·基础理论与方法·

空间流行病学中的偏倚与混杂

周艺彪 姜庆五 赵根明

【摘要】 **目的** 探讨空间流行病学中的偏倚与混杂。**方法** 结合实例分析空间流行病学中可能存在偏倚与混杂及其对研究结果的可能影响。**结果** 空间流行病学研究中存在选择性偏倚，确证、分子和分母偏倚，由疾病诱导期/潜隐期的选择和暴露-疾病模式的错误载明所致的偏倚，暴露不准确偏倚，空间相关性，显著性检验，生态学偏倚和社会-经济混杂等 8 种偏倚与混杂。**结论** 空间流行病学研究中的偏倚来源众多且较为复杂，由此可以夸大或掩盖研究结果，故对研究结果的解释应慎重。

【关键词】 流行病学；空间分析；偏倚与混杂

Study on bias and confounding in 'Spatial Epidemiology' *ZHOU Yi-biao, JIANG Qing-wu, ZHAO Gen-ming. The Department of Epidemiology, School of Public Health, Fudan University, Shanghai* 200032, China

【**Abstract**】 **Objective** To explore the biases and confoundings in Spatial Epidemiological studies. **Methods** Possible bias and confounding and their impact on study results in Spatial Epidemiology were analyzed in given examples. **Results** In Spatial Epidemiology, biases related to ascertainment/numerator/denominator induced by the choice of the disease induction/latency period and mis-specification of exposure-disease model, exposure inaccuracy, spatial dependency, significance tests etc. were involved, as well as to ecological, socio-economic confoundings factors. **Conclusion** The sources of bias in 'Spatial Epidemiology' were both numerous and complex, that might be overestimated or underestimated on the study results. Hence, careful interpretation of such studies was needed.

【**Key words**】 Epidemiology; Spatial analysis; Bias and confounding

空间流行病学(spatial epidemiology)是流行病学的一个分支，是一门描述、定量和解释疾病在地理上分布变化的学科，特别是针对小面积范围内环境暴露中疾病分布的变化，其涉及到流行病学、统计学、地理学、人口统计学和环境卫生学等多个领域，主要研究内容有疾病地图、地理相关研究、点源或线源危险性评价和聚集性检测与疾病的聚集性等。随着现代计算机软硬件(如地理信息系统，GIS)、遥感技术(RS)、全球定位系统(GPS)和统计学方法(如贝叶斯模型的 MCMC 技术)的发展，以及近年来公众、政府和媒体对环境与健康问题的普遍关注，极大地促进了空间流行病学的发展与应用。空间流行病学和其他流行病学研究一样，也存在偏倚与混杂的问题，影响研究结果的有效性。虽然偏倚和混杂不是空间流行病学所特有的，但也存在几种偏倚形式，有些可以明确地识别，另一些却不是很明显。空间流行病学中的偏倚和混杂主要有[1]：①选择性偏倚(特别是特殊的地域、时间、疾病和人口的选择)；②确证、分子和分母偏倚(ascertainment, numerator and denominator bias)；③由疾病诱导期/潜隐期的选择和暴露-疾病模式的错误载明所致的偏倚(bias induced by the choice of the disease induction/latency period and mis-specification of exposure-disease model)；④ 暴露不准确偏倚(exposure inaccuracy bias)；⑤空间相关性(spatial dependency)；⑥ 显著性检验(significance tests)；⑦生态学偏倚(ecological bias)；⑧ 社会-经济混杂(socio-economic confounding)等。

1. 选择性偏倚：在空间流行病学中，选择偏倚除了在高危人群的选择中出现之外，还可以在研究地区、时段和疾病分类的选择上出现。特别是在疾病的聚集研究中，有几点是值得注意的，因存在轶事证据而被关注的疾病地方聚集性，并不是经常与该地区所关注的危险因素相吻合，这样就没有专门的病因假说，且该地区危险性升高是与某一参考值比较后得出的，但这结果是由数据本身产生，传统的检验方法则无效。这样会存在许多问题，例如为什么

