精神卫生科研如何严格遵守试验设计四原则之重复原则

张效嘉[1]，胡良平[1,2*]

（1. 军事医学科学院生物医学统计学咨询中心，北京 100850；
2. 世界中医药联合会临床科研统计学专业委员会，北京 100029
* 通信作者：胡良平，E-mail：lphu812@sina.com）

【摘要】 明确阐释在进行精神卫生临床试验设计时，应正确把握“重复原则”的意义和要领。从基本常识出发，并基于精神卫生科研的特点，寻找和发现在此研究领域中，怎样做才能被称为严格遵守了“重复原则”。通过结合本专业的特点，并结合实例，获得如下的结果，即在进行精神卫生临床试验设计时，必须把握好以下四个方面：①正确领悟重复的三层涵义；②正确领悟什么是独立重复试验；③应清楚估计样本含量需要满足哪些前提条件；④应至少学会用一种统计软件方便快捷地估计出合适的样本含量。在如何严格遵守重复原则问题上，正确把握好前述提及的四个方面，就是抓住了问题的本质，是提高临床试验研究质量的一个重要环节。

【关键词】 精神卫生；临床试验设计；重复原则；样本含量；检验效能
中图分类号：R195.1 文献标识码：A doi：10.11886/j.issn.1007-3256.2016.04.003

Approaches to strictly observe four principles in the clinical trial design of mental health research: the "replication principle"

ZHANG Xiao-jia[1], *HU Liang-ping*[1,2*]

(1. *Consulting Center of Biomedical Statistics, Academy of Military Medical Sciences, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
* *Corresponding author: HU Liang-ping, E-mail: lphu812@sina.com*)

【**Abstract**】 This paper illustrated that the implication and essential of replication principle should be taken into consideration in the clinical trial design of mental health. By combining common sense with characteristics of mental health research, we explored approaches to strictly observe replication principle in the field. The paper suggested that in mental health clinical trial design, researchers should hold four aspects in terms of replication principle as follows. First, researchers should realize the implication of replication. Second, researchers should recognize the meaning of independent repeated trials correctly. Third, researchers should know clearly that which prerequisites may be satisfied to estimate sample size. Last, researchers may master a statistical software package to perform the sample size estimation. The article concluded that the implement of four aspects mentioned above is a significant element to enhance the quality of clinical trial study.

【**Key words**】 Mental health; Clinical trial design; Replication principle; Sample size; Power

1 概 述[1-5]

1.1 重复原则与样本含量

试验设计中的重复原则是指在特定的条件下应做足够多次数的独立重复试验，以便使随机变量的变化规律能充分地显露出来。而人们习惯把整个试验研究中所用到的受试对象个数称为样本含量。事实上，把每种特定条件下用到的独立受试对象个数称为样本含量更贴切、更有实际意义。因为有时一个试验研究项目仅把受试对象分为两组，但有时需要把受试对象分为 12 组、24 组、48 组甚至更多组。

1.2 重复的三层含义

在生物医学和临床研究中，除了“独立重复试验”之外，“重复”还有其他两层含义：其一，重复取样，即从每位受试对象身上获得一个样品，将其均分成若干份，在同一时间点上对其进行逐一观测，其目的是看各标本中某定量观测指标值的分布是否均匀或检测方法是否具有重现性；其二，重复测量，即在不同时间点或不同部位，对同一位受试对象或取自其身上的样品进行观测，其目的是看定量指标随时

间推移的动态变化情况或部位改变条件下定量指标取值的分布情况。

2 与样本含量估计有关的概念与基础[1-5]

2.1 估计样本含量和检验效能的意义

2.1.1 估计样本含量的意义

正确估计样本含量体现了统计研究设计中的重复原则，可以降低研究中的抽样误差。同时足够的样本含量也是保证试验研究中组间均衡性的基础。若样本含量过小，评价指标的平均值不稳定，意味着抽样误差大，推论总体的精密性与准确性都比较差，统计检验的效能低，实际存在的差别不易真实地显露出来；样本含量过大，会增加实际工作的困难，浪费人力、物力、财力和时间，虽然减少了抽样误差，但由于过分追求数量，可能引入更多的混杂因素，或因工作粗枝大叶导致科研资料不准确，对研究结果造成不良影响。

