中华流行病学杂志2010年1月第31卷第1期 Chin J Epidemiol, January 2010, Vol.31, No.1

·基础理论与方法·

阶梯设计在随机对照试验中的应用

杨祖耀 詹思延

【导读】 介绍阶梯设计的基本原理和实施方法，探讨其在随机对照试验中的应用。当研究者希望在一定范围内全面推行某项通常来说"利大于弊"的措施，又想通过随机对照试验来对该措施的效果进行评价，尤其是当资源有限使得本来就只能分步实施干预的时候，阶梯设计非常适用。另外，该方法还可用来发现或控制时间趋势对效果评价的影响。但与传统设计相比，阶梯设计的试验周期更长，统计学问题也要复杂得多。因此，在设计和实施的过程中应严格把握有关注意事项，以保证研究结果的稳健性。

【关键词】 阶梯设计；随机对照试验；设计延迟；分阶段引入；分阶段实施

Application of "stepped-wedge design" methodology in randomized controlled trials *YANG Zu-yao, ZHAN Si-yan. Department of Epidemiology and Biostatistics, Peking University Health Science Center, Beijing* 100191, China

Corresponding author: ZHAN Si-yan, Email: siyan-zhan@bjmu.edu.cn

This work was supported by a grant from the National Key Technology R&D Program of China's 11th Five-Year Plan (No. 2007BA124B03)

【Introduction】 In this article, two research cases are employed to show the rationale of the stepped-wedge design, under what situations that such a design is desirable, and how it can be implemented. Stepped-wedge design seems to suit to randomized controlled trials in which the entire study population will receive intervention programs as they would "provide more advantages than harm". When intervention can not be given to all the targets simultaneously due to limited resources, this design is particularly useful. The stepped-wedge design is also relevant when there is a hope to detect or control the time trend effect on the effectiveness of the intervention strategy. On the other hand, however, this design requires longer trial duration and presents a number of statistical challenges. Hence, careful planning and monitoring are essential to ensure that a robust evaluation is undertaken.

【Key words】 Stepped-wedge design; Randomized controlled trials; Designed delay; Staged/phased introduction; Phased implementation

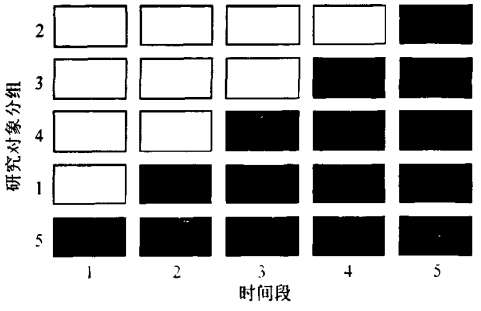
随机对照试验(RCT)是目前评估医学干预措施效果最严谨、最可靠的科学方法[1]。通常情况下，不管是平行设计还是交叉设计，均要求在大约一半的研究对象中同时实施所研究的干预措施。而在实际操作的过程中，这样的要求并不总是能得到满足。例如，当所要研究的干预属于一般被认为是"利大于弊"的措施时，只在部分研究对象中实施干预(平行设计)，或是从正在接受干预的研究对象中撤出干预(交叉设计)，容易引发伦理问题。又比如，当需要接受干预的研究对象较多而人力、物力或财力等资源有限时，研究者可能无法在这么多研究对象中同时给予干预措施。为此，Tom Chalmers在1968年首先建议了一种后来被称为"设计延迟"(designed delay)的方法[2]；1979年Cook和Campbell[3]提出了与之类似的"实验式分阶段引入"(experimentally staged introduction)的方法。1986年冈比亚肝炎干预研究(the Gambia Hepatitis Intervention Study, GHIS)第一次将"实验式分阶段引入"的思想付诸实践[4]，并把它称之为"阶梯设计"(the stepped-wedge design)[5]。近几年来，采用这种设计的研究明显增多[4]。鉴于此，本文介绍该设计的基本原理，并通过研究实例加以说明。

