BMC Genetics

BioMed Central

Research article

Open Access

Retrospective analysis of main and interaction effects in genetic association studies of human complex traits

Qihua Tan*[1,2], Lene Christiansen[1,2], Charlotte Brasch-Andersen[2,3], Jing Hua Zhao[4], Shuxia Li[1], Torben A Kruse[2] and Kaare Christensen[1]

Address: [1]Epidemiology, Institute of Public Health, University of Southern Denmark, Denmark, [2]Department of Biochemistry, Pharmacology and Genetics, Odense University Hospital, Denmark, [3]Clinical Pharmacology, Institute of Public Health, University of Southern Denmark, Denmark and [4]MRC Epidemiology Unit, The Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK

Email: Qihua Tan* - qihua.tan@ouh.regionsyddanmark.dk; Lene Christiansen - lchristiansen@health.sdu.dk; Charlotte Brasch-Andersen - charlotte.b.andersen@ouh.regionsyddanmark.dk; Jing Hua Zhao - jhz22@medschl.cam.ac.uk; Shuxia Li - sli@health.sdu.dk; Torben A Kruse - torben.kruse@ouh.regionsyddanmark.dk; Kaare Christensen - kchristensen@health.sdu.dk

* Corresponding author

Published: 16 October 2007

BMC Genetics 2007, **8**:70 doi:10.1186/1471-2156-8-70

Received: 17 April 2007
Accepted: 16 October 2007

This article is available from: http://www.biomedcentral.com/1471-2156/8/70

Abstract

Background: The etiology of multifactorial human diseases involves complex interactions between numerous environmental factors and alleles of many genes. Efficient statistical tools are demanded in identifying the genetic and environmental variants that affect the risk of disease development. This paper introduces a retrospective polytomous logistic regression model to measure both the main and interaction effects in genetic association studies of human discrete and continuous complex traits. In this model, combinations of genotypes at two interacting loci or of environmental exposure and genotypes at one locus are treated as nominal outcomes of which the proportions are modeled as a function of the disease trait assigning both main and interaction effects and with no assumption of normality in the trait distribution. Performance of our method in detecting interaction effect is compared with that of the case-only model.

Results: Results from our simulation study indicate that our retrospective model exhibits high power in capturing even relatively small effect with reasonable sample sizes. Application of our method to data from an association study on the catalase -262C/T promoter polymorphism and aging phenotypes detected significant main and interaction effects for age-group and allele T on individual's cognitive functioning and produced consistent results in estimating the interaction effect as compared with the popular case-only model.

Conclusion: The retrospective polytomous logistic regression model can be used as a convenient tool for assessing both main and interaction effects in genetic association studies of human multifactorial diseases involving genetic and non-genetic factors as well as categorical or continuous traits.

Background

The changing disease pattern has brought complex diseases as one of the significant challenges for the 21st century medicine. As the etiology of complex diseases involves both multiple genetic and environmental factors combined with their interactions, statistical methods for

efficiently measuring the main and interaction effects are demanded. In the literature of genetic epidemiology, the linkage disequilibrium (LD) based genetic association study, advantaged by the recent development of high-throughput SNP genotyping technology, has been the workhorse and holds the promise of mapping out susceptibility genes to complex diseases [1]. The case-control design, a retrospective design by nature, has been popular in establishing the genetic associations in single locus and haplotype analyses [2] as well as in assessing gene-environment interactions [3,4]. Recently, this approach has been extended to handle both dichotomous and continuous traits by introducing the retrospective logistic regression model [5] that treats alleles or genotypes as dependent variables. For example, the idea has been used by Waldman et al. [6] to model the probability of allele transmission as a function of offspring's trait value in family-based transmission disequilibrium test (TDT). The same idea has been used for single locus analysis in both unmatched and matched case-control studies [7] and for haplotype analysis [8].

