



随机对照临床试验与因果推断

战义强

公共卫生学院 (深圳)

中山大学

2025 年春



- ▶ 当你抬头望向天空时，其他行人也会抬头吗？
- ▶ 因果问题：我们想知道一个行为（你的抬头）是否影响了一个结果（其他人的抬头）
- ▶ 设计一个试验：可以站在人行道上，每当有人接近时，我就抛硬币。如果正面朝上，我就抬头；如果反面朝上，我就直视前方
- ▶ 重复这个实验几千次
- ▶ 如果在我抬头后 10 秒内抬头的行人的比例大于在我没有抬头时抬头的行人的比例，我将得出结论，我的抬头对其他人的抬头有因果影响
- ▶ 我可能会雇佣一个助手来记录人们在我抬头时的行为。” 在进行这个研究后，你发现当抬头时，55% 的行人会抬头，但当你直视前方时，只有 1% 的行人会抬头



- ▶ 也可以在一个男人接近时抬头，在一个女人接近时直视前方
- ▶ 那么行动的分配将遵循一个确定性的规则（男人抬头，女人直视），而不是一个随机机制
- ▶ 然而，如果进行了一个非随机实验，研究发现将不会那么有说服力
- ▶ 如果你的行动是由行人的性别决定的，可能会说男人和女人的“抬头”行为不同（女人可能在你抬头后抬头的频率低于男人），因此你的研究比较了本质上“不可比较”的人群
- ▶ 为什么随机化导致有说服力的因果推断？

- ▶ 流行病学的研究目的之一是理解人群中疾病的分布、病因以及非病因的预测因素
- ▶ 为了实现此目的，研究的设计通常包括：实验性和非实验性研究
- ▶ 对于流行病学来说，“实验”这个词通常意味着研究者对研究对象施加某一个或者某些具体的干预措施
- ▶ 因此实验流行病学通常局限在一些暴露因素可以被改变（manipulated）的研究中，因为所有的“试验” / “实验”措施都假定不会对研究对象造成大的伤害（不一定有效，但肯定无害）
- ▶ No causation without manipulation

- ▶ 当流行病学“试验”满足以上基本的伦理学和可行性要求时，选择何种研究的设计，都要以此为指导原则：降低外源性因素所引起的研究结局 outcome 的变异 variation(方差 variance)，或者其他外源性因素所造成的变异 variation(方差 variance) 可以被考虑到（统计分析中可以进行控制或者调整）。
- ▶ 试验方式的分配通常采取随机的方式
- ▶ 随机的目的是：在分配试验方式的当时，潜在的研究结局 potential outcome 在两个（几个）组之间的差异是随机的（一样的，可互换的）
- ▶ 随机分组使得统计推断（点估计和可信区间）变得有意义

流行病学试验研究是队列研究中的一种

► 临床试验



流行病学试验研究是队列研究中的一种

- ▶ 临床试验
- ▶ 现场试验: 选择社区中的某些人



流行病学试验研究是队列研究中的一种

- ▶ 临床试验
- ▶ 现场试验: 选择社区中的某些人
- ▶ 实用性试验 pragmatic trials



流行病学试验研究是队列研究中的一种

- ▶ 临床试验
- ▶ 现场试验: 选择社区中的某些人
- ▶ 实用性试验 pragmatic trials
- ▶ 社区干预试验: 整个社区的全部人



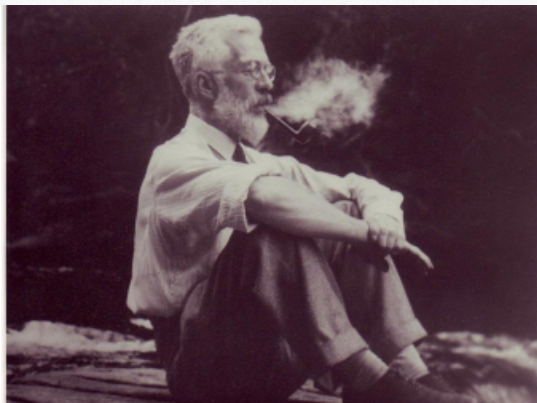
当“试验”因伦理或者可行性等原因，无法进行时，流行病学研究就会选择非试验性研究设计或者观察性研究设计 (non-experimental /observational)，用来模拟 (emulate) 如果试验可以进行的时，可以发生什么样的结果/结局。这些研究设计包括：

