均衡组间差异的有效方法:倾向评分

赵守军　张勇　汪萱怡　高燕宁

【摘要】 目的　实例介绍倾向评分法的基本原理和适用条件,设计适用于分析二分类资料的SAS宏程序。方法　运用倾向评分比较平衡前后两组间差异的改变情况,评价放弃心肺复苏急救与充血性心力衰竭患者院内死亡的联系。结果　采用分层法和匹配法都可以有效地平衡两组各个特征变量间所存在的高度差异,三种分析方法获得相近的估计结果。结论　倾向评分法是均衡组间差异的有效方法,能够匹配和平衡各个特征变量的作用,并用于分析各种观察性研究资料。

【关键词】 倾向评分法;组间均衡;计算机重新抽样

An effective method to reduce bias between two compared groups: propensity score　*ZHAO Shou-jun**, *ZHANG Yong*, *WANG Xuan-yi*, *GAO Yan-ning*. **Department of Molecular Virology*, *Medical Center of Fudan University*, *Shanghai* 200032, China

【**Abstract**】 **Objective**　Through introduction of principal theory and algorithm of propensity score to design SAS macro programs for binary data. **Methods**　Propensity score method was used to compare the differences of character variables between two groups, and the association of DNR (Do Not Resuscitate) with the mortality of congestive heart failure was evaluated with different methods. **Results**　Significant differences among the character variables between two groups were effectively balanced with stratification or matching method. The odds ratios of DNR with the in-hospital mortality rate of congestive heart failure were estimated identical with different algorithms and to find that the association of DNR to in-hospital mortality was highly significant. **Conclusion**　Propensity score was a good algorithm that could be used to analyze any kind of observational data for matching the effects among the character variables.

【**Key words**】 Propensity score; Homogeneity between groups; Re-sampling with computer

完全随机化的对照分组试验,在理论上是最完善的试验设计,是医学科学发展的基石。但在实际工作中,许多医学或临床试验研究,难以实施完全随机化的对照分组试验。一些突发性疾病,如心脑血管意外,或危险性较大的治疗措施,如开颅或开胸等一些临床手术,都难以进行完全随机化的对照分组试验。如果再考虑到试验对象的志愿选择,以及一些医学伦理问题,完全随机分组的对照试验就成为某些医学研究课题无法逾越的难题。如何评价非随机化分组的观察结果,一直是流行病学研究中探讨的问题。按照完全随机化分组对照的实验设计理论,随意或不完全随机分组的观察结果,由于处理组间的不均衡性,难以确切验证出干预措施的真实作用。如何利用临床积累的大量观察性资料,评价或比较某种处理或治疗措施的效果,是一个值得探索的问题。尤其是近年来社会信息化的发展,医学研究资料的收集方式也伴随着发生了重大的变革。医院信息系统的创立和发展,以及逐年数据信息量的递增和累积,提供了越来越多的资料。针对这种客观需求,一些新的资料分析方法也逐渐发展完善。我们介绍的倾向评分(propensity score, PS)就是近年来广泛应用的一种资料分析方法[1,2]。文中将借助于医院登记资料,介绍该方法的基本原理,并给出实例分析的SAS程序,以便读者参考试用。

作者单位:200032 上海,复旦大学医学院分子病毒研究室(赵守军、汪萱怡) 河北医科大学流行病学教研室(张勇) 复旦大学公共卫生学院(高燕宁)

通讯作者:高燕宁

基本原理

倾向评分法是Rosenbaum和Rubin 1984年首次提出的[3-5],其主要目的是均衡各对比组间各个特征变量的可比性。Rosenbaum和Rubin定义的倾向评分为研究对象 $i(i=1,\cdots,N)$ 按照给定的一组特征变量(x_i)划分到处理组($Z_i=1$)或对照组($Z_i=0$)的条件概率,可以表达为:

$$e(x_i)=P(Z_i=1\mid X_i=x_i)$$

假定分组变量 Z_i 和特征变量 x_i 相互独立,则

$$P(Z_1 = z_1 \cdots\cdots, Z = z_N \mid X_1 = x_1, \cdots\cdots, X_N = x_N) = \prod_{i=1}^{N} e(x_i)^{Z_i} \{1 - e(x_i)\}^{1 - Z_i}$$

