

· 中医药临床疗效评价 ·

临床试验效应值的分析和解释

朱 泉 郭新峰 赖世隆

摘要 临床试验效应值的分析应包括效应值大小和精度两个方面,而对试验效应的解释应遵循准确、清晰和定量的原则。对于主要临床结局,可使用相对危险度、比数比、相对危险度降低率、绝对危险度降低率、减少 1 例重大临床事件需要治疗的人数等指标;而替代指标和中间指标的解释方法仍需要加强研究。国内中医药研究中的试验效应表达尚有很多问题有待解决。

关键词 临床试验;效应值;中医药;疗效评价

How to Analyze and Interpret the Effective Values in Clinical Trials ZHU Quan, GUO Xin-feng, and LAI Shi-long DME Center, Guangzhou University of TCM, Applied Epidemiology Department of The Second Affiliated Hospital, Guangzhou (510405)

Abstract The analysis of the effective value obtained in clinical trails should take account of both sides of its size and precision, for its interpretation, the general principles of: accuracy, explicitness and quantification should be complied. Criteria such as relative risk (RR), relative risk reducing rate (RRR), absolute risk reducing rate (ARR) and number needed to treat (NNT) an important clinical event, etc. are commonly used in the primary outcomes interpretation, while the methods to express the effective size and precision of surrogate outcomes and intermediary outcomes are needed to be further studied. Many problems for give expression to the outcome of TCM clinical trail still remain to be solved.

Key words clinical trial; effective value; traditional Chinese medicine; outcome assessment

临床试验是检验和证实医疗干预措施效果的科学方法,而恰当的疗效评价是临床试验的重要环节。疗效评价的问题可以包括两个方面:一方面是选择什么结局指标(即从哪些方面)进行疗效评价、另一方面是如何分析效应值和如何表达其反映出的实际意义。关于前者,我们已经做了相关的讨论^[1],本文主要讨论采用什么方法来表达和解释试验结果。

我们认为结局指标分析和表达应遵循的原则是:准确、清晰和定量。由于概率与统计学知识的普及和应用,在临床试验的分析中,研究者习惯于应用概率的推断方法,一般使用特定统计量结合 P 值进行表述。概率统计的推断是建立在对事件发生可能性大小的分析基础之上的,不能够代替临床思维和临床意义的判断。随着循证实践(evidence-based practice, EBP)模式的兴起,医学研究的结论成为患者、医生、医疗保健决

策者共同关心的问题 and 决策依据的来源。临床医生关心的是某种疗法到底有多大把握可以成功地应用于自己的患者;卫生决策者关心的是某种医疗措施的推广应用价值如何;患者关心的是减少发病或重要临床事件发生的可能性有多大、对生存质量的改善程度等,而不是关心组间有差异的概率是多大。所有这些都向我们提出了问题:用什么方法计算和表达试验效应更为合适?

分析和解释试验效应应正确理解和处理好两个关系:(1)效应的直观表述和统计学意义的关系;(2)统计学意义和实际意义的关系。对于前者,由于一切临床试验均是通过抽样的方法,用样本信息推断总体的特征(多为无法直接测得的真值),直观上组间有差异效应可能只是由于抽样误差导致的,只有具有统计学意义,才可以表明试验效应的作用不太可能是由于抽样误差所致。对于后者,仅具有统计学意义,并不一定就具有推广应用的实际价值。同时,统计量和 P 值的大小并不直接表示效应的大小,只表明作出相应推断的把握大小, $P=0.01$ 并不表示比 $P=0.05$ 具有更大的效应。

衡量医疗干预措施的效应需要特定的指标分析和

基金项目:广东省科技厅资助的循证医学方法研究(No. 2KM05801s)

作者单位:广州中医药大学 DME 中心,第二附属医院临床流行病学应用研究室(广州 510405)

通讯作者:朱 泉, Tel: 020 - 36585404, E-mail: drzhuquan @ 163.com

解释方法,一般包括对试验效应的大小和精度两个方面的内容。临床试验效应值大小的表示方法,可以根据资料的类型进行分类,如可将资料分为定性资料、定量资料等。因为资料的性质往往与结局指标的类型相关,所以我们可以从结局指标类型的角度来理解。临床结局指标可以分为主要结局指标(primary outcome)、替代指标(surrogate outcome)、中间指标(Intermediary outcome)等。

