·临床研究与循证医学·

临床研究结局指标选择与样本量估计

曾于珍，陈世耀

复旦大学附属中山医院消化科，上海 200032
复旦大学循证医学中心，上海 200032

通信作者：陈世耀　电话：021-64041990，E-mail：chen.shiyao@zs-hospital.sh.cn

【摘要】结局指标的选择应考虑研究设计、预期结果及可用资源。目前常用的结局指标包括生物学指标、卫生经济学评价及生活质量评价。结局指标、观察方法、评价方法与研究设计相关并影响样本量估计。样本量估计需同时考虑检验效力、差异显著水平、效应量及连续变量中的标准差值。临床研究中，合理的结局指标选择与样本量估计可显著提高临床研究结果的可靠性及可行性。

【关键词】结局指标；样本量；临床研究
【中图分类号】R-1　【文献标志码】A　【文章编号】1674-9081(2018)01-0087-06
DOI：10.3969/j.issn.1674-9081.2018.01.016

Outcome Measure Selection and Sample Size Estimation for Clinical Research

TSENG Yu-jen, CHEN Shi-yao

Department of Gastroenterology, Zhongshan Hospital, Fudan University, Shanghai 200032, China
Evidence-based Medicine, Fudan University, Shanghai 200032, China

Corresponding author: CHEN Shi-yao　Tel: 021-64041990, E-mail: chen.shiyao@zs-hospital.sh.cn

【**Abstract**】The selection of outcome measures should be based on the study design, anticipated results, and available resources. Biological parameters, analysis of economic benefits, and life-quality assessment are commonly employed outcome measures. Outcome measures, observation method, and evaluation method correlate with the study design and affect the sample size estimation. The sample size should be determined according to various factors, including power, significance level, expected effect sizes and standard deviation in the case of continuous variables. In clinical research, rationally selecting outcome measures and scientifically estimating sample sizes might markedly improve the reliability and feasibility of the study results.

【**Key words**】outcome measures; sample size; evidence-based medicine

Med J PUMCH, 2018,9(1):87-92

临床研究与基础研究的区别在于研究目的直接回答临床问题、研究对象为患者或患者来源的生物学标本、研究场地多在医院、研究主体为临床医生且常涉及多学科参与。临床研究可以理解为探索暴露与结局之间的关系。暴露包括自然暴露（病因研究）和人为暴露（干预-疗效或预后因素评价）。通常意义的疗效比较通过随机对照临床试验实现，疗效评价可以是短期或固定时间；广义疗效则包括长期结局，同时考虑产生结局的时间，即预后研究。需要指出的是，疗效不仅指治疗效果，也包括诊断效率。

基金项目：上海市科学技术委员会科技创新项目（15411950500）

临床研究的基础是临床问题，临床问题的提出需遵循 PICO 原则："P"指特定的患病人群（population/participants），也是研究的目标人群；"I"指干预或暴露（intervention/exposure）；"C"指对照或另一种可用比较的干预措施（comparator/control）；"O"为结局（outcome）。临床研究中，结局指标的选择对疗效评估起重要作用。采用不同的结局指标，可能会对相同的干预手段或暴露得出迥然不同的结论。由于结局指标选择涉及测量和评价，因此在临床研究中，如何选择结局指标，如何获得结果，如何评价结果，如何合理估计样本量，是临床医生在临床研究设计中需要面临的重要问题。

临床研究案例：肝硬化患者首次出血后 1 年内再出血的发生率高达 60%～80%，采取干预措施可能减少再出血，改善预后。卡维地洛是新的 β-受体阻滞剂，通过降低门脉压力减少再出血，延长患者生存时间，提高生存率。如何评价卡维地洛在肝硬化门脉高压预防静脉曲张出血治疗中的作用？

在这项评价卡维地洛治疗效果的研究中，选择怎样的研究结局指标将直接影响临床研究设计中样本量估计和随访时间，也决定了测量方法及结局评价方法[1]。

1 临床研究结局指标

临床研究的结局指标包括生物学指标、卫生经济学评价以及生活质量评价。采用哪些结局指标将直接影响研究设计。目前对结局指标的选择正从基本的治疗安全与疗效向精准医学转变，更关注疾病症状的改善、生活质量的提升以及诊疗手段的费用。上述转变对临床研究的设计和开展提出前所未有的要求，也增加了临床研究需要考察变量的多样性和复杂性。