P 就是所定义的倾向评分。

这里倾向评分 P 是评价两组间特征变量 x_i 均衡性的近似函数。如果从治疗组选出研究对象 i，则 $P_i(z_i = 1 \mid X_i = x_i)$，再从对照组选出一个研究对象 j，那么 $P_j(z_j = 0 \mid X_j = x_j)$ 如果 $P_i = P_j$，则必然有 $x_i = x_j$，如果我们尽量使 $P_i \approx P_j$，则 x_i 和 x_j 必然十分接近。由此可见，倾向评分 P_i 最大限度地概括了特征变量 x_i 的作用，因而可以有效地保持处理组和对照组间 x_i 的均衡性，使两组间各个特征变量均衡一致。多数情况下 Z_i 均为二分类变量，因此可以运用判别分析或 logistic 回归的方法，估计出各个研究对象的倾向评分 P_i。如果特征变量 x_i 均为正态分布的计量数值，宜于选用判别分析法估计出各个观察对象的倾向评分 P_i；在大多数情况下，尤其是医学研究资料，x_i 中都包含有一些二分类变量或等级变量，多选用 logistic 回归的方法，即

$$\hat{P}_i = e^{(\alpha + \beta_i' x_i)} / \{1 + e^{(\alpha + \beta_i' x_i)}\}$$

这里 $\hat{P}_i$ 就是根据分组特征变量 x_i 估计出的倾向评分，α 和 β_i 是运用 logistic 回归估计出的模型参数。这时根据每个观察对象的 x_i 和估计的模型参数计算出倾向评分 $\hat{P}_i$。

目前文献报道的运用倾向评分 $\hat{P}_i$ 均衡处理分组的方法主要有三种[1,3,4]：

1. 变量调整法（adjustment）：即直接将倾向评分 $\hat{P}_i$ 作为自变量引入模型，分析结果变量与分组处理变量的联系。

2. 分层法（stratification）：是应用较多且简单易行的方法。按照倾向评分 P_i 将全部观察对象分为 5～10 层，依次分析各个观察结果变量（因变量）和处理变量及分层变量的关系，即在均衡了各个特征变量的条件下分析结果变量与分组处理的关系。

3. 配比法（matching）：是最能均衡组间样本分布和构成的方法，也最能体现流行病学研究各观察组间均衡可比的思想。首先将包含有倾向评分 P_i 的全部观察对象按照处理措施有无划分为两个数据文件，并分别按照倾向评分 P_i 的数值大小排序；然后依次从试验组选出一个个体，并从对照组寻找出和该个体的倾向评分 P_i 最为接近的全部个体（小于设定的选择标准），再随机从这些选定的对象中抽取一个或 R 个作为对照（1∶1，一个试验组对象配一个对照 或 1∶R，一个试验组对象配 R 个对照）；依次抽取，直至符合选择标准的观察对象全部抽取。最后按照抽取好的样本分析结果变量和处理变量的联系。这里选择标准的设定和观察资料的利用是一个值得关注的问题。选择标准定得越高，能够完成匹配的对子数就越少，就可能浪费掉许多已得到的信息；反之，则观察组间样本的匹配效果就差一些。

实例分析

本资料来源于美国加州 1999 年医院病例数据库（Office of Statewide Health Planning and Development，OSHPD），研究者欲分析入院时或入院 24 h 内充血性心力衰竭患者签署放弃人工呼吸器急救同意书（do not resuscitate，DNR）与否和患者院内死亡率的关系[6,7]。按照数据库中的变量相关诊断组（diagnosis related group，DRG）筛选出充血性心力衰竭全部入出院记录，剔除非加州居民、邮编缺失或种族缺失的记录，选出78 651人次作为研究观察对象。再将这些观察对象按照病例记录联系编码（record linkage number，RLN）和入院时间排序，以首次入出院时病例签署 DNR 与否和患者出院时生存与否为分析变量。由于 DNR 签署与否和种族、年龄、病程、疾病的严重程度、并发症等一系列影响因素有关，所以 OSHPD 数据库中存在的这些变量均选为特征变量进行分析，具体分析步骤和结果如下。