基本原理

阶梯设计多用于整群随机试验，评价"利大于弊"的干预措施，例如免疫接种、筛检、健康教育、医护人员培训等[4]。这种设计有两个基本特点，一是它通常不设置专门的对照组，随着试验的进行，各个组(每组可有一个或多个群)都将接受干预。之所以这样做，是因为最终达到"有益"措施的全面覆盖，既符合卫生决策者的目标，也能在很大程度上减轻试

DOI: 10.3760/cma.j.issn.0254-6450.2010.01.023
基金项目："十一五"国家科技支撑计划(2007BA124B03)
作者单位：100191 北京大学公共卫生学院流行病与卫生统计学系
通信作者：詹思延，Email: siyan-zhan@bjmu.edu.cn

验的伦理学负担。阶梯设计的第二个特点是各个组不在同一时间，而是按照随机的顺序相继接受干预。以一个分5次进行干预的阶梯设计为例（图1），其实施步骤：先把研究对象分成5组，编号为1～5，并将干预的时间划分为5个时间段，一般来说，这5个时间段的长度应该大致相等。通过查阅随机排列表或抽签等方式，获得数字1～5的一个随机排列顺序，如5→1→4→3→2。按照该顺序，在第1个时间段，编号为5的组将开始接受干预，其余的组则处于"等待干预"状态；在第2个时间段，编号为1的组将开始接受干预，此前已经接受干预的5组仍继续接受干预或保持干预后状态，其余的三个组"等待干预"；同理，在第3个时间段，编号为5、1、4的三个组接受干预或保持干预后状态，其余的两个组"等待干预"；以此类推，至第5个时间段结束时，所有组都成了干预组。



注：黑色框表示正在或已经接受干预，空白框表示尚未接受干预

图1 阶梯设计原理示意图

由于各组的状态都发生了从"未干预"到"干预"的单向转换，因而可以把阶梯设计看成是交叉试验的一种特殊形式[6]。当各组研究对象为动态队列时，有两种结局比较方式可以实现"随机（同期）对照"：①分别对各个时间段内接受干预的对象与其他组未接受干预的对象进行比较；②将达到全面干预之前的各个时间段中接受与未接受干预者的结局数据分别汇总（对应于图1中所有白色框和第1～4时间段的所有黑色框），然后再对两者进行比较，这种方式兼有同期对照和前后对照的属性。当各组研究对象为固定队列时，会导致退出观察或失访的事件，如严重疾病、死亡等，不宜作为结局指标，否则可能会引入健康幸存者偏倚（healthy survivor bias）[4]。如采用上述第一种方式进行结局比较，选择的结局指标应该是在给予干预后一个时间段以内就能观察到的，不用通过更长时间的随访来获得；如要采用上述第二种方式，则要求在第1～4时间段都有可能观察到结局。

采用阶梯设计的常见原因是，既想在一定的人群范围内全面实施某"有益"的干预，又想通过RCT评价其效果，有的时候还可能因为受资源（人力/物力/财力）的限制，本来就难以在所有组同时给予干预措施。在此设计中，虽然各组最终都将变成"干预组"，但并不意味着所有的研究对象都能接受干预。当研究对象为动态队列时，在某一组中，接受了干预的就只是那些自该组引入干预时起进入或仍在队列中的人。阶梯设计还可用来发现及控制"时间趋势"（time trend）对效果评价的影响。在开展一些为期较长的试验时，随着时间的推移或某些因素（如季节、卫生条件等）的改变，疾病的发生、发展等情况可能本身就会发生变化，从而给干预效果的评价造成偏倚。采用阶梯设计有助于展示是否存在这种效应，并对其进行调整。

近年来已有研究者将阶梯设计应用于个体随机试验[7-9]，在决定接受干预的先后顺序时，随机化的单位不再是"组"，而是个体，每个个体都可看作是一个"固定队列"，因此其原理与上述研究对象为固定队列的整群随机试验相似。