- ▶ 队列研究/纵向研究/随访研究
- ▶ 病例对照研究
- ▶ 横断面研究
- ▶ 生态学研究



- ▶ 假设试验只有 2 个组，理想的状态下，2 个组中影响结局的所有外源性因素都相同（分布相同），所以，如果干预措施没有效果/作用，那么 2 个组的结局也是完全一致的
- ▶ 当然，这只是理想的理论情形；因为，我们不可能控制所有的可能影响结局的因素
- ▶ 影响结局的因素很多，也非常复杂，并且大多数情况下是未知的，所以，我们不可能使 2 个组之间的所有因素完全一致
- ▶ 所以，结局 outcome 也会存在变异 variation（即使干预措施没有效果）
- ▶ 这种变异 variation 会使我们评价干预效果的时候非常的困难
- ▶ 如果某因素既影响研究结局又影响干预措施的分配，那么，直接比较两个组的结局之间的差异，是由两部分组成的：（可能的）干预措施的效果 + 某因素对干预措施和结局的效果

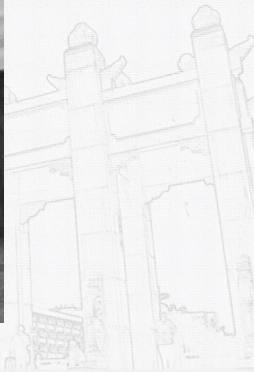
- ▶ 一个减少外部因素影响的方法是：是让外部因素在不同的干预组之间达到平衡 (分布平衡)
- ▶ 20 世纪初，R. A. Fisher 和他的同事发展了可以处理外部因素的方法
- ▶ 随机化



R. A. Fisher 1890-1962



中山大學 公共衛生學院 (深圳)
SUN YAT-SEN UNIVERSITY SCHOOL OF PUBLIC HEALTH, SHENZHEN



主要有两点：

- ▶ 控制混杂！为什么？



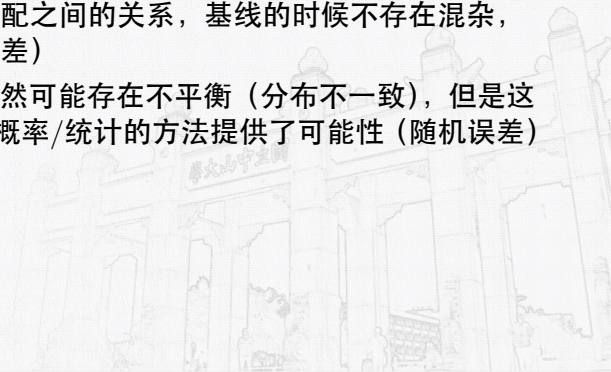
主要有两点：

- ▶ 控制混杂！为什么？
- ▶ 随机化打断了混杂因素和干预措施分配之间的关系，基线的时候不存在混杂，所以解决了混杂偏倚的问题（系统误差）



主要有两点：

- ▶ 控制混杂！为什么？
- ▶ 随机化打断了混杂因素和干预措施分配之间的关系，基线的时候不存在混杂，所以解决了混杂偏倚的问题（系统误差）
- ▶ 随机化之后，基线时外部因素之间仍然可能存在不平衡（分布不一致），但是这完全是偶然的、随机的，这就为使用概率/统计的方法提供了可能性（随机误差）



主要有两点：

- ▶ 控制混杂！为什么？
- ▶ 随机化打断了混杂因素和干预措施分配之间的关系，基线的时候不存在混杂，所以解决了混杂偏倚的问题（系统误差）
- ▶ 随机化之后，基线时外部因素之间仍然可能存在不平衡（分布不一致），但是这完全是偶然的、随机的，这就为使用概率/统计的方法提供了可能性（随机误差）
- ▶ 社会学和流行病学领域，大多数情况下称“试验”为“随机试验”，有时候，称“非随机的”干预试验为准试验（类试验）研究 (quasi-experiment)

Statistical Science
1990, Vol. 5, No. 4, 465-480

On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.

