Gene Cell Tissue. 2015 January; 2(1): e25572. DOI: 10.17795/gct-25572

Published online 2015 January 21. **Research Article**

Assessment of SNP Interactions Affecting Total Cholesterol Over Time Using Logic Mixed Model: TLGS Study

Yadollah Mehrabi [1]; Parvin Sarbakhsh [2,*]; Jeanine J. Houwing-Duistermaat [3]; Farid zayeri [4]; Maryam Sadat Daneshpour [5]

[1]Department of Epidemiology, School of Public Health, Shahid Beheshti University of Medical Sciences, Tehran, IR Iran
[2]Department of Public Health, Statistics and Epidemiology, Tabriz University of Medical Sciences, Tabriz, IR Iran
[3]Department of Medical Statistics and Bioinformatics, Leiden University Medical Centre, Leiden, Netherlands
[4]Proteomics Research Center, Shahid Beheshti University of Medical Sciences, Tehran, IR Iran
[5]Cellular and Molecular Endocrine Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, IR Iran

*Corresponding author: Parvin Sarbakhsh, Department of Public Health, Statistics and Epidemiology, Tabriz University of Medical Sciences, Tabriz, IR Iran. Tel: +98-9141068513, Fax: +98-4113340634, E-mail: p.sarbakhsh@gmail.com

Received: November 30, 2014; **Revised:** December 31, 2014; **Accepted:** January 3, 2015

Background: Serum total cholesterol is an established risk factor for coronary heart diseases. In addition to environmental factors and main effects of polymorphisms, genetic interactions can influence cholesterol level. Furthermore, polymorphisms effects can be time-dependent. So, they could be valuable to study of genetic interactions over time.
Objectives: In this study we proposed logic mixed model to assess the association of Single-nucleotide polymorphism and other risk factors with longitudinal quantitative cholesterol level with respect to their interactions.
Materials and Method: Data of 329 subjects with age ≥ 20 years that participated in Tehran Lipid and Glucose Study with complete data for three phases of study was analyzed using proposed model.
Results: The results showed that the cholesterol level of male or subjects with GG genotype for ApoAIV and normal waist circumstance and normal blood pressure was 19.8 mg/dL less than other subjects without this combination (β = -19.8, CI 95%: -13.9, -25.69). Also, having "high triglyceride or allele e2 for ApoE" was associated with an increased effect on cholesterol (β = 16.1, CI 95%: 11.64, 20.55). The level of the cholesterol in phase one of study was 21.4 mg/dL (CI 95%: 18.35, 24.44) more than other phases. The variance component of the random effect was statistically differing with zero.
Conclusions: In this study, we extended logic regression to the longitudinal quantitative response and applied it to the TLGS data. We identified some interactions among SNPs and other covariates related to cholesterol level using logic mixed model.

*Keywords:*Cholesterol; Coronary Disease; Polymorphism, Single Nucleotide; Regression Analysis

1. Background

Serum total cholesterol is an established risk factor for coronary heart diseases (1, 2). Many factors can affect the cholesterol levels in blood such as diet, weight, exercise, age and gender. Also, the role of genetics in affecting the level of total cholesterol is significant and genes partly determine the amount of cholesterol and high blood cholesterol can run in families. Recent studies have identified major loci associated with total cholesterol (3). However, most these associated SNPs (Single-nucleotide polymorphism) show a small effect size (4), and collectively explain only 25–30% of heritable variation for total cholesterol level (5).

Gene–gene interactions might have a significant role in complex traits and are one of the many possible factors contributing to the missing heritability (6). Because of gene-gene interaction effects on missing heritability, we would like to find SNPs that interact in their effects on the total cholesterol level.

To identify these SNPs interactions, we can use logic regression to search for sets of SNPs that are jointly associated with the phenotype in the form of Boolean combinations of them (7). Logic regression is an adaptive statistical method that is used to model an outcome with combinations of multiple binary predictors, such as indicators of SNP genotypes (i.e. dominant or recessive gene (7). Logic regression has been applied successfully to a number of SNP data analyses with selected candidate genes in case control, matched case control or cohort study (8-10).

