·问题与探讨·

观察与实验 效力与效果

唐金陵 杨祖耀

【导读】 作者以2012和2013年在中国医师协会循证医学年会上的两次讲座内容为基础，重点分析“观察”与“实验”的区别以及“效力”与“效果”的关系，讨论评估效力和效果时研究类型的选择，强调现实世界研究的大数据增加的精确性不能替代实验研究的真实性，并依此阐述大数据观察性现实世界研究在评估疗效中的作用。文中的讨论也有助于决策者根据研究设计判断证据的真实性，以便更好地进行医学实践。文后附有特邀专家对该文的点评。

【关键词】 疗效；实验研究；随机对照试验；观察性研究；现实世界研究

Observation versus experiment, efficacy versus effectiveness *Tang Jinling, Yang Zuyao. Division of Epidemiology, JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong SAR, China; Shenzhen Municipal Key Laboratory for Health Risk Analysis, Shenzhen Research Institute of The Chinese University of Hong Kong, Shenzhen* 518055, China
Corresponding author: Tang Jinling, Email: jltang@cuhk.edu.hk
This work was supported by a grant from the National Natural Science Foundation of China (No. 81273171).

【Key words】 Effectiveness; Experimental study; Randomized controlled trial; Observational study; Real world research

1. 观察与实验：现代流行病学是在人群中定量地研究有关健康、疾病以及医疗卫生服务一般规律的方法论[1-3]。其常见的研究设计包括病例系列、横断面研究、病例对照研究、队列研究、随机对照试验和系统综述。按照设计特征，又可分为观察性研究和实验研究，或简称观察和实验。顾名思义，观察性研究是在无研究者影响或控制的“自然条件下”进行的研究，而实验研究则是在研究者完全或部分控制的“非自然条件下”进行的研究，可获得比观察性研究更可靠的结论。

流行病学研究中研究者可控的条件有两方面，一是对暴露状态的分配，二是对其他研究条件（如组间可比性）的控制。研究者对暴露分配的介入就是干预，即研究对象是否处于某种暴露状态（例如是否接受某种治疗）不是自然条件下形成的，而是研究者的主动行为。由研究者主动施加干预的研究称干预研究，在临床研究里称做临床试验，即对一项治疗措施作用的试验（trial）。干预研究常被等同于实验研究。这样区分实验研究和非实验研究（观察性研究），本质上是依据研究问题而进行的分类。然而，实验和观察的本质区别在于其科学性，而非研究的问题。因此，严格意义上讲，如果不具备通过对其他实验条件的控制而获得的组间可比性，这样的“实验研究”与“观察性研究”在结果真实性上无本质区别，同样存在观察性研究常见的混杂偏倚和其他偏倚。只有严格控制其他研究条件和具有组间可比性的干预研究，才可能与观察性研究区分开来，成为真正意义上的实验研究，而随机分组则是实现组间可比性的关键。因此实验和观察的根本区别在于是否采用了随机分组，区别的本质是组间的可比性，而非施加的干预，后者仅仅使研究者实施随机分组和形成真正的实验条件成为可能，并不是实验条件形成的必然因素（图1）。

现代流行病学已将实验研究等同于随机对照试验[2,4-7]，而其他流行病学研究设计均属于观察性研究。在流行病学里常被冠为“实验”研究的目的及其与随机对照试验的区别见表1。特别值得一提的是，并非所有的“临床试验”都是“随机对照试验”或真正意义上的实验，“试验”与“实验”的区别在于前者是测试一个干预的效果，而后者则强调对研究条件的控制。据此可知，临床试验既可以是实验性的，也可以是观察性的。例如交叉试验、序贯试验、析因试验和单人交叉试验，如果分组和交叉是随机形成的，则属于“实验”，否则等同于观察，或称类实验。

DOI: 10.3760/cma.j.issn.0254-6450.2014.03.001
基金项目：国家自然科学基金（81273171）
作者单位：香港中文大学公共卫生及基层医疗学院流行病学部 香港中文大学深圳研究院 深圳市卫生风险分析重点实验室
通信作者：唐金陵，Email: jltang@cuhk.edu.hk

