

第五章

系统评价和 Meta 分析

本章要点

- 系统分析和 Meta 分析的基本步骤。
- 掌握 PICOS 的含义及应用。
- 掌握森林图的含义。
- 预后评价的 EBM 框架,解释如何判定预后研究结果的真实性和研究对象的代表性,资料的完整性,分析的合理性;结果估计的精确性;和结果的应用性。

教学目的

- 掌握: ① PICOS 原则;② 森林图的含义;③ 文献评阅的方法。
- 熟悉: ① 系统分析和 Meta 分析的基本步骤和注意事项;② 漏斗图的含义。
- 了解: Meta 分析异质性意义及模型选择。

一、系统综述的概念

系统综述(systematic review, SR),又称系统评价,是对某一特定的临床问题(可涉及临床实践中任何方面,包括预防、医疗、护理、政策等)系统而全面地收集全世界所有已发表或未发表的相关研究文献,根据研究目的和临床流行病学原则,对文献进行评价并筛选出符合质量标准的文章,进行定量或定性的综合,得出综合可靠的结论。

对有争议或相互矛盾的小样本研究,系统评价可通过严格的、科学的方法进行分析和合成,探究其分歧的原因,很大程度上提高了医学信息的可靠性和科学性,为临床实践提供决策依据和今后的研究导向。但由于系统评价是对原始文献资料的二次分析和综合,必然会受原始文献的质量、系统评价的方法以及评价者本人的主观认识的制约,因此在阅读与应用系统评价的证据时,仍需要持谨慎的批判态度,不能盲目接受,避免误导。

二、系统评价和 Meta 分析的步骤与方法

Cochrane 系统评价常被认为是评价干预措施效果的最好的资源。在 Cochrane 系统

综述包括 9 个步骤,包括:① 构建进行系统评价的临床问题,并确立纳入研究的标准;② 文献检索与研究选择;③ 选择研究和收集需要分析的数据;④ 按循证实践的原则和方法进行评价纳入研究存在偏倚的风险;⑤ 分析数据并进行 Meta 分析;⑥ 阐述报告偏倚;⑦ 呈现结果和发现;⑧ 解释结果并提出结论;⑨ 对系统综述结果改进与更新。但通常的系统综述至少 7 个主要步骤。

(一) 系统评价的步骤

1. 构建系统评价的问题 提出的临床问题应具有科学性和临床适用性,并有可能通过研究得到回答。形成问题将决定系统评价的结构和评价过程。系统综述解决的问题比较专一,涉及的研究对象、治疗措施、结果因素相似或相同。因此,在确定题目时,应围绕研究问题明确四个要素,即定义研究对象、干预措施、结果因素和研究设计类型四个要素。常采用 PICO 格式简洁明了地阐明问题,P 为特定的人群(population),I 为干预或暴露(intervention/exposure),C 为对照组或另一种可用于比较的干预措施(control/comparator),O 为结果因素(outcome)。在具体进行系统综述时,特别是确定纳入研究的标准时,还要考虑研究设计类型(study design),变成 PICOS。

(1) 确定研究对象:明确要研究的患者的特征,如患者的性别、年龄、种族、是否伴随疾病等。当然,可以对患者特征进一步限制,但一定要有合理的生物学或社会学意义。如研究分子靶向治疗药物治疗恶性肿瘤的疗效时,需要对肿瘤细胞类型、分化程度、肿瘤分期,甚至是基因表型等都要进行限制。

(2) 干预措施:有一个明确的干预措施。表面上看,干预措施比较容易定义,但在实际应用中需要谨慎。比如同样的药物,要明确其药物的剂型、剂量、给药的途径、疗程长度。有时,会定义一类药物的疗效,比如研究血管紧张素受体阻断剂类药物抗高血压的疗效时,更要注意药物应用的细节。

