RESEARCH COMMUNICATION

A Multifactor Dimensionality Reduction-logistic Regression Model of Gene Polymorphisms and an Environmental Interaction Analysis in Cancer Research

Yazhou Wu[1], Ling Zhang[2], Ling Liu[1], Yanqi Zhang[1], Zengwei Zhao[1], Xiaoyu Liu[1], Dong Yi[1*]

Abstract

Background: Analysis of interactions between genes and the environment with complex multifactorial human disease faces important challenges. Limitations of parametric-statistical methods for detection of gene effects that are dependent solely or partially on interactions with other genes or environmental exposures are key problems. The aim of the study was to investigate the use of multifactor dimensionality reduction (MDR) and logistic regression models to analyze the effects of interactions between complex disease genes with other genes and with environmental factors and to compare the results of these two methods in interaction analysis. Methods: In this case-control study, the two methods were applied to analog data of samples from 486 cancer patients and 514 control individuals by computer simulation, including 4 environment factors (E1~E4) and 8 gene polymorphism factors (G1~G8). Non-conditional logistic regression was used to analyze risk factors for cancer, and MDR and logistic regression were employed to analyze interactions under various conditions. Results: MDR could find high-level interactions between genes and the environment (E3*G1*G7), but it could not find a main effect; conversely, logistic regression better analyzed the main effects (E3, G1, and G4) but was limited in its analysis of high-level interactions (E3*G1*G7). The results of these two methods with analog data show that the gene G1 site, the G4 site, E3, and the E3*G1*G7 interaction may be risk factors for occurrence of cancer. Conclusions: MDR and logistic regression, which are the two complementary methods, can be combined to analyze gene-gene (gene-environment) interactions with good results. This approach should help to determine the causes of diseases, such as chronic non-transmittable diseases like cancer.

Keywords: Multifactor dimensionality reduction (MDR) - logistic regression - SNP - cancer - interaction

Asian Pacific J Cancer Prev, **12**, 2887-2892

Introduction

Obtaining a more accurate and more systematic grasp of the occurrence and development of complex diseases influenced by multiple genes is the most important research topic in modern genetic epidemiology. Complex diseases, such as colorectal cancer and bladder cancer, are affected by genetic factors, such as the number of single nucleotide polymorphisms (SNP), and environmental factors (Bosetti et al., 2005; Parkin et al., 2008; Reeves et al., 2008; Yang et al., 2009). This study shows that the occurrence and development of complex diseases are not entirely due to genetic factors. Rather, these are the combined result of genetic variation and environmental factors. There may only be a weak association between each gene and disease, with no major gene effect. This effect is more vulnerable to being overcome by the external environment. If both gene-gene and gene-environment interactions are ignored, the effects of genetic variation may not be present or accurately described. Thus, how to properly analyze and evaluate genetic and environmental interactions with diseases is very significant for disease prevention and forming public health policies (Nilanjan et al., 2006; Wong et al., 2010). How to effectively process and analyze multi-gene distributions and the related environmental factors of complex diseases are the major issues that current genetic epidemiologists and bioinformaticians face. Especially in cases of gene-gene and gene-environment interactions, currently, the greatest challenge is choosing the proper statistical method to analyze these interactions to evaluate their role in disease.

In recent years, rapid progress has been made in the development of statistical methods for analyzing gene-gene and gene-environment interactions, such as logistic regression models, stratified analysis, crossover analysis, general relative risk models, composite linkage and disequilibrium methods (Wu et al., 2008). These methods have their own advantages but also have many other

*[1]Department of Health Statistics, College of Preventive Medicine, [2]Department of Health Education & Medical Humanities, Third Military Medical University, Chongqing, China *For correspondence: yd_house@hotmail.com, asiawu5@sina.com*

issues, such as multiple testing, dimensional problems, and model dependence. In contrast, most traditional methods can only analyze single SNPs or low-level SNP interactions associated with the disease but cannot analyze the impact of high-level interactions of SNPs on disease generation. Thus, many samples are required for studies of high-level SNP interactions.