实例分析

1. 整群随机试验：20世纪80年代，冈比亚儿童的HBV感染率非常高，构成重大的公共卫生问题。当时已有一些试验证明接种乙肝疫苗在高危人群中预防急性HBV感染及携带状态的有效性及安全性，但少有关于其长期效果的研究。1986年GHIS工作组希望利用在全国开展乙肝疫苗接种的机会对出生后9个月内接种乙肝疫苗（0、2、4、9个月各一针）预防HBV感染、慢性肝病和肝细胞癌的效果进行评价[5]。但由于当时乙肝疫苗较昂贵，且供应有限，不可能一开始就在全国范围内同时接种；此外研究者希望能够通过与同期对照的比较来评价疫苗接种的预防效果；如果进行个体随机试验的话，需要一个庞大的工作队伍，以现有的力量很难做到，且容易引发伦理问题。因此，GHIS工作组采用了阶梯设计整群随机试验的方法。为叙述方便，本文仅以第一针（0个月）为例，说明其实施过程，其余的三针依此类推。

乙肝疫苗接种工作组根据冈比亚全国卫生服务中心的分布，将划分的17个地区由17个团队按照随机的顺序在不同的时间启动疫苗接种工作，相继开始接种的两个团队启动的时间相差10～12周（即前文所说的一个时间段的长度）。接种对象为在整个干预阶段时间内该负责的地区出生的所有新生儿（与图1中黑色框的含义近似）。就接种团队而言，

试验期间每个地区的新生儿都是一个动态队列，因为每个时间段接种的新生儿与其他时间段的新生儿不一致。除了第一个启动的团队以外，在其他团队所负责的地区内，从第一个时间段开始到相应团队启动之前出生的新生儿都不接种乙肝疫苗（与图1中空白框的含义近似），从而形成"同期对照"。经过3年多的时间，接种工作将覆盖到全国范围。在此期间出生的新生儿中，接种和未接种乙肝疫苗者大约各占50%。研究者拟对这些人进行30～40年的随访，然后分别比较各个时间段出生的人中接种者和未接种者HBV感染率、携带率、慢性肝病发生率和肝细胞癌发生率。这样比较的好处是，即使不同时间段出生的新生儿发生上述各种疾病的风险不同（时间趋势），对乙肝疫苗效果的评价也不会受到影响。GHIS的随访预计在2017—2020年结束，因此其最终结果如何尚不清楚。

2. 个体随机试验：临床试验的结果表明，异烟肼预防性治疗（IPT）能够降低HIV感染者中结核的发生率。但由于实际操作中存在的困难，IPT并没有得到广泛的实施，其在实际的临床环境中效果如何不得而知。南非某金矿公司的诊所常规提供IPT服务，为评价此诊所对HIV感染者结核发生率的影响，Grant等[9]开展了一项阶梯设计个体随机试验。采用这种设计的原因是：①无需专门设置不去该诊所就诊的对照组，合乎伦理学要求；②该设计可以控制"时间趋势"对效果评价的影响。

研究者给金矿公司的所有2135名HIV感染者每人分配一个计算机生成的随机数字，按照该数字大小决定先后顺序，分别与他们面谈（1999年7月），并应邀到上述诊所就诊。对于同意就诊者，研究者根据他们参加面谈的顺序与之预约相应的就诊日期，顺序仍然按照上述计算机生成的随机数字排列，即就诊顺序随机化；对于拒访者，也可按研究人员的联系方式预约。凡是预约者，从1999年7月1日起至就诊日期（对于未去诊所的人，则为自1999年7月1日起至因各种原因退出或失访的日期）称为"诊所前阶段"，其意义与图1中空白框近似。凡是预约就诊者，从第一次去诊所之日起至随访结束或截尾日期称为"诊所后阶段"，其意义与图1中的黑色框近似。医生对就诊者进行HIV、结核相关检查，符合标准者将给予IPT。定期随访"诊所前阶段"和"诊所后阶段"的结核发生情况，同一人的多次结核发作全部计算在内，以避免一次发作后即停止随访而造成健康幸存者偏倚。分别汇总所有"诊所前阶段"和"诊所后阶段"的数据，用泊松随机效应模型比较两者的结核例次发生率，从而评价该诊所对HIV感染者结核发生率的影响。

