学习·发现·交流

流行病学论文数据分析思路

王琳琳 陈常中

【导读】 数据分析是根据现有认识对数据从不同角度进行统计分析，对结果进行解释、归纳，逐步形成新认识的过程，也就是将数据转换为证据的过程。分析结果最终以论文方式发表。准确、清晰、全面严谨的数据分析思路对形成有说服力的结论非常关键。本文将就如何快速建立流行病学论文数据分析思路与分析方法进行探讨。

【关键词】 流行病学；数据分析

Way of 'analytical thinking' on data from epidemiological studies *Wang Linlin*[1], *Chen Changzhong*[2]. 1 *Institute of Reproductive and Child Health, Key Laboratory of Reproductive Health of Ministry of Health, Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing* 100191, China; 2 *Dana-Farber Cancer Institute, Harvard Medical School*
Corresponding author: Chen Changzhong, Email: changzhong_chen@dfci.harvard.edu
This work was supported by grants from the National Natural Science Foundation of China (No. 81202215) *and the Specialized Research Fund for the Doctoral Program of Higher Education* (No. 20100001120132).

【Introduction】 Analysis on data from epidemiological studies is the sequencing process of applying statistical methods to collected data from different angles, interpreting intermediate results, drawing statistical conclusions and forming scientific findings based on existing knowledge. This is also called the 'process of converting data to evidence'. Final results from the analysis are expressed through scientific papers. Process of an accurate, clear and comprehensive data analysis is critical to form a convincing conclusion on a paper. This article discusses how to form the analytical thoughts for conducting a thorough data analysis in order to draw a convincing evidence from epidemiological data.

【Key words】 Epidemiology; Data analysis

流行病学是医学科学研究领域一门方法学。流行病学论文则是医学科研工作者展示和传播医学科研成果的重要途径。医学科研工作者智慧的结晶主要集中于论文的结果及结论，而论文结果及结论的获得需要经过科学的数据分析。因此，一条准确、清晰、严谨的数据分析思路对于科研工作者至关重要，它贯穿于整个课题的设计、实施和数据分析过程中。虽然经过系统培训的科研工作者已经略知或熟知流行病学原理和方法，但是如何透过复杂深奥的教科书，真正掌握切实可行的流行病学数据分析方法，对于一些科研工作者，尤其是初入门的研究者，并非易事。本文将就如何快速建立流行病学论文数据分析思路及高效的统计分析进行探讨。

1. 科研假设：在分析数据之前，需要先了解科研的过程（图1）。首先要有假设，然后设计课题，收集资料，再就是数据分析，数据分析的目的是验证假设。另外一条路线是，从现有数据，如临床日积月累的大量病历资料中，分析提取科学信息。同样先要有个假设，然后根据假设，提取资料，再做数据分析验证假设。数据分析的关键是要抓住假设。抓住了假设，就有了方向。

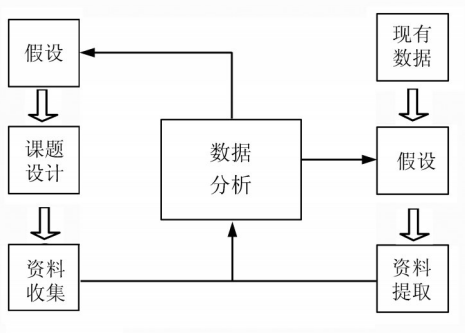

图1 科研过程示意图

科研假设，简单来说，就是有一个X、一个Y，Y是结果变量，X是危险因素。假设是X的变化导致Y的变化，X与Y有联系。如分析肥胖是否会引起高血压，肥胖是X，高血压是Y。有些论文分析是描述性的，目的是对一个地方或一组人某种结局指标的分布情况及其主要影响因素进行描述与分

DOI: 10.3760/cma.j.issn.0254-6450.2014.06.029
基金项目：国家自然青年基金（81202215）；高等学校博士学科点专项科研基金（20100001120132）
作者单位：100191 北京大学生育健康研究所和卫生部生育健康重点实验室、北京大学公共卫生学院流行病与卫生统计学系（王琳琳）；美国哈佛大学医学院Dana-Farber癌症研究所（陈常中）
通信作者：陈常中，Email：changzhong_chen@dfci.harvard.edu

