

对数线性模型在病例-父母对照研究中的应用

朱明 聂绍发 严薇荣

【导读】 介绍应用对数线性模型分析病例-父母对照研究设计的方法。以亚甲基四氢叶酸还原酶基因(MTHFR)C677T与唇腭裂的关联研究为例,应用对数线性模型分析母亲、子代基因及其交互作用与唇腭裂的关系。结果显示变异型纯合子母亲的子代发生唇腭裂的风险低于野生型纯合子母亲的子代, $S_2=0.43$ (95%CI:0.19~0.95),未发现唇腭裂与子代基因及母子交互作用相关。应用对数线性模型分析病例-父母对照研究设计的方法尤其适用于妊娠期疾病与源于胚胎时期疾病等的病因学研究。

【关键词】 病例-父母对照研究;对数线性模型;基因多态性;母子交互作用

Application of log-linear model in the case-parent triad study ZHU Ming, NIE Shao-fa, YAN Wei-rong. Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China
Corresponding author: YAN Wei-rong, Email: weirongy@gmail.com
This work was supported by grants from the National Natural Science Foundation of China (No. 30700681/H2610, 81172679/H2605).

【Introduction】 We were introducing a log-linear model for case-parent triad study. Data from a previous study of the association between MTHFR C677T and cleft lip with or without cleft palate (CL/P) was analyzed, in order to investigate the maternal effect, fetal effect and their interaction, using a series of log-linear models. Our results showed that mothers who were carrying two copies of MTHFR C677T variant alleles appeared to have reduced the risk of CL/P in offspring, comparing to those with homozygous of wild-type allele. With $S_2=0.43$ (95% CI: 0.19-0.95). No significant association was found for fetal genotype and maternal-fetal interaction with CL/P. Log-linear model method seemed to be useful in the estimation of maternal effect, fetal effect and maternal-fetal interaction, in the case-parent triad study design. This approach showed specific benefit in studies that related to genetic effects on complex diseases such as pregnancy complications and diseases originated from fetus.

【Key words】 Case-parent triad; Log-linear model; Gene polymorphisms; Maternal-fetal interaction

病例-父母对照研究是指以无关联的病例及其父母作为研究对象,对患者及其父母进行基因分型,收集一般信息和环境暴露资料,以检测与疾病发生有关联的遗传标记或与其相邻位点上连锁不平衡的等位基因,评价环境因素与基因的交互作用^[1]。目前常用于病例-父母对照研究的分析方法有:单体型相对风险(HRR),基因型相对风险(GRR)和传递/不平衡检验(TDT)等。这些方法虽可用于病例-父母对照研究的资料分析,但仅从子代角度考虑易感基

因与疾病的关系。对于某些疾病,如出生缺陷等源于胚胎时期的疾病,母亲的易感基因可通过影响宫内胎儿的发育环境对子代的发病产生影响^[2];先兆子痫等妊娠期疾病,胎儿的易感基因可能通过作用于母体从而影响母亲的发病^[3];母子间交互作用也可能与疾病发生相关^[4]。因此探讨这些疾病的病因时,应同时分析子代、母亲易感基因以及母子交互作用在疾病发生中的作用。应用对数线性模型分析此类研究设计的数据具有独特优势^[5],本文将通过实例分析介绍其原理和方法。

基本原理

应用对数线性模型分析病例-父母对照研究设计的资料,通过检测病例家庭中病例-父母基因型频

DOI: 10.3760/cma.j.issn.0254-6450.2012.04.020

基金项目:国家自然科学基金(30700681/H2610, 81172679/H2605)

