《Cochrane 干预措施系统评价手册》
中文翻译版

The Translation of Cochrane Handbook for Systematic Reviews of Interventions

总审校　李静　张鸣明

审译单位

四川大学华西医院中国 **Cochrane** 中心

兰州大学循证医学中心

2014．12．1

审译者名单（按拼音顺序排列）

四川大学团队

岑　啸　陈群飞　陈燕玲　成　岚　崔晓华　杜　亮　高　霑　郭　琴　郭　琴

何　佳　何　林　贾鹏丽　蒋兰慧　康德英　李　静　李　玲　李雨璘　秦天强

王　雱　王　莉　王　凌　文　进　吴红梅　袁　强　张龙浩　张鸣明　钟大可

兰州大学团队

陈耀龙　崇　乐　柯法勇　李　江　李　伦　林晓亭　刘雅莉　柳文杰　马　彬

齐国卿　申希平　田金徽　汪泽皓　汪泽皓　王小琴　肖晓娟　袁金秋

总审校　　李　静　　张鸣明

前　言

“证据”及其质量是实践循证医学的关键。高质量证据因采用了足够防止偏倚的措施，保证了结果的真实性。国际 Cochrane 协作网制作的系统评价，因具有严密的组织结构和质量控制系统，是实践循证医学最好的证据来源之一。

《Cochrane 干预措施系统评价手册》（Cochrane Handbook for Systematic Reviews of Interventions，简称手册）英文版首次于 1996 年 10 月发行（Cochrane Collaboration Handbook）。1998 年中国 Cochrane 中心因举办首届 Cochrane 系统评价培训班需要，在原华西医科大学附一院李幼平副院长、王家良教授的大力支持下，由李静、罗德诚、方芳和张鸣明老师在完成极其繁重的科研和临床工作同时，利用业余时间完成了 1997 年发行的 3.02 版本的全书翻译工作，为中国的系统评价者制作高质量的系统评价提供了第一本参考教材。最新的《Cochrane 干预措施系统评价手册》（第 5 版）是 Cochrane 协作网成立以来逐步形成的重大修订版本，共有 22 章，可在 Cochrane 协作网网页（http://www.cochrane.org/handbook）获取，旨在指导和帮助 Cochrane 系统评价作者系统、知证、明确地了解所提出的问题及解决问题的方法。Cochrane 系统评价作者严格遵循《手册》要求，采用固定的格式与内容，统一的系统评价软件（RevMan）录入和分析数据、撰写计划书、完成系统评价全文，定期更新系统评价，为临床实践提供高质量的研究证据。因而《手册》具有极高的权威性，成为系统评价作者必备的方法学指南。

目前《Cochrane 干预措施系统评价手册》已被授权翻译为西班牙语、德语、法语等。中国 Cochrane 中心是国际 Cochrane 协作网的 14 个中心之一，是中国唯一被授权翻译《手册》的中译机构。中心自 2009 年获 Cochrane 协作网授权，组织了相关研究生、流行病学、统计学、信息科学等专家同步翻译、审校。《手册》的翻译难度远超出了我们的预想，常为一个单词、一个术语、一个句子、一个段落的理解翻译讨论达数十次，旨在力求忠实于原文，并为中国读者接受。

本书的翻译审校以中国 Cochrane 中心为主，同时邀请兰州大学循证医学中心的部分同行参与，排版由《中国循证医学杂志》袁媛同志完成。Cochrane 协作网为本书的翻译、审校提供了启动经费支持。对他们的辛勤劳动和支持深表谢意。

提交之际，再阅《手册》中译本，仍有诸多瑕疵，恳请读者不吝指正，以便我们今后改进。

中国 Cochrane 中心　李静　张鸣明

2014 年 12 月 1 日 中国 成都

目　录

Cochrane 干预措施系统评价手册 i
第一章　导论 1
1.1　Cochrane 协作网 2
1.1.1　引言 2
1.1.2　Cochrane 协作网组织结构 3
1.1.3　Cochrane 评价的发表 3
1.2　系统评价 4
1.2.1　系统评价的需求 4
1.2.2　什么是系统评价 4
1.3　关于本手册 5
1.4　手册参编者 6
1.5　本章信息 7
1.6　参考文献 7
第二章　系统评价的准备 9
2.1　计划书的原则 10
2.2　Cochrane 系统评价的格式 10
2.2.1　Cochrane 系统评价格式的原则 10
2.2.2　Cochrane 系统评价计划书的框架 11
2.2.3　Cochrane 系统评价的大纲 13
2.3　制作系统评价的流程 15
2.3.1　制作系统评价的动机 15
2.3.2　规划系统评价的主题和范围 15
2.3.3　注册计划书 15
2.3.4　系统评价工作组 16
2.3.5　Cochrane 评价使用的软件 18
2.3.6　培训 19
2.3.7　Cochrane 系统评价小组的编辑过程 19

2.3.8 系统评价的资源 20
2.3.9 申请经费 22
2.4 在杂志或图书上发表 Cochrane 系统评价 22
2.5 将已发表的系统评价作为 Cochrane 系统评价发表 24
2.6 声明利益和商业资助 24
2.7 本章信息 26
2.8 参考文献 26
第三章 系统评价的维护：更新、修改和反馈 28
3.1 引言 29
3.1.1 为什么要维护系统评价 29
3.1.2 系统评价更新的频率 29
3.2 重要定义 30
3.2.1 引言 30
3.2.2 更新和修订 30
3.2.3 Cochrane 系统评价或计划书的引文版本 30
3.2.4 Cochrane 计划书相关术语的应用 32
3.2.5 Cochrane 系统评价相关术语的应用 33
3.3 与 Cochrane 系统评价相关的重要日期 36
3.3.1 引言 36
3.3.2 系统评价更新日期需要作者输入（仅在系统评价全文中，不包括计划书） 36
3.3.3 检索的日期 37
3.3.4 预期进入下一个阶段的时间 37
3.3.5 最后一次编辑的时间 37
3.3.6 声明系统评价不再需要更新的时间 37
3.4 更新系统评价需要考虑的方面 38
3.4.1 更新从何开始 38
3.4.2 更新不修改研究问题的系统评价 38
3.4.3 修改研究问题和纳入标准 40
3.4.4 分割系统评价 40

3.4.5 系统评价方法的修订 41
3.4.6 系统评价其他更改 41
3.4.7 编辑过程 41
3.5 “新内容”和历史事件表格 42
3.5.1 “新内容”事件 42
3.5.2 完成“新内容”表格 42
3.5.3 历史事件表格 43
3.6 Cochrane 系统评价反馈的归纳综合与处理 43
3.7 本章信息 44
3.8 参考文献 44
第四章 Cochrane 计划书及系统评价内容指南 46
4.1 引言 47
4.2 标题与系统评价信息（或计划书信息） 47
4.2.1 标题 47
4.2.2 作者 48
4.2.3 通讯作者 49
4.2.4 日期 49
4.2.5 新内容和历史 50
4.3 摘要 51
4.4 通俗语言总结 51
4.5 正文 51
4.6 表格 64
4.6.1 纳入研究特征 64
4.6.2 偏倚风险 64
4.6.3 排除研究特征 65
4.6.4 待分类研究特征 65
4.6.5 在研研究的特征 65
4.6.6 结果的总结 66
4.6.7 附加表格 66
4.7 研究和参考文献 66

4.7.1 研究的参考文献 66
4.7.2 其他参考文献 67
4.8 数据和分析 68
4.8.1 比较 68
4.8.2 结果 68
4.8.3 亚组 69
4.8.4 研究数据 69
4.9 图形 69
4.9.1 RevMan 软件的图和表格 70
4.9.2 其他图形 71
4.10 支持系统评价的资源 71
4.11 反馈 71
4.12 附录 72
4.13 本章信息 72
4.14 参考文献 72
第五章：立题与制定纳入研究标准 74
5.1 问题与合格标准 75
5.1.1 精心构建问题的原理 75
5.1.2 合格标准 75
5.2 确定受试者类型：哪些人和人群？ 76
5.3 确定干预措施类型：与哪个对照？ 77
5.4 确定结局类型：哪个结局指标最重要？ 78
5.4.1 列出相关结局 78
5.4.2 优化结局指标：重要、主要和次要结局指标 79
5.4.3 不良结果指标 80
5.4.4 经济学数据 81
5.5 确定研究类型 81
5.6 确定系统评价问题的范畴（广义与狭义） 82
5.7 研究问题的变更 84
5.8 本章信息 84

5.9 参考文献 84
第六章 文献检索 86
6.1 引言 87
6.1.1 一般问题 88
6.1.2 要点总结 89
6.2 检索信息源 89
6.2.1 书目数据库 89
6.2.2 期刊和其它非书目数据库源 98
6.2.3 未发表和在研的研究 103
6.2.4 要点总结 108
6.3 规划检索过程 109
6.3.1 邀请试验检索协调员和卫生保健图书馆员参与检索过程 109
6.3.2 协作网检索倡议 109
6.3.3 CENTRAL，MEDLINE 和 MEDLINE 检索：特殊问题 115
6.3.4 要点总结 117
6.4 设计检索策略 117
6.4.1 设计检索策略-简介 117
6.4.2 检索策略架构 118
6.4.3 服务提供商和检索界面 118
6.4.4 检索敏感度与精确性 119
6.4.5 受控词表和文本词 119
6.4.6 同义词、相关词、不同拼写、截词和通配符 121
6.4.7 布尔运算符（与、或、非） 121
6.4.8 相邻运算符（NEAR，NEXT and ADJ） 122
6.4.9 语言、日期和文献格式的限制 122
6.4.10 识别欺诈性研究、其它撤回发表物、勘误和意见 123
6.4.11 检索过滤器 123
6.4.12 检索更新 128
6.4.13 检索策略示范 129
6.4.14 要点总结 130

6.5 参考文献管理 131
6.5.1 书目文献管理软件 131
6.5.2 下载哪些字段 132
6.5.3 要点总结 133
6.6 记录和报告检索过程 133
6.6.1 记录检索过程 133
6.6.2 报告检索过程 134
6.6.3 要点总结 135
6.7 本章信息 136
6.8 参考文献 136
第七章 选择研究报告和收集数据 141
7.1 引言 142
7.2 选择研究 142
7.2.1 研究（非报告）作为兴趣单位 142
7.2.2 识别同一研究的多个研究报告 142
7.2.3 选择研究的典型过程 143
7.2.4 选择过程的实施 143
7.2.5 选择“排除研究” 144
7.2.6 测量一致性 144
7.3 收集何种数据 146
7.3.1 什么是数据 146
7.3.2 方法和潜在的偏倚源 148
7.3.3 受试者和实施场地 148
7.3.4 干预措施 149
7.3.5 结局测量 150
7.3.6 结果 151
7.3.7 收集其它信息 152
7.4 数据来源 152
7.4.1 报告 152
7.4.2 联系研究者 153

7.4.3 单个患者数据 153
7.5 数据提取表 153
7.5.1 数据提取表的基本原理 153
7.5.2 电子与纸质数据提取表 154
7.5.3 数据提取表设计 154
7.5.4 编码和解释 156
7.6 从研究报告中提取数据 156
7.6.1 引言 156
7.6.2 谁应提取数据 156
7.6.3 数据提取的准备 157
7.6.4 从同一研究的多个报告中提取数据 157
7.6.5 可靠性和达成共识 158
7.6.6 总结 158
7.7 提取研究结果和转化成需要的格式 159
7.7.1 引言 159
7.7.2 二分类结局的数据提取 159
7.7.3 连续性结局数据提取 159
7.7.4 等级结局数据的提取 165
7.7.5 计数数据的提取 166
7.7.6 时间事件结局数据的提取 167
7.7.7 效应估计值数据的提取 168
7.8 数据管理 169
7.9 本章信息 170
7.10 参考文献 170
第八章 纳入研究的偏倚风险评价 174
8.1 引言 175
8.2 什么是偏倚？ 175
8.2.1 偏倚和偏倚风险 175
8.2.2 偏倚风险和质量 176
8.2.3 评估偏倚的实证证据 177