析。虽然目的不是具体回答某个危险因素与结局变量有无关系。实则，如果对这些主要影响因素逐一分析，即看每个X与Y是否有联系，其思路相同。

2. 论文的科研价值：在讨论分析思路之前，需要先比较论文中可能的几种结论？①胖与瘦相比，SBP差异有统计学意义（$P<0.005$）。可以想象，这个结论是用t检验比较两组均数得出的。②BMI与SBP显著相关（$P<0.001$），这是用相关分析做出的。以上两个结论，统计学上都有显著的意义，但没有多大临床指导意义。因为均未回答降低BMI能否降低血压？能降多少？③BMI每增加1 kg/m^2，SBP增加0.01 mmHg，95%*CI*：0.007～0.013 mmHg，$P<0.001$。这是采用回归方程且在统计学上有显著的意义。提示BMI每降低1 kg/m^2，能降低血压0.01 mmHg。但因为控制BMI能导致的降压幅度太小，而无多大临床意义。④控制了其他因素的作用，BMI每增加1 kg/m^2，SBP增加1 mmHg，95%*CI*：0.7～1.3 mmHg。这是采用多元回归方程并控制了其他因素的作用，得出的回归系数1 mmHg，BMI对SBP的独立作用，统计学上有显著意义。BMI每降低1 kg/m^2，能降低血压1 mmHg，这就很有临床价值。

比较这些结论，可以帮助理解统计意义与临床意义的关系，从而理解如何提高一篇论文的科学价值。上面这些结论中，分别有t检验、相关分析和回归分析得出的。其中回归分析，给出有临床意义的回归系数，而且可以控制其他因素的作用，评价X对Y的独立作用。医学数据中，大部分问题，都可用回归分析解决，掌握回归分析非常有必要。

3. 全面的数据分析要回答哪些问题：围绕X与Y是否有联系这样一个假设，在数据分析过程中需要回答以下3个问题：①表面上看，X对Y有无作用？是什么样的作用？②有哪些因素影响X对Y的作用？③X对Y有无独立作用？独立作用的大小是多少？所谓独立作用，是排除了其他因素的混杂作用后，X对Y的作用。

4. 流行病学分析流程（4步）：

（1）研究对象基本特征描述：研究对象基本特征描述是论文数据分析的第一步，通常在论文中以表格形式出现。对读者了解研究人群、解释研究结果及判断结果是否可以外推到其他人群非常必要。在作研究对象特征描述分析时，应考虑①需要描述哪些指标？②是否要分层及如何分层描述？回答第一个问题需要从课题设计入手，可以简单地把课题设计分成试验性研究与观察性研究两大类。所谓试验性研究是指人为控制其他条件使得有X与无X两组研究对象除X外其他条件相同，因此所得出Y的不同，仅仅是因为X不同造成的。对于观察性研究，不能人为地控制有X与无X两组其他因素的分布，所观察到Y的不同则可能不完全是X的不同造成的。试验性研究设计与实施相对较难，但数据分析就相对简单多了。观察性研究设计与实施相对容易，但在数据分析过程中排除其他因素的干扰与混杂就非常重要，其第一步是要描述研究人群中那些可能干扰X与Y关系的因素。一般来说研究人群描述需要包含以下3类变量：①人群的基本特征，如性别、年龄、种族、文化程度等；②与Y可能有关的变量；③与X可能有关的变量。如一项病例对照研究，分析DDE与自然流产关系的研究[1]，作者在表1中列出的吸烟、饮酒、轮班作业、粉尘、噪音、振动这些与自然流产可能有关的变量，同时列出哺乳及哺乳时间，因为哺乳能排泄体内有机氯农药所以与X可能有关。哪些变量与X和Y可能有关，需要参考现有专业知识及从文献中获得，这些因素要尽可能收集到，并在研究人群中加以描述。是否要分层及如何分层描述研究人群，一般根据研究类型，如果是队列研究，分暴露组与非暴露组；病例对照研究，分病例组与对照组；也可不分组。

