J Hum Genet (2003) 48:520–524
DOI 10.1007/s10038-003-0069-4

ORIGINAL ARTICLE

Noriko Tanaka · Toru Kinoshita · Takashi Asada
Yasuo Ohashi

Log-linear models for assessing gene–age interaction and their application to case-control studies of the apolipoprotein E (apoE) gene in Alzheimer's disease

Received: 20 May 2003 / Accepted: 2 August 2003 / Published online: 23 September 2003

Abstract Case-control studies provide a powerful approach for detecting disease-susceptibility genes or assessing gene–environment interactions. We investigated the situation in which the gene being studied plays a role in several diseases, and the allele frequency among subjects free of the disease of interest consequently decreases with age as subjects die from other diseases. The logistic model is one approach frequently used for analyzing case-control data, but it cannot accommodate this dependence of genotype and age. Using a log-linear model, we therefore proposed a hierarchical procedure that could be used as a valid method for assessing interactions in such situations. We then applied this procedure to observed data on Alzheimer's disease and the apolipoprotein E gene in Japan. We were able to derive an appropriate inference on whether the interaction was a gene–age interaction or merely a bias due to death from other diseases.

Keywords Log-linear model · Case-control study · Allele frequency · Alzheimer's disease

N. Tanaka (✉) · Y. Ohashi
Department of Biostatistics,
School of Health Sciences and Nursing,
The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-0033, Japan
E-mail: tanaka@epistat.m.u-tokyo.ac.jp
Tel.: +81-3-58413520
Fax: +81-3-38142779

T. Kinoshita
Department of Psychiatry,
National Center of Neurology and Psychiatry,
Tokyo, Japan

T. Asada
Department of Psychiatry,
Institute of Clinical Medicine,
University of Tsukuba,
Ibaraki, Japan

Introduction

When researchers verify the effects of genes or gene–environment interactions on common diseases or age-related diseases, such as diabetes, cancer, or coronary heart disease, they usually regard age as a confounding factor and calculate an age-specific or pooled odds ratio or age-adjusted odds ratio by stratification. However, if the investigators aim to study how disease and genetic risk factors or their associations are related to age, they must assess the interactive effects of age and genes or age and the disease. If the gene being studied plays a role in several diseases, and the allele frequency among subjects free of the disease of interest consequently decreases with age as subjects die of other diseases, then it becomes possible that bias due to death from other diseases can result in a quasi-association between age and the gene, and that the Hardy-Weinberg equilibrium would not apply in the control subjects.

One example is that of the apolipoprotein E (apoE) gene at locus 19q13.2, which is associated with Alzheimer's disease (AD) and coronary heart disease (CHD). ApoE has three common alleles: e2, e3, and e4, with varying frequencies in populations around the world (Corbo et al. 1999). Carriers of the e4 allele have a higher risk of CHD and AD than people with the most common genotype, e3/e3, and carriers of the e2-allele have a lower risk (Ou et al. 1998; Wilson et al. 1994; Farrer et al. 1997).

In many studies, apoE genotype has been found to be associated with age of onset of AD. The more e4 alleles there are, the younger the age at disease onset. Onset tends to occur later among persons with the e2/e3 genotype (Corder et al. 1993; Corder et al. 1994; Borgaonkar et al. 1993). This tendency has been reported in various ethnic populations, and Farrer et al. (1997) concluded that the apoE e4 effect is evident at all ages between 40 and 90 years but diminishes after age 70. However, in the meta-analysis of Farrer et al. (1997) and in many other studies, it is reported that the apoE e4 effect is greater among Caucasian than among Japanese older people (especially older than 70 years); this difference is not as obvious in younger populations.

In populations aged 80 years or more, the frequency of occurrence of e4 carriers is lower, and that of e2 is higher, than in younger people (Asada et al. 1996; Gerdes et al. 2000). In the Caucasian population, this difference in e4 allele frequency between older and younger people is greater than in the Japanese population. The prevalence of CHD is much higher in Caucasians (except in the French) than in Japanese, but that of AD is almost the same in both populations (Health and Welfare Statistics Association, 2002).

