技术与方法 文章编号:0253-3626(2010)11-1697-03

应用 Logistic 回归探讨 MDR 交互的具体方式与效应大小

石修权[1],王增珍[2]

(1. 遵义医学院预防医学系流行病与卫生统计教研室,遵义 563003;2. 华中科技大学同济医学院公卫学院流行病与卫生统计学系,武汉 430030)

【摘 要】目的:探讨在分析基因、环境交互作用时,用 Logistic 回归分析多因子降维(Multifactor dimensionality reduction,MDR)所得交互模型的具体形式和效应大小是否可行。方法:以多环芳烃(Polycyclic aromatic hydrocarbons,PAHs)与肺癌风险的病例对照研究为例,采用 MDR 分析其基因 - 基因及基因 - 环境的交互效应,找出最优模型并用 Logistic 回归对交互效应的形式和效应大小进行补充分析。结果:在 MDR 发现的最佳三阶交互模型的基础上发现"r_s4646782*r_s2158041"基因多态性具有相乘效应,"r_s4646782+r_s2158041+smoking year"具有相加效应,前者乘积及后者的和每增大 1 个取值单位,其肺癌发病风险分别增加 0.097 和 0.199 倍。结论:结果提示 Logistic 回归可用于对 MDR 法所得到的交互模型的具体形式和效应大小进行补充分析。

【关键词】多因子降维法;交互作用;Logistic 回归

【中国图书分类法分类号】R181.25 【文献标识码】A 【收稿日期】2010-04-08

Application of Logistic regression to explore the form and effect size of the interaction in MDR

SHI Xiu-quan, et al

(*Department of Epidemiology and Health Statistics, Falcclty of Preventive Medicine, Zunyi Medical College*)

【Abstract】*Objective*: To explore how to use Logistic regression model to analyze the form and effect size of the gene-gene-environment interaction after MDR analysis. *Methods*: Data of a case-control study which focused on the relationship of PAHs and lung cancer were used. After the gene-gene interactions as well as the gene-environment interactions models were built by the MDR software, logistic regression was carried out to further search the form and the effect size of the interaction as a supplement to MDR. *Results*: The combination of "r_s4646782*r_s2158041" gene polymorphisms had the multiplication effect, and "r_s4646782+r_s2158041+smoking year" had the additive effect; lung cancer risk were increased 0.097 or 0.199 times as values of the product or the additive of the interaction terms were increased one unit, respectively. *Conclusions*: Our findings suggested that the logistic regression could be used as a supplementary method after MDR to analyze the form and the effect size of the interaction.

【Key words】MDR; Interaction; Logistic regression

近年来,在分析与环境和遗传因素均有关的多因子病时,发现其不但受环境暴露、遗传易感性及个体因素等的单独影响,各种因素间还可能存在着交互作用,如基因 - 基因和基因 - 环境交互作用。但在分析基因 - 基因和基因 - 环境交互作用这些复杂关系上,传统的统计学方法存在较多局限,如很难处理数据的共线性,对数据的假设过于严格,分析前提比较多等。

而由 Ritchie 等[1]于 2001 年提出的多因子降维(Multifactor dimensionality reduction,MDR)则在一定程度上克服了上述某些缺点,它可在有限样本含量条件下,分析高维数据之间的交互作用。MDR 是一种非参数、无需知道很明确的遗传模式和交互作用的模式(线性或非线性模型,加法还是乘法模型)的分析方法,在一定程度上解决了多个因子之间交互作用研究时的"维度困扰"问题[2]。

但 MDR 自身也存在一些缺陷,如 MDR 法很难得到最终的模型,如不知交互的是全部还是部分因子?也不知交互的具体形式[3]。本研究从寻找解决这一缺点的简单易行的方法出发,采用非条件 Logistic 回归模型对交互的具体形式、效应大小等进行分析,探讨 Logistic 回归在补充分析 MDR 法遗留的交互效应的具体问题等方面的可行性。

