面对真实世界中残缺的数据：“ 队列 ”还是“ 病例-对照 ”？

作者：李楠 赵一鸣 来源：临床流行病学和循证医学

本文原地址：https://www.iikx.com/news/statistics/1551.html

使用临床现有数据开展研究，这的确是一条捷径，但是前提是：数据质量能够满足研究的需求。我们常说，不少临床常规的记录，本身就可以看做一个纵向数据，假设我们调取了患者从入院开始到出院这个过程的纵向数据，“ 入院→治疗决策A
or决策B→院内治疗→效果评价→出院 ”就可以看做是一个小型的队列。

在一些情况下，我们的评价终点不只是院内结局事件，还需要继续随访一段时间。比如我们给患者进行了关节置换手术之后，比院内结局更重要的是患者的后续功能恢复情况。这就需要通过后续的门诊随访获得患者结局信息了。此时当我们再转向临床数据，试图提取纵向数据信息的时候，很可能数据的完整性就没有住院病例这么好了。

来个看图说话吧：

1、左侧是每个病人观察的起点，比如我们选取了2015-2017年进行手术的患者，那么左侧就是这期间每个进行手术患者各自的手术时间点。

2、右侧是每个患者半年时的随访信息，假说我们关注的是患者有没有进行二次手术，黄色星星代表进行了二次手术。

3、每个箭头都表示1名患者。灰色箭头表示患者失访，也就是在6个月这一时点没有门诊信息(对应后面的红色×);蓝色箭头代表6个月有信息，或者6个月前已经收集到结局事件(二次手术)。当一个箭头对应的右侧终点没有出现图标的时候，意味着这名患者经过6个月随访没有二次手术。

问题来了!~假如我们要评价“ 2中不同术式，患者二次手术的风险 ”。根据我们的尝试，一定首先想到，这么现成的纵向数据，不做个队列研究的设计就浪费了。是的，一般来说的确如此。但是如果同时出现了下面两种情况，您就要多加小心了：

1、当6个月随访点失访患者过多时

。这在临床中是很常见的情况，不少患者觉得病好了，很可能半年就不来复查了;当然也有些患者觉得手术效果不好，想换个更好的医院试试，也不来复查了。对于水平比较高的医院来说，可能第一种情况出现的更多一些。失访有可能会同时带来混杂和偏倚，当然如果失访的数量不多，我们往往直接忽略偏倚，再进一步通过控制患者的基线特征来控制混杂，从而做出推论。

如果失访过多，偏倚的力量就不能被忽略了。最让人发愁的是，偏倚本身无法通过统计学手段去除。此时我们是费更多力气，整理出所有患者的数据，完成一个可能证据等级更高的“ 回顾性队列研究 ”呢?还是干脆省点儿力气，从所有的患者中找出发生结局的患者，在找一些半年没发生结局的患者作为对照，完成一个“ 病例-对照研究 ”就算了呢?也许您就应该做个平衡了。毕竟此时，研究设计类型已经不是证据等级的短板了。介于偏倚存在的可能性，此时的研究无论怎样都只是一个探索，摸索一下大致的方向而已。因此，选择队列还是病例-对照的设计类型，对研究效力的提升就很有限了。此时最理智的设计类型，一定是最省时省力的设计类型。

2、结局发生数量比较少

。比如以二次手术作为结局时，也许我们2年收集了400名患者，其中300名在半年附近有随访(失访率25%，暂且认为失访带来的便宜影响不太大)，而300人中只有20人进行了二次手术。这时候如果还像之前所说的，试图进行一个回顾性队列设计的研究似乎就比较困难了。困住并不在于数据的收集本身，而是因为结局事件发生较少，当我们把所有的患者纳入多因素分析的时候，模型的统计学效能就不够了。毕竟在很多以分类变量作为结局的模型中，样本够不够更多的取决于较少的那个类别。所以这时候我们期望通过统计学模型控制混杂因素的想法就落空了。

但是别忘了，我们控制混杂因素，除了多因素分析之外还有一手，就是匹配。如果这时候我们暂时放弃所有研究对象，只拿出20个发生结局事件的患者作为“ 病例组 ”，通过匹配的方法，找到和他们基线足够相似的20例(20-80例均可)患者作为“ 对照组 ”。这样有匹配的病例-对照设计反而能让我们更好的控制混杂，得到更接近真实情况的效应估计。

不难看出，当我们手头有很多现有的数据可用时，并不是一股脑丢进统计模型就能解决所有问题了。更多时候我们面对的并不是一个理想数据集，而是一个存在各种问题的“ 千疮百孔 ”的真实数据。在数据资源条件有限的情况下，合理的方案选择有可能能帮助我们更好的配置人力资源，甚至是帮助我们得到更接近真实情况的结论。总之，当我们追逐真实世界这一理念的时候，并不意味着“ 真实数据=真实世界 ”，路还很长，需要一步一步走。