中华流行病学杂志2010年1月第31卷第1期 Chin J Epidemiol, January 2010, Vol.31, No.1

·基础理论与方法·

弹性匹配策略在分析基因与环境交互作用中的应用

梁霞 张勇晶 刘冰 金明娟 陈坤

【导读】 以HER-2原癌基因Ile655Val多态性、吸烟与乳腺癌之间的关联研究为例，运用大样本近似原理计算检验效能，并通过逐步提高对照组中匹配因素的比例，探索弹性匹配策略在环境与基因交互作用分析中的应用价值及其效能计算方法。HER-2基因多态和吸烟交互作用的检验效能在非匹配的病例对照研究中为30%，应用传统的频数匹配则提高为56%，进一步增加对照组的吸烟率，则能获得更高的效能值(power=74%)。结论：与非匹配或频数匹配的病例对照研究相比，应用弹性匹配的病例对照研究，能够显著增加环境与基因交互作用的检验效能和研究效率，此匹配策略尤其适用于人群中环境暴露率较低、环境暴露与基因易感性呈负向关联或匹配对照例数较少等情况。在研究设计时可充分权衡检验效能的提高与匹配成本的增加，从而选择最佳的匹配策略。

【关键词】 弹性匹配；检验效能；病例对照研究；交互作用

Application of flexible matching strategy to detect gene-environment interactions for increasing the study power *LIANG Xia, ZHANG Yong-jing, LIU Bing, JIN Ming-juan, CHEN Kun. Department of Epidemiology and Health Statistics, Zhejiang University School of Medicine, Hangzhou* 310058, China

Corresponding author: CHEN Kun, Email: ck@zju.edu.cn

【**Introduction**】 Flexible matching has recently been proposed as a method of improving interactions efficiency. In this study, the concept of flexible matching has been introduced, and the applicability of this strategy has also been described based on the power calculation of interaction between HER-2 polymorphism and smoking with breast cancer. A large-sample approximation method is used to estimate the power and efficiency of gene-environment interactions. In the basic scenario, power of interaction between HER-2 polymorphism and smoking of unmatched case-control study appears to be 30% while in the frequency matching case-control study it is 56%. However, when increasing the smoking prevalence in controls, greater power can be obtained (power=74%). Conclusions: Flexible matching strategies can increase the power and efficiency of case-control studies to detect and estimate the gene-environment interactions when compared with traditional frequency matching and it is especially useful under those scenarios when low environmental exposure of population, adverse gene-environment interactions or less paired controls are seen. Optimal matching design should be made available by weighing the benefits and loss due to flexible matching.

【**Key words**】 Flexible matching; Power; Case-control study; Interaction

在慢性复杂性疾病的病因学研究中，剖析影响因素之间的交互作用，譬如基因-环境因素交互作用，对探明致病因素的联合作用机制和确定高危人群具有重要意义。回顾性病例对照研究被广泛应用于探索各种风险因素对疾病的影响，具有经济、快速、评价主效应的效能较高等优势[1]。然而，病例对照设计在分析交互作用时所表现出的检验效能却相对较低，大大影响了其在评价基因-环境交互作用中的应用。虽然对混杂因素进行匹配可以提高病例对照研究的效能，但对交互作用检验效能的影响却非常有限。而使用弹性匹配策略可以充分增强病例对照研究在分析基因-环境交互作用时的检验效能[2]。以下介绍弹性匹配的基本概念、计算方法及其在流行病学研究设计中的应用价值。

DOI: 10.3760/cma.j.issn.0254-6450.2010.01.024

作者单位：310058 杭州，浙江大学公共卫生学院流行病与卫生统计学系

通信作者：陈坤，Email：ck@zju.edu.cn

基本原理

Sturmer和Brenner[3]在2001年首次提出“弹性匹配”的概念，即在选取对照时以环境暴露作为匹配

因素，使对照组的环境暴露率在病例组的暴露率和人群的暴露率之间（或之外）弹性变化，然后计算不同匹配情形下的检验效能，选择其中检验效能最高者作为分析交互作用最佳的匹配度（degree of match，D_M）。

