·专家论坛·

大型随机对照试验:精准流行病学研究的典范与陷阱

唐金陵 杨祖耀 毛琛
999077 中国香港中文大学公共卫生及基层医疗学院流行病学系
通信作者:唐金陵, Email:jltang@cuhk.edu.hk
DOI:10.3760/cma.j.issn.0254-6450.2017.10.001

【摘要】 现代流行病学是医学应用型研究的方法论,是在人群中定量地研究有关健康、疾病和医疗服务实践问题一般规律的科学和艺术。流行病学研究结果的准确度主要取决于研究的设计类型,精确度主要取决于样本量大小。大型随机对照试验是最精、最准的流行病学研究设计类型,但是由于伦理的限制,只能用于评估医学干预效果。一项研究需要的设计严谨性和样本量与预期效果的大小成反比:效果越小,所需的研究设计就越严谨,需要的样本量就越大。因此,只有当疗效比较小时,才需要大型随机对照试验,当疗效十分明显时,中小型随机对照试验甚至观察性研究就足以证明其有效性。从研究阶段上看,它是确认性、终结性研究,而不是提出假设的原创性研究。然而,研究的价值最终取决于研究问题的意义和原创性,而不是研究方法和*P*值的大小。过度推崇大型随机对照试验会引发:①对中、小疗效干预的过度强调;②对确认性研究的过度重视,以及对项目大小和经费多少而不是科学问题的追逐,进而弱化原创性研究工作;③增加研究资源、医学活动和患者利益被制药公司绑架的风险。

【关键词】 研究设计; 临床试验; 大型随机对照试验; 偏倚; 样本量
基金项目:国家自然科学基金(81273171)

The use and pitfalls of large randomized controlled trials *Tang Jinling, Yang Zuyao, Mao Chen*
Division of Epidemiology, School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong SAR 999077, China
Corresponding author: Tang Jinling, Email: jltang@cuhk.edu.hk

【**Abstract**】 Modern epidemiology is the art and science of investigating quantitatively regularities or general laws regarding applied healthcare issues. The validity of epidemiological studies is primarily determined by the study design and the precision by the sample size. Large randomized controlled trial (RCT) is thus the most rigorous and most precise epidemiological study design. Due to ethical concerns, RCTs can however be used only to evaluate medical interventions. Rigorousness of study design and sample size required for a study are inversely related to the anticipated size of effect to be evaluated: the smaller the effect, the more rigorous the study design and larger the sample size are required. Thus, large RCTs are necessary and called upon when and only when the effectiveness to be proved is relatively small; large effectiveness can be verified with small or medium-sized RCTs or even observational studies. In the stages of scientific research, large RCTs are confirmatory rather than original investigations on new hypotheses, whereas the value of a study is ultimately determined by the importance and novelty of the research question rather than methodology and the P value. Overemphasis on large RCTs has been causing: 1) overemphasis on interventions of small or moderate effect; 2) overemphasis on confirmatory studies and on size of study and funding and weakening original creative work; 3) increasing the risk of research resources, medical activities, and patients' well-being being hijacked by pharmaceutical companies.

【**Key words**】 Study design; Clinical trial; Large randomized controlled trial; Bias; Sample size
Fund program: National Natural Science Foundation of China (81273171)

现代流行病学是医学应用型研究的方法论,是在人群中定量地研究有关健康、疾病和医疗服务实践问题一般规律的科学和艺术。流行病学研究是与实验室基础研究分庭抗礼的两大医学研究阵营之一,可以用来研究疾病的病因和危险因素、疾病诊断的准确性、疾病负担、疾病的转归及其影响因素、医

学干预措施的效果和副作用、医学措施的成本效果等与医学实践直接相关的问题，其发现可直接用来指导和改善医学实践[1-3]。

一、大型随机对照试验是最精准的流行病学研究

1. 研究设计与研究结果的准确度：科学研究的目的是揭示真相，流行病学研究也不例外。以干预措施效果的评估为例，如果一项具体的研究观察到的结果为观察值，真实的效果为真实值，则二者的关系如图1所示。观察值和真实值之间的差别或距离，在流行病学中称为偏倚，也称作系统误差。一个研究的偏倚越大，其观察值离真实值越远，就越不准确，反之亦然。控制偏倚的首要措施是研究设计。研究设计越严谨，偏倚越小，得到的结果就越准确。