整个试验期间，共有1655人预约就诊，贡献的"诊所前阶段"为1595.1人年，期间共发生结核190例次，例次发生率为11.9/100人年。1016人至少去一次诊所，纳入分析的"诊所后阶段"共为709.0人年，期间共发生结核64例次，例次发生率为9.0/100人年，由此得到粗 *RR* 值为0.78（95% *CI*：0.58～1.05）。该试验持续两年余，此期间患者相继就诊，随HIV感染病情不断进展，所有人发生结核的机会都会比试验开始时有所增大，因此诊所对结核发生率的作用就可能受到患者本底风险变化的影响而被低估。为了控制这种时间趋势对效果评价造成的偏倚，研究者对随访时间（自1999年7月起，以6个月为单位转换成有序分类变量）进行调整，调整后 *RR* 值为0.68（95% *CI*：0.48～0.96）。在多因素分析中，调整了随访时间、年龄和矽肺评分后，*RR*=0.62（95% *CI*：0.43～0.89）。可见，设立一个常规提供IPT的诊所后，HIV感染者的结核发生率降低了38%，但仍然保持在较高水平。

讨　论

阶梯设计的优点是在没有设立专门的对照组，且允许分多步进行干预的情况下，仍能保持传统RCT"随机（同期）对照"的优势。当卫生决策者希望在一定范围内全面推行某措施，又想通过RCT来对该措施的效果进行评价，尤其是当资源有限只能分步实施干预时，阶梯设计非常适用。另外，该方法还可用来发现及控制时间趋势对效果评价的影响。但该方法也存在统计学和组织实施等问题。

1. 统计学问题：在RCT的各种类型中，等比例分配平行设计的统计效能是最高的，其他变体与之相比都有或多或少的下降[1]。就阶梯设计而言，可能影响统计效能的因素包括：

（1）阶梯设计效应（stepped wedge design effect）：Moulton等[10]进行数据模拟的结果显示，在样本量一定、群数一定且每个群大小相等的条件下，阶梯设计的统计量 *Z* 值总是比两组平行设计的 *Z* 值小，从而导致拒绝零假设的可能性降低，即没有足够的把握度来发现可能存在的差异。这就是"阶梯设计效应"。由于这个缘故，GHIS的统计效能只能达到传统设计的70%[5]。如要使两者接近，阶梯设计需纳入更大的样本量。

(2)干预的步数:Hussey和Hughes[6]的研究表明,在群数一定的情况下,干预的步数越少,即每一步同时开始接受干预的群数越多,统计效能越低。例如,假设有24个群将接受干预,如果试验分8步进行,即3个群/步,其效能可以达到0.9以上;在其他条件不变的情况下,当随机化的步数分别减少到4、3、2步,即每步的群数分别变成6、8、12个时,效能将分别降低到0.7~0.8、0.6、0.4左右,这种降低主要是由测量次数变少所致,而不是因为随机化的步数减少本身。但实际上,一步只在一个群中实施干预并不可行,因为这样会延长整个试验所需的时间,尤其是在群数较多的时候。因此,在设计阶段应注意权衡两者的关系,以保证较高的统计效能。

(3)治疗效应延迟(treatment effect delay):是指在某一时间段实施的干预,其效果要在一个或多个时间段以后才能完全显现出来的现象。治疗效应延迟会显著地影响统计效能,延迟的越多则效能的损失越大。在干预阶段结束以后延长随访时间可以恢复部分效能,但通常不能完全恢复。最好的做法是在设计阶段即考虑到这一点,将每一个时间段的长度定得足够长,以使在该时间段实施的干预效果能够在下一步开始之前完全显现出来[6]。