Another contribution to genetic epidemiology by the retrospective case-control design is the introduction of non-traditional case-only design [9] for assessing gene-environment and gene-gene interactions. Measuring the interaction effects in complex disease study is important because many of the susceptibility genes act through modification of disease risk associated with other genes or environmental factors. Unfortunately, application of the case-only method is restricted to dichotomous or binary traits. Otherwise one needs to set up a cut-off on a continuous trait to define cases [10]. Although other statistical models for measuring interactions [11-13] exist, Glaser et al. [14] recently reported that different methods can give different results for the same data due to underlying assumptions.

In this paper, we introduce a retrospective polytomous logistic regression model to measure both the main and the interaction effects in genetic association studies of human complex traits which can be discrete or continuous. In this model, combinations of genotypes at two interacting loci or of environmental exposure and genotypes at one locus are treated as nominal outcomes of which the proportions are modeled as a function of the disease trait assigning both main and interaction effects. The performance of our method in detecting interaction effect is compared with that of the case-only model. A limited simulation study is performed to assess the power and type I error rate for given parameters settings. Application of our method is exemplified using data from our association study on catalase -262C/T promoter polymorphism and aging phenotypes [15] to look for both main and interaction effects of genetic variation and aging in affecting cognitive function in the Danish population.

Methods

Suppose we are interested in assessing the main and interaction effects of one genetic variant G (allele or genotype, G = 1 for carriers and 0 for non-carriers) and environmental exposure E (E = 0 for non-exposed and 1 for exposed). The combination of G and E leads to four nominal categories. The purpose of our retrospective polytomous logistic regression model is to model the proportion of each of the categories, *p*, as a function of the disease trait by treating the proportions as responses for given trait value *x*, i.e. we model

$$Logit[p(G = I, E = J|x)] = a_G I + a_E J + (b_G I + b_E J + b_{G\times E} IJ)x \quad I, J = 0,1 \qquad (1)$$

where *a* and *b* are the intercept and slope parameters assigned to G and E respectively. In (1), the association of G and E with trait *x* are measured by the slope parameters with b_G and b_E for the main effects from G and E and $b_{G\times E}$ for their interaction effect. Rewriting (1) as the conditional probability of each outcome category given the trait value *x*, we have

$$p(G = I, E = J \mid x) = \frac{\exp[a_G I + a_E J + (b_G I + b_E J + b_{GxE} IJ)x]}{\sum_{I'}\sum_{J'} \exp[a_G I' + a_E J' + (b_G I' + b_E J' + b_{GxE} I' J')x]} \quad I, J, I', J' = 0,1 \qquad (2)$$

When $I = J = 0$, the numerator of (2) becomes 1 so that we have

$$p(G = 0, E = 0 \mid x) = \frac{1}{\sum_{I'}\sum_{J'} \exp[a_G I' + a_E J' + (b_G I' + b_E J' + b_{GxE} I' J')x]} \quad I', J' = 0,1 \qquad (3)$$

Since (3) is for the group of individuals who are neither carriers of the genetic variant nor exposed, it can serve as the reference or baseline. With that, we are able to derive the relative risks (*RR*) for the main and interaction effects at a given trait value *x* and then define the relative risk ratios (*RRR*) for comparing *RR* at two given trait values x_1 and x_2. To obtain RR for the main genetic effect, we set $I = 1$ and $J = 0$ so that (2) becomes

$$p(G = 1, E = 0 \mid x) = \frac{\exp(a_G + b_G x)}{\sum_{I'}\sum_{J'} \exp[a_G I' + a_E J' + (b_G I' + b_E J' + b_{GxE} I' J')x]} \quad I', J' = 0,1 \qquad (4)$$

The *RR* for the main genetic effect is then calculated as

$$RR_G(x) = \frac{p(G = 1, E = 0 \mid x)}{p(G = 0, E = 0 \mid x)} = \exp(a_G + b_G x) \qquad (5)$$

Based on (5), we can calculate *RRR* for comparing *RR*s at two given trait values x_1 and x_2 as

$$RRR_G = \frac{RR_G(x_2)}{RR_G(x_1)} = \frac{\exp(a_G + b_G x_2)}{\exp(a_G + b_G x_1)} = \exp[b_G(x_2 - x_1)] = \exp(kb_G) \quad (6)$$

Note that, when $k = 1$ such as in a case-control study, we have $RRR_G = \exp(b_G)$. In the same manner, we obtain $RRR_E = \exp(kb_E)$.