Jerzy Splawa-Neyman

Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original, which appeared in *Roczniki Nauk Rolniczych Tom X (1923) 1-51 (Annals of Agricultural Sciences)*

Abstract. In the portion of the paper translated here, Neyman introduces a model for the analysis of field experiments conducted for the purpose of comparing a number of crop varieties, which makes use of a double-indexed array of unknown potential yields, one index corresponding to varieties and the other to plots. The yield corresponding to only one variety will be observed on any given plot, but through an urn model embodying sampling without replacement from this doubly indexed array, Neyman obtains a

Jerzy Neyman 1894-1981



中山大學 公共衛生學院 (深圳)
SUN YAT-SEN UNIVERSITY SCHOOL OF PUBLIC HEALTH, SHENZHEN



- ▶ 假设 id1 是暴露 ($X = 1$), 那么 $Y = Y_1$, Y_0 是反事实的
- ▶ 假设 id2 和 id3 是非暴露 ($X = 0$), 那么 $Y = Y_0$, Y_1 is 是反事实的

Subject	X	Y	Y_0	Y_1
id1	1	1	?	1
id2	0	0	0	?
id3	0	1	1	?

- ▶ 在 RCT 中, 所有处理/治疗之前 (pre-treatment) 的变量都与处理/治疗措施 (treatment) 无关, 例如: 年龄、性别、遗传、环境等因素均与治疗本身无关;
- ▶ “潜在结果” (Y_0, Y_1) 是处理前变量, 他们是“虚拟的/假设的”, 不影响处理/治疗措施, 也不受处理/治疗措施的影响;
- ▶ 因此, $(Y_0, Y_1) \perp X$
- ▶ 这意味着,

$$E(Y_0) = E(Y_0|X=0) = E(Y_0|X=1) \quad (1)$$

$$E(Y_1) = E(Y_1|X=0) = E(Y_1|X=1) \quad (2)$$

- ▶ 另外, 基于“一致性的假设 consistency assumption”

$$E(Y_0) = E(Y_0|X=0) = E(Y|X=0) \quad (3)$$

$$E(Y_1) = E(Y_1|X=1) = E(Y|X=1) \quad (4)$$



$$E(Y_0) = E(Y_0|X=0) = E(Y|X=0) \quad (5)$$

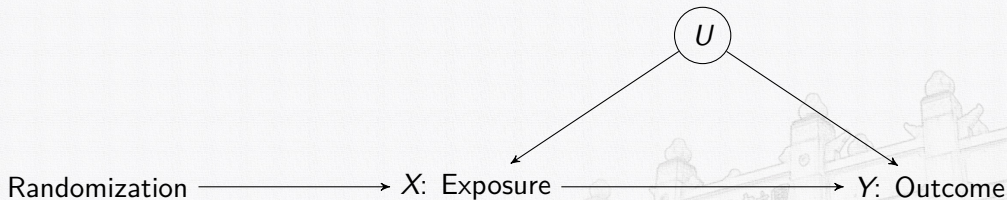
$$E(Y_1) = E(Y_1|X=1) = E(Y|X=1) \quad (6)$$

▶ 可以计算差值

$$E(Y_1) - E(Y_0) = E(Y|X=1) - E(Y|X=0) \quad (7)$$

▶ 可以计算比值

$$E(Y_1)/E(Y_0) = E(Y|X=1)/E(Y|X=0) \quad (8)$$



- ▶ 临床试验以病人为研究对象，目的是评价疾病的治疗、预防效果等
- ▶ 干预措施的分配方案应该以减少“可能影响组间比较的外部因素”为目的
- ▶ 如果参加研究的医生（研究发起者）特别偏爱某一种治疗方式，那么这个医生的患者就会被“干预措施的分配”所影响
- ▶ 另外，如果具有某些特征的病人更倾向于接收到某种治疗方式，那么研究结果的可靠性也会打折扣
- ▶ 以上两种情况，都有一个专有名词：指征混杂（confounding by indication）

- ▶ 为了避免出现类似指征混杂的情况出现，在临床试验中，最好采取随机的方法，来控制因不同治疗组的不同所造成的影响
- ▶ 这些因素包括：特别是患者的个体临床特征、医生的偏好等
- ▶ (医生的偏好，可以用来做其他的分析)
- ▶ 而，随机，是处理混杂（患者的特征、医生的偏好等）最好的方法
- ▶ 临床研究中的有效性 validity，很大程度上取决于基线时随机化的执行情况



