中国中西医结合杂志 2003 年 7 月第 23 卷第 7 期 CJITWM, July 2003, Vol. 23, No.7

·临床试验方法学·

临床试验样本含量的计算

刘建平

1 样本量估计的重要性

临床试验报告中有无预先的样本量估计是评价试验质量的重要依据之一。在试验设计阶段需要确定研究所需的病例数(通常称为样本含量)。理论上,验证某一干预措施与对照之间的差异,样本量越大,试验结果越接近于真值,即结果越可靠。但由于资源的限制和伦理的原因,临床试验的对象数量不可能做到无限大,而需要确定统计学显著性检验要求的最适样本大小。试验需要从研究对象(如患病人群)中选择具有代表性的合适样本,将符合纳入标准的病例在经过其本人的知情同意后进行分组。临床试验的样本量过小,无论试验结果是否存在差异,均不能排除因机遇(随机误差)因素造成的假阳性或假阴性错误;因而,小样本临床试验不能下肯定或否定的结论。目前国内发表的临床试验大多为小样本的试验。我们对中医药随机临床试验的评价中也发现大多数试验样本量都不大,且极少有报道计算样本量的试验[1]。近 10 多年来,开始强调大样本临床试验的重要性,通常认为上千例的试验为大样本试验,国际上已出现了所谓的"兆级试验"(Mega-trial),即病例数超过万例的临床试验。如阿司匹林治疗心脏病的临床试验,样本量超过 10 万例。大样本试验的优越性除了可靠地验证总疗效外,还有助于探讨亚组疗效,如药物在男性与女性、老年与青年、轻型与重型等亚组病例中作用的差异[2]。有的药物综合疗效虽不明显,但亚组分析可能发现其对某一特定患病人群有效。另外,大样本也有助于发现那些罕见的重要结局,如某药罕见的副作用。当然,并非所有的临床试验都需要大样本,尤其对于医疗资源十分有限的发展中国家。样本量过大,由于费用增加会造成资源浪费,有时会因患者持续接受较差的治疗如安慰剂对照而涉及伦理问题;此外,一些临床意义不大的微弱疗效最终也可能会出现统计学上的显著性差异,而这种差异是没有实际应用价值的。只有那些疗效差异既具有临床意义又有统计学显著性差异的结果才有应用价值。

英国利物浦大学热带医学院国际健康研究组,挪威国家另证医学研究中心

Tel: + 441517053185; Fax: + 441517053364; E-mail: Jpliu@liverpool.ac.uk

2 样本量估计的方法

试验设计时怎样知道合适的样本大小呢?一般可以通过统计学方法估计样本量。根据试验的目的和测量结局指标的不同,计算方法也不同。对于优劣性临床试验(superiority trial 或 inferiority trial),其目的是要验证试验干预与对照干预效果之间是否存在差异,通常是验证试验干预效果优于对照,如验证某一新药的疗效优于老药或安慰剂。对于等效性临床试验(equivalence trial),其目的是验证试验干预与对照之间效果相当,即差异不显著。通常见于不同的有效治疗如抗生素之间的比较,也用于比较同一种药物的不同剂型、不同给药途径的疗效。上述两种试验评价疗效的指标通常可分为两类:一类为计数(定性)指标,如死亡与存活,阳性与阴性,正常与异常;另一类为计量(定量)指标,如血压、血糖值、血清酶水平等实验室检测指标。有时临床试验评价的结局指标有多个,估计样本含量时需要选择其中最重要的结局指标。下面分别介绍两种试验的样本量计算方法。