In order to estimate the risk of interaction effect, we set $I = J = 1$ so that (2) becomes

$$p(G = 1, E = 1 \mid x) = \frac{\exp[a_G + a_E + (b_G + b_E + b_{GxE})x]}{\sum_{I'}\sum_{J'} \exp[a_G I' + a_E J' + (b_G I' + b_E J' + b_{GxE} I' J')x]} \quad I', J' = 0,1 \quad (7)$$

The RR for comparing exposed carriers to unexposed non-carriers at x is

$$RR(x) = \frac{p(G = 1, E = 1 \mid x)}{p(G = 0, E = 0 \mid x)} = \exp[a_G + a_E + (b_G + b_E + b_{GxE})x] \quad (8)$$

Likewise, the *RRR* for $k = x_2 - x_1$ is

$$RRR = \frac{RR(x_2)}{RR(x_1)} = \exp[k(b_G + b_E + b_{GxE})] = \exp(kb_G)\exp(kb_E)\exp(kb_{GxE}) = RRR_G RRR_E \exp(kb_{GxE}) \quad (9)$$

From (9) we obtain $RRR_{G\times E}$ as the departure from the multiplicative effects of $RRR_G RRR_E$, i.e. $RRR_{G\times E} = \exp(kb_{G\times E})$.

In order to estimate the parameters, we construct the following likelihood function [16]

$$L(a_G, a_E, b_G, b_E, b_{GxE}) = \prod_n \prod_{I,J} \pi_{IJ}(x_n)^{I(G_n = I \& E_n = J)} \quad I, J = 0,1 \quad (10)$$

Here, n denotes individual observations from 1 to N, $\pi_{IJ}(x_n) = p(G = I, E = J|x_n)$, $I(\cdot)$ is an indicator function. Since our interest is only in the slope parameters, the two intercept parameters are just nuisance parameters. To obtain an overall significance of both main and interaction effects, we use the log-likelihood ratio test with $df = 3$,

$$LRT = -2[\ln L(a_G, a_E) - \ln L(a_G, a_E, b_G, b_E, b_{G\times E})] \quad (11)$$

where $L(a_G, a_E)$ is the likelihood of the intercept only model. Statistical test on single slope parameters can be done likewise or by introducing the Wald statistic [16].

Simulation

In order to examine the performance of our method, we perform a limited computer simulation study to assess the power and type I error rate (α) when given different parameter settings. The data were simulated using a linear model, i.e. for individual i, we have $y_i = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 G*E + e_i$. Here y_i is a continuous phenotype value for individual i. G and E are the indictors for the genetic (1 for carriers and 0 for non-carriers, we set frequency of carriers to 0.2) and environmental (1 for exposed and 0 for non-exposed, exposure rate set to 0.3) variants. β_0 is the intercept which we set to 0.1. β_1, β_2 and β_3 are the slope parameters for the genetic, environmental and their interaction effects respectively. For simplicity, we assume that there is no main effect in the model, but there is an interaction effect that lowers the phenotype value for carriers for the genetic variant and who are exposed to the environment. The power for capturing the interaction effect is estimated as the frequency for rejecting a null hypothesis of $\beta_3 = 0$ for a given type I error rate (we set $\alpha = 0.05$). The last term e_i is the error part for individual i which follows a standard normal distribution $N(0,1)$. To assess the model performance, we specify different sample sizes and assign different values for β_3. We set β_3 to -0.65, -0.95 and -1.2 so that the interaction effect accounts for about 5, 10 and 15 percent of the total variance in the data.