On the other hand, the development of genotype association tests for repeatedly measured phenotypes, such as the longitudinal total cholesterol level, has received little attention in the literature. Especially, studies that have investigated interaction effects of SNPs on phenotype over time are rare. The present paper was initially motivated by the SNP dataset with longitudinal total cholesterol level as response variable and potential genetic

interactions affecting total cholesterol level. We know interactions may contribute to missing heritability in total cholesterol (6). Furthermore, because there is some evidence of time-dependent genetic influences on total cholesterol (11), it seems that genetic interaction may be time-dependent as well.

2. Objectives

We would like to investigate interaction effects during time to find out whether there is any interaction with various effects on response in different time points.

3. Materials and Methods

In this study we propose an approach to study interaction effects on response over time. In particular, we propose "logic mixed model" method, in which we fit a linear mixed-effects model in Logic Regression framework to find the best Boolean combinations of the binary predictors such as SNPs affecting longitudinal total cholesterol level. Such Boolean combinations contain interaction effects among predictors. In this method, we use a general random-effects structure to capture the correlation between the longitudinal total cholesterol level and negative likelihood as the score function of the model and Annealing algorithm as the search algorithm of the model to find best combinations.

3.1. Logic Regression

Logic Regression is a generalized regression and classification method that enables identification of interactions by using Boolean combinations as new independent variables of the original binary variables ($X_1, \ldots, X_k$). This model was initially developed to reveal interacting SNPs in genetic association studies by searching for Boolean combinations of SNPs that reveals SNPs and interactions that are associated with the response.

Logic Regression models are of the form:

$g(E(\mathbf{Y})) = \beta_0 + \sum^{p}_{i=1}\gamma_i W_i + \sum^{t}_{j=1}\beta_j \mathbf{L}_j$

Where g is a link function for the response, are quantitative covariates and are Boolean combinations of the binary predictors $(X_1,\ldots,X_k)$ such as $(X_1 \wedge X_3) \vee (X_5 \wedge X_c^7)$

For every model type, a score function is defined that shows the "quality"' of the model under consideration. We try to find Boolean statements involving the binary predictors that enhance the prediction for the response and minimize the scoring function associated with this model type.

The number of logic expressions that can be built from a given set of binary predictors is very high, and there is no straight method to enlist all logic terms that yield different score. So, it is infeasible to do an exhaustive assessment of all different logic terms and select the best model. To solve this problem in Logic Regression method, simulated annealing as a stochastic search algorithm is used to search for Boolean combinations and estimate the β_j (7, 12).

3.2. Linear Mixed Effects Model

Let y_i be a vector of n_i observations for sample unit i , i = 1,. . .,m. y_i follows the general linear mixed model:

$Y_i = X_i\beta + Z_i b_i + e_i$

Where X_i $(n_i \times p)$ and Z_i $(n_i \times q)$ are known covariate matrices related to fixed and random effect respectively. b_i and e_i are random errors distributed as:

$e_i \sim N(0, R_i)$, $R_i = \sigma^2 I_{ni}$, $b_i \sim N(0, \sum)$

Independently for $i = 1, \ldots, m$

Linear mixed models make specific assumptions about the variation in observations attributable to variation within a subject and to variation among subjects. The within-subject variation is the deviation between individual longitudinal observations (13). There are some unobserved factors that cause this variation between subjects. For counting this variation, random effect models include random coefficients in the model. Thus, each subject has their own coefficient. Because there are many coefficients as same as the number of subjects, it is difficult to estimate random coefficients of all subjects>Instead, the random effect models estimate the one variance component for these coefficients. If this variance component significantly differs with zero, we can say that there is an unobserved variation between subjects.

