· 基础理论与方法 ·

配对病例对照研究基因与基因交互作用样本量的确定

王志萍 李会庆

【摘要】 目的 采用配对病例对照研究探讨基因与基因交互作用所需的样本量。方法 logistic回归分析原理建立两基因与一种疾病相关性模型，以模拟参数值回代计算样本量，计算过程以QUANTO软件完成。结果 (1)基因交互作用越大，所需样本量越少；基因型分布频率 P_r 为35%(相当于显性遗传模型易感基因频率等于0.20、阴性遗传模型易感基因频率等于0.60)时，基因交互作用的样本量最少：如两基因主作用均为2，基因交互作用为2时样本量为700例；基因交互作用为5时样本量为200例。(2)探讨基因主作用与探讨基因交互作用所需的样本量不同。结论 提供了可供查阅的两基因交互作用样本含量值。

【关键词】 病例对照研究；基因交互作用；样本量

Sample size requirements for association studies on gene-gene interaction in case-control study *WANG Zhi-ping**, *LI Hui-qing*.* *College of Public Health, Shandong University, Jinan* 250012, China

Corresponding author: *LI Hui-qing*. E-mail: huiqing4192@hotmail.com

【**Abstract**】 **Objective** Sample size requirements for association studies on gene-gene interaction in case-control study. **Methods** Selecting different parameters (such as inheritance mode, susceptibility frequency, frequency of allele for disease, *OR* of gene main effect) and infilling them into QUANTO software based on conditional logistic regression mode. **Results** (1) The main parameters influencing the sample size requirements were the levels of interaction between genes and the susceptibility frequency. The numbers of sample were the same between recessive and dominant when susceptibility frequency were the same. (2) Sample size for testing of gene-gene interaction was different from that for testing of genetic effects. **Conclusion** It was convenient to use the numbers of sample size from the results for gene-gene interaction in case-control study.

【**Key words**】 Case-control study; Gene-gene interaction; Sample size requirements

目前认为约5%的疾病属于基因缺陷性遗传病，绝大多数疾病的发生与基因的易感性有关。易感基因及基因与基因之间交互作用的研究为新热点之一，病例对照研究是最常被采用的研究方法之一。由于基因研究花费高，用最小的样本获取最佳结果是设计的重要问题。本文旨在提供采用匹配的病例对照设计方法研究基因与基因间交互作用，所需要的样本量大小。

基本原理

1. 符号注释：设D代表疾病；g与h代表两个要研究的等位基因，分别有三个基因型(AA，Aa，aa和BB，Bb，bb)，其中A和B分别为易感基因位点，q_A 和 q_B 分别表示g和h等位基因位点上A和B的携带频率。

假定两基因位点不连锁。根据Hardy-Weinberg遗传平衡定律，g基因的各种基因型在人群中分布频率 $P_r(g/q_A)$ 分别为 $AA=q_A^2$，$Aa=2q_A(1-q_A)$，$aa=(1-q_A)^2$。同理，h基因的各种基因型在人群中分布频率 $P_r(h/q_B)$ 分别为 $BB=q_B^2$，$Bb=2q_B(1-q_B)$，$bb=(1-q_B)^2$。

定义致病基因模型：当g=AA或h=BB时G(g)或G(h)=1；当g=aa或h=bb时G(g)或G(h)=0；当g=Aa或h=Bb时，G(g)或G(h)=δ，其极限形式δ=0为隐性遗传模型，而δ=1为显性遗传模型。易感基因g与h相对危险度近似值(*OR*)以 R_g 和 R_h 表示，$R_g=\exp(\beta_g)$ 和 $R_h=\exp(\beta_h)$，易感基因g与h交互作用(为相乘模型)*OR*值以 R_{gh} 表示。

疾病的本底发生概率为 P_O。

2. 样本量的计算：采用条件logistic回归分析

作者单位：250012 济南，山东大学公共卫生学院(王志萍)；山东省医学科学院基础医学研究所(李会庆)

