Original Article

Identification of Genetic Polymorphism Interactions in Sporadic Alzheimer's disease Using Logic Regression

Najimeh Tarkesh Esfehani, MSc; Mahdi Rahgozar*, PhD; Akbar Biglarian, PhD;
University of Social Welfare and Rehabilitation, Tehran, Iran

Hamidreza Khorram Khorshid, PhD;
University of Social Welfare and Rehabilitation, Genetic Research Center, Tehran, Iran

Objectives: Genetic polymorphism interactions are among the important factors in affliction with complex diseases like Alzheimer's disease. The important goal of genetic association studies is to identify a combination of polymorphisms and measure their importance in increasing the risk of occurrence of such diseases. In this study, feature selection approach of logic regression was used to identify the interactions among genetic polymorphisms influential in patients affected with Alzheimer's disease.

Method and Materials: 101 Alzheimer's cases and 109 control subjects from Iranian population were recruited in a case-control study. The evaluation of genes in two groups was performed using molecular technique methods; in particular, the PCR-RFLP technique was used to evaluate the intended polymorphisms in APOE, ABCA1, CALHM, CCR2, GSK3β, SAITOHIN, TAU, TNF-α and VDR genes, and then the feature selection approach was used to detect the significance polymorphisms and interactions between them.

Results: Based on feature selection approach, the two-way interaction between the polymorphisms of SAITOHIN and APOE genes were significant on occurrence of Alzheimer's disease.

Conclusion: Logic regression approach is recommended to detect interaction in the genetic association studies.

Keywords: Logic regression, Feature selection, Interactions, Genetic Polymorphisms, Alzheimer's disease

Submitted: 12 Dec 2010
Accepted: 04 Mar 2011

Introduction

Single nucleotide polymorphism (SNP) is a minor genetic variation which can occur in DNA sequence. SNP occurs when a nucleotide is replaced by one of the other three nucleotides in nucleotide chain. On the average SNPs in human populations occurs more than %1 of the times. Individual SNP usually has small to medium effects in occurrence of diseases, particularly in complex or multi-factorial diseases. Therefore when dealing with complex diseases, the purpose of association studies is to specify the combined effects of SNPs and the interaction among them on the increase in risk of disease (1). Alzheimer's disease (AD) is one of such diseases. AD is the most common cause of dementia in middle and old age in western societies; therefore, aging increases the risk of affliction with the disease. AD is the fourth important cause of death in the United States (2), and is one of the most important factors of disability and health endangering in the world. In 2006, nearly 26.6 million people in the world were suffering from AD. With elevation of life expectancy, it is quite possibly anticipated that until 2050, more than 100 million people will suffer from AD, which shows that one per 85 people in the world will be affected with the disease (3). Due to the increasing trend in Alzheimer's disease it is crucial paying more attention to its early diagnosis and detection. Over 95% of patients suffering from Alzheimer are sporadic and late-onset type, the diagnosis of which is based on clinical and neuropsychological evaluations and is time consuming and costly. Consequently, the diagnosis of disease by a genetic marker could be a good solution for this problem, so as to be used for quick diagnosis of disease in early stages or for treatment

* All correspondences to: Mahdi Rahgozar, E-mail:<m.rahgozar@uswr.ac.ir>

aims (4). Alzheimer is a complex disease because it lacks any specific hereditary pattern and is heterogenic, since a variety of mutations and polymorphisms in several genes are responsible for the disease along with non-genetic factors. Individual SNPs have small to middle effects in the occurrence of such complex disease and it seems necessary to specify the combined effect of SNPs and the interactions between them in increasing of the risk of this disease (1).

Thus far, many genes have been investigated as risk factor for Alzheimer's disease, the most well-known of is the APOE on chromosome 19. This gene has been identified as the most important risk factor in 65% of sporadic Alzheimer cases (5). The APOE gene in human has the three allele e2, e3, e4. These alleles are differently influential in the risk of occurrence of Alzheimer's disease (6). Also, there are evidences on the relationship between Alzheimer and SNPs from genes such as *ABCA1*, *CALHM*, *CCR2*, *GSK3β*, *SAITOHIN, TAU*, *TNF-α*, and *VDR*.

Since the human genome is diploid, that means it has pairs of chromosomes, 2 bases explained each SNP. Thus, each SNP can have one of the following 3 forms:

– "Homozygous reference (wild type) genotype": both explaining bases of the SNP are the variant which is more frequent.
– "Heterozygous variant genotype": one of the bases is more frequent variant and the other is the less.
– "Homozygous variant genotype": both bases are the less frequent variant.