2.1 优劣性临床试验的样本量计算

例如,一临床试验拟验证某中药治疗慢性乙型肝炎的疗效优于安慰剂对照,属于优劣性试验,结局测量为计数指标如抗病毒作用和肝功能恢复正常。估计样本量之前研究者需要考虑 3 个要素:试验干预与对照干预效应差异的大小、对试验精确度的要求和试验对象的依从性。效应差异的大小需要研究者根据该药物前期的临床研究和临床的实际意义决定,如试验组生存率比对照组提高 10% 就可认为有临床意义。临床试验的精确度也称为试验的把握度(power)。在此需要掌握两个基本概念,即统计学上的Ⅰ型错误和Ⅱ型错误,前者又称为假阳性(α)错误,后者又称为假阴性(β)错误,把握度 = 1 - β。对于计数指标的结局通常用四格表的形式来表示(见表 1),即试验结果可能出现的 4 种情况。那么,就此例来说计算样本量之前需要明确所用的疗效评价指标。如果想验证该中药的抗病毒作用,则可选择乙型肝炎病毒复制的指标,如

表 1 临床试验计数资料结局根据假设检验可能出现的结果

组间出现统计学上显著性差异	组间存在的真实差异	
	有	无
有	正确(1 - β)	Ⅰ型错误(α)
无	Ⅱ型错误(β)	正确(1 - α)

HBeAg 的阴转率；如果是了解中药的保肝作用，可选择肝功能指标，如血清转氨酶或胆红素的复常。以抗病毒作用为例，通过查阅文献，我们知道慢性乙型肝炎在不治疗的情况下，每年血清 HBeAg 的自然阴转率为 15%，可作为安慰剂对照组的本底资料（有些情况下没有安慰剂对照的资料时，可参照非特异性治疗，如维生素、肌苷、葡萄糖等作为对照的资料，即对照组的阴转率）。通过以往的文献报道或经过小样本的预试验，假设我们要验证的中药具有抗病毒作用，以 HBeAg 阴转率为指标，在原有基础上可提高 15%，即使阴转率达到 30%。此外，我们还需要确定两个参数，一个是 α 值，它的含义是当试验结果呈阳性时，我们下结论犯错误（假阳性错误）的可能性。通常将 α 值控制在 5% 以内，使试验有 95% 的可信性对一个阳性结果下肯定的结论；另一个参数是 β 值，它的意义为当试验结果呈阴性时，我们下结论犯错误（假阴性错误）的可能性，通常控制在 10% 以内。研究人员也可以根据对试验结果的精确性要求不同来确定 α 值和 β 值，如要求精确度极高，则可能设定 α 值为 1%，β 值为 5%；反之，如果要求的精度不高，则可设定 α 值为 10%，β 值为 20%。本文选择中间值，即 α 值为 5%（0.05），β 值为 10%（0.1）。第三个要素是需估计试验中病人退出的比例。如试验治疗的时间（或治疗结束后随访的时间）较长，则病人退出或失访的可能性较大。但是按照国际惯例，当试验病例退出或失访超过病例总数的 20% 时，试验结果将不可靠。假设本试验预计的病例退出率为 10%。考虑了以上 3 个因素后，可按以下公式计算样本含量[3]。

$$n = (U_\alpha + U_\beta)^2 2P(1-P)/(P_1 - P_0)^2$$

n 为每一治疗组所需的样本量，一般各组样本数应均等；U_α、U_β 为 α、β 所对应的 U 值，当 α 为 0.05，β 为 0.1 时，查正态分布分位数表得到：$U_{\alpha(0.05)} = 1.65$，$U_{\beta(0.1)} = 1.28$；P_0 和 P_1 分别代表原有的疗效和预计可达到的疗效，本例为自然阴转率和预计中药可达到的阴转率，分别为 15% 和 30%，$P = (P_1 + P_0)/2 \times 100\%$。将上述参数和数值代入公式：

$$P = (P_1 + P_0)/2 \times 100\% = (30 + 15)/2 \times 100\% = 22.5\%$$

$$n = (1.65 + 1.28)^2 \times 2 \times 0.225(1 - 0.225)/(0.3 - 0.15)^2 = 133$$

即每组需 133 例，两组共计 266 例，加上 10% 的退出病例（约 26 例），最后估计的试验样本量为 292 例，即每组约各需 146 例。

对于试验评价的结局为计量资料的临床试验，其样本量的计算方法有所不同。例如，某试验用中药治疗糖尿病，观察对血糖水平的影响。同样我们需要知道几个本底资料，包括试验前患者的基础血糖水平（包括均值和标准差），假设根据以往资料或预试验，测得空腹血糖水平为 9.7mmol/L（标准差为 2.1），现采用中药治疗，期望能将血糖水平降至 8.3mmol/L。假设 $\alpha = 0.05$，$\beta = 0.1$。计量指标的样本量公式如下[4]：