1.1 生物学指标

生物学指标指反映患者病理变化过程的临床结局或者结局替代指标。最合适的生物学指标应该与患者健康状况直接相关，通常包括临床疗效和安全性。若研究对象为严重或致死率较高的疾病，该指标可以是重大临床事件（如死亡或再出血的发生）。部分临床研究中，结局替代指标可能是疾病严重程度的分级或评分，被称为"有序分类数据"，如肝硬化患者肝功能 Child-Pugh 分级，也可能是连续资料，如肝静脉楔压（hepatic venous pressure gradient，HVPG）、动脉血压或血胆固醇水平[2]。

采用生物学指标作为结局指标时，其测量需要选择合适的方法，制定测量标准，避免主观与人为因素带来的测量偏倚。要求测量手段可靠，具有足够的敏感性，能测量出患者健康状态的变化，测量标准统一，尽可能采用盲法判断，同时考虑各种因素（包括测量的时间、地点、人员、方法、条件及记录方法等）对测量的影响。

评价结局指标通常为比较治疗前后变化的绝对量或者百分比。

临床研究案例：在卡维地洛的疗效评价中，临床结局指标包括生存/死亡，是否静脉曲张再出血，食管胃静脉曲张严重程度变化；结局替代指标包括门脉压力、HVPG 以及可以直接在用药后观察的临床指标包括心率变化。

结局的测量：死亡尽管容易判断，但需要判断是否为出血相关死亡，在结局评价中需要考虑非出血相关死亡是否计入结局评价，是否与卡维地洛治疗相关。静脉曲张破裂出血可以临床诊断，但准确的应该是内镜诊断，因为肝硬化患者上消化道出血也可以来源于消化性溃疡，而临床上急诊出血的内镜检查存在较大风险，研究实施的依从性存在问题，即使行内镜检查，排除一般溃疡出血，直接观察到静脉曲张出血也很困难。门脉压力测定可以反映药物治疗的病理生理改变，但直接测压存在穿刺出血风险，除了拟行介入治疗或者手术患者，临床难以实施。HVPG 可以部分替代直接测压，但受很多因素影响。无创门脉压力测定是一种简便的方法，但反映患者最终疗效（再出血或生存）的可能性低于 HVPG。心率变化最简便易行，但心率下降一定伴随门脉压力下降，反映临床结局的代表性低于 HVPG 或无创评价指标。

结局的评估：在生存与死亡的结局中，需要考虑非静脉曲张出血相关死亡；在静脉曲张破裂出血的结局中，需要考虑无出血的死亡、肝移植处理，也需要考虑发生肝癌、门脉血栓、腹水感染、肝性脑病等其他并发症的结局评价。在评价门脉压力或 HVPG 作为结局替代指标时，需要考虑 HVPG 下降多少有临床意义（临床结局会出现显著差异），或者将 HVPG 转化为治疗反应率（多少比例患者 HVPG 有明显下降）。

1.2 卫生经济学评价

治疗过程中的经济因素越来越受重视。患者消耗的医疗资源可以与诊疗过程中的花费结合起来进行经济学分析，用以更好地评价耗用资源后得到的"三效"，即效果（effectiveness）、效用（utility）和效益

(benefit)。对于两项措施的比较，在关注新措施疗效的同时，也要注意成本变化。卫生经济学指标以独特的方式被纳入临床研究中，将患者的治疗开支及医疗资源消耗作为评价结局指标的重要方面[3]，在收集成本数据时需考虑研究目的。目前基于卫生经济学考虑的结局指标尚未达成共识，但随着越来越多的临床研究开始将卫生经济学因素纳入其中，这一方法也在实践中逐渐得到发展。

1.3 生活质量评价

临床研究中患者对自身生活质量的评价越来越受重视，其中部分原因由世界卫生组织（World Health Organization，WHO）的最新健康定义所导致。WHO认为，健康不仅是无病和不虚弱，而且应当是身体、心理、社会功能3方面的完满状态[4]。生活质量可以用患者报告结局（patient-reported outcome，PRO）来估计，常用量表包括WHO-QOL、SF-36、SF-12、EuroQOL（EQ-5D）、Nottingham Health Profile（NHP）、Rosser index、Sickness Impact Profile（SIP）等。