（2）单因素分析：单因素分析对每个结果变量与每个自变量一一组合进行回归分析，如果需要则可以调整一些基本协变量，如性别、年龄等，或按协变量水平分层。单因素分析的目的是要回答以下3个问题：①不管其他因素的干扰或混杂，看X与Y表面上有没有联系？是什么样的联系？如X是连续性变量，则要看X对Y的作用是单纯的直线性关系，还是呈分段的线性关系？有没有阈值效应或饱和效应？②在收集的其他变量中，哪些因素与Y有联系？是什么样的联系？③在收集的其他变量中，哪些因素与X有联系？是什么样的联系？单因素分析是对所研究的X与Y联系的初步分析，同时发现潜在的混杂因素。单因素分析结果通常也在论文中以表格形式出现。在单因素分析过程中，寻找与发现X与Y的非直线性关系，如阈值效应或饱和效应的关系，非常重要，因为这类信息蕴含着很重要的科研价值[2,3]。如图2所示，X与Y的关系，直线回归X的回归系数为0.2304，$P=0.002$。分析平滑曲线拟合：曲线分为两段，前一段上升，后一段略下降。用两条直线拟合数据，折点为22.73，当X<22.73时，回归系数为0.33，$P<0.001$；当X>22.73时，回归系数为−0.54，$P=0.163$。这两个回归系数差是−0.86，$P=0.043$，表示差异有统计学意义。

（3）分层分析：分层分析主要是解决混杂与交互作用问题，有哪些因素影响所研究X与Y的关系？①混杂：如果某因素Z与研究因素X和研究疾病Y都有关系，且Z不是X与Y因果链上的中间变量，但该因素Z在人群中分布不均，可能

表1　单因素分析与多因素分析中不同回归模型比较

自变量	单因素分析		多因素分析					
			方程一		方程二		方程三	
	β(95%*CI*)	*P*值	β(95%*CI*)	*P*值	β(95%*CI*)	*P*值	β(95%*CI*)	*P*值
X1	0.30(−0.28～0.87)	0.311	0.11(−0.42～0.65)	0.679				
X2	0.47(0.36～0.59)	<0.001	0.47(0.36～0.58)	<0.001	0.47(0.36～0.57)	<0.001	0.47(0.36～0.58)	<0.001
X3	0.41(0.13～0.68)	0.004	0.28(0.01～0.55)	0.044	0.28(0.01～0.55)	0.046	0.36(0.11～0.61)	0.005
X4	3.32(0.37～6.27)	0.028	2.30(−0.59～5.19)	0.119	2.28(−0.60～5.17)	0.122		
X5	5.22(2.91～7.53)	<0.001	4.81(2.60～7.02)	<0.001	4.93(2.80～7.06)	<0.001	4.77(2.65～6.90)	<0.001

会掩盖或夸大X与Y之间的真正联系，这个Z就是混杂因素。分层分析时，如果按Z的不同水平分层，在Z的每个亚层内，X与Y都无关系，这个混杂因素Z就很容易被发现。如图3所示，单因素分析X1对Y的回归线是中间的虚线，回归系数为0.489 9，$P<0.006$，但如按X2分层，每层内X1对Y的回归系数都不显著，这表示原来X与Y的关系是由X2这个混杂因素引起的。②交互作用：如果X对Y的作用在有Z与无Z的存在时有显著不同，则Z就是效应修饰因子，或称Z与X有交互作用。如图4所示，X1对Y的回归系数，在X2=0时为0.23，X2=1为0.51，检验两者差别的显著性，得到$P=0.093$，提示X2与X1可能有交互作用。寻找与发现交互作用因素，对进一步理解X对Y的作用机制非常重要，也是数据分析中很重要的一步[4,5]。