Logistic regression is a conventional method for processing interactions between classified variables. In a situation with fewer independent variables, the impact of the interaction can be deduced by examining whether the interaction is statistically significant, but when processing higher-level interactions, logistic regression has many limitations. Multifactor dimensionality reduction (MDR) is a new method for interaction analysis that has been developed in recent years. Its main advantage is its ability to simultaneously detect and characterize the combined effects of multiple disease factors and that it is a parameter-free approach. Therefore, it does not require a predefined genetic model to handle dimension reduction. This paper uses MDR in combination with logistic regression to analyze the analog case-control data of the samples for 1000 by the computer simulation, and find the factors affecting the cancer and the interactions between genes and the environment. First, it introduces the basic principles of the MDR method and the major analysis steps. Then it contrasts the results from applying MDR to those of logistic regression. Next, it introduces the implementation and principles behind applying these two methods. Finally, it explores the conditions and advantages of these methods.

Materials and Methods

The logistic regression model

Assuming a dependent variable Y and m independent variables, when Y is a binary variable,

$$Y=\begin{cases}1 & \text{positive results (effective, die, etc.)}\\ 0 & \text{negative results (before the illness, ineffective, living, etc.)}\end{cases}$$

Recorded in a group of independent variables, the probability of obtaining a positive result is, which we will denote as P; then it follows that the Logistic regression model can be expressed as

$$P=\frac{1}{1+\exp\left[-(\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m)\right]} \quad (1)$$

where α is a constant, and $\beta_1,\beta_2,....,\beta_{20}$ are the regression coefficients.

If formula (1) is transformed, then there is another linear expression for the Logistic regression model:

$$\ln\left(\frac{P}{1-P}\right)=\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m \quad (2)$$

The left side of this formula is the natural logarithm of the ratio of the positive occurrence probability to the negative occurrence probability, which we will denote as Logit(P). It then follows that

$$\text{Logit}(P)=\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m \quad (3)$$

This formula is called the Logit model.

If P in the Logistic model is seen as the probability of a disease state, then the constant represents the exposure to a dose of 0 or the natural logarithm of the ratio of the disease incidence probability to the probability of no disease. The regression coefficient β_j (j =1,2...,m) indicates the changes in Logit(P) when the X_j factor changes by a unit, and its corresponding relationship when measuring the risk factors odds ratio (OR) is the following:

$$OR=\exp(\beta_j) \quad (4)$$

When $\beta_j = 0$, OR = 1, which indicates that Xj has no effect on the disease; when $\beta_j > 0$, OR > 1, which indicates that is a risk factor. When $\beta_j < 0$, OR < 1, which indicates that Xj is a protective factor.

Note: when the function Y of an independent variable Xi changes with another independent variable X_j, the interaction between the independent variables X_i and X_j should be considered. To address this issue, one of the most common methods is to include the product term (X_i x X_j) in the equation. Consider the following fitting equation:

$$\text{Logit}(P)=a+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m+\gamma X_iX_j \quad (5)$$

If the product (X_i x X_j) is significant, then there is an interaction between X_i and X_j.

Whether this interaction must considered in the regression equation mainly depends on one's professional knowledge. In reality, the interaction and main effects have certain roles. To avoid impacting the total, the mean can be subtracted from each variable before multiplying the variables together.

The basic principles and analysis steps of the MDR method

Ritchie et al. (2001) and Coffey et al. (2004) proposed a multifactor dimensionality reduction (MDR) method to study case-control data, gene-gene interactions, and gene-environment interactions. Here, the "factor" is a variable in the study of the interaction, and the "dimension" refers to the number of multi-factor combinations. This method uses a disease susceptibility classification to create its models, multiple factors are treated as multi-factor combinations, and the objects of study are divided into two levels (high-risk and low-risk) to reduce the dimensionality according to the disease susceptibility. Finally, through a 10-fold cross-validation and a permutation test, the factor combination identification and disease prediction abilities were evaluated to calculate the classification and prediction errors.