量表分为普适性量表与疾病特异性量表。普适性量表关注与生活质量多个方面相关的问题，适用于无法确定治疗影响或者对多个疾病进行比较时；但其同时也存在缺点，如对健康状况改变的测量不够敏感。疾病特异性量表仅针对某种疾病相关的生活质量，能够准确测量出机体健康状态的变化，也更为患者所接受；适用于症状性疾病，患者易于察觉出因治疗而产生的疾病严重程度变化[5]。

2 结局指标的选择与临床研究设计

选择何种临床研究设计，一是与研究目的有关，回答需要研究的问题，二是与研究可行性有关，能否筛选到合适的患者、随访时间、样本量、研究经费等因素会直接影响方案能否顺利实施。此外，对研究结局的把握也是选择研究设计方案的重要依据。

2.1 病例-对照研究

病例-对照研究的基本原理是以确诊患有某种疾病的患者作为病例，以不患有该病但具有可比性的个体作为对照，通过询问、实验室检查或复查病史，搜集既往各种可能的危险因素暴露史，测量并比较病例组和对照组中各因素的暴露比例，经统计学检验，若两组差别有意义，则可认为因素与疾病之间存在统计学关联。在评估各种偏倚对研究结果的影响之后，借助病因推断技术，推断某个或某些暴露因素是疾病危险因素，从而达到探索和检验疾病病因假说的目的。一般策略是比较来自同一总体、具有代表性的病例组与非病例组间潜在危险因素的频率或水平高低。

临床研究案例：在肝硬化门脉高压预防再出血的相关临床问题中，寻找药物治疗（如传统的心得安治疗、内镜治疗、新药卡维地洛治疗）失败的原因，可以采用病例-对照研究设计。研究人群是采用心得安预防再出血的肝硬化患者，病例组是治疗失败（再出血）的患者，对照组是治疗有效（无再出血）的患者，危险因素考虑门脉血栓或合并肝癌，结局变量（观察评价的指标）为患者在开始心得安治疗时是否存在门脉血栓或者肝癌。指标的测量可以采用增强CT检查（患者开始治疗时有相关检查），CT评价可以邀请放射科专家重新盲法读片（不知患者是否有再出血）。

病例-对照研究的结局指标可包括多个，但样本量估计应关注主要结局指标，如门脉血栓比例或者合并肝癌概率，研究得出的结论也只能说明疗效差异是否与门脉血栓存在有关。如要评价心得安疗效是否与门脉血栓有关，还需要加入未采用心得安治疗的人群，采用其他设计评价。

2.2 队列研究

队列研究是在一个特定人群中选择所需的研究对象，根据目前或既往某时期内是否暴露于某项待研究的危险因素或不同暴露水平，将研究对象分组，观察随访一段时间后，检查并登记各组人群待研究的预期结局发生情况（如疾病、死亡或其他健康状况），比较各组结局发生率，评价和检验危险因素与结局关系[6]；队列研究也可根据不同干预措施进行分组。由于队列研究需观察随访一段时间，因此失访不可避免，在研究开始前要考虑失访率，按照估计样本量增加10%作为实际样本量。

临床研究案例：在评估卡维地洛和内镜治疗预防肝硬化静脉曲张再出血的临床研究中，可以采用前瞻性队列研究设计。研究人群及目标人群是肝硬化门脉高压静脉曲张出血后需要采取措施预防再出血的患者群，评估措施是卡维地洛，对照措施为内镜治疗，结局指标可以是生存或者死亡，也可以是再出血是否发生。因为内镜治疗预防再出血不是通过降低门脉压力，因此，采用门脉压力/HVPG等中间结果不适合作为结局观察指标。样本量估算需考虑两种措施预防再出血的效果，并且与入选人群是否存在高危因素有关（高危人群结局事件发