作者介绍:石修权(1978-),男,副教授,博士,
研究方向:环境健康危害的评价。

通讯作者:王增珍,女,教授,E-mail: wzzh@mails.tjmu.edu.cn。

1 对象与方法

1.1 资料来源

研究对象均来自南方某市，其中 500 例肺癌病例来自该市肿瘤医院和医大附属医院，均经过组织学方法的明确诊断，517 例对照则源自同期周边社区进行的健康筛查的无肿瘤个体，两个组人群在年龄、性别、居住区域等方面进行了成组匹配。经收集研究对象的生活习惯(烟、酒史)、家族肿瘤史等信息后，采集血样抽提 DNA，进行可能影响 PAHs 代谢的 16 种基因的 65 个位点单核苷酸多态性(Single nucleotide polymorphism，SNP)的检测[4,5]。

1.2 分析方法

1.2.1 分析的基本步骤 将 MDR 方法建模得出的最优基因－环境交互模型，用非条件 logistic 回归模型探讨交互的具体形式、阶数与效应大小。

1.2.2 统计分析软件 基因－环境交互分析采用 MDR Software 程序包(2.0_Beta 版)以及 MDR Permutation Testing Module(v0.4.6 版)。其余分析均采用 SPSS16.0 进行，logistic 逐步回归入选水准 0.05，剔除水准 0.10。

2 结 果

2.1 MDR 最佳模型的确定与总交互效应大小的估计

将初筛得到的 14 个变量、1017 条记录资料导入 MDR 软件，经拟合 1～6 阶交互的所有模型，MDR 软件给出了 1～6 阶不同交互情况下的模型见表 1。

选择其中包含 r_s4646782，r_s2158041，smoyear 3 个因子的交互模型为最佳模型，因其 Cross-validation Statistics 为：Training Balanced Accuracy：0.619 0，Testing Balanced Accuracy：0.601 3(经 1 000 次置换检验，按 α=0.001 时的界值为 0.585 9，故 $P<0.001$，且更改随机种子，也可得到一致的结果)，十重交叉验证的一致率为 10/10，明显优于其他模型。

按照不同的基因－环境交互组合将研究对象重新划分为肺癌发病的“高危、低危”人群，此时肺癌发病风险交互效应的危险度估计 OR 值 2.639，95%CI(2.047～3.403)。

2.2 非条件 logistic 回归对 MDR 交互效应的补充分析

按照 MDR 得到的最优模型里面包含的 3 个因子 r_s4646782，r_s2158041 和 smoyear，采用非条件 Logistic 逐步回归，逐一分析二、三阶相乘和相加交互作用项的效应大小，得到如下结果：

2.3.1 二阶、三阶相乘交互作用 r_s4646782，r_s2158041 和 smoyear 的二阶、三阶相乘交互作用效应大小的估计以及检验结果如表 2 所示。

2.3.2 二阶、三阶相加交互作用 r_s4646782，r_s2158041 和 smoyear 的二阶、三阶相加交互作用效应大小的估计以及检验结果如表 3 所示。

以上结果表明，“r_s4646782*r_s2158041”具有相乘效应，乘积每增大一个取值单位，其肺癌发病风险增加 0.097 倍(OR 值为 1.097，95%CI 为 1.046～1.150)。“r_s4646782+r_s215 8041+smoyear”具有相加效应，其和每增大一个取值单位，其肺癌发病风险增加 0.199 倍(OR 值为 1.199，95%CI 为 1.099～1.308)。

3 讨 论

3.1 探讨交互作用的意义

一般认为，交互作用大致可分为两大类，一类是指统计学上的交互作用，另一类则是生物学上的交互作用。统计交互是指暴露于多个单一风险因素作用的乘积或总和是否等于同时暴露于两个或多个风险因素的作用[6]。而生物学交互则主要指从生物学的角度出发，各因素之间的交互作用(如协同作用、拮抗作用等)的可能生物机制是有理论基础的，可以被合理解释的。