作者单位：200032 上海，复旦大学公共卫生学院流行病学教研室

选择该地区，而不是别的地区？为什么是这个时间段，这种疾病等？由于总是缺乏有关这些选择机理或方法的相关信息，故统计检验所需的合适分布是无法知道的。

为什么选择该地区进行研究？通常是因为该地某疾病的率较高，然而很少有人去关注那些疾病率较低的地方，但是由于存在抽样变异、人口学特征以及漏报率不相同等，疾病率往往会在地区之间波动，有些地区的疾病率较高，而另一些较低。特别是小范围罕见病的聚集调查中，这个问题就更加突出。还有一些地区的疾病率较低，可能是一个重要的保护因素在起作用，但是这很少被注意与调查。例如在心肌梗死的死亡率研究中[2]，一般用于分析的数据是来自于医院住院病例的报告（因为非住院病例的数据很难获得或低报或不完全）。如果使用这样数据对心肌梗死的死亡进行分析，即使在心肌梗死死亡率无地区分布差异并在排除年龄、社会经济等因素的影响下，也存在离医院越近死亡率越高，越远死亡率越低的现象，即存在向医院附近聚集的假象，或是向大城市（医院多、交通方便）聚集的假象，这主要是由于离医院越近交通越方便，发生心肌梗死时，去医院的机会较大，且死在医院外的机会也较路途远、交通不便的地方小所致。又例如使用 GIS 对乳腺癌进行空间分析时[3]，某些研究对象由于住址缺失或不详细的原因而排除在研究之外，但这些被排除的对象往往是和某些特定的人口学特征（如种族、出生地等）、居住地（如城镇的收入水平等）及其疾患（诊断的年份）相联系，造成在都市"过度"抽样和在农村抽样"不足"。当这些被排除的对象占较大比例时（如 15% 时），就可能产生选择性偏倚。

"边界收缩"（boundary shrinkage）或"texas sharpshooter"效应也干扰疾病聚集性研究[4]。在调查中，边界（地理、时间、人口和疾病）往往紧缩在病例的周围，使得观察到的病例呈聚集性，而期望病例数是减少的，这样显然会增加观察到的病例与期望的病例之比（如 SMR）的比值。例如在 Dounreay 核工厂附近有关年轻人白血病和淋巴瘤的发病研究中，不同时间段和地理边界的选择产生了不同危险性估计的结果[5]。

2. 确证、分子和分母偏倚：确证是指确定某事件在某一人群或研究组中是否发生的过程，确证偏倚是将某一个样本的所有病例或个人的多个级别当作同一级别的系统错误，该偏倚可由资料来源的特性产生，如病例来源于某一个特定的诊所或病例来源于某一受文化、习惯等影响的诊断过程[6]。例如异性恋在 HIV 传播中的作用研究[7]，研究对象中的应答者多数来自于门诊部，通过比较应答者与无应答者的社会人口学特征发现[8]，在男性应答者中，白种人较多，非洲裔美国人较少，有妓女性伴侣者多；应答者与无应答者在地理上的分布约有 70% 重叠，无论男性还是女性非洲裔美国人，应答者与无应答者在地理的分布上有实质性差异，而在"小面积"人群中往往具有相似的社会种族的背景、收入水平、职业阶层和教育程度等[9,10]，这样会对研究结果产生确证偏倚。

在危险度估计中，有许多问题影响着分子和分母的计数。对于分子（病例），存在低估和高估，这主要依赖于疾病诊断的准确性、注册的完整性以及可能的病例重复。例如在早期流产研究中，估计约有 25% 的怀孕是在临床出现之前就损失了[11]，同时诊断标准、诊断技术和注册系统的变化也影响着病例的计数。对于分母（人口），通常只能获得某一个时点的人口数（例如人口普查），因此在非人口普查年需要对人口计数进行估计。人口迁移或流动在空间流行病学中是一个严重的问题，它不仅影响人口的计数，而且影响人群的暴露时期，特别是后者估计起来很困难。分子和分母计数不准确都能导致研究结果在方向上产生偏倚，这样在给定的地区/时段需要详细地了解所研究的疾病、注册系统和人口数据，以解释偏倚产生的可能方向及大小。当分子和分母不准的计数在地理上分布不均匀时，在危险度估计中就会产生偏倚，即使不准的计数在不同地理上对等地分布，尽管相对危险度的估计未产生偏倚，但仍会对发病率或患病率的估计产生偏倚。例如对苏格兰 Dalgety Bay 地区的癌症发病研究中[12]，该地区从 20 世纪 60 年代起，人口快速增长，在 1971 年人口为 1575 人，1981 年增加到 5572 人，3 年（1971、1981 和 1991 年）人口普查资料显示，该地人口数量和年龄构成都发生了实质性的改变。如果以 1981 年的人口普查数据作为 1975 - 1990 年期间人口的点估计值（这在小面积卫生研究中是非常普遍的），实际上低估了期望病例数，明显高估观察病例数与期望病例数的比例（SMR 或 SIR）。