In Table 1, we show the power estimates using 500 replicates for given type I error rate of $\alpha = 0.05$. It can be seen that for an interaction effect that explains only 5 percent

Table 1: Power and empirical type I error rate for given $\alpha = 0.05$

Sample size	Power			Type I error
	β_3 = -0.65	β_3 = -0.95	β_3 = -1.20	β_3 = 0
150	0.348	0.642	0.780	0.052
200	0.540	0.668	0.894	0.048
250	0.604	0.852	0.898	0.054
300	0.664	0.884	0.960	0.046
400	0.760	0.948	1.000	0.050
600	0.876	1.000	1.000	0.052

of the total variance (β_3 = -0.65), a sample size of above 400 is needed in order to achieve reasonable power. For an effect responsible for 15 percent of the overall variance (β_3 = -1.2), a sample size of 150 will give acceptable power (about 0.8). By setting β_3 to zero, we further our simulation study to assess the type I error rate for a given nominal α = 0.05 using again 500 replicates. The estimated empirical type I error rate is shown in the right most column in Table 1. It can be seen that, although there is a slight fluctuation, the estimates of empirical type I error rate are centered at the nominal α of 0.05. Overall, results from our simulation study indicate that our retrospective model exhibits high power in capturing even relatively small interaction effect with reasonable sample sizes.

Results

The effect of catalase -262C/T promoter polymorphism on human aging phenotypes (cognitive and physical functioning) has been investigated by Christiansen et al. [15]. In this study, a modest protective effect of the T allele on cognitive and physical function was observed although a statistical significance was not reached. Here we apply our retrospective logistic regression model to the data to look for both main and interaction effects on individual's cognitive score (a continuous trait measuring fluency, for- and backward digit span and a modified 12-word learning test) by the genetic variant (T allele carrier = 1, non-carrier = 0) and age-group (equal or above age 65 = 1, below age 65 = 0) in males (N = 789). A combination of the two variants forms four nominal response categories among which non-carriers below age 65 serve as the reference group. In Table 2, we show the parameter estimates for the main and interaction effects by our logistic regression model. The model identified a highly significant effect of age-group that is negatively correlated with individual's cognitive function (RRR = 0.630, p-value = 0.001). Moreover, we found a modest main effect of the T allele (RRR = 0.948, p-value = 0.037) and a modest interaction effect between the T allele and age-group (RRR = 1.083, p-value = 0.033). It is interesting to see that, although the overall effect of allele T reduces carrier's cognitive score, the interaction effect indicates that the effect of the allele is age-dependent which means that the T allele conveys beneficial effect that improves carries' cognitive performances at old ages.

By dichotomizing the cognitive score it is possible to apply the case-only model to assess the interaction effect of allele T and aging on cognitive functioning. To do that, we selected all individuals with cognitive score above 4 (about 24% of the top scores) and defined them as cases (186 individuals). The case-only model gave an odds ratio of 2.386 with a p-value of 0.008 indicating that allele T significantly enhances carrier's cognitive function at old ages. For comparison, we applied our retrospective logistic regression model to the dichotomized cognitive score. Parameter estimates in Table 2 also reveal the negative association with cognitive functioning by aging (RRR = 0.060, p = 0.000) and allele T (RRR = 0.663, p = 0.041). Meanwhile our model also reports a highly significant interaction effect even with exactly the same estimate of the risk parameter (RRR = 2.386, p = 0.009) as from the case-only model (OR = 2.386, p = 0.008) meaning that our retrospective logistic regression model yields valid estimate of the interaction effect. Consistent estimates on the interaction effect by the case-only and our models were also obtained when varying the cut-off for dichotomizing the cognitive score. This is understandable since the case-only model measures the deviation from the multiplication of main effects [9] which is exactly the definition of interaction effect in our model. However, since the maximum likelihood from the dichotomized trait is lower than the continuous trait (Table 2), the model using cognitive score as a continuous trait should be preferred.

