Statistical Machine Learning in Longitudinal Studies

Student: Kuijun Wang; Faculty Advisor: Colin O. Wu, PhD

Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University, Washington, DC

Introduction

In the United States and most other developed countries, atherosclerosis is the leading cause of disease and death [4,5]. Coronary artery calcium (CAC) is a highly specific feature of coronary atherosclerosis [1]. The discovery of potential risk factors affecting coronary atherosclerosis may prevent the development of cardiovascular diseases (CVD).

Using data from the CARDIA (Coronary Artery Risk Development in Young Adults) Study, this study aims to find out the important risk factors for cardiovascular outcomes by statistical machine learning methods.

Methods

- **Participants** The CARDIA study is a longitudinal cohort study that enrolled 5,115 healthy Black and White men and women who were aged 18 to 30 years from 1985 to 1986 (baseline year 0: Y0) in 4 U.S. urban areas (Birmingham, Alabama; Oakland, California; Chicago, Illinois; and Minneapolis, Minnesota) [2,3]. Participant samples were approximately balanced in subgroups of by race, sex, education (less or more than high school) status, and age at Y0 for each site. CARDIA participants have taken 9 in-person examinations at Y0, Y2, Y5, Y7, Y10, Y15, Y20, Y25, and Y30. Only initial CARDIA participants were allowed to participate in the follow-up examination.
- **Outcome and Covariates** The primary outcome for this study was the CAC score. We created binary variables for CAC score=0 and >0 (at 0 cutoff). The detailed information about the outcome variable and the covariates are listed in the table below.

Table 1: Variables Descriptions.

Variable	Description	Type
Male	0=female, 1=male	Binary
Race	0=white, 1=black	Binary
Education	Years of Education	Continuous
NeverMeetPA	Never meet physical activity recommendations	Binary
Creatinine	Creatinine, mh/dl	Continuous
CHOL	Cholesterol, mg/dl	Continuous
Patotalscore	Physical activity total score	Continuous
EverSmoked	Ever smoked	Binary
dislipedima	Dyslipidemia	Binary
HTN	Hypertension	Binary
BMI	Body mass index	Continuous
GLUCOSE	Fasting glucose, mg/dl	Continuous
SBP	Systolic blood pressure	Continuous
DBP	Diastolic blood pressure	Continuous
EGFR	eGFR, ml/min/1.73m^2	Continuous
NTRIG	Triglyceride, mg/dl	Continuous
HDL	HDL cholesterol, mg/dl	Continuous
LDL	LDL cholesterol, mg/dl	Continuous
DM	Diabetes	Binary

- **Statistical Analysis** Figure 1 shows the statistical analysis procedure followed in this study. The original data (n=5115, p=290) include all the covariates and the outcomes during the 9 exams. CAC scores at Y15, Y20, Y25 and covariates at Y5 and Y15 were extracted for the further procedure. The data from the patients that already have CVD events or death before the exam data were removed.

Methods

CART algorithm (Classification and Regression Trees)

We first applied the CART algorithm [7] on 4 sets of covariates and outcomes. Here we use CART with our covariates and outcomes to show the tree results for most important covariates.

Random forest (RF)

We applied the Random Forest (RF) algorithm [8] on Y15 covariates for Y15 CAC>0, and Y15 and Y5 covariates for Y15 CAC>0 to find out the important predictors. The importance of the variable was sorted by Mean Decrease Accuracy in descending order.

Logistic regression

We choose top 10 important predictors to include in the logistic regression model to see the covariate effects between predictors and CAC score. Two variable selection approaches were used in our analysis: Lasso-regularized logistic regression model [9], Forward and Backward Stepwise logistic regression model.

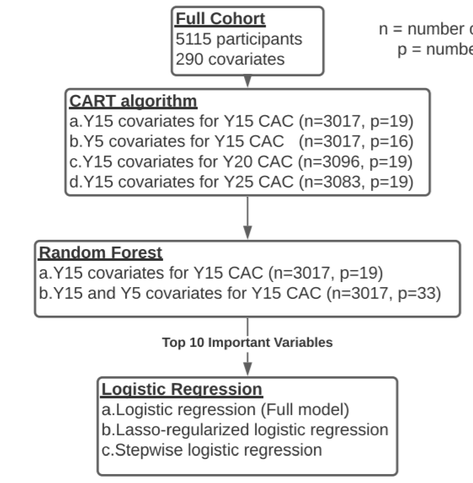

Figure 1: Flowchart Describing the Statistical Procedure of the Study.