与通常以混杂因素作为匹配对象不同，弹性匹配以需研究的环境暴露因素作为匹配因素，使其分析基因-环境交互作用的检验效能得到一定提高[4]。在此基础上，假设所研究的环境暴露和基因易感性为二分类变量，人群中的环境暴露率为P_E，基因易感性为P_G，环境暴露与基因易感性之间的比数比为OR_{EG}，在无易感性人群中环境暴露与疾病之间关联的比数比为$OR_{ED|g}$，在易感性人群中环境暴露与疾病之间关联的比数比为$OR_{ED|G}$，基因-环境交互作用为INT，在非暴露人群中基因易感性与疾病之间关联的比数比为$OR_{GD|e}$，对照例数与病例例数比为CC_{RATIO}，匹配度$D_M=(P_{E0}-P_E)/(P_{E1}-P_E)$，其中$P_{E1}$、$P_{E0}$、$P_E$分别为病例组、对照组和人群中的环境暴露率。

根据以上参数设定，分别计算人群中、病例组中及不同匹配度对照组中环境暴露和基因易感性的理论联合分布。设p_{ij}，p_{ijc}和p_{ijm}分别为人群、病例组及匹配对照组中具有不同属性个体的比例，其中i代表环境暴露（$i=1$为暴露状态，$i=0$为非暴露状态），j代表基因易感性（$j=1$为具有易感性，$j=0$为不具有易感性）。

由于$1=p_{11}+p_{00}+p_{10}+p_{01}$；$P_E=p_{11}+p_{10}$；$P_G=p_{11}+p_{01}$；

$$OR_{EG}=\frac{(p_{11}\times p_{00})}{(p_{10}\times p_{01})}$$

解以上方程组得出人群分布：

$$p_{11}=-\frac{[1-(1-OR_{EG})\times(P_E+P_G)]}{[(1-OR_{EG})\times 2]}\pm\sqrt{\left\{\frac{[1-(1-OR_{EG})\times(P_E+P_G)]}{[(1-OR_{EG})\times 2]}\right\}^2+\frac{OR_{EG}\times P_E\times P_G}{(1-OR_{EG})}}$$

$p_{01}=P_G-p_{11}$；$p_{10}=P_E-p_{11}$；$p_{00}=1-(p_{11}+p_{10}+p_{01})$

又由于

$1=p_{11c}+p_{00c}+p_{10c}+p_{01c}$；$OR_{ED|g}=\frac{(p_{10c}\times p_{00})}{(p_{00c}\times p_{10})}$；

$OR_{ED|G}=\frac{(p_{11c}\times p_{01})}{(p_{01c}\times p_{11})}$；$OR_{GD|e}=\frac{(p_{01c}\times p_{00})}{(p_{00c}\times p_{01})}$

解以上方程组得出病例组中分布：

$$p_{00c}=\frac{1}{(OR_{ED|G}\times p_{11}/p_{01}+1)\times OR_{GD|e}\times p_{01}/p_{00}+OR_{ED|g}\times p_{10}/p_{00}+1}$$

$p_{01c}=p_{00c}\times OR_{GD|e}\times p_{01}/p_{00}$；$p_{10c}=p_{00c}\times OR_{ED|g}\times p_{10}/p_{00}$

$p_{11c}=p_{11c}\times OR_{ED|G}\times p_{11}/p_{01}$

令

$P_{E1}=p_{10c}+p_{11c}$；$P_{E0}=p_{10m}+p_{11m}$

则

$$D_M=\frac{(p_{10m}+p_{11m})-P_E}{(p_{10c}+p_{11c})-P_E}；\ OR_{GD|e}=\frac{p_{01c}\times p_{00m}}{p_{00c}\times p_{01m}}；\ OR_{EG}=\frac{p_{11m}\times p_{00m}}{p_{10m}\times p_{01m}}$$

$1=p_{11m}+p_{00m}+p_{10m}+p_{01m}$

解方程组得对照组中分布：

$$p_{00m}=\frac{OR_{GD|e}\times p_{00c}\times\{1-[D_M\times(p_{10c}+p_{11c}-P_E)]-P_E\}}{(p_{01c}+OR_{GD|e}\times p_{00c})}$$

$$p_{01m}=1-\{[D_M\times(p_{10c}+p_{11c}-P_E)]+P_E\}-p_{00m}$$

$$p_{11m}=\frac{OR_{EG}\times p_{01m}\times(1-p_{00m}-p_{01m})}{OR_{EG}\times p_{01m}+p_{00m}}$$

$$p_{10m}=1-(p_{11m}+p_{00m}+p_{01m})$$

以大样本近似（large sample approximation）原理为基础进行检验效能的计算。使用Woolf法进行$OR_{ED|G}$和$OR_{ED|g}$之间的齐性检验，在假定两者齐性前提下估算交互作用的标准误。在无交互作用存在的情况下（$\alpha=0.05$），用所得标准误乘以1.96确定其可信上限，该上限值与理论交互作用对数值之差，除以交互作用理论方差的平方根，可生成一个服从标准正态分布的新统计量。最后，检验效能就等于1减去标准正态曲线在该统计量数值下的面积。然后，以非匹配设计作为比较对象，通过计算两者交互作用的方差之商进行相对效率的计算。