作者单位:430030 武汉,华中科技大学同济医学院公共卫生学院流行病学与卫生统计学系

通信作者:严薇荣, Email: weirongy@gmail.com

数分布,采用极大似然法对实际频数与人群中满足孟德尔遗传定律的理论频数进行迭代拟合,分别估计母亲、子代的基因型效应及其交互作用在疾病中的作用。病例-父母基因型(case-parent triad genotype)共15种(表1第1列)。按父母携带的易感等位基因数将婚配类型(mating type)分6种(表1第2列)。随机婚配(random mating)指父母的婚配类型与基因型无关,父母的基因型分布对称(mating symmetry),如杂合子母亲和纯合子父亲婚配的概率与纯合子母亲和杂合子父亲婚配的概率相同。基于婚配类型分析可避免由人群分层现象或不满足H-W平衡产生的偏倚^[5]。根据Bayes定理,人群中病例-父母基因型的概率为该婚配类型的概率与父母按孟德尔遗传定律传递给子代易感等位基因概率的乘积, $Pr[M, F, C] = Pr[C | M, F]Pr[M, F]$, 这15种病例-父母基因型概率之和为1。病例家庭中病例-父母基因型分布概率为 $Pr[M, F, C | D] = Pr[D | M, F, C]Pr[C | M, F] Pr[M, F]/Pr[D]$, 即病例家庭病例-父母基因型的概率为人群中父母婚配类型的概率、父母按孟德尔遗传定律的传递给子代易感基因的概率与相应相对危险度的乘积。进行该分析需满足两个前提条件:第一,孟德尔遗传定律;第二,随机婚配。利用病例-父母对照研究可分为以下5种情况进行分析:

1. 子代基因型效应:分析子代的基因型效应时 $Pr[D | M, F, C] = Pr[D | C]$ (病例-父母基因型分布概率仅与子代基因型的相对危险度相关,提供有效信息的婚配类型为2、4、5)。病例-父母基因型期望频数为:

$$Exp[\omega_n + \beta_1 I_{|C=1|} + \beta_2 I_{|C=2|} + \ln(2) I_{|M=F=C=1|}] \quad (1)$$

$\omega_n = \ln(\gamma_n)$ 为婚配类型参数;子代易感等位基因数为1时, $I_{|C=1|}$ 取值为1,否则为0, $I_{|C=2|}$ 的取值与之同理; $\ln(2) I_{|M=F=C=1|}$ 指当 $M=F=C=1$ 时, $I_{|M=F=C=1|}$ 为1,由于子代以相同的概率分别从父母遗传得到一个易感等位基因,故系数为2。对数线性模型通过最大似然法对理论频数与观察频数进行迭代拟合,估计参数 β_1 和 β_2 , $R_1 = \exp(\beta_1)$, $R_2 = \exp(\beta_2)$,按照方差分析思想进行参数检验,并根据标准误计算置信区间(CI)。

2. 母亲基因型效应:如果疾病的发病风险仅与母亲携带的易感基因相关,提供有效信息的婚配类型为2、3、5,其期望频数为:

$$Exp[\omega_n + \alpha_1 I_{|M=1|} + \alpha_2 I_{|M=2|} + \ln(2) I_{|M=F=C=1|}] \quad (2)$$

表1 病例-父母对照研究理论频数分布

病例-父母基因型分布(MFC)	婚配类型	理论频数				
		1	2	3	4	5
222	1	$R_2\gamma_1$	$S_2\gamma_1$	$R_2S_2\gamma_1$	$I_M I_F R_2 S_2 \gamma_1$	$\mu_{22}\gamma_1$
212	2	$R_2\gamma_2$	$S_2\gamma_2$	$R_2S_2\gamma_2$	$I_M I_F R_2 S_2 \gamma_2$	$\mu_{22}\gamma_2$
211	2	$R_1\gamma_2$	$S_2\gamma_2$	$R_1S_2\gamma_2$	$I_M R_1 S_2 \gamma_2$	$\mu_{21}\gamma_2$
122	2	$R_2\gamma_2$	$S_1\gamma_2$	$R_2S_1\gamma_2$	$I_M I_F R_2 S_1 \gamma_2$	$\mu_{12}\gamma_2$
121	2	$R_1\gamma_2$	$S_1\gamma_2$	$R_1S_1\gamma_2$	$I_M R_1 S_1 \gamma_2$	$\mu_{11}\gamma_2$
201	3	$R_1\gamma_3$	$S_2\gamma_3$	$R_1S_2\gamma_3$	$I_M R_1 S_2 \gamma_3$	$\mu_{01}\gamma_3$
021	3	$R_1\gamma_3$	γ_3	$R_1\gamma_3$	$I_F R_1 \gamma_3$	$\mu_{01}\gamma_3$
112	4	$R_2\gamma_4$	$S_1\gamma_4$	$R_2S_1\gamma_4$	$I_M I_F R_2 S_1 \gamma_4$	$\mu_{12}\gamma_4$
111	4	$2R_1\gamma_4$	$2S_1\gamma_4$	$2R_1S_1\gamma_4$	$(I_M/I_F)R_1S_1\gamma_4$	$2\mu_{11}\gamma_4$
110	4	γ_4	$S_1\gamma_4$	$S_1\gamma_4$	$S_1\gamma_4$	$\mu_{10}\gamma_4$
101	5	$R_1\gamma_5$	$S_1\gamma_5$	$R_1S_1\gamma_5$	$I_M R_1 S_1 \gamma_5$	$\mu_{01}\gamma_5$
100	5	γ_5	$S_1\gamma_5$	$S_1\gamma_5$	$S_1\gamma_5$	$\mu_{10}\gamma_5$
011	5	$R_1\gamma_5$	γ_5	$R_1\gamma_5$	$I_F R_1 \gamma_5$	$\mu_{01}\gamma_5$
010	5	γ_5	γ_5	γ_5	γ_5	γ_5
000	6	γ_6	γ_6	γ_6	γ_6	γ_6