1 主要结局指标效应大小的表达方法

主要临床结局指标多为对疾病自然史最终终点的测量,如死亡、残疾、功能丧失和某些重要的临床事件的发生,如心肌梗死、骨折、失明等。此类指标多为二分变量:“发生”和“不发生”。常用的表达方式为相对危险度(relative risk, RR),比数比或称优势比(odds ratio, OR),相对危险度降低率(relative risk reduction, RRR),绝对危险度降低率或称归因危险度(absolute risk reduction, ARR; attributed risk, AR),减少 1 例临床事件的发生需要治疗的人数(number needed to treat, NNT)等。RR、OR、RRR 均为相对性的危险度计算指标,而 ARR 和 NNT 则是绝对性的危险度计算指标,表示由于干预所致的净效应。

1.1 RR

RR 的概念首先由 Cronfield J 提出,RR 的意义是用于衡量因素和发病(或死亡)的联系强度的指标。其定义为暴露者(即暴露于该研究因素的人群)的发病率与非暴露者的发病率之比。RR = 1 表示研究因素与发病(或死亡)无关;RR > 1 说明研究因素为危险因素;RR < 1 说明研究因素为保护因素^[2]。RR 值早期多应用于流行病学的队列研究中,在引入临床试验的解释后,其意义一般为试验组与对照组某重要临床事件发生率的比值,由于试验措施常为潜在有效的干预,在与安慰剂对照或与疗效低于试验措施的阳性对照措施比较时,RR 值经常 < 1,表明试验措施是一种保护措施。当把有利结局作为试验效应,且试验措施为有效时,RR 值常 > 1。

RRR 也是常用的一种危险度衡量指标,计算公式见表 1。表示某因素导致的危险度改善值占对照组或基线情况的百分比,即有多大比例的改变是由此因素导致的。临床试验中,高的 RRR 表示干预措施更加有效。究竟 RRR 多大具有临床意义呢? David L 等^[3]认为 RRR > 50% 总是表示具有临床意义,RRR > 25% 经常被认为是具有临床意义的。当然,这些标准并非绝对的,我们认为仍需要结合统计学的显著性检验和其他临床效应衡量指标,尤其是 NNT 等来综合判断。

试验组临床事件发生率: $P_A = a/(a + b)$; 对照组临床事件发生率: $P_C = c/(c + d)$; 相对危险度的计算公式为: $RR = P_A/P_C$; 相对危险度降低率的计算公式为: $RRR = (P_C - P_A)/P_C$ 。

表 1 计算 RR 值的四格表资料

组别	临床事件		合计
	发生	未发生	
试验	a	b	a + b
对照	c	d	c + d
合计	a + c	b + d	n

1.2 OR

或称比值比、优势比。OR 值是用来表示暴露因素与疾病联系强度的一个十分有用的指标,OR 值与 RR 值的趋势是一样的。在某些情况下,如在病例对照研究中,无法得到发病率的数据,或在采用 Logistic 回归模型对多个因素进行估计时,无法计算 RR 值,可以采用 OR 对 RR 值进行估计。但通过衡等变换可得 $RR = (ac + ad)/(ac + bc)$,从数值上讲显然 RR 值是 OR 值分子分母同加 $ac(ac > 0)$,故在 OR = 1 时,即 $ad = bc$ 时,RR 显然也等于 1;当 $0 < OR < 1$ 时, $RR > OR$;当 $OR > 1$ 时, $RR < OR$ 。则 $|OR - 1| \geq |RR - 1|$,即在数轴上 OR 值比 RR 值更加远离 1。而 OR 或 $RR = 1$ 表示某因素(临床试验中常指干预措施)的保护作用或危险性与对照措施或基线相同,故 OR 值总是夸大了因素的保护作用或危险性。故只有两组事件的发生率都十分低时, $RR = a(c + d)/c(a + b)$ 中 a 和 c 可以忽略不计,此时用 OR 估计 RR 才比较准确。