生率与普通人群不同），随访时间也是影响结局事件发生高低的因素，影响样本量估算。

2.3 随机对照试验

随机对照试验将研究对象随机分组，对不同组的研究对象实施不同干预，观察不同干预手段对结局事件的影响。这种方法可有效剔除混杂因素的影响，为临床实践提供高级别证据。

临床研究案例：在卡维地洛通过降低肝硬化门脉压力预防患者静脉曲张再出血的研究中，研究者通过单盲随机对照试验，将受试者分为卡维地洛组和心得安组，观察患者 HVPG 下降程度，比较两种药物降低门脉压的疗效。HVPG 是客观可信、重复性高的结局替代指标，反映药物对门脉压力的改善情况，可以提高研究结果可信度。盲法设计可以减少资料收集、分析阶段的信息偏倚，使结果更真实可靠。结果评价可以直接使用 HVPG 下降程度进行组间比较，也可根据临床意义即 HVPG 下降带来的再出血减少定义 HVPG 有效下降，比较两组疗效差异。

结局指标如果是再出血或生存/死亡，不仅需要随访，且随访时间至少在 12 个月以上，而且结局事件出现率低，样本量会显著增加，而 HVPG 缺失则不能从机制上说明药物作用靶点。选择中间替代结果还是最终临床结果各有利弊和特点，与样本量、研究目的、研究可操作性及随访时间等因素密切相关[7]（图 1）。

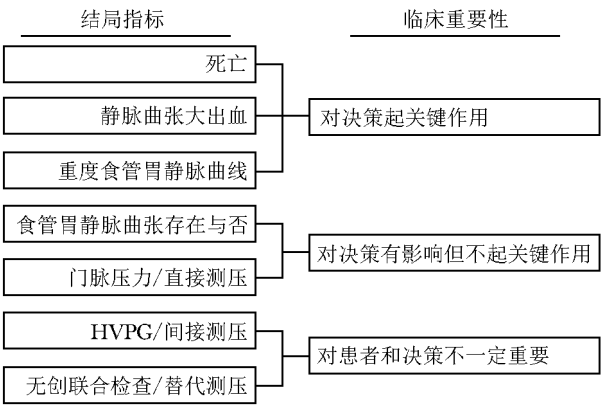

图 1 药物治疗肝硬化门脉高压可选择的结局指标及其临床重要性
HVPG：肝静脉楔压

3 样本量估计与结局指标的选择

样本量估计需要考虑以下 4 方面因素：（1）效应量，在效应量及其差异大小之间，通常两组之间效应量差异更重要，差异越大所需样本量越小；（2）检测指标结局评价的标准差（只适用于连续变量）；（3）数据的检验效力，检验效力（1−β）越大，所需样本量越大；（4）数据的显著水平，α 值越小，所需样本量越大。前两个数据是由研究对象及实验设计而定，且与结局指标选择直接相关，研究者必须定量说明需要检测的效应量，而检测指标结局评价的标准差可以通过预实验结果、其他类似实验结果、本实验室数据或查阅文献得出。后两个数值根据惯例有一般约定，如大多数研究不能接受效能低于 80%，而显著水平则应该控制在 5%以下[8]。

针对不同类型结局指标及不同研究设计，样本量的估计方法存在差异，主要体现在以下几方面。

3.1 二元数据的样本量估计

二元数据主要用来判断随访期内事件是否发生。估计样本量时通过效能、显著水平、效应量即可判断。例如，若过去试验证明某个群体疾病状态（结局指标）的发生率为 20%，当某种暴露因素存在时，有 80%的可能性观察到发生率上升到 50%，显著水平为 0.05，利用公式或计算机程序可以计算得出，评价暴露因素至少需要 44 个研究对象。需要注意的是，除非特殊设定，临床研究应该是双侧检验，因为效应量的改变既可能增加，也可能减小。

3.2 连续数据的样本量估计

较二元数据样本量估计的公式简单，但需要两组数据的平均数与标准差。

3.3 关注研究事件发生时间的样本量估计

很多临床研究关注事件发生时间，涉及数据模型较复杂，样本量可以采用预后研究设计的生存曲线比较进行估计。

临床研究案例：采用随机对照试验设计，比较卡维地洛和心得安对降低门脉压力减少再出血的效果。如果选择 HVPG 下降作为观察指标，治疗后两组 HVPG 值或两组下降的差值（计量数据）是计算样本量的依据。当然，如果定义 HVPG 下降 20%为有效，两组反应率（计数数据）则是计算样本量的依据，可能会增加样本量。