表 1 MDR 拟合 1～6 阶基因－环境交互模型情况

交互阶数	Model 包含的因子	Training Bal.Acc.	Testing Bal.Acc.	CV Consistency
1	smoyear	0.576 8	0.575 8	9/10
2	r_s2158041 smoyear	0.589 8	0.555 7	6/10
3	r_s4646782 r_s2158041 smoyear	0.619 0	0.601 3	10/10
4	r_s4646782 r_s2291635 r_s2158041 smoyear	0.640 1	0.555 5	3/10
5	r_s4767944 r_s2291635 r_s7811989 r_s2854462 smoyear	0.682 7	0.549 6	5/10
6	r_s4767944 r_s4149239 r_s2291635 r_s7811989 r_s2854462 smoyear	0.729 1	0.565 0	7/10

表 2 非条件 Logistic 逐步回归结果(相乘交互)

Factors	df	Wald χ^2	P 值	OR 及其 95%CI
模型*中纳入二阶相乘交互项 r_s4646782*r_s2158041；r_s4646782*smoyear；smoyear*r_s2158041				
r_s8190845	1	4.742	0.029	1.293 (1.026, 1.629)
r_s4646782*r_s2158041	1	14.374	0.000 1	1.097 (1.046, 1.150)
模型中纳入以上 3 个二阶相乘交互项和一个三阶相乘交互项 rs4646782*rs2158041*smoyear				
r_s8190845	1	4.742	0.029	1.293 (1.026, 1.629)
r_s4646782*r_s2158041	1	14.374	0.000 1	1.097 (1.046, 1.150)
r_s4646782*r_s2158041*smoyear	1	3.203	0.074▲	0.886 (0.776, 1.012)

注：*模型中还包括吸烟量、家族肿瘤史等一阶非交互项，其结果略，表 3 同；▲逐步回归该交互项被剔除了模型，这是采用强制纳入模型所得结果

表 3 非条件 Logistic 逐步回归结果(相加交互)

Factors	df	Wald χ^2	P 值	OR 及其 95%CI
模型中纳入二阶相加交互项 r_s4646782+r_s2158041; r_s4646782+smoyear; smoyear+r_s2158041				
r_s8190845	1	5.563	0.013	1.321 (1.048, 1.664)
r_s4767944	1	7.490	0.006	1.206 (1.055, 1.379)
r_s2158041+smoyear	1	9.677	0.002	1.197 (1.069, 1.340)
模型中纳入以上 3 个二阶相加交互项和 1 个三阶相加交互项"rs4646782+rs2158041+smoyear"				
r_s8190845	1	5.561	0.018	1.320 (1.048, 1.664)
r_s4646782+r_s2158041+smoyear	1	16.551	0.000 1	1.199 (1.099, 1.308)

因此,探讨统计学交互的生物学意义就具有现实意义与价值。本研究选 r_s4646782,r_s2158041 和 smoyear 3 因子交互模型为最佳模型,不但从统计学角度认为有意义(因该模型交叉验证和置换检验均有显著意义,且符合统计学简约性原则),而且这 3 个因子从生物学角度考虑,均是具有相应的实际意义的(因前两个位点的基因多态性会直接或者间接影响 PAHs 的代谢,加上烟龄,共同都可影响肺癌的发病风险)。正如我们用 MDR 软件分析比较降维后的"高危"人群相对于"低危"人群的肺癌发病风险一样,基因 - 基因以及基因 - 环境的交互效应导致的危险度升高都是十分明显的,其累积效应通常都能达到 2 倍以上。

3.2 交互作用模式的判断

在流行病学研究中,交互作用常采用加法和乘法交互模型。基因 - 环境交互作用的分析依赖于相加或相乘模型的选择,而基于不同模型的分析其意义是不同的。因此,在 MDR 方法分析存在交互效应后, 如何判断交互作用的具体模式,对结果的合理解释意义重大。

多数学者认为,统计学交互必须要建立在有生物学意义的基础上才更加利于结果的解释,比如很多学者就认为相加模型的交互作用往往比相乘模型的交互作用更具有公共卫生学意义。

本研究采用非条件 Logistic 回归对交互模型的具体形式与效应大小等进行了探讨,结果发现,"r_s4646782*r_s2158041"基因多态性具有相乘效应,而"r_s4646782+r_s2158041+smoyear"具有相加效应,乘积或和每增大 1 个取值单位,其肺癌发病风险分别增加 0.097 倍和 0.199 倍。