The MDR method combines attribute selection, attribute construction, and classification with cross-validation and permutation testing to provide a comprehensive and powerful data mining approach to detecting nonlinear interactions. The concrete steps of the MDR method are as follows:

1)Use a 10-fold cross-validation method to divide the data into 10 sets, in which 9 subsets are training sets, and 1 subset is a test set; 2) Determine the number of combined factors n according to the total number of factors; 3) For each training set and test set, screen the best n-factor combination (the screening criteria are based

Table 1. Variables and Their Assignment for a Cancer in this Study

Variable		Variable assignment		
Environment variables	E1	no = 0	yes = 1	
	E2	no = 0	yes = 1	
	E3	low = 1	middle = 2	high = 3
	E4	low = 1	middle = 2	high = 3
Gene variables	G1	aa = 1	aA = 2	AA = 3
	G2	aa = 1	aA = 2	AA = 3
	G3	aa = 1	aA = 2	AA = 3
	G4	aa = 1	aA = 2	AA = 3
	G5	aa = 1	aA = 2	AA = 3
	G6	aa = 1	aA = 2	AA = 3
	G7	aa = 1	aA = 2	AA = 3
	G8	aa = 1	aA = 2	AA = 3
Dependent variable	class	control = 0	case = 1	

on the minimum classification error in the training set); 4) Repeat this process 10 times and then screen the best n-factor combinations based on the average minimum and maximum cross-validation; 5) Calculate the ratio of cases to controls for each n-factor combination. If the ratio is equal to or exceeds the threshold, then the genotype combination is identified as high-risk; otherwise, it is identified as low-risk (if the number of cases and controls equal to the domain is 1); 6) For various values of n, we obtain the best n-factor combinations. Some n-factor combinations may have a smaller prediction error, while others may have the greatest cross-validation consistency; we take the model with the smaller value of n. The model with the minimum prediction error and the maximum cross-validation consistency is the optimal factor model.

Results

Simulated data

This study used a case-control method and conducted an analysis on analog data of the samples for 1000 by the computer simulation. There are 486 patients with a cancer and 514 control individuals, which include 4 environmental factors (E1, E2, E3 and E4) and 8 gene polymorphism factors (G1, G2, G3, G4, G5, G6, G7 and G8). The assignment of variables is shown in Table 1.

Data processing and statistical analysis

First, a Hardy-Weinberg equilibrium fitting test was performed on each genotype distribution for cases group (Salanti et al., 2005), where a P-value of > 0.05 shows that the genotype meets the HW equilibrium condition. This was followed by the use of a SPSS 18.0 to conduct non-conditional multivariate logistic regression analyses to estimate the odds ratios (OR) of the cancer risk factors; MDR (version 2.0_beta_8.4 (Lance et al., 2003; Moore et al., 2006) software was used to analyze the gene-gene and gene-environment interactions, and we compared these results to those of the logistic regression analysis. A P-value of < 0.05 indicates a statistically significant difference.

The analytical results of the factors that affect the logistic regression model (main effects)

SPSS was used to perform a non-conditional multivariate logistic regression analysis on the 12 factors that may affect the cancer, and the results are detailed in Table 2. The results show that the environment variable E3, the G1 gene site, and the G4 site are possible risk factors for the cancer.

The results of the interaction of the multi-factor dimensionality reduction model analysis (interaction effect)

MDR software was used to analyze the interaction of the 12 factors that may affect the cancer, and the results detailed in Table 3. We found that the cross-validation (CV) consistency of the three-factor model (E3*G1*G7) was maximal (10/10), and the accuracy of the test samples was the highest (0.6941), the accuracy of the training samples was the 0.7029. Thus, the three-factor interaction model was the best model, which shows that there was an interaction between the environmental factors E3, the genetic G1 site, and the G7 site (p<0.0001). However, the MDR method used here did not find the main effect that the logistic regression analysis found.

Figure 1 shows the interaction model for the three factors for the cancer (E3, G1, and G7). In the cell in the figure, the left bands represent the disease case, and the right bands represent the control case. It can be seen in the figure that the G1 gene site represents the heterozygote AA, the G7 gene site represents the homozygous aA genotype, and the relative risk for the cancer is increased in the E3 environmental population. The OR value was 5.1971 (2.2121, 12.2104) (P < 0.0001), which was calculated by the MDR software.