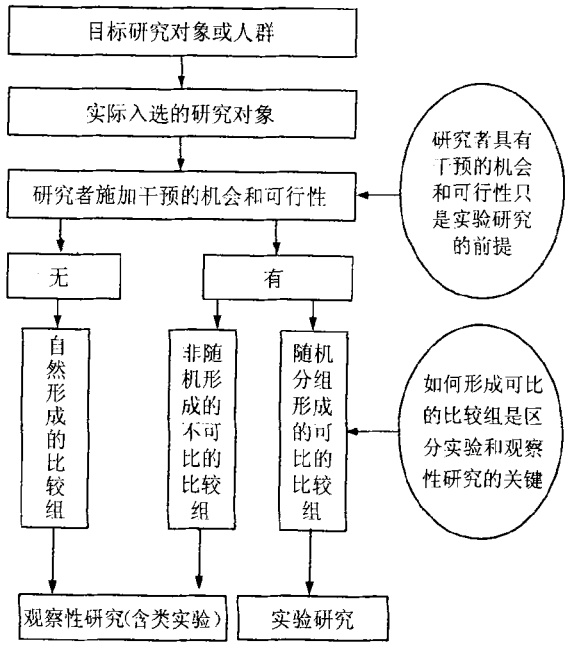


图1　观察性研究与实验研究的区别

用研究的科学性而不是研究目的对流行病学研究进行分类，还在于不同研究类型可用于同一研究目的，同一研究类型也可用于不同的研究目的（表2）[8]。例如在人群中研究干预的作用，并非必须使用实验研究，也可以是观察性的，尤其是初期测试和后期对慢性毒副作用的调查。另外，由于伦理的限制，研究者不能在研究对象人群中人为施加可能有害的病因或危险因素，所以在人类中进行的实验研究只能用于评估可能有益的干预措施[2]。因此，不同的研究问题均有适合自身的最佳、可行的研究设计（表3）[8]。研究类型的区别及其设计的严谨程度决定了流行病学研究结果和结论的真实性。就可用于评估干预效果的研究设计而言，随机对照试验结果的真实性一般高于观察性队列研究（图2）[9-11]，有盲法的随机对照试验结果的真实性又高于无盲法的随机对照试验。因此在研究同一问题时，不同的研究设计就具有不同的定位和作用。本文重点讨论在评估干预作用时，如何根据研究目的选择适合的研究设计。

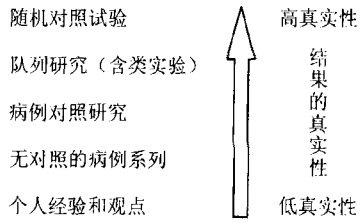


图2　可用于评估干预效果研究的真实性[9]

因此按研究的真实性进行分类，有助于研究者在评估一项干预措施的不同阶段合理地选择研究设计，并利于决策者根据研究设计判断其结果的真实性，更好地开展医学实践[8-11]。尽管如此，本文并不排除流行病学研究其他分类方法的合理性，虽然有些分类同时使用两种或以上的研究特征，易引起逻辑上的混乱。

2. 实验与实践条件的差异：研究设计决定结果的真实性。而研究结果的意义和价值，则取决于研究的问题，即研究的问题越重要，结果的价值就越高。随机对照试验的研究问题由PICOS［即patient/population（患者或接受干预的人群）、intervention（测试的干预措施）、comparator（对照组的干预措施）、outcome（使用的结局指标）和setting（实施干预的环境和条件）］决定，其结果的实践意义也就取决于PICOS的设置[12]。例如，在常规治疗基础上，三甲医院急性心肌梗死住院患者预防性使用利多卡因是否可以降低患者死亡的机会？该研究问题中，P＝急性心肌梗死患者，I＝利多卡因，C＝无利多卡因治疗，O＝死亡，S＝服务水平最高的一类医院。

出于可行性、安全性和科学性的考虑，随机对照试验的PICOS常常与实践存在差异，尤其是初期探索疗效的试验。实验一般是在理想的控制条件下进行的，而实践则发生于现实的自然条件下。比如实验研究多采用诊断明确、病情稳定、依从性高的典型病例，而实践中必须包括非典型病例的治疗；实验研究多采用安慰剂作为对照，而实践中往往需要在不

表1　流行病学中常被冠以“实验”的研究与随机对照试验的区别

研究种类	研究目的及其与随机对照研究（真正的实验研究）的主要区别
干预研究	评估医学干预措施的研究，常被等同于流行病学实验；但其研究设计既可以是实验研究，也可以是观察性研究
临床试验	评估临床治疗措施效果的干预研究，其设计可以是实验研究，也可以是观察性研究
类实验	属于评估干预效果的研究，使用的是非随机分组形成的对照组，或不设平行对照，以自身前后作为对照，其设计更接近观察性研究，又叫半实验或准实验
社区试验	属于评估干预效果的研究。一般流行病学研究的观察和分组单位是一个个体，在社区试验里，观察和分组单位为一个群组（如社区、工厂和学校），等同于临床试验里的群组试验（cluster trial），其设计可以是随机对照试验，但由于受研究群组数量和随机分组可行性的限制，此类研究多是观察性研究
现场试验	属于评估干预效果的研究，但现场试验是一个定义十分模糊的词汇，一般指在公共卫生和预防医学领域、在非临床环境的现场进行的有关干预的研究，可为社区试验，多为观察性研究