综上所述，当P_E、P_G、OR_{EG}、$OR_{ED|g}$、$OR_{ED|G}$、INT、$OR_{GD|e}$、CC_{RATIO}和N_{CASE}等参数已知（或假定）时，可得到不同D_M情况下的检验效能和相对效率。那么令P_{E0}从1%递增至99%，分别计算相应的检验效能和相对效率，从中选择检验效能最大的就可作为最佳D_M。

由于其计算过程较为繁复，Sturmer等已将其编写成名为“pe0int.sas”的SAS宏程序便于使用和推广（http://www.imbe.med.uni-erlangen.de/issan/SAS/pe0int/pe0int.htm）。

实例分析

目前关于HER-2原癌基因Ile655Val多态性、吸烟与乳腺癌之间的关系已有较多研究，对其产生的主效应已有较为一致的结果，但是涉及这两个危险因素间交互作用对乳腺癌发病风险影响的报道却极少。在此以HER-2 Ile655Val多态性、吸烟与乳腺癌易感性之间的关联研究设计为例，说明如何运用弹性匹配策略在分子流行病学中研究基因-环境交互作用时选择合适的对照人群。

本设计将吸烟和HER-2基因型均设为二分类变量，P_E即女性吸烟率约为3.1%[5]，引用来自2002年我国人群出生、死亡和行为危险因素监测数据。

HER-2易感基因型为Ile/Val和Val/Val基因型，对照人群中Ile/Val和Val/Val基因型的频率分别为23%和2%[6]，即HER-2的P_G频率为25%，该数据由Qu等在上海开展的大规模基于人群的乳腺癌病例对照研究获得。同时在参考大量相关文献的基础上可以假设吸烟与HER-2基因易感性之间相互独立（$OR_{EG}=1.0$）；在非易感人群中即携带Ile/Ile基因型者吸烟与乳腺癌之间的比值比为2（$OR_{ED|g}=2$），HER-2基因多态性——吸烟交互作用为3（INT=3）；在非吸烟人群中，HER-2易感基因型与乳腺癌之间的比值比为2（$OR_{GD|e}=2$）。

根据以上参数设置，计算不同D_M时所得的检验效能（power）和相对效率（relative efficiency），结果见表1。

由上述结果可知，当$D_M=0$，对照组吸烟率等于

表1　对照组在不同吸烟率时分析交互作用的检验效能

P_{E0}	P_{E1}	N_{000}[a]	N_{010}[b]	N_{001}[c]	N_{011}[d]	检验效能(%)	相对效率(%)
0.01	0.10	297	99	3	1	15	40
0.03	**0.10**	**291**	**97**	**9**	**3**	**30**	**100**
0.10	**0.10**	**270**	**90**	**30**	**10**	**56**	**214**
0.24	0.10	228	76	72	24	70	296
0.33	0.10	201	67	99	33	73	316
0.39	0.10	183	61	117	39	74	324
0.45	0.10	165	55	135	45	74	327
0.50	**0.10**	**150**	**50**	**150**	**50**	**74**	**328**
0.55	0.10	135	45	165	55	74	327
0.62	0.10	114	38	186	62	73	323
0.70	0.10	90	30	210	70	72	311
0.80	0.10	60	20	240	80	67	281
0.90	0.10	30	10	270	90	56	214
0.97	0.10	9	3	291	97	30	100
0.99	0.10	3	1	297	99	15	40

注：[a]对照组中未携带易感基因且不吸烟者的例数；[b]对照组中携带易感基因且不吸烟者的例数；[c]对照组中未携带易感基因且吸烟者的例数；[d]对照组中携带易感基因且吸烟者的例数

人群中吸烟率（$P_{E0}=3\%$），即采用非匹配病例对照设计时，其分析交互作用的检验效能仅为30%，并以此时的相对效率作为对比组。当$D_M=1$，对照组吸烟率等于病例组中吸烟率（$P_{E0}=P_{E1}=10\%$），即采用通常的频数匹配设计时，其检验效能提升至56%，相对效率为214%，与非匹配设计相比有明显提高。然而，此时的检验效能并未达到其最大值。若在选择对照人群时采用一定的挑选方法，人为使得对照组吸烟率在24%时其分析交互作用的检验效能就可以达到70%，相对效率为296%，将比通常采用的频数匹配设计（214%）进一步增高，计算此时的匹配度$D_M=3$。当对照组的吸烟率再提高到40%～60%之间时，交互作用的检验效能就可以达到最大值（power=74%），相对效率也在最大值（328%），由此可见，相对于对照组中匹配因素频率“固定”的频数匹配设计，应用弹性匹配策略选择对照人群可以获得更高的检验效能和相对效率。