Data analysis was performed using R software, using publically available libraries for CART algorithm [7], random forest [8], logistic regression, and LASSO methods [9].

Results

Figure 2 shows the tree results by CART algorithm. For each set of covariates and outcomes, Gender is the most important predictor in every set of covariates and outcomes. The proportion in Male groups is greater than it in Female groups (Male vs Female is 0.15 vs 0.053 in Y15 CAC; 0.4 vs 0.17 in Y25 CAC). Since there is a lack of information in the female group, CART algorithm was applied in male and female groups separately. The gender-specific trees show that LDL and Hypertension are important in the male group, and the smoking status and the Blood Pressure played an important role in the female group.

Figure 3 shows the ranked variables by the variable importance from the random forest method in descending order.

Result

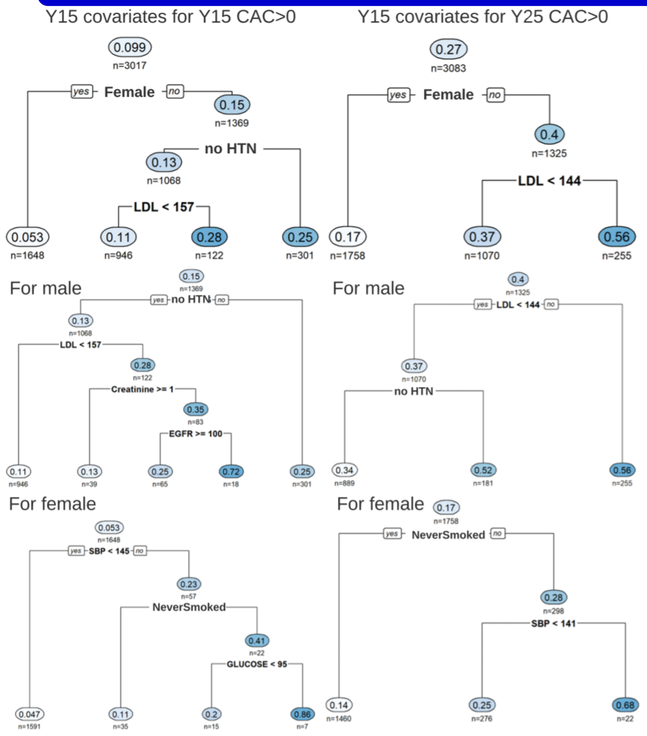

Figure 2: Results From Classification and Regression Trees.

Table 2: Logistic Regression Results Using Y15 covariates.

	Full model		Lasso		Backwards		Forwards	
	Coeff	p-value	Coeff	p-value	Coeff	p-value	Coeff	p-value
Intercept	-6.8182	<2e-16	-6.7820	<2e-16	-6.7400	<2e-16	-6.7400	<2e-16
Male	0.8139	2.97e-07	0.9069	8.78e-11	0.8070	5.05e-08	0.8070	5.05e-08
CHOL	0.1712	0.4507			0.0104	3.73e-09	0.0104	3.73e-09
LDL	-0.1618	0.4760	0.0093	1.57e-06				
HDL	-0.1736	0.4448			-0.0161	0.0020	-0.0161	0.0020
Creatinine	0.3288	0.0156			0.3291	0.0160	0.3291	0.0160
DBP	-0.0063	0.4417						
NTRIG	-0.0312	0.4926	0.0034	0.0004				
SBP	0.0262	4.13e-05	0.0224	4.68e-08	0.0222	4.68e-08	0.0222	4.68e-08
BMI	-0.0054	0.6616						
Dislipedima	0.0813	0.6551						

Table 3: Logistic Regression Results Using Y15 and Y5 covariates.

	Full model		Lasso		Backwards		Forwards	
	Coeff	p-value	Coeff	p-value	Coeff	p-value	Coeff	p-value
Intercept	-7.4064	<2e-16	-6.7495	<2e-16	-7.2523	<2e-16	-7.3116	<2e-16
Male	0.8215	1.16e-06	0.8460	7.44e-08	0.8087	3.45e-07	0.8035	4.34e-07
CHOL15	0.0094	0.1446			0.0043	0.0888	0.0045	0.0768
LDL15	-0.0059	0.4053						
CHOL5	0.0089	0.2473	0.0078	0.0022	0.0124	0.0717		
SBP5	0.0259	0.0023	0.0335	3.86e-08	0.0326	1.06e-07	0.0323	1.61e-07
LDL5	-0.0013	0.8741	0.0108	1.73e-07	-0.0053	0.4721		
HDL15	0.0002	0.9837						
HDL5	-0.0271	0.0036	-0.0170	0.0034	-0.0243	1.74e-05	-0.0276	0.0001
BMI15	-0.0020	0.8721						
DBP5	0.0108	0.2550						