Table 2. The Results of a Logistic Regression (Forward LR) Analysis on the Factors that Affect a Cancer

Factors	β value	SE	Wald	P-value	OR	95% CI	
						Lower	Upper
E3			137.6205	0.0000			
E3(1)	1.3981	0.1769	62.4537	0.0000	4.0475	2.8616	5.7251
E3(2)	1.9575	0.1708	131.3468	0.0000	7.0816	5.0669	9.8974
G1			18.2768	0.0001			
G1(1)	-0.3342	0.1816	3.3888	0.0656	0.7159	0.5015	1.0219
G1(2)	0.4076	0.1647	6.1212	0.0134	1.5031	1.0884	2.0759
G4			23.8720	0.0000			
G4(1)	0.6157	0.1695	13.2001	0.0003	1.8510	1.3279	2.5803
G4(2)	-0.1727	0.1796	0.9246	0.3363	0.8414	0.5917	1.1964
Constant	-0.1119	0.0714	2.4534	0.1173	0.8942		

Table 3. The Results of the Interaction of the Impact Factors as Determined by an MDR Analysis of a Cancer

Factor number	Factors	Training sample		Test sample		Cross-validation
		Accuracy	p	Accuracy	p	consistency
1	E3	0.6785	<0.0001	0.6785	0.0002	10/10
2	E3 G3	0.6855	<0.0001	0.6762	0.0004	8/10
3	E3 G1 G7	0.7029	<0.0001	0.6941	<0.0001	10/10
4	E3 G1 G4 G7	0.7314	<0.0001	0.6849	0.0002	10/10
5	E3 E4 G1 G4 G8	0.7881	<0.0001	0.6140	0.0214	3/10
6	E3 G1 G4 G5 G7 G8	0.8740	<0.0001	0.5874	0.0661	7/10
7	E3 G1 G2 G4 G5 G7 G8	0.9473	<0.0001	0.5469	0.2265	8/10
8	E1 G1 G2 G4 G5 G6 G7 G8	0.9813	<0.0001	0.5114	0.6968	3/10
9	E1 E3 E4 G1 G2 G4 G6 G7 G8	0.9913	<0.0001	0.4916	0.6980	3/10
10	E2 E3 E4 G1 G2 G3 G4 G6 G7 G8	0.9976	<0.0001	0.4994	0.9665	3/10
11	E1 E2 E3 E4 G1 G2 G3 G4 G6 G7 G8	0.9991	<0.0001	0.4999	0.9858	9/10
12	E1 E2 E3 E4 G1 G2 G3 G4 G5 G6 G7 G8	0.9991	<0.0001	0.4990	0.7449	10/10

Table 4. The Analysis Results of Logistic Regression (Forward LR) Performed after Adding an Interaction (in the Model) for a Cancer

Factors	β value	SE	Wald	P-value	OR	95% CI	
						Lower	Upper
E3			136.3681	0.0000			
E3(1)	1.4416	0.1788	64.9957	0.0000	4.2275	2.9777	6.0020
E3(2)	1.9595	0.1729	128.4338	0.0000	7.0955	5.0560	9.9576
G1			18.0619	0.0001			
G1(1)	-0.3446	0.1835	3.5259	0.0604	0.7085	0.4945	1.0152
G1(2)	0.4016	0.1662	5.8399	0.0157	1.4942	1.0788	2.0695
G4			22.9417	0.0000			
G4(1)	0.6040	0.1708	12.5023	0.0004	1.8294	1.3089	2.5569
G4(2)	-0.1780	0.1813	0.9645	0.3261	0.8369	0.5867	1.1939
E3 * G7			15.2983	0.0041			
E3(1) by G7(1)	-0.8255	0.4222	3.8228	0.0506	0.4380	0.1915	1.0020
E3(1) by G7(2)	-0.3306	0.4475	0.5458	0.4600	0.7185	0.2989	1.7272
E3(2) by G7(1)	0.7742	0.4033	3.6855	0.0549	2.1688	0.9839	4.7807
E3(2) by G7(2)	0.7279	0.4216	2.9809	0.0843	2.0707	0.9063	4.7315
Constant	-0.1017	0.0722	1.9828	0.1591	0.9033		