Table 2: Parameter estimates for main and interaction effects on cognitive score by the logistic regression model

	Slope	SE	p-value	Risk			logMLK*
				RRR	95% CI		
Continuous							
Age-group	-0.463	0.036	0.000	0.630	0.587	0.676	
Allele T	-0.054	0.026	0.037	0.948	0.901	0.997	
Interaction effect	0.079	0.037	0.033	1.083	1.006	1.164	
							-862.657
Dichotomous							
Age-group	-2.822	0.253	0.000	0.060	0.036	0.098	
Allele T	-0.411	0.202	0.041	0.663	0.446	0.984	
Interaction effect	0.870	0.333	0.009	2.386	1.241	4.588	
							-928.454

*MLK = maximum likelihood

Discussion

The etiology of multifactorial human disease involves complex interactions between numerous environmental factors and alleles of many genes. Efficient statistical tools are demanded for identification of the genetic and environmental variants that affect the risk of disease development. Through example application, we have shown that our retrospective polytomous logistic regression model can capture both main and interaction effects and produce consistent results in estimating the interaction effect as compared with the popular case-only model. The distinct feature in our model is that the disease trait is treated as an independent variable so that our model is capable of accommodating both categorical and continuous traits. Different from the existing models [12,13], no assumption of normal distribution of the trait value is needed in our method. Furthermore, genotype or allele effects can be easily estimated by coding 1 and 0 to carriers and non-carriers to assess dominant or recessive effects.

Since our relative risk ratio is estimated from a retrospective model, it is necessary to study its connection with the relative risk parameter in a general prospective model. In additional file-1, we derive the relationship between the risk parameters in the retrospective model and that in the prospective model when studying a binary disease trait. It is shown that, when the disease is rare, the relative risks in a prospective model can be approximated by the relative risk ratios estimated from our model. This is important because, as long as the disease incidence is low in the population, our model estimates the risk parameters that can be interpreted in terms of trait penetrance as in a prospective model. As shown by equation (6), testing the null hypothesis of $b = 0$ is equivalent to testing $H_o : RRR = 1$. This is also shown by the 95% confidence intervals for the estimated RRRs in Table 2. Since the slope parameters for the main and interaction effects are all statistically different from zero, none of the 95% confidence intervals of RRR covers the null risk of one.

It is necessary to point out that, as in any interaction model, it is critical that the interacting variants be independent. By independent we mean that the interacting variants are not correlated or in association. This is especially relevant in studying gene by gene interactions. It is important to make sure that the two loci under testing are not in LD if they reside on the same chromosome. In case of LD between the two genetic variants, a haplotype-based analysis is more appropriate [8]. Our experience showed that violation of independence can result in unreliable estimates on the risk parameters. Independence between interacting variables is also required by the case-only model to ensure reliable estimates [17]. For case-control studies, if the main interest is interaction effect, the case-only model should be preferred because it is more efficient than the traditional case-control model [18]. Note also that our model is limited to discrete exposure variables when applied to gene-environment interaction although it is no longer a problem for measuring gene by gene interactions because all genotypes are discrete.

Finally, although our model is proposed for genetic association studies (gene by gene or gene by environment), the same model can be applied to study the main and interaction effects of non-genetic variants. Perhaps the biggest advantage of our approach is that it can easily be implemented by using any programming statistical package to fit the multinomial logistic regression model. Considering all these advantages, we hope that our proposed method can be of use for epidemiologists who are interested in studying multifactorial or complex human diseases.

Conclusion

Our proposed retrospective polytomous logistic regression model can be used as a convenient tool for assessing both main and interaction effects in genetic association studies of human complex diseases involving both genetic and non-genetic factors.