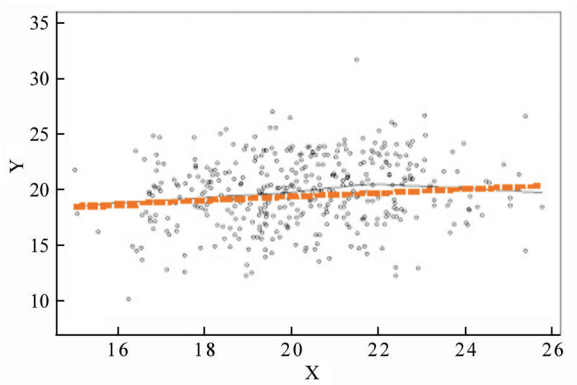

图2　单因素分析中X与Y的直线性及非直线性关系

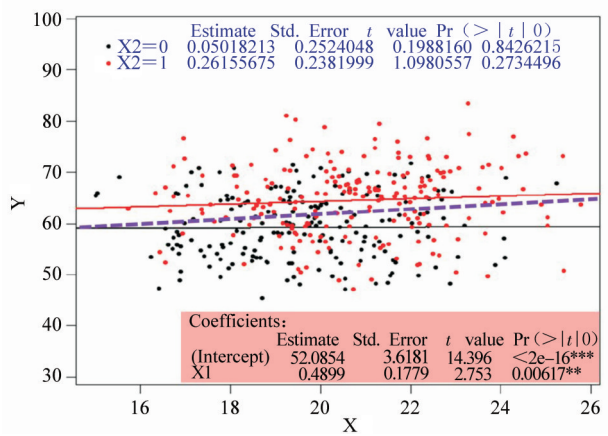

图3　混杂作用分析中X与Y的关系

（4）多元回归分析：多元回归分析的目的是控制、调整其他因素的混杂作用，评价X对Y有无独立作用？独立作用的大小？建立多元模型的关键是要确立哪些因素要纳入模型进行调整。很多分析人员用逐步回归法筛选变量，发现向前与向后的方法，所得出的结果不同；变量顺序不同，结果也不同。如果把过多不必要的因素纳入模型，则增加了模型的自由度，降低检验效率；反之，如果应该调整的因素未纳入模型，则所观察到的X的回归系数，即X对Y的作用中包含了其他因素的作用。

如表1所示，分别对X1、X2、X3、X4、X5与结局变量Y的

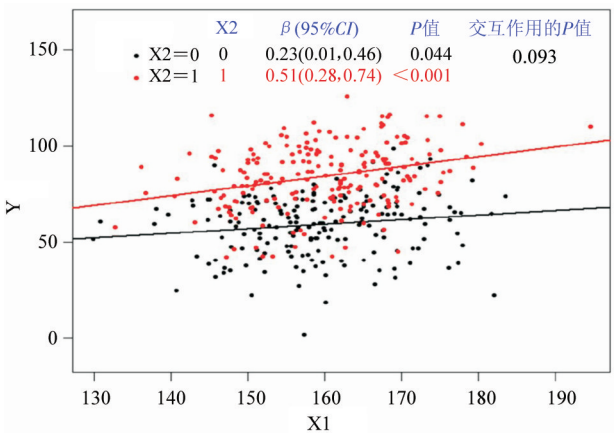

图4　交互作用分析中X与Y的关系

关系逐个做单因素分析，得出除X1外其他4个X与Y均有显著性关系。再做多元回归模型，把5个X同时放入模型中，结果X1还是不显著，X4变得不显著，X2、X3、X5仍然显著。因为X1不显著，多元模型二中不含X1，这个方程二中X4仍不显著。如把X4剔除，得方程三。现在假设要分析的危险因素是X3，结果变量是Y，单因素分析X3对Y的作用是显著的，作用大小是0.41，即X3每增加一个单位，Y增加0.41。多元回归方程一、二、三控制了其他因素，X3的回归系数仍然显著，很明显X3对Y有独立作用。但独立作用大小究竟是多少呢？也即X3对Y的回归系数该用哪个方程呢？是方程二中的0.28，还是方程三中的0.36？比较方程一、二、三可见，调整X4与不调整X4，X3的回归系数变化很大。原因是X3与X4相关。不调整X4，X4的作用就加到X3身上去了；调整了X4，就是把X4的作用从X3中剥离出来，这时X3的回归系数就变小了。所以要正确评价X3对Y的独立作用，要用方程二。按照传统逐步回归方法，仅依据调整变量的P值是否显著，决定是否调整某变量，是不够的。如上例中，根据X4回归系数的P值X4将被剔除，得出的最后方程是方程三，而不是方程二，而方程二中X3的回归系数才是正确地反映X3对Y的独立作用。因此，在建立多元模型评价X对Y的独立作用大小时，如何确定是否需要调整协变量Z？需要对引进Z后所带来X的回归系数的变化进行评价，可以从两个方面进行：①在基本模型中引进Z，也就是比较模型Ⅰ与Ⅱ中X的回归系数β_1的变化。模型Ⅰ：$Y=\beta_0+\beta_1*X$；模型Ⅱ：$Y=\beta_0+\beta_1*X+\beta_2*Z$。②在完整模型中剔除Z，即比较模型Ⅲ与模型Ⅳ中X的回归系数β_1的变化。模型Ⅲ：$Y=\beta_0+\beta_1*X+\beta_2*Z+\beta_2*Z_2+\beta_2*Z_3+\cdots$；模型Ⅳ：$Y=\beta_0+\beta_1*X+\beta_2*Z_2+\beta_2*Z_3+\cdots$；一般认为回归系数变化在10%范围内变化不大，可以不调整Z。在建立多元模型确定要调整哪些变量时，另外一个需要考虑的因素是对有些变量的调整与否，结果的解释完全不同。如分析某因素X与出生体重的关系，可以想象X的作用途径之一，可以是通过引起早产来降低出生体重，如调整出生孕周，则控制了通过引起早产来降低出生体重这条通路，看是否是通过其他通路影响出生体重。因此，在最终结果表达时，常看到很多论文列出了多个模型，以呈现不同的调整方案所观察到的X的作用大小。