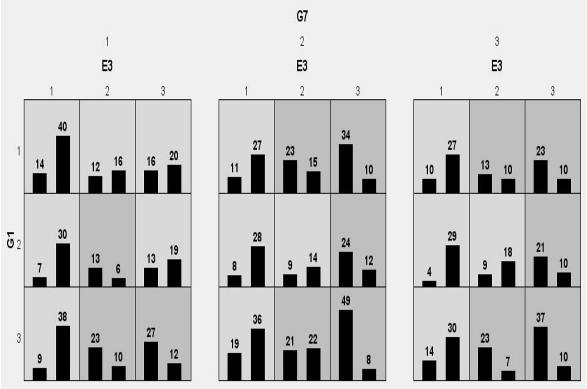

Figure 1. An MDR Analysis of the Three-factor (E3, G1, and G7) Interaction Model of a Cancer. In the cell in the figure, the left bands represent the disease case, and the right bands represent the control case

Figure 2 shows a tree diagram of the interaction system for the three factors for the cancer (E3, G1, and G7). Blue and green represent a synergetic interaction, where the blue intensity is greater than the green intensity. Red and orange represent an antagonistic interaction effect, where the red intensity is stronger than the orange intensity. From the color in the diagram, we know that the interaction of three factors (E3 and the G1-G4 gene site) shows a strong synergistic effect. However, after adding the G3 and G7 gene site, it showed a strong antagonistic effect, indicating that the interaction among all three factors (E3, the G1 gene site, and the G7 site) restrain the occurrence of the cancer.

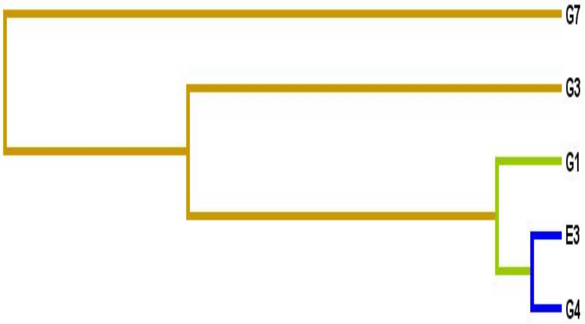

Figure 2. A Tree Diagram of the Interactions Among Three Factors for a Cancer (smoking, XRCC194, and XRCC7), as Analyzed by MDR. Blue and green represent a synergetic interaction, where the blue intensity is greater than the green intensity. Red and orange represent an antagonistic interaction effect, where the red intensity is stronger than the orange intensity

The results of the logistic regression model analysis after adding the interaction effect

A forward logistic regression method was used to analyze the aforementioned data, and an MDR approach was introduced to analyze the three-factor interaction (E3*G1*G7). The factors remaining in the model were E3, G1, G4, and the low interaction among the two

Table 5. The Analysis Results of Logistic Regression (Forward Lr) Performed after Adding an Interaction (not in the Model) for a Cancer

Variables	Score	df	P
E1	0.0001	1	0.9934
E2	0.7943	1	0.3728
E4	0.8228	2	0.6627
E4(1)	0.0342	1	0.8533
E4(2)	0.4858	1	0.4858
G2	0.3402	2	0.8436
G2(1)	0.1293	1	0.7192
G2(2)	0.3327	1	0.5641
G3	0.4285	2	0.8071
G3(1)	0.2360	1	0.6271
G3(2)	0.3832	1	0.5359
G5	0.2907	2	0.8647
G5(1)	0.0913	1	0.7626
G5(2)	0.0661	1	0.7971
G6	2.1901	2	0.3345
G6(1)	0.8430	1	0.3585
G6(2)	0.3044	1	0.5811
G7	3.9158	2	0.1412
G7(1)	2.0025	1	0.1570
G7(2)	0.2269	1	0.6338
G8	4.4424	2	0.1085
G8(1)	2.7226	1	0.0989
G8(2)	3.4335	1	0.0639
E3 * G1	2.4527	4	0.6531
E3(1) by G1(1)	0.0001	1	0.9939
E3(1) by G1(2)	0.0147	1	0.9034
E3(2) by G1(1)	0.0159	1	0.8997
E3(2) by G1(2)	1.7223	1	0.1894
G1 * G7	4.6854	4	0.3211
G1(1) by G7(1)	0.1420	1	0.7063
G1(1) by G7(2)	1.9010	1	0.1680
G1(2) by G7(1)	0.6533	1	0.4189
G1(2) by G7(2)	1.2446	1	0.2646
E3 * G1 * G7	13.5362	8	0.0947
E3(1) by G1(1) by G7(1)	0.0164	1	0.8980
E3(1) by G1(1) by G7(2)	3.0196	1	0.0823
E3(1) by G1(2) by G7(1)	4.4176	1	0.0356
E3(1) by G1(2) by G7(2)	2.8012	1	0.0942
E3(2) by G1(1) by G7(1)	0.0981	1	0.7541
E3(2) by G1(1) by G7(2)	4.0622	1	0.0439
E3(2) by G1(2) by G7(1)	0.3523	1	0.5528
E3(2) by G1(2) by G7(2)	4.4301	1	0.0353
Overall Statistics	33.4696	32	0.3959