注:理论频数1代表子代基因型效应、2代表母亲基因型效应、3代表母亲与子代基因型效应、4代表父母基因印记效应、5代表母子基因交互作用

分别计算 $\alpha_1, \alpha_2, S_1 = \exp(\alpha_1), S_2 = \exp(\alpha_2)$ 及其CI,并进行拟合优度检验。

3. 同时考虑母亲和子代的基因型效应:病例-父母基因型期望频数为:

$$Exp[\omega_n + \beta_1 I_{|C=1|} + \beta_2 I_{|C=2|} + \alpha_1 I_{|M=1|} + \alpha_2 I_{|M=2|} + \ln(2) I_{|M=F=C=1|}] \quad (3)$$

基于最大似然法进行参数估计、检验并计算其CI。值得注意的是应用对数线性模型同时分析母亲与子代的基因型效应,不会受母子遗传关系的影响而致偏倚。

4. 亲本印记(parental imprinting):指一对同源染色体或等位基因之间存在功能或活性的差异,其表达活性取决于亲本来源。印记基因在胚胎生长、行为发生、肿瘤生成、神经发育等过程中起着重要的作用。分析易感等位基因是否存在亲本印记效应^[6],其理论频数为:

$$Exp[\omega_n + \beta_1 I_{|C=1|} + \beta_2 I_{|C=2|} + \alpha_1 I_{|M=1|} + \alpha_2 I_{|M=2|} + \varepsilon_F I_{|F|} + \varepsilon_M I_{|M|} + \ln(\eta) I_{|M=F=C=1|}] \quad (4)$$

$I_F = \exp(\varepsilon_F)$ 和 $I_M = \exp(\varepsilon_M)$ 表示子代的易感等位基因分别来自父亲、母亲的相对危险度, $\eta = \exp(\varepsilon_F) + \exp(\varepsilon_M)$,由于 $I_{|C=1|} + 2 I_{|C=2|} = I_{|F|} + I_{|M|}$,导致该模型参数间存在共线性,解决方法为减少参数,如单独分析 I_M 或 I_F ,或假定遗传模型为显性或隐性等。

5. 母子交互作用:研究表明母子的交互作用可影响母亲或子代的患病风险。对模型(3)进行衍变,

理论频数为(表 1 第 8 列):

$$\begin{aligned} & \text{Exp}[\omega_n + \mu_{22}I_{|M=2,C=2|} + \mu_{21}I_{|M=2,C=1|} + \mu_{12}I_{|M=1,C=2|} \\ & + \mu_{11}I_{|M=1,C=1|} + \mu_{10}I_{|M=1,C=0|} + \mu_{01}I_{|M=0,C=1|} \\ & + \ln(2)I_{|M=F=C=1|}] \end{aligned} \quad (5)$$

其中 $\mu_{00}=1$, 根据最大似然法对理论频数与实际频数进行迭代拟合, 比较模型(5)与模型(3)的最大似然差值, 估计 6 个参数及其 CI 以评价是否存在母子交互效应及大小。

