

• 专家讲座 •

如何洞察临床科研设计中的陷阱

胡良平 刘惠刚

统计学思想是用辩证的思维去观察事物,用透视的眼光去洞察事物,即为透过现象看本质的统计学思维模式,概括起来为“八性”和“八思维”^[1-3]。“八性”的含义为“系统性与代表性”、“随机性与均衡性”、“概括性与延展性”及“自悖性与相合性”;“八思维”的含义为“弱化静态思维,强化动态思维”,“突破正向思维、巧用逆向思维”,“跳出简单思维、步入复杂思维”,以及“活用横向思维、发展纵向思维”。

统计学三型理论概述如下^[3-5]:任何与统计学有关的实际问题均存在表现型、原型和标准型,在解决实际问题时,不为表现型所惑,能准确揭示出原型,并将原型正确转变为标准型,就不易出错。这就是笔者创立的统计学三型理论。

下面笔者给大家讲几个科研方面的事例,希望大家掌握临床科研设计的精髓,及时准确地洞察出临床科研设计中可能出现的陷阱。

1 与诺贝尔奖失之交臂

痢特灵原本用来治疗痢疾,但很多医师发现其治疗痢疾的同时也治好了溃疡病。其作用机制是什么呢?中国有关专家猜想:使用痢特灵后体内可能产生了某种抗体,阻止了溃疡病的发展。经过两年的实验,专家们未发现相应抗体则放弃了研究。后来,澳大利亚的一位医师就此问题进行了深入地研究,改变了研究思路,不是寻找抗体,而是探究溃疡病的病因,结果发现导致此病的主要原因是幽门螺旋杆菌(Hp),并且进一步的实验发现,痢特灵可以高效地抑制 Hp,从而控制和治愈溃疡病。该医师因此项研究而一举获得诺贝尔奖,中国有关学者则与诺贝尔奖失之交臂,令人痛心疾首。

2 SCI 露出羞愧的面孔

国际上学术造假的丑闻不断出现,《科学》和《自然》等一流杂志一次又一次蒙受不白之冤;《新英格兰医学杂志》、《美国医学会杂志》和《柳叶刀》杂志上发表的高影响因子和高引用率的学术论文中,有 1/3 经不起时间和实践的检验!显然,这些杂志都是被《科学引文索引》(SCI)收录的高影响因子杂志。毋庸置疑,SCI 并不能被视为“科学严谨”和“学术水平高”的代名词^[6-9]!很

多单位和个人把 SCI 奉为评价科研单位和科技工作者学术水平高低的“金标准”,不考察学术成果和论文本身的科学性、严谨性、创新性和实用性,令人难以接受。

3 哈佛大学校长为何在北大公开认错

美国哈佛大学人体研究计划始于 1995 年,是以哈佛大学公共卫生学院和安徽医科大学合作的名义进行的,对象是安徽大别山区岳西县的农民。所采用的知情同意书是用英文撰写的,农民根本不知道上面写的具体内容,让他们在哪签字就在哪签字,并且他们被多次抽血检查和做多项其他医学检查,其血样被送往何地检测、有何用途,某些检查对身体有多大伤害,他们都一概不知。严格地说,这项人体研究计划是严重违背伦理道德的^[10]!

4 Simpson 悖论给人何启示

Simpson^[11]于 1951 年提出了一种悖论,即对同一种调查资料(表 1)采用 3 种不同的统计分析方法处理,会得出两种自相矛盾的统计结果和专业

表 1 按年龄和性别分层后吸烟与是否患肺癌的调查结果

年龄 (岁)	吸烟与否	例数			
		患病(男)	未患病(男)	患病(女)	未患病(女)
≤40	吸烟	5	5	40	50
	不吸烟	60	55	5	5
>40	吸烟	30	10	5	55
	不吸烟	30	5	5	35

结论。一种结论是吸烟有利于健康,另一种结论是吸烟有害于健康。3 种分析策略分别是:第一,仅考察吸烟与不吸烟者患肺癌概率的差别;第二,将患者按性别分为两组,再考察吸烟与不吸烟患肺癌概率的差别;第三,先将患者按性别分为两组,再将每组患者按年龄分为两组,然后,再在特定性别、年龄组中考察吸烟与不吸烟患肺癌概率的差别。这种悖论给学者们什么启示呢?