(3) 对照措施:研究一个干预措施的效果,通常是相对其对照而言。对照干预可以是空白对照,即安慰剂,也可以阳性对照。就阳性对照而言,可以是标准疗法,也可以是某一特定的对照疗法。对照不同,其相对效应也会不同,所以必须在计划书中明确说明对照是什么,以避免混淆。

(4) 疾病的结局因素:循证医学尤其强调终点结果因素。评价者不要遗漏重要的研究结果,但又要避免罗列过多不重要的结果因素。有时不同的文献所采用不同的结果因素,这时有必要时与文献作者联系,经过转换得到一个统一的结果因素,不然只能舍弃。

(5) 研究设计的类型:不同的问题有其特殊的研究设计类型。随机对照试验被认为是研究疗效最佳设计方案。而对诊断试验的评价则完全不同。同是病因研究,也因为发病率不一而选择队列研究或病例对照研究。不过在临床实践中,并不是绝对的,要回答一个临床问题,可能有多种研究方法。比如疗效的研究,理想的方法是应用随机对照研究设计,也可以应用非随机对照研究,队列研究,病例对照研究等设计。在进行 Meta 分析时不要将不同设计的结果混为一谈,可以独立分析,相互印证。应用非随机的对照研究需要注意避免两方面的偏倚。其一,设计是否考虑主要混杂因素的控制,如很多已知的影响效果的因素。同时,由于没有分组隐藏和双盲,可能会夸大疗效,下结论时应慎重。其次,特别要重视发表偏倚的问题,进行相关的分析。

确定 PICOS 后,系统综述需要制定相应的研究入选标准和排除标准。这些标准与研

究质量、研究对象、干预措施和对照、结果因素和随访时间相关。临床试验质量和特征能影响结果。理想地,SR 只对那些设计和实施良好的随机对照研究进行综述,也就是入选标准应要求患者随机分组、根据意向性治疗(intent-to-treat, ITT)原则进行分析、客观并采用盲法。评价研究质量过程往往是主观的,而且许多研究的报道常不完整,所以,在实际进行 SR 时,通常先定一个基本的入选标准,然而通过作敏感性分析来校正研究质量。

2. 文献的检索与研究选择 明确了系统综述所要研究的问题,结合文献纳入标准和排除标准,创建合理的文献检索策略。检索时宁可多选也尽量避免遗漏。制定检索策略、进行全面无偏的检索也是系统评价的特点。最好不要有语言和时间的限制。检索策略的制定可参照 Cochrane 协作网各专业组制定的检索策略。检索的过程最好由两个以上的作者独立完成,再相互印证。

系统综述中的相关研究的查找策略应该是明确描述的,尤其对对照研究的搜索变得越来越直观。检索策略的制定需要充分考虑研究问题所涉及的四个方面: PICO,即研究对象、干预措施、对照措施和研究结局。检索的手段有电子检索和手工检索,发表及未发表的资料。常用的英文数据库包括 MEDLINE、EMBASE、Cochrane 图书馆等,常用的中文数据库包括中国生物医学期刊库(CBM)、中国知网(CNKI)、万方数据库等。不同数据库有自己的侧重,比如 Medline 和 Embase 两个数据库重复率只有 34%,所以在作系统综述时应该同时查找多个数据库。很多的对照研究会 The Cochrane Controlled Trials Register 系统中注册,这个数据库也是所有作系统综述研究必须用的数据库。当然,文献的搜寻应包括那个没有发表的研究,这些研究的结果往往与发表的研究不一致。发表偏倚就发生在这种情景。

所有检索到文章(包括以前发表的综述)的参考文献是一个获得完整数据库的有效而简单的补充。此外,Cochrane 图书馆中的数据库之一,疗效综述的摘要数据库(Database of Abstracts of Reviews of Effectiveness, DARE)也能提供以往发表的临床试验效果评价的信息,可作为参考。其他文献如技术总结报告、会议文章汇编、研究生论文等可能是获得那些未发表文献的重要途径。若涉及药物或医疗器械,甚至可以向生产厂家索取有关资料。与同领域中专家交流,从中获得已经完成而尚未发表,或从未发表的临床试验也是获得完整数据库的补充。