以上几种分析也可根据遗传方式变换为显性 ($1 < R_1 = R_2$) 和隐性遗传模型 ($1 = R_1 < R_2$) 等进行分析。

实例分析

以唇腭裂与亚甲基四氢叶酸还原酶基因 (MTHFR) C677T 多态性的关系研究为例^[7](表 2), 应用对数线性模型对病例-父母对照研究设计的资料进行分析。采用 LEM 软件(程序语言参考 Weinberg 提供的相关资料)^[8, 9]。

表 2 173 例唇/腭裂患者 MTHFR C677T 基因多态性的病例-父母基因型频数分布

病例-父母基因型分布(MFC)	婚配类型	实际频数
222	1	0
212	2	1
211	2	2
122	2	2
121	2	3
201	3	6
021	3	14
112	4	8
111	4	18
110	4	12
101	5	16
100	5	9
011	5	17
010	5	19
000	6	46

分析结果表明母亲 MTHFR C677T 的基因型为变异型纯合子时, 与野生型纯合子母亲相比, 其子代的发病风险降低, $S_1=0.43$ (95% CI: 0.19 ~ 0.95), 而母亲为杂合子时未见关联, $S_2=0.70$ (95% CI: 0.43 ~ 1.13)(表 3)。分析未发现子代的基因型与唇腭裂发

病存在关联, $R_1=1.03$ (95% CI: 0.68 ~ 1.56), $R_2=0.76$ (95% CI: 0.36 ~ 1.63), 见表 3, 且母子间交互作用无统计学意义(表 4)。

讨 论

病例-父母对照研究通过收集病例家庭父母及子代的遗传信息和环境暴露资料, 分析易感基因是否与疾病相关并评价基因与环境之间的交互作用。由于病例-父母对照研究能避免由人群分层现象产生的偏倚, 使其在复杂性疾病的研究中得到广泛应用。而源于胚胎时期疾病和妊娠期疾病等复杂性疾病, 母亲和子代间可能通过免疫反应、母子基因型不匹配等机制影响疾病发病风险。目前常用的分析方法仅考虑到子代与疾病关系, 忽略母亲及其与子代交互作用对疾病的影响, 使其应用存在一定的局限性。应用对数线性模型分析病例-父母对照研究设计的资料具有以下优点: ①可分别分析母子基因型效应、交互效应以及是否存在亲本印记效应等; ②分析子代的遗传效应时, 考虑到对数线性模型的自由度为 2, TDT 为 1, 其统计学效率高于 TDT, 仅当 $R_2=R_1^2$ 时, TDT 的统计效能高于对数线性模型^[5]。

采用对数线性模型分析病例-父母对照研究也有一定的局限性, 主要表现为: ①需检测父母子三者的基因型。要提高分析的统计学效能和可靠性, 需尽量保证数据的完整^[10]; ②需满足两个前提条件——孟德尔遗传定律和随机婚配; 单独采用病例-父母对照研究设计无法检验这两个条件, 而结合病例-对照研究设计, 如加入对照父母的样本^[11]、对照人群^[12]或对照-母亲的样本^[13], 可检验这两个前提条件, 同时增加研究的统计学效能。

应用对数线性模型分析病例-父母对照研究设计的数据, 常用的软件有 SAS、R 和 GLIM 等。此外一些特殊的软件可用于分析这类数据, 如使用简便的 LEM, 适用于分析 MFG 效应的 Mendel 软件包等以及可用于分析多种设计的数据的 EMIM 软件包等^[14, 15]。

目前国内应用对数线性模型分析病例-父母对照研究的实例较少, 而如妊娠期疾病、源于胚胎时期的疾病、遗传性疾病以及如精神分裂症等复杂性疾

表 3 母亲、子代 MTHFR C677T 与唇腭裂的关系

基因型效应模型	InL ₀₀	InL _{0n}	ν	P 值	R ₁ 值(95%CI)	R ₂ 值(95%CI)	S ₁ 值(95%CI)	S ₂ 值(95%CI)
子代	-399.58	-399.21	2	0.70	1.03 (0.68 ~ 1.56)	0.76 (0.36 ~ 1.63)	-	-
母亲	-399.58	-396.69	2	0.05	-	-	0.43 (0.19 ~ 0.95)	0.70 (0.43 ~ 1.13)
母、子	-399.58	-396.32	4	0.16	1.03 (0.68 ~ 1.56)	0.76 (0.36 ~ 1.63)	0.43 (0.19 ~ 0.95)	0.70 (0.43 ~ 1.13)