8.3 研究质量和偏倚风险评估工具 178
8.3.1 工具类型 178
8.3.2 报告与实施 178
8.3.3 质量量表和 Cochrane 系统评价 178
8.3.4 收集偏倚风险评价的信息 179
8.4 临床试验偏倚来源介绍 180
8.4.1 选择性偏倚 180
8.4.2 实施偏倚 181
8.4.3 测量偏倚 181
8.4.4 随访偏倚 181
8.4.5 报告偏倚 181
8.4.6 其他偏倚 181
8.5 Cochrane 协作网偏倚风险评估工具 182
8.5.1 概述 182
8.5.2 判断依据 184
8.5.3 判断 185
8.6 风险评估描述 189
8.7 偏倚风险评估小结 191
8.8 将评估结果纳入分析 193
8.8.1 引言 193
8.8.2 偏倚风险影响 193
8.8.3 在分析中纳入对偏倚风险的评估 195
8.8.4 其他解决偏倚风险的方法 196
8.9 随机序列生成 197
8.9.1 偏倚相关理论 197
8.9.2 有关随机序列生成是否恰当的偏倚风险评估 198
8.10 分配序列隐藏 200
8.10.1 偏倚相关理论 200
8.10.2 分配隐藏是否完善的偏倚风险评价 201
8.11 对受试者和工作人员的盲法 202

8.11.1 偏倚相关理论 202
8.11.2 对受试者及研究人员的盲法是否完善的偏倚风险评价 204
8.12 对结局盲法的评估 204
8.12.1 偏倚相关的理论 204
8.12.2 结局评估中是否正确实施盲法的偏倚风险评估 205
8.13 不完整结果数据 206
8.13.1 偏倚相关原理 206
8.13.2 结局数据不完整所致偏倚风险的评估 207
8.14 选择性的结果报告 211
8.14.1 偏倚相关原理 211
8.14.2 选择性报告结果所致偏倚风险的评估 214
8.15 对有效性的其他潜在威胁 215
8.15.1 偏倚相关理论 215
8.15.2 其它来源偏倚风险的评估 218
8.16 本章信息 219
8.17 参考文献 219
第九章 数据分析和 Meta 分析 230
9.1 引言 231
9.1.1 请勿从此处开始 231
9.1.2 制定分析计划 231
9.1.3 为什么在系统评价中做 Meta 分析？ 233
9.1.4 何时在系统评价中不使用 Meta 分析 233
9.1.5 系统评价承担了什么？ 234
9.1.6 应做哪些比较 234
9.1.7 撰写计划书的分析部分 235
9.2 数据类型和效应值测量 236
9.2.1 数据类型 236
9.2.2 二分类结局指标的效应值 236
9.2.3 连续结局效应指标 240
9.2.4 有序结局和量表的效应指标 242

9.2.5 计数和率的效应指标 243
9.2.6 时间-事件（生存）结局的效应指标 244
9.2.7 以对数形式表述干预效应 244
9.3 研究设计和确定分析单元 245
9.3.1 分析单元问题 245
9.3.2 整群随机试验 245
9.3.3 交叉试验 245
9.3.4 受试者的重复观察 246
9.3.5 可重复发生的事件 246
9.3.6 多次治疗 246
9.3.7 身体多个部分（一）：身体多个部分接受相同的干预 246
9.3.8 身体多个部分（二）：身体多个部分接受不同的干预 247
9.3.9 多个干预组 247
9.4 汇总研究效应 247
9.4.1 Meta 分析 247
9.4.2 Meta 分析的原则 247
9.4.3 Meta 分析的倒方差法 248
9.4.4 二分类结局的 Meta 分析 249
9.4.5 连续性结局的 Meta 分析 252
9.4.6 合并二分类和连续性结局 255
9.4.7 有序结局和测量量表的 Meta 分析 255
9.4.8 频数和率的 Meta 分析 256
9.4.9 时间-事件结局的 Meta 分析 258
9.4.10 RevMan 中可用的 Meta 分析方法小结 258
9.4.11 Meta 分析投票计数的使用 259
9.5 异质性 260
9.5.1 什么是异质性？ 260
9.5.2 识别和测量异质性 261
9.5.3 解决异质性的策略 262
9.5.4 通过随机效应模型综合异质性 263

9.6 研究异质性 265
9.6.1 交互作用和效应修正 265
9.6.2 什么是亚组分析 265
9.6.3 进行亚组分析 266
9.6.4 Meta 回归 266
9.6.5 用于亚组分析和 Meta 回归的研究特征的选择 267
9.6.6 亚组分析和 Meta 回归的解释 269
9.6.7 分析基线风险效应 270
9.6.8 剂量-效应分析 271
9.7 敏感性分析 271
9.8 本章信息 273
9.9 参考文献 274
第十章 论述报告偏倚 280
10.1 引言 281
10.2 报告偏倚的类型及支持证据 282
10.2.1 发表偏倚 282
10.2.2 其它报告偏倚 287
10.3 避免报告偏倚 291
10.3.1 报告偏倚证据的影响 291
10.3.2 系统评价中纳入未发表研究 291
10.3.3 试验注册与发表偏倚 292
10.4 发现报告偏倚 293
10.4.1 漏斗图 293
10.4.2 造成漏斗图不对称的不同原因 296
10.4.3 对于漏斗图不对称的检验方法 299
10.4.4 敏感性分析 303
10.4.5 小结 307
10.5 本章信息 307
10.6 参考文献 308
第十一章 结果报告和“结果汇总”表（SoFs 表） 323

11.1 引言 324
11.2 研究的检索和筛选结果 324
11.2.1 研究流程图 324
11.2.2 “纳入研究特征”表 325
11.3 数据和分析 326
11.3.1 系统评价的“数据和分析”部分 326
11.3.2 森林图 327
11.3.3 其它数据表 329
11.4 图 330
11.4.1 图的类型 330
11.4.2 以图片形式选择 RevMan 分析 330
11.4.3 附加图 330
11.5 “结果总结”表（SoFs 表） 331
11.5.1 “结果总结”表简介 331
11.5.2 “结果总结”表的结局指标筛选 332
11.5.3 “结果总结”表的通用模板 332
11.5.4 制作“结果总结”表 335
11.5.5 “结果总结”表中的统计学因素 335
11.5.6 “结果总结”表的详细内容 336
11.6 附加表格 339
11.7 在正文中呈现结果 339
11.7.1 Meta 分析结果 339
11.7.2 无 Meta 分析的结果 339
11.8 撰写摘要 340
11.9 撰写通俗语言摘要 343
11.9.1 关于通俗语言摘要 343
11.9.2 通俗语言摘要的标题 343
11.9.3 摘要正文 343
11.10 本章信息 344
11.11 参考文献 345

第十二章　解释结果和得出结论 346
12.1　引言 347
12.2　评价一组证据的质量 348
12.2.1　GRADE 法 348
12.2.2　降低一组证据质量的因素 349
12.2.3　增加证据质量水平的因素 353
12.3　适用性问题 354
12.3.1　系统评价作者的角色 354
12.3.2　生物学差异 355
12.3.3　文化背景差异 355
12.3.4　依从性差异 356
12.3.5　价值和偏好的差异 356
12.4　解释统计分析的结果 356
12.4.1　可信区间 356
12.4.2　P 值和统计学意义 358
12.5　二分类结局结果解释（包括 NNT） 359
12.5.1　相对和绝对风险降低 359
12.5.2 关于 NNT 的更多内容 359
12.5.3　绝对风险降低的表达 360
12.5.4　计算 361
12.6　通过连续性结局解释结果（包括标准化均数差） 363
12.6.1　连续性结局的 Meta 分析 363
12.6.2　使用效应量经验法则对 SMD 进行表述 364
12.6.3　通过转化为 OR 对 SMD 进行表述 364
12.6.4　使用常用工具对 SMD 进行表述 365
12.7　得出结论 365
12.7.1　Cochrane 系统评价的结论部分 365
12.7.2　对实践的意义 365
12.7.3　对研究的意义 366
12.7.4　得出结论中的常见错误 367

12.8 本章信息 368
12.9 参考文献 368
第十三章 纳入非随机研究 375
13.1 引言 376
13.1.1 本章主要内容 376
13.1.2 为什么考虑纳入非随机研究？ 377
13.1.3 将非随机研究纳入 Cochrane 系统评价存在的关键问题 379
13.1.4 计划书对于纳入了 NRS 的 Cochrane 系统评价的重要性 379
13.1.5 本章的章节内容安排 380
13.2 制定纳入非随机研究的标准 380
13.2.1 纳入非随机研究有何不同？ 380
13.2.2 用于支持系统评价作者的指导意见和可用资源 383
13.2.3 总结 385
13.3 非随机试验检索 388
13.3.1 纳入非随机试验有什么不同 388
13.3.2 支持系统评价作者的指南和可利用资源 390
13.3.3 总结 391
13.4 选择研究和收集数据 391
13.4.1 当包括非随机研究时，检索策略会有何不同？ 391
13.4.2 支持系统评价作者的指南和可获取的资源 392
13.4.3 总结 394
13.5 非随机研究的偏倚风险评估 395
13.5.1 纳入非随机研究时会有什么不同？ 395
13.5.2 支持系统评价作者的指南和可用的资源 397
13.5.3 总结 401
13.6 从非随机研究中整合数据 401
13.6.1 当纳入非随机研究时会有什么不同？ 401
13.6.2 支持系统评价作者的指南和可用资源 402
13.6.3 总结 406
13.7 解释与讨论 406

13.7.1 解释纳入非随机研究的有效性的 Cochrane 系统评价结果时所面临的挑战 406
13.7.2 评估纳入非随机试验的系统评价的证据强度 408
13.7.3 对潜在系统评价作者的指导 409
13.8 本章信息 409
13.9 参考文献 410
第十四章 不良反应 416
14.1 引言 417
14.1.1 研究不良反应的必要性 417
14.1.2 概念和术语 417
14.1.3 何时需考虑干预措施的不良反应 418
14.2 系统评价中不良反应的研究范围 419
14.2.1 用同种方法研究有利结果和不良反应 419
14.2.2 用不同方法研究有利结果和不良反应 419
14.2.3 针对不良反应的独立系统评价 420
14.3 选择纳入哪些不良反应 420
14.3.1 狭小关注与广泛关注 420
14.3.2 退出或脱落作为不良反应的结局测量指标 421
14.4 研究类型 421
14.5 不良反应的检索方法 422
14.5.1 药物不良反应的信息源 422
14.5.2 不良反应检索策略 423
14.6 不良反应偏倚风险评估 425
14.6.1 临床试验 425
14.6.2 病例对照和队列研究 426
14.6.3 病案报告 426
14.7 本章信息 428
14.8 参考文献 429
第十五章 整合经济学证据 433
15.1 经济学证据在 Cochrane 评价中的作用和联系 434

15.1.1 引言 434
15.1.2 经济学和经济学评价 435
15.1.3 Cochrane 评价中涵盖的经济学问题 436
15.2 计划 Cochrane 评价中经济学构成要素 437
15.2.1 构建经济学问题 437
15.2.2 纳入资源利用、成本和成本-效果指标作为结局指标 440
15.2.3 卫生经济学研究的具体类型和系统评价中经济学内容的范围 441
15.3 查找研究 442
15.3.1 电子检索过滤器的使用 442
15.3.2 专题数据库的使用 444
15.4 筛选研究和收集数据 445
15.4.1 评价与研究主题的相关性 445
15.4.2 数据收集 445
15.5 偏倚风险的处理 446
15.5.1 按研究设计进行研究分类 446
15.5.2 方法学质量的严格评价 447
15.6 结果分析和描述 451
15.6.1 以表格形式描述结果 451
15.6.2 结果的描述性总结 451
15.6.3 资源利用和成本数据的 Meta 分析 453
15.6.4 建立经济学模型 454
15.7 解决报告偏倚 455
15.8 结果的解释 456
15.9 结论 457
15.10 本章信息 457
15.11 参考文献 458
第十六章 统计学中的特殊问题 464
16.1 缺失数据 465
16.1.1 缺失数据的类型 465
16.1.2 数据缺失的一般处理原则 467

16.1.3 标准差缺失 468
16.2 意向性分析相关问题 472
16.2.1 引言 472
16.2.2 二分类数据的意向性处理 473
16.2.3 连续性数据的意向性处理 475
16.2.4 适用于部分受试者的结局观察指标 475
16.3 整群随机对照试验 476
16.3.1 引言 476
16.3.2 整群随机对照试验的偏倚风险评估 477
16.3.3 整群随机对照试验的分析方法 478
16.3.4 整群随机试验 Meta 分析的近似方法：有效样本含量 478
16.3.5 整合整群随机试验的实例 479
16.3.6 整群随机试验 Meta 分析的校正分析：标准误调整法 479
16.3.7 整合整群随机试验需注意的问题 480
16.3.8 个体随机试验中的整群抽样问题 480
16.4 交叉试验 480
16.4.1 引言 480
16.4.2 交叉试验的适用性评价 481
16.4.3 交叉试验的偏倚风险评估 482
16.4.4 交叉试验的分析方法 483
16.4.5 将交叉试验纳入 Meta 分析的方法 484
16.4.6 Meta 分析中交叉试验的近似分析 485
16.4.7 纳入交叉试验应注意的问题 488
16.5 多个干预组的研究 488
16.5.1 引言 488
16.5.2 确定哪些干预组与系统评价相关 489
16.5.3 评价多个干预组研究的偏倚风险 490
16.5.4 如何从一个研究中纳入多个干预组 490
16.5.5 多臂试验中的异质性考虑 492
16.5.6 析因试验 493