表2 流行病学研究设计及其主要应用领域[8]

研究设计	主要应用领域
随机对照试验	干预、筛查、诊断和管理模式效果，副作用
队列研究	病因，副作用，疾病预后和转归，疾病负担
病例对照研究	病因，副作用，诊断
现况调查	疾病负担，卫生服务需求评估，诊断准确性
系统综述	各种研究结果的总结和整理

表3 常见医学问题的最佳研究设计[8]

医学问题	研究设计
疾病负担	现况和队列研究
病因和危险因素	队列研究
预后和预后因素	队列研究
干预（含筛检）的效果	随机对照试验
诊断的准确性	现况研究
罕见的病因	病例对照研究
罕见的药物慢性副作用	病例对照研究
患者和服务现状	现况研究
总结现有的研究证据	系统综述

同的有效治疗中进行选择；实验研究采用的结局指标可能是中间替代变量（如血压），而实践中更关心的是终末变量（如心肌梗死）；研究多是在优越或高于一般的医疗条件下进行，实施者多是有经验的高水平的医生，对各项条件的严格控制使得干预措施的作用趋向最大化，而实际治疗中，由于条件和资源限制，不可能像做研究那样严格控制这些条件，从而使干预措施的作用被多种影响因素“稀释”甚至完全消除。因此研究在理想条件下显示的疗效的大小往往不能在实际医疗环境中得到验证和实现。干预措施还需要在实际或至少是接近实际实践的条件下进一步评估。

3. 效力与效果：为便于区别，通常将理想医疗环境下显示的疗效称作效力（efficacy）、最大效果或理论效果，把评估效力的研究叫做效力试验（efficacy trial）、探索性试验（exploratory trial）或解释性试验（explanatory trial）[11-13]。相反，在实际或接近实际的医疗环境下显示的疗效称为效果（effectiveness），把评估效果的研究称作效果试验（effectiveness trial）、实用性试验（pragmatic trial）或验证性试验（confirmatory trial）[13-16]。由于需要征募大量患者，大规模（多中心）试验（large multi-centre trials，mega trials）多属于评估干预效果试验[17]。因此，效力试验应尽可能在理想的环境中进行，效果试验则应尽可能模仿实际治疗环境。

与效力试验比较，效果试验多采用患者和医生最关心的终末结局（如伤残、死亡和生活质量）来估计治疗作用，并选择现行最好的或常规的治疗作为对照。使用终末结局评估干预措施效果的研究也称结局研究（outcome research），即关于医学实践活动最终结果的研究[18]。而采用现行治疗作为对照研究则属于比较疗效研究（comparative effectiveness research）[19]，比较疗效研究也可以通过系统综述来实现，如网络Meta分析[20]。重要的是，效果试验多是在代表一般实践条件的医疗场所进行的。

然而，在医疗服务条件和医生素质方面，理想和实际的医疗环境是相对的，美国的理想医疗环境高于我国，我国实际医疗环境在不同地区也不尽相同。因此，理想和实际条件并非截然分开的两种独立的情况，而是一个从最好到最差的连续现象。好和差是相对的，一个医院在一个地区可能是比较好的，而在另一个地区则可能是比较差的。因此，从理论上讲，“效力”还存在最大效力和最小效力，“效果”也存在最大效果和最小效果，从最大效力到最小效果应是一个连续现象。当研究效力的服务条件优于研究效果的服务条件时，效力应大于或等于效果。当效力等于或接近效果时，说明该项治疗对医疗环境要求不高，易推广，且使用面宽（如各种同类患者），很多药物治疗都是如此；相反，当效果远远小于效力时，说明该项治疗需要一定医疗条件的保障，不易推广，很多复杂的外科手术即属于此类。

4. 观察和实验的互补作用：关于观察和实验以及效力和效果的概念，在临床流行病学领域已有广泛讨论和认识。在人群中进行药物治疗测试的过程中，为保障安全、缩短测试时间和降低费用，以及降低未来大规模试验失败的风险，首先采用观察性研究初步了解治疗毒副作用和效果是合理的，然后再进行更严谨的随机对照试验以确定疗效。然而，即使是大规模的试验，也不足以检出罕见的慢性毒副作用。因此在药物上市后，再次求助于观察性研究，调查药物可能引起的罕见慢性毒副作用，同样是合理选择。可见观察性研究在药物评估中不可或缺，但观察性研究的重要性并不能取代实验研究（即随机对照试验）在确定治疗效果中的关键性作用。在不同评估阶段，根据不同研究目的，选择不同的研究设计和PICOS，正是西医药物测试的四期临床试验的概念（表4）[4,21-23]。