3.3. Our Proposed Method: Logic Linear Mixed Model

In this paper, mentioned linear mixed model with negative likelihood of the model as the score function, was used to extend Logic Regression to analyze quantitative longitudinal response. Therefore, logic linear mixed model was defined as:

$Y_i = W_i\gamma + L_i\beta + z_i b_i + e_i$, $e_i \sim N(0, R_i)$, $R_i = \sigma^2 I_{ni}$, $b_i \sim N(0, \sum)$

Where Y_i is vector of quantitative responses of individual, W_i is quantitative covariates with fixed effect, Z_i is quantitative covariates with random effect, and L_i is matrix of Boolean expression from binary predictors with fixed effect. γ, β are vectors of unknown parameters.

Searching to find best L_i - so that the fitted model has the lowest score - was done using Annealing algorithm. Therefore, Annealing algorithm searched for Boolean combinations which according to the negative likelihood statistic had the lowest score and therefore had the best fitting in logic mixed model. The program of logic mixed model was written in FORTRAN 77 and added to "LogicReg" package (7). Modified "LogicReg" package was recompiled an installed in R (2.15.3) to analyze data.

3.4. Data Set and Variables Definition

Subjects in this longitudinal study were selected from among participants of the Tehran Lipid and Glucose Study (TLGS). TLGS is a prospective study to determine the risk factors and outcomes of non-communicable disease. The structure of this study includes some major components; In the TLGS, a random sample of 15005 people aged

3 years and over, living in district 13 of Tehran participated in the cross-sectional phase as a prevalence study of cardiovascular disease and associated risk factors (14); Other phases are prospective follow-up studies with median follow-up of 3.6 years, in which subjects were categorized into the cohort and intervention groups, the latter to be educated for implementation of lifestyle modifications. The TLGS design has been explained elsewhere (15).

A total of 329 subjects - 127 (38.6%) men and 202 (61.4%) women - who were present in phase I ,II and III of TLGS study with age ≥ 20 years and have complete data in candidate SNPs and other evaluated variables for all phases were included in the current study.

High waist circumference (WC) was defined as WC ≥ 95 cm for Iranian men and women (16). High triglyceride (TG) level was defined as TG ≥ 150 mg/dL. Subjects who had blood pressure (BP) ≥ 130/85 mmHg or used antihypertension drug and subjects with fasting blood sugar (FBS) ≥ 110 mg/dL or users of anti- diabetic drugs were considered as high BP and high FBS respectively (17). Subjects who smoke daily or occasionally were considered as smokers. Phase of study was considered as time.

The polymorphisms of ApoA1M1, ApoA1M2, ApoB, ApoAIV, ApoCIII, ABCA1, SRB1 and ApoE genes that have been shown to be associated with total cholesterol disorder (18-22) were investigated.

Each SNP was considered as a random variable, taking values 0, 1 and 2 corresponding to the nucleotide pairs. We coded each of these variables into two dummy binary variables corresponding to a dominant (if at least one variant allele exist) and a recessive (if two variant alleles exist) effect. By this approach, we generated 2p binary predictors out of p SNPs to perform interaction terms for Logic Regression (7).

With considering Z = 1 and W = age , Logic mixed model with three Boolean combination consist of seven binary predictor variables was fitted to assess the association between the candidate polymorphisms, age, sex, high TG, high WC, high BP, high FBS and smoking with cholesterol level over time in longitudinal data of the TLGS study.

4. Results

Table 1 pictures the summary of demographic characteristic and clinical and lipid profiles of these subjects in three phases of study. The highest level of the total cholesterol was seen in phase 1 of study (211.31 mg/dL). This amount has decreased by about 17 mg/dL in phase 2 and this value remained stable in phase 3.

Risk factors studied included high TG, high FBS and smoking were almost fixed during follow-up period but age and high WC had increased trend and high BP risk factor had decreased trend in subjects of study during follow-up time.

Allele frequencies are given in Table 2, which shows genotype distributions that were in Hardy–Weinberg equilibrium. The +/+ genotype of Apo A1M2 gene had the highest prevalence (91.2%) and TT genotype of Apo AIV gene had the lowest frequency (0.3%).

