Using Logistic Regression with Single-nucleotide Polymorphism Data to Determine Significant SNPs Associated with Coronary Artery Disease

CEC- K. Aleem, C. Chia, L. Lau, S. Mcguire, Q. Tupker
November 2018

***Abstract*— We attempt to predict the single-nucleotide polymorphisms (SNPs) most likely to be responsible for Coronary Artery Disease (CAD), making attempts to control for sex, age, high density lipoproteins. We accomplish this by passing individual SNPs into a linear model to obtain a p-value, and determine the most significant SNPs**

I. INTRODUCTION

The motivation for using machine learning models in this context, is to try and correlate clinical features and specify Single-nucleotide polymorphisms, owing to the complex associations between triglyceride, high density lipoprotein & low density lipoprotein concentration. The aim is to identify potential areas of research and development for geneticists and other health care professionals, and potentially offer preventative measures towards individuals that may be at greater risk of coronary disease.

II. THEORETICAL BACKGROUND

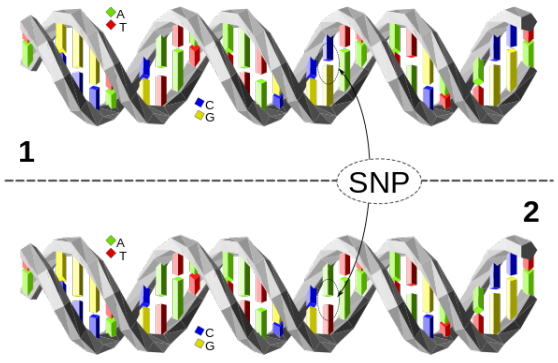

Fig. 1. Illustration of a single-nucleotide polymorphism (SNP) between two sequences of DNA. Image courtesy to David Eccles under the C.C. 4.0.

A SNP is a variation of a nucleotide at a single base-pair (e.g. A-G) location on DNA. Differences in SNPs are conjectured to have significance in determining the likelihood of having certain genetic factors, such as the likelihood of carrying a disease, or being the carrier of certain characteristics or traits. SNPs have a wide applications in genomic assocation studies and determining correlations between specific phenotypes and genotypes.

From literature, it has been often concluded that plasma triglycerides (TG) are the univariate [1] cause of coronary artery disease (CAD). However, contrary results have also illustrated that TG levels are not independent to High Density Lipoproteins (HDL) and Low Density Lipoproteins (LDL) concentrations. This is a sensible conclusion as HDL and LDL are responsible for the transportation of hydrophobic lipid molecules- an example is TGs- in water, for example blood. It has further been discovered that the increase in relative risk of coronary heart disease associated with a concentration of small, dense LDL-III in excess of 100mg/ 100ml is retained after correction for plasma triglyceride. (**B. A. Griffin et al. 1994**[3]) which suggests that TGs are not the sole factor responsible for CADs.

This intrinsic connection between TGs, LDLs and HDLs shows that there is a complex relationship between these variables and predicting CAD in a patient. Thus multivariate clinical trials that generally produce conclusions of no effect of TGs on CAD are not conclusive and TGs cannot be removed from the analysis. Particularly interesting is the effect of low concentration of HDL on CAD, suggesting that high HDL concentration in the blood may act as prevention of CAD.

The clinical feature-space consisting of age, sex etc. cannot be ignored; according to **Freedman et al. (2004)**[2]lipoproteins sizes are strongly correlated with lipid and lipoprotein concentrations. The paper

also found that the difference in sizes of LDL and HDL between men and women persisted after adjusting for lipoprotein concentrations, which were seen to vary between the sexes. However the paper was clear that these sex differences are consistent with differences in CAD risk, yet they were unable to account for the differences between the lipoprotein subclasses would account for this difference. Therefore this ambiguity requires age and sex to be considered to ensure no bias or misleading bias are induced.