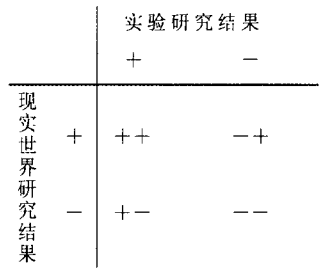
5. 现实世界研究及其作用：由于随机对照试验费用昂贵和实施困难，以及其PICOS组合与实际情况的差异，加之互联网时代大量常规医学数据的出现，有人开始怀疑和挑战随机对照试验在确定干预效果中的作用[24,25]，并提出使用观察性研究最终验

表4　四期临床试验的比较[4,21,23]

项目	Ⅰ期	Ⅱ期	Ⅲ期	Ⅳ期
目的	了解新药的临床药理及人体安全性，观察药物代谢动力学、人体对新药的耐受程度、是否有不可接受的急性毒副作用，为制定给药方案提供依据	初步评价有效性及一般的毒副作用，建立剂量反应关系	确定有效性及常见毒副作用	新药上市后被广泛使用的情况下，开展现实世界研究，并监测罕见严重的慢性毒副作用
研究设计	观察性研究：类实验	观察性研究或实验研究：类实验、小样本的随机对照试验、交叉试验、序贯试验等	实验研究：严格设计的随机对照试验，如大规模多中心试验	观察性研究：队列研究、病例对照研究或常规数据分析
人群/患者	通常是健康志愿者；样本量常为十几至几十	高度选择的目标疾病患者；样本量常为几十至几百	目标疾病的各类患者；样本量常为几百至几千	大批使用过和未使用过该药者，入选标准一般较宽；样本量常为几千以上
干预	多次给药，从初始安全剂量开始，逐渐加大，以确定人体可耐受而无毒副作用的合适剂量	Ⅰ期试验中确定的剂量及给药方式，可能包括≥2个不同剂量组	与Ⅱ期试验类似，一般给药方案比较固定	与Ⅱ期和Ⅲ期试验类似，具体方案可能根据临床实际做调整
对照	无对照组	可以无对照组。如有，可用无治疗、安慰剂、同一药物的不同剂量或已被证明有效的标准治疗作对照	多为安慰剂或已被证明有效的标准治疗	在现实世界中接受测试用药之外的任何可能处理
结局	常见的实验室检测指标，如血球计数、肝肾功能检查、心电图等	疗效的替代结局；常见的急性毒副作用	疗效的终末结局；常见的急性毒副作用	罕见严重的慢性毒副作用，如死亡

证现实条件下的疗效，而且认为这类研究完全可以通过分析常规收集的资料来完成[26,27]。该设想最重要的理由是临床治疗是基于患者实际情况而定，不可能是随机的，尽管随机对照试验可以在十分接近现实的环境中进行，但是就随机分配治疗这一点，它无法反映实际情况，因此随机对照试验的结果不可能反映实际疗效（实效），后者只能通过观察加以验证。如果把现实情况下进行的研究称为实效研究或现实世界研究（real world research）[28,29]，那么现实世界研究就是对疗效的最后测试。“现实”与“理想”相对，强调实际环境与理想环境的差别。亦有学者称之为真实世界研究，但是“真实”与“不真实”或“虚假”相对，随机对照试验的条件并非不真实，只是在常规实践中较难达到而已，因此我们认为用“现实”更妥当。广义地讲，现实世界研究包括观察性研究，也包括在接近现实世界环境中进行的随机对照试验。但在非随机决定治疗的现实世界中，治疗效果只能通过观察进行验证。因此狭义的现实世界研究排除了任何随机对照试验，仅指观察性研究。

那么，现实世界研究（或观察性实效研究）在评估疗效中的意义何在？比较实验研究和现实世界研究的结果，有四种可能：①两种研究均显示治疗有效；②均显示治疗无效；③实验研究显示无效，现实世界研究显示有效；④实验研究显示有效，现实世界研究显示无效（图3）。众所周知，实验研究的真实性高于观察性研究，当两者不一致时，观察性研究结果更可能是错误的。由此推论，在实验研究显示无效时，现实世界研究的进一步验证是没有意义的，只有当实验研究显示有效时，现实世界观察性研究的