分析与解答:提出这种悖论的本意是好的,希望告诫学者们在分析科研数据前一定要慎重选择统计分析方法。但原作者的好意不仅没有起到任何正面作用,反而误导其他学者拿到数据就千方百计地想办法去分析,而不是先考察数据是否值得分析。原文及相关文献^[12]并没有给出选用统计分析方法的正确策略。事实上,本例数据是不值得分析的!因为这些数据很可能不是真实的调查数据,很有可能是造假的结果。这些数据的个位数不是 0 就是 5,整齐划一。这种情况在调查数据中出现的可能性实在太小。从表 1 可知:有两种条件下均只有 10 个被调查对象,调查结果居然有一半被调查者患了肺癌,这

是什么样的人群? 抽出如此小的样本, 其代表性有多大? 即便这是真实的调查数据, 原作者的 3 种分析策略都是错误的, 因篇幅所限, 此处从略, 详细解答请看有关文献^[13]。

5 狂想不等于创新

为了研究由射线引起染色体畸变的量效关系, 传统的做法是选用外周血淋巴细胞作为研究的体系, 而创新者却选用骨髓细胞, 并声称是重大的发明创造。您有何看法?

分析与解答: 外周血淋巴细胞比较均匀, 能够比较准确地反映出剂量与效应之间的关系; 而骨髓细胞恰恰相反, 其内含有多种细胞体系, 细胞所处的周期也不尽相同, 这样的细胞体系对照射剂量并不敏感。因此, 很难达到专业上的需要, 其结论的可信度会大打折扣。创新必须立足于基本常识和专业知识; 其次, 结果和结论对人类的科技进步或对世界的认识有重要促进作用且经得起时间和实践的检验。

6 研究中的陷阱是什么

例 1: 某研究拟探讨某药物治疗妇女更年期综合征的有效性和安全性, 将 240 例更年期妇女完全随机地均分入试验组与安慰剂对照组, 并观察受试者的各种症状和感觉的改善情况, 如出汗、失眠、乏力等。请问: 此项新药临床研究中的陷阱是什么?

分析与解答: 该研究的主要问题为以下两个方面。第一, 在样本含量并非十分大(至少 1000 例以上)的前提下, 仅靠完全随机化分组, 很难保证试验组与对照组之间在很多重要非试验因素方面均衡一致。就本例而言, 重要的非试验因素有: 更年期妇女所处的状态(围绝经期与绝经后期), 心理状态(差、中、好), 不良事件(是否出现下岗、离婚、亲人去世等)。这些重要的非试验因素都会严重影响对试验结果的正确评价。第二, 所选定的评价疗效的指标均是主观的, 其调查结果受调查者、被调查者的心理状态和倾向性的影响, 应该选择一些定量的客观指标(如雌激素水平)评价疗效更为科学、准确。

例 2: 某项研究希望评价某药物治疗骨折的有效性与安全性。在临床试验中, 研究者选取最低药物剂量, 并选择骨密度、骨钙素、胫骨缺省处周长作为主要疗效指标。请问: 此项新药临床研究中的陷阱是什么?

分析与解答: 该研究选取最低剂量的依据不足! 在不知此药物的量效关系时, 盲目地选定最低剂量是不妥当的, 万一此最低剂量无效, 所有的研究工作都将徒劳无益。主要疗效指标的选择也很不恰当! 该研究者所选用的这些定量指标与评价药物治疗骨折效果没有明确的对应关系, 即指标的特异性很

差,都是反映骨质疏松程度的定量指标。正确的疗效指标应该是影像学检查结果(如 X 光片)、功能恢复情况、骨形态、病理检查结果等。

例 3:某项研究希望评价某药物治疗胆囊炎的有效性与安全性。该研究中,有些患者属于单纯胆囊炎患者,有些合并胆结石。将 240 例该病患者随机均分为两组,试验组用新药,对照组用老药。这样设计结果可信吗?