3. 由疾病诱导期/潜隐期的选择和暴露-疾病模式的错误载明所致的偏倚：合适的暴露-疾病模式的载明是流行病学中疾病危险性估计的一个重要问

题，它影响所有流行病学研究，也是空间流行病学调查必须面对与处理的一个问题。如果疾病的诱导期或潜伏期是非特异的，那么由诱导期/潜隐期选择所致的偏倚趋向于零，但是各种疾病的诱导期/潜隐期长短不一，一般具有特异性，如果研究期间选择不当，就可能产生偏倚。一般来说，空间流行病学比较适合研究那些诱导期和潜隐期短的疾病，这是因为由失访、移民或人口流动和疾病的竞争病因所致的偏倚少有发生，以及个体在这段时间的变异较少和比较容易确定一个合适的研究期间。也正是这些原因，胎儿和婴幼儿对环境损伤特别敏感，空间流行病学对与出生有关事件（如低出生体重、先天性异常和儿童期癌症）的研究是富有成效的[13-15]。对于其他疾病，如大多数癌，其诱导期/潜隐期很长，有的长达30年以上。使用空间流行病学对这些疾病进行研究时，必须考虑所研究疾病的基本生物学模式，选择合适的暴露-疾病模式，进而确定适宜的研究期间，避免因暴露-疾病模式的错误载明所致的偏倚。例如，在英国都市固体废弃物焚化炉附近癌症发病研究中[16]，研究时间段为 1974 - 1986 年（英格兰）、1974 - 1984 年（威尔士）和 1975 - 1987 年（苏格兰），这是根据暴露-疾病模式确定，对实体肿瘤（solid tumor），10 年的诱导期/潜隐期是特异的，而淋巴和造血系统肿瘤 5 年的诱导期/潜隐期是特异的[17]。

4. 暴露不准确偏倚：暴露测量是流行病学研究中的一项重要内容，但往往难以对暴露进行直接或准确的测量，空间流行病学研究尤其如此，因此缺乏准确的暴露和混杂信息是流行病学研究（特别是空间流行病学研究）的一个主要的困难。这种测量不准确的问题在流行病学中被称为暴露错误分类或暴露错分。最为理想的是能够对暴露因素的终生生物学暴露剂量进行测量，但这往往不可能，因此在流行病学研究中，常采取多种间接方法对暴露进行测量，比如对空气污染的测量，测量的方法有测量某一点的暴露因素的浓度、用离污染源的距离作为指标替代对暴露的测量和用个体采样器对个体暴露进行测量等，此外对个体的混杂因素（饮食、吸烟和饮酒）的测量，也不可避免地出现测量误差，所有这些都会影响研究结果有效性，产生暴露不准确偏倚。例如在杀虫剂使用与健康关系研究中[18]，有许多因素能引起暴露错分，如暴露资料的来源（来源于杀虫剂使用报告和某一年的现场使用调查）、以居民接近使用杀虫剂环境的距离来划分暴露与否、人口流动、以杀虫剂每年的暴露代替季节的暴露、气候和风速及风向等。如果单独以杀虫剂使用报告或某一年的使用调查资料对 *OR* 值进行估计，其 *OR* 值会低估57%～59%；以距离来划分暴露与否，会产生无差异的暴露分类错误，如以较大距离（如 1000 m）来划分暴露与否，所得 *OR* 值的低估程度大于小距离（如 500 m）者；人口流动对效应估计影响较为复杂，暴露地人口可以迁出或迁入或两者都有，如果人口只从暴露地迁出，随着迁移率的增大，效应估计值的低估程度也越大。对暴露错分可以用误差变量模型（errors-in-variables）和测量误差（measurement error）模型进行处理与校正[19,20]。