In light of the above research, we considered the possibility of bias due to death from CHD. Population-based, case-control studies are often conducted to evaluate the association between genes and AD, and the odds ratio is used as a measure. Usually, case and control subjects are sampled from people who are still alive. Hence, if the prevalence or incidence of CHD does indeed affect e4 allele frequencies, then data collected from only AD cases would be biased because of deaths from CHD.

In such a situation where the e4 allele frequency among disease-free subjects decreases with age, we need to assess the parameters governing the control group, because we need to determine whether the estimated interaction effect is due to a decrease in allele frequency in the control group or a decrease in the risk associated with the allele. One approach frequently used to analyze case-control data is the logistic model, but it cannot be used to assess the parameters governing the control group. Umbach and Weinberg (1997) pointed out the above problem and proposed maximum likelihood methods based on log-linear models that explicitly impose the independence of genotype and exposure to assess the gene–environmental interaction.

We used the model proposed by Umbach and Weinberg (1997) to devise a hierarchical procedure using log-linear models to estimate genetic effects and the effects of gene–age interaction and to assess possible bias. Next we applied realistic data on Alzheimer's disease and apoE in Japan to the model. We also briefly discuss the strengths and weaknesses of the method.

Proposed methods

Consider a simple scenario: a population with two genotypes (G), 0–t levels of age group, and no confounders but age. Let G be 1 for the disease-susceptible genotype and 0 for the 'common genotype'; let the T_t (age group = t) be 1 for a group of t subjects and 0 otherwise; let disease status (D) be 1 for cases and 0 for controls.

A typical analysis of such case-control data would use the logistic regression model:

$$\mathrm{logit}(P(D|G,T_t)) = \mu_1 + \alpha_{1t}T_t + \beta_1 G + \gamma_{1t}GT_t$$

Interest is focused on the unknown parameters α_1, β_1, and γ_1, which assess the effects of age level, genotype, and genotype-by-age interactions, respectively: exp (α_1) is the odds ratio relating disease to age among the common genotype subjects; exp (β_1) is the odds ratio relating disease to genotype among the lowest age group; and the interaction parameter exp (γ_1), is the ratio of the odds ratio relating disease to genotype among the age group $T_t = t$ versus that among the age group $T_t = 0$.

$$\log\mu_{dgt} = \mu_0 + \alpha_{0t}T_t + \beta_0 G + \gamma_{0t}GT_t + \mu_1 D + \alpha_{1t}DT_t + \beta_1 DG + \gamma_{1t}DGT_t \quad (1)$$

An alternative and equivalent analysis for this 2×2×t table, employs a linear model for the logarithm of the expected count μ_{dgt} (expected count of disease status D = d, genetic status G = g, and age group $T_t = t$) that fully parameterizes the all cells, namely:

Here, μ_0, $\alpha_{0\,t}$, β_{0t}, and $\gamma_{0\,t}$ parameterize the joint distribution of genotype and age among the controls, and including μ_1 constrains the fitted marginal totals for both cases and controls to match those observed. The parameters of interest are α_1, β_1, and γ_1, as same as those of the logistic model, which assess the effects of age levels, genotypes, and genotype-by-age interactions, respectively. The logistic model and the log-linear model provide the same parametric description of disease risk: α_{1t}, β_{1t}, and γ_{1t} have exactly the same interpretations in both models. The log-linear model, however, explicitly models the control parameters, μ_0, α_{1t}, β_{1t}, and γ_{1t}.

When we have assessed whether or not the genotype is associated with disease, we can then make a comparison between a reduced model and the full model by the maximum likelihood ratio test. Using log-linear models, we propose the following procedure (Fig. 1) for inferences regarding the main genetic effect and its interactions.