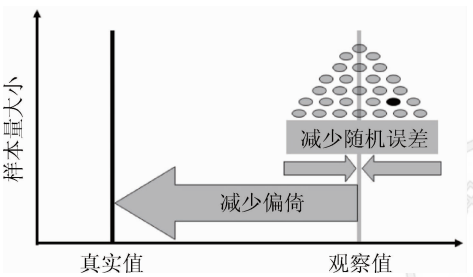

图1　观察值、真实值、偏倚和随机误差的关系

常见的可用于评估干预措施效果的流行病学研究设计包括单病例报告、无对照的病例系列、对照研究（定群研究）和随机对照研究，在临床试验里，它们被分别称为单一病例试验、多病例无对照试验、非随机分组的对照试验，以及随机对照试验[4-5]。所谓试验，就是测试干预措施效果的试验性研究。不同的研究类型就像不同准确度的尺子，有些可以测量到厘米，有些可以测量到毫米，随机对照试验是流行病学研究里可测量到最小差别的尺子，是评估干预措施效果最严谨、最准确、最可信的流行病学研究设计（图2）[1,6]。

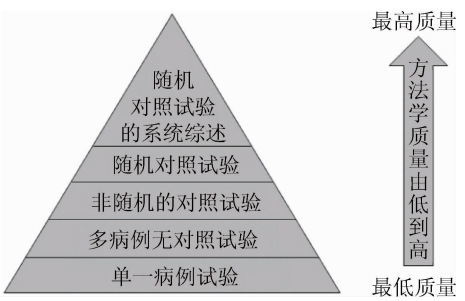

图2　评估干预措施效果的临床试验的设计种类和证据可信度

2. 样本量与研究结果的精确度：如图1所示，即使研究没有任何偏倚，观察值还是不可能与真实值完全重叠。由于随机误差的存在，不同研究得到的观察值不是一个固定的值，而会因样本量的不同围绕一个中心值左右摆动。摆动的幅度与样本量成反比：样本量越大，研究结果围绕中心值摆动的平均幅度就越小，平均来说会更接近中心值，或者说精确度越高，反之则摆动幅度越大，精确度越低。由此可见，精确度与随机误差成反比，主要由样本量的大小决定，增加样本量是提高精确度的主要措施[7-8]。

什么是合适的样本量？一项研究需要的样本量大小与其欲评估的干预绝对效果的大小有关[9]。以简单的两组比较的临床试验为例，估计样本量需要知道对照组的有关事件发生率、治疗效果的大小，统计学的第Ⅰ类和第Ⅱ类错误的概率。第Ⅰ类和第Ⅱ类错误是人为规定的几个固定可选的常数值，随机对照试验需要的样本量主要取决于对照组结局事件的发生率及其与治疗组事件发生率之间的差别（即效果的大小）。一般来讲，干预的绝对效果越小，需要的样本量就越大。如果结局事件的发生率很低，比如在心血管病和癌症的预防研究中，干预的绝对效果无论如何都不会太大，因此需要的样本量都会很大[1,10-12]。

3. 大型随机对照试验：大型随机对照试验又称大规模随机对照试验（large randomized trials 或 mega-trials），样本量可以大至几千，甚至上万。与小型研究或其他设计类型的研究相比，大型随机对照试验结果的精确度、准确度都是最高的，因此，大型随机对照试验就是最精、最准的流行病学研究设计。由于样本量很大，大型随机对照试验在设计上往往还具备一些其他共同的特征，例如多中心、程序简单、病例特征和医疗条件宽泛、治疗环境贴近实际医疗环境、可以估计实际效果、结果有利于推广等。