16.6 间接比较和多臂试验 Meta 分析 494
16.6.1 引言 494
16.6.2 间接比较 494
16.6.3 多臂试验 Meta 分析 495
16.7 多重比较及机遇的作用 496
16.7.1 引言 496
16.7.2 系统评价中的多重比较 497
16.8 Meta 分析中的贝叶斯和分层方法 498
16.8.1 贝叶斯方法 498
16.8.2 分层模型 500
16.9 罕见事件（包括 0 频数） 500
16.9.1 罕见事件的 Meta 分析 500
16.9.2 格子计数为零的研究 501
16.9.3 无事件发生的研究 501
16.9.4 无事件发生研究的可信区间 502
16.9.5 罕见事件 Meta 分析方法的有效性 503
16.10 本章信息 504
16.11 参考文献 504
第十七章 病人报告的结局 513
17.1 什么是病人报告的临床结局？ 514
17.2 病人报告的临床结局和 Cochrane 系统评价 515
17.3 作为病人报告临床结局的健康状况和生命质量 516
17.4 测量病人报告的临床结局中的问题 519
17.4.1 工具的有效性 519
17.4.2 一个工具测量变化的能力 520
17.5 定位并选择有病人报告的临床结局的研究 520
17.6 评估和描述病人报告的临床结局 521
17.7 病人报告的临床结局不同测量指标间的可比性 522
17.8 结果解释 524
17.8.1 关注单个病人报告的临床结局的研究总结 524

17.8.2 运用不止一个病人报告的临床结局进行研究总结 525
17.8.3 当研究并没有涉及病人报告的临床结局 525
17.9 本章信息 526
17.10 参考文献 526
第十八章 个体病人数据的系统评价 530
18.1 引言 531
18.1.1 什么是 IPD 系统评价? 531
18.1.2 何时需制作 IPD 系统评价？ 531
18.1.3 IPD 系统评价方法有何不同？ 532
18.1.4 如何组织一项 IPD 系统评价？ 532
18.1.5 那些卫生保健领域使用过 IPD 分析方法？ 532
18.1.6 制作 IPD 系统评价的首要步骤 533
18.2 IPD Meta 分析的协作性质 533
18.2.1 协作组 533
18.2.2 协商合作 533
18.2.3 保密 533
18.3 数据处理 534
18.3.1 确定需要收集的资料 534
18.3.2 数据格式 534
18.3.3 变量的重新编码与定义 534
18.3.4 核对数据 535
18.4 数据分析 536
18.4.1 数据分析优势 536
18.4.2 一般方法 536
18.4.3 时间-事件分析 536
18.4.4 长期随访结果分析的更新 537
18.4.5 亚组分析 537
18.4.6 其他分析 537
18.4.7 软件支持 537
18.5 局限性与注意事项 538

18.5.1 IPD 系统评价不能解决的问题 538
18.5.2 不能获得的研究 538
18.5.3 何时制作 IPD 系统评价 539
18.6 本章信息 539
18.7 参考文献 540
第十九章 前瞻性 Meta 分析 542
19.1 引言 543
19.1.1 什么是前瞻性 Meta 分析？ 543
19.1.2 前瞻性 Meta 分析和大型多中心试验的区别是什么？ 544
19.1.3 哪些卫生保健领域使用过前瞻性 Meta 分析方法？ 545
19.1.4 我们需要什么资源？ 545
19.2 前瞻性 Meta 分析的协作本质 545
19.2.1 协作小组 545
19.2.2 协商合作 546
19.2.3 保密 546
19.3 前瞻性 Meta 分析的计划书 547
19.3.1 计划书应包括的内容？ 547
19.3.2 计划书的发表 549
19.4 前瞻性 Meta 分析的数据收集 550
19.5 前瞻性 Meta 分析的问题 550
19.5.1 一般方法 550
19.5.2 中期分析和数据监测 550
19.6 本章信息 551
19.7 参考文献 552
第二十章 定性研究与 Cochrane 系统评价 555
20.1 引言 556
20.2 Cochrane 系统评价中整合定性研究证据：概念和问题 557
20.2.1 定性研究的定义 557
20.2.2 Cochrane 系统评价中定性研究证据的使用 557
20.2.3 考虑那些随机对照试验内部或与之并列的定性研究 559

20.2.4 关于资源 560
20.3 定性研究证据的综合 560
20.3.1 综合定性研究证据来补充 Cochrane 干预性系统评价的实例：直接督导疗法与结核病（TB） 560
20.3.2 方法学问题 562
20.4 本章信息 566
20.5 参考文献 568
20.6 定性研究方法 573
20.6.1 一般的定性研究 573
20.6.2 定性研究方法 573
20.6.3 定性研究文献检索 574
20.6.4 定性研究证据综合 574
20.6.5 定性定量证据合并 575
20.6.6 定性研究严格评价 576
20.6.7 相关网址（Accessed 1 January 2008） 577
第二十一章 公共卫生与健康促进系统评价 578
21.1 引言 579
21.2 纳入的研究类型 579
21.3 检索 580
21.4 研究质量和偏倚风险评估 582
21.5 伦理和不平等问题 582
21.6 背景 584
21.7 可持续性 585
21.8 适用性和可转移性 587
21.9 本章信息 588
21.10 参考文献 589
第二十二章 系统评价再评价 594
22.1 引言 595
22.1.1 系统评价再评价的定义 595
22.1.2 Cochrane 再评价的原理 595

22.2 制作 Cochrane 再评价 595
22.2.1 组织事宜 595
22.2.2 方法学 599
22.2.3 更新 Cochrane 再评价 600
22.3 Cochrane 再评价的格式 600
22.3.1 标题和再评价的内容（或计划书内容） 600
22.3.2 摘要 601
22.3.3 通俗语言摘要（plain language summary） 602
22.3.4 Cochrane 再评价正文 602
22.3.5 系统评价和参考文献 608
22.3.6 表格 608
22.3.7 图 613
22.4 本章信息 614
22.5 参考文献 614
附录 A Cochrane 方法学研究方案和系统评价 615
内容的指南 615
A.1 简介 615
A.2 标题和系统评价信息（或方案信息） 615
A.2.1 标题 615
A.2.2 作者 615
A.2.3 联系人 616
A.2.4 日期 616
A.2.5 新内容和旧内容 617
A.3 摘要 618
A.4 简明的言语总结 618
A.5 文章正文 618
A.6 表格 631
A.6.1 纳入研究的特征 631
A.6.2 偏倚风险 632
A.6.3 排除研究的特征 632

A.6.4 待分级研究的特征 632
A.6.5 进行中研究的特征 633
A.6.6 结果总结表 633
A.6.7 附加的表格 633
A.7 研究和参考文献 633
A.7.1 研究的参考文献 633
A.7.2 其他参考文献 634
A.8 数据和分析 635
A.9 图形 636
A.9.1 RevMan 图形和表格 636
A.9.2 其他图形 636
A.10 系统评价的资助来源 637
A.11 反馈 637
A.12 附件 637
A.13 附录信息 637
A.14 参考文献 638

Cochrane 干预措施系统评价手册

本节选自指导手册5.0.1版本。这些材料还刊登于Higgins JPT和Green S编辑的《Cochrane干预措施系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

引言

《Cochrane干预措施系统评价手册》简称《手册》旨在为准备制作Cochrane干预措施系统评价(包括再评价)的作者提供指导。

不断更新

本《手册》定期更新是为了反映系统评价方法学前言以及回应用户的反馈信息。请通过以下网址参考最新的版本，如《手册》的临时更新和先前版本的详细信息。

www.cochrane.org/resources/handbook、

我们鼓励<手册>的使用者向本《手册》的编辑进行反馈和更正。网页上有详细联系信息。

主要支持来源

目前支持来源

The Cochrane Collaboration

Medical Research Council, United Kingdom

Department of Health and Ageing, Australia

Monash University, Australia

以前支持来源

National Health Service Research and Development Programme, United Kingdom

Health Research Board, Ireland

National Institute of Public Health, Norway

Copenhagen Hospital Corporation, Denmark

Health Services Research and Development Service and the University of Texas Health Science Center, San Antonio, USA

US Veterans Health Administration, USA

Oxford Regional Health Authority, UK

Nuffield Provincial Hospitals Trust, UK

LW Frohlich Fund, USA

Norwegian Ministry of Health and Social Affairs, Norway

Norwegian Research Council, Norway

Glaxo Wellcome, Norway

致谢

感谢《手册》顾问组以前和现在成员的讨论和反馈，特别感谢 Doug Altman, Chris Cates, Mike Clarke, Jon Deeks, Donna Gillies, Andrew Herxheimer, Harriet MacLehose, Philippa Middleton, Ruth Mitchell, David Moher, Donald Patrick, Ian Shemilt, Lesley Stewart, Jessica Thomas, Jane Tierney and Danielle Wheeler.

许多人为同行评审做出了建设性和及时的贡献。我们要感谢 Phil Alderson, Claire Allen, Judith Anzures, Chris Cates, Jonathan Craig, Miranda Cumpston, Chris Del Mar, Kay Dickersin, Christian Gluud, Peter Gøtzsche, Frans Helmerhorst, Jini Hetherington, Sophie Hill, Sally Hopewell, Steve McDonald, David Moher, Ann Møller, Duncan Mortimer, Karen New, Denise O'Connor, Jordi Pardo, Rob Scholten, Simon Thompson, Jan Vandenbroucke, Janet Wale, Phil Wiffen, Hywel Williams, Paula Williamson, Jim Wright and Diana Wyatt.

Jane Lane 为本版《手册》提供了特别的管理支持。此外，Claire Allen, Dave Booker, Jini Hetherington,Monica Kjeldstrøm, Cindy Manukonga, Rasmus Moustgaard, Jane Predl, Jacob Riis 为我们提供了熟练和慷慨的管理和技术支持。同时，Verena Roloff 为本《手册》的准备和协调工作做出了巨大的贡献。我们也要感谢 Wiley-Blackwell 的 Lucy Sayer, Fiona Woods and Laura Mellor 的耐心，支持和建议。

如果没有英国剑桥大学公共卫生学院 MRC 统计系（MRC Biostatistics Unit and the Institute of Public Health in Cambridge, UK ）及澳大利亚墨尔本 Monash 大学，澳大利亚 Cochrane 中心（ the Australasian Cochrane Centre, Monash University, Australia ）同事对编辑的慷慨支持，本《手册》也不可能完成。

《手册》编辑

Julian Higgin：英国剑桥大学公共卫生学院 MRC 统计系资深统计学家及英国牛津的英国 Cochrane 中心访问学者。

Sally Green：澳大利亚墨尔本 Monash 大学,卫生服务及研究院学者及澳大利亚 Cochrane 中心主任。

主要贡献者

Acquadro, Catherine
MAPI Research Institute
Lyon
France

Alderson, Philip
National Institute for Health and Clinical
Excellence
London/Manchester
United Kingdom

Altman, Douglas G
Centre for Statistics in Medicine
University of Oxford
Oxford
United Kingdom

Altman, Douglas G
Centre for Statistics in Medicine
University of Oxford
Oxford
United Kingdom

Armstrong, Rebecca
The McCaughey Centre: VicHealth Centre for
the Promotion of Mental Health and Community Wellbeing
University of Melbourne
Melbourne
Australia

Askie, Lisa M
NHMRC Clinical Trials Centre
University of Sydney
Camperdown
Australia

Becker, Lorne A
Department of Family Medicine
SUNY Upstate Medical University
Syracuse, NY
United States of America

Becker, Lorne A
Department of Family Medicine
SUNY Upstate Medical University
Syracuse, NY
United States of America

Booth, Andrew
School of Health and Related Research
University of Sheffield
Sheffield
United Kingdom

Byford, Sarah
Centre for the Economics of Mental Health
Institute of Psychiatry
King's College London
London
United Kingdom

Clarke, Mike
UK Cochrane Centre
National Institute for Health Research
Oxford
United Kingdom

Deeks, Jonathan J
Department of Public Health and Epidemiology
University of Birmingham
Birmingham
United Kingdom

Doyle, Jodie
The McCaughey Centre: VicHealth Centre for
the Promotion of Mental Health and Community Wellbeing
University of Melbourne
Melbourne
Australia