$$n = 2\sigma^2 \times f(\alpha, \beta)/(\mu_1 - \mu_2)^2$$

μ_1 为基础空腹血糖值（本例 $\mu_1 = 9.7$），μ_2 为拟降低的血糖水平（本例 $\mu_2 = 8.3$），σ 为标准差（本例 $\sigma = 2.1$）；$f(\alpha, \beta)$ 为一常数，根据不同的 α 和 β 值，可查表获得（见表 2），当 $\alpha = 0.05$，$\beta = 0.1$ 为 10.5。代入公式得：

$$n = 2(2.1)^2 \times 10.5/(9.7 - 8.3)^2 = 47$$

即每组需 47 例，两组共计 94 例，如考虑退出与失访 10%（约 10 例），则该试验所需样本总例数应为 104 例（每组各需 52 例）。

有时很难得到基础均数和标准差，也可将计量资料转换为计数指标进行样本量计算。如假设安慰剂组血糖水平可平均降低 5%，而中药组可降低 20%，$\alpha = 0.05$，$\beta = 0.1$，则可用以上计数资料公式计算样本量。

2.2 等效性临床试验的样本量计算

如前所述，有的临床试验试图验证两种治疗方法之间差异无显著性。也就是说，即使存在差异，该差异也是在可接受的范围之内，且不具有统计学上的显著性意义。以下介绍等效性试验计数资料结局的样本量计算方法。例如，欲验证某中药治疗慢性乙型肝炎的抗病毒作用与西药干扰素的效应相当，试验以血清 HBeAg 阴转率作为评价指标。已知干扰素治疗慢性乙型肝炎 HBeAg 阴转率可达 50%，预计该中药的阴转效果不低于干扰素的 5%（95% 的可信性），要验证两者的 HBeAg 阴转效果相当，可按以下公式计算试验所需样本量[4]：

$$n = 2p \times (100 - p) \times f(\alpha, \beta)/d^2$$

p 代表标准治疗所预期的疗效（本例为 50%），d 代表试验药物与标准治疗比较可接受的差异（本例为 5%），取 $\alpha = 0.05$，$\beta = 0.2$，则（见表 2）$f(\alpha, \beta)$ 为 7.9。

代入公式，$n = 2 \times 50 \times (100 - 50) \times 7.9/5^2 = 158$

表 2 用于样本量计算公式中的 $f(\alpha, \beta)$ 值

		β（Ⅱ型错误）			
		0.05	0.1	0.2	0.5
α（Ⅰ型错误）	0.1	10.8	8.6	6.2	2.7
	0.05	13.0	10.5	7.9	3.8
	0.02	15.8	13.0	10.0	5.4
	0.01	17.8	14.9	11.7	6.6

即每组需要 158 例，两组共计 316 例。可见，疗效差异越接近，所需样本量将会越大。小样本的等效性试验如果没有样本量的预先估算，往往不能轻易下两者疗效无差异(即等效)的结论[5]。等效性临床试验也可根据预计治疗的成功率和临床等效性差异的可接受范围，通过查表获得每组所需的样本量(见表 3)[6]。

表 3 等效性临床试验各组样本含量

期望的成功率	可接受的临床差异		
	5%	10%	15%
50%	2102(1570)*	526(393)	234(175)
60%	2018(1507)	505(377)	225(168)
70%	1766(1320)	442(331)	197(148)
80%	1346(1006)	337(252)	150(113)
90%	757(566)	190(142)	85(64)