2.1.2 估计检验效能的意义

检验效能（即能发现客观上存在的差别的能力，也称为把握度，英文用 power 表示）由犯第二类错误概率 β 的大小决定，等于 $1-\beta$。估计检验效能，其意义是，当所研究的两个总体确有差别时，按检验水准 α 能够发现它的能力。如果 $1-\beta=0.9$，则意味着当 H_0（通常为对比组的效应值相等）不成立且检验水准为 α 时，理论上在每 100 次抽样中，平均有 90 次能拒绝 H_0。

当假设检验出现“阴性”结果（$P>0.05$）时，有必要复核样本含量和检验效能是否偏低，以便正确分析假设检验“阴性”结论的正确性。若检验效能偏低，说明试验的样本含量不够，应进一步增大样本含量，继续进行试验；若检验效能足够大，可下“接受 H_0”的阴性结论，尽管此时仍有可能犯第二类错误，但犯第二类错误的概率 β 在可接受的范围之内。

2.2 估计样本含量需要的前提条件

2.2.1 问题的提出

在统计教学、咨询和培训工作中，统计工作者经常被问到“我希望做××试验，请问该选择多少受试对象”。这样的问题往往使统计工作者无言以对。要想确定一项试验的样本含量，需要很多前提条件，凭空想象是没有科学依据的。以下列出估计样本含量需要的前提条件，实际工作者遇到类似的问题，可以先提供这些信息，通过相应的公式计算样本含量。当然，在统计软件日趋普及的今天，人们基本上都是借助统计软件包，如 SAS 软件[3] 直接估计样本含量。

2.2.2 前提条件

检验水准 α：研究者应事先规定本次试验允许犯第一类（也称假阳性）错误的概率 α，通常规定 $\alpha=0.05$，同时还应明确是单侧检验还是双侧检验。α 定得越小，所需的样本含量越大。同一问题，在其他条件不变的情况下，用单侧检验比用双侧检验所需要的样本含量要少。需要注意的是，选择单侧检验还是双侧检验不是随意确定的，而取决于专业知识和比较类型。一般差异性检验，存在选择单侧还是双侧检验的问题；而非劣效性检验和优效性检验只能进行单侧检验；等效性检验要求进行双单侧检验[2]。在一般差异性检验的场合下，当有专业知识为依据时，才可选择单侧检验；否则，一律选择双侧检验。

期望的检验效能或把握度（$1-\beta$）：要求的检验效能越大，所需要的样本含量就越大。在科研设计中，一般检验效能不宜低于 75%，通常设定为 80%，否则很可能出现非真实的阴性结果，从而不能反映出总体的真实差别。

先验知识：所谓先验知识，就是根据专业知识、文献资料或预试验结果获得的由样本推断总体的一些信息，包括容许误差、总体标准差、总体均数、总体率等。容许误差是指研究者要求的或客观实际存在的样本统计量与总体参数间或样本统计量间的差值。例如，比较两总体均数的差别时，不仅要知道总体均值间差值 $\delta=\mu_1-\mu_2$ 的信息，还要知道总体标准差 σ 的信息；若比较两总体率间的差别，应当知道总体率间差值 $\delta=\pi_1-\pi_2$ 的信息。有时研究者很难得到总体参数的信息，可以用专业上认为有意义的差值代替。例如，研究某种降压药的疗效，可以将临床上认为有意义的血压降低值作为先验知识，也可以人为规定试验药物的有效率超过标准药物有效率的 20%，这 20% 就是先验知识。若仔细推敲，这个 20% 就可能有两个名称。第一个名称叫做“右单侧检验的容许误差”，对应于一般差异性检验且为右单侧或上单侧检验；第二个名称叫做临床上有意义的“优效性界值”，对应于优效性检验。