Inference regarding the main gene effect

Using the log-linear model, we first test the null hypothesis H_{01}: the genotype is not associated with the disease [thus, $\beta_1 = \gamma_{1\,t} = 0$ for any *t* in model (1)] by a likelihood ratio test that compares model (2)

$$\log\mu_{dgt} = \mu_0 + \alpha_{0t}T_t + \beta_0 G + \gamma_{0t}GT + \mu_1 D + \alpha_{1t}DT \quad (2)$$

and the full model (1).

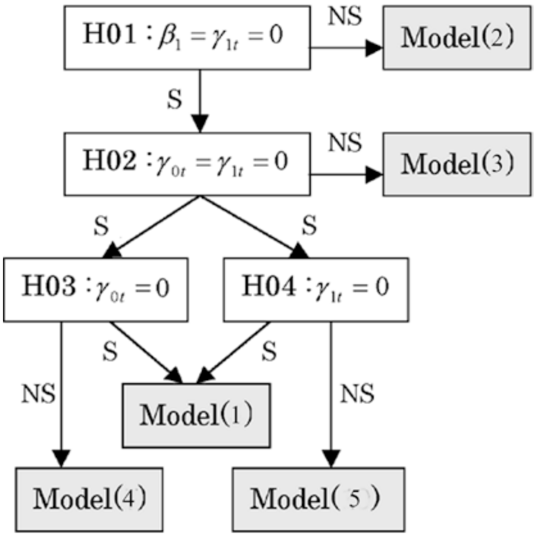

Fig. 1 Proposed hierarchical procedure. In this figure, *S* denotes 'significant' and *NS* denotes 'not significant' of the maximum likelihood ratio test. For example, if one tests the deviance for model (2), namely, H_{01}, and the result is not significant, then the model (2) is considered as the adequate model for the data

Inference regarding the association between genotype and age

If model (2) does not fit the data, we can then make a comparison between a conditional independence model (3)

$$\log \mu_{dgt} = \mu_0 + \alpha_{0t}T_t + \beta_0 G + \mu_1 D + \alpha_{1t}DT_t + \beta_1 DG \quad (3)$$

and the full model (1), to test the null hypothesis H_{02}: the genotype is not associated with age (thus, $\gamma_{0t} = \gamma_{1\,t} = 0$ for any t). In this case, 'conditional independence' means that age and genotype are mutually independent in both the case and the control groups.

If the null hypotheses H_{01} and H_{02} are significantly contradicted—in other words, the genotype is associated with disease and age—we then assess whether the effect of genotype varies according to age.

Inference regarding gene–age interaction

Enforced independence model

To assess gene–environment interactions, Umbach and Weinberg (1997) proposed maximum likelihood methods based on log-linear models that explicitly imposed independence between genotype and exposure, because we can estimate multiplicative gene–environment interactions only when the environmental factor and the genotype are independent in the population. We focused on the likelihood ratio testing of each model, whereas they focused on estimates of the parameters. Here we can consider age to be an environmental factor. If the gene being studied plays a role in only one disease, and the allele frequency among disease-free subjects does decrease greatly with age, then we can loosely consider age and genotype to be independent in the control group. Thus, γ_{0t} is constrained to zero in model (1) to get:

$$\log \mu_{dgt} = \mu_0 + \alpha_{0t}T_t + \beta_0 G + \mu_1 D + \alpha_{1t}DT_t + \beta_1 DG + \gamma_{1t}DGT_t \quad (4)$$

Here, $\mu_{0,}$ α_{0t}, and β_0 parameterize the independent distribution of genotype and age among the controls, and the other four parameters represent the disease risk. In particular, α_0 is the log-odds of age among the controls, and β_0 is the log-odds of having the variant genotype among the controls. The parameters α_1, β_1, and $\gamma_{1\,t}$ assess the effects of age, genotype, and gene–age interaction, respectively, and μ_1 constrains the fitted marginal totals for cases and for controls to match those observed. Using model (4), we can test H_{03}: $\gamma_{0\,t} = 0$ for any t, by comparing models (1) and (4).