通讯作者：李会庆 E-mail: huiqing4192@hotmail.com

方法建立两种基因(g 和 h)与疾病 D 相关性及基因间交互作用的研究模型[1,2]。则 N 个配对样本其相应的条件似然值

$$L(\beta_g,\beta_h,\beta_{gh})=\prod_{i=1}^{N}\frac{e^{\beta_g G_{i1}+\beta_h H_{i1}+\beta_{gh}G_{i1}H_{i1}}}{\sum_{j\in M(i)}e^{\beta_g G_{ij}+\beta_h H_{ij}+\beta_{gh}G_{ij}H_{ij}}} \quad (1)$$

式中,"1"代表病例,M(i)为每一匹配的对子中病例和对照的个数。当 H=0 时 G 的相对危险度 $R_g=\exp(\beta_g)$,当 G=0 时 H 的相对危险度 $R_h=\exp(\beta_h)$。当 G>0 且 H>0 时相对危险度用 $R_g\times R_h$ 计算:$R_{gh}=\exp(\beta_{gh})$。$R_{gh}\neq 1$ 表示存在交互作用。对数似然函数的最大预期值 L^1 的估计值 $\hat{L}^1$ 计算方法

$$\hat{L}^1=\ln[L(\hat{\beta}_g,\hat{\beta}_h,\hat{\beta}_{gh})] \quad (2)$$

$$\hat{L}^0=\ln[L(\hat{\beta}_g,\hat{\beta}_h)] \quad (3)$$

公式(3)表示基因 g 和 h 之间没有交互作用。样本量的计算公式(双侧检验)为

$$N=\frac{(z_{a/2}+z_b)^2}{2(\hat{L}^1-\hat{L}^0)} \quad (4)$$

无效检验假设(Ho):$\beta_{gh}=0$,即两个基因之间无交互作用。采用 QUANTO 软件(该程序可从 http://hydra.usc.edu/gxe 下载)计算样本量大小。运行该程序时,首先在 study design 栏选择 case-control study;在 hypothesis 栏选择 gene-gene interaction;在 gene G 和 H 两栏分别设定基因 G 和 H 的 allele frequency、inheritance mode: dominant or recessive、susceptibility frequency 参数值;在 disease risk model 中设定 baseline risk 水平、gene G 和 gene H 的主作用以及交互作用水平;在 power 栏设定 power、type Ⅰ error、one-sided or two sided;最后在 calculation 栏得到基因主作用和交互作用样本含量。

实例分析

1. 显性模型基因交互作用研究所需样本量:

(1)易感基因携带率 q_A 与 q_B 相同时交互作用样本量:在显性模型中,R_{gh}、R_g 和 R_h 各设定 4 个水平。结果提示:①不同 q_A 或 q_B 对应不同样本量。当 q_A 或 q_B 为 0.05(基因型频率 P_r 为 10%)时,样本量最大;q_A 或 q_B 为 0.15~0.25(P_r 为 28%~44%)时,样本量最低。②相同 R_g 和 R_h 水平,随交互作用水平增加,样本量明显下降。例如,q_A 或 q_B 为 0.2、R_g 和 R_h 取值为 2,$R_{gh}=2$ 时样本量为 696;$R_{gh}=3$ 时样本量为 322。③当基因间交互作用水平不同时,基因主作用 R_g 和 R_h 取值越大,样本量相应增大(表 1)。

