·基础理论与方法·

反向配比设计在流行病学中的应用

胡筱芸 王建华

【摘要】 **目的** 探讨反向配比设计在流行病学研究中的应用。**方法** 以评价基因-环境交互作用的乳腺癌病因学研究为实例，介绍反向配比的设计方法和统计分析方法。**结果** 该实例说明当人群中基因突变和感兴趣的环境暴露因素均为罕见因素时，使用反向配比设计可提高研究基因与环境交互作用的潜能。**结论** 反向配比设计比传统的流行病学方法更适合研究包括罕见因素在内的交互作用。

【关键词】 流行病学；反向配比设计；巢式病例对照研究

The application of 'counter-matching design' in epidemiological research *HU Xiao-yun, WANG Jian-hua. Department of Epidemiology and Biostatistics, College of Public Health, Tianjin Medical University, Tianjin* 300070, China

【**Abstract**】 **Objective** To explore the application of counter-matching design in epidemiological research. **Methods** Through elaboration of the study about gene-environment interactions in the etiology of breast cancer, methodology regarding counter-matching design and statistic methods was introduced. **Results** This design improved the potential for detecting gene-environment interactions for diseases when both gene mutations and the environmental exposures of interest were rare in the general population. **Conclusion** Counter-matching appeared to be more appropriate than most traditional epidemiologic methods for the study of interactions involving rare factors.

【**Key words**】 Epidemiology; Counter-matching design; Nested case-control study

配比(match)又称匹配，即要求对照在某些因素或特征上与病例保持一致，目的是对两组进行比较时排除配比因素的干扰。配比的形式有反向配比、频数配比、部分配比、个体配比和边缘配比等等。对于多数慢性复杂性疾病来说，基因与环境交互作用的研究日益受到人们的重视。近年来，反向配比(counter matching)被广泛用于评价基因与环境交互作用的研究中。

基本原理

反向配比是 Langholz, Clayton[1] 在 1994 年首先提出的。它是从队列或者具有某种危险因素的人群中获取对照以进行巢式病例对照研究的一种方法，其主要原则是按暴露因素(或暴露因素的替代变量)进行分层抽样以评价那些较为少见的感兴趣因素。回归参数估计可用 Cox 回归或条件 logistic 回归等。

反向配比的具体实施方法就是在巢式病例对照研究抽样方法的基础上进行分层抽样。进行反向配比的关键在于如何选择抽样危险集(sampled risk set)，所以要了解反向配比的方法，我们必须首先弄清抽样危险集和危险集(risk set)的概念[2]。在巢式病例对照研究中每一个危险集包括一个病例和与之在某些因素上配比的全部对照。顾名思义抽样危险集就是从危险集中随机抽样的结果，但这种随机抽样仅仅是针对对照而言的，病例是危险集中的那个病例，故每一个抽样危险集包括一个病例和从危险集中随机抽取的对照。所以危险集和抽样危险集的数量与病例数量是一致的(即抽样危险集内研究对象的数量与病例数量的乘积等于该研究的样本量)。抽样危险集中的统计资料是根据危险集中协变量的变化而获得的。下面介绍一下反向配比中抽样危险集的选择方法。

首先，确定病例和与之配比的所有潜在对照(potential control)即危险集(若该危险集内有 N_S 个研究对象包括 1 个病例和与该病例配比的 N_S-1 个潜在对照)，按暴露因子替代变量(surrogate variable)和/或对混杂因子调整后的暴露情况分为 S 层，每个研究对象根据其暴露信息均被划分到 S 层

作者单位：300070 天津医科大学公共卫生学院流行病与卫生统计教研室

中的某一层内(病例也在 S 层中的某一层内),假如每层中有 N 个研究对象,从各层内的 N 个研究对象中随机抽取 M 个研究对象(病例所在层随机抽取 $M-1$ 个研究对象),则共抽取 M_S 个研究对象,M_S 即抽样危险集中的全部研究对象(包括 1 个病例和与之配比的 M_S-1 个对照)。依此类推,每个病例都按该方法选择对照,故抽样危险集的数量与病例数量一致。