Drummond, Michael
Centre for Health Economics
University of York
York
United Kingdom

Egger, Matthias
Institute of Social and Preventive Medicine
University of Bern Switzerland

Eisenstein, Eric
Duke Clinical Research Center
Duke University
Durham, NC
United States of America

Ghersi, Davina
Department of Research Policy and Cooperation
World Health Organization
Geneva
Switzerland

Glanville, Julie
Centre for Reviews and Dissemination
University of York
York
United Kingdom

Glasziou, Paul P
Department of Primary Health Care
University of Oxford
Oxford
United Kingdom

Green, Sally
Australasian Cochrane Centre
Monash University
Melbourne
Australia

Guyatt, Gordon H
Departments of Clinical Epidemiology and
Biostatics
McMaster University
Ontario
Canada

Hannes, Karin
Belgian Centre for Evidence-Based Medicine
Leuven
Belgium

Hannes, Karin
Belgian Centre for Evidence-Based Medicine
Leuven
Belgium

Higgins, Julian PT
MRC Biostatistics Unit
Cambridge
United Kingdom

Knapp, Martin
Institute of Psychiatry
King's College London
and
London School of Economics
London
United Kingdom

Lefebvre, Carol
UK Cochrane Centre
National Institute for Health Research
Oxford
United Kingdom

Loke, Yoon K
School of Medicine, Health Policy and Practice
University of East Anglia
Norwich
United Kingdom

Mallender, Jacqueline
Matrix Knowledge Group Ltd.
London
United Kingdom

Manheimer, Eric
Center for Integrative Medicine
University of Maryland School of Medicine
Baltimore, MA
United States of America

McDaid, David
Personal Social Services Research Unit
London School of Economics and Political
Science
London
United Kingdom

Moher, David
Chalmers Research Group, Children's Hospital
of Eastern Ontario Research Institute;
Department of Epidemiology and Community
Medicine, University of Ottawa
Ottawa
Canada

Mugford, Miranda
Health Economics Group
School of Medicine, Health Policy and Practice
University of East Anglia
Norwich
United Kingdom

Noyes, Jane
Centre for Health-Related Research
School of Healthcare Sciences
Bangor University
Bangor
Wales
United Kingdom

O'Connor, Denise
Australasian Cochrane Centre
Monash University
Melbourne
Australia

Oxman, Andrew D
Preventive and International Health Care Unit
Norwegian Knowledge Centre for the Health
Services
Oslo
Norway

Patrick, Donald L
Department of Health Services and Seattle
Quality of Life Group
University of Washington
Seattle, WA
United States of America

Pearson, Alan
Joanna Briggs Institute
University of Adelaide
Adelaide
Australia

Popay, Jennie
Institute for Health Research
Lancaster University
Lancaster
United Kingdom

Price, Deirdre
Department of Clinical Pharmacology
University of Oxford
Oxford
United Kingdom

Reeves, Barnaby
Bristol Heart Institute
University of Bristol
Bristol
United Kingdom

Scholten, Rob
Dutch Cochrane Centre
Academic Medical Center
Amsterdam
The Netherlands

Schünemann, Holger J
INFORMA/CLARITY
Research/Department
of Epidemiology
National Cancer Institute Regina Elena
Rome
Italy

Shemilt, Ian
Health Economics Group
School of Medicine, Health Policy and Practice
University of East Anglia
Norwich
United Kingdom

Sterne, Jonathan AC
Department of Social Medicine
University of Bristol
Bristol
United Kingdom

Stewart, Lesley A
Centre for Reviews and Dissemination
University of York
York
United Kingdom

Tierney, Jayne F
MRC Clinical Trials Unit
London
United Kingdom

Vale, Luke
Health Economics Research Unit
University of Aberdeen
Aberdeen
United Kingdom

Vist, Gunn E
Preventive and International Health Care Unit
Norwegian Knowledge Centre for the Health Services
Oslo
Norway

Walker, Damian
Health Systems Program
Department of International Health
Johns Hopkins Bloomberg School of Public Health
Baltimore, MA
United States of America

Waters, Elizabeth
The McCaughey Centre: VicHealth Centre for the Promotion of Mental Health and Community Wellbeing University of Melbourne
Melbourne
Australia

Wells, George A

Department of Epidemiology and Community

Medicine University of Ottawa

Ottawa

Ontario

Canada

第一章　导论

作者：Sally Green, Julian PT Higgins, Philip Alderson, Mike Clarke, Cynthia D Mulrow, Andrew D Oxman。

本文选自工作手册 5.0.1版本。有关如何引用它的指南，见1.5节。这些材料还刊登于Higgins JPT和Green S编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 系统评价是寻求收集及整理所有符合预先规定的纳入标准的证据，以解决特定的研究问题。
- 系统评价旨在使用明确、系统的方法降低偏倚。
- Cochrane协作网制作、维护和传播Cochrane系统评价，为医疗卫生保健决策提供依据。
- Cochrane评价发表于Cochrane图书馆的Cochrane系统评价数据库。
- 关于干预措施的Cochrane系统评价手册包含了制作和维护干预措施Cochrane评价及Cochrane评价再评价的方法学指南。

1.1 Cochrane协作网

1.1.1 引言

Cochrane协作网（www.cochrane.org）是一个国际性组织，旨在通过制作、维护和提高决策所需的系统评价证据可及性，帮助人们制定遵循证据的（知证）卫生决策。为可靠合成针对某特定问题的当前所有证据，系统评价考虑到某一干预措施效果的所有证据，始终坚持科学可积累且可促进决策的原则。自1993年成立以来，Cochrane协作网已拥有来自100多个国家共15000余名成员，成为该领域最大的组织（Allen 2006，Allen 2007）。由Iain Chalmers爵士及其同事共同创建、以英国流行病学家Archie Cochrane命名的牛津大学Cochrane中心（现为英国Cochrane中心）成立一年后，成立国际Cochrane协作网。Cochrane协作网如今已成为一个国际知名组织（Clarke 2005，Green 2005）。

Cochrane协作网的工作以表1.1.a中列举的10个主要原则为支撑。

表1.1.a Cochrane协作网的原则

1. 相互合作：内、外部良好沟通，公开决策及团队合作。
2. 热心奉献：吸收并支持不同技能及背景的热情人士。
3. 避免重复：通过优良的管理及协调工作，避免重复使工作效益最大化。
4. 减少偏倚：采用严谨的科学分析、确保广泛参与及避免利益冲突等多种途径，使偏倚最小化。
5. 及时更新：承诺通过检索和纳入新的证据，确保Cochrane评价的更新。
6. 力求相关：采用与人们制订卫生保健决策息息相关的结局，促进对医疗保健干预措施的评估。
7. 推动实践：广泛传播协作网的产出成果，利用战略联盟优势及提供合理的价格、内容和媒介以满足全世界用户需求，提高可及性。
8. 确保质量：公开回应批评，运用先进的方法学并制订质量完善系统。
9. 持续发展：维护和更新评价、编辑过程及其它主要职能的责任。
10. 广泛参与：减少合作障碍，鼓励多样性。

1.1.2 Cochrane协作网组织结构

Cochrane协作网以52个Cochrane系统评价小组（Cochrane review groups，CRGs）为核心开展工作，负责特定卫生保健领域内系统评价的生产和维护。CRGs小组的成员包括研究人员、医护人员和使用医疗服务的人群（消费者或用户），所有人都抱有同样的热情，即生产有关预防和治疗某一健康问题或某一类健康问题的可靠、最新的证据。

Cochrane评价小组在评价的制作过程中有方法学小组、Cochrane中心及相关领域的支持。Cochrane方法学小组为方法学家提供了一个论坛，讨论制作Cochrane系统评价所采用的方法的进展、评估及应用方面问题。该方法学组对制定干预措施的Cochrane系统评价手册（手册）发挥了重要作用，有关章节包含了相关方法学小组的信息。Cochrane中心设在不同国家，除宣传和推广Cochrane评价外，还代表所在区域，为评价员和Cochrane小组提供培训和支持。Cochrane领域对卫生保健的关注层面较广，比如保健方面（如初级保健）、消费者类型（如儿童）或干预类型（如疫苗）。各领域的相关人员有助于确保Cochrane评价小组在他们感兴趣的相关领域更好的选择研究课题。

1.1.3 Cochrane评价的发表

Cochrane评价全文在Cochrane系统评价数据库（Cochrane Database of Systematic Reviews，CDSR）在线发表，后者是Cochrane图书馆的核心组成部分。Cochrane图书馆由Wiley-Blackwell在网上（www.thecochranelibrary.com）及用光盘发行，一些获取全国许可的国家可免费使用，而在资源最贫乏的地方，Wiley-Blackwell提供免费使用。其它地方则采取订购，或按次收费的方式。除了Cochrane系统评价数据库，Cochrane图书馆还包含其它几个知识库，见表1.1.b.

Cochrane系统评价数据库每年出版四次，每次都有新的评价和更新后的评价。2008年第1期的Cochrane系统评价数据库载有3000余篇Cochrane评价和1700余篇评价计划书。

表1.1.b　Cochrane图书馆包含的数据库

- Cochrane系统评价数据库（Cochrane Database of Systematic Reviews，CDSR），提供Cochrane评价全文（包括方法，结果和结论）以及研究方案。
- 效果评价文摘数据库（The Database of Abstracts of Reviews of Effects，DARE），由英国约克的评审传播中心整理和维护，提供严谨的评价和符合明确质量标准的其它系统评价的结构式摘要。
- Cochrane临床对照试验中心注册数据库（The Cochrane Central Register of Controlled Trials，CENTRAL），提供成百上千研究的引文信息，包括会议论文和目前其他文献数据库中未列出的其他来源的论文。
- Cochrane方法学注册资料数据库（The Cochrane Methodology Register，CMR），提供与系统评价研究领域相关的文章和书籍的文献信息及方法学研究的预先注册。
- Cochrane协作网提供Cochrane评价小组及Cochrane协作网中其他小组的联系方式及其他信息。

1.2　系统评价

1.2.1　系统评价的需求

医疗服务提供者、消费者、研究者和政策制定者淹没在难以处理的信息海洋中，包括来自卫生保健领域的研究证据。并非所有人都有时间、技术和资源去查找、评价和解释证据并将其用于卫生决策。Cochrane评价通过查找、评价和综合这些研究证据并以容易获得的格式发表，以应对这一挑战（Mulrow 1994）。

1.2.2　什么是系统评价

系统评价旨在收集所有符合预定纳入标准的研究证据并进行整理评价来回答某一具体的研究问题。其采用明确、系统的方法降低偏倚，提供更为可靠的结果，促进决策（Antman 1992，Oxman 1993）。系统评价的主要特征为：

- 目的明确、预设的文献纳入标准清晰；
- 方法明确、且可重复；
- 系统检索所有符合纳入标准的研究文献；

- 评估纳入研究结果的真实性，如评估偏倚风险；
- 系统描述及整合纳入研究的特点和结果。

许多系统评价包含Meta分析。Meta分析是采用统计学方法总结独立研究的结果（Glass 1976）。与单个研究的评价相比，Meta分析通过整合所有相关研究，可更精准的估计卫生保健的效果（见第9章，9.1.3节），并有利于探索各研究证据的一致性及研究间的差异性。

1.3 关于本手册

合成研究的科学正在迅速发展，Cochrane评价的实施方法也随着时间推移得到了已有的发展。干预措施的Cochrane系统评价手册（手册）旨在帮助Cochrane评价员选择合适的方法，而非提出强制性标准。所推荐的建议都尽可能来自经验证据。这里提供的指南旨在帮助评价员系统、知证和明确地（但不是机械的）了解他们所提出的问题及解决问题的过程。本指南的释义及实施需与Cochrane评价小组编辑共同完成。

本手册重点是干预措施效果的系统评价。大部分建议面向临床试验的综合，特别是随机试验，因为相对于其它关于卫生保健干预相对效果的研究设计，随机试验提供了更可靠的证据（Kunz 2007）。但有些章节也提供了纳入其它类型研究证据的建议，特别是考虑到安全性或副作用影响，随机试验可能无法进行或不恰当。2003年，Cochrane协作网扩大了范围，包括Cochrane诊断试验准确性的系统评价。这些评价的实施指南包含在一个单独的文件：诊断试验准确性的Cochrane系统评价手册。