Sturmer和Brenner[2]根据同样参数设置，利用SAS统计软件产生10 000例病例与对照均为400例的模拟数据集，采用非条件logistic回归模型进行统计分析，计算其检验效能和相对效率与上述结果相似，即在D_M等于2或3时就可以增加病例对照研究在分析交互作用时的检验效能。

在保持其他参数值不变的前提下，分别改变P_E、OR_{EG}、CC_{RATIO}的值，探讨交互作用在不同匹配度下的检验效能和相对效率。与上述的常见情况比较可见，应用弹性匹配策略增加对照人群中的环境暴露率，能更大程度的提高交互作用，在人群中环境暴露率较低（$P_E=0.05$）、环境暴露与基因易感性呈负向关联（$OR_{EG}=0.5$）及匹配对照例数较少（$CC_{RATIO}=0.5$）的情况下检验效能和相对效率。具体结果见表2。

讨　论

病例对照研究中采用匹配是为了提高检验效能和研究效率及控制混杂因素。但匹配设计也有不足，如无法分析匹配因素对疾病影响的主效应，只能用相乘模型评价交互作用，同时增加选择对照的难度等。然而，上述不足均与所选对照中环境暴露率的范围无关，因此，应当以最佳检验效能和研究效率作为匹配的主要标准，而不是以病例组中的环境暴露率作为参考[7]。弹性匹配策略正是基于最大检验效能来确定最佳匹配度D_M，从而选择适当的对照人群。

弹性匹配实质是一种针对分析交互作用的过度抽样（oversampling）策略，适用于当环境暴露因素与疾病之间的关系已经明确，研究目的着重于探讨两因素交互作用的情况。具体操作是将环境因素作为匹配变量，弹性改变对照组中的匹配因素暴露率，即高于、等于或低于病例组中的环境暴露率，以探求一种最优的匹配度来达到交互作用的最高检验效能和研究效率。它尤其适用于：①人群中环境暴露率较低时（$P_E=0.05$）如辐射等罕见的职业性环境暴露；②环境暴露与基因易感性呈负向关联（$OR_{EG}=0.5$）；③对照与病例1:2匹配（$CC_{RATIO}=0.5$）时，例如在肿瘤病因学研究中，比较容易获得病例的基本资料和血或组织样本，而对照人群的血或组织样本却较难获得，这种情况下采用弹性匹配策略选择对照可获得更高的检验效能。虽然，在实际应用中可能无法得到所有参数数据，但通过查阅文献等手段还是可以

表2 病例组和对照组中不同匹配度下环境暴露和遗传易感性的理论联合分布

项目		病例组(%)		对照组(%)							
				D_M=0.0		D_M=1.0		D_M=2.0		D_M=3.0	
Basic											
G[a]		0	1	0	1	0	1	0	1	0	1
E[b]	0	61.0	15.3	72.0	18.0	60.7	15.3	41.5	10.5	22.5	5.5
	1	13.5	10.2	8.0	2.0	19.3	4.7	38.5	9.5	57.5	14.5
检验效能(%)				62		81		87		83	
相对效率(%)				100		159		188		168	
P_E=0.05											
G[a]		0	1	0	1	0	1	0	1	0	1
E[b]	0	69.8	17.5	76.0	19.0	69.5	17.5	59.2	14.8	48.7	12.3
	1	7.2	5.5	4.0	1.0	10.5	2.5	20.8	5.2	31.3	7.7
检验效能(%)				39		62		74		78	
相对效率(%)				100		181		239		261	
OR_{EG}=0.5											
G[a]		0	1	0	1	0	1	0	1	0	1
E[b]	0	62.0	16.5	71.2	18.8	61.8	16.2	44.3	11.7	27.0	7.0
	1	15.5	6.0	8.8	1.2	19.5	2.5	38.8	5.2	58.3	7.7
检验效能(%)				47		68		79		79	
相对效率(%)				100		163		212		213	
CC_{RATIO}=0.5											
G[a]		0	1	0	1	0	1	0	1	0	1
E[b]	0	61.0	15.3	72.0	18.0	61.0	15.0	41.5	10.5	22.5	5.5
	1	13.5	10.2	8.0	2.0	19.0	5.0	38.5	9.5	57.5	14.5
检验效能(%)				40		62		72		65	
相对效率(%)				100		175		218		188	