注：+ 治疗有效；− 治疗无效；第一、第二个符号分别代表实验研究结果和现实世界研究结果；++示治疗在两种条件下效果相当，容易推广；+−示①现实世界研究可能有误；②现实世界研究可能正确，但效果随PICOS变化的可能性很大，治疗很可能在特殊条件下或在严格质量控制下才会有效

图3　实验研究与现实世界研究结果不同的4种可能性

验证才具有意义。在情况①，现实世界研究验证了效果的存在，说明疗效受实际医疗条件的限制较小，易推广；相反，情况④有两种可能的解释，一是现实世界研究的阴性结果有误；二是其结果正确，但并不能否定疗效，可能说明治疗生效还应满足一定的医疗条件和环境，如医生素质、诊断和护理质量、患者依从性等。

由此可见，在评估干预措施效果时，现实世界研究的特殊作用在于检验理想条件下已经证明有效的措施在一般环境下是否仍然有效。当然，一些效果非常显著的干预措施如断指再植手术、胰岛素降血糖及乙醚麻醉的效果，通过现实世界研究就足以可靠地证明，不必再开展费时费力的实验研究[30]。现实世界研究本质是观察性的，因此存在所有观察性比较研究共同的问题——混杂。在观察性研究里控制混杂的方法有很多，例如在设计阶段可以采用匹配（matching）和限制（restriction），但两者在前瞻性研究

和病例对照研究中控制混杂的能力有限[2,3]。“匹配”在前瞻性研究中费时费力，在病例对照研究中是无效的[2,3]，“限制”在两种研究中都会使入组人群大大减少。在数据分析阶段，可采用标化(standardization)、分层分析(stratified analysis)和多元回归分析(multivariate analysis)[2,3]，其中使用最多的是最后一种，可以同时有效控制多个混杂因素。此外还可使用倾向评分(propensity score)以及控制未知变量潜在混杂的手段(如difference in differences、instrumental variables和regression discontinuity designs等方法)同时控制多个已知或未知的混杂因素[31]。但在控制“混杂”的意义上，上述方法均没有随机分组有效。因此观察性研究的真实性一般低于实验研究，更无法替代实验研究。关于利用常规资料评估干预效果的其他方法学问题，请参见文献[3]。

如果现实世界研究的结果可信，当实验研究显示某治疗有效而现实世界研究显示其无效时，其原因大致有两种。一是现实世界研究的某些要素与实验研究明显不同，例如实验研究中实施某种干预者均为高水平医生，并有高端医疗设备辅助，而现实世界研究的人员及环境不具备此条件；二是现实世界研究的某些要素与实验研究部分不同，例如实验研究纳入的是病情单一、依从性高的患者，而现实世界研究除此之外还纳入了患有其他疾病的复杂病例和依从性低的患者。对于第一种情况，做好过程评价及对治疗条件的评估，有助于探索现实世界研究结果与实验研究存在差异的原因；对于第二种情况，在现实世界研究中可以通过亚组分析，比较各亚组(例如依从性不同的患者)中的治疗效果是否有区别，以解释现实世界研究和实验研究结果的差异。图4总结了评估一项医学干预措施的全过程，显示了不同阶段的研究目的以及干预条件和研究设计的选择。