1. 设计影响 DNR 分组的特征变量，建立估计倾向评分的 logistic 模型：首先根据患者入院时存在的并发症（即其他诊断的 ICD-9 编码），设计并发症评分（co-morbidity score）；然后以 DNR 为因变量，年龄、性别、种族、入院来源、医疗保险种类、并发症评分为特征变量，建立倾向评分的 logistic 模型，并进一步估计出各个研究对象的倾向评分，各特征变量的估计参数和其他统计量见表 1。表 1 中各个特征变量，除 HMO 外均与选择 DNR 与否有高度统计学联系，说明在分析 DNR 与患者院内死亡的联系时，必须消除这些因素的影响。

2. 倾向评分分层前后 DNR 组和对照组的均衡性比较：各个观察对象根据其各自的特征变量运用 logistic 模型估计出的 PS 分层，再比较 DNR 组和非 DNR 组各层 PS 分布情况。表 2 给出了各层的均数和标准差（$\bar{x} \pm s$），数值均十分接近；但各组总计的 PS $\bar{x} \pm s$ 相差却十分明显，如图 1 所示。表 3 为运用 PS 均衡分层后各组观察对象的年龄分布，分析结果

显示各层两组间的 $\bar{x} \pm s$ 都接近于相等，而分层前两组间的 $\bar{x} \pm s$ 差异悬殊，说明所建立的分层变量可以代表各个特征变量的均衡作用。

表1　倾向评分的各个特征变量和估计参数

特征变量	β	$s_{\bar{x}}(\beta)$	χ^2 值	P 值
截矩 β_0	−7.323	0.156	2 202.42	<0.000 1
年龄 β_1（岁）	0.060	0.002	1 122.18	<0.000 1
性别 β_2（女性 = 1）	0.144	0.034	18.42	<0.000 1
亚裔 β_3	−0.812	0.090	81.13	<0.000 1
黑人 β_4	−0.715	0.071	102.27	<0.000 1
西班牙裔 β_5	−0.591	0.066	81.06	<0.000 1
急诊入院 β_6	0.260	0.043	36.75	<0.000 1
转院入院 β_7	0.662	0.068	95.28	<0.000 1
护理机构入院 β_8	0.846	0.103	67.16	<0.000 1
medi-cal/in* β_9	−0.234	0.076	9.55	0.002
HMO* β_{10}	−0.102	0.066	2.36	0.125
non-HMO* β_{11}	−0.237	0.089	7.08	0.008
并发症评分 β_{12}	0.135	0.013	112.69	<0.000 1

* 不同等级的医疗保险

表2　分层后 DNR 与否的均衡性比较

按 PS 分层	倾向评分（$\bar{x} \pm s$）	
	非 DNR 组（n = 45 239）	DNR 组（n = 4 532）
1	0.009 1 ± 0.005 1	0.009 6 ± 0.004 8
2	0.026 9 ± 0.005 5	0.027 2 ± 0.005 5
3	0.046 5 ± 0.006 0	0.047 6 ± 0.006 0
4	0.067 5 ± 0.006 1	0.067 6 ± 0.006 1
5	0.090 1 ± 0.007 0	0.090 7 ± 0.007 0
6	0.115 7 ± 0.008 2	0.116 5 ± 0.008 3
7	0.149 3 ± 0.012 0	0.151 6 ± 0.011 7
8	0.219 0 ± 0.043 4	0.229 5 ± 0.049 2
分层前	0.085 6 ± 0.064 9	0.143 5 ± 0.074 3