虽然大部分研究都是对"干预"和"对照"的数据进行比较,但所用的分析方法却因为具体的研究目的、设计思路、研究对象的形式(固定/动态队列)及疾病本身特点等因素的不同而有很大差异。如前所述,GHIS通过比较每一个时间段中接受和未接受干预者的结局来评价接种乙肝疫苗的效果,并考察在不同时间段干预的效果是否相同[5]。Priestley等[11]关于重症监护培训效果的研究则同时采用了三种分析方法互相验证。当所得结果近似时,以其中的某一种为准;若所得结果相悖,则认为不宜下定论。Allegri等[12]关于社区卫生保险的研究对不同时间段的数据采用了不同的比较分析方法,相对而言更为复杂。即使出于同样的分析目的,不同研究运用的统计模型也可能有所差别。例如,尽管都是为了控制时间趋势,Grant等[9]采用泊松分布模型,而Ciliberto等[13]采用的则是线性和logistic模型。

2. 组织实施问题:

(1)试验周期:与平行设计的RCT相比,阶梯设计的试验周期要长得多。在不进行长期随访的情况下,阶梯设计所需的时间是平行设计的数倍。

(2)盲法:在阶梯设计中,几乎不可能对研究对象和实施干预者设盲,因为从"无干预"到"有干预"是一个很明显的过程。所以,要想减少信息偏倚,对评估者实施盲法就显得尤为重要。

(3)沾染:采用阶梯设计试验,其干预措施往往是"利大于弊",且研究对象很容易获知哪些组正在接受干预。所以,应注意防止"等待干预"组和"正在接受干预"组之间发生沾染,以免对结果的真实性造成歪曲。

总之,阶梯设计为由于各种限制而不适合采用平行设计或交叉设计的随机对照试验提供了新的选择,但与传统设计相比,该方法所要求的试验周期更长,统计学问题也要复杂得多,多种因素如阶梯设计效应、干预步数的多少、治疗效应延迟、盲法的实施等,都可能会对结果的稳健性产生影响。因此,要想利用阶梯设计较好地评价干预措施的效果,必须进行谨慎、严格的计划和实施。

参 考 文 献

[1] Tang JL, Jiang Y, Zhang HW. Randomized controlled trial//Li LM. Epidemiology. 6th ed. People's Medical Publishing House, Beijing, 2007: 129-163. (in Chinese)
唐金陵,江宇,张宏伟. 随机对照试验//李立明. 流行病学. 6版. 北京:人民卫生出版社,2007:129-163.

[2] Maclure M. Tom Chalmers 1917-1995, Part 2: The tribulations of a trialist. Can Med Assoc J, 1996, 155: 986-988.

[3] Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Boston, MA: Houghton Mifflin Company, 1979.

[4] Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. BMC Med Res Methodol, 2006, 6: 54-62.

[5] The Gambia Hepatitis Study Group. The Gambia hepatitis intervention study. Cancer Res, 1987, 47: 5782-5787.

[6] Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemporary Clinical Trials, 2007, 28: 182-191.

[7] Wilmink TBM, Quick CRG, Hubbard CF, et al. The influence of screening on the incidence of ruptured abdominal aortic aneurysms. J Vasc Surg, 1999, 30: 203-208.

[8] Levy RW, Rayner CR, Fairley CK, et al. Multidisciplinary HIV adherence intervention: a randomized study. AIDS Patient Care STDs, 2004, 18: 728-735.

[9] Grant AD, Charalambous S, Fielding KL, et al. Effect of routine isoniazid preventative therapy on tuberculosis incidence among HIV-infected men in South Africa. JAMA, 2005, 22: 2719-2725.

[10] Moulton LH, Golub JE, Durovni B, et al. Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. Clin Trials, 2007, 4: 190-199.

[11] Priestley G, Watson W, Rashidian A, et al. Introducing critical care outreach: a ward-randomised trial of phased introduction in a general hospital. Intensive Care Med, 2004, 30: 1398-1404.

[12] Allegri MD, Pokhrel S, Becher H, et al. Step-wedge cluster-randomised community-based trials: an application to the study of the impact of community health insurance. Health Res Policy and Systems, 2008, 6: 10.

[13] Ciliberto MA, Sandige H, Ndekha MJ, et al. Comparison of home-based therapy with ready-to-use therapeutic food with standard therapy in the treatment of malnourished Malawian children: a controlled, clinical effectiveness trial. Am J Clin Nutr, 2005, 81: 864-870.

(收稿日期:2009-06-16)
(本文编辑:张林东)