5. 空间相关性：在空间流行病学研究中存在一个特殊的问题——空间相关性，即邻近地区的变量值（如发病率、死亡率等）不是互相独立的，而是一个地区的变量值依赖于其邻近地区的变量值，也就是通常所说的，存在空间自相关（spatial autocorrelation）的关系，距离较近地区的变量值之间的相关倾向于正相关，而距离较远地区之间的相关多为负相关或非相关。例如在西欧癌症的空间自相关研究中[21]，空间研究单位之间距离在342 km之内，所有 40 种癌症表现为明显的正自相关，当距离超过1747 km时，有 75% 的癌症表现为明显的负自相关。空间相关是由于相邻地区具有某些共同的特征（包括社会和物理环境等）所致，如果所有这些特征都能测量，这些特征就可以在统计分析中得到合理的分析，并合理地选择统计模型，但在实际中却很难达到，不可测的因素可能引入空间相关中。此外所选择的空间研究单位的大小和形状不合理，也会导致研究结果偏离真实值和绘制出有偏倚的疾病分布图[22]。因此，在研究分析中如果没有考虑空间的相关性，就有可能使研究结果产生偏倚[23]。

6. 显著性检验：对空间流行病学研究的资料进行统计学检验与评估时，存在着一些问题。如以上所述：①研究对象的选择有时不是随机的，其选择机理也不很清楚，因而无法知道数据的分布。②由于相邻地区的变量常不是独立的（即空间相关），对具有空间相关的数据不合理地使用传统的统计学方法[24]。③空间流行病学研究中常存在多重比较，如不同地区、病种、时间段、年龄和性别等之间的两两比较，只有最“显著性”的才被选择报告，这类似于论文的发表偏倚（publication bias），同时也不清楚进行

了多少次的有效检验。使用统计模型对混杂因素进行控制时(如回归模型变量选择的向前、向后和逐步选择法),由于不清楚混杂因素的选择机制并只有最"显著性"变量才保留在模型中,这样会对主效应的估计产生偏倚[25]。④空间研究单位(如地区)上的人口数量不同,也会使研究结果产生偏倚,比如大人口数的地区倾向有一个较小的 P 值,即使它们的相对危险度(RR)相同。因此对空间流行病学数据进行分析与结果的报告,应小心谨慎。

7. 生态学偏倚:空间流行病学往往以地区为研究单位,是从群体水平上分析暴露与疾病的关系,这和生态研究学研究一样,也会产生生态学偏倚,导致生态学谬误。生态学偏倚一般有下列几种形式[26-29]。

(1)载明偏倚(specification bias)[27,28]:分析暴露与疾病关系的回归模型,有直线回归模型、对数直线回归模型和非直线回归模型。模型的不合理选择以及模型中没有包括交互项(效应修饰)和需要控制的变量,都有可能产生载明偏倚。

(2)混杂(confounding)[27,29]:无论地区之间还是地区内存在混杂,都有可能发生生态学偏倚。地区之间的混杂类似于个体水平研究中的混杂。例如,如果研究因素在地区内是随机分布的(即该研究因素在个体水平上与其他危险因素无联系),但由于背景危险因素(除研究因素之外的其他危险因素)在地区之间分布不同并与研究因素有联系,即使该研究因素与所研究的疾病无关,也有可能导致它与所研究的疾病有关的虚假结论,这时"地区"就成了混杂因素。地区内混杂的效果一般很难测定与控制,因为这需要知道混杂和暴露因素在地区内的分布。例如,研究因素在地区内分布不是随机的,即在地区内跟背景危险因素有关,即使研究因素在地区之间与背景危险因素无关,也可以导致生态偏倚的发生。因此在个体水平上无混杂,并不能保证在群体水平上也无混杂;即使群体上无混杂偏倚,个体水平上仍可能发生混杂。