3.3 MDR 的不足之处及部分解决方法

在 MDR 的主要缺陷中,关于"MDR 法很难得到最终的模型,如不知交互的是全部还是部分因子? 也不知交互的具体形式[3]",通过本研究的方法与思路,采用 Logistic 回归模型对后续交互效应的具体形式程度等进行探讨,可在一定程度上补充一些交互效应的具体信息,在一定程度上克服了这个缺点。但关于 MDR 的其余缺点如面对大量的 SNPs 时分析能力仍显不足,尽管已有部分解决办法,如平行多因子降维等方法[7,8],使得 MDR 分析基因位点的能力提高了很多,但这些解决 MDR 缺点的方法仍待今后进一步深入研究与完善。

3.4 复杂性疾病分析面临的挑战

关于复杂性疾病中基因 - 基因,基因 - 环境交互作用的统计学分析方法,除了 MDR 分析方法以外,还有分层分析,多元 Logistic 回归,广义相对危险度模型,分类回归树法,人工神经网络等分析法。虽然方法众多,但是多基因的复杂性疾病的分析面临着许多问题与挑战[9]。

首先随着实验手段的不断提高,检测出的有多态性的基因和位点成几何级数增长,而纳入研究的样本数始终是有限的,"维度困扰"问题仍不能彻底避免。其次是如何保证筛选的阳性位点和基因真正与所研究疾病有关,这项工作量很大。一般认为应从生物学功能上来选择有意义的位点进行分析。然而,交互作用的模型中很难纳入这种生物学信息。第三个挑战即如何合理的解释基因 - 基因,基因 - 环境交互作用模型。应注意既要有统计学意义上的交互作用,更要具有生物学和医学方面的意义等。

综上所述,本研究结果提示 Logistic 回归在对 MDR 方法分析多因子疾病基因 - 基因、基因 - 环境交互作用的交互具体形式以及效应大小的补充分析中易于操作,结果也较易解释,因而具有可行性和实用性。

(致谢:实例资料来自复旦大学卢大儒教授和南京医科大学沈洪兵教授领衔的研究团队资料与文献)

参 考 文 献

[1] Ritchie M D, Hahn L W, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer[J]. Am J Hum Genet, 2001, 69(1): 138-147.

[2] 唐 迅, 李 娜, 陈大方, 等. 多因子降维法分析基因 - 基因交互作用的应用进展[J]. 中华流行病学杂志, 2007, 28(9): 918-921.

[3] Coffey C S, Hebert P R, Ritchie M D, et al. 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[J]. BMC Bioinformatics, 2004, 5: 49.

[4] Hu Z, Wang H, Shao M, et al. Genetic variants in MGMT and risk of lung cancer in Southeastern Chinese: a haplotype-based analysis[J]. Hum Mutat, 2007, 28(5): 431-440.

[5] Chen D, Tian T, Wang H, et al. Association of human aryl hydrocarbon receptor gene polymorphisms with risk of lung cancer among cigarette smokers in a Chinese population[J]. Pharmacogenet Genomics, 2009, 19(1): 25-34.

[6] 耿 直, 胡永华. 交互作用的统计推断[J]. 中华流行病学杂志, 2002, 23(3): 221-224.

[7] Bush W S, Dudek S M, Ritchie M D. Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions[J]. Bioinformatics, 2006, 22(17): 2173-2174.

[8] Edwards T L, Lewis K, Velez D R, et al. Exploring the performance of Multifactor Dimensionality Reduction in large scale SNP studies and in the presence of genetic heterogeneity among epistatic disease models[J]. Hum Hered, 2009, 67(3): 183-192.

[9] Moore J H, Ritchie M D. The challenges of whole-genome approaches to common diseases[J]. JAMA, 2004, 291(13): 1642-1643.

(责任编辑:冉明会)