与试验设计有关的其他因素：除了上述信息，研究者还需要明确试验设计的其他信息。包括研究者拟开展的研究类型是调查研究、试验研究还是临床试验研究，研究中将涉及到的因素个数及其水平数、

试验设计类型和比较类型，观测的效应指标的性质是定性的还是定量的，研究结果适用范围的大小等。

2.3 估计样本含量和检验效能方法的分类

我们知道，经典的统计推断可以分为两大类，一类是参数估计，即由样本信息推测总体参数，如用样本均数估计总体均数，用样本率估计总体率，这些都属于参数估计；另一类是假设检验，即对所估计的总体首先提出一个假设，然后通过样本数据去推断是否拒绝这一假设。例如，要探讨某种药物的疗效，假设这种药物疗效与标准药物疗效相同，然后通过试验数据验证这一假设。

根据研究目的可将样本含量估计分为参数估计时的样本含量估计和假设检验时的样本含量估计，因为这部分内容所占篇幅比较大，将在后续的专题中详细介绍，此处从略。

3 精神卫生学术论文中估计样本含量方面存在的问题

翻阅本刊发表的学术论文[5-15]，在如何确定样本含量这个问题上，绝大多数学术论文都是这样做的：先给出“入组标准”和“排除标准”，然后直接说：研究组（或试验组）入组 n_1 例；对照组入组 n_2 例。

显然，以上面的方式确定样本含量，属于毫无根据地确定样本含量。得出的检验结果，究竟有多大的检验效能是不知道的。一旦出现了阴性检验结果，研究者不知道是因为样本含量不够多，使检验效能过低所致；也不知道是因为所观测的评价指标的特异性和灵敏度都很低所致；当然，还可能由于其他种种原因，导致试验误差过大、试验数据不准确所致。

建议：开展任何试验研究、临床试验研究和调查研究，甚至包括做 Meta 分析的课题研究，在制订研究设计方案时，都应有根据地估计出合适的样本含量。这样做不仅保证了研究工作的科学性与严谨性，还充分体现了经济性。因为一旦盲目地确定了过多的样本含量，会浪费很多人力、物力、时间和财力；当然，若盲目地确定了过少的样本含量，就很容易得出假阴性结果。从而导致研究工作的失败。

4 对重复原则方面的错误案例进行辨析与释疑[1,16-18]

【例 1】某研究者为了证明 A（HP-1000 型超声诊断仪）、B（研究者自制的成像系统）两台仪器测定的结果无差别，作了如下的试验设计：选一个健康人作为受试对象，用 A、B 两台仪器前后两次（间隔为 1 个月）对此人分别重复测定 4 次，观测的定量指标分别是：①二尖瓣前叶 EC 幅度，②左室后壁运动幅度，③R-R 间期。数据处理方法：每个指标下有 4 组数据，既作了方差齐性检验，又作了配对比较的 t 检验，P 均 >0.1。结论：两台仪器的测定结果无差别，可用自制的成像系统取代费用很高的同类进口仪器。

试问：原研究者这样的设计试验，其结论可信吗？

【对差错的辨析与释疑】要得出两台仪器测定结果的差别无统计学意义的结论，仅凭对 1 个健康受试者 4 次重复测定数据进行比较，证据不足。因为在实际操作中，每台仪器每天要测定多个受试者，由于不同受试者之间存在很大的个体差异，两台仪器对某一个人的测定结果之间无差别，并不能推出在多数人身上测定结果的差别也一定很小。