(3)标准化偏倚(standardization bias)[29]:当对混杂进行不完全的调整时,如只对结局率标准化,没有对暴露因素标化,就可能产生偏倚,即标准化偏倚。例如食管癌与饮酒的关系,如果只对食管癌死亡率进行标准化,没有对饮酒者的比例标准化,则食管癌死亡率与饮酒者比例的直线回归方程为 $Y = -0.61 + 22.4\ X$, RR 的估计值为负值,即 $(-0.61 + 22.4 \times 1)/(-0.61 + 22.4 \times 0) = -36$,而使用对数直线回归模型,则 RR 的估计值为11.8,这将明显使研究结果产生偏倚。因此,在研究分析中应对结局率和暴露因素用相同的方法进行标准化,以避免这种偏倚的产生,但在实际应用时往往难以实现,因为这需要知道暴露和混杂变量在研究地区内的分布情况。

(4)效应修饰(effect modification)[29]:生态学偏倚可由纯效应修饰产生。例如在食管癌地区变异与吸烟的关系研究中,如果存在一个协变量(如某营养不足),其在不同地区的分布有较小的变异,并当研究因素(吸烟)不存在时它本身不是一个危险因素(即没吸烟的人群,其跨地区的食管癌发病率是一个常数),但研究因素的效果在地区之间是变化的。在这种情况下用地区食管癌发病率作因变量、地区人群吸烟比例作自变量进行直线回归的生态学分析,就可能产生较大的生态学偏倚,甚至完全颠倒研究因素的效应。

8. 社会-经济混杂[16,30-32]:在小面积流行病学研究中,特别是对环境污染与健康的关系研究,社会-经济混杂是偏倚产生的一个主要潜在来源,这是因为社会经济因素(如住房条件、收入、文化程度、医疗服务、饮食等)不仅与疾病的率有较强的联系(这种联系是可以独立于污染的效应),也可以与高工业化水平和污染的地区有联系,这样,即使污染对所研究的疾病没有作用,仍可得出该疾病与污染有联系的虚假结论。对社会-经济混杂的处理方法有间接标准化法和回归模型法。间接标化法是指先对小面积上的"贫困(deprivation)"从群体水平上进行测量,得到"贫困"指标,然后将"贫困"指标分五层对期望病例数进行调整。但这种方法也有不足之处,存在过度调整的现象。

总之,空间流行病学研究中的偏倚来源众多且较为复杂,它可以夸大或掩盖研究因素的效应,甚至得出相反的结论。我们并不是反对进行空间流行病学研究,只是想强调在实施空间流行病学研究时应仔细考虑这些问题,并对研究结果慎重解释。

参 考 文 献

1 Elliott P, Wakefield JG, Best NG, et al. Spatial epidemiology: methods and applications. Oxford University Press, New York, 2000. 1-84.