Table 2 and Table3 shows the logistic regression results from 2 variable selection approach. Variables selected by backwards and forwards are similar. Using the Y15 covariates, the gender, LDL or HDL, Blood Pressure (SBP) are all significant in all approaches. Using the Y15 and Y5 covariates, the gender, Y5 blood pressure, Y5 HDL are all significant in all approaches. Cholesterol or LDL is significantly using LASSO or Backwards.

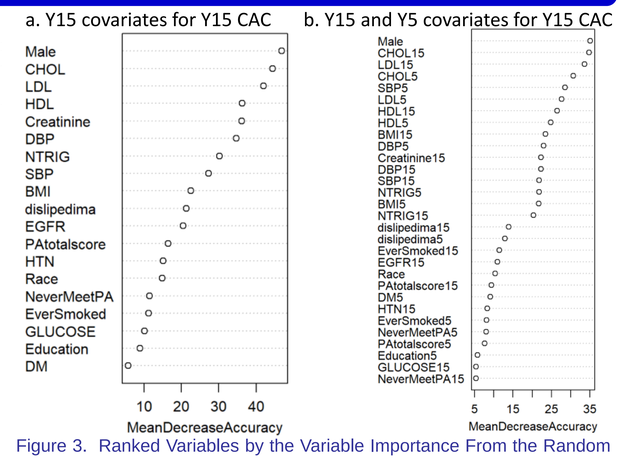

Figure 3. Ranked Variables by the Variable Importance From the Random Forests Method.

Conclusions

Using the CAC scores at Y15, Y20, Y25 and covariates at Y5 and Y15, we show that male have more risk of getting the CAC score >0, and the importance variables are different in male and female group. Male's CAC score is influenced by cholesterol (CHOL, HDL, or LDL), and female's CAC score will have more probability to be greater than zero if they smoked or have a high blood pressure. The cholesterol indicators and the blood pressure also appeared in top-10 important variables from the random forest method.

Our study shows that cholesterol indicators are important risk factors that influence the CAC scores. Further studies may be designed to confirm the findings and explore the impacts of risk factors for different subgroups.

ACKNOWLEDGMENTS The data used in this study was derived from the CARDIA study. Available data can be found at: https://biolincc.nhlbi.nih.gov/studies/cardia/. The author thank the investigators, staff, and participants of the CARDIA study for their valuable contributions.

REFERENCES

[1] Greenland, P., Blaha, M. J., Budoff, M. J., Erbel, R., & Watson, K. E. (2018). Coronary Calcium Score and Cardiovascular Risk. Journal of the American College of Cardiology, 72(4), 434–447. [2] Cutter, G. R., Burke, G. L., Dyer, A. R., Friedman, G. D., Hilner, J. E., Hughes, G. H., Hulley, S. B., Jacobs, D. R., Jr, Liu, K., & Manolio, T. A. (1991). Cardiovascular risk factors in young adults. The CARDIA baseline monograph. Controlled clinical trials, 12(1 Suppl), 1S–77S. https://doi.org/10.1016/0197-2456(91)90002-4 [3] Friedman, G. D., Cutter, G. R., Donahue, R. P., Hughes, G. H., Hulley, S. B., Jacobs, D. R., Jr, Liu, K., & Savage, P. J. (1988). CARDIA: study design, recruitment, and some characteristics of the examined subjects. Journal of clinical epidemiology, 41(11), 1105–1116. https://doi.org/10.1016/0895-4356(88)90080-7 [4] Cardiovascular diseases (CVDs). Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed November 29, 2020. [5] Atherosclerosis. Available at: https://www.merckmanuals.com/home/heart-and-blood-vessel-disorders/atherosclerosis/Atherosclerosis. Accessed November 29, 2020. [6] Cardiac Calcium Scoring (Heart Scan). Available at: https://www.umms.org/ummc/health-services/imaging/diagnostic/cardiac-calcium-scoring. Accessed November 29, 2020. [7] Breiman, L., Friedman, J., Olshen, R., & Stone, C.J. (1983). Classification and Regression Trees. [8] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). [9] Tibshirani, R. (1996), Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58: 267-288.

Poster produced by Faculty & Curriculum Support (FACS), Georgetown University School of Medicine