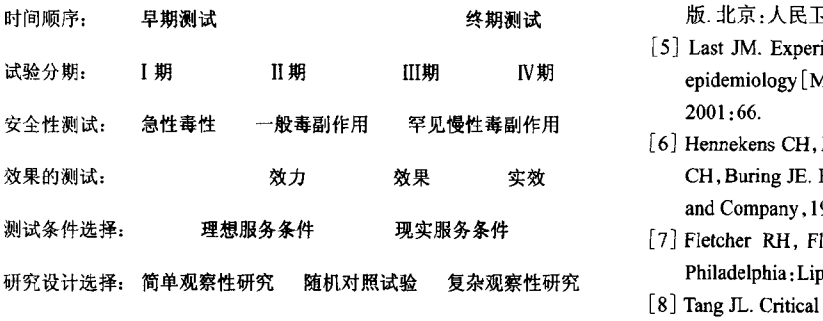


图4 医学干预措施在人群中的测试：测试阶段和测试目的与服务条件和研究设计的选择

6. 结论：实验和观察的根本区别在于是否采用了随机分组，实验研究的真实性高于观察性研究。“效力”是一项干预措施可能的最大有益作用，“效果”是指该措施在实际条件下的作用。“实验和观察”是指不同的研究类型，“效力和效果”是指不同的研究目的。不同研究类型可用于同一研究目的，同一研究类型可用于不同研究目的。干预措施的评估应循序渐进(观察-实验-观察)。实验研究只能用于评估干预，但评估干预的研究并不都是实验研究。实验条件下展示的效果能否在实际条件下得到重复，只能利用现实世界的观察性研究予以验证。但在现实世界中验证疗效的重要性不能否定实验研究的必要性。同理，当实验研究的效果无法在现实情况下重复时，仅凭此不足以否定干预的效果，因为观察性研究的结果可能有误，也可能干预需要高质量的服务条件方能生效。重要的是，现实世界并不是一个单一的状况，而是千变万化，如果一项治疗的效果对治疗条件的依赖性很高，我们不可能在所有具体的现实情况下进行测试，然后决定在什么条件下可以推荐。更实际、可取的方法是，根据研究中的服务标准，改善当地实际的诊疗和服务质量，以获得应有的疗效。

参 考 文 献

[1] Feinstein AR. Clinical epidemiology: the architecture of clinical research[M]. 2nd ed. Philadelphia: WB Saunders, 1985.

[2] Rothman KJ, Greenland S, Lash TL. Modern epidemiology[M]. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008.

[3] Tang JL. Analysis and interpretation of clinical epidemiological data [M]//Li LM. Clinical epidemiology. Beijing: People's Medical Publishing House, 2011: 235-261. (in Chinese)
唐金陵. 临床流行病学数据的分析与结果解释[M]//李立明. 临床流行病学. 北京：人民卫生出版社，2011：235-261.

[4] Tang JL, Jiang Y, Zhang HW. Randomized controlled trial//Li LM. Epidemiology[M]. 6th ed. Beijing: People's Medical Publishing House, 2007: 128-163. (in Chinese)
唐金陵，江宇，张宏伟. 随机对照试验//李立明. 流行病学[M]. 6版. 北京：人民卫生出版社，2007：128-163.

[5] Last JM. Experimental epidemiology//Last JM. A dictionary of epidemiology[M]. 4th ed. New York: Oxford University Press, 2001: 66.

[6] Hennekens CH, Buring JE. Intervention studies[M]// Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown and Company, 1987: 178-212.

[7] Fletcher RH, Fletcher SW. Clinical epidemiology[M]. 4th ed. Philadelphia: Lippincott Williams and Wilkins, 2005.

[8] Tang JL. Critical appraisal of research evidence//Tang JL, Glasziou P. Essentials in evidence-based medicine [M]. Beijing: Peking University Medical Press, 2010: 45-58. (in Chinese)
唐金陵. 医学文献评估概论//唐金陵，Glasziou P. 循证医学基础[M]. 北京：北京大学医学出版社，2010：45-58.

[9] Tang JL, Glasziou P. Introduction to evidence-based medicine// Tang JL, Glasziou P. Essentials in evidence-based medicine[M]. Beijing: Peking University Medical Press, 2010: 1-16. (in Chinese)
唐金陵, Glasziou P. 循证医学概论//唐金陵, Glasziou P. 循证医学基础[M]. 北京:北京大学医学出版社, 2010:1-16.

[10] Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables[J]. J Clin Epidemiol, 2011, 64: 383e94.

[11] Guyatt GH, Oxman AD, Vist GE, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias) [J]. J Clin Epidemiol, 2011, 64: 407e15.

[12] Haynes RB, Sackett DL, Guyatt GH, et al. Clinical epidemiology: How to do clinical practice research[M]. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2006.

[13] Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials [J]. J Chron Dis, 1967, 20 (8): 637-648. [Reprinted in J Clin Epidemiol, 2009, 62(5): 499-505.]

[14] MacRae KD. Pragmatic versus explanatory trials[J]. Int J Technol Assess Health Care, 1989, 5(3): 333-339.

[15] Haynes B. Can it work? Does it work? Is it worth it? The testing of healthcare interventions is evolving [J]. BMJ, 1999, 319 (7211): 652-653.

[16] Roland M, Torgerson DJ. Understanding controlled trials: what are pragmatic trials?[J]. BMJ, 1998, 316(7127): 285.

[17] Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials?[J]. Stat Med, 1984, 3(4): 409-420.

[18] Jefford M, Stockler MR, Tattersall MH. Outcomes research: what is it and why does it matter? [J]. Intern Med J, 2003, 33(3): 110-118.

[19] Golub RM, Fontanarosa PB. Comparative effectiveness research: relative successes[J]. JAMA, 2012, 307(15): 1643-1645.

[20] Mills EJ, Thorlund K, Ioannidis JP. Demystifying trial networks and network meta-analysis[J]. BMJ, 2013, 346: 2914.