表 4 母亲、子代 MTHFR C677T 基因交互作用与唇腭裂的关系

母亲基因型	子代基因型	OR 值(95%CI)
CC	CC	1.00
TT	TT	0.70(0.40 ~ 1.21)
TT	CT	0.92(0.47 ~ 1.78)
CT	TT	1.00(0.69 ~ 1.46)
CT	CT	0.78(0.57 ~ 1.06)
CT	TT	1.05
TT	CT	0.95(0.69 ~ 1.31)

注: $\ln L_{\text{母}} = -396.32$, $\ln L_{\text{子}} = -396.48$, $\nu = 1$, $P = 0.57$

病,母亲和子代的易感基因可能均与疾病的发生相关,此类实例很少见有国内学者的公开报道。虽然这种方法目前主要应用于分析单基因多态性与疾病的关联研究,随着全基因组关联研究(GWAS)方法的应用,对数线性模型也可对 MI-GWAS 数据进行分析^[16]。

参 考 文 献

[1] Yi HG, Chen F. Case-parental control study. Chin J Epidemiol, 2004, 25(5): 439-444. (in Chinese)
 易洪刚,陈峰. 病例父母亲对照研究. 中华流行病学杂志, 2004, 25(5): 439-444.

[2] Gjessing HK, Lie RT. Case-parent triads: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes. Ann Hum Genet, 2006, 70(Pt 3): 382-396.

[3] Petry CJ, Ong KK, Dunger DB. Does the fetal genotype affect maternal physiology during pregnancy? Trends Mol Med, 2007, 13(10): 414-421.

[4] Liang M, Wang X, Li J, et al. Association of combined maternal-fetal TNF-alpha gene G308A genotypes with preterm delivery: a gene-gene interaction study. J Biomed Biotechnol, 2010: 396184.

[5] Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads". Am J Epidemiol, 1998, 148(9): 893-901.

[6] Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act

either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet, 1998, 62(4): 969-978.

[7] Jugessur A, Wilcox AJ, Lie RT, et al. Exploring the effects of methylenetetrahydrofolate reductase gene variants C677T and A1298C on the risk of orofacial clefts in 261 Norwegian case-parent triads. Am J Epidemiol, 2003, 157(12): 1083-1091.

[8] van Den Oord EJ, Vermunt JK. Testing for linkage disequilibrium, maternal effects, and imprinting with (In)complete case-parent triads, by use of the computer program LEM. Am J Hum Genet, 2000, 66(1): 335-338.

[9] LEM scripts. [2012-1-19]. <http://www.niehs.nih.gov/research/resources/software/biostatistics/lem/index.cfm>.

[10] Weinberg CR. Allowing for missing parents in genetic studies of case-parent triads. Am J Hum Genet, 1999, 64(4): 1186-1193.

[11] Weinberg CR, Umbach DM. A hybrid design for studying genetic influences on risk of diseases with onset early in life. Am J Hum Genet, 2005, 77(4): 627-636.

[12] Epstein MP, Veal CD, Trembath RC, et al. Genetic association analysis using data from triads and unrelated subjects. Am J Hum Genet, 2005, 76(4): 592-608.

[13] Vermeulen SH, Shi M, Weinberg CR, et al. A hybrid design: case-parent triads supplemented by control-mother dyads. Genet Epidemiol, 2009, 33(2): 136-144.

[14] EMIM: Estimation of Maternal, Imprinting and interaction effects using Multinomial modelling. 2011.

[15] Ainsworth HF, Unwin J, Jamison DL, et al. Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. Genet Epidemiol, 2011, 35(1): 19-45.

[16] Agopian A, Mitchell LE. MI-GWAS: a SAS platform for the analysis of inherited and maternal genetic effects in genome-wide association studies using log-linear models. BMC Bioinformatics, 2011, 12: 117.

(收稿日期: 2011-10-23)

(本文编辑: 万玉立)