治疗组比对照组 HVPG 多下降 20%有临床意义，标准差为 2%，成组比较（m=1∶1），$\alpha=0.05$，$1-\beta=0.8$，估计单组样本量 $n=17$；对照组 HVPG 达标 50%，治疗组优于对照组（1.5 倍，75%），成组比较（m=1∶1），$\alpha=0.05$，$1-\beta=0.8$，估计单组样本量 $n=58$。

如果选择再出血作为阳性结局指标，那么在药物治疗后，需要一定的观察时间让足够的阳性事件出现，且再出血出现的概率远低于 HVPG 无效的概率（即使 HVPG 未下降，患者也不一定出现再出血），因此需要更多样本量。如果用死亡作为结局指标，随访时间更长，结局事件出现的概率则更低（因为再出血不一定死亡），所需观察的样本量需更大；此外，再出血后给予不同干预还会影响死亡结局（有效的重复干预或者补救措施的干预使死亡率更低），同样需要在估计样本量时统筹考虑。

选择再出血作为结局观察指标，对照组中位再出血时间 16 个月，R = 1.5，入组时间 A = 12 个月，延长随访 F = 12 个月，成组比较（m = 1∶1），$\alpha = 0.05$，$1-\beta = 0.8$，估计单组样本量 $n = 208$；选择生存与死亡作为结局指标，假设对照组中位生存时间 24 个月，R = 1.5，入组时间 A = 12 个月，延长随访 F = 12 个月，成组比较（m = 1∶1），$\alpha = 0.05$，$1-\beta = 0.8$，估计单组样本量 $n = 282$。

样本量可以通过专业软件进行计算，常见软件包括：SPSS，MINITAB 和 SAS 等[9]，也可由网络下载小程序 PS 直接计算（http：//biostat. mc. vanderbilt. edu/wiki/Main/PowerSampleSize）。

4 结局测量中需要注意的问题

结局测量是评价研究结果、选择合适测量值的重要部分，须注意可靠性、真实性与多样性。（1）可靠性指测量的分数或指标在多次重复试验中保持不变。统计学中有专门的数据来分析可靠性，例如 kappa 值、组内和组间相关系数等。（2）真实性指结局测量真正获取到研究者想要的信息，具体体现为结局测量是否能测量出研究者真正关注的变量、测量准确性如何、测量结局是否全面、抽象变量测量是否准确，如患者焦虑与抑郁程度。（3）多样性是指某一种测量值在研究人群的分布情况。如果研究变量是由研究者自行汇报，如生活质量等，还需要关注调查表反馈情况，以确保数据准确可靠[10]。

4.1 多终点结局测量

部分临床研究将多个终点事件同时包括在结局变量中，这种做法会增加Ⅰ类错误的发生概率，提高假阳性的发生率。在评价结局时，尽量选择单个评价指标；验证多种假说时才选择多个终点事件。选择多终点事件作为结局测量指标时，首先要考虑主要结局指标，其他作为次要终点指标，样本量依据主要结局指标估计；其次，任何一项终点事件出现均作为结局，更符合临床实践，减少所需样本量及随访时间，但需要注意不同临床事件对结局影响的程度或性质可能不同，下结论或者推广应用时需权衡。在预防性抗生素使用的研究中，发热、白细胞升高等可作为可能出现感染的相同结局判断，处理的结果是增加抗生素应用，与研究结果推广到临床实践的选择相同；在肝硬化患者应用抗生素改善预后的研究中，是否将死亡、出血、肝性脑病、自发性腹膜炎、肝肾综合症、门静脉血栓、肝癌等事件设定为联合终点需要考虑其合理性，将肝癌发生作为联合终点并不合理，死亡与自发性腹膜炎的比例亦明显不一致，这些不合理的选择和设定会导致错误的研究结论。