Thus, in an association study concerned with SNPs data, it is thus of interest to construct classification rules of the following type:

> "If SNP A is of the heterozygous variant genotype AND SNP B is of the homozygous variant genotype OR both SNP C AND D are NOT of the homozygous reference genotype, then a person has a higher risk for the disease of interest".

Classic parametrical statistical methods such as logistic regression are unable to detect such interactions and in most problems a regression model can only investigate the relationship the main effects of predictors on the response and the interaction between variables, in case considered in the model, does not go beyond two-way and, at most, three-way. A procedure developed for solving exactly these types of problems is logic regression which was introduced by Ingo Ruczinski, and attempts to identify Boolean combinations of binary variables for the prediction of case–control status in an observation (7). After the first introduction on logic regression model by Ruczinski, several models were proposed to improve the model, among which the Feature selection logic regression (logicFS) can be noted (8). Feature selection is a combination of bootstrap and logic regression that can be used for quantifying the importance of interactions for classification. In order to detect the interactions of genetic polymorphisms of the noted genes and genotypes of *APOE* gene in affliction with Alzheimer, the Feature selection approach of logic regression was used.

Materials and methods

This study was a case-control one in which the required samples for the case and control groups were received from the Genetic Research Center-university of Social Welfare and Rehabilitation Science, in which Alzheimer cases and control subjects were included if they were older than 65 years old and the informed consent was signed by them or their legal care takers. The criteria for inclusion as a case were existence of Alzheimer diagnosed by an expert psychiatrist based on DSM IV criteria and lacking any neurologic or psychiatric disorders for control group according to medical report or responsible physician statements. Subjects were excluded if they had any family history of dementia or neurologic diseases. Alzheimer and control subjects were recruited from Alzheimer's society of Iran and Geriatric centers Mehrvarzan, Kahrizak, Shayestegan, Farzanegan,Hashemi nezhad and Rheumatism Center in Tehran, Iran from 2007 to 2008. The evaluation of genes in the two groups was performed using molecular techniques; The PCR-RFLP technique was used to particularly evaluate the intended polymorphisms in *APOE*, *ABCA1*, *CALHM1*, *CCR2*, *GSK3B*, *SAITOHIN* and *TAU, TNF-α* and *VDR* genes. Afterwards, the information related to 316 people were received from the lab. From these people 106 observations had one or more missing polymorphisms and with deletion of these observations 210 observations were analyzed by logic FS and the important interactions were specified by the calculation of the two indexes of VIM_{single} and $VIM_{multiple}$. In order to find the best logic combination the algorithm Simulated Annealing was used (7). For this purpose the R statistical software version 2.13.2 was used.

Results

Present study was conducted on data obtained from 210 participants above 65 years of age including 101 afflicted with Alzheimer's disease in the case group and 109 in the control group. The primary information about the *APOE* genotypes and other polymorphisms investigated is given in table 1 and 2.

For fitting the logic regression model and using feature selection method, the input variables are changed into binary variables in the following form. Regarding the *APOE* gene, the information related to the six genotypes (e2e2, e2e3, e2e4, e3e3, e3e4, e4e4) is at hand and the binary variables of X_1 to X_6 are defined as follows:

Table1: The *APOE* genotype frequencies were compared between Alzheimer cases and control subjects

Genotype	control number (percent)	case number (percent)
e2e2	1 (0.9)	1 (1.0)
e2e3	14 (12.8)	5 (5.0)
e2e4	1 (0.9)	1 (1.0)
e3e3	81 (74.3)	78 (77.2)
e3e4	11 (10.1)	15 (14.9)
e4e4	1 (0.9)	1 (1.0)

$$X_i = \begin{cases} 1 & \text{the person have } i \text{ genotype} \\ 0 & \text{the person does not have } i \text{ genotype} \end{cases}$$

Each SNP S_i, is split into two variables as defined in below:

S_{i1}: "At least one of the bases explaining Si is the less frequent variant."

S_{i2}: "Both bases explaining Si are the less frequent variant."

These made variables are used instead of the SNPs themselves.