Partial association model

We can test whether the degree of association between age and genotype is the same in both the case and the control groups, namely, H_{04}: $\gamma_{1t} = 0$ for any t, by using the log-linear model

$$\log \mu_{dgt} = \mu_0 + \alpha_{0t}T_t + \beta_0 G + \gamma_{0t}GT_t + \mu_1 D + \alpha_{1t}DT_t + \beta_1 DG \quad (5)$$

Interpretation

When H_{04} is not contradicted significantly, we can infer that the genetic effect does not vary with age. Furthermore, if not H_{03}, but H_{04}, is significantly contradicted, then we can interpret the estimator of γ_{1t} in model (4) as the coefficient of gene–age interaction. When both H_{03} and H_{04} are significantly contradicted, however, it is not obvious whether the apparent interaction is a gene–age interaction or merely a bias caused by death from other diseases. In such cases, we have to check which model—model (4) or model (5)—fits the data better. These two models are not nested, so the likelihood ratio test is not directly applicable. However, the two models can be compared by using the ratio of the likelihood for models (4) and (5) given data D, because

$$\begin{aligned} K &= \log\frac{\Pr(D|Model(4))}{\Pr(D|Model(5))} \\ &= \log\frac{\Pr(D|Model(4))}{\Pr(D|Model(1))} - \log\frac{\Pr(D|Model(5))}{\Pr(D|Model(1))} \\ &= \frac{1}{2}\times[D(Model(5) - D(Model(4))] \end{aligned} \quad (6)$$

where D(*Model*(4)) and D(*Model*(5)) are, respectively, the deviances for models (4) and (5), with the same degree of freedom because they have the same number of parameters.

K can reinforce the interpretation of gene–age interaction. If the data fit model (4) better than they fit model (5), then $K \gg 0$. Contrary, if the data fit model (5) better than model (4), then $K \ll 0$.

Application of the proposed method to data on Alzheimer's disease and the apoE gene

To illustrate how to apply the log-linear model and assess bias due to death from other known or unknown diseases, we examined case-control study data on the association between Alzheimer's disease and apoE in Japanese subjects, with additional subject to the published data (Asada et al. 2000). The data in Table 1 show that the disease group consists of people aged 45 to 91 years with Alzheimer's disease, and the controls are healthy people aged 45 to 93 years. The genotype categories are presence (either heterozygous or homozygous) or absence of the apoE allele e4, and there are six age categories.

We calculated the logistic estimate of the common odds ratio in each age group and the age-adjusted odds ratio, and we performed a Breslow-Day test for homogeneity of the odds ratios. The results of these tests are summarized in Table 1. They show that there is evidence that age modifies the risk of AD related to the presence of the e4 allele. In addition, the test for homogeneity was significant

Table 1 Data of subjects classified by age, Alzheimer's disease (AD) and the variant allele of apoE gene. *CI* confidence interval, + genotype with e4, − genotype without e4

Age group (years)	Odds ratio	(95% CI)	Control		AD cases		Total
			Genotype				
			−	+	−	+	
< 60	1.8	(1.1–2.9)	109	54	79	71	313
61–65	4.0	(2.1–7.4)	65	31	29	55	180
66–70	3.7	(2.3–6.0)	90	41	61	103	295
71–75	3.6	(2.2–6.0)	82	32	77	108	299
76–80	2.6	(1.4–4.7)	57	20	82	75	234
81–85	1.6	(0.6–4.3)	19	8	39	27	93
86 >	0.5	(0.1–2.1)	9	10	9	5	33

Age-adjusted odds ratio (95% CI) = 2.7 (2.1–3.4)
Test for homogeneity $\varkappa^2 = 14.2$ ($p < 0.03$)