选择研究是指根据事先拟定的入选标准和排除标准,从收集到的所有文献中检出能够回答研究问题的文献资料。评估所有可能合格的文献是否满足系统评价的入选标准和排除标准。文献的判断过程有一定的主观性,一般要求两人独立进行,若出现分歧的情况,由双方讨论协商解决或第三方裁定。

3. 纳入文献质量的评阅(critical appraisal) 研究质量是指研究的结果是否真实地反映了客观实际情况,主要是指内部真实性(internal validity)。其评价方法主要看该研究在设计、实施和分析中的可靠程度。相对应的是研究的外部真实性(external validity),指研究中的对象、干预措施和结局测量等特征对问题和结果的解释所产生的影响。

系统综述需要对纳入研究的质量进行严格文献评阅(critical appraisal),目前有不同的质量评价工具用于研究的质量评价,对不同设计类型也有不同评价方法。比如独立评价临床试验方法学质量的工具,强调研究设计、实施和评价三个阶段的管理。常用的有改良 Jadad 评分,但近年来较多地采用 Cochrane 推荐的质量评价标准——偏倚风险评估

(risk of bias),通过评价其可能存在的偏倚来反映研究的真实性,将随机对照试验的质量评为低风险、高风险和风险不确定三类。

在Cochrane偏倚风险评估中,包括了常见的偏倚类型有:①随机分配方案的产生(random sequence generation,归属于选择性偏倚 selection bias),产生的原因是随机不充分,包括伪随机。导致影响预后或疗效的研究对象特征在治疗组间存在的系统差异。②随机分配方案的隐藏(allocation concealment,归属于选择性偏倚 selection bias),产生的原因是分组前没有做到分配方案的隐藏,研究人员有意不遵循分配方案,导致影响预后或疗效的研究对象特征在治疗组间存在的系统差异。③受试对象和研究人員设盲(blinding of participants and personnel,归属于实施偏倚 performance bias),产生的原因是受试对象和研究人員在研究过程中获知治疗分组的情况,导致除干预措施外,治疗组间其他医疗措施存在的系统差异。④结局测量时设盲(blinding of outcome assessment,归属于检测偏倚 detection bias),产生的原因是研究结局测量者知道治疗分配方案,导致结局测量时治疗组间存在的系统差异。⑤结局数据不完整(incomplete outcome,归属于退出偏倚 attrition bias),产生的原因是不完整结局数据的数量、性质和处理方法导致研究结果的系统差异。通常辨别的方法作者是否应用了意向性分析(intent to treat, ITT)。⑥选择性报告(selective reporting,归属于报告偏倚 reporting bias),产生原因是作者只选择某些结局进行报告。

系统综述中研究质量评价有重要的意义:①作为系统综述的研究纳入标准;②在Meta分析中根据质量高低进行敏感性分析,甚至可以赋予每个研究的权重;③探索质量与研究结果异质性之间的相关性,有助于结果的推论;④有助于综述结论的建议分级;⑤根据现有研究的不足,提出进一步研究的建议。

4. 资料提取 资料提取就是将能反映原始研究概况的资料提取出来,将原始资料条理化、数据化,既便于集中分析,也便于读者了解资料的情况。数据提取是进行系统评价的基础工作,是原始研究与系统综述分析之间的桥梁。数据提取原则上应能尽量详尽而准确地反映原始资料情况。资料的收集需要根据具体研究特点设计标准化的资料提取表,提取表通常包括文献来源、研究的合格性、研究的设计与方法、研究的特征和供综合分析的数据等四部分内容。这些资料的收集通常有一定的规范和细化。特别是研究质量要素,如随机分配方案的产生、随机方案隐藏、盲法、样本量估计;研究特征包括研究对象、干预措施、对照措施和研究结局;和供系统综述资料分析数据的收集往往与资料分析结果的解释,异质性的探索等密切相关,需要十分重视。