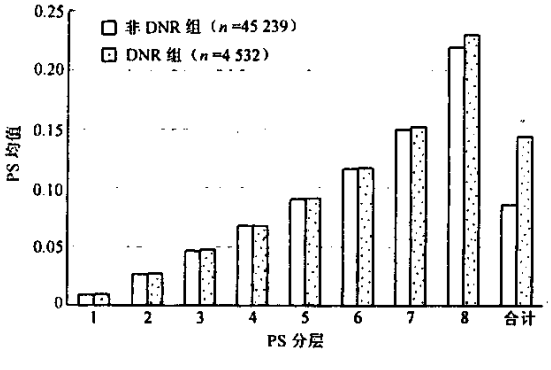

图1　分层前后 PS 均衡性比较

3. 配比法重新抽样后 DNR 组和对照组的均衡性比较：按照设计的 SAS 宏程序，选择精度从高到低的匹配方法，分别以0.000 01、0.000 1和0.001的精度作为标准，DNR 组共抽取4 370例病例，非 DNR 组也抽取相同的对照，病例组资料利用比例为96.4%。原始观察数据和按照设计程序重新抽样后各个特征变量的均数或构成比比较见表 4。DNR 组和非 DNR 组原始数据的均数或构成比的差别均有高度统计学意义（$P < 0.000\ 1$）；重新抽样后两组间年龄、并发症评分的 $\bar{x} \pm s$ 十分接近，性别、种族、入院来源和付费部门的构成均基本一致，其差异均无统计学意义。

表3　按 PS 分层前后年龄分布的比较

按 PS 分层	年　龄　（$\bar{x} \pm s$）	
	非 DNR 组（n = 45 239）	DNR 组（n = 4 532）
1	42.9 ± 16.4	43.4 ± 14.6
2	62.0 ± 6.8	62.0 ± 6.6
3	69.8 ± 6.2	69.3 ± 6.8
4	74.2 ± 5.7	74.0 ± 6.1
5	77.7 ± 5.3	77.8 ± 5.4
6	81.0 ± 4.8	80.9 ± 4.9
7	84.4 ± 4.4	84.9 ± 4.5
8	89.5 ± 4.6	90.5 ± 4.5
分层前	71.8 ± 16.0	81.8 ± 10.8

4. DNR 和充血性心力衰竭患者院内死亡率的关系：从前述表 1 得知，所选入的特征变量均与 DNR 有较强的联系；表 2 和表 4 的分析也清楚说明，分层法和配比法都可以很好地平衡两组间特征变量的分布或构成；所以可以选用直接 PS 调整、分层法和配比法分析 DNR 和充血性心力衰竭患者院内死亡率的关系。表 5 是三种方法所得结果的比较，数值十分接近，说明 DNR 和充血性心力衰竭患者的院内死亡有密切关系，即入院时或入院 24 h 内签署 DNR 显著增加患者的院内死亡率，值得临床医师高度注意和重视。

讨　论

分析结果显示：倾向评分法能够有效地均衡各对比组间特征变量的分布和构成，并在组间均衡的基础上评价干预措施或危险因素与结果变量间的联系或作用。目前一些研究者已经将倾向评分法运用于各种流行病学研究资料，如病例对照研究、队列分析和干预性研究。和经典研究设计一样，倾向评分法适用的前提是研究者所选入的特征变量应包涵可能存在的全部影响因素或混杂因素；由于倾向评分法应用于资料搜集之后，只要分析者能够在分析过程中意识到过度匹配的问题，分析中及时调整，原则上似可减少匹配过头的问题。本文中的应用实例，