表1 显性模型中易感基因不同携带率的交互作用样本含量

组别	R_{gh}	R_g	R_h	易感基因携带率 q_A 或 q_B					
				0.05	0.1	0.15	0.2	0.25	0.3
1	2	2	2	2505	1026	750	696	738	855
	2	3	3	2453	1113	897	910	1046	1300
	2	4	4	2533	1254	1092	1182	1435	1869
	2	5	5	2665	1424	1319	1503	1900	2555
2	3	2	2	930	415	326	322	360	436
	3	3	3	942	477	418	453	548	712
	3	4	4	1001	563	535	618	788	1067
	3	5	5	1082	663	671	813	1076	1498
3	4	2	2	564	270	224	232	269	337
	4	3	4	587	325	303	343	430	574
	4	4	4	639	396	401	484	637	882
	4	5	5	706	479	515	650	886	1259
4	5	2	2	411	209	182	195	232	298
	5	3	3	438	261	255	299	385	523
	5	4	4	487	327	346	431	580	818
	5	5	5	548	402	453	589	818	1178
5	5	2	3	421	231	214	240	297	392
	5	2	4	438	254	245	283	360	486
	5	2	5	456	277	275	326	423	578
	5	3	4	460	291	296	358	471	653
	5	3	5	483	319	336	415	556	781
	5	4	5	515	365	395	503	688	980

(2)两易感基因携带率 q_A 与 q_B 不同时交互作用样本量:q_A 和 q_B 可有很多种取值,形成多种组合方式。从稀有基因和常见基因中,分别取 7 个位点组合,选取 R_{gh}、R_g 和 R_h 等于 2 和 5 两个水平,交互作用的样本量如表 2 所示。结果表明:①与表 1 比较,当 q_A 或 q_B 不同时,其样本量位于较低携带率样本量与较高携带率样本量(当两基因位点携带率相同时)之间。②R_{gh}、R_g 和 R_h 相同,q_A 或 q_B 从稀有基因到常见基因,样本量逐渐减少,q_A 或 q_B 为 0.15~0.25 时,样本量最小。③R_{gh} 对样本量的影响远大于 R_g 和 R_h。

2. 隐性模型中两基因间交互作用研究所需样本量:该模型两基因位点携带率相同时的样本量如表 3 示:①随 q_A 或 q_B 增加,样本量减少。q_A 或 q_B 为 0.5~0.6(P_r 为30%~36%),样本量最少。②随交互作用水平增加,样本量明显下降。③相同的 R_{gh},R_g 和 R_h 增大时样本量相应增大。④第五组表示 R_g 和 R_h 不同组合时样本量。

3. 显性模型与隐性模型样本量的相关性：按照 Hardy-Weinberg 平衡定律，如果隐性遗传方式$q_A=0.25$，基因型频率 P_r 为0.25的开平方0.5，如果显性遗传方式，P_r 则为0.4375。尽管隐性遗传方式中 q_A 或 q_B 与显性遗传方式中 q_A 或 q_B 不同，但表4示相同的 P_r，显性和隐性遗传方式所对应的样本含量一致，并随着 P_r 的增加样本量明显减少。$P_r<0.2$ 时，样本量>1000；当 $P_r=0.35$ 时，样本量最低；P_r 在0.35以上，样本量逐渐增加。

4. 基因主作用的样本量：表5结果显示，当基因间无交互作用（$R_{gh}=1$）时：①研究基因主作用所需的样本量明显小于研究交互作用的样本量。②与交互作用中样本量和 P_r 之间规律相同，当 $P_r<0.35$ 时基因主作用所需的样本量随 P_r 增加而明显减少；$P_r=0.35$ 时样本量最小；$P_r>0.35$ 样本量增加。③与交互作用中样本量和 R_g、R_h 之间规律不同，当基因间无交互作用时，主作用样本量随 R_g 和 R_h 增加逐渐降低。

5. 疾病本底发生概率与样本含量：上述样本含量计算前提示该基因在人群中发病风险 $Po=0.0001$。当 Po 增加到0.005时，对样本量影响较小。

讨　论

不同的设计方法达到同样检验效率所需的样本量大小不同[3-5]。采用病例对照设计方法研究基因与基因交互作用，主要影响因素为疾病遗传模型、基因的易感性、基因的相对危险度、疾病本底发病风险、检验水准和把握度等。检验水准和把握度在设计中往往为定值。