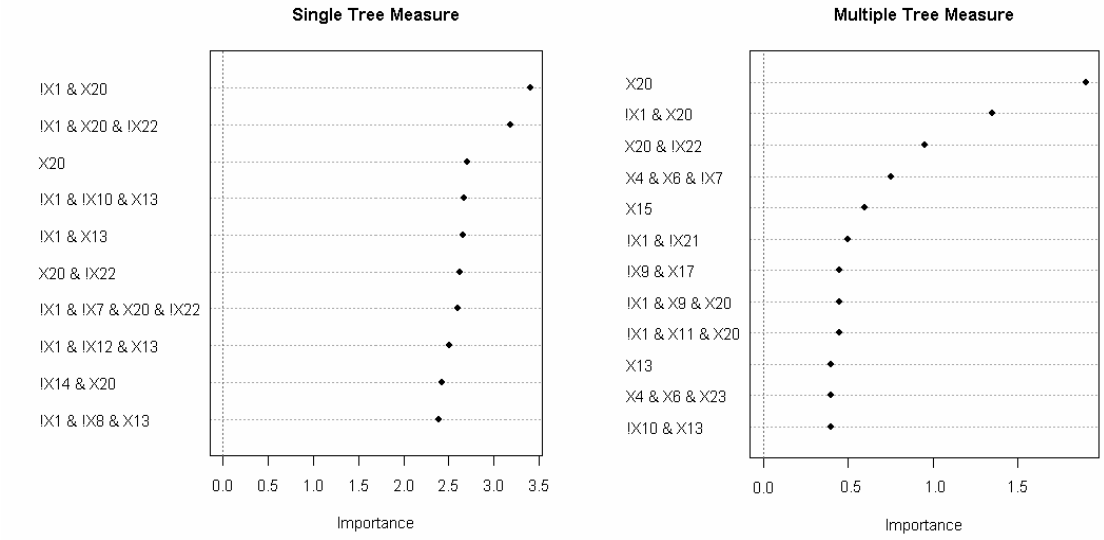

Figure1: VIMSingle (left panel) and VIMMultiple (right panel) of the interactions identified in analysis of Alzheimer data set. Since the SNP names are too long for graphical representation, they are coded.

Consequently, having six genotypes from *APOE* gene and eleven SNPs from other genes possibly related to Alzheimer, the 28 binary variables as predictors are available as input for the logic regression model. Of these variables all observations showing more than 5 missing values are removed from the analysis leading to a total of 22 variable and 210 observations.

Logic FS is applied to this data set twice using 10000 iterations in each run of simulated annealing and 200 bootstrap samples,—once with a single-tree and a maximum of 6 variables contained in this tree and the other time allowing 2 trees to grow with a maximum of 10 variables in all the 2 trees combined. In the single-tree case, this leads to the detection of 449 potentially interesting polymorphisms interactions, whereas in the

multiple-tree case, 562 SNPs and SNP interactions are identified. however, just one interaction, namely !X1&X20 or decoded e2e3&STH (HinfI(A/G))$_1$, consisting of 2 polymorphisms from the gene *APOE* and *SAITOHIN* seems to be associated with the case–control status since in it both indexes VIM_{single} and $VIM_{multiple}$ have high values. If the *STH* (A/G) is not of homozygous reference genotype and there is no e2e3 genotype in the person, there will be a little higher risk of developing Alzheimer.

STH (A/G) itself has the highest value of $VIM_{multiple}$ and third highest VIM_{single}. Therefore, the *SAITOHIN* gene may itself influence in Alzheimer and AG and GG genotypes of this gene are risk factors for the disease.

Discussion

Alzheimer's disease is one of the complex diseases which lacks specific hereditary pattern and is heterogenic and mutations and polymorphisms existent in several genes along with environmental factors, are influential in it. Since there is no definite treatment for Alzheimer at the time, the identification of risk factors leading to this disease and prevention of their occurrence is of high significance. Based on the studies conducted, one-late Alzheimer occurs under the influence of a number of genetic and environmental factors. Controlling genetic factors is impossible but it is highly possible, through the identification of genetic factors influential in Alzheimer, to identify people at risk and trying to control the influential environmental factors in the occurrence of Alzheimer such as low level of mental activities, social-psychological stress, diet, smoking and drinking, pesticides environmental factors over use of some medicines,". Also it is possible to use these genetic markers to detect diseases in elementary stages and reduce the speed and disabilities resulted from the disease by its early diagnosis of it.