注：[a]是否携带易感基因：0=未携带，1=携带；[b]是否暴露于环境因素：0=未暴露，1=暴露

获得P_E、P_G和OR_{ED}的估计值，并且在多数分子流行病学研究中OR_{EG}均假设为1，而其他未知参数如INT和OR_{GD}的数值则对最优匹配度的影响不大[2]。因此，在研究设计阶段或从已建立的数据库中抽取对照时（如进行巢式病例对照研究），应用弹性匹配策略切实可行并可获得较高效率。而当所需参数可以获得时，该方法同样适用于环境-环境、基因-基因等其他交互作用的分析。

近年来，为解决分析基因-环境交互作用研究效率低的问题，涌现出许多新颖的研究设计方法，其中常见的有病例-病例设计和反向匹配设计等。病例-病例设计具有省时省力、费用较少、易于实施等优点，而且在大样本量的情况下检测基因-环境交互作用的检验效能较高。但其前提条件是在人群中环境暴露与基因易感性之间相互独立，如果此前提不成立则会使结果的假阳性率大大提高[8]。而弹性匹配在运用时不受此限制，并且在环境暴露与基因易感性存在关联的情况下，能大大降低检验交互作用所需的样本量[9]。反向匹配，即在巢式病例对照研究中按照环境暴露因素（或其替代变量）进行分层抽样选取对照，其目的在于使不一致的匹配对子的数量最大化。在高灵敏性和特异性的替代变量存在情况下，反向匹配既可以估计主效应又可以估计交互作用，并且与弹性匹配一样可以明显提高研究功效[10]。但不足是对于低外显性基因导致的常见慢性复杂性疾病，反向匹配的研究功效也并不高，而弹性匹配策略无此限制。

弹性匹配策略的应用也有其局限性，本文介绍的检验效能的计算方法是基于大样本近似原理，当样本量较大时其结果与模拟数据集测算结果非常接近，而当样本量偏小时则变异较大。Saunders和Barrett[9]在Sturmer和Brenner的基础上计算出多种情况下所需的样本量，均达到几千例以上。目前该方法只用于研究因素是二分类变量的情况下，尚不能处理多分类、多水平或连续性变量，有待在此基础上进一步开发应用。另外，如果增加阳性暴露的对照所需经费很高，那么应该将研究成本和检验效能联合考虑，从中找到最适宜的匹配度。总之，各种研究设计方法都存在各自的使用范围和优缺点，应当根据研究的具体问题进行权衡比较或综合应用，选择最适宜的对照抽样策略。

参 考 文 献

[1] Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet, 2001, 358 (9290): 1356-1360.

[2] Sturmer T, Brenner H. Flexible matching strategies to increase power and efficiency to detect and estimate gene-environment interactions in case-control studies. Am J Epidemiol, 2002, 155 (7): 593-602.

[3] Sturmer T, Brenner H. Degree of matching and gain in power and efficiency in case-control studies. Epidemiology, 2001, 12 (1): 101-108.

[4] Sturmer T, Brenner H. Potential gain in efficiency and power to detect gene-environment interactions by matching in case-control studies. Genet Epidemiol, 2000, 18 (1): 63-80.

[5] Yang GH, Ma JM, Liu N, et al. Smoking and passive smoking in Chinese, 2002. Chin J Epidemiol, 2005, 26 (2): 77-83. (in Chinese)
杨功焕，马杰民，刘娜，等. 中国人群2002年吸烟与被动吸烟的现状调查. 中华流行病学杂志，2005，26(2)：77-83.

[6] Qu S, Cai Q, Gao YT, et al. ERBB2 genetic polymorphism and breast cancer risk in Chinese women: a population-based case-control study. Breast Cancer Res Treat, 2008, 110 (1): 169-176.

[7] Sturmer T, Gefeller O, Brenner H. A computer program to estimate power and relative efficiency of flexibly matched case-control studies. Methods Inf Med, 2005, 44 (5): 693-696.

[8] Albert PS, Ratnasinghe D, Tangrea J, et al. Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol, 2001, 154 (8): 687-693.

[9] Saunders CL, Barrett JH. Flexible matching in case-control studies of gene-environment interactions. Am J Epidemiol, 2004, 159 (1): 17-22.

[10] Andrieu N, Goldstein AM, Thomas DC, et al. Counter-matching in studies of gene-environment interaction: efficiency and feasibility. Am J Epidemiol, 2001, 153 (3): 265-274.

（收稿日期：2009-06-03）
（本文编辑：张林东）