表2　显性模型中易感基因携带率不同组合时交互作用样本含量

g 基因 q_A	R_{gh}	R_g	R_h	h 基因携带频率 q_B						
				0.02	0.05	0.1	0.15	0.20	0.25	0.3
0.02	2	2	2	11680	5395	3404	2848	2672	2673	2790
	5	2	2	1645	818	566	507	501	522	563
	2	5	5	10675	5212	3516	3084	2997	3078	3279
	5	5	5	1652	916	707	674	692	738	808
0.05	2	2	2	5395	2505	1594	1343	1268	1275	1338
	5	2	2	818	411	289	263	264	278	302
	2	5	5	5212	2665	1892	1718	1713	1796	1945
	5	5	5	916	548	454	451	477	521	580
0.10	2	2	2	3404	1594	1026	873	832	843	891
	5	2	2	566	289	209	194	197	210	231
	2	5	5	3516	1892	1424	1349	1390	1496	1655
	5	5	5	707	454	402	419	459	515	588
0.15	2	2	2	2848	1343	873	750	720	735	781
	5	2	2	507	263	194	182	187	202	224
	2	5	5	3084	1718	1349	1319	1394	1513	1723
	5	5	5	674	451	419	453	510	586	681
0.20	2	2	2	2672	1268	832	720	696	715	764
	5	2	2	501	264	197	187	195	212	237
	2	5	5	2997	1713	1390	1394	1503	1678	1915
	5	5	5	692	477	459	510	589	688	812
0.25	2	2	2	2673	1275	843	735	715	738	793
	5	2	2	522	278	210	202	212	232	262
	2	5	5	3078	1796	1496	1513	1678	1900	2193
	5	5	5	738	521	515	586	688	818	976
0.30	2	2	2	2790	1338	891	781	764	793	855
	5	2	2	563	302	231	224	237	262	298
	2	5	5	3279	1945	1655	1723	1915	2193	2555
	5	5	5	808	580	588	681	812	976	1178

表3　隐性模型中易感基因不同携带率的交互作用样本量

组别	R_{gh}	R_g	R_h	易感基因携带率 q_A 或 q_B						
				0.2	0.3	0.5	0.55	0.6	0.65	0.7
1	2	2	2	11468	2840	801	720	696	724	814
	2	3	3	10647	2760	930	886	910	1010	1214
	2	4	4	10463	2829	1105	1100	1182	1372	1725
	2	5	5	10490	2957	1311	1351	1503	1803	2338
2	3	2	2	4019	1047	341	319	322	349	410
	3	3	3	3791	1049	423	422	453	525	657
	3	4	4	3767	1105	528	551	618	747	975
	3	5	5	3816	1184	650	701	813	1013	1359
3	4	2	2	2319	631	231	223	232	259	314
	4	3	4	2213	649	301	310	343	409	527
	4	4	4	2220	699	389	419	484	600	801
	4	5	5	2270	765	491	546	650	830	1136
4	5	2	2	1617	457	185	183	195	223	276
	5	3	3	1559	481	250	264	299	364	478
	5	4	4	1577	528	332	365	431	545	739
	5	5	5	1627	588	426	485	589	763	1059
5	5	2	3	1580	465	213	218	240	283	361
	5	2	4	1575	481	241	252	283	342	444
	5	2	5	1583	499	269	286	326	400	527
	5	3	4	1564	502	287	309	358	444	593
	5	3	5	1579	525	322	354	415	523	707
	5	4	5	1599	556	375	420	503	644	884