本手册分为3部分，共22章。第1部分介绍了Cochrane评价，包括计划、制作、维护、更新，结尾是Cochrane评价或计划书的撰写指南。第2部分提供了所有Cochrane评价的一般性方法学指导，包括问题的提出、纳入标准、检索、收集数据、研究偏倚、分析数据、报告偏倚、描述和解释结果。第3部分涉及到若干（并非全部）Cochrane评价相关的特定主题，包括不良反应的处理、非标准研究设计的Meta分析和使用单个病人数据的Meta分析。这部分纳入了经济学评价、非随机研究、定性研究、采用病人报告结局的评价、前瞻性Meta分析和健康促进和公共卫生评价等章节。最后一章介绍了新的评价类型，即系统评价再评价。

每章都列出了重点，为评价员总结和提取主要信息。

该手册主要由Cochrane协作网的方法学组编写，其成员进行了大量的方法学及实证研究来指导指南的编写。

虽然该手册的主要阅读对象是Cochrane干预措施评价的作者，但许多原则和方法同样适用于其它类型研究的系统评价及由其它评价员进行的干预类系统评价（Moher 2007）。

1.4 手册参编者

“如果我看得更远，是因为站在了巨人的肩膀上”——艾萨克 牛顿（Isaac Newton）

该干预措施的Cochrane系统评价手册（第5版）是Cochrane协作网成立初期以来逐步形成的重大修订版本。第5版手册的许多章节是建立在以前版本的基础之上，其它章节则是新的创作。这体现了真正的协作，反映了Cochrane协作网的原则。很多人对本手册都有着直接的贡献，如章节的作者、编辑、审稿，Cochrane手册咨询组成员，及其他方式参与进来的人员。该手册也反映了以下人员曾给予的宝贵贡献：以前版本的编辑，过去和现在的Cochrane方法学组成员、评价员、Cochrane评价小组、RevMan软件咨询小组、Cochrane中心和Cochrane领域。

最初为Cochrane评价员制作方法学指南的有Andy Oxman、Iain Chalmers、Mike Clarke、Murray Enkin、Ken Schulz、Mark Starr、Kay Dickersin、Andrew Herxheimer和Chris Silagy，其中Sally Hunt提供行政工作支持。该指南作为协作网综合指南的“第六部分：制作和维护系统评价（Cochrane协作网工具包）”于1994年3月出版。它描述了Cochrane评价最初的结构形式，该Cochrane评价结构由Mike Clarke、Murray Enkin、Chris Silagy和Mark Starr开发。

该指南在新成立的手册顾问小组支持下，于1996年10月成为了一个独立文件，即以Andy Oxman and Cynthia Mulrow作为编者的Cochrane协作网手册（第3版）。名为Cochrane评价员手册的第4版手册于1999年配合新推出的RevMan4软件发行，该版本由主编Mike Clarke和Andy Oxman，编者Phil Alderson, Julian Higgins 和 Sally Green，从1999年始至2003年12月完成。Cochrane诊断试验准确性评价的引入及对指导这类评价的新手册需求促成了2005年3月发行的4.2.4版本，改名为Cochrane干预措施系统评价手册，由Julian Higgins和Sally Green编写。

本手册的编者得到了手册顾问小组的建议支持。目前手册顾问小组成员有：Lisa Askie、Chris Cates、Jon Deeks、Matthias Egger、Davina Ghersi、Donna Gillies、Paul Glasziou、Sally Green（会议召集人）、Andrew Herxheimer、Julian Higgins（会议召集人）、Jane Lane（行政人员）、Carol Lefebvre、Harriet MacLehose、Philippa Middleton、Ruth Mitchell、David Moher、Miranda Mugford、Jane Noyes、Donald Patrick、Jennie Popay、Barney Reeves、Jacob Riis、Ian Shemilt、Jonathan Sterne、Lesley Stewart、Jessica Thomas、Jayne Tierney、Danielle Wheeler.

除了上述的编者外，以下人员也为此前版本的手册做出了大量的贡献：Christina Aguilar、Doug Altman、Bob Badgett、Hilda Bastian、Lisa Bero、Michael Brand、Joe Cavellero、Mildred Cho、Kay Dickersin、Lelia Duley、Frances Fairman、Jeremy Grimshaw、Gord Guyatt、Peter Gøtzsche、Jeph Herrin、Nicki Jackson、Monica Kjeldstrøm、Jos Kleijnen、Kristen Larson、Valerie Lawrence、Eric Manheimer、Rasmus Moustgaard、Melissa Ober、Drummond Rennie、Dave Sackett、Mark Starr、Nicola Thornton、Luke Vale 和Veronica Yank.

1.5 本章信息

作者：Sally Green, Julian PT Higgins, Philip Alderson, Mike Clarke, Cynthia D Mulrow and Andrew D Oxman.

本章引用格式：Green S, Higgins JPT, Alderson P, Clarke M, Mulrow CD, Oxman AD. Chapter 1: Introduction. In: Higgins JPT, Green S (editors), Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.

1.6 参考文献

Allen 2006

Allen C, Clarke M. International activity in Cochrane Review Groups with particular reference to China. Chinese Journal of Evidence-based Medicine 2006; 6: 541-545.

Allen 2007

Allen C, Clarke M, Tharyan P. International activity in Cochrane Review Groups with particular reference to India. National Medical Journal of India 2007; 20: 250-255.

Antman 1992

Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of Meta-analyses of randomized control trials and recommendations of clinical experts: Treatments for myocardial infarction. JAMA 1992; 268: 240-248.

Clarke 2005

Clarke M. Cochrane Collaboration. In: Armitage P, Colton T (editors). Encyclopedia of Biostatistics (2nd edition). Chichester (UK): John Wiley & Sons, 2005.

Glass 1976

Glass GV. Primary, secondary and Meta-analysis of research. Educational Researcher 1976; 5: 3-8.

Green 2005

Green S, McDonald S. The Cochrane Collaboration: More than systematic reviews? Internal Medicine Journal 2005; 35: 4-5.

Kunz 2007

Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials. Cochrane Database of Systematic Reviews 2007, Issue 2. Art No: MR000012.

Moher 2007

Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Medicine 2007; 4: e78.

Mulrow 1994

Mulrow CD. Rationale for systematic reviews. BMJ 1994; 309: 597-599.

Oxman 1993

Oxman AD, Guyatt GH. The science of reviewing research. Annals of the New York Academy of Sciences 1993; 703: 125-133.

Mulrow 1994

Mulrow CD. Rationale for systematic reviews. BMJ 1994; 309: 597-599.

Oxman 1993

Oxman AD, Guyatt GH. The science of reviewing research. Annals of the New York Academy of Sciences 1993; 703: 125-133.

（蒋兰慧译，岑啸、张龙浩初审）

第二章　系统评价的准备

本文选自工作手册5.0.1版本。有关如何引用它的指南，见2.7节。这些材料还刊登于Higgins JPT和Green S编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley ＆ Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 发表 Cochrane 系统评价前，应先发表系统评价计划书，减少作者偏倚的影响，增加方法和制作过程的透明性，避免重复，允许同行评审其方法。
- Cochrane 系统评价及其计划书可在 Cochrane 协作网的 RevMan 软件中完成，并有统一的格式。
- Cochrane 系统评价的框架在本章中讨论
- Cochrane 系统评价小组负责管理系统评价及计划书的编辑和出版，干预措施系统评价的题目必须经 Cochrane 系统评价小组同意并注册。
- Cochrane 系统评价应由多人组成的小组完成。

- 在其他杂志上发表 Cochrane 系统评价的指南。
- Cochrane 协作网有避免潜在利益冲突的行为准则

2.1 计划书的原则

创作 Cochrane 系统评价是一个复杂的过程，需从多方面判断。在制作系统评价的过程中为减少偏倚，应不依赖纳入研究的结果做出判断。系统评价员事先了解潜在纳入研究的结果可能会影响以下几方面：系统评价问题的界定、纳入研究的标准、需分析的对照干预措施的选择或者系统评价应报告结局指标的确定等等。Cochrane 系统评价是回顾性的评价（除前瞻性 Meta-分析外，在第 19 章介绍），所以应该提前确定并记录所采用的方法。在了解可能纳入的研究之前发表系统评价的计划书，可减少系统评价员的主观偏倚，促进系统评价方法的透明性，减小重复的可能性，允许同行评审其方法（Light 1984）。

系统评价应该与计划书保持一致，但有时也需做必要的改变。这与随机对照试验的计划书有些相似，随机对照试验计划书有时需要改变以应对未预料到的问题，比如受试者招募、数据收集或者非期望的结局事件发生率。虽然与计划书应尽可能保持一致，但这也不总是可能的或者恰当的。然而，重要的是不能基于对研究结果的影响来改变计划书。当已知某些改变会影响研究结果（如排除已经纳入系统评价的研究）时，再去改变计划书，这可能会引起高度偏倚，应该避免这种事后决策的情况。

计划书要在完成系统评价之前发表在 Cochrane 系统评价数据库（CDSR）。计划书改变的地方应该在系统评价全文“计划书与系统评价的不同”部分说明，必要时还应使用敏感性分析（见第 9 章，第 9.7 节）说明这些改变对研究结果的影响。

2.2 Cochrane 系统评价的格式

2.2.1 Cochrane 系统评价格式的原则

所有干预措施的Cochrane系统评价都有相同的格式，采用这种统一格式的优点包括：

1. 帮助系统评价读者快速找到结果，评估真实性和实用性，并应用这些结果；
2. 指导系统评价作者以最小的努力简明扼要地报告结果；

3. 更容易发表和维护；
4. 产生衍生产品（如系统评价再评价，见第22章）和基于多个系统评价的经验性研究。

Cochrane系统评价的格式很灵活，足以适合不同类型的系统评价，包括单一比较、多重比较以及应用单个病人数据。RevMan软件提供了标准的标题和表格格式以指导系统评价作者制作系统评价，也使读者更容易找到他们感兴趣的内容。RevMan中的标题格式在2.2.2节和2.2.3节中介绍，其内容在第4章中详细阐述。

2.2.2 Cochrane系统评价计划书的框架

表2.2.a列举Cochrane系统评价完整的计划书的内容，并说明在CDSR上如何显示（可能与RevMan中的形式不一样）。如果计划书中带有“*”的必填部分没有内容，则不会被发表。

表2.2.a Cochrane系统评价计划书的大纲

题目*
计划书一般信息：
作者*
通讯作者*
时间
新内容
历史
计划书：
背景*
目的*
方法：
纳入标准：
研究类型*
受试者的类型*

干预措施的类型*

结局指标的类型*

检索策略*

数据收集与分析*

致谢

参考文献

其他参考文献

附加参考文献

该评价发表的其他版本

表格和图

附加表格

图

补充信息

附录

反馈：

题目

摘要

回复

贡献

本研究信息

作者贡献

声明利益冲突*

资源支持

内部资源

外部资源

发表备注

2.2.3 Cochrane系统评价的大纲

表2.2.b列举了完整的系统评价内容，展示了Cochrane系统评价如何在CDSR上呈现（与RevMan中的不一样）。如果计划书中带有“*” 的必填部分没有内容，则不会被发表。

表2.2.b Cochrane系统评价的大纲

题目

系统评价一般信息

作者*

通讯作者*

时间*

更新

历史

摘要

背景*

目的*

检索方法*

数据收集与分析*

结果*

结论*

通俗语言摘要

通俗语言摘要题目*

摘要内容*

系统评价

背景*

目的*

方法：

纳入标准

研究类型*

受试者类型*

干预措施类型*

结局指标类型*

检索策略*

数据收集和分析*

结果：

纳入研究特征描述*

纳入研究偏倚风险*

干预措施效果*

讨论*

作者的结论：

对临床实践的意义*

对研究的意义*

致谢

参考文献：

纳入研究参考文献

纳入研究

排除研究

待分类研究

在研研究

其他参考文献

附加参考文献

本系统评价发表的其他版本

图表

研究特征：

纳入研究特征（包括“偏倚风险”表）

排除研究特征

待评估研究特征

在研研究特征

结果概述

附加表格

图

附加信息

数据与分析

附件

反馈：

题目

摘要

回复

贡献

本研究信息

对本研究有贡献的作者

声明利益冲突

计划书和全文之间的不同

资助来源：

内部资助来源

外部资助来源

发表备注

2.3 制作系统评价的流程

2.3.1 制作系统评价的动机

完成一个系统评价的动机很多，比如制作系统评价解决证据不一致的问题，解决临床实践中尚不确定的问题，探索临床实践中的差异，明确目前临床实践的合适性，强调哪些需要进一步开展研究。Cochrane 系统评价的首要目的是总结并帮助人们理解证据，做出临床决策。此目的在决定是否需要制作 Cochrane 系统评价，如何形成系统评价需要回答的问题，如何根据研究问题确定纳入标准，如何制作计划书，如何表示系统评价的结果方面有重要的意义。