One of the important and common goals in genetic association studies in such diseases is the determination of SNPs and their interactions which are related to the occurrence of the disease. The previous studies have shown the relationship between several single SNPs and Alzheimer (4, 9-15). Since the interaction between SNPs ore more influential than single SNPs in the occurrence of complex disease, it seems necessary to have some methods to identify these influences. Moreover, in order to have a suitable prediction and classification for the intended response, these methods should be able to quantify the significance of these interactions. In this article, feature selection which is a combination of bootstrap and logic regression methods was used to determine potential individual and interaction effects between genetic polymorphisms influential in affliction with one-late Alzheimer and then the two indexes of VIM_{single} and $VIM_{multiple}$ were used to quantify the importance of the specified effects and based on it, one interaction effect of the polymorphisms of *APOE* and *SAITOHON* genes was determined. The results revealed that if the polymorphism A/G in the *SAITOHIN* gene is not of homozygous reference type and in case of non-existence of e2e3 genotype in the person, the risk of Alzheimer increases. Based on the previous studies in non-Iranian populations, the e2 allele increases the age of onset of the disease and protects from it (16-18). In the studies conducted in Iran by Vaisi Reygani and et.al (19) and Gozalpour (20), the frequency of ε2 allele and ε2ε3 genotype in healthy people was reported to be more than in no significant the patients, but there was difference between the two groups. The only study investigating the *SAITOHIN* gene in Iranian population was conducted by Veisi (21); in his study the AA genotype was introduced as having a protective role and the AG genotype as being the risk factor in affliction with Alzheimer. Moreover, in the investigation of interaction of this gene in AG polymorphisms and minus *APOE*ε2 subjects a significance difference was reported but the interaction of GG genotype and minus *APOE*ε2 subjects was not meaningful.

One of the advantages of this method is that unlike other regression models, in order to investigate the existence of an interaction,

Table2: The SNP frequencies were compared between Alzheimer's cases and control subjects

Gene	SNP	genotype	Control	case
			number (percent)	number (percent)
ABCA1	R219K (G/A)	GG	41 (37.6)	34 (33.7)
		GA	50 (45.9)	48 (47.5)
		AA	18 (16.5)	19 (18.8)
CALHM1	P86L	CC	93 (85.3)	79 (78.2)
		CT	12 (11.0)	17 (16.8)

Gene	SNP	genotype	Control	case
			number (percent)	number (percent)
CCR2	CCR2- (V64I) (G/A)	TT	4 (3.7)	5 (5.0)
		GG	91 (83.5)	86 (85.1)
		GA	16 (14.7)	14 (13.9)
		AA	2 (1.8)	1 (1.0)
GSK3â	*Alu*I(T/C)	TT	33 (30.3)	27 (26.7)
		TC	52 (47.7)	53 (52.5)
		CC	24 (22.0)	21 (20.8)
	*ALu*I(G/A)	GG	77 (70.6)	63 (62.4)
		GA	24 (22.0)	30 (29.7)
		AA	8 (7.3)	8 (7.9)
TAU	*Alw*26I(C/G)	CC	80 (73.4)	60 (59.4)
		CG	21 (19.3)	32 (31.7)
		GG	8 (7.3)	9 (8.9)
	*Sep*I(A/G)	AA	77 (70.6)	66 (65.3)
		AG	27 (24.8)	31 (30.7)
		GG	5 (4.6)	4 (4.0)
TNF-α	-308(G/A)	GG	93 (85.3)	77 (76.2)
		GA	15 (13.8)	24 (23.8)
		AA	1 (0.9)	0 (0)
SAITOHIN	*Hinf*I(A/G)	AA	89 (81.7)	67 (66.3)
		AG	17 (15.6)	34 (33.7)
		GG	3 (2.8)	0 (0)
VDR	*Taq*I(C/T)	CC	50 (45.9)	45 (11.9)
		CT	46 (42.2)	46 (42.2)
		TT	13 (11.9)	10 (9.9)
	*Apa*I(G/T)	GG	23 (21.1)	19 (18.8)
		GT	52 (47.7)	46 (45.5)
		TT	34 (31.2)	36 (35.6)

Interactions do not need to be known in advance and used as input variables in the model, but the detection of important variable interactions is the main aim of logic regression; and this way it is possible to concentrate on the most important effects specified by this approach. Since in case-control studies, the goal is to make a classification rule based on the minimum possible number of variables, the identification of the interactions of SNPs influential in predicting the response is the first, and the same time a very important, stage. In the next stage, it is possible to, for example, consider K number of the most important interactions which are higher than specific level of significance and use the form of binary variables in logic regression or any other classification and prediction models. Therefore, conducting studies based on *APOE* and *SAITOHIN* genes with a larger sample is recommended.

Conclusion

Feature selection approach is a new method for the detection of interaction in genetic association studies with many variables. In the present study, the two-way interaction between polymorphisms in *APOE* and *SAITOHIN* genes was detected using this method.