一套完整的数据分析思路是要通过上述4步分析，对所研究的X对结局变量Y有无作用？是什么样的作用？哪些因素影响X对Y的作用？X对Y独立作用的大小是多少？进行全面系统的阐述。而且在分析过程中，还需要根据阶段分析结果，进行相应的调整，例如，当发现所研究的X与Y不是直线性关系，而呈现有拐点的阈值效应关系或饱和效应关系时，则在后面的分层分析与多元模型分析时，宜用分段线性模型拟合数据，而不能用直线模型[2,3]。

5. 流行病学数据分析软件：从上述分析流程中，可以看出要完整地实现这些分析步骤，需要分析人员有系统的统计软件应用与编程能力。目前最流行的统计软件有SAS与R软件，这两套软件功能强大，能完成所有的统计分析，但需要用户有很强的编程能力，入门不易。另一个较为普及的软件是SPSS，不需编程，容易入门，但不一定能完成所有要做的分析。如①需要确定拐点使用分段线性模型拟合数据；②需要对不同调整的模型进行比较，确定需要调整哪些变量时，如果要检查的变量很多则很费时；③对多应变量同时进行回归分析时需要对数据结构进行重组以便适用所用模型，等等。这时仅靠手工调用统计模块，不一定能实现。目前由X&Y Solutions软件公司推出了一套全新设计的流行病学分析软件EmpowerStats（中文版命名为：易倾统计）[6]。该软件一改传统的按统计方法设计软件模块的思路，采用全新的根据流行病学分析思路设计分析模块，使用户能按照分析思路调用相应的分析模块。使用该软件，不需编程，不需摘录统计结果，直接输出论文中所要用的图表，使用户能集中精力思考分析结果，调整分析思路，深挖数据中的科学信息，专门适用于流行病学数据分析。

参 考 文 献

[1] Korrick SA, Chen C, Damokosh AI, et al. Association of DDT with spontaneous abortion: a case-control study [J]. Ann Epidemiol, 2001, 11(7): 491-506.

[2] Yu X, Cao L, Yu X. Elevated cord serum manganese level is associated with a neonatal high ponderal index [J]. Environ Res, 2013, 121: 79-83.

[3] Yu X, Zhang J, Yan C, et al. Relationships between serum 25-hydroxyvitamin D and quantitative ultrasound bone mineral density in 0-6 year old children [J]. Bone, 2013, 53(1): 306-310.

[4] Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction [J]. Int J Epidemiol, 2012, 41(2): 514-520.

[5] Sagiv SK, Thurston SW, Bellinger DC, et al. Neuropsychological measures of attention and impulse control among 8-year-old children exposed prenatally to organochlorines [J]. Environ Health Perspect, 2012, 120(6): 904-909.

[6] Empower Stats [OL]. [2014-01-07]. http://www.empowerstats.com/cn/index.html. (in Chinese)
易倾统计 [OL]. [2014-01-07]. http://www.empowerstats.com/cn/index.html.

（收稿日期：2014-01-07）
（本文编辑：王岚）