注：* 可信性水平为 95%，把握度为 90%；( )内数据为把握度 80%时的样本量

等效性试验以定量指标为测量结局的样本量计算，通常采用可信区间的途径来评估等效性，即在多大的差异范围内可认为是“等效”的。从统计学意义上作结论可能发生两类错误：一类是两种治疗实际有明显差异，而我们得出等效的结论，即发生Ⅰ型(假阳性)错误；另一类是当两种治疗实际为等效的，而我们却得出有差异的结论〔Ⅱ型(假阴性)错误〕。因此，研究者需要从专业角度和临床意义上确定一个界值范围，超出这个范围则认为是不等效的。等效性试验计量资料结局样本量计算(双侧检验)的计算公式如下[7]：

$$n = 2s^2/\delta^2[z(1-\alpha) + z(1-\beta/2)]^2$$

n 代表每个治疗组的样本量；s 表示均值的标准差；δ 表示对照组均值(μR)与试验组均值(μT)的差($\delta = |\mu R - \mu T|$)，也就是研究人员认为可接受的差值范围；常数 $z(1-\alpha) = z(1-0.025) = 1.96$，常数 $z(1-\beta/2) = z(1-0.2/2) = z(0.9) = 1.28$。

例如：试验两种气雾吸入剂缓解哮喘发作的疗效是否相等。采用 95% 可信区间(双侧检验)作为判断等效的允许变异范围。结局测量指标为晨间呼出气流峰值(L/min)。根据以往的试验估计，15L/min 为允许的差异范围(即正负值均不超过 15 为等效)，晨间呼出气流峰值的标准差(s)为 40L/min。该试验的把握度($1-\beta$)为 80%，则每组的样本量根据上述公式计算得：

$$n = 2\times 40^2/15^2[1.96 + 1.28]^2 = 149.3 \approx 150$$

即该试验每组所需样本量为 150 例哮喘患者。

样本量估计是任何一个前瞻性临床试验所必须的步骤，尤其是随机对照临床试验。没有样本量估计和报告试验把握度，临床试验结果很难让读者判断其真实性和可靠性。尤其是在小样本的临床试验，不能排除假阳性和假阴性错误，轻率地下肯定或否定的结论，推荐给临床医生或患者使用，或以此作为制定决策的依据，是不科学和不道德的。任何科学研究都是在前人研究基础上的延续。因此，如果没有广泛地查阅文献，了解试验领域国内外研究的现状，是不可能提出一个好的研究问题、作出一个严格的设计、形成一篇规范的试验报告，这样的研究缺乏创新性，属于低质量的重复，结果将不为人们所用或使用后造成误导，最终损害患者的利益，同时也造成有限医疗资源的浪费。

参 考 文 献

1 刘建平，林 辉，刘理礼，等.病毒性肝炎治疗随机对照试验文献方法学评价.华西医学 1999;14(2):126—128.

2 Warlow C. Advanced issues in the design and conduct of randomized clinical trials: the bigger, the better? Statistics in Medicine 2002; 21:2797—2805.

3 Carlin JB, Doyle LW. Statistics for clinicians. 7:Sample size. J Paediatr Child Health 2002;38:300—304.

4 Pocock SJ. Clinical trials: A practical approach. 1st edition. Chichester: John Wiley & Sons Ltd. England, 1983:123—141.

5 Millar JA, Burke V. Relationship between sample size and the definition of equivalence in non-inferiority drug studies. Journal of Clinical Pharmacy and Therapeutics 2002;27:329—333.

6 Smith C, Burley C, Ireson M, et al. Clinical trials of antibacterial agents: a practical guide to design and analysis. Journal of Antimicrobial Chemotherapy 1998;41:467—480.

7 Jones B, Jarvis P, Lewis JA, et al. Trials to assess equivalence: the importance of rigorous methods. BMJ 1996;313:36—39.

(收稿：2003-03-20)

·消 息·

一市(北京市)二省(福建省、广东省)中西医结合学会男科专业委员会成立

2002 年 12 月北京市、福建省、广东省中西医结合学会男科专业委员会成立，并举办了第一次学会会议，贾金铭、崔学教、张敏建教授分别当选为主任委员。

(郭 军)