factors (E3*G7), with statistically significant P-values of P=0.0004. Therefore, E3, G1, G4, and E3*G7 can be considered impact factors for the cancer; the results are shown in Table 4. However, the high interaction among the three factors (E3*G1*G7) obtained by the MDR method did not remain in the model, as P>0.05 (P=0.0947), which is not statistically significant. The results of this analysis are displayed in Table 5.

The results described above and the results from the MDR analysis are not identical, possibly because MDR reduced the multidimensional model to a one-dimensional model, and the quality of analog data is relatively poor. This method was created by differentiating the high-risk and low-risk sample groups. This is flawed because, to create this dimensionality reduction, MDR had to search for the multidimensional space. To detect interactions, MDR was more sensitive to high-level interactions.

Discussion

Exploring the interactions among risk factors (gene-environment) for complex diseases is important for disease epidemiology because these interactions exist, and various interaction patterns have differential epidemiological significances (public health significance). An understanding of gene-environment interaction also has important implications for public health. It aids in predicting disease rates and provides a basis for well-informed recommendations for disease prevention (Ruth, 1996).

The cancer (such as colorectal cancer and bladder cancer) is caused by multifactorial, multi-step, and complex pathological changes. It has both intrinsic genetic factors and external environment factors, and it is especially affected by the interactions between genetic and environmental factors. This study used MDR and a logistic regression method to analyze analog data of the samples for 1000 by the computer simulation, the possible impact factors of a cancer and the interactions among these factors. The logistic regression results show that environmental factors, E3, genetic factors, G1, and G4 may be risk factors, but their interaction has not been observed. However, the MDR method found a third-order interaction (E3*G1*G7). The results of combining these two methods showed that the G1 gene site, the G4 site, E3, and the E3*G1*G7 interaction may be impact factors for the cancer. MDR combination with logistic regression method will further be applied to a concrete cancer disease (such as colorectal cancer and bladder cancer) to verify the characteristics of two methods on gene-gene (environment) interactions, which is the next plan of ours research.

Logistic regression is a conventional method for processing the interactions among classified response variables. In a small number of independent variables, by examining whether the interaction is statistically significant, the impact of the interaction can be extrapolated. The difference between it and MDR is that logistic regression can distinguish main effects the interactive effects, and for low-dimensional data, logistic regression is more appropriate than MDR. However, logistic regression has its own shortcomings. When there are interactions among multiple factors, the model can often produce more parameters. If the sample size is relatively small (due to over-fitting), many concerns (e.g., “dimension distress”, etc.) will be raised about the statistical method. When dealing with high-level interactions, logistic regression has significant limitations, as it is difficult to discover these interactions.

MDR is a method that was recently discovered for analyzing interactions. Its main advantage is its ability to simultaneously detect and characterize the impact of the combined effects of multiple disease factors. It does not consider main effects when analyzing various factors. Therefore, when the main effects are not statistically significant, it still can find high-level interactions. Using a tree diagram of the interaction system, researchers can visually determine the effects and strengths of interactions to provide a basis for further studies. However, when the main effects are significant, MDR cannot find the main effect, and logistic regression can be used in conjunction. MDR is first used to detect the interaction, and the interaction is then forced into the main effects and the interaction effects of a logistic regression analysis. MDR finds that the capacity for interactions decreases with a decreased number of factors and levels because the number of parameters is exponentially positively correlated with the number of factors. When the real interaction is effective and has a low dimension, MDR finds that the interaction capacity is greatly reduced.