分析与解答:将全部患者完全随机均分为两组,不能确保两组中单纯胆囊炎患者及胆囊炎合并胆结石患者的人数相等、大中小胆结石的人数构成在两组中彼此相等、患病时间在两组的分布相同或接近。这些都是本问题中的重要非试验因素,它们必然会影响对药物疗效的正确评价。因此,本研究所得的结论令人难以置信。

7 对照组设置不当

为说明单纯性老年性白内障会导致眼内测量到的某些指标异常,某研究者进行了如下试验设计:

治疗组 单纯性老年性白内障患者 22 例(共 22 只眼,男 13 只,女 9 只),48~83 岁,平均 66.5 岁。

对照组 意外死亡的健康青壮年人(共 10 只眼,男 9 只,女 1 只),年龄 25~35 岁。

请问:基于这样的设计所得结论有说服力吗^[14]?

分析与解答:与其说这是在做科研,不如说这是在浪费国家的科研经费,在浪费宝贵的时间和生命!该研究者所设置的两个组根本没有可比性!两组受试者在例数、性别构成、年龄构成等方面都不具有可比性,其数据是不值得进行任何统计分析的。

8 应慎用拉丁方设计

某研究者用 6×6 的拉丁方设计安排试验,共选 6 名飞行员,每人服用 6 种降压药(含安慰剂),每种药连续用 1 周,两药之间间隔 1 周。其研究目的是比较药物的疗效。通过正确的统计分析发现,5 种降血压药物与安慰剂之间的疗效差别无统计学意义。显然,此结论与临床知识不符。请问:是什么原因导致此项临床试验研究得出错误的结论?

分析与解答:这是盲目套用拉丁方设计安排多因素试验的结果。这里涉及两个主要的试验因素(药物种类和测定时间)和一个次要因素(个体)。不同药物在每位受试者身上使用的顺序不同,对每种药物疗效的评价必然会有影响,而且这种影响一般来说是非线性的,是不可简单分割开的,况且,药物种类和测定时间之间可能还存在不可忽视的交互作用(但无法真实地显露出来),因此,得出了违反专业知识的结论。一般来说,拉丁方设计适用于一种主要试验因素的各水

平对受试者观测指标影响短暂且指标的取值可以恢复到原先的水平,而且试验中所涉及的三个因素之间的交互作用可以忽略不计。否则,盲目套用拉丁方设计很容易出错!

9 两种手术方法的比较

某医师拟比较 A、B 两种手术方法对同一种疾病的有效性和安全性且希望证明 A 手术法优于 B 手术法(已知本院擅长于 A 法,且很多患者指定要用 A 法)。问:应如何安排,其结论才具有说服力。

分析与解答:若仅在本医院进行此项临床研究,显然是不合适的!因为本院擅长 A 手术不擅长 B 手术,多数患者慕名而来选择要做 A 手术,无法用随机化方法分配患者;况且,对分入 B 手术组的患者是不公平的,而且是违背伦理道德的!若一定要做此项临床试验,应在全国或全世界范围内,选择做 B 手术技术水平最高的医院作为对照医院,在不违背伦理道德、制定了合理的纳入和排除标准、有根据地给出足够样本含量且有很好质量控制的前提下,将全部该病患者按病情轻重、患病时间长短、性别、年龄等重要非试验因素进行分层随机化,确保分入两所医院的该病患者在一切重要非试验因素上具有很好的可比性,选定客观性强的评价疗效和安全性指标,这样做出来的结论才能令人置信。