资料提取表可以采用纸质也可以应用电子表格,由2名评价员对符合纳入标准的试验独立进行提取数据,完成后交叉核对。不一致处通过讨论协商确定,或由第三方判定。数据提取后还应就数据提取和质量评价的一致性进行评价。

5. 资料的定性与定量综合 系统综述的资料综合包括考虑定性与定量两方面的分析。定量分析,即Meta分析比较直观,近年来Meta分析的方法学上进展也较多,受到广泛的接受,有时甚至被滥用。定性分析其实一直是系统综述的一个重要方法,尤其适合对新的临床问题分析,近年来逐渐受到临床医生的重视。决定定量还是定性分析,需要考虑几方面问题:纳入研究是否存在异质性,数据是否适合定量分析,定量综合后结果的意义是什么,定性分析是否能更全面地了解临床问题等。

定量分析包括三个方面：①效应的合并：应根据资料的类型及评价目的选择效应量和统计方法。对于分类变量，可选择比值比(OR)、相对危险度(RR)、危险度差值(RD)等效应指标。对于连续性变量，当结果因素的度量单位相同时应选择加权均数差值(weighted mean difference)；而当结果因素的度量衡单位不同时，应选择标化的均数差值(standardized mean difference)。进行效应合并时，根据各研究间是否同质，相应地选择固定效应模型和随机效应模型。Meta分析的结果常用森林图(forest plot)来表示。②同质性(homogeneity)检验：或称为异质性(heterogeneity)检验，是对原始研究之间结果的变异程度进行检验。如果有显著性差异，同样不宜将不同研究的结果进行合成，或选择随机效应模型合成结果。当研究结果间存在异质性时，不要轻易下结论，应慎重分析异质性的来源和临床应用时的解释。③敏感性分析：系统评价综合了很多单项研究，故当研究总数不多，而某项研究又特别强势时，应考虑加入该研究是否会极大地改变结论，从而判断结果的稳定性。④偏倚的检测和处理：减少偏倚是结论真实的可靠保证。同样在系统评价中也要尽量减少偏倚。发表偏倚(publication bias)是系统评价中常见的偏倚。它是指阳性结果的研究容易得到发表，而阴性结果不容易获得发表。因而人为地造成治疗效果的片面夸大。

6. 对结果的解释 系统综述需要对综合定量分析(Meta分析)或定性分析的结果进行解释。对所有纳入研究中质量、干预措施的效应进行概述和分析，对不同条件下干预措施效应进行分析、综合结果与其他类型研究的结果进行比较，最后对证据的强度、推广应用性、治疗措施的利、弊、费用的权衡等方面进行讨论，形成对临床实践有指导意义的建议，以及指导进一步临床研究的线索。

(1) 证据的强度：首先应对最后纳入文献的方法学质量和系统评价本身质量进行讨论，这部分内容将决定系统评价所得结论的强度。其次是对未纳入文献加以讨论，有不同意见时应引起重视，尤其是那些副反应的信息。有时研究很难提供治疗与重要结局之间的因果联系，特别是对少见病的危险因素研究只能提供间接证据，如采用中间指标，或称替代指标，这需要长时期的观察来证明其疗效。

(2) 推广应用性：临床工作者需要根据系统评价所提供的信息为临床服务，但在应用时必须判断这些信息是否适合，如患者是否相似，研究场所是否相似等。若系统评价的结论来自成人，那该结论应用于儿童将会十分冒险。所以要求评价者对证据可能应用的情况及影响效果因素进行讨论，包括生物学或文化的差异、患者对治疗的依从性、患者特征、治疗费用及患者的态度。