OR 值的 Logistic 回归估计非常有用。在 Logistic 回归模型中,可以将干预措施作为一个变量引入模型,同样可以将年龄、性别、疾病亚型等因素(在基线不均衡的情况下,往往会成为混杂因素)同时引入。各变量可以独立引入,也可引入交互作用项。以 e 为底,以回归系数为指数的幂即为 OR 值,可以得到各变量对结果变量,即干预结局的效应,而此时的 OR 是校正了其他因素的仅由此因素所致的效应大小。Logistic 回归模型可以计算自变量为二分变量、等级变量和连续变量的 OR 值,但在解释上有所不同^[4],对于等级变量 OR 值表示每增加一个等级,结果出现的可能性的倍数;而自变量为连续性变量时则表示每增加一个单位,结局出现的可能性的倍数。

同样,在临床疗效判定标准可靠、观察准确的前提下,我们可以将有效(或痊愈、显效)作为临床事件,计算 RR 或 OR 值,国内 Meta 分析的研究文献中常采用这种效应大小的表达方式。

1.3 ARR 或 AR

ARR 的概念由 Levin 于 1953 年提出,表示净由某危险因素所致发病率(或死亡率)增加的部分^[2]。引入临床试验后可指净由干预措施所致不利临床结局减少的部分。

绝对危险度降低率的计算公式为:ARR = P_C - P_A
1.4 NNT

与 RR 和 OR 不同,NNT 一开始就是为解释临床干预试验的结果而提出的,是一种绝对性的危险度测量方法,由于危险度的相对测量方法难以考虑基线的问题,ARR 和 NNT 的测量可以弥补 OR、RR 等相对测量方法的不足。

减少 1 例临床事件需治疗人数的计算公式为:
NNT = 1/(P_C - P_A) = 1/ARR

Cook RJ 等^[5]引用了一篇公开发表的有关抗高血压疗法对轻、中度高血压患者益处的文献数据来阐明 NNT 在效应表达上的优势,见表 2。两组患者是互不包含的,中度高血压患者对照组 5 年内中风发生率为 20%,试验组为 12%,则 ARR 为 0.20 - 0.12 = 0.08。ARR 的倒数即 NNT 为 13,表明医生需要治疗 13 例中度高血压患者 5 年,相对于对照组或基线,可以防止 1 例中风的发生。同样也可以计算轻度高血压患者相应的 NNT 为 167。两组的 RRR 相同,均为 40%,似乎提示应对两类患者的治疗付出同样的力气,但 NNT 的结果显然提示了应对两类患者区别对待,同时也表明 RRR 应结合 NNT 进行综合评价。ARR 虽然也为绝对性的危险度计算指标,但其直观性显然不如 NNT。

表 2 使用 RRR 和 NNT 计算高血压病患者试验效应					
高血压	5 年内中风发生率		RRR ($P_C - P_A$)/ P_C	ARR $P_C - P_A$	NNT $1/(P_C - P_A)$
	对照组	试验组			
中度(舒张压≤115mmHg)					
事件率(P)	0.20	0.12	0.40	0.08	13
总例数	16778	16898			
轻度(舒张压≤110mmHg)					
事件率(P)	0.015	0.009	0.40	0.006	167
总例数	15165	15238			

2 替代指标或中间指标效应大小的表达方式
在经费和时间有限的情况下,很多研究者采用血脂、血糖、血压、骨密度等作为替代指标来衡量试验效应,而属于中间指标的重要临床症状和患者主观满意度等也常常用来衡量试验效应。健康相关生存质量(health-related quality of life, HRQL)视疾病具体情况有时也可以被认为是中间指标。这些指标多为连续变量或等级变量,临床研究中经常采用分析指标改变值的方法,如差值、改善程度、改善等级等测量效应大小。