The results of first Boolean combination of fitted logic mixed model show that total cholesterol level of the males or subjects with normal WC and normal BP and GG genotype for ApoAIV gene, was 19.8 mg/dL less than other subjects without this combination.

Also, according to second Boolean combination of SNPs and risk factors found by Annealing Algorithm, it was seen that having high TG or e2 allele for ApoE was associated with increased total cholesterol level; so compared to subject without this combination, we would expect total cholesterol level of subjects having this combination to be 16.1 mg/dL more, on average at the same value of other combinations of the model.

According to the latest combination of the fitted model, at the same level of other combinations of the model, mean of total cholesterol level in phase I is 21.4 mg/dL higher than other phases.

Age as the quantitative adjusted variable, has significant positive effect on cholesterol level over time; so that total cholesterol is predicted to increase 0.73 mg/dL (CI 95%: 0.49, .96) when the age variable goes up by one year.

Table 1. Demographic Characteristic and Clinical and Lipid Profiles of Subjects in Phases of Study [a]

Variables	Phase 1, (n = 329)	Phase 2, (n = 329)	Phase 3, (n = 329)	P Value
Age, y [b]	41.09 ± 15.82	44.84 ± 15.71	47.46 ± 15.62	< 0.001
Gender (female) [c]	202 (61.4)	202 (61.4)	202 (61.4)	1
Total cholesterol level [b]	211.31 ± 48.80	193.62 ± 40.92	194.94 ± 40.92	< 0.001
High WC[b]	90 (27.4)	134 (40.7)	151 (45.9)	< 0.001
Hypertension[b]	111 (33.7)	97 (29.5)	82 (24.9)	0.002
High TG[b]	139 (42.2)	140 (42.6)	141 (42.9)	0.97
High FBS [b]	39 (11.9)	39 (11.9)	41 (12.5)	0.85
Smoker [b]	27 (8.2)	30 (9.1)	27 (8.2)	0.66

[a] Entries are mean ± SD for Age and number (%) for the rest categorical variables.
[b] is time dependent variable.
[c] is time independent variable.

Table 2. Genotype and Allele Frequencies (percentage) of apolipoprotein E (Apo E), Apo A1M1, Apo A1M2, Apo B, Apo AIV, Apo CIII, and SRB1 in the Study Population

Polymorphisms			
Apo E Alleles	e2	e3	e4
	34 (1.3)	258 (78.4)	37 (11.2)
Apo A1M1 Genotypes	+/+	+/-	-/-
	233 (70.8)	90 (27.4)	6 (1.8)
Apo A1M2 Genotypes	+/+	+/-	-/-
	300 (91.2)	23 (7)	6 (1.8)
Apo B Genotypes	X+X+	X+X-	X-X-
	28 (8.5)	126 (38.3)	175 (53.2)
Apo AIV Genotypes	TT	GT	GG
	1 (0.3)	56 (17)	272 (82.7)
Apo CIII Genotypes	CC	CG	GG
	232 (70.5)	87 (26.4)	10 (3)
ABC A1Genotypes	GG	GA	AA
	112 (34)	171 (52)	46 (14)
SRB1Genotypes	GG	GA	AA
	268 (81.5)	58 (17.6)	3 (0.9)

Table 3. Results of Logic Mixed Model With 3 Boolean Combinations of 7 Binary Predictor Variables to Study Interaction Effects of SNPs and Other Risk Factors on Total Cholesterol Level Over Time

Boolean Combination	**Coefficient**	**Standard Error of Coefficient**	**CI 95% for Coefficient**
(Normal WC and normal BP and ApoAIV = GG) or Gender = male	-19.8	3.01	(-25.69, -13.90)
(High TG or ApoE = e2)	16.1	2.27	(11.64, 20.55)
Phase 1compared to other phases	21.4	1.55	(18.35, 24.44)
Standard deviation of random effect		30.89	P Value < 0.001

5. Discussion

An issue that often arises in genetic association studies is the investigation for interaction effect. Moreover, genetic interactions can be time-dependent, so considering of interactions among SNPs and other environmental risk factors in modeling of longitudinal SNP data is important. In this study, we developed a mixed model approach in Logic Regression framework to test associations between gene sets and linear type of longitudinal phenotypes, which is able to test for the effect of interaction terms in candidate gene studies.