(3) 副反应：评价者应当充分说明治疗可能带来的副反应。高度重视副反应的发生频率及严重程度。

(4) 结论的意义：评价者需要对系统评价的发现对临床实践的意义进行总结，说明该结果对临床实践和进一步研究具有什么样的意义。

7. 对系统综述结果的修订与更新 循证医学所要求的最佳研究证据具有时效性，随着临床研究新结果的产生，系统综述的结果需要作相应的更新。全球每年发表大量的临床随机对照试验结果，系统评价中位生命周期为5.5年，因此有必要定期对系统评价进行更新。其次，系统综述是一种科学的严谨的系统方法，有助于临床决策、卫生政策制定等，但对系统综述不能过于迷信，其质量的高低仍依赖于自身方法学的完善及所入选的单个试验的质量。更重要的是，系统综述结果最终需要接受临床实践的检验，当系统综述提

出建议与临床实践的结果不符合时或存在不足时,系统综述需要作出修订或纠正。

(二) Meta 分析

Meta 分析的统计基础属于多中心研究中的统计理论范畴,但由于 Meta 分析的资料往往来自文献报道结果而不是原始资料,它所用的统计方法具有特殊性,但其原理与其他统计方法是相同的。

1. Meta 分析中资料的异质性和统计模型的选择

(1) 评价资料的异质性:按 Meta 分析的原理,只有同质资料的合并、比较才有意义。异质性检验(heterogeneity testing)是 Meta 分析的一个重要内容。Meta 分析过程必须对多个研究的结果进行异质性分析,尽可能地找到导致异质的原因,并有效解释。文献间的异质性主要表现为三个方面:生物学、方法学和统计学异质性。

生物学异质性主要是指研究对象、干预措施及研究终点的差异。产生生物学异质性的主要原因有:研究对象的差异(纳入和排除标准不一致造成研究人群的特征不同,如年龄、性别和种族、疾病严重程度的差异)、干预措施的差异(如剂量、剂型、用药途径、疗程、依从性等不同)、结局评估的差异(结局的定义和结果的表达方式、不同的测量方法和测量时点)。方法学异质性主要是指由于研究设计、实施、分析的不同引起的差异。产生方法学异质性的主要原因有:分配隐藏的方法不同、盲法的方法不同、研究过程中对结局的定义和测量方法的不一致、统计分析方法不同,尤其是对退出或失访病例、缺失资料病例的处理方式不同。生物学异质性和方法学异质性主要根据专业知识判断,一般采用描述性方法,若各研究间的两种异质性较大,就不能轻易合并。

统计学异质性主要是指各个研究之间效应值的差异,它是生物学和方法学上多样性的直接结果,可以进行定量的、客观的评价。评价统计学异质性采用异质性检验(tests for heterogeneity)有:① 图示法:森林图(forest plot)和拉贝图(Labbe plot)。② 统计学检验法:异质性的检验方法有 Q 统计量的卡方检验, I² 检验。Q 统计量的卡方检验检验效能较低,判断异质性设定的 P 值与效应值的统计学检验判断标准不同,通常是以 P 值 < 0.1 作为具有异质性的判断标准,而不是通常设定的 P < 0.05 为判断标准。如果纳入研究较多时,即使这些研究间的效应量是同质的,由于抽样误差的存在,也可能出现异质性检验结果有统计学意义,即 P < 0.05,所以在应用 Q 检验法结果时应慎重。I² 统计量定量地描述异质性的大小, I² = 0, 表示完全同质; I² < 25%, 表示低异质性; I² 为 25% ~ 50%, 表示中等异质性; I² > 50%, 表示高度异质性; I² > 70%, 建议分开描述,不合并。缺乏统计学的异质性,也不代表没有生物学异质性。

(2) 固定效应模型和随机效应模型:Meta 分析所用的统计方法分为固定效应模型(fixed effects model)和随机效应模型(random effects model)的统计方法。固定效应模型的统计假设个体研究间的效应无差异的情况(具有统计学同质性),即假设各个研究是一个总体的随机样本,具有相同的总体效应。而随机效应模型假设干预措施效应对不同的对象或使用不同的对象将产生不同的效应,即该措施的效应不具统计学同质性。以上已经提及,固定效应模型的统计方法常用的包括 Mantel-Haenszel 法、Peto 法和 General Variance-Based 统计方法。随机效应模型的统计方法常用的是 Der Simonian and Laird 方法。