III. Reference Methodology

A. Data Preprocessing

Reed et al. [5] proposes a method for GWAS. Data preprocessing is first conducted to remove features with low variance, being SNPs that have more than 5% missing data and those that have the 2nd most frequent allele, the Minor Allele Frequency (MAF), the occuring less than 1%. Then, sample level filtering is conducted to remove rows, subjects for which more than 5% data is missing and to account for subjects who have have heterozygosity, both expressions of an alelle. Similar samples are then removed using identity by descent (IBD), which accounts for subjects with an identical nucleotide segment from a common ancestor, and Linkage Disequilibrium (LD) pruning to identify non-random association between alleles, where the conjoined frequency of the of alleles is higher than what they would be independently, and setting a threshold for the Hardy-Weinberg equilibrium, which reveals genotyping errors or population substructures.
New data is then generated. Population substructure, being a presence of genetic diversity in seemingly homogeneous, necessitating the use of Principal Component analysis, usually an arbitrary 10 Principal Components are selected as confounders. Missing data is then imputed, by obtaining a best estimate for the genotype based on posterior odds of each genotype at a giving location. The imputed data is then filtered to eliminate high uncertainty (using R^2)low MAF, and failed imputations.
Genome Wide Association Analysis (GWAS) can then be conducted. A typical method is to regress each SNP seperately on a given trait, adjusted for patient-level clinical, demographic, and environmental factors. A Bonferonni correction is used, to reduce the significance threshold to 510^{-8} (experimentally obtained number) and reduce Type-I errors. Then,

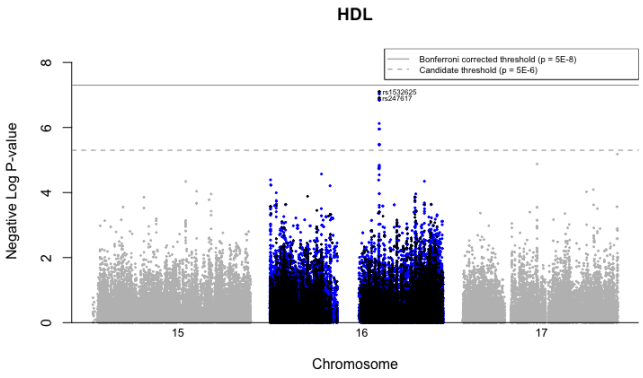

Fig. 2. Previous linear regression model for HDL

visualization plots such as a Q-Q plot or Manhattan plot can be used to determine insights from the data.

IV. METHOD

A. Problems We Encountered

1) ***Importing Data****:* The raw data was provided by Harrison Zhu and Thomas Wong and was sourced from PennCATH study [6]of coronary artery disease (CAD) where individuals were surveyed between July 1988 and March 2003. The raw data contained 22 csv files each representing one chromosome, together totalling to 861,473 SNPs, saved as columns. These took values 0, 1, 2 and 3. 0 meant no data was recorded, 1 being homozygous recessive, 2 being heterozygous recessive and 3 being homozygous dominant. Anonymised clinical data from 1401 patients was also provided.

The initial idea was to remove unnecessary columns such as "Fam ID" and to perform regression on the whole data set, which is a standard process in GWAS. Early on, there were problems with loading the whole data set into python using the pandas library due to the memory limitations of our machines. Therefore other methods were explored to load the whole data set, including writing scripts in C, and C++ to read the data line by line. Finally a script in R was written with multi-core threading and was ran on a remote machine with better hardware via ssh. However there were troubles storing the p-values due to the multi-threading process. The method of using the whole data set was abandoned and we decided to focus on a subset of the data, particularly chromosome 16, this was chosen due to chapter 1 of *"The Heart in Rheumatic, Autoimmune and Inflammatory Diseases"* (**Nussinovitch and**

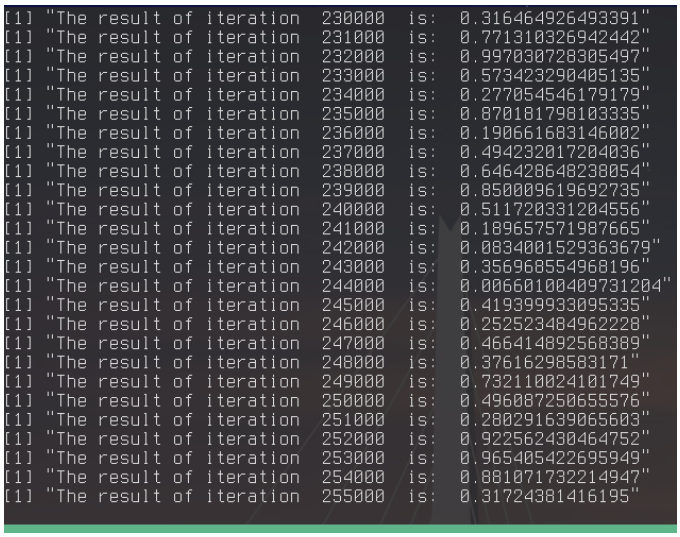

Fig. 3. Terminal output for the R script, calculating the p-values per SNP for the whole data set. However due to multi core threading not able to save to the same value, the data was not saved to an array.

