; if \; an \; individual \; is \; not \; sampled \end{cases}$$

Let X be the independent variables, and then define

$$\prod_0 = P[Z = 1 | the \; individual \; is \; diseased]$$

$$\prod_1 = P[Z = 1 | the \; individual \; is \; not \; diseased]$$

Then,

$$P[D|Z = 1, X] = \frac{P[Z = 1|D, X]P[D|X]}{P[Z = 1|D, X] + P[Z = 1|D', X]P[D'|X]}$$

$$= \frac{\pi_0 \exp(\alpha + \beta' x)}{\pi_1 + \pi_0 \exp(\alpha + \beta' x)}$$

$$= \frac{\pi_0 \exp(\alpha^* + \beta' x)}{1 + \pi_0 \exp(\alpha^* + \beta' x)}$$

So we had

$$P[D|X] = \frac{\exp(\alpha + \beta' x)}{1 + \exp(\alpha + \beta' x)}$$

Now we need

$$\alpha^* = \alpha + \log\left(\frac{\pi_1}{\pi_0}\right)$$

Note that, when this ratio $\frac{\pi 1}{\pi 0}$ is 1, we have the Naïve Bayes method. However for our analysis it is not advisable to assume this ratio to be one. This ratio contains a term that is the probability of a person being sampled given he is disease and probability of a person being sampled given he is not diseased. Note that both these values are very difficult to determine by the available data. Hence we need some way to either estimate it or have sufficient logical reasoning to assume a specific value. So far in the project we have not been able to come up with this crucial answer but are working on it to find the 'right' estimate of the ratio.

Therefore we fit the appropriate GLM to obtain our logistic model. To this model we then apply StepAIC[10]to get the variables among these that have a significant effect on the outcome. This also helps us to obtain a "best possible model", where only variables that contribute significantly to the outcome are considered. We also have a full model, i.e. model consisting of all the variables and another null model where the outcome is modeled on a constant. From the three models we can now perform an Analysis of Deviance and obtain the approximate effect of a variable on the output.

Let

D_F: Deviance of the full model

D_N: Deviance of the null model

D_{Fitted}: Deviance of the fitted model, model based on selected variables

Clearly the full model should give the least deviance and the null model should give the maximum deviance out of all possible models.

Then let us define

$$D = \frac{D_N - D_{Fitted}}{D_N - D_F} X100$$

This will give us a certain percentage which increases as the contribution of the variables in the fitted model affects the outcome increases. Hence we can compare two models with the help of D.

Results

We consider the following variables for our analysis:

- SNP 1-5
- All possible combinations that can be made out of these SNP's
- Smoking
- Blood aggregation on addition of 2.5mμ ADP
- Blood aggregation on addition or 2.5mμ ADP
- Age
- Gender
- Systole
- Diastole

[10] http://isites.harvard.edu/fs/docs/icb.topic750732.files/model_selection_revised.pdf

We first perform some basic elementary data analysis. From the plots of the data versus their frequencies and by performing the student t-test we get the hunch that SNP4 may have a significant role in the prediction rule. More importantly, individuals with SNP4 as CT were mostly ACS patients.