2.3.2 规划系统评价的主题和范围

计划系统评价和制作计划书的注意事项：

- 系统评价问题应针对人们在医疗决策中面临的实际问题；
- 系统评价应采用对卫生决策有意义的结局指标；
- 系统评价作者应同样关注不良反应和疗效；
- 系统评价选择的方法应最大限度地为决策提供当前最佳证据，并在计划书中详细描述以帮助读者充分理解计划步骤；
- 让人们知道对决策者可能非常重要的某个结局指标尚无可靠证据或缺乏证据，是十分重要的。应区分没有有效的证据和证据显示无效；
- 系统评价纳入有高度偏倚风险的文献是无益的，即使目前尚无更佳证据（参考第 8 章关于偏倚风险的内容）；
- 同样，关注一些不重要的结局指标也是无益的，仅因为这些就是研究人员在单个研究中选择测量的结局指标（参考第五章）；
- 尽可能保持国际视角也很重要。没有充分的理由，收集的证据不应该限制在某个国家或某种语言，背景中的信息如患病率和发病率应有全球观念，尽量使系统评价的结果放在更宽泛的情景中。

2.3.3 注册计划书

系统评价的第一步是与相应的 Cochrane 评价小组（Cochrane Review Group，CRG）

联系，获得他们对某个系统评价题目的批准。52 个系统评价小组在 CDSR 上描述了相应的立题范围。许多 CRG 根据系统评价的重要性确立了优选主题，并要求填写“题目注册表格”。一个题目经过 CRG 的编辑讨论后才可能被注册，如果注册成功就要求提交计划书。计划书完成后要送给 CRG 的编辑和工作人员进行同行评审。当计划书接受后（可能要经过反复几个过程）才可在 CDSR 上发表和传播。如果系统评价作者不承诺按时发表和及时更新，计划书将不会被发表。

Cochrane 协作网规定，如果计划书发表后两年全文没有完成，其计划书将会从 CDSR 撤销。一个计划书除被全文取代外，因任何原因被撤销，都应该在 CDSR 一期上发表撤销声明。之后，计划书被撤销的信息将记录在相关系统评价小组的模块（module）上。

2.3.4 系统评价工作组

2.3.4.1 工作组的重要性

Cochrane 系统评价必须由一人以上完成。这样可以确保筛选纳入研究或提取数据时是由两人以上独立完成，增加错误的识别率。如果同一个题目有多个工作组申请注册，CRG 则鼓励多个工作组合作完成。

系统评价工作组应该包括相关领域专家，以及系统评价的方法学家（包括统计学专家）。鼓励第一次做系统评价的作者与有经验的系统评价员合作并参加 Cochrane 协作网组织的培训（见 2.3.6 节）。Cochrane 协作网坚持用户参与的原则（协作网原则第十条为保证协作网工作的广泛参与，见第一章表 1.1），并鼓励系统评价作者在制作计划书或全文的过程中吸取系统评价用户的观点，包括用户、临床医生、其他领域的专家。某些题目涉及特别的地区或环境（如发展中国家的疟疾），Cochrane 协作网也鼓励该地区的人参与。

2.3.4.2 用户参与

Cochrane 协作网鼓励卫生保健用户参与系统评价的制作过程或编辑过程。用户参与有助于确保系统评价：

- 针对人们关心的重要的问题；
- 考虑有重要影响的结局指标；
- 决策者易于获得；

- 充分考虑人们不同的价值观和条件以及不同国家的卫生环境。

以不同方式让用户参与系统评价或者卫生研究，其效果如何还知之甚少（Nilsen 2006）。尽管如此，Cochrane 协作网还是坚持广泛参与的原则。这是基于我们的原则和逻辑性，因为卫生保健用户的观点及视角与卫生保健的提供者和研究者常常不一致（Bastian 1998）。

用户可以通过以下几种途径参与：

- 帮助 CRG 确定优先研究的领域；
- 参与完成系统评价；
- 制作系统评价时提供用户咨询；
- 同行评审计划书或系统评价。

如果在制作系统评价或计划书的过程中咨询了用户，就需要在致谢部分予以说明；若有实质性贡献，像其他贡献者一样正式列为系统评价的作者也是合适的（见第四章 4.2.2 节）。

2.3.4.3 顾问团

系统评价与终端用户更为相关，如果有相关领域专家，包括临床专业或方法学家的指导就可以保证系统评价的质量（Khan 2001，Thomas 2004，Rees 2004）。决策者及用户关注的重点可能与作者不一样，系统评价作者有必要将所述问题的重要性阐释清楚，包括干预措施、结局指标和受试者。成立一个有相关兴趣和技能的利益相关代表的顾问团可能有用，这对影响力大或者涉及复杂干预措施的系统评价更为重要。表 2.3.a 是一个顾问团作用的例子。

系统评价小组需要与顾问团协作解决系统评价的关键问题。加拿大公共卫生计划发现 6 个成员组成的顾问团基本涵盖了所有领域，且易于管理（Effective Public Health Practice Project 2007）。但是系统评价的范围越宽，需要的顾问团成员也就越多。

系统评价过程中有必要考虑资源匮乏国家的需求。为增强系统评价的适用性，系统评价作者应该咨询发展中国家的需求，找出应该优先考虑的系统评价题目（Richards 2004）。顾问团中也应包含弱势及边缘群体（Stell 2001），以确保系统评价结论可应用到社会各个阶层。

顾问团成员的工作应该具体到个人以满足每一项任务的需求，例如 Hanley 2000 或者 INVOLVE 网（www.invo.org.uk）所描述。顾问团成员可能涉及以下一种或多种任务：

- 提出或精炼系统评价涉及的干预措施、纳入的受试者、优先考虑的结局指标，或者亚组分析。
- 提供或建议重要的背景资料，从不同角度说明问题。
- 帮助解释系统评价的结果。
- 设计传播计划，并协助系统评价向相关群体的传播。

表2.3.a　计划过程中应用顾问团的优势举例

一项关于男同性恋HIV预防的系统评价（Rees 2004）咨询了临床医生、政府官员和研究人员，采用共识法（Consensus methods）解决系统评价的相关问题。顾问团成员来自研究机构、行政管理部门及HIV/AIDS感染者的慈善机构代表。在系统评价过程中，顾问团共开了3次会议。

顾问团提供了系统评价相关的背景信息：范围、概念、研究目的、研究问题、研究阶段和方法。讨论集中在系统评价相关的政策及政治背景；研究纳入标准（干预措施、结局指标、男性亚组）；传播策略；时间表。两轮投票决定优先分析的结局指标。开放讨论找出易感亚组人群。通过顾问团讨论提炼了干预措施的特征框架。

该系统评价遵循此指导建议，采纳经讨论后决定的干预措施、受试者和结局指标提炼了纳入标准，并进行Meta分析和亚组分析。随后的产品包括合成的证据直接与卫生资源不均衡有关。

2.3.5　Cochrane评价使用的软件

Cochrane 协作网采用 Cochrane 信息管理系统（Cochrane information management system，IMS）支持制作和编辑监察 Cochrane 系统评价。这个系统主要包括两个部分，系统评价写作软件 RevMan 和负责管理文件和联系的中心服务器 Archie。IMS 作为 Cochrane 协作网一个电子组织结构，有效的促进了工作在不同地区的 Cochrane 系统评价小组工作人员与作者之间的协作。

RevMan 是 Cochrane 系统评价作者的必备使用工具，用于系制作和维护 Cochrane 系统评价计划书或全文（格式见 2.2 节）。该软件经反复咨询用户与 Cochrane 方法学家后逐渐成熟，可支持 Cochrane 系统评价的标准和指南，并提供分析方法、在线帮助和错误检查等功能。

除支持 Cochrane 干预性系统评价，RevMan 软件也支持方法学系统评价、诊断试验准确性系统评价和系统评价再评价（见第 22 章）。

RevMan 软件对 Cochrane 系统评价作者和研究机构是免费的，商业机构如购买了许可权也可使用，但仅为注册了 Cochrane 系统评价的作者提供相关技术支持。

RevMan 用于制作和编辑系统评价，Archie 则用于储存系统评价的草稿和发表的版本。集中储存系统评价所有相关的版本，当系统评价更新时，该系统可以帮助获取最新发表的版本。通过 Archie，系统评价作者也可以获取之前的版本，并比较不同版本，发现哪些地方进行过修改。除此之外，系统评价作者可以维护他们的联系信息并获得合作者或编辑部的联系信息，Cochrane 系统评价作者通过联系 CRG 的编辑部后才能使用 Archie。

IMS 由北欧 Cochrane 中心开发和运行。在咨询小组的指导下由 Cochrane 信息管理小组监管其发展。更多关于 Cochrane 协作网软件的信息如最新版本和发展计划，可在 IMS 网上获取（网址：www.cc-ims.net）。

2.3.6 培训

确保 Cochrane 系统评价的相关工作人员具有足够的知识、技能和支持来出色地完成工作是很重要的。所以需要对系统评价作者、编辑、评论编辑、审稿人、Cochrane 系统评价小组协调员和检索人员、手工检索人员、培训者和用户进行培训。这里我们主要强调系统评价作者和编辑的培训需求，以便他们完成高质量的系统评价。

少数加入 CRG 的系统评价作者经过了培训，并有系统评价的经验，但多数系统评价作者则没有。除了 CRG 为系统评价作者提供了培训材料和相关支持，Cochrane 中心负责与方法学组一起制作基于 Cochrane 手册的培训材料，并为 CRG 小组成员组织培训。CRG 则负责确保系统评价作者经过完整的培训，并有方法学支持。为了反映协作网的发展需求、标准及指南，培训材料和培训机会也在不断地改进和更新。

许多国家是由 Cochrane 中心、方法学组、CRG 负责培训系统评价作者。培训的时间表公布在 Cochrane 协作网的网站上（www.cochrane.org/resources/training.htm）。网站上有多种培训资源，包括 Cochrane 协作网开放学习资料。Cochrane 中心的详情公布在 www.chochrane.org。

2.3.7 Cochrane系统评价小组的编辑过程

CRG 编辑小组负责决定 Cochrane 系统评价最终发表。发表之前需进行同行评审并进行修改，这可能要重复几次。

每个 CRG 编辑小组负责维护一个模块，包括系统评价小组的信息和编辑过程。一些未在手册中介绍，但 CRG 在采用的特定方法应在其模块中说明，包括：

- 评价计划书的方法；
- 系统评价纳入研究的标准入选标准；
- 建立和维护 CRG 使用的专业注册库的检索方法和详细检索策略，及向系统评价作者传送可能相关的研究报告引文信息或全文的方法；
- 评价员需要常规使用的附加检索方法；
- 筛选文献的标准方法，纳入研究评估表格的模板；
- 除“偏倚风险”表格之外评估纳入研究的标准和方法；
- 提取数据的标准方法和数据提取表格的模板；

CRG 采用的特殊方法作为评价小组模块的一部分发表在 Cochrane 图书馆上，作者应该熟知这些信息。

2.3.8 系统评价的资源

单个 Cochrane 系统评价由其作者和 CRG 共同完成。每个 CRG 有一个专门的编辑小组负责编辑系统评价并通过 Cochrane 图书馆 CDSR 发表。

因为 Cochrane 协作网是围绕 CRG 建立起来的，所以每个作者在系统评价之前都需要与相应的 CRG 取得联系。除了能确保 Cochrane 系统评价能正确实施外，这种组织结构可以减轻作者的工作量，编辑小组负责提供以下大部分或全部帮助：

- 系统检索相关文献，并将潜在相关文献告知作者；
- 建立 CRG 的具体标准和方法；
- 确保系统评价作者获得需要的方法学支持；

系统评价作者需要的主要资源是他们的时间。大多数作者需要牺牲他们的业余时间去更新专业知识。某些情况下，作者需要其他的资源或向不了解系统评价过程或重要性的合作者说明其所需要花费的时间。

根据系统评价的题目、研究的数量、采用的方法（如为获取未发表研究所作的努力）、作者的经验、编辑小组提供的支持等，系统评价需要的时间是不一样的。因此系统评价的工作量也不一样。综合考虑每项任务的时间有助于作者预计系统评价所要花费的时间。这些任务包括培训、会议、撰写计划书、检索、评估研究报告的引文信息及全文的合格

性、评估纳入研究偏倚风险、收集数据、获取缺失数据和未发表研究、分析数据、解释结果、研究报告撰写和更新。

列出一个完成主要任务的时间表对按时完成系统评价是有帮助的。这些目标对不同系统评价而言存在很大差异。作者需要与 CRG 编辑小组一起商讨具体的时间表，例如表 2.3.b 所示：