Livneh, 2017[4]).

However,when importing the data, it was noted that there was change in the ordering of the data, meaning chromosomes 1 and 2 were investigated.

B. Implementing Logistic Regression for Predicting CAD

A logistic regression model was implemented using the data of chromosome 16. This was to test if using the extra SNP data has any significance to modelling CAD for patients. To do this a likelihood ratio test was performed on a logistic regression using our model with SNPs and without SNPs, p-values for individual SNPs were then calculated.

It should be noted that the p-values throughout the 22 chromosomes were calculated but due to time limitations, weren't included in the initial draft of this paper.

Our model:

$$P(Y_i = y) = \frac{e^{\zeta_i y}}{1+e^{\zeta_i}}$$

$$\zeta_i = \alpha_i + \beta_1 x_{SNP_{k,i}} + \sum_{j=1}^{N} PC_j \beta_{2,j} + \beta_3 x_{sex_i} + \beta_4 x_{LDL,i} + \beta_5 x_{HDL,i} + \beta_6 x_{TG,i} + \beta_7 x_{age,i}$$

The normal model is that without SNPs. The number of principle components N is determined by the data set used. These are the components of the data when projected on to the N most significant eigenvectors of the vector space. This was determined by the largest corresponding eigenvalues.

V. RESULTS

A. Figures and Tables

Figures shown on the next page

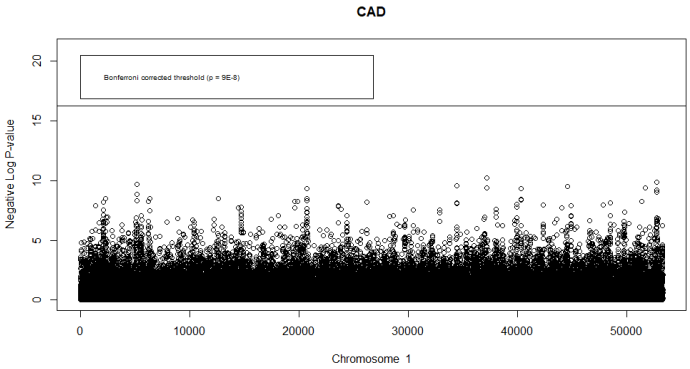

Fig. 4. p-value distribution for Chromosome 1, note how none of the -log(p) values are above the threshold, denoted by the line.

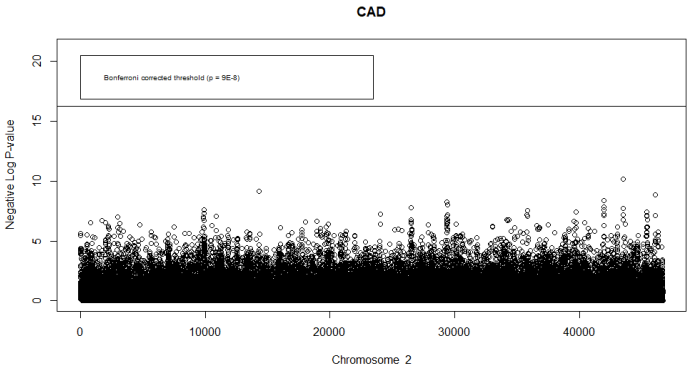

Fig. 5. p-value distribution for Chromosome 2, note how none of the -log(p) values are above the threshold, denoted by the line.

VI. DISCUSSION

From literature, it can be seen that more clinical trials have to be performed to help us understand more about coronary artery disease. How specific SNPs in the genome affects the concentration of TGs, HDLs and LDLs have to be fully understood before their specific roles in affecting CAD can be deduced. The motivation for this paper is to help narrow the range of SNPs that seem to have an effect on CAD.