It is worth noting that, for a high-level interaction between genes and the environment, MDR better handles multidimensional data. However, the MDR method also has shortcomings. First, when the sample size of the disease cases with multi-site genotypes and the sample size of the control cases are small, it is easy to obtain false positive or false negative results. Second, for each genotype with a number of sites, it can only be classified as high- or low-risk, and we are unable to obtain indicators for quantification. Thirdly, the MDR method cannot perform comparisons between different genotypes, but in practice, we need to understand the risk ratios of various genotypes. Finally, when the number of factors is more than 15, it brings out a large amount of calculation, and the speed of is slower and the time is longer for MDR software.

In summary, when MDR and logistic regression analyze interactions, there are various conditions and advantages, and in the case of an interaction analysis, a combination of both is superior. By comparing examples, this study introduces the principles behind and the application of these two methods to analyze interactions. This study also explores the applicable conditions and advantages of each method. It provides new ideas for medical research in the field of interaction analysis.

Currently, many researchers have explored other methods for statistically analyzing gene/environment interactions. For example, a neural network method to process the interactions between genes and the environment also has unique advantages (Motsinger et al., 2006; Frauke et al., 2009), and it obtains the internal relationships between variables by studying samples. With no prior knowledge of the relationship between the variables, it is not affected by interference, and it obtains a richer and more realistic relationship between the input and output to analyze the complex interactions between the factors. Genome-wide associations are the current research focus for studying complex disease susceptibility genes, and many new breakthroughs have been made (Elbers et al., 2009; Roukos, 2009).

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No.30901242, 81172773). The authors would like to thank the reviewers and researchers who participated in this study. The author(s) declare that no potential conflicts of interest were disclosed.

References

Bosetti C, Pira E, Vecchial CL (2005). Bladder cancer risk in painters: a review of the epidemiological evidence, 1989-2004. *Caneer Causes Control*, **16**, 997-1008.

Coffey C, Hebert P, Ritchie M, et al (2004). An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. *BMC Bioinformatics*, **5**, 49.

Elbers CC, van Eijk KR, Franke L, et al (2009). Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genetic Epidemiology*, **33**, 419-31.

Günther F, Wawro N, Bammann K (2009). Neural networks for modeling gene-gene interactions in association studies. *BMC Genetics*, **10**, 87.

Hahn LW, Ritchie MD, Moore JH (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, **19**, 376-82.

Moore JH, Gilbert JC, Tsai CT, et al (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol*, **241**, 252-61.

Motsinger AA, Lee SL, Mellick G, et a1 (2005). GPNN-Power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. *BMC Bioinformatics*, **7**, 39.

Nilanjan C, Zeynep K, Roxana M, et al (2006). Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment Interactions. *Am J Hum Genet*, **79**, 1002-16.

Ottman R (1996). Theoretical epidemiology gene-environment interaction: Definitions and study designs. *Prev Med*, **25**, 764-70.

Parkin DM (2008). The global burden of urinary bladder cancer. *Scand J Urol Nephrol Suppl*, **218**, 12-20.

Reeves SG, Mossman D, Meldrum CJ, et al (2008). The 149 C/T SNP within the DDNMT3B gene, is not associated with early disease onset in hereditary non-polyposis colorectal cancer. *Cancer Letters*, **265**, 39-44.

Richie MD, Hahn LW, Roodi N, et al (2001). Multifactor dimensionality reduction reveals high-order interaction among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, **69**, 138-47.

Roukos DH (2009). Genome-wide association studies and aggressive surgery toward individualized prevention, and improved local control and overall survival for gastric cancer. *Ann Surg Oncol*, **16**, 795-8.

Salanti G, Amountza G, Ntzani E, et al (2005). Hardy-Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power. *Eur J Hum Genet*, **13**, 840-8.

Wong HL, Ulrike P, Richard BH, et al (2010). Polymorphisms in the adenomatous polyposis coli (APC) gene and advanced colorectal adenoma risk. *Eur J Cancer*, **46**, 2457-66.

Wu X, Jin L, Xiong M (2008). Composite measure of linkage disequilibrium for testing interaction between unlinked loci. *Eur J Hum Genet*, **16**, 644-51.

Yang H, Zhou Y, Zhou Z, et al (2009). A novel polymorphism rs1329149 of CYP2E1 and a known polymorphism rs671 of ALDH2 of alcohol metabolizing enzymes are associated with colorectal cancer in a southwestern Chinese population. *Cancer Epidemiol Biomarkers Prev*, **18**, 2522 -7.