2.1 关于使用指标改变值测量试验效应

在一般情况下,单纯的治疗前后比较是没有意义的,因为难以区分治疗前后的差异是干预措施所致还是疾病自然病史所致。为了控制基线特征的影响和疾病自然病程的影响,往往采用治疗前后改变量的组间比较方法。国内广泛使用的治疗后比治疗前减少或增加多少个等级,然后将差值进行检验的方法。这种方法在统计学上是没有问题的,但问题是基线特征轻的病例减轻一个等级与基线严重的病例减轻一个等级的意义是相差很远的。解决的方法可以采用事后分层分析的方法,既可以分析差别,又可以对不同类型的患者改善情况加以了解。

2.2 关于使用综合指标测量试验效应
当一个指标难以反应治疗的效应或主要疗效指标测量不可行时,可以选择某些替代指标、中间指标及其构成的综合指标作为试验效应的测量。这种思路显然是合理的,但综合指标的形成及其合理性的评价方法需要经过一系列严格的程序。国内中医药临床研究中常用的疗效判定标准或疗效指数,如分为痊愈、显效、有效、无效或恶化等多可以认为是综合指标。这种指标的效应大小非常易于被患者、医生和卫生决策者所理解,但由于此类指标的构成要素仍然属于替代指标或中间指标,一则多数疾病的替代指标或中间指标的合理性尚未得到完整科学的评价,二则多个要素构成综合指标的加权方法非常复杂。显然未经过严格评价的综合指标对客观反映干预措施的真实效应仍然存疑。

2.3 关于试验效应实际意义的理解
连续性资料的效应大小显然也要从临床实际意义来考量,并非有统计学意义即可。Guyatt GH 等^[6]论述了连续性资料效应大小的解释方法,并举例:一个研究评价呼吸肌训练对慢性换气功能受限的患者的效应,采用封闭道路的 6min 步行距离试验(6-minute walk test),结果显示试验组平均最远距离从 406m 提高到 416m,而对照组平均最远距离从 409m 提高到 429m。与对照组比较,试验组呼吸肌训练导致的提高为 -10m(点估计),95% CI 为(-26,5)。由此可知即使是最好情况下,在试验基线 400m 的基础上增加 5m 对患者来说并不是有重要意义的。

使用 HRQL 测量的效应多表现为分值的改变,其大小如何去理解和解释呢? Guyatt GH 等^[7]也认为理解生存质量的研究结果是一个特殊的挑战,并举例用 Oswestry Back-Disability Index 测量急性背痛患者的结果为试验组低于对照组 3.9 分;在应用环孢霉素治疗风湿性关节炎的患者中,使用功能障碍计分评价试验

效应,结果为试验组好于对照组 0.28 个单位。从表面上看,我们很难理解这些分值的差异是微不足道的、小而重要的、中等的还是大而绝对重要的。Guyatt GH 等提出了理解生存质量测量结果的 4 种策略:(1)将 HRQL 分值与众所周知的功能测量(如 NYHA 功能分类)和临床诊断(如抑郁症的计分诊断)联系在一起;(2)与重要事件的发生联系在一起;(3)与患者经历的重要改变联系在一起;(4)与患者对好转或恶化的自我评价联系在一起。我们也可以从另外一个角度理解生存质量分值改变的临床意义,当构成生存质量的条目自身的等级是按照临床意义划分的,则可以由之推断总分改变量的临床意义。在实际运用中可以简化处理:平均提高 1 个等级。如欧洲癌症研究与治疗组织(EORTC)的生存质量量表 QLQ-C30^[8] 每条目的等级均是 4 级(从病症最轻的 1 级到最重的 4 级、活动完全不受限的 1 级到完全受限的 4 级)、总分为 100 分,可以简单地认为生存质量评分的改善达到 1 级(总分改善 33 分)才算是有重要临床意义的改善;同理,一个条目均是 5 级的量表(假如总分亦为 100 分)其得分改善要达到 25 分才有临床意义。当然,这样做的前提是量表内条目的等级较一致,且其等级划分较为科学。现在多数量表的等级划分在 5 级左右,故临床常用 25% 作为参考标准也是有一定道理的。