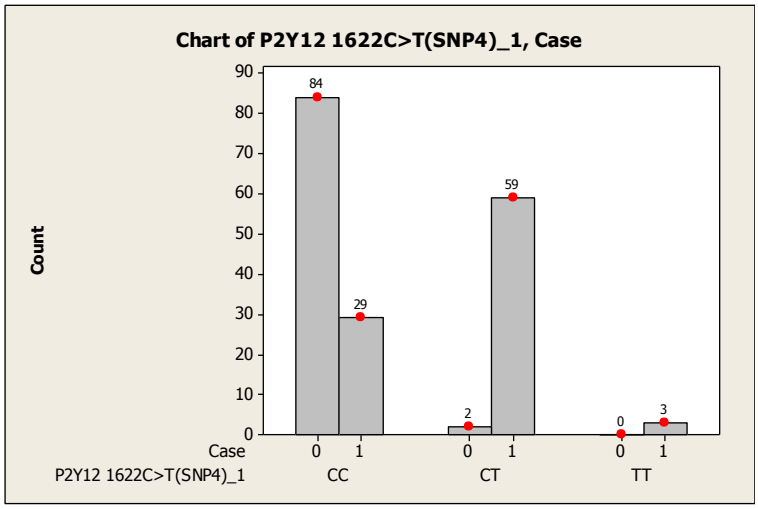

Figure 2: Histogram Plot of SNP4 Case and Control

Also ACS patients seemed to smoke more than healthy individuals

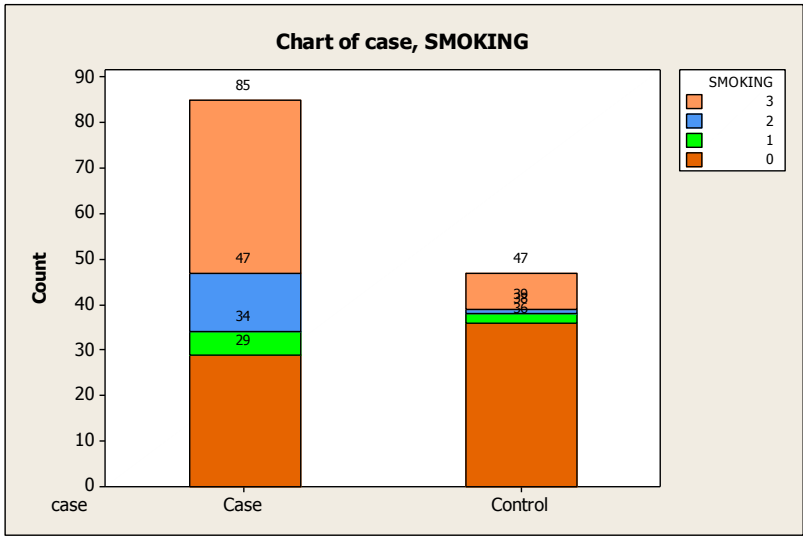

Figure 3: Chart showing number of smokers in the case and control group

A t-test also confirmed that the ACS patients smoked more than the non-ACS individuals at 95% confidence interval. As per the exploratory data analysis, more men were affected than not and more women were healthy than not.

There did not seem to be any significant difference between the systole and the diastole of the two groups. Also, the blood aggregation was more for healthy individuals than the patients. These two results are contradictory to general medical belief and the reason for this could be the fact that the individuals of the case group were already administered some drug because of their heart attack before we could collect their blood samples at the hospital. We need to look into this ambiguity with a little more care in terms of not only statistical analysis but also biological reasoning.

We then considered looking for significant variables in 3 ways:

a) Combination of both SNPs and phenotype variables
b) Combinations of only phenotype variables
c) Combination of only genotype variables.

Since our data set is small, we implemented bootstrapping[11] (give some ref on Bootstrap) in all of our analyses.

We first tried the Naïve Bayes classifier for these three types of models and they gave an accuracy ranging from 70-80%, which can be considered decent. However, as discussed earlier, the Naïve Bayes classifier qualifies only as a special case for our case-control study.

We then tried to filter out the more important variables. We fit the GLM for logistic regression and of all the SNP combinations. We then perform a StepAIC and only 1 out of possible 31 combinations of 5 SNPs, namely combination SNP2 & SNP4 of the kind CCCT or CCTT or TCCC seemed to be most important.

We then tried to fit a GLM on all variables under consideration. We then performed a StepAIC and significant genotype SNP combinations and the significant phenotype variables are:

- Aggregation on addition of ADP +Nano particles
- Systole
- Smoking
- Gender
- Diastole
- Age
- Combination of SNP2
- SNP4

FITTED GLM:

$$P(ACS = 1|Z = 1, X) = \frac{\pi_0 \exp(\alpha^* + \beta' x)}{\pi_1 + \pi_0 \exp(\alpha^* + \beta' x)}$$

where α*+β' x=-52.6 - 12.4{ADP+NP} + 0.6{ Diastole} - 0.6{ I(Smoking =1)}+25.8{ I(Smoking=2)}
-1.8{I(Smoking=3) }+ 0.4{ Age}+ 16.3{ I(Gender=M)} + 21.0 {I(SNP2==CC,SNP4=CT)}
+27.5{I(SNP2=CC,SNP4=TT) }+ 8.7{ I(SNP2=TC,SNP4=CC)}

The accuracy of prediction on cross validation is 78%, which too can be considered good.

We then performed the analysis of deviance and obtained the following table for our best fitted model:

[11] Duda,Hart,Stork; Pattern Classification;2nd edition; Chapter 9; pg 24

ANALYSIS OF DEVIANCE TABLE:

	Df	Deviance	Resid. Df	Resid. Dev
NULL			80	111.684
snp24	6	48.717	74	62.968
ADP.NP	1	11.152	73	51.816
Systole	1	0.167	72	51.649
Diastole	1	1.949	71	49.701
SMOKING	1	5.965	70	43.735
Age	1	4.254	69	39.482
Sex	1	6.518	68	32.964

By comparing the values of D obtained from this table, we get the approximate ranking as:

1. Combination of SNP 2 and 4
2. Aggregation of blood on addition of ADP and Nano particles
3. Gender
4. Smoking habit
5. Age
6. Diastole
7. Systole

Hence these are the variables among all the variables that help shed light on the vulnerability of a person towards ACS.

Limitations

The data set, as mentioned above, is indeed a very small one. Hence we require more data, especially from different regions, as that will help us to determine if the obtained results hold good for individuals across different regions or only for the community considered. More importantly a larger data set will provide us with more stable results.

As also mentioned before, due to the technical faults we have many cells of our data missing. Hence we have some individuals with incomplete or missing data. One way to tackle that is to avoid taking into account all details of any individual with even a single cell missing. However, this will further reduce the size of our already small data set.

Future Scope and Next Few Steps

This is an ongoing project and a lot of the work is yet to be done.

We have already considered classification techniques, some clustering techniques such as the k-modes, which is an analogy of k-means for categorical data, should give good results. The issue of the sample size can be brought to the advantage of the project by considering some missing data analysis techniques and hence eventually tackle the problem of the small datasets. We are also looking for similar dataset to validate our study results and establish new.

Acknowledgement

India-SA Bilateral project, CSIR South Africa, Indian Statistical Institute, University of Calcutta

All analysis and graphs done using R, Minitab® and MS Excel®