To improve on our model, one could implement one-hot encoding which would produce a new β value per SNP. Our original model does not take into account that the data is categorical. Therefore this may explain the low likelihood of our original model, thus calculating lower p values from logistic regression.

This would be the slightly improved model.

$$\zeta_i = \alpha_i + \beta_{1,SNP_{k,i}} + \sum_{j=1}^{N} PC_j \beta_{2,j} + \beta_{3,sex_i} + \beta_4 x_{LDL,i} + \beta_5 x_{HDL,i} + \beta_6 x_{TG,i} + \beta_7 x_{age,i}$$

The number of principle components will have to be altered, here number used was 9. PCA is used to remove multicollinearity from a set of possibly correlated observations. Since the number of principle components was much smaller than the dimension of component space, leading to under constraining.

To further this study, the significant SNPs (the SNPs with p values above the calculated threshold of roughly **5e-7**, negative-log value of **14.51** to 4s.f.) could be taken and fed into a new model with non-linearity. The Bonferroni correction is applied to get this p-threshold, which takes the total number of tests into account. Non-linearity would be interesting to see the effect of taking a non linear relationship, such as investigating if the elderly might suffer from CAD more greatly than the linear model would predict.

It would also be interesting to see how several SNPs together affect CAD, instead of individual SNPs. A parallel

VII. CONCLUSIONS

In conclusion, none of the SNPs were found to be significant with respect to predicting Coronary Arterial Disease, although there are possible improvements to the model which could change this result.

ACKNOWLEDGMENT

Thanks to Thomas and Harrison for their untiring helpfulness. We would like to thank Luis Leal Ayala (lgl15@imperial.ac.uk). We would also like to thank the whole AI Hack team for putting on a great and well-run event.

REFERENCES

[1] MELISSA A. AUSTIN. PLASMA TRIGLYCERIDE AS A RISK FACTOR FOR CORONARY HEART DISEASE. *American Journal of Epidemiology*, 129(2):249–259, 2 1989.

[2] D. S. Freedman, James D Otvos, Elias J Jeyarajah, Irina Shalaurova, L Adrienne Cupples, Helen Parise, Ralph B D'Agostino, Peter W F Wilson, and Ernst J Schaefer. Sex and Age Differences in Lipoprotein Subclasses Measured by Nuclear Magnetic Resonance Spectroscopy: The Framingham Study. *Clinical Chemistry*, 50(7):1189–1200, 4 2004.

[3] Bruce A. Griffin, Dilys J. Freeman, Graeme W. Tait, Jim Thomson, Muriel J. Caslake, Christopher J. Packard, and James Shepherd. Role of plasma triglyceride in the regulation of plasma low density lipoprotein (LDL) subfractions: relative contribution of small, dense LDL to coronary heart disease risk. *Atherosclerosis*, 106(2):241–253, 4 1994.

[4] Udi Nussinovitch. *The heart in rheumatic, autoimmune and inflammatory diseases : pathophysiology, clinical aspects and therapeutic approaches*.

[5] Eric Reed, Sara Nunez, David Kulp, Jing Qian, Muredach P Reilly, and Andrea S Foulkes. A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in medicine*, 34(28):3769–92, 12 2015.

[6] Muredach P Reilly, Mingyao Li, Jing He, Jane F Ferguson, Ioannis M Stylianou, Nehal N Mehta, Mary Susan Burnett, Joseph M Devaney, Christopher W Knouff, John R Thompson, Benjamin D Horne, Alexandre FR Stewart, Themistocles L Assimes, Philipp S Wild, Hooman Allayee, Patrick Linsel Nitschke, Riyaz S Patel, Nicola Martinelli, Domenico Girelli, Arshed A Quyyumi, Jeffrey L Anderson, Jeanette Erdmann, Alistair S Hall, Heribert Schunkert, Thomas Quertermous, Stefan Blankenberg, Stanley L Hazen, Robert Roberts, Sekar Kathiresan, Nilesh J Samani, Stephen E Epstein, Daniel J Rader, Stephen E Epstein, and Daniel J Rader. Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies. *The Lancet*, 377(9763):383–392, 1 2011.