10 应尽量避免犯对照不全的错误

某项前瞻性研究纳入 90 例食管癌或贲门癌术后吻合口狭窄的住院患者,排除吻合口复发癌和息肉所致的狭窄。其中,男 52 例,女 38 例,年龄 40~76 岁,病程 2~12 个月。狭窄部位于食管上段者 66 例,中段者 5 例,下段 19 例。随机分为微波组、探条组、联合组各 30 例,治疗后跟踪观察吻合口狭窄改善情况。结果:治疗 1 周后,微波组有效率(CR+PR)为 93.3%,探条组为 100%,联合组为 100%(与前两组比较 $P > 0.05$);1 个月后微波组为 70.0%,探条组为 16.7%,联合组为 93.3%(与前两组比较分别为 $P > 0.05$, $P < 0.01$);3 个月后微波组为 53.3%,探条组为 6.7%,联合组为 86.7%(与前两组比较 $P < 0.01$)。结论:内镜下微波高温凝固联合探条扩张是食管癌或贲门癌术后吻合口狭窄的最佳选择。请问:在此项临床研究中,对照组设置有何不妥之处^[15]?

分析与解答:(1)原文作者将患者分为 3 个实验组,分别为微波组、探条组、联合组,不难看出此分组中涉及到两个实验因素,即“使用微波与否”和“使用探条与否”,每个因素中涉及到两个水平,故本应该采用析因设计,即应该有 4 组,原文中缺少了一般治疗的对照组(被称为对照不全)。若在临床上不存在一般治疗组,考虑到伦理道德问题,只能设计成目前的这 3 组,但在数据处理时应慎重,因为这 3 组不是一个标准的单因素三水平设计。(2)原文中所选

患者在性别、年龄、病程、狭窄部位等重要非试验因素方面都差别很大,在分组中原作者只说随机分组,但是没有考虑到这些重要的非试验因素在各组之间是否均衡,应该对这些重要非试验因素进行分层随机化,使每组中患者的基本情况大致相同。这样各组之间才有可比性,才能得出更有说服力的结论。

【关键词】 科研设计;医学研究

【中图法分类号】 R3

【文献标识码】 A

参考文献

- [1] 胡良平. 临床科研工作者呼唤正确的统计学思想. 基础医学与临床, 2007, 27: 228-232.
- [2] Hu LP, Zhang TM. The Statistics thought and its value in biomedical research. Adv Syst Sci Appl, 2008, 8: 430-436.
- [3] 胡良平, 刘惠刚. 统计学思想与三型理论在生物医学科研中的应用. 中西医结合学报, 2007, 5: 216-219.
- [4] 胡良平, 刘惠刚. 统计学的三型理论及其在生物医学科研中的应用. 中华医学杂志, 2005, 85: 1936-1940.
- [5] Hu LP, Liu HM. Triple-type theory of statistics and its role of guidance for scientific research work. J US-China Med Sci, 2009, 6: 56-62.
- [6] 蒋文. 临床研究结果面临时间的考验. 中国医学论坛报, 2005-07-21 (1).
- [7] 胡良平. “学术造假”给科学界敲响了警钟. 中华医学杂志, 2006, 86: 507-509.
- [8] 胡良平. 学术期刊如何防假. 中国医学论坛报, 2006-1-5 (27).
- [9] Ioannidis JP. Why most published research findings are false. Plos Med, 2005, 2: 124.
- [10] 胡良平. 医学统计实用手册. 北京: 人民卫生出版社, 2004: 3-18.
- [11] Simpson EH. The interpretation of interaction in contingency tables. J Roy Statist Soc B, 1951, 13: 238-241.
- [12] 耿直, 金华. 统计因果推断//方积乾, 陆盈. 现代医学统计学. 北京: 人民卫生出版社, 2002: 512-534.
- [13] 胡良平, 张天明. 影响我国科研成果和学术论文质量的要因分析. 科学观察, 2006, 1: 9-19.
- [14] 胡良平. 统计学三型理论在实验设计中的应用. 北京: 人民军医出版社, 2006: 28-33.
- [15] 胡良平. 科研课题的研究设计与统计分析错误案例辨析与释疑(第一集). 北京: 军事医学科学出版社, 2008: 248-255.

(收稿日期: 2009-04-28)

(本文编辑: 罗承丽)

胡良平, 刘惠刚. 如何洞察临床科研设计中的陷阱[J/CD]. 中华乳腺病杂志: 电子版, 2010, 4(3): 282-287.