- ▶ 当以上这些条件满足时，随机化做的很好，任何未被测量的其他的混杂因素所造成的影响，可以认为是随机的
- ▶ 随机，可以被统计学模型处理
- ▶ 并且，随着样本量的增大，其影响越小
- ▶ 当以上条件不满足时，未被测量的混杂因素就对研究结果造成影响
- ▶ 然而，当这种情况发生的时候，通常我们没有采集到足够的信息来定义指征偏倚，其对结果的影响就会很大



- ▶ 虽然随机化的方法，通常按照相同的概率将研究对象分配到不同的组，然而，这不是必须的，有时候也不是最优的选择
- ▶ 通常情况下，我们基于文献、生物学机制、或者前期的试验研究，会假设一种干预方法比另外一种干预方法好
- ▶ 这种情况下，适应性临床试验 Adaptive randomized trials 或者非比例分配的临床试验可能会更节约成本（比如，纳入更多的干预组患者）
- ▶ 只要知道分配的概率，任何分配比例都可以进行分析

- ▶ 研究对象对干预措施的不依从导致了，随机分配的干预措施与实际接受到的（实施的）干预措施之间不一致
- ▶ 通常的分析方法是比较“随机分配”对结局的差别，而不是“实际接收”到的干预措施对结局的差别。这种分析方法称为意向性分析（intent-to-treat），因为这种分析方法是基于原来最初的治疗意愿，而不是实际接受到的治疗。
- ▶ 虽然这种分析方法有助于保留“治疗效果评价检验”的有效性，但是它仅仅是估计了“随机分配”对结局的影响，而不是“治疗方式”对结局的影响
- ▶ 如果要计算“治疗方式”对结局的影响，使用的统计方法是遵循研究方案分析（per protocol 或者 as treated）分析，这种分析方法其实就是一般的观察性研究的分析方法，收到混杂因素的影响，对结果的有效性有影响（结果是有偏倚的）

- ▶ 如果对研究对象进行了几年的干预，尚在研究中的研究对象可能仅仅是最初随机分配到某一个组的一小部分人群，这部分人群与其他治疗组的研究对象在外部因素方面（基线特征、随访中的变量等）各个方面，都有可能（极大的可能）是不一样的
- ▶ 研究对象是否遵守研究方案（依从性），可以通过直接询问、看看使用的药品盒子等方式获知。最直接、准确的方法是测量血液中血药浓度
- ▶ 测量依从性可以辅助分析，甚至是必须的，比如工具变量的检验等
- ▶ 可以使用 IPW (inverse probability weighting) 或者 G-estimation 的方法调整依从性，但是需要对混杂因素和依从性进行重复测量 (time-varying variables)

- ▶ 即使 RCT 的所有条件/假设/实施都非常完美，研究结果的外推性也较为局限
- ▶ 所以，RCT 的研究结果的外推需谨慎
- ▶ 在临床试验领域，“efficacy”用在所研究的人群中 study population，“effectiveness”用在目标人群中 target population



- ▶ 实用性临床试验的目的是完善 RCT 的不足之处（外推性），在真实世界中，研究药物/治疗方式的效果，
- ▶ 基本上不对研究对象做出限制
- ▶ 与 RCT 类似，也有纳入和排除标准，但是其涵盖的人群更广泛（多种族、性别、年龄）
- ▶ 因其研究的人群具有广泛的代表性，故结果外推性很好
- ▶ 然而，因其不是严格的随机对照，混杂、选择偏倚仍然是个问题

- ▶ 传统的临床试验与实用性试验研究通常互相比较
- ▶ 传统的临床试验通过纳入非常严格定义的一群人作为研究对象，其强调研究的内部有效性
- ▶ 实用性临床试验通过纳入非严格定义的一群人作为研究对象，其强调外部有效性（外推性），研究所需要的变量通常通过：电子病历、电子健康档案、注册数据库获得
- ▶ 然而，注册数据库等数据库中包含的有效信息很少，可能无法完全模拟一个 RCT 研究

- ▶ 当伦理上和实际上可以给予 assign 研究对象某暴露因素, 在理论上就可以人为的设计 (make) 一个研究队列, 当暴露因素不存在时, 两个组 (暴露组和非暴露组) 的疾病发生率是一样的; 这样的队列是消除了混杂因素的队列
- ▶ 如果研究者知道哪些因素会影响疾病的发生, 研究者同样可以设计一个队列, 使得导致疾病发生的因素在没个组之间是平衡的
- ▶ 随机化在大样本中可以很好的控制混杂, 但是在小样本中不是特别有效