当存在明显异质性时,应首先检查提取与输入数据的环节是否存在错误,然后分析异质性的来源,针对原因寻求解决办法。如果异质性太强,需检查生物学异质性,若同时存在生物学异质性,可以放弃资料合并,改为单个研究的效应分析;若生物学异质性在可以

接受的范围,进一步分析可采用随机效应模型,并与固定效应模型结果相比较对不同临床情况。可以应用 Meta 回归等高级的方法进行分析,或进行敏感性分析和亚组分析探讨异质性产生的原因,以及去除异质性以后的效应变化情况。

(3) 评价效应的指标选择:测量的指标常分为二分变量或连续变量。当比较两组患者的治疗效果时,可应用二分变量。这些变量包括相对危险度(risk ratio, RR)、比数比(odds ratio, OR)、率差(risk difference, RD)。较常用的是相对危险度和比数比,用它们来估计治疗的相对效应;而率差则是用来估计治疗的绝对效果。连续资料可以是治疗组和对照组之间的粗平均数差值和相关系数等。在人群中进行的病例对照研究中,效果大小的估计可以是 OR 或归因危险度。对非实验性研究,常用的效果估计是 OR 或 RR;而在随机对照研究中,最常用的 RD。各种指标的选择均有其优缺点,在进行 Meta 分析之前应仔细考虑。

Meta 分析的方法和特殊的公式受效应测量的影响。表 2-5-1 总结了效应测量形式和统计方法的选择。

表 2-5-1 效应测量形式和统计方法的选择

模型假设	统计方法	效应测量形式
确定性效应模型	Mentel - Haenszel 法	比(OR, RR)
	Peto 法	比数比(OR)
	General Variance - Based 法	比(所有类型)、差值(RD、平均数差值等)和回归系数
随机效应模型	Dersimonian and Laird 法	比(所有类型)、差值(RD、平均数差值等)

2. 发表偏倚的检测 系统综述由于是既往已经完成的临床试验进行分析,这些资料通常来自期刊上发表的文章。而那些不具有统计学意义的阴性结果的研究,相较于有统计学意义的阳性结果的研究更难或没有机会获得发表。换句话说,就是阳性结果的研究比阴性结果的研究更容易被投稿和被发表。这个矛盾就使系统综述容易产生了发表偏倚(publication bias)。检测发表偏倚的统计学方法很多,但最常用的是应用漏斗图(funnel plot),如图 2-5-1。该方法假设样本量越大的研究,其结果越接近于干预措施

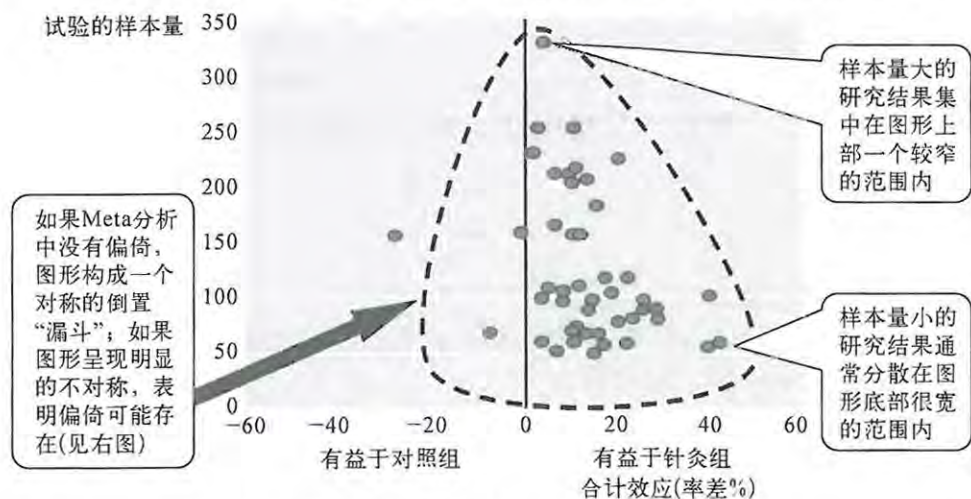


图 2-5-1 漏斗图示例及标注(引自 Tang JL, 1999)