大型研究为了快速征募病例并完成研究，研究者多会联合很多单位，一个单位就是一个中心，各中心分别征募和治疗自己中心的病例，研究在不同地方同时进行。因此，大型随机对照试验一般都是多中心试验，中心可以局限于一个国家或地区，也可以遍布全世界。由于研究中心遍布各地，每个地区在医疗环境、医疗水平、病例特征等方面存在差异，大型试验多是“简单试验”。所谓简单，就是在病例入选、治疗安排和数据收集等方面采取比小型试验更为简单易行的方法。因为需要征募的病例数量很大，而且涉及的单位很多，各单位的条件和水平参差不齐，研究的各种要求需要考虑效率，还需要兼顾中下水平的参与单位的条件和能力，使它们能够完成

研究的各种要求。比如，病例入选条件应该简单，不添加很多限制（如病程、病情、治疗史和合并症），不使用临床上不常见的仪器，尽可能使临床上同类病例都可以入选，这样才会有更多的合格病例入选，加速病例征募的程序。

由于参与的地区和中心很多，以及病例入选条件和临床治疗环境比较宽泛，如果最终证明有效，一般认为其结果更容易在实际医疗条件下得到重复，更容易推广和普及。由于大型随机对照试验的这些特点，以及它们结果的高准确性和高精确性，大型随机对照试验是目前最有影响的临床研究，医学杂志会争相发表，如*NEJM*、*Lancet*、*JAMA*、*BMJ*几乎每期都有报道，这又进一步推升了大型随机对照试验在临床研究中的地位和影响力。

二、随机对照试验的适用条件

1. 研究问题与随机对照试验的选用：在流行病学研究的范式里，针对同一研究问题，有多种研究设计可以选用，不同研究设计提供的证据可信度高低有别，只有一种是最优的切实可行的研究类型（表1）。例如，欲研究吸烟与肺癌的关系，可选择的研究设计包括：病例系列、横断面研究、生态学研究、病例对照研究、队列研究。队列研究是研究病因切实可行的可信度最高的研究。又如，评估干预的效果，单病例试验、多病例无对照试验、非随机分组对照试验都可以用于评估干预的效果，随机对照试验是可信度最高的切实可行的研究。但是，在研究药物严重、罕见的慢性不良反应时，最可信、切实可行的研究往往是病例对照研究。

表1 医学实践问题与最优可行的研究设计

常见研究问题	最优可行的研究设计
患病率	横断面研究
发病率	队列研究
常见疾病的原因	队列研究
极其罕见疾病的病因和药物不良反应	病例对照研究
不常见疾病的病因和药物不良反应	队列研究
干预效果和常见不良反应	随机对照试验
诊断方法的准确性	横断面研究
疾病的预后及其决定因素	队列研究
系统综述适合于对上述所有研究结果的回顾总结	

评估医学干预的效果，至少有4类研究可以使用，但是它们不是等同的。从设计上讲，主要区别为3个方面：观察从暴露到结局的时间走向，对照的特征，比较组间的可比性。研究设计的核心区别就是在处理这些问题上的不同，研究设计的进步就是在处理这些问题上的进步。不同研究提供的证据的可信度有高有低，随机对照试验最高，单个病例试验最低。提高可信度是有代价的，一般来讲研究需要的资源（包括时间）与结果的可信度成正比，回答一类问题最可信、切实可行的研究往往也是最昂贵耗时的研究。据估计，一个典型的大型随机对照试验平均需要费用高达3 000万美元[13]。进行大型随机对照试验必须谨慎。

其实，评估干预效果时，不是都需要使用随机对照试验的。在效果极其明显时（如白内障手术、断肢再植、麻醉和输血），通常不需要随机对照试验的确认[13]。因为效果极其明显的干预措施毕竟是少数，绝大多数干预措施还是需要经过随机对照试验的验证，但不是所有情况下都需要使用大型随机对照试验。效果明显时（如抗生素），中、小型试验就足以，只有当效果较小时，才需要大型随机对照试验。