[21] Di MY, Tang JL. Adaption and application of the four phase trials to Traditional Chinese Medicines [J]. Evid Based Complement Alternat Med, 2013, 2013: 128030.

[22] Piantadosi S. Clinical Trials[M]. New York: John Wiley & Sons, INC, 1997.

[23] Cancer Research UK. Phase of trials[R/OL]. Accessed at http://www.cancerresearchuk. org/cancer-help/trials/types-of-trials/phase-1- 2-3-and-4-trials on 27 December 2013.

[24] Grapow MT, von Wattenwyl R, Guller U, et al. Randomized controlled trials do not reflect reality: real-world analyses are critical for treatment guidelines! [J]. J Thorac Cardiovasc Surg, 2006, 132(1): 5-7.

[25] Wang YJ. Actively carry out the evaluation of real world effectiveness of drugs[J]. Chin Drug Eval, 2012, 29(1): 1-3. (in Chinese)
王拥军. 积极开展真实世界的药物疗效评价[J]. 中国药物评价, 2012, 29(1): 1-3.

[26] Liu BY. Paradigm of real-world clinical research of traditional Chinese medicine[J]. J Tradit Chin Med, 2013, 54(6): 451-455. (in Chinese)
刘保延. 真实世界的中医临床科研范式[J]. 中医杂志, 2013, 54(6): 451-455.

[27] Gilbody SM, House AO, Sheldon TA. Outcomes research in mental health[J]. Br J Psychiatry, 2002, 181: 8-16.

[28] Tian F, Xie YM. Real-world study: a potential new approach to effectiveness evaluation of traditional Chinese medicine interventions [J]. J Chin Integr Med, 2010, 8(4): 301-306. (in Chinese)
田峰, 谢雁鸣. 真实世界研究: 中医干预措施效果评价的新理念[J]. 中西医结合学报, 2010, 8(4): 301-306.

[29] Robson C. Real world research[M]. 3rd ed. United Kingdom: John Wiley and Sons Ltd, 2011.

[30] Glasziou P, Chalmers I, Rawlins M, et al. When are randomised trials unnecessary? Picking signal from noise[J]. BMJ, 2007, 334 (7589): 349-351.

[31] Craig P, Cooper C, Gunnell D, et al. Using natural experiments to evaluate population health interventions: new medical research council guidance[J]. J Epidemiol Community Health, 2012, 66 (12): 1182-1186.

(收稿日期: 2013-11-20)
(本文编辑: 张林东)

【点评】 循证医学强调证据分级，在评价疗效时“最佳证据”主要来自随机对照试验(randomized controlled trial, RCT)及其系统综述和Meta分析。但经典的RCT通常要求研究对象患单一疾病，采用标准治疗和单一干预措施，从而评价干预措施在理想状态下所能达到的最大效果，即理论疗效或效力(efficacy)。而在临床实际中，患者通常罹患多种疾病，同时接受多种治疗措施，最终疗效是欲研究的干预措施与其他各种处理因素(如治疗方式、管理、辅助治疗等)的综合效果。为了帮助临床医生、患者和管理者更好地进行诊疗决策，仅有理论疗效是不够的，还需要提供这些疗法在“真实世界”中的实际疗效或效果(effectiveness)。因此，近年来疗效比较研究(comparative effectiveness research, CER)得到前所未有的重视。CER的兴起不仅为临床研究开启了新天地，也为流行病学，尤其是观察性流行病学研究方法用于临床疗效评价带来机遇，但观察性研究能否取代实验性研究，二者有何联系与区别，如何根据不同的临床问题选择最佳研究设计等，均是当前迫切需要思考和回答的问题。香港中文大学唐金陵教授和杨祖耀博士撰写的“观察与实验　效力与效果”一文针对上述问题进行了充分讨论，并提出一些独到见解，不仅为读者了解国际此领域新的研究进展提供大量信息，且有助于推动内地学者对流行病学一些本质问题的思考和探讨。

观察性研究主要的特点是研究对象客观存在的各种特征，研究者并不能将研究因素随机分配予观察对象，只能靠全面、客观的描述或精心设计的方案对人群现象进行分析、比较、归纳和判断，以揭示事物之间的

联系。观察法相对于实验法而言，易实施，且伦理学问题相对较少。但研究中也存在多种偏倚，影响结果的真实性。实验性研究将人群随机分为实验组和对照组，人为地给实验组施与措施，如待评价的新药、疫苗接种等，对照组则给予安慰剂，或不给任何措施。在相同的条件下，随访并比较两组人群的结果以判断措施的效果。由于实验研究对象对处理因素的暴露是由研究者随机分配的，能够控制各种外部因素的影响，因此结论可靠，可以论证因果关系假说。当然，有时难以做到严格的随机分组，即为非随机对照试验。