效应的真实值;反之,当样本小的试验,由于机遇的作用,其结果就会比较分散,有的效应值可以很大,有的可以很小。因此,对于一个干预措施的评价从倒漏斗来看应该是对称分布的,当出现了不对称的情况,则可以推断可能存在发表偏倚。应该注意的是倒漏斗图形不对称还可能因为纳入的试验总体质量较差,试验数少(机遇的作用),或效应变异性过大引起。当纳入系统综述试验数太少时(如低于10个),进行倒漏斗图分析对结果的解释需要慎重,这时候的判断往往不准确。

对系统综述来说,目前要求的临床研究登记、注册制度,并建立相应的数据库有望降低这种偏倚。因为临床研究一旦注册,不管发表与否都应该提供研究结果。

3. Meta分析森林图的解读 Meta分析通过对多个研究的效应值进行合并分析,并报告其效应值得可信区间,可以用森林图(forest plots)的形式呈现分析的结果,如图2-5-2,该图以一条中心垂直线为无差异线为界(二分变量对应的数值为1,连续变量对应的为0)。试验结局的效应值横向排列,每一横线代表效应值的分布(可信区间)情况,中间的方块代表该试验效应的点估计值,若该横线触及或跨越中线,则表示该试验干预与对照比较的结局效应差异无统计学意义。横线中央的方框代表组建比较的效应值,方块的大小代表该试验在Meta分析中所占的权重。综合效应值为一菱形方块表示,中心为点估计值,两端表示可信区间。若菱形方块在中线的右边表示干预措施的效果优于对照组。

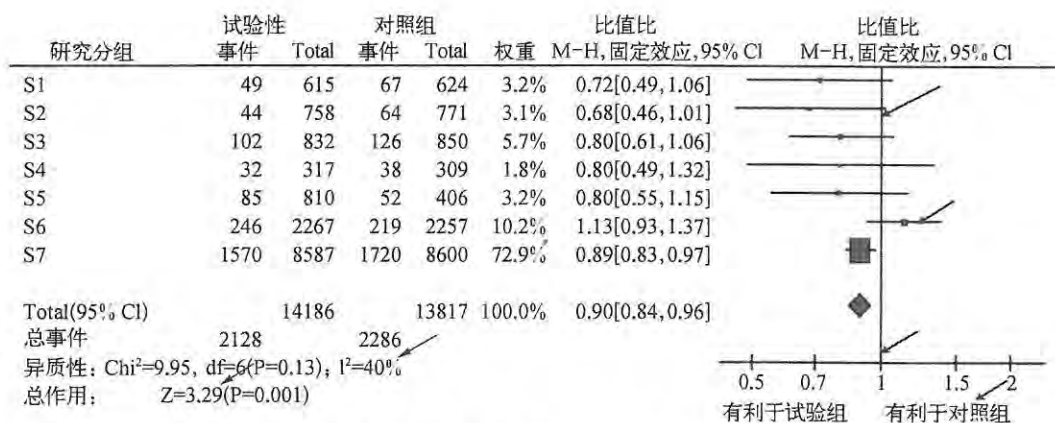


图2-5-2 Meta分析森林图例及关键点标注

(张博恒)

【思考题】

1. 简述系统综述的步骤。
2. 系统综述中如何提出临床问题? 简述关于临床疗效问题的四要素。
3. 如何处理Meta分析资料的异质性?