2. 研究阶段与随机对照试验的选用：决定研究设计选择的另一个重要因素是评估的阶段。对于同一类研究问题，研究设计的选择主要取决于研究的阶段。

科学研究可大致分为3个阶段：产生假设、检验假设和确认假设。从时间上看，产生假设是研究的初期阶段，检验假设是中期阶段，确认假设是终末阶段。就评估医学干预效果而言，早期阶段应使用安全、快速、简单、省钱的研究，但结果可信度比较低，中期应使用可信度较高的研究，最后再经过随机对照试验的确认。

以新药的临床验证为例，一般分为4个阶段，即常说的Ⅰ～Ⅳ期临床试验（图3）[14-16]。研究的问题包括毒副作用和效果2个方面，对二者的研究也需要循序渐进，研究设计选择也需灵活多变。Ⅰ期临床试验是一个新药第一次在人体上的测试，需十分谨慎，评估的不是疗效而是急性毒性作用，也包括对药物代谢动力学的考察。研究往往是在仔细挑选的健康人中进行的，且没有对照组。如果一个药物通

图3 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

过了Ⅰ期试验，没有明显的急性毒性作用，可进入Ⅱ期试验，对疗效进行初步评估。根据评估的进展时段，可选的研究设计很多，如无对照试验、前后对照试验、交叉试验、非随机的平行对照试验和小样本随机对照试验等。Ⅲ期试验是对疗效最严格的验证，需要使用样本量足够大的随机对照盲法试验。Ⅳ期试验是药物上市后的研究，主要是对严重罕见慢性不良反应的监察，主要使用的是病例对照研究和队列研究。

三、大型随机对照试验的作用和陷阱

1. 大型随机对照试验的作用：为什么要进行大型随机对照试验呢？最根本的理由是研究问题的需要，具体来说是欲证明的治疗的效果很小。由此可见，大型随机对照试验主要是用于评估中小疗效（尤其是小疗效）的干预措施的效果[10]。

换言之，如果一个随机对照试验需要很大的样本量，那么它所估计的绝对效果就一定很小。白内障手术，立竿见影，3～5例患者就足以证明其效果，对照都不需要。抗菌素治疗大叶性肺炎，在几十例病例中测试就足够了。但是，要证明一个抗高血压药物是否可以预防心血管事件，一般需要几千甚至上万病例，随访观察几年的时间。

上述3种治疗之间，第一种肯定是最有效的，第三种是效果最低的。换言之，临床观察就可以确认的效果一定大于小型随机对照试验可确认的效果，小型随机对照试验可确认的效果一定大于大型随机对照试验可确认的效果。从这个意义上讲，我们绝不能说临床观察和小型随机对照试验确认的疗效没有大型随机试验证明的效果重要，其实恰恰相反。

事实上，抗生素、麻醉术、胰岛素、夹板正骨、疖痈引流、大失血后的输血、出血的压迫或包扎等十分有效的医学措施，在随机对照试验诞生以前，就被证明有效且广泛应用了[17]。相反，随机对照试验诞生以后，我们并没有产生多少新的比这些治疗更有效的措施，现在临床上使用最多的抗高血压药物、抗血脂药物、抗肿瘤药物（表2）等，虽然它们的疗效经都过大型随机对照试验的确认，但是它们的效果远远没有上述医学措施的效果明显而快速[18]。

认识到这一点十分重要，在讨论大型随机对照试验结果的临床实践意义时，不能盯在样本量上，不能把样本量和统计学显著性与干预效果大小混为一谈，干预的临床价值在于其疗效的大小，不是样本量的大小。增加随机对照试验的样本量，可以增加测量的稳定性，但是只有当效果很小时才需要大型随机对照试验。换言之，对研究要求越是精、准，说明欲证明的效果就越小。相反，如果效果很大，小型随机对照试验甚至非随机分组的对照试验就足以形成确认性研究。