表2.3.b　Cochrane评价时间表

时间（月）	任务
1-2	准备计划书
3-8	检索已发表和未发表的研究
2-3	纳入标准预试验
3-8	纳入文献评估
3	“偏倚风险”评估预试验
3-10	真实性评估
3	数据提取预试验
3-10	收集数据
3-10	数据录入
5-11	获取缺失信息
8-10	分析数据
1-11	准备系统评价报告
12-	保持系统评价更新

为完成这些任务，除时间外，系统评价作者需要的资源还包括：

- 检索（检索文献主要是 CRG 编辑小组的责任：然而，作者可能需分担该责任且对某些特别的系统评价可能需要额外检索其他的数据库）；
- 得到图书馆的帮助，如国际图书馆的下载许可和影印权限；
- 需要 2 名作者筛选文献、评估纳入研究“偏倚风险”、提取数据、录入数据和分析；

- 需要统计学支持进行纳入研究数据的综合（如果合适的话）；
- 设备，如电脑硬件和软件；
- 支持与服务，如长途电话费用、互联网、传真、打印纸、打印机、复印机、视频和电脑耗材；
- 办公室；
- 差旅费；

2.3.9 申请经费

现许多机构为优选的系统评价提供研究经费。包括研究资助部门，这些部门为卫生保健、卫生技术评估和开发临床指南提供资金。

Cochrane 协作网的一项政策是系统评价的制作以及 CRG 的基础设施的花费都不接受商业机构的资助。

2.4 在杂志或图书上发表Cochrane系统评价

作者有时可能会寻求在同行评审的医学期刊上发表 Cochrane 系统评价，特别是一些乐意接受 Cochrane 系统评价的杂志。对此，Cochrane 协作网合作发表的一个基本条件是：Cochrane 系统评价可以自由在众多媒体上传播而不受任何杂志的限制。为确保传播，Cochrane 系统评价作者授予协作网这些活动的许可，且不向其他杂志或出版社签署独家版权。杂志可自由要求发表或再发表 Cochrane 系统评价的一个非独家版权，但不能限制 Cochrane 协作网以其它任何合适的形式发表 Cochrane 系统评价。重复发表 CDSR 上的材料，特别是在杂志发表，作者需要完成一个“发表许可”表格，这个表格可以在 Cochrane 协作网指南上下载（www.cochrane.org/admin/manual.htm），并附有详细的说明。

强烈建议作者在 CDSR 上发表 Cochrane 系统评价之前不要在其他杂志上发表，特别针对 Cochrane 中心的主任和 CRG 的编辑。但有些杂志有时坚持认为杂志发表应该在 CDSR 之前。这种情况下，作者应该取得 CRG 编辑同意后和在 CDSR 上发表前方能向杂志投稿。纸质版的发表不能受制于漫长的出版时间，作者不能因杂志的延期或为了向另一杂志再投稿而过度延误 Cochrane 系统评价的发表。

杂志可以因编辑或内容的原因要求修改系统评价。欢迎杂志提供的外部同行评审，

这可增加系统评价的价值。杂志一般要求篇幅比在 CDSR 上发表的短。缩短篇幅是合理的，但在杂志上发表的和 CDSR 上发表的不应该有实质性的差异。如果在杂志上发表了，应该申明在 CDSR 上可以获取更详细的全文。例如在背景部分可以这样描述“一个更详细的系统评价将在 Cochrane database of systematic reviews 上发表和更新”。应将杂志上发表的参考文献在 CDSR 发表的系统评价计划书中引用。如果系统评价在 CDSR 的发表先于其他杂志，在前言中也应有同样的申明。Cochrane 系统评价在杂志上发表后， 应在条目“系统评价其他发表版本”中注明。也鼓励在杂志上发表的 Cochrane 系统评价版本上有如下描述：

‘这篇论文是基于首次发表在 Cochrane 图书馆 YYYY，X 期（见 http://www.thecochranelibary.com）的 Cochrane 系统评价［或者做了适当的修改］。Cochrane 系统评价随着新证据的出现定期更新并对反馈意见进行回复，应该查看 Cochrane 图书馆中最新的版本。’

下面的描述应该在杂志上发表的系统评价中描述：

‘Cochrane 系统评价的结果可以有不同的解释，受个人的观点和环境的影响。需要慎重考虑系统评价的结论。这些是作者的观点，不一定是 Cochrane 协作网所共识的。’

下面描述的内容是在向杂志投稿时提交给杂志编辑的信，递交这封信时应该给 CRG 编辑小组一份复印件。这个过程对某些杂志编辑来说可能还不熟悉，需要与其他编辑一起讨论。实际进行时 CRG 小组可能会遇到很多问题。下面的内容可以写在给杂志编辑的信中：

‘这篇系统评价是在 Cochrane 协作网的指导下完成的。Cochrane 协作网是一个通过制作、保存和传播医疗干预措施的系统评价来帮助人们更好地知证决策的国际性组织。Cochrane 图书馆的出版政策是如果征得 Cochrane 协作网的同意，其他杂志也可以发表，但 Cochrane 协作网同时也可以出版和传播这个系统评价。Cochrane 协作网不能为某些杂志提供他们要求的版权。’

2.5 将已发表的系统评价作为Cochrane系统评价发表

大部分由 Cochrane 协作网之外的作者完成的系统评价（这里我们称作“已发表的系统评价”）如果要作为 Cochrane 系统评价在 CDSR 上发表，需要增加大量额外的工作。考虑到增加的额外工作，并且与原系统评价有很大不同，所以该 Cochrane 系统评价可以作为新的系统评价发表。先前发表的系统评价必须在 Cochrane 系统评价中引用，列在“系统评价其他版本”部分。一般不需要取得先前出版商的许可。

有时 Cochrane 系统评价与先前发表的系统评价可能很相似，仅格式不一样。这种情况下在 CDSR 上发表之前就需取得原出版商的同意。如果作者不确定是否需要征求同意时，我们一般鼓励作者取得许可后再行发表。这样在向 CDSR 提交系统评价时就不会出现什么问题。如果早有在 CDSR 上发表的意愿，作者不能向杂志签独家版权（见 2.4 节）。Cochrane 协作网不要求独家版权。如果发表时未被称为 Cochrane 系统评价，并且声明是在 Cochrane 系统评价的基础上完成的，那么在 CDSR 上发表后再在杂志上发表也是没有问题的（见 2.4 节）。

2.6 声明利益和商业资助

Cochrane 系统评价不能因接受任何对评价结果感兴趣的机构或组织提供的现金或款待、津贴等形式的资助或好处而引入真实的或可察觉的偏倚，从而影响系统评价的结果。在 Cochrane 系统评价结论中必须和与评价结论有经济利益的商业资助严格划清界限。因此 Cochrane 系统评价禁止接受商业资助。可接受其他类型的资助，但资助者不能推迟或拒绝 Cochrane 评价在 Cochrane 协作网上发表，在系统评价的制作过程中不能干扰作者的独立性。在计划书中应该明确提到资助者不会影响系统评价的结果。

这一规则也同样适用 Cochrane 的其他衍生产品（包括 Cochrane 系统评价），所以商业资助者也不能干涉这些产品。任何来源的资助及其利益冲突都必须在 CDSR 和 Cochrane 其他出版物上声明。

Cochrane 协作网避免潜在利益冲突的准则都列举在表 2.6.a 中。如果一个系统评价存在严重的利益冲突就应该提交给协作网的基金仲裁者（fundingarbiter@cochrane.org）。在接受资助前，不必向当地 Cochrane 中心或督导小组提交融资方案。但必须对某些受限制或似乎与上述一般原则冲突的资助情况予以说明。

除非人们对系统评价的课题一无所知，否则不能彻底消除利益冲突（Smith 1994）。商业利益会导致很多问题，能够且应该避免，如果存在的话必须声明。任何次要利益（secondary interest）如个人的利益也可能严重影响系统评价过程中的判断（比如纳入或排除研究，评估纳入研究偏倚或解释结果）也应该声明。最常见的例子是系统评价的作者也可能是潜在纳入研究的作者。这种情况也应该在系统评价中进行说明，如果可能的话应该由另一名没有利益冲突的作者独立评估研究的合格性和偏倚风险。

声明利益冲突并不意味着降低了系统评价的价值和不诚实。然而利益冲突也潜移默化地影响判断。所以即使系统评价作者确信利益冲突不会影响他们的判断也应该让 CRG 编辑知道潜在的利益冲突。编辑会决定声明是否必要，或者决定读者是否该知道利益冲突，让他们自己判断利益冲突的重要性。由系统评价作者和编辑共同决定是否需要在系统评价中提供相关信息。

为帮助确定系统评价的完整性，所有的作者必须签署允许 Cochrane 协作网发表他们的系统评价的相关声明，包括利益冲突声明。CRG 小组也应该在他们的模块或相关系统评价中声明他们所涉及的利益冲突。

表2.6.a　Cochrane协作网避免潜在利益冲突的准则

一般原则

Cochrane 协作网的主要工作是协助系统评价作者按照协作网指定的方法和步骤制作和维护医疗干预措施的系统评价。Cochrane 系统评价不能因接受任何对评价结果感兴趣的机构或组织提供的现金或款待、津贴等形式的资助或好处而引入真实的或能察觉的偏倚，从而影响系统评价的结果。Cochrane 协作网的所有成员都应该遵守这一基本原则。

策略

1. 任何来源的资助及其利益冲突都必须在 CDSR 及 Cochrane 其他出版物中声明。
2. 如果纳入研究的作者也是系统评价的作者，那么在系统评价中也应该声明，这也被视为潜在的利益冲突。
3. 如果一个系统评价存在严重的利益冲突问题，应该提交给当地 Cochrane 中心评估（或者督导小组），如果 Cochrane 中心也涉及利益冲突，就应该让督导小组处理。
4. 在系统评价接受之前，不是必须向当地 Cochrane 中心或督导小组提交融资方案。但对某些受限制或易出现利益冲突的资助情况，这是必须的
5. 督导小组至少每年对 Cochrane 各机构接受的资助基金进行一次审核并发布 Cochrane 协助网使用资助的潜在利益冲突年度报告。
6. 督导小组是伦理组的一个下属机构，审查系统评价潜在的利益冲突并提供处理建议，对违反基本原则者进行制裁。

2.7 本章信息

作者：Sally Green, Julian PT Higgins.

本章引用格式如下：Green S, Higgins JPT(editors). Chapter 2: Preparing a Cochrane review. In: Higgins JPT, Green S (editors), Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.

参编者（自2005年3月）：Ginny Brunton, Sally Green, Julian Higgins, Monica Kjeldstron, Nicki Jackson 和Sandy Oliver.

致谢：本节以手册之前的版本为基础，后者详情见第一章1.4节。感谢Chris Cates, Carol Lefebvre, Philippa Middleton, Denise O`Connor和Lesley Stewart自2005年3月来对草稿的意见。

2.8 参考文献

Bastian 1998

Bastian H. Speaking up for ourselves: the evolution of consumer advocacy in health care. International Journal of Technology Assessment in Health Care 1998; 14: 3-23.

Effective Public Health Practice Project 2007

Effective Public Health Practice Project. Effective Public Health Practice Project [Updated 25 October 2007]. Available from: http://www.city.hamilton.on.ca/PHCS/EPHPP (accessed 1 January 2008).

Hanley 2000

Hanley B, Bradburn J, Gorin S, Barnes M, Goodare H, Kelson M, Kent A, Oliver S, Wallcraft J. Involving Consumers in Research and Development in the NHS: Briefing Notes for Researchers. Winchester (UK): Help for Health Trust, 2000. Available from www.hfht.org/ Consumersin NHSR esearch/ pdf/ involving_consumers_in_rd.pdf.

Khan 2001

Khan KS, ter Riet G, Glanville J, Sowden AJ, Kleijnen J (editors). Undertaking Systematic Reviews of Research on Effectiveness: CRD's Guidance for those Carrying Out or Commissioning Reviews (CRD Report Number 4) (2nd edition). York (UK): NHS Centre for Reviews and Dissemination, University of York, 2001.

Light 1984

Light RJ, Pillemer DB. Summing Up: The Science of Reviewing Research. Cambridge (MA): Harvard University Press, 1984.

Nilsen 2006

Nilsen ES, Myrhaug HT, Johansen M, Oliver S, Oxman AD. Methods of consumer involvement in developing healthcare policy and research, clinical practice guidelines and patient information material. Cochrane Database of Systematic Reviews 2006, Issue 3. Art No: CD004563.