Authors' contributions

QT designed the study, analyzed the data and drafted the manuscript. LC and SL contributed to the study design and provided data. CBA and JHZ contributed to the formulation and design of the study. TAK and KC directed the study. All authors approved the final manuscript.

Additional material

Additional file 1

The relationship between risk estimates in the retrospective and the prospective models. In this additional file, we derive the relationship between risk estimates in the retrospective and the prospective models under the assumption of low incidence for a binary disease trait.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2156-8-70-S1.doc]

Acknowledgements

The study was jointly supported by the US National Institute on Aging (NIA) research grant NIAP01AG08761 and the microarray center project under the Biotechnological Research Program financed by the Danish Research Agency and the Danish Medical Research Council.

References

1. Wang WY, Barratt BJ, Clayton DG, Todd JA: **Genome-wide association studies: theoretical and practical concerns.** *Nat Rev Genet* 2005, **6(2):**109-118.
2. Epstein MP, Satten GA: **Inference on haplotype effects in case-control studies using unphased genotype data.** *Am J Hum Genet* 2003, **73:**1316-1329.

3. Witte JS, Gauderman WJ, Thomas DC: **Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs.** *Am J Epidemiol* 1999, **149(8):**693-705.
4. Weinberg CR, Umbach DM: **Choosing a retrospective design to assess joint genetic and environmental contributions to risk.** *Am J Epidemiol* 2000, **152(3):**197-203.
5. Prentice R: **Use of the logistic model in retrospective studies.** *Biometrics* 1976, **32(3):**599-606.
6. Waldman ID, Robinson BF, Rowe DC: **A logistic regression based extension of the TDT for continuous and categorical traits.** *Ann Hum Genet* 1999, **63(Pt 4):**329-340.
7. Zou GY: **Statistical methods for the analysis of genetic association studies.** *Ann Hum Genet* 2006, **70(Pt 2):**262-276.
8. Tan Q, Christiansen L, Christensen K, Bathum L, Li S, Zhao JH, Kruse TA: **Haplotype association analysis of human disease traits using genotype data of unrelated individuals.** *Genet Res* 2005, **86(3):**223-231.
9. Khoury MJ, Flanders WD: **Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls!** *Am J Epidemiol* 1996, **144(3):**207-213.
10. Tan Q, De Benedictis G, Ukraintseva SV, Franceschi C, Vaupel JW, Yashin AI: **A centenarian-only approach for assessing gene-gene interaction in human longevity.** *Eur J Hum Genet* 2002, **10(2):**119-124.
11. Cordell HJ, Todd JA, Hill NJ, Lord CJ, Lyons PA, Peterson LB, Wicker LS, Clayton DG: **Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes.** *Genetics* 2001, **158(1):**357-367.
12. Cordell HJ: **Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans.** *Human Molecular Genetics* 2002, **11:**2463-2468.
13. Hahn LW, Ritchie MD, Moore JH: **Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions.** *Bioinformatics* 2003, **19:**376-382.
14. Glaser B, Nikolov I, Chubb D, Hamshere ML, Segurado R, Holmans P, Moskvina V: **Does interaction between candidate genes influence susceptibility to Rheumatoid arthritis?** *The 15th Genetic Analysis Workshop, Nov. 12-15,2006, St. Pete Beach, FL, USA* .
15. Christiansen L, Petersen HC, Bathum L, Frederiksen H, McGue M, Christensen K: **The catalase -262C/T promoter polymorphism and aging phenotypes.** *J Gerontol A Biol Sci Med Sci* 2004, **59(9):**B886-B889.
16. Hosmer DW, Lemeshow S: *Applied Logistic Regression* John Wiley & Sons, Inc; 2000.
17. Albert PS, Ratnasinghe D, Tangrea J, Wacholder S: **Limitations of the case-only design for identifying gene-environment interactions.** *Am J Epidemiol* 2001, **154(8):**687-693.
18. Piegorsch WW, Weinberg CR, Taylor JA: **Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies.** *Stat Med* 1994, **13:**153-162.