表2　大型随机对照试验证明有效的医学干预措施典型举例

干预措施	效果
降血压药预防心血管病事件(治疗5年)	NNT≈30
降血脂药预防心血管病事件(治疗3年)	NNT≈90
阿司匹林预防心血管病事件(治疗5年)	NNT≈320
冠心病介入治疗减少死亡(随访2年)	NNT≈20
肺癌术后辅助放化疗减少死亡(随访5年)	NNT≈20
表皮生长因子受体酪氨酸激酶抑制剂(靶向治疗的一种)治疗晚期非小细胞肺癌(不论何种治疗环境和随访期)	延长的生存期≈0
粪便潜血试验筛查结直肠癌(随访15年)	NNT≈700
人乳头瘤病毒基因检测筛查宫颈癌(随访8年)	NNT≈1 000

注：NNT（number needed to treat）：需治人数，表示为了实现1例获益者而需要治疗的人数。例如：NNT＝30，意思是为了使1个人获益，需要治疗30个人；换言之，30个人接受治疗，只有1个人可以从该治疗获益，其他人不但不能获益，还要承受潜在的毒副作用及治疗费用。NNT越大，说明治疗效果越小

如此看来，大型随机对照试验只是评估中、小疗效的金标准，不是评估所有干预措施的金标准。大型随机对照试验也可能用来比较两个疗效差别不大的治疗，证明疗效不存在，或是证明药物间中、小交互作用的存在。但无论是哪种情况，需要大型随机对照试验证明的作用或差别都是比较小的，因此它们的实践意义也是值得拷问的。

2. 大型随机对照试验不是原创性研究：如前所述，对一个科学问题的探索可分为提出假设、验证假设和确认假设3个阶段。从评估干预效果的程序上看，大型随机对照试验不属于早期提出科学问题的原创性研究，而是中后期的确认性研究[16,19]，是最后的研究，是终结性研究，一旦证明有效，将不再需要新的研究继续验证，一般也不会衍生出新的科学发现和发明。

如果把流行病学研究分为创新性研究和验证性研究，前者在于提出崭新的研究问题或假设，后者在于验证初步探索过的假设或理论的正确性。前者是科学发展的基础，是科学的灵魂，因为没有前者就不需要后者；但是如果没有后者，科学就只有一堆不知错对的想象和假设而已，科学将不会扎实地进步。前者更多需要的是灵感和想象力，是科学家的核心价值所在；而后者更多依赖于研究资源和执行能力，是常规科学和科学工匠的工作范围[20]。

由此可见，科学研究（尤其是医学应用型研究）的意义和价值最终取决于研究问题，而不是回答问题的方法的优劣和结果的精准。哲学家Laude

Levi-Strauss甚至说，真正的科学家是那些提出关键问题的人，而不是可以正确回答问题的人。普林斯顿大学统计学系创始人John Tukey也说过：数据分析最重要的原则，也是很多统计学家想规避的原则：对一个正确问题的不精准回答，远远好于对一个错误问题的精准回答。统计分析和流行病学研究的正确运用，其实异曲同工。再好再大的研究，如果没有一流的研究问题，也不可能成为一流的研究。

3. 大型随机对照试验的陷阱：大型随机对照试验是最精、最准的流行病学研究，因此人们经常把大型随机对照试验视为确认医学干预措施效果的金标准[13]。*NEJM*、*Lancet*、*JAMA*、*BMJ*等医学杂志已成为大型随机对照试验热衷的发布平台，它们的宣传和肯定进一步推升和稳固了大型随机对照试验在医学实践中的“霸主”地位。因此，大型随机对照试验成了国际临床指南所追捧的旗帜，通过国际指南巨大而又快速影响着世界范围内的医学实践活动。

然而，大型随机对照试验只是确认中、小疗效的金标准，不是确认十分有效的干预的金标准。在医学实践中，过度推崇大型随机对照试验，会导致对中、小疗效干预的过度强调。所谓中、小疗效的临床意义是值得商榷的[11]，而且在多数情况下中、小型随机对照试验的系统综述完全可以替代大型随机对照试验[19,21-22]。另外，如果设计上出了问题，例如未使用重要的终末结局[23-25]，大型随机对照试验的浪费是巨大的[26]。

绝大多数大型随机对照试验主要是评估新的药物。过度推崇大型随机对照试验，还会导致对新药的过度重视和依赖，弱化很多老的甚至更有效的药物，弱化预防、非药物干预（例如生活方式）和传统医学（例如中医）的作用[13]。