3 试验效应大小的可信区间表示

临床试验的结果是对真值的估计,前述 RR、OR 等称为点估计(point estimate),一般来讲,我们相信真值应出现在这附近,但不太可能恰好是真值。真值与点估计的差别显然是由于抽样误差的存在。而这个抽样误差到底有多大,可以用统计学方法估计,在考虑了抽样误差(标准误)和点估计的综合因素后,我们便可以得到一种估计方法,即区间估计(interval estimate),相应的结果称为可信区间或置信区间(confidence interval, CI),通常我们可以使用 95% CI 或 99% CI。以 95% 可信区间为例,意味着从同一总体中作样本含量相同的 100 次大样本重复抽样,可得 100 个可信区间,其中有 95 个可信区间包含真值(估计正确),只有 5 个可信区间不包含真值(估计错误)。5% 是小概率事件,实际发生的可能性小,因此,在实际应用中可以认为一次抽样的可信区间已包含了真值^[9-11]。正态分布数据的 CI 为点估计值加减相应倍数的标准误(如 95% 可信区间取 1.96 倍的标准误)得到。非正态分布的数据首先通过数据变换转为某种特定分布,根据分布的概率密度函数,确定标准误的倍数,得到两侧界值,再通过前述数据变换的反变换得到 CI。前述 RR、

OR、RRR、ARR、NNT 等常用指标的 CI 计算公式,有兴趣的读者可以参考相应的书籍和文章^[12],而 NNT 的 CI 可以通过 ARR 的 CI 取倒数得到。

可信区间表达试验效应有什么优势呢?

首先,CI 提供给我们一种不看 P 值就可以得到是否有统计学意义的直观方法。如上述举例中慢性肺功能疾病采用呼吸肌训练导致的 6min 步行试验成绩提高的 95% CI 为 $(-26, 5)$,由于提高成绩为 0 是没有意义的,而 95% CI 包含 0,则可以认为在 $P=0.05$ 的水平上,是没有统计学意义的。同样 OR、RR 等于 1 为两组危险度相同,故 95% CI 包含 1 时,表示在 $P=0.05$ 的水平没有统计学意义。

其次,CI 表示了一种估计精度,如果试验效应大小相同(点估计值一样),宽的可信区间精度相对较低,而窄的可信区间精度相对较高。CI 的宽度受标准误的影响,从标准误的计算公式 $SE = S/\sqrt{n}$ 可知:标准误受样本量和标准差两方面的影响。在临床试验中,同质性越好的样本,数据离散程度越小,即标准差越小,则相应的 CI 越窄。而样本的同质性直接受研究的纳入标准影响,纳入标准越严格,研究的内部真实性越好,但外推范围就会受到影响,所以多同时考虑内部真实性和外部真实性权衡确定纳入标准。在样本同质性既定的情况下,显然标准误在很大程度上受样本含量的影响,样本含量越大标准误越小,相应的 CI 越窄。前述例子中 CI 的宽度为 31,假如通过提高样本量的前提下,仍然得到 $-10m$ 的点估计,而 95% CI 为 $(-18, -2)$,CI 的宽度为 16,则表明估计的精度更高。

再者,CI 可以提示我们临床试验结果效应的实际意义。如前述例子中对 6min 步行距离试验结果的实际意义的解释。

从以上论述可以看出,对试验效应的解释已经不局限于统计学和数学的范畴,一个试验效应的实际意义涉及临床意义、推广性、实用性等问题。在考虑干预措施带来的益处时,必须同时充分考虑其是否也带来危害,是否利大于弊,花费是多少。主要结局指标效应的表述方法非常直观和明确,如病死率的减少是一个非常直观的指标,人们都很清楚减低多少病死率的意义有多大。而替代指标或中间指标如血糖、血压或生存质量的得分究竟改变多少才具有明确的临床意义,仍然是一个学术界探讨的问题。一般认为,只有在有明确证据表明替代指标和中间指标可以可靠而敏感地预测主要结局的前提下,它们的临床意义才会比较明晰。