表4　原始数据和配比法抽取样本均衡性的比较

特征变量	原始观察数据			按 PS 重新抽样数据		
	DNR(n = 4 532)	非 DNR(n = 45 239)	P 值	DNR(n = 4 370)	非 DNR(n = 4 370)	P 值
年龄(岁,$\bar{x} \pm s$)	81.8 ± 10.8	71.8 ± 16.0	< 0.000 1	81.43 ± 10.79	81.38 ± 10.68	0.84
性别						
男性	1 701(37.6)	21 054(46.5)		1 649(37.7)	1 655(37.9)	
女性	2 822(62.4)	24 194(53.5)	< 0.000 1	2 721(62.3)	2 715(62.1)	0.89
种族						
白人	3 870(85.6)	30 696(67.8)		3 742(85.6)	3 751(84.8)	
黑人	235(5.2)	6 198(13.7)		226(5.2)	244(5.6)	
西班牙裔	279(6.1)	5 400(11.9)		268(6.1)	248(7.7)	
亚裔	139(3.1)	2 954(6.5)	< 0.000 1	134(3.1)	127(2.9)	0.65
入院来源						
电话预约	762(16.8)	10 011(22.1)		750(17.2)	744(17.0)	
急诊入院	3 210(71.0)	32 456(71.7)		3 153(72.1)	3 179(72.7)	
转院	403(8.9)	2 200(4.9)		367(8.4)	357(8.2)	
护理中心转入	148(3.2)	578(1.3)	< 0.000 1	100(2.3)	90(2.1)	0.85
付费部门						
madicare	3 866(85.7)	31 147(69.3)		3 757(86.2)	3 801(87.3)	
medi-cal/ind	214(4.7)	6 389(14.2)		202(4.6)	182(4.2)	
HMO	284(6.3)	4 747(10.6)		265(6.1)	252(5.8)	
non-HMO	147(3.3)	2 638(5.9)	< 0.000 1	134(3.1)	117(2.7)	0.43
并发症评分	1.58 ± 1.81	1.72 ± 2.06	< 0.000 1	1.62 ± 1.97	1.60 ± 1.87	0.72

注 :括号内数据为构成比(%),括号外数据为例数

表5　三种方法估计 DNR 和患者院内死亡率关系的比较

方　法	OR 值	OR 值 95% CI
直接 PS 调整	6.37	5.73 ~ 7.07
按 PS 分层调整	6.28	5.66 ~ 6.97
按 PS 配比调整	5.96	4.94 ~ 7.17

将全部特征变量引入估计分组变量 DNR 的模型,实际上也可以保留某些混杂变量引入处理因素的分析模型。对于具体的分析资料,则取决于研究者对课题的掌握程度和数据库资料的完整性。如果数据库中仅有足够的观察对象而缺乏部分危险因素变量或结果变量,研究者也可以利用倾向评分法从数据库中随机抽取出均衡可比的足够样本,再进行适当的调查或随访,这样就等同于把倾向评分法应用于经典的流行病学设计阶段。因此可以说,倾向评分法既是流行病学研究设计和资料分析的有效工具,也是当前数据库资料科学利用的一种方法。

如果仅从资料分析的角度,与分层和多变量回归模型调整的方法相比,倾向评分法也可以减少或避免分层或回归方法存在的一些问题,如分层分析中产生的样本含量不足和回归分析中的共线问题。在医学生物学研究中,自变量间的相关或共线问题,是产生有偏估计的重要来源。倾向评分综合了全部混杂因素的共同作用,将众多的因素综合为一个变量,使估计因果联系的模型简单化,也可以最大限度地减少共线作用所导致的偏差。当然,倾向评分法也存在一定的局限性,笔者并不认为它可以象有人预言的那样,能够替代随机化临床试验 ;例如将倾向评分法用于干预措施效果的评价时,不管如何均衡,志愿参加或拒绝参加试验这一因素是无法均衡的。但运用倾向评分法,可以平衡已经认知的各个混杂因素,控制和消除它们在分析因果时的作用,为清楚辨析病因联系或进行效用评价提供了一个有效工具。

参　考　文　献

1 D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in Medicine, 1998,17:2265-2281.

2 Joffe MM, Rosenbaum PR. Invited commentary:propensity scores. Am J Epidemiol, 1999,150:327-333.

3 Rubin DB. Matching using estimated propensity scores:relating theory to practice. Biometrics, 1996,52:249-264.

4 Paul R, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc, 1984,79:516-524.

5 Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med, 1997,127:757-763.

6 Shepardson LB, Yongner SJ, Speroff T, et al. Increased risk of death in patients with Do-Not-Resuscitate orders. Medical Care, 1999,37:722-726.

7 Connors AF, Speroff T, Dawson NV, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. JAMA, 1996,276:889-897.

(收稿日期:2002-08-13)

(本文编辑:张林东)