在医学研究中，过度推崇大型随机对照试验，会导致对确认性研究的过度重视，因而间接地弱化开创性研究工作的重要性，导致对大型验证性项目的盲目崇拜，对大样本和统计学显著性而不是临床意义的重视[27]，以及对科研经费而不是科学问题的追逐。长期下去，势必会导致研究领域原创性能力的下降，丢了科学灵魂（表3）。

大型随机对照试验的病例入选条件和临床治疗环境比较宽泛，增加了代表性，提高了结果的外推性，似乎是一个优点。但实际上，这恰恰是大型随机对照试验的另一个更隐秘的陷阱。为了尽快招募到足够的病例，大型随机对照试验经常会放宽纳入标准。在疗效很小的情况下，病例纳入标准越宽泛，从治疗中受益的病例的比例可能就越低，需要的样本量就越大，最后认为在更广泛的病例中治疗有效就越不合理，因为会使更多不会受益的人接受治疗。大型随机对照试验所谓增加了研究的外部真实性，其实是一种误导。

表3 大型随机对照试验的主要应用和陷阱

项目	内 容
应用	确认干预措施中、小疗效的存在 确认不同干预措施之间疗效中、小差别的存在
陷阱	导致对仅具中、小疗效措施或中、小疗效差异的过度重视导致对统计学显著性和样本量而非临床意义的重视 导致对终末确认性研究的过度追逐，弱化原创性研究

另外，值得注意的是，很多大型随机对照试验是由药厂资助的，它们已经成为推广药物最有力的市场工具[28-31]。例如，*NEJM*曾多次报道有关癌症靶向治疗药物的临床试验，但是这些极其昂贵的药物最好也只不过能延长几个月的生存，经常只是延缓肿瘤进展几个月的时间，但并不能延长生存，也不能提高生命质量[23-24,32-33]。这使得我们更应该对大型随机对照试验持客观、谨慎、警觉的态度。

Ioannidis[34]曾说，医学研究中有很多所谓的“阳性结果”和“重要发现”只是具有高度统计学显著意义但没有太大临床意义的结果，或者是在各种动机的驱使下玩弄数字的结果，很难得到重复。大型随机对照试验的结果被重复验证的机会几乎是零，因为大型随机对照试验极其昂贵，针对同一研究问题开展两个或多个如此昂贵的研究的可能性极小。即使有后续研究，由于研究间诸多因素的不同，它们的结果都不大可能否定之前大型试验的结论，从而使得现实中不大可能出现有效的验证研究。这种几乎不可验证的特性，增加了大型随机对照试验被不良动机利用、甚至出现欺骗性结果的风险。

四、大型随机对照试验与大型队列研究的区别

20世纪中叶以后，随机对照试验的兴起以及医学对大型随机对照试验的尊崇和膜拜，正在导致对中、小干预效果的过度重视，对验证性研究的过度重视，对样本量和研究经费的过度追逐，进而导致对原创性研究的淡化。但是，这样的结论似乎不适用于大型前瞻性队列研究。

除了研究设计严谨性之外，大型前瞻性队列研究与大型随机对照试验之间存在两个重要区别。一是随机对照试验一般只能用来回答一个简单的研究问题，即在干预和结局方面都必须做严格的限定，如某药与安慰剂比较是否可以在某特定病例中改变某重要临床结局。队列研究则不然，它可纳入的暴露

往往有几十种几百种甚至更多，对照有很多种可能，这些暴露可能影响的结局又有很多种，可以包括多种常见和罕见疾病[35-37]。因此，通常一个随机对照试验一般只产生一个核心研究报告，而大型队列研究可以产生无数个重要性相当的研究报告。例如，美国的佛明翰心脏研究1950年以来共发表超过3 000篇文章，涉及很多危险因素和疾病，是历史上贡献最大的流行病学研究之一[37]。