本研究涉及三个因素，其中一个是试验因素（即仪器种类），另外两个是区组因素（即测定时间和受试对象），故可以选用交叉设计安排试验。若重复测定的结果之间变异度较小，样本含量 $n=6$ 或 8 即可；反之，n 应取 10 例或更多一些为宜。若从文献上查到交叉设计样本含量估计公式，并提供所需要的基本信息，按公式计算出 n 值，则更为妥当。从另一个角度看，上述问题属于对两台仪器测定定量指标的“一致性”评价，若基于最合适的一致性评价方法来估计样本含量，则更合适。因篇幅所限，此处就不深究了。

【例 2】为了观察甲紫注入小型猪正常腮腺后组织病理变化情况，有人选择 6 个月龄、体重 20～25 kg 的中国试验用小型猪 15 只，雄性 9 只、雌性 6 只。每只动物任选一侧腮腺为试验侧，另一侧作为正常对照，以消除个体差异及增龄对试验结果的影响。按注入甲紫后 1 周、2 周、1 个月、3 个月及 6 个月将 15 只动物随机分为 5 组，每组 3 只，每组的 3 只动物分别随机注入 0.6 mL、1.0 mL 及 4.0 mL 1% 甲紫溶液，然后观察组织病理变化情况。试问：此项试验研究中违背了试验设计的什么原则？

【对差错的辨析与释疑】本试验研究共用了 15 只小型猪，初看起来，“15”这个数目不算太小。但仔细看一下不难发现，试验中共涉及两个试验因素，第一个因素是“甲紫作用时间”，它有“1 周”、“2 周”、“1 个月”、“3 个月”及“6 个月”5 个水平；第二个因素是“甲紫剂量”，它有“0.6 mL”、“1.0 mL”及“4.0 mL”3 个水平。这两个因素的全面组合共有 15 种情况，每种情况构成一个特殊的试验条件，每个条件下仅有一只动物，所以本试验若列表表示，各

组的样本大小 $n=1$。这就违背了试验设计中的“重复原则”。因为生物医学研究的现象常常带有变异性，只有在相同试验条件下进行多次独立重复试验，随机现象的变化规律才能正确地显露出来。

那么，各小组究竟应该用几只动物合适呢？严格地说，需要根据预试验或文献资料提供的信息，结合研究者对试验精确度的要求，并根据拟采用的试验设计类型，按估计样本大小的相应公式计算为宜。一般情况下，若不便用公式估算时，如果是小动物试验（来源方便，花费不太大），各小组动物数不少于 10 只为宜；若是较大动物试验，各小组动物数不少于 5 只为宜。这里所讲的“各小组”，是指试验中独立的试验条件所决定的每个小组，如本例中是指在一个特定的甲紫作用时间下同时在一个特定的剂量下所形成的试验组，即本例共有 15 个小组。

【例 3】某项尺骨相对桡骨位移的试验研究中，取两个肱骨中段以上的上肢标本，去除所有的前臂伸屈肌腱，前臂中立位下固定桡骨，对尺骨小头施加 20 N 的掌背侧拉力，测量尺骨相对于桡骨的位移。然后先后切断背侧和掌侧桡尺韧带，相同的作用力下测量尺骨相对于桡骨的位移（DRUJ）。结果：在 20 N 拉力作用下，切断背侧桡尺韧带（DRUL），保持完整的掌侧桡尺韧带（PRUL），尺骨相对桡骨的背侧位移明显；切断掌侧桡尺韧带，保持完整的背侧桡尺韧带，尺骨相对桡骨的掌背侧位移都明显增加；而掌背侧桡尺韧带（DPRUL）都切断导致 DRUJ 明显下降。

【对差错的辨析与释疑】该研究违反了重复原则，致使研究结果科学性不强，结论部分的证据不足。在大多数情况下，要在相同试验条件下进行足够多次的独立重复试验，从而找出比较可靠的客观规律。原文试验只取了两个样本，试验涉及“是否切断 DRUL”、“是否切断 PRUL”两个试验分组因素和“在背侧还是掌侧测量位移”这一重复测量因素，只利用两个样品完成一个涉及三因素的试验研究项目，严重违反试验的重复原则，对照、随机、均衡原则也无从谈起，使得原文的研究结论“苍白无力”。这项研究可形成具有一个重复测量因素的三因素设计，试验设计及数据记录格式如表 1 所示。表中的每个“×”代表每个试验点（即全部因素不同水平组合所构成的一个特定试验条件）下所得数据，应在每个试验点下做两次以上的重复试验。也就是说这项研究要得到比较准确的结果，至少需要 8 个标本（样品）。当然也可以在每个试验点多做几次独立重复试验。