Table 2 Log likelihood and residual deviance for models fitted to the Alzheimer's disease (AD) case-control data in Table 1. *DF* degree of freedom

Model	Log likelihood	Hypothesis	Deviance	DF	*P* >
Equation (1)	4528.3				
Equation (2)	4469.2	H_{01}	118.2	7	0.00
Equation (3)	4510.9	H_{02}	34.9	12	0.00
Equation (4)	4527.3	H_{03}	2.0	6	0.93
Equation (5)	4522.1	H_{04}	12.4	6	0.05

H_{01} The genotype is not associated with the disease
H_{02} The genotype is not associated with age
H_{03} The genotype is not associated with age in controls
H_{04} The degree of association between age and genotype is the same in both groups

(Breslow-Day $P < 0.03$). From the results of our analyses, we suggest that the risk of the apoE gene varies with age. However, we cannot conclude whether the risk truly varies with age or whether this was merely a bias due to variation in genotype frequency with age in the population, such that the frequency of genotypes with the e4 allele in control subjects decreased slightly with age.

Table 2 presents the results of application of the logistic models and the log-linear models to this data, according to our proposed procedure. We used SAS (Ver. 8.1) to fit all models. H_{01} and H_{02} were significantly contradicted, so that the genotype was associated with AD and age in our data (Table 2). However, the likelihood ratio test of hypothesis H_{03} that $\gamma_{0\,t}=0$ gave $P=0.93$, and that of hypothesis H_{04} that $\gamma_{1t}=0$ gave $P=0.05$. The *K* value was 5.2. These results suggest that the effect of the apoE gene varies with age. Furthermore, the results show that there was very little evidence of age-related bias in genotype frequencies in the control group in this study.

Discussion

The hierarchical procedure proposed herein provides an answer to the annoying problem of how to avoid misreading the results of analyses of samples from populations with variations in allele frequency. Logistic models are applied to most case-control study data. Our result shows that this procedure using log-linear models can not only measure the association between a gene and disease but can also assess the bias of the estimator. Logistic models cannot do this.

We used the model proposed by Umbach and Weinberg (1997), but we focused particularly on likelihood ratio testing of each model, whereas they focused on estimates of parameters. Furthermore, we proposed a measure, '*K*', to compare models (4) and (5) for assessment of age-related bias. As mentioned above, *K* can reinforce the interpretation of gene–age interaction, but its standard calibration—like that of Akaike's Information Criterion (AIC) (Akaike 1970) or Bayes factor—cannot be defined explicitly. We therefore needed to interpret the results in the light of both the *K* value and hypothesis testing.

Table 3 Log likelihood and residual deviance for models fitted to the hypothetical case-control data.*DF* degree of freedom

Model	Log likelihood	Hypothesis	Deviance	DF	*P* >
Equation (1)	94131.4				
Equation (2)	93655.3	H_{01}	476.1	7	0.00
Equation (3)	93608.5	H_{02}	522.9	12	0.00
Equation (4)	94018.1	H_{03}	113.3	6	0.00
Equation (5)	94118.7	H_{04}	12.7	6	0.05

H_{01} The genotype is not associated with the disease
H_{02} The genotype is not associated with age
H_{03} The genotype is not associated with age in controls
H_{04} The degree of association between age and genotype is the same in both groups

We use as our example the apoE gene that is associated with variations in the risk of AD and CHD; however, there does not seem to be much of a bias component in our data. We thought that this might be because of a lower rate of CHD death in Japan. Table 3 shows the results of application of the logistic models and the log-linear models to the hypothetical data based on the result of Fig. 2 in Gerdes et al (2000) and Fig. 4 in Farrer et al (1997). We assumed the data is sufficiently large to detect gene–age interaction in Caucasian populations, so we set each age group as having 1,200 cases and 1,200 controls. The result is that both H_{03} and H_{04} are significantly contradicted, so it is not obvious whether the apparent interaction is a gene–age interaction or merely a bias caused by death from other diseases. To compare the model (4) and the model (5), we calculated *K* and get $K=-50.6$. Then, we can infer that the data fit model (5) better than model (4), and it suggests that the effect of gene–age interaction might be biased.