表4　基因型频率 P_r、不同遗传方式与样本量关系*

P_r (G=1)	q_A 或 q_B		样本量	
	显性	隐性	显性	隐性
0.01	0.0051	0.1000	155 454	160 789
0.02	0.0101	0.1415	41 604	41 905
0.05	0.0254	0.2238	7 632	7 651
0.10	0.0514	0.3163	2 403	2 407
0.15	0.0780	0.3873	1 354	1 355
0.20	0.1056	0.4473	972	972
0.25	0.1340	0.5000	801	801
0.30	0.1634	0.5478	722	722
0.35	0.1938	0.5916	696	696
0.40	0.2255	0.6325	708	708
0.50	0.2929	0.7072	834	834

* 数字计算条件为 $R_g=2$, $R_h=2$, $R_{gh}=2$, $Po=0.0001$

表5　基因主作用样本量

R_g	R_h	基因型频率 P_r								
		0.01	0.02	0.05	0.10	0.20	0.30	0.35	0.40	0.50
2	2	2317	1197	509	282	173	143	137	135	139
3	3	760	396	173	100	66	57	56	56	60
4	4	417	219	98	59	41	37	37	37	41
5	5	279	148	68	42	30	28	28	29	32
6	6	207	111	52	33	25	24	24	25	28
7	7	164	88	42	28	21	21	21	22	25
8	8	135	63	36	24	19	19	19	20	23
9	9	115	73	31	21	17	17	18	19	21
10	10	100	55	28	19	16	16	17	18	20

1. 易感基因携带率、基因在人群中分布频率与样本含量：在基因相对危险度、基因交互作用水平一定时，样本含量主要取决于易感基因携带率、基因型频率。易感基因携带率对样本量的影响受疾病模型的影响。尽管显性基因与隐性基因携带率不同，按基因型频率所计算的样本量在两疾病模型中的规律相同。当基因型频率 P_r 小于 0.01 时，样本量较大，研究难于实施。当基因型频率 P_r 在 0.25～0.40（显性模型和隐性模型基因携带率 q_A 或 q_B 分别为 0.13～0.23 和 0.5～0.63）左右，基因交互作用样本量较少，以 P_r 为 35% 时样本量最低。如果交互作用 $R_{gh}=2$（$R_g=2$, $R_h=2$）时，所需样本量低于 1000；如果交互作用 $R_{gh}=5$（$R_g=2$, $R_h=2$）时，所需样本量在 200 左右。当两个易感基因携带频率不同时，交互作用样本量介于较低携带率所需样本量与较高携带率样本量之间。

2. 基因交互作用、基因主作用与样本量：与基因主作用增加时样本量减少的规律相同，随基因交互作用增大，交互作用的样本量明显减少。研究基因主作用的样本量远低于研究基因交互作用所需要的样本量。在探讨两个基因与疾病关系时，按主作用参数确定样本量以研究两个基因各自对疾病的影响。如果研究条件许可，可按研究基因交互作用所需的样本量进行基因之间交互作用的研究。不能用研究基因主作用的样本量研究基因与基因之间的交互作用，否则将得出错误的结论。

3. 病例对照设计适用于罕见疾病危险因素的研究，疾病的发生频率变化对样本量计算影响很小[1]。

4. 该软件计算研究设计样本量是采用人群中的无关正常者作为对照，根据 *OR* 值的交互作用乘法模型，同时假设 AA 等位基因型的危险度为 $(R_A)^2$ 的前提下计算的，且仅适用于两基因存在相乘交互作用的情况。

参　考　文　献

1 Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol, 2002, 155: 478-484.

2 Gauderman W, Morrison J. QUANTO documentation. [Version O. 4. 2 (Beta)] Los Angeles, CA: Department of Preventive Medicine, University of Southern Calif Rnia, 2001.

3 Witte JS, Gauderman WJ, Thomas DC. Asymptotic bias and dfficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am J Epidemiol, 1999, 149: 693-705.

4 Garcia-Closas M, Lubin J. Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. Am J Epidemiol, 1999, 149: 689-692.

5 倪宗瓒，主编. 卫生统计学. 第 4 版. 北京：人民卫生出版社，2000. 159-167.

（收稿日期：2003-09-02）

（本文编辑：张林东）