实例分析

Jonine 等[3]为研究辐射与基因易感性对乳腺癌的联合作用,设计了一个以人群为基础的巢式病例对照研究。他们假设携带任何一种乳腺癌易感基因的妇女受到辐射所致乳腺癌的危险性都比不携带该基因的妇女高。在该研究中,选择了 700 例不同时间发病的双侧乳腺癌妇女为病例组,对照采用个体配比的形式,按其确诊的日期和年龄,种族及登记地均与病例相配比。在放射治疗上采用反向配比,按病例登记放射治疗(RRT)情况分层,分为两层,即放射治疗(RRT +)层和非放射治疗(RRT −)层。此外,抽样危险集是由每一个病例和与之配比的并符合条件的 2 个对照组成的。本次研究共有 700 个抽样危险集。每个病例与两个配比的对照组成 3 人一组,2 人接受放疗,1 人未接受放疗,这样对照共选择 1400 例单侧乳腺癌妇女。实际上,反向配比比个体配比仅仅多了一步,那就是将潜在对照按 RRT 情况分层。该研究中反向配比的具体实施过程是,若要配比的病例是 RRT +,首先选择一组潜在对照,该对照确诊的日期和年龄,种族及登记地均与病例个体配比即确定危险集,并按 RRT 情况分层,然后分别从这两层中各随机抽取一个对照与病例配比组成抽样危险集;若要配比的病例是RRT −,选择一组RRT + 的潜在对照然后从中随机抽取 2 个对照与之配比。

病例与对照为1:2反向配比设计的似然度贡献率(likelihood contribution)是根据标准的 Bayes 方程计算而得的。比如,本次研究中第三个反向配比危险集由编号为 9、18 和 24 三个研究对象组成。则似然度由下列公式计算

$$\lambda_9 \mathrm{pr}(18,24|9)/[\lambda_9 \mathrm{pr}(18,24|9)+\lambda_{18}\mathrm{pr}(9,24|18)+\lambda_{24}\mathrm{pr}(9,18|24)] \quad (1)$$

若第 9 号研究对象为病例那么 λ_9 为病例的概率,pr(18,24|9)是在第 9 号研究对象为病例的条件下抽取第 18 和 24 号研究对象为对照的概率,计算第三危险集似然度的各组成部分如表 1。

表1　第三危险集似然度的各组成部分

病例	对照	病例概率 pr	病例 RRT 情况	RRT 层研究对象的数量	病例条件下对照概率 pr*	权数
9	18,24	$\lambda r(Z_9;\beta)$	+	8	$1/12\times1/7$	8/2
18	9,24	$\lambda r(Z_{18};\beta)$	−	12	$2/(8\times7)$	12
24	9,18	$\lambda r(Z_{24};\beta)$	+	8	$1/7\times1/12$	8/2

* 对照人数/患者例数

假定 λ 值是根据比例风险模型和模型中每个个体协变量的值计算而来的,如 $\lambda_9=\lambda r(Z_9;\beta)$,这里 λ 为基础风险,$r(Z_9;\beta)$ 是根据协变量值 Z_9 计算的率比,也是率比参数,β 为回归系数。对照的概率和权数是根据危险集内 RRT − 和 RRT + 层中研究对象的数量计算的。将公式(1)取消公因子后产生加权的条件 logistic 似然度贡献率为

$$(8/2)r(Z_9;\beta)/[(8/2)r(Z_9;\beta)+12r(Z_{18};\beta)+(8/2)r(Z_{24};\beta)] \quad (2)$$

反向配比设计与简单抽样设计在资料分析上只有一个不同之处,那就是对模型的加权。这一点大多数统计软件都可以做到。

由于在实际应用中我们经常使用对数线性模型如 $r(Z;\beta)=\exp(Z\beta)$,上述公式可演变为

$$\exp[Z_9\beta+\log(8/2)]/\{\exp[Z_9\beta+\log(8/2)]+\exp[Z_{18}\beta\log(12)]+\exp[Z_{24}\beta+\log(8/2)]\} \quad (3)$$

因此加权值成为模型中一个固定系数。

讨　论

1. 替代变量的选择:反向配比中提高检验效率最重要的参数是替代变量的敏感性与特异性,因此选择一个高敏感性和高特异性的替代变量是至关重要的。而选择这种替代变量的前提是该替代变量在处于某种危险因素的队列/人群中可以测量。比如,现在测量整个队列人群的基因型是不可能的。在这种情况下,可以考虑用某病的家族史作为该因素的替代变量。但是,在研究设计前,研究者必须弄清疾病家族史与该基因是否存在联系。也就是它能否代表此基因。对少量强易感性基因用家族史作为替代变量,其敏感性和特异性比普通的低外显率基因高。因此在多数慢性复杂性疾病中,在下列前提下替代

变量不会有较高的敏感性与特异性：①所研究的疾病是异源性，即由一个以上基因导致的(引起低敏感性)；②大多数基因携带者并不发病(外显率低)；③家庭成员数较少(②③引起低特异性)。在这种情况下，家族史是基因的一个较弱的替代变量，反向配比的功效也不高。故当感兴趣的基因外显率较低时，可应用生理学上的替代变量，如尿、唾液、毛发等等的表现型试验[4]。