4.2 主观与客观结局指标

大多临床结局需要临床医生或研究人员进行诊断或评估，具有一定的主观性；实验室检查也可能被视为主观结局指标。部分临床结局，如任何原因导致的死亡，可以被轻易确认，无需经过人为解释或判断，被称作客观结局指标。多种方法可以纠正主观结局指标中的不确定性，例如对疾病严重程度和发病范围采用标准化评价，尽可能采用客观结局指标可有效提高研究结局测量的可靠性[11]。盲法检测、重复检测、统一标准化检测等方法也是增加检测可靠性的重要措施。

4.3 医生与患者评价的结局指标

美国食品药品监督管理局将临床试验结局分为 3 种，分别为患者评价、医生评价和旁观者评价。患者评价直接来自患者疾病和自身健康状况报告，其原始数据未经过修改或解释，由患者直接提供，或由数据收集者记录。医生评价最常见，由受过专业培训的医生进行结局评定。旁观者评价指由不具备医学背景的人员记录结果和评价结局变量（例如观察者可以是教师或保姆）；旁观者评价只能在变量可被直接观察记录的情况下采用（例如某种征兆或行为），不可用作评估症状（例如疼痛）或其他无法观察的概念；当患者本人无法汇报时可采用旁观者评价（例如婴幼儿），是患者评价的补充形式。在实际测量中，应尽量实现多种评价同时进行[12]。

4.4 中间结局与最终结局

临床试验中采用中间结局或者替代结局较为常见，而在观察性试验中则相对少见。采用中间结局可能是为了减少随访时间，或者是为了阐述干预措

施的作用机制及环节。例如，研究他汀类药物控制血脂的研究，最终结局指标应是冠心病的发病率，但偶尔可将血脂水平作为中间结局指标，间接推断药物疗效。采用中间结局可大大缩短随访时间，增加研究效率，选取的客观评价指标要与药物治疗的病理生理机制一致。当然，中间结局可能会导致无法观察到真实完整的治疗效果或副作用，导致过分夸大临床疗效[2]。

5 总结

结局指标的选择受多种因素影响。选择合理、清晰、尽量客观的结局指标，并根据研究问题和预期结果，明确研究目的，计算样本量，至关重要。结局指标不局限于传统生物学指标，应该综合考虑，结合患者与临床研究的实际情况而定。为了增加试验结果的可靠性，应尽量选择客观结局指标、长期随访的最终结局指标以及患者与医生同时评价的指标。

参 考 文 献

[1] Banares R, Moitinho E, Matilla A, et al. Randomized comparison of long-term carvedilol and propranolol administration in the treatment of portal hypertension in cirrhosis [J]. Hepatology, 2002, 36: 1367-1373.

[2] Heyse JF. Outcome measures in clinical trials [M]. Hoboken: John Wiley & Sons, Inc., 2005: 1-7.

[3] Robinson R. Cost-effectiveness analysis [J]. BMJ, 1993, 307: 793-795.

[4] Grad FP. Constitution of the World Health Organization [J]. Bull World Health Organ, 2002, 80: 983-984.

[5] Dawson J, Doll H, Fitzpatrick R, et al. The routine use of patient reported outcome measures in healthcare settings [J]. BMJ, 2010, 340: c186.

[6] Mann CJ. Observational research methods. Research design Ⅱ: cohort, cross sectional and case-control studies [J]. Emerg Med J, 2003, 20: 54-60.

[7] Zhong B. How to calculate sample size in randomized controlled trial? [J]. J Thorac Dis, 2009, 1: 51-54.

[8] Dell RB, Holleran S, Ramakrishnan P. Sample size determination [J]. ILAR J, 2002, 43: 207-213.

[9] Chernick MR, Friis RH. Software packages for statistical analysis [M]. Hoboken: John Wiley & Sons, Inc., 2003: 356-361.

[10] Sinha IP, Altman DG, Beresford MW, et al. Standard 5: selection, measurement, and reporting of outcomes in clinical trials in children [J]. Pediatrics, 2012, 129 Suppl 3: S146-S152.

[11] Porter ME, Larsson S, Lee TH. Standardizing patient outcomes measurement [J]. N Engl J Med, 2016, 374: 504-506.

[12] McLeod LD, Coon CD, Martin SA, et al. Interpreting patient-reported outcome results: US FDA guidance and emerging methods [J]. Expert Rev Pharmacoecon Outcomes Res, 2011, 11: 163-169.

（收稿日期：2017-12-05）