Acknowledgement:

We wish to express our special thanks to all colleagues at the Genetic Research Center - University of Social Welfare and Rehabilitation Sciences, especially Dr. Koorosh Kamali, for their helps in the data collection.

References

1. Garte S. Metabolic susceptibility genes as cancer risk factors: time for a reassessment? Cancer Epidemiol Biomarkers Prev. 2001;10:1233-7.
2. Rahkonen T, Eloniemi-Sulkava U, Rissanen S, Vatanen A, Viramo P, Sulkava R. Dementia with Lewy bodies according to the consensus criteria in a general population aged 75 years or older. J Neurol Neurosurg Psychiatry 2003;74(6):720-4.
3. Hooijmans C, KIliaan A, Fatty acids, lipid metabolism and Alzheimer pathology. Eur J Pharmacol . 2008; 585: 176-96.
4. Shibata N, Kawarai T, Lee JH, Lee H-S, Shibata E, Sato C, et al. Association studies of cholesterol metabolism genes (CH25H, ABCA1 and CH24H) in Alzheimer's disease. Neurosci Lett. 2006; 391(3): 142-6.

5. Reinshagen VH-, zhou S, Burgess B, Bernier L, Mclsaac S, Chan J. Deficiency of ABCA1 Impairs Apolipoprotein E Metabolism in Brain. J Biol Chem. 2004; 279(39): 4119-207.
6. Puglilli L, Tanzi R, Kovasca D. Alzheimer's disease: Cholestrol connection. Neuroscience. 2003; 6(4): 345-51.
7. Ruczinski I, Kooperberg C, Leblanc M. Logic Regression. J COMPUT GRAPH STAT. 2003; 12(3): 475-511.
8. Schwender H, Ickstadt K. Identification of SNP Interaction Using Logic Regression. Biostatistics. 2008; 9: 187-98.
9. Smith MW, Dean M, Carrington M, Winkler C, Huttley GA, Lomb DA, et al. Contrasting Genetic Influence of CCR2 and CCR5 Variants on HIV-1 Infection and Disease Progression. Science. 1997; 277(5328): 959-65.
10. Dreses-Werringloer U, Lambert J-C, Vingtdeux V, Zhao1 H, Vais H, Siebert A, et al. A polymorphism in CALHM1 influences Ca2+ homostasis, AB levels, and Alzheimer Disease risk. Cell. 2008; 133(7): 1149-61.
11. Luo J. Glycogen synthase kinase 3 in tumorigenesis and cancer chemotherapy. Cancer Lett. 2009; 273(2): 194-200.
12. Ezquerra M, Gaig C, Ascaso C, Muñoz E, Tolosa E. Tau and saitohin gene expression pattern in progressive supranuclear palsy. Brain Res. 2007; 1145: 168-76.
13. Candore G, Balistreri CR, Colonna-Romano G, Lio D, Caruso C. Major histocompatibiblity complex polymorphisms and sporadic Alzheimer's disease: a critical reappraisal. Exp Gerontol. 2004; 39(4): 645-52.
14. Zuo L, Dyck C, Luo X, Kranzler H, zhu Yang B, Gelernter J. Variation at APOE and STH loci and Alzheimer's disease. Behav Brain FUNCT. 2006; 2(1).
15. Poduslo S, Yin X. Chromosome 12 and late onset Alzheimer's disease. Neurosci Lett. 2001; 88(310): 188-90.
16. Corder EH, Saunders AM, Risch NJ, Strittmatter WJ, Schmechel DE, Jr PCG, et al. Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. Nat Genet. 1994; 35(7): 180-4.
17. Scott W, Saunders A, Gaskell P, Locke PA, Grow J, Farrer L. Apolipoprptein E e2 does not increase risk of early-onset sporadic Alzheimer's Disease. Ann Neurol. 1997; 36(42): 376-38.
18. VaisiRaygania A, Zahraia M, VaisiRaygania A, Doostia M, Javadic E, Rezaeid M. Association between apolipoprotein E polymorphism and Alzheimer disease in Tehran, Iran. Neurosci Lett. 2005; 58(375): 1-6.
19. Gozalpour E, Kamali K, Mohammd K, Khorram Khorshid HR, Ohadi M, Karimloo M, et al. Association between Alzheimer's Disease and Apolipoprotein E Polymorphisms. Iranian J Publ Health. 2010; 39(2): 1-6.
20. Veisi K. Association study between MAPT, GSK3b and STH genes polymorphisms with sporadic Alzheimer disease in Iranian population. Tehran: University of social welfare and rehabilitation Sciences; 1388.