An important feature of our method is that it can capture correlations between longitudinal correlated responses and allows for inference about interactions affecting phenotypes over time.

The finding of the current study is that ApoAIV, ApoE, TG, WC, BP and sex are significantly associated with cholesterol.

Although association of main effects of these variables with total cholesterol have previously been shown in several studies (20, 23, 24), investigation for interaction effects among them on longitudinal total cholesterol level using logic mixed model approach was done for first time in this study.

Results of this study show that there is a combination of WC, BP and ApoAIV with significant effect on total cholesterol level. It means that having normal WC and normal BP and GG genotype for ApoAIV gene had a decreasing effect on total cholesterol so that total cholesterol level of subjects with this combination was 25.88 mg/dL lower than subjects without this combination (184.83 versus 210.71 mg/dL).

Also, according to the result of Logic mixed model, it seems that there is an interaction among sex and "normal WC and normal BP and ApoAIV = GG" combination. In other words, males have lower level of total cholesterol than females by itself (192.24 compared to 204.82 mg/dL) or by accompanying with "normal WC and normal BP and ApoAIV = GG" combination. Total cholesterol level in

the male with "normal WC and normal BP and ApoAIV = GG" was 189.07 mg/dL and in the other subjects was 221.54 md/dL.

It has been proposed that APO AIV may play different roles in lipids modulation. The Human apoAIV is a glycoprotein constituent of triglyceride-rich and high-density lipoproteins (HDL) and in reverse cholesterol transport, is mainly synthesized by the intestine and is secreted with chylomicrons.

Moreover, there is an interaction between Apo E gene and high TG, which having high TG or e2 allele for ApoE can increase the level of total cholesterol during time and consequently increase the risk of cardiovascular disease.

The differing risks for high blood cholesterol are the result of the enzymatic activity associated with the variants. The APOE gene makes Apo E, which is involved in the production, delivery, and utilization of cholesterol in the body.

Furthermore, time by itself has an important effect on the level of total cholesterol so that compared to other phases, mean of total cholesterol in phase I of study was higher. This might be due to the effect of interventions for developing a healthy lifestyle after phase I in TLGS. According to the result, there was no interaction between time and SNPs or other risk factors affecting total cholesterol level.

There were some limitations in this study, such as small sample size and low number of SNPs, so we could not identify more interactions between SNPs affecting level of total cholesterol.

Logic mixed model makes us able to investigate interactions between polymorphisms and other risk factors on longitudinal phenotype that we may not be able to find by using conventional methods of data analysis.

Acknowledgements

We would like to thank the staff and participants in the TLGS study for data collection. The research presented in this paper was carried out on the High Performance Computing Cluster supported by the computer science department of Institute for Research in Fundamental Sciences (IPM).