实际上，关于观察与实验在群体医学研究中作用的辨析由来已久。世界上第一个严格意义的RCT设计者Hill[1]爵士于1953年在*New England Journal of Medicine*发表文章曾讨论该问题，认为实验的科学性高于观察，但二者并非对立的关系，好的观察性研究结合恰当的统计学分析可以得出因果的结论，如伦敦宽街的霍乱研究、孕期风疹感染与出生聋儿关系研究以及吸烟与肺癌的研究；而缺乏严格设计的某些疫苗现场试验虽然称为实验研究，其结果未必真实可信。因此，即使难以开展实验研究，至少要有严谨的实验性思维。当然，也有学者持不同观点，如Herman[2]1994年在*Lancet*发表文章指出，过于强调实验限制了临床医学研究的发展，尤其当RCT在不必要、不适当、不可能或者伦理法律不允许的情况下，需要观察性研究来评价保健的效果[3]。1998—2000年发表的几篇系统综述更是将观察与实验的争论再次点燃[4-6]。这几篇基于文献的研究提示尚未发现某种设计可比另一种设计得到更大的效果，而设计良好的观察性研究与同一主题的RCT相比不会系统地高估干预的效果。由此说来，观察性研究似乎可以代替RCT。事实的确如此吗？2001年McMahon[7]通过一个队列研究说明，在药物效力的评价上，即使采用严格的队列设计，尚有混杂问题无法克服，因此观察性研究的作用有限。Ioannidis等[8]的系统综述也支持该结论。什么时候观察性研究像RCT一样可信呢？Vandenbroucke[9]2004年提出，只有兼顾选题、设计和分析三方面的限制(three-pronged restriction)才能帮助观察性研究获得可信的结果。因此，实验与观察都是科学，但应恰当使用[10]。随着大数据时代的到来，电子医疗记录(electronic medical record, EMR)数据库用于疗效评价的设想也被提了出来。鉴于EMR是临床管理工作的记录，而非专门为科学研究所设计，因此数据的质量以及非随机观察性研究的特点就是主要的挑战，需要开发新的统计学分析指标和方法处理已知及未知的混杂，如倾向评分、工具变量、本底事件率比值(prior event rate ratio, PERR)调整等[11]。

当然，如何清晰划分观察性研究与实验性研究，传统上是按干预措施是否由研究者决定和分配加以区分，目前的百科全书，包括国内外大部分教材也是如此分类；而该文作者将“干预”作为前提，是否随机分组作为区分的主要依据，可视为一种学术观点。其实，没有“干预”这个大前提，泛泛地探讨是否随机分组没有意义；但有了“干预”，如果不是严格意义的随机分配，可能会影响对比组之间的可比性。因此，二者的本质区别应该是得到的关系(relationship)不同，观察性研究结果显示的是相关或统计学关联(correlation, association)，而实验性研究结果显示的是因果关系(causation)。总之，如何分类实验与观察是学术界需要不断探讨和逐步达成共识的问题。

参 考 文 献

[1] Hill AB. Observation and experiment[J]. N Engl J Med, 1953, 248(24): 995-1001.

[2] Herman J. Experiment and observation[J]. Lancet, 1994, 344(8931): 1209-1211.

[3] Black N. Why we need observational studies to evaluate the effectiveness of health care[J]. BMJ, 1996, 312(7040): 1215-1218.

[4] McKee M, Britton A, Black N, et al. Methods in health services research. Interpreting the evidence: choosing between randomised and non-randomised studies[J]. BMJ, 1999, 319(7205): 312-315.

[5] Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials[J]. N Engl J Med, 2000, 342(25): 1878-1886.

[6] Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs[J]. N Engl J Med, 2000, 342(25): 1887-1892.

[7] McMahon AD. Observation and experiment with the efficacy of drugs: a warning example from a cohort of nonsteroidal anti-inflammatory and ulcer-healing drug users[J]. Am J Epidemiol, 2001, 154(6): 557-562.

[8] Ioannidis JPA, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies[J]. JAMA, 2001, 286: 821-830.

[9] Vandenbroucke JP. When are observational studies as credible as randomised trials?[J]. Lancet, 2004, 363(9422): 1728-1731.

[10] Chakravarty EF, Fries JF. Science as experiment, science as observation[J]. Nat Clin Pract Rheumatol, 2006, 2(6): 286-287.

[11] Tannen RL, Weiner MG, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings[J]. BMJ, 2009, 338: b81.