2. 反向配比的优缺点：近年来，复杂疾病的基因与环境交互作用问题日益受到人们的关注。但是，有些用于研究基因-环境交互作用的检测方法由于其统计功效较低，不适用于研究罕见基因(G)和未知环境暴露因素(E)的作用，故提出了反向配比设计。在巢式病例对照研究中反向配比在估计主效应功效方面比传统的随机抽样方法高 25%[5]。Andrieu 等[4]认为，反向配比设计不但可评价二者的主效应而且可用于估计基因与环境交互作用。实际上，当两个因素都是罕见因素时(频率<0.1)，且交互作用较为合适时($RR_{int} \leqslant 5$)，反向配比设计功效较高，但相应的样本量很大。当基因频率很小(如 0.001)时(在癌症和慢性病中多见)，若暴露因子较普遍且交互作用较明显时，样本量是能够达到的。巢式病例对照设计可以在不同危险集中重复选择研究对象，甚至作为病例的个体在疾病发生前可以作为对照出现。在研究中，如果潜在对照的数量远远大于抽样危险集的数量，则潜在对照被重复选择的概率要大大缩小。对于配比来讲，由于暴露因素未分层，潜在对照的数量是有限的可导致重复选择对照。而应用反向配比，由于是按暴露剂量分层，重复选择发生的可能性不大，除非对照与病例比率较大时才可能发生。总之，在研究危险因素交互作用的巢式病例对照研究中，当对照：病例≥2:1时，对已知连续的暴露因子进行配比与反向配比时，其检验效能是相同的。然而，通过调整一个或多个混杂因素和效应修正因素后，对暴露因素的配比可影响该因素的研究效果。另一方面反向配比的研究效率高于随机抽样，所以它优于上述两种方法[6]。在巢式病例对照研究中，为研究某因素 X 与 Z 的交互做用，可以用反向配比和个体配比选择对照。Cologne，Langholz[7]也比较了上述两种配比的研究效率，指出反向配比优于配比和随机抽样。此外，Andrieu 等[4]为评价反向配比对于研究基因(G)与环境(E)交互作用的功效比较了以下几种以人群为基础的设计，包括①队列研究。②1:3巢式病例对照研究。③2-2巢式病例对照研究[以环境因素的替代变量 E(E_{sur})做反向配比即 2 个个体暴露于 E_{sur} 和 2 个个体未暴露于 E_{sur} 组成的 4 人危险集，每个危险集包括 1 个病例和 3 个对照。如：病例暴露于 E_{sur} 则 1 个对照暴露于 E_{sur} 2 个对照未暴露于 E_{sur}；病例未暴露于 E_{sur} 则 1 个对照未暴露于 E_{sur} 2 个对照暴露于 E_{sur}]。④2-2巢式病例对照研究[以基因的替代变量 G 做反向配比(G_{sur})，方法同③]。⑤1-1-1-1 巢式病例对照研究(以环境和基因的替代变量 E、G 做反向配比，1 个人暴露于 E_{sur} 和 G_{sur}；1 个人未暴露于 E_{sur} 和 G_{sur}；1 个人暴露于 E_{sur} 而未暴露于 G_{sur}；1 个人暴露于 G_{sur} 而未暴露于 E_{sur})。结果表明：应用1-1-1-1反向配比设计比用传统的流行病学配比方法更适合研究基因与环境的交互作用，当然也包括那些基因和环境为罕见因素的研究。

在研究设计阶段，选择合适的配比方法是研究的关键，因为它决定了对照的选择，样本量的确定及研究是否能达到预期效果等问题，但是会失掉对配比因素分析的机会。所以在研究中，设计者要根据研究目的充分考虑配比因素，以防配比过度。

参 考 文 献

1 Langholz B, Clayton D. Sampling strategies in nest case-control studies. Environ Health Perspect, 1994, 102 suppl 8: 47-51.

2 Langholz B, Goldstein L. Risk set sampling in epidemiologic cohort studies. Statistical Science, 1996, 11: 35-53.

3 Jonine LB, Langholz B, Robert WH, et al. Study design: evaluating gene-environment interactions in the etiology of breast cancer-the WECARE study. Breast Cancer Res, 2004, 6: R199-R214.

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

5 Steenland K, Deddens JA. Increased precision using counter-matching in nested case-control studies. Epidemiology, 1997, 8: 238-242.

6 John BC, Gerald BS, Kazuo N, et al. Improving the efficiency of nested case-control studies of interaction by selecting controls using counter matching on exposure. Int J Epidemiol, 2004, 33: 485-492.

7 Cologne J, Langholz B. Selecting controls for assessing interaction in nested case-control studies. J Epidemiol, 2003, 13: 193-202.

(收稿日期：2005-05-16)

(本文编辑：张林东)