表 1 切断 DRUL、PRUL 对 DRUJ 的影响

是否切断 DRUL	是否切断 PRUL	尺骨相对桡骨的位移（mm）	
		背侧	掌侧
是	是	××	××
	否	××	××
否	是	××	××
	否	××	××

参考文献

[1] 胡良平. 统计学三型理论在实验设计中的应用[M]. 北京：军事医学科学出版社，2006：20－43，215－264.

[2] 胡良平. 科研设计与统计分析[M]. 北京：军事医学科学出版社，2012：129－227.

[3] 胡良平. SAS 常用统计分析教程[M]. 北京：电子工业出版社，2015：235－252.

[4] 柳伟伟，胡良平，贾元杰，等. 实验设计中的重复原则[J]. 药学服务与研究，2010，10(5)：330－334.

[5] 胡良平，关雪. 如何正确把握实验设计的重复与均衡原则[J]. 中华脑血管病杂志（电子版），2010，4(6)：43－47.

[6] 施玉梅，许小梅，李淑芬，等. 草酸艾司西酞普兰合并艾地苯醌对脑卒中后抑郁的临床疗效观察[J]. 四川精神卫生，2015，28(4)：336－338.

[7] 周平，张瑶，谭庆荣. 电针联合舍曲林治疗创伤后应激障碍的效果观察[J]. 四川精神卫生，2015，28(6)：504－506.

[8] 王雪，罗炯，李晓虹，等. 低频重复经颅磁刺激对难治性精神分裂症的增效作用[J]. 四川精神卫生，2016，29(1)：41－45.

[9] 刘萍，郭杰峰，吴郁丽，等. 变应性鼻炎儿童的智力结构与个性特征分析[J]. 四川精神卫生，2016，29(2)：172－175.

[10] 周燕玲，张杰，黄伟杰，等. 精神分裂症患者住院天数对自知力的影响[J]. 四川精神卫生，2015，28(4)：291－294.

[11] 金毅琼. 帕罗西汀联合重复经颅磁刺激治疗女性更年期抑郁症的对照研究[J]. 四川精神卫生，2015，28(6)：515－518.

[12] 郭新宇，杨媛，田丽. 草酸艾司西酞普兰联合重复经颅磁刺激对改善难治性抑郁症患者执行功能的疗效研究[J]. 四川精神卫生，2016，29(1)：26－30.

[13] 郭永芳，周小东，付华斌，等. 改良电痉挛治疗对精神分裂症患者血清细胞因子及 C－反应蛋白水平的影响[J]. 四川精神卫生，2015，28(2)：119－122.

[14] 平军辉，仲照希，王东平. 齐拉西酮注射液治疗儿童精神分裂症急性激越症状 35 例[J]. 四川精神卫生，2015，28(4)：314－316.

[15] 刘若楠，戴立磊，邹韶红. 不同家庭类型抑郁症患者自杀意念的差异性研究[J]. 四川精神卫生，2016，29(1)：35－40.

[16] 胡良平. 医学统计应用错误的诊断与释疑[M]. 北京：军事医学科学出版社，1999：5－21.

[17] 胡良平，李子建. 医学统计学基础与典型错误辨析[M]. 北京：军事医学科学出版社，2003：242－259.

[18] 胡良平，赵铁牛，李长平. 医学综合统计设计与数据分析[M]. 北京：电子工业出版社，2014：12－67.

（收稿日期：2016－07－26）

（本文编辑：陈 霞）