现在国内中医药研究中尚较少见到以主要结局为

评价指标的研究,也较少使用 OR、RR、AR、NNT 等危险度指标进行试验效应大小的解释。中西医的理论体系虽然不同,但无论是中医还是西医所关心的重点都应是疾病的主要结局。除此之外,在中医药的疗效评价中,十分重要的是合理选择重要临床症状、若干症状的组合、患者的主观感觉,对于治疗的总体满意度和生存质量等中间指标和综合指标。提到中间指标和综合指标,就不得不提到中医软指标量化的问题,软指标的量化及其综合方法不能草率行事,简单地将症状分为轻、中、重处理以及不顾指标的重要性不同进行加分的方法都显得过于粗糙,由这样的方法反映出的试验效应的准确性难以令人满意。如何将临床表现进行量化需要做很多细致深入的工作,只有经过科学方法建立的指标才能正确反应出中医药应有的试验效应。有关这方面的内容,将在后面的文章中具体阐述。

参 考 文 献

- 1 郭新峰,朱 泉,赖世隆. 替代指标、中间指标及其在中医药疗效评价研究中应用价值的思考. 中国中西医结合杂志 2005;25(7):585—588.
Guo XF, Zhu Q, Lai SL. Surrogate outcomes, intermediary outcomes and their application in TCM clinical effectiveness studies. Chin J Integr Tradit West Med 2005;25(7):585—588.
- 2 徐天和,田考聪主编. 中国医学统计百科全书——描述性统计分册. 北京:人民卫生出版社,2004:37.
Xu TH, Tian KC, editors. China statistical encyclopedia—descriptive statistics. Beijing: People's Medical Publishing House, 2004:37.
- 3 David L, Sackett R, Brian Haynes, et al. Clinical epidemiology—A basic science for clinical medicine. 2nd ed. Little: Brown and Company, 1991:203.
- 4 Mitchell H. Katz 著. 式变量分析临床实用指南. 北京:中国科学技术出版社,2000:818—850.
- Mitchell HK, editor. Multivariable analysis: a practical guide for clinicians. Beijing: China Science and Technology Publishing House,2000:818—850.
- 5 Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995; 310 (6977):452—454.
- 6 Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature: II. How to use an article about therapy or prevention: B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994;271(9):703—707.
- 7 Guyatt GH, Naylor CD, Juniper E, et al. Users' guides to the medical literature. III. How to use articles about health-related quality of life. AJMA 1997;277(15):1232—1237.
- 8 Sprangers MA, Cull A, Bjordal K, et al. The European organization for research and treatment of cancer. Approach to quality of life assessment: guidelines for developing questionnaire modules. EORTC study group on quality of life. Qual Life Res 1993;2(4):287—295.
- 9 徐勇勇,孙振球,颜 虹主编. 医学统计学. 第 2 版. 北京:高等教育出版社,2004:107.
Xu YY, Sun ZQ, Yan H, editors. Medical Statistics. 2nd ed. Beijing: Higher Education Press, 2004:107.
- 10 杨树勤主编. 卫生统计学. 北京:人民卫生出版社,1993:27.
Yang SQ, editor. Biostatistics. Beijing: People's Medical Publishing House,1993:27.
- 11 Kelvin J, Bie NO, Peter C. Mastering statistics a guide for health service professionals and researchers. United Kingdom: Stanley Thornes (Publishers) Ltd, 109—113.
- 12 Bernard Rosner 著. 生物统计学基础. 北京:科学出版社,2004:554—563.
Bernard Rosner, editor. Fundamentals of biostatistics. Beijing: Science Press, 2004:554—563.

(收稿:2005-05-28)

· 消 息 ·

由中国中西医结合学会主办,广东省中医院承办的“全国围手术期中西医结合研究专题研讨会”将于 2005 年 11 月 28 日—12 月 1 日在广州举行。著名中西医结合专家吴咸中院士、陈可冀院士等将应邀出席大会,并做专题讲座。大会同时邀请了多位国内著名中西医结合专家举办精彩讲座,学习结束授予国家级 I 类继续教育学分 6 分。大会征文现已开始,截稿日期 2005 年 9 月 10 日。

联系电话:(020)81887233 转 31225 或 31226;传真:(020)81874903。

联系人:毛炜 周罗晶 老膺荣

地址:广州市大德路 111 号广东省中医院科研部(邮政编码:510120)。