References

1. Tanabe N, Iso H, Okada K, Nakamura Y, Harada A, Ohashi Y, et al. Serum total and non-high-density lipoprotein cholesterol and the risk prediction of cardiovascular events - the JALS-ECC. *Circ J.* 2010;**74**(7):1346–56.
2. Nippon Data Research Group . Risk assessment chart for death from cardiovascular disease based on a 19-year follow-up study of a Japanese representative population. *Circ J.* 2006;**70**(10):1249–55.
3. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature.* 2010;**466**(7307):707–13.
4. Ma L, Brautbar A, Boerwinkle E, Sing CF, Clark AG, Keinan A. Knowledge-driven analysis identifies a gene-gene interaction affecting high-density lipoprotein cholesterol levels in multi-ethnic populations. *PLoS Genet.* 2012;**8**(5).
5. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet.* 2010;**11**(6):446–50.
6. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A.* 2012;**109**(4):1193–8.
7. Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. *J Comp Graph Stat.* 2003;**12**(3):475–511.
8. Li Q , Fallin MD, Louis TA, Lasseter VK, McGrath JA, Avramopoulos D, et al. Detection of SNP-SNP interactions in trios of parents with schizophrenic children. *Genet Epidemiol.* 2010;**34**(5):396–406.
9. Sarbakhsh P, Mehrabi Y, Daneshpour MS, Zayeri F, Zarkesh M. Logic regression analysis of association of gene polymorphisms with low HDL: Tehran Lipid and Glucose Study. *Gene.* 2013;**513**(2):278–81.
10. Schwender H, Ickstadt K. Identification of SNP interactions using logic regression. *Biostatistics.* 2008;**9**(1):187–98.
11. Middelberg R, Heath AC, Martin NG, Whitfield JB. Evidence of age-dependent genetic influences on plasma total cholesterol. *Eur J Cardiovasc Prev Rehabil.* 2005;**12**(4):380–6.
12. Schwender H, Ruczinski I. Logic regression and its extensions. *Adv Genet.* 2010;**72**:25–45.
13. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics.* 1982;**38**(4):963–74.
14. Azizi F, Rahmani M, Emami H, Mirmiran P, Hajipour R, Madjid M, et al. Cardiovascular risk factors in an Iranian urban population: Tehran lipid and glucose study (phase 1). *Soz Praventivmed.* 2002;**47**(6):408–26.
15. Azizi F, Ghanbarian A, Momenan AA, Hadaegh F, Mirmiran P, Hedayati M, et al. Prevention of non-communicable disease in a population in nutrition transition: Tehran Lipid and Glucose Study phase II. *Trials.* 2009;**10**:5.
16. Azizi F, Khalili D, Aghajani H, Esteghamati A, Hosseinpanah F, Delavari A, et al. Appropriate waist circumference cut-off points among Iranian adults: the first report of the Iranian National Committee of Obesity. *Arch Iran Med.* 2010;**13**(3):243–4.
17. Grundy SM, Cleeman JI, Daniels SR, Donato KA, Eckel RH, Franklin BA, et al. Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. *Circulation.* 2005;**112**(17):2735–52.
18. Daneshpour MS, Faam B, Hedayati M, Eshraghi P, Azizi F. ApoB (XbaI) polymorphism and lipid variation in Teharnian population. *Eur J Lipid Sci Tech .* 2011;**113**(4):436–40.
19. Brown CM, Rea TJ, Hamon SC, Hixson JE, Boerwinkle E, Clark AG, et al. The contribution of individual and pairwise combinations of SNPs in the APOA1 and APOC3 genes to interindividual HDL-C variability. *J Mol Med (Berl).* 2006;**84**(7):561–72.
20. Daneshpour MS, Hedayati M, Eshraghi P, Azizi F. Association of Apo E gene polymorphism with HDL level in Tehranian population. *Eur J Lipid Sci Tech .* 2010;**112**(7):810–6.
21. McCarthy JJ, Lehner T, Reeves C, Moliterno DJ, Newby LK, Rogers WJ, et al. Association of genetic variants in the HDL receptor, SR-B1, with abnormal lipids in women with coronary artery disease. *J Med Genet.* 2003;**40**(6):453–8.
22. Frikke-Schmidt R. Context-dependent and invariant associations between APOE genotype and levels of lipoproteins and risk of ischemic heart disease: a review. *Scand J Clin Lab Invest Suppl.* 2000;**233**:3–25.
23. Daneshpour MS, Zarkesh M, Hedayati M, NaminMesbah SM, Halalkhor S, Faam B, et al. The G360t Polymorphism in the APO AIV Gene and its Association with Combined HDL/LDLCholesterol Phenotype: Tehran Lipid and Glucose Study. *Int J Endocrinol Metab.* 2010;**8**(1):32–8.
24. Prospective Studies C, Lewington S, Whitlock G, Clarke R, Sherliker P, Emberson J, et al. Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths. *Lancet.* 2007;**370**(9602):1829–39.