We applied our method to data on AD and the apoE gene. Not only in the case of AD but also in many common diseases, there is potential for the gene being studied to play a role in several other diseases, and the allele frequency among subjects free of the study disease will consequently decrease with age as patients die from these other diseases (example: breast and ovarian cancer and the BRCA1 gene). We suggest that our method could be extended to this problem of other diseases and genes. In addition, this model can be extended to any number of loci, any number of alleles, or any number of age categories and environmental risk factors, the only major practical limitation being the sample size needed to estimate an increasing number of effects with high precision.

Our procedure can determine only the likelihood of bias of the estimator caused by death from other diseases. If bias is likely and the investigator would like to assess the bias quantitatively, then he or she will have to conduct a prospective study and apply a competing risks analysis. However, prospective studies, such as cohort studies, are difficult to conduct because of time and cost constraints. If a case-control study is to be used instead to infer gene–age interactions, the cases and controls must be sampled adequately and as much data as possible on other diseases in which the gene is involved must be collected.

Acknowledgements We thank Drs. Shuji Hashimoto, Takuhiro Yamaguchi, Jun Ohashi, and Seiichiro Yamamoto for their helpful discussions.

References

Akaike H (1970) Statistical predictor identification. Ann inst Stat Math 22:203–217

Asada T, Kariya T, Kinoshita T, Asaka A (1996) Apolipoprotein E allele in centenarians. Neurology 46: 1484–84

Asada T, Motonaga T, Yamagata Z, Uno M and Takahashi K (2000) Associations between retrospectively recalled napping behavior and later development of Alzheimer's Disease: Associations with APOE genotypes. Sleep 23:629–634

Borgaonkar DS, Schmidt LC, Martin SE, et al (1993) Linkage of late-onset Alzheimer's disease with apolipoprotein E type 4 on chromosome 19. Lancet 342:625

Corbo RM, Scacchi R (1999) Apolipoprotein E (ApoE) allele distribution in the world. Is APOE*4 a 'thrifty' allele? Ann Hum Genet 63:301–310

Corder EH, Saunders AM, Strittmatter WJ, et al (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261:921–923

Corder EH, Saunders AM, Risch NJ, et al (1994) Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. Nat Genet 7:180–184

Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, Pericak-Vance MA, Risch N, van Duijn CM (1997) Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. JAMA 278: 1349–1356

Gerdes LU, Jeune B, Ranberg KA, Nybo H, Vaupel JW (2000) Estimation of apolipoprotein e genotype-specific relative mortality risks from the distribution of genotypes in centenarians and middle-aged men: Apolipoprotein e gene is a "frailty gene", not a "longevity gene". Genet Epidemiol 19: 202–210

Health and Welfare Statistics Association (2002) Journal of Health and Welfare Statistics

Ou T, Yamakawa-Kobayashi K, Arinami T, Amemiya H, Fujiwara H, Kawata K, Saito M, Kikuchi S, Noguchi Y, Sugishita Y, Hamaguchi H (1998) Methylenetetrahydrofolate reductase and apolipoprotein E polymorphisms are independent risk factors for coronary heart disease in Japanese: a case-control study. Atherosclerosis 137: 23–28

Umbach DM, Weinberg CR (1997) Designing and analyzing case-control studies to exploit independence of genotype and exposure. Stat Med 16:1731–1743

Wilson PW, Myers RH, Larson MG, Ordovas JM, Wolf PA, Schaefer EJ (1994) Apolipoprotein E alleles, dyslipidemia, and coronary heart disease. The Framingham Offspring Study. JAMA 272:1666–1671