2 O'Neill L. Estimating out-of-hospital mortality due to myocardial infarction. Health Care Manag Sci, 2003, 6: 147-154.

3 Gregorio DI, Cromley E, Mrozinski R, et al. Subject loss in spatial analysis of breast cancer. Health Place, 1999, 5: 173-177.
4 Rothman KJ. A sobering start for the cluster Buster' conference. Am J Epidemiol, 1990, 132 suppl: s6-s13.
5 Wakeford R, Binks K, Wilkie D. Childhood leukaemia and nuclear installations. J R Stat Soc Ser A, 1989, 152: 61-86.
6 施侣元, 主编. 流行病学词典. 第 1 版. 北京: 科学出版社, 2001. 180.
7 Woodhouse DE, Rothenberg RB, Potterat JJ, et al. Mapping a social network of heterosexuals at high risk for human immunodeficiency virus infection. AIDS, 1994, 8: 1331-1336.
8 Muth SQ, Potterat JJ, Rothenberg RB. Birds of a feather: using a rotational box plot to assess ascertainment bias. Int J Epidemiol, 2000, 29: 899-904.
9 Stoto MA. Public health assessment in 1990s. Ann Rev Public Health, 1992, 13: 59-78.
10 Elliott P, Martuzzi M, Shaddick G. Spatial statistical methods in environmental epidemiolgy: a critique. Stat Met Med Res, 1995, 4: 137-159.
11 Wilcox AJ, Weinberg CR, O' Connor JF, et al. Incidence of early loss pregnancy. N Eng J Med, 1988, 319: 189-194.
12 Black RJ, Sharp L, Finlayson AR, et al. Cancer incidence in a population potentially exposed to radium-226 at Dalgety Bay. Br J Cancer, 1994, 69: 140-143.
13 Feychting M, Kaune WT, Savitz DA, et al. Estimating exposure in studies of residential magnetic fields and cancer. Importance of short-term variability, time interval between diagnosis and measurement, and distance to power line. Epidemiology, 1996, 7: 220-224.
14 Swan SH, Waller K, Hopkins B, et al. A prospective study of spontaneous abortion: relation to amount and source of drinking water consumed in early pregnancy. Epidemiology, 1998, 9: 126-133.
15 Gakos SW, Wesser BM, Zelen M. An analysis of contaminated well water and health effects in Woburn, Massachusetts (with discussion). J Am Stat Assoc, 1986, 81: 583-614.
16 Lliott P, Shaddick G, Kleinschmidt I, et al. Cancer incidence near municipal solid waste incinerators in Great Britain. Br J Cancer, 1996, 73: 702-710.
17 Rothman KJ. Modern Epidemiology. Boston: Little-Brown, 1986. 58.
18 Rull RP, Ritz B. Historical pesticide exposure in California using pesticide use reports and land-use surveys: an assessment of misclassification error and bias. Environ Health Perspect, 2003, 111: 1582-1589.
19 Carroll RJ, Ruppert D, Stefanski LA. Measurement error in nonlinear models. Chapman and Hall, London, 1995.
20 Fuller WA. Measurement error models. Wiley, New York, 1987.
21 Rosenberg MS, Sokal RR, Oden NL, et al. Spatial autocorrelation of cancer in western Europe. Euro J Epidemiol, 1999, 15: 15-22.
22 Srividya A, Michae E, Palaniyandi M, et al. A geostatistical analysis of the geographic distribution of lymphatic filariasis prevalence in southern India. Am J Trop Med Hyg, 2002, 67: 480-489.
23 David S, Remontet L, Bouvier AM, et al. How to choose in practice the model of spatial variation of cancer incidence? Example of digestive cancers from Côte-d' or "department"-France. Rev Epidemiol Sante Publique, 2002, 50: 413-425.
24 Richardson S. Statistical methods for geographical correlation studies. In: Elliott P, Cuzick J, English D, et al. eds. Geographical and environmental epidemiology: methods for small-area studies. Oxford University Press, New York, 1992. 181-204.
25 Miller AJ. Subset selection in regression. Chapman and Hall, London, 1990.
26 Greenland S. Divergent biases in ecologic and individual-level studies. Stat Med, 1992, 11: 1209-1223.
27 Greenland S, Robins J. Ecologic studies-biases, misconceptions and counterexamples. Am J Epidemiol, 1994, 139: 747-760.
28 Plummer M, Clayton D. Estimation of population exposure in ecological studies. J R Stat Soc Ser B, 1996, 58: 113-126.
29 Greenland S, Morgenstern H. Ecologic bias, confounding and effect modification. Int J Epidemiol, 1989, 18: 269-274.
30 Carstairs V. The use and interpretation of deprivation indices in ralation to health. J Epidemiol Community Health, 1995, 49 suppl 2: s3-s8.
31 Dolk H, Mertens B, Kleinschmidt I, et al. A standardisation approach to the control of socioeconomic confounding in small area studies of environment and health. J Epidemiol Community Health, 1995, 49 suppl 2: s9-s14.
32 Bithell JF, Dutton SJ, Neary NM, et al. Controlling for socioeconomic confounding using regression methods. J Epidemiol Community Health, 1995, 49 suppl 2: s15-s19.

(收稿日期: 2004-04-26)
(本文编辑: 张林东)