大型队列研究与大型随机对照试验的第二个重要区别在于它们可引发新的发现。大型随机对照试验是终结性研究，所谓终结性研究，就是完成以后不再需要新的验证，一般也不会引发出新的科学问题。队列研究则不同，主要用于发现病因，病因是预防和治疗疾病的开始，因此队列研究是控制一个疾病的开端，而不是结束。发现了病因，就可以找到预防的方法，以及治疗的线索。例如，发现细菌是传染病的病因，这个发现导致了后来抗生素和疫苗的发现。再如，发现高血脂可引起冠心病，控制高血脂的危险因素就可以预防冠心病，而很多治疗冠心病药物也是受到血脂和动脉粥样硬化和血栓关系的启发而产生的。

利益冲突　无

参 考 文 献

[1] 唐金陵，Glasziou P. 循证医学概论//唐金陵，Glasziou P. 循证医学基础[M]. 2版. 北京：北京大学医学出版社，2016：1-18.
Tang JL, Glasziou P. Introduction to evidence-based medicine// Tang JL, Glasziou P. Essentials in evidence-based medicine[M]. 2nd ed. Beijing: Peking University Medical Press, 2016: 1-18.

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

[3] Feinstein AR. Clinical epidemiology: the architecture of clinical research[M]. 2nd ed. Philadelphia, PA: W.B. Saunders, 1985.

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

[5] Fletcher RH, Fletcher SW. Clinical epidemiology: the essentials [M]. 5th ed. Baltimore, MD: Lippincott Williams & Wilkins, 2012.

[6] Straus SE, Glasziou P, Richardson WS, et al. Evidence-based medicine: how to practice and teach it[M]. 4th ed. Edinburgh: Churchill Livingstone, 2010.

[7] Altman DG. Practical statisticsfor medical research[M]. London: Chapman & Hall, 1990.

[8] Armitage P, Berry G, Matthews JNS. Statistical methods in medical research[M]. 4th ed. Oxford: Wiley-Blackwell Science, 2001.

[9] Day SJ, Graham DF. Sample size and power for comparing two or more treatment groups in clinical trials[J]. BMJ, 1989, 299(6700): 663-665. DOI: 10.1136/bmj.299.6700.663.

[10] Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials[J]. J Clin Epidemiol, 1995, 48(1): 23-40. DOI: 10.1016/0895-4356(94)00150-O.

[11] Worrall J. Do we need some large, simple randomized trials in medicine//Suárez M, Dorato M, Rédei M. EPSA philosophical issues in the sciences[M]. Dordrecht: Springer, 2010: 289-301. DOI: 10.1007/978-90-481-3252-2_27.

[12] Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials?[J]. Stat Med, 1984, 3(4): 409-420. DOI: 10.1002/sim.4780030421.

[13] Bothwell LE, Greene JA, Podolsky SH, et al. Assessing the gold standard-lessons from the history of RCTs[J]. N Engl J Med, 2016, 374(22): 2175-2181. DOI: 10.1056/NEJMms1604593.

[14] Sedgwick P. What are the four phases of clinical research trials?[J]. BMJ, 2014, 348: g3727. DOI: 10.1136/bmj.g3727.

[15] 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. DOI: 10.1155/2013/128030.

[16] 唐金陵，杨祖耀. 观察与实验效力与效果[J]. 中华流行病学杂志，2014, 35(3): 221-227. DOI: 10.3760/cma.j.issn.0254-6450.2014.03.001.
Tang JL, Yang ZY. Observation versus experiment, efficacy versus effectiveness[J]. Chin J Epidemiol, 2014, 35(3): 221-227. DOI: 10.3760/cma.j.issn.0254-6450.2014.03.001.

[17] Glasziou P, Chalmers I, Rawlins M, et al. When are randomised trials unnecessary? Picking signal from noise[J]. BMJ, 2007, 334(7589): 349-351. DOI: 10.1136/bmj.39070.527986.68.

[18] 唐金陵. 证据选集导读//唐金陵，Glasziou P. 循证医学基础[M]. 2版. 北京：北京大学医学出版社，2016：226-228.
Tang JL. A summary of evidence on the effectiveness of selected medical interventions//Tang JL, Glasziou P. Essentials in evidence-based medicine[M]. 2nd ed. Beijing: Peking University Medical Press, 2016: 226-228.

[19] Lau J, Antman EM, Jimenez-Silva J, et al. Cumulative Meta-analysis of therapeutic trials for myocardial infarction[J]. N Engl J Med, 1992, 327(4): 248-254. DOI: 10.1056/NEJM199207233270406.

[20] Kuhn TS. The structure of scientific revolutions[M]. Chicago, IL: University of Chicago Press, 1962.

[21] Shrier I, Platt RW, Steele RJ. Mega-trials *vs.* Meta-analysis: precision *vs.* heterogeneity?[J]. Contemp Clin Trials, 2007, 28(3): 324-328. DOI: 10.1016/j.cct.2006.11.007.

[22] Antman EM, Lau J, Kupelnick B, et al. A comparison of results of Meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction[J]. JAMA, 1992, 268(2): 240-248. DOI: 10.1001/jama.1992.03490020088036.

[23] Mok TS, Wu YL, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma[J]. N Engl J Med, 2009, 361(10): 947-957. DOI: 10.1056/NEJMoa0810699.

[24] Mok TS, Wu YL, Ahn MJ, et al. Osimertinib or platinum-pemetrexed in *EGFR* T790M-positive lung cancer[J]. N Engl J Med, 2017, 376(7): 629-640. DOI: 10.1056/NEJMoa1612674.

[25] Guyatt GH, Oxman AD, Kunz R, et al. What is "quality of evidence" and why is it important to clinicians?[J]. BMJ, 2008, 336(7651): 995-998. DOI: 10.1136/bmj.39490.551019.BE.

[26] Ioannidis JPA. Mega-trials for blockbusters[J]. JAMA, 2013, 309(3): 239-240. DOI: 10.1001/jama.2012.168095.

[27] Horton R. Offline: what is medicine's 5 sigma?[J]. Lancet, 2015, 385(9976): 1380. DOI: 10.1016/S0140-6736(15)60696-1.

[28] Goldacre B. Bad pharma: how drug companies mislead doctors and harm patients[M]. London: Fourth Estate, 2012.

[29] Moynihan R, Cassels A. Selling sickness: how the world's biggest pharmaceutical companies are turning us all into patients [M]. New York, NY: Nation Books, 2006.

[30] Smith R. Medical journals are an extension of the marketing arm of pharmaceutical companies[J]. PLoS Med, 2005, 2(5): e138. DOI: 10.1371/journal.pmed.0020138.

[31] Macleod MR, Michie S, Roberts I, et al. Biomedical research: increasing value, reducing waste[J]. Lancet, 2014, 383(9912): 101-104. DOI: 10.1016/S0140-6736(13)62329-6.

[32] Kim C, Prasad V. Cancer drugs approved on the basis of a surrogate end point and subsequent overall survival: an analysis of 5 years of US Food and Drug Administration approvals[J]. JAMA Intern Med, 2015, 175(12): 1992-1994. DOI: 10.1001/jamainternmed.2015.5868.

[33] Rupp T, Zuckerman D. Quality of life, overall survival, and costs of cancer drugs approved based on surrogate endpoints[J]. JAMA Intern Med, 2017, 177(2): 276-277. DOI: 10.1001/jamainternmed.2016.7761.

[34] Ioannidis JPA. Why most published research findings are false [J]. PLoS Med, 2005, 2(8): e124. DOI: 10.1371/journal.pmed.0020124.

[35] The Nurses' Health Study. About NHS[EB/OL]. [2017-04-27]. http://www.nurseshealthstudy.org/about-nhs.

[36] Oxford Clinical Trial Service Unit and Epidemiological Studies Unit. British doctors study[EB/OL]. [2017-04-27]. https://www.ctsu.ox.ac.uk/research/british-doctors-study.

[37] Framingham Heart Study. Framingham heart study bibliography [EB/OL]. [2017-04-27]. https://www.framinghamheartstudy.org/fhs-bibliography/index.php.

（收稿日期：2017-05-23）
（本文编辑：王岚）