Rees 2004

Rees R, Kavanagh J, Burchett H, Shepherd J, Brunton G, Harden A, Thomas S, Oakley A. HIV Health Promotion and Men who have Sex with Men (MSM): A Systematic Review of Research Relevant to the Development and Implementation of Effective and Appropriate Interventions. London (UK): EPPI-Centre, Social Science Research Unit, Institute of Education, University of London, 2004.

Richards 2004

Richards T. Poor countries lack relevant health information, says Cochrane editor. BMJ 2004; 328: 310.

（王凌译，岑啸、张龙浩初审）

第三章　系统评价的维护：更新、修改和反馈

作者：Julian PT Higgins, Sally Green 和 Rod JPM Scholten。

本文选自工作手册 5.0.1 版本。有关如何引用它的指南，见 3.7 节。这些材料还刊登于 Higgins JPT 和 Green S 编辑的《关于干预措施的 Cochrane 系统评价手册》（书号 978-0470057964）。该手册由 John Wiley＆Sons 出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 系统评价如没有持续更新就可能会滞后或误导。
- Cochrane 协作网政策规定，Cochrane 干预性系统评价要么在两年内更新，要么解释为何没有更新。
- Cochrane 系统评价的修改视为更新或修订，更新需包括检索新的研究，否则称为修订。
- Cochrane 系统评价有引文版本，本章将介绍一些标准判断何时需要新的引文版本。
- 更新 Cochrane 系统评价除了检索新的研究，还包括修改系统评价问题和采用新的方法。
- 对 Cochrane 系统评价的反馈促进系统评价的更新和保存。

- 更新日期由系统评价作者在系统评价开始部分说明，判断系统评价是否需要更新的标准将在本章中介绍。

3.1 引言

3.1.1 为什么要维护系统评价

Cochrane 系统评价的目的是为用户、临床医师和决策者的卫生决策提供“最佳”且最新的临床证据。由于某个特定问题的证据是动态发展的，因此纳入新的研究可能会改变系统评价的结果（Chalmers 1994）。因此，未维护的系统评价可能导致证据滞后或者误导。Cochrane 系统评价的一个重要特征就是作者不仅制作系统评价，还要常规保持系统评价更新。

3.1.2 系统评价更新的频率

尽管目前有一些指南（Moher 2007，Shojania 2007a，shojania 2007b），但如何合理有效地更新系统评价的经验很少。Cochrane 协作网规定系统评价应该在两年内更新，或者解释为何没有更新。“更新”的定义见 3.2.2。两年的间期是从系统评价刚更新时开始算（见 3.3.2）。

除了新出现的证据外，其他的原因也使系统评价需要更新。例如临床上，出现更好的描述亚组特征的标记物或工具，可使用更好的治疗方法，或采用了新的结局指标（或完善了已有结局指标的测量方法）。此外，Cochrane 系统评价方法的发展也需进行系统评价的更新。

在制作系统评价时，作者如果发现相关研究正在频繁发表，那就应该提出更频繁更新的需要。与之相反，如果某些领域证据新生缓慢或没有，多年前的系统评价依然是现时最新的且有价值。这种情况下每两年更新一次可能没有必要且太浪费（Chapman 2002）。如果作者认为不需要每两年更新一次，建议其与 CRG 讨论。如不能按 Cochrane 协作网的规定按时更新系统评价就应该在“发表说明”部分给出原因。

3.2 重要定义

3.2.1 引言

下面介绍 Cochrane 协作网常采用的与系统评价维护相关的重要定义以及它们在发表系统评价中的应用。3.3 节专门介绍了描述与系统评价相关事件时间的定义及应用。详细信息大部分是技术性的，作者需要理解它们以便在系统评价中正确应用并在 Revman 中完成相关内容。

3.2.2 更新和修订

对 Cochrane 系统评价任何修改均称作更新或修订。

更新必须包括检索新的研究。在将系统评价标为“更新”之前，如果存在新的研究就需要将其添加到纳入研究、排除研究或正在进行中的研究部分（不能归入上述部分的就列入“等待分类”部分）（见 3.2.5.1）。

对 Cochrane 系统评价或计划书的其他改变称作“修订”，工作量可大可小。更多术语以及何时应用它们详见 3.2.4。

3.2.3 Cochrane系统评价或计划书的引文版本

发表的系统评价或计划书都有当前引文版本。对系统评价而言，引文版本被视为主要的新出版物并被 MEDLINE 和 SCI 等数据库收录，而计划书则没有。需创建引文版本的情况见表 3.2.a。

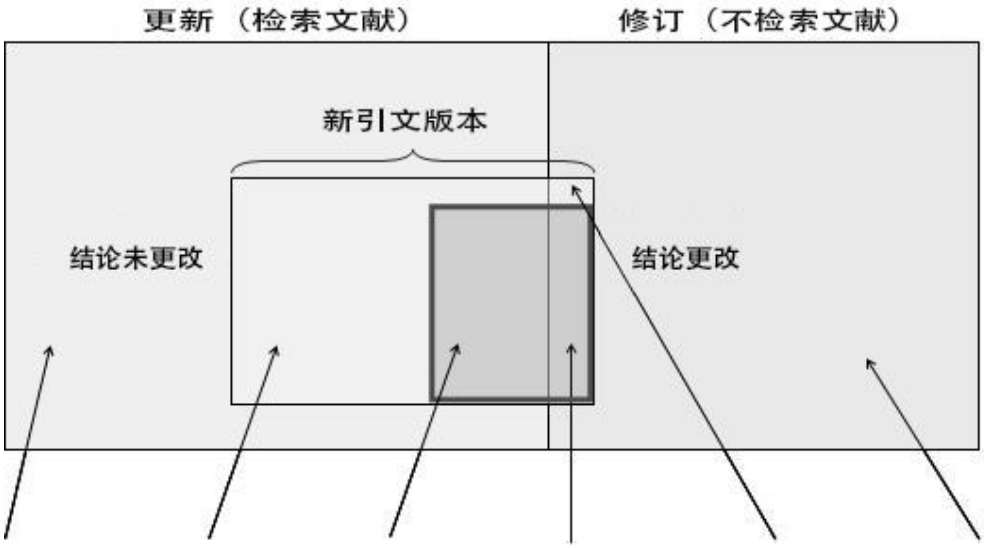
系统评价经过重大修改（更新或修订）后就应该在 CDSR、MEDLINE、SCI 中有新的引文（如系统评价的结论修改、作者改变和纠正严重错误），称为新引文版本。此外一些新引文版本需要在 CDSR 中标出（如用一个旗帜符号），特别是结论改变后应该提示重新阅读，这些特殊的引文版本称为结论修正。尽管有些更新的系统评价没有达到创建新引文版本的标准，但他们依旧很重要，因此所有更新的系统评价在 CDSR 中都应该标明（如用旗帜符号标明“新检索”）。

计划书有重大修改（如作者和纳入标准发生变化）时也可有新引文版本。但计划书未被 MEDLIEN 和 SCI 收录，所以只影响到 CDSR 内部引用。重大修改后的计划书也应该为有兴趣重读的读者标出（如用旗帜符号）。这些计划书称作大修。

图 3.2.a 总结了 Cochrane 系统评价的修改类型，图 3.2.a 总结了 Cochrane 计划书的修改类型。

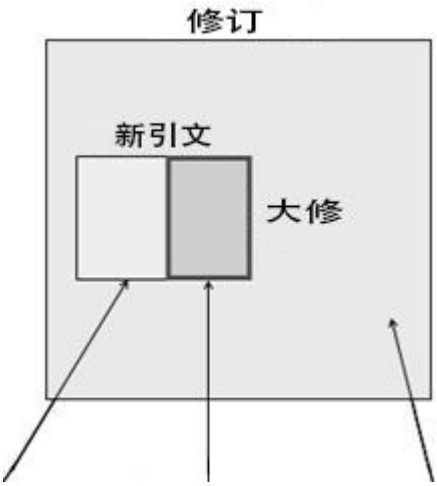
表3.2.a　系统评价或计划书建立新引文版本的情况

计划书首次发表；
确定为新引文版本后计划书再次发表；
系统评价首次发表（如从计划书转为系统评价）；
确定为新引文版本后系统评价再次发表（修订或更新）；
系统评价撤销后再次发表，分割或合并已有的系统评价或计划书。



更新，无新引文版本如未更改结论或作者	更新，有新引文版本，未更改结论如更改作者	更新，有新引文版本，更改结论如目前对疗效有充分的证据	修订，有新引文版本，更改结论即修正结论中的严重错误（勘误）	修订，有新引文版本，未更改结论即修正引文中的严重错误（勘误）	修订，无新引文版本如修正小错误或修改方法

图3.2.a　总结Cochrane系统评价的修改类型



修订， 有新引文版本， 没有大修计划 如改变作者	修订， 有新引文版本， 大修计划 如修改范围	修订， 没有新引文版本 如修正小错误

图3.2.b　总结Cochrane计划书的修改类型

3.2.4　Cochrane计划书相关术语的应用

3.2.4.1　计划书的修订

发表的计划书进行任何修改或编辑（包括撤销）均使计划书为修订状态。计划书不可能“更新”，修订后在 CDSR 上再次发表。计划书可在任何时间都接受修改。修订的工作量可大可小，可使计划书发生或大或小的变化。

3.2.4.2　计划书的新引文版本

计划书修订后，由 CRG 根据标准（见表 3.2.b）判断是否作为新引文版本发布。尽管计划书引文未被 MEDLINE 和 SCI 收录，但在 CDSR 中正式的题录也应被修改。

计划书新引文版本进一步分为大修和未大修，前者将在 CDSR 中标出。

表3.2.b　Cochrane计划书建立新引文版本的标准

Cochrane计划书建立新引文版本的标准：大修（major change） 如果研究目的和研究范围有重要的改变，如修改研究纳入标准，则分为大修的新引文版本，这样的计划书在CDSR上再次发表时标注为“大修”。 Cochrane计划书建立新引文版本的标准：未大修（no major change） 如果系统评价小组发生重大调整，则分为未大修的新引文版本，这样的计划书不会在CDSR中标注。

3.2.4.3　计划书的修改不被视为新引文版本的实例

除非满足表 3.2.b 中一个或两个标准，否则下列修改不会使计划书归为新引文版本。对发表的计划书会进行修改，但仍采用原引文。

- 修改计划书的文字内容（如背景部分）
- 修改原方法
- 改变原有作者的排序（除第1作者的更改），或者删除作者
- 改错

3.2.5　Cochrane系统评价相关术语的应用

3.2.5.1　系统评价的更新

系统评价更新指在已发表系统评价的基础上检索、纳入新的研究(也可没有)，修改原系统评价。在 CDSR 中更新的系统评价被标注为“新检索”。新检出的研究必须整合到更新的系统评价中（除非确实不能归入纳入、排除或正在进行中的研究组才能归入“等待归类”组）。如未检出最新研究也可视为已更新。

这个定义表明系统评价更新是‘检索并筛选出新的证据整合到已发表的系统评价中（Moher 2006）’。更新系统评价的工作量因检索结果而大小不等，原则上至少每两年更新一次。

3.2.5.2　系统评价的修订

除更新以外对系统评价所做的修改、编辑（包括撤销）称为修订。系统评价未进行新检索且有以下一点及以上者则认为已修订：（ i ）方法学的改变；（ ii ）纠正拼写错误；

（iii）重写背景部分；（iv）纳入等待分类的研究；（ｖ）因重大的编码错误而修改结论。Cochrane 系统评价在任何时候均可接受修订。修订系统评价的工作量大小不等，对原系统评价的影响可大可小。

3.2.5.3　系统评价的新引文版本

更新或修订的系统评价才可能以新引文版本再发表。作者和 CRG 共同决定是否将系统评价归为新引文版本。表 3.2.c 列举了将评价归为新引文版本的 6 条明确标准。除三种特殊情况外（结论本质性修正，紧急纳入新的证据，引文本质性改动），仅更新的系统评价符合新建引文版本的标准。

新引文版本又进一步分为“结论更改”和“结论未更改”两类。前者在 CDSR 上标注为“结论更改”。

系统评价在两个新引文版本之间可能进行更新或修订，但这些更新情况会在 CDSR 上发表但没有新引文。因此在“系统评价更新日期”栏目反映系统评价更新程度就非常重要。

表3.2.c　Cochrane系统评价新引文版本分类的标准

系统评价新引文版本的标准：结论更改
1. 更新后结论更改
如果系统评价在更新后结论发生变化以至于读者需重新阅读该评价，则必须归为结论更改的新引文版本。
增加或剔除研究、方法改变或者系统评价范围变化（如采用新的结局指标、对照、受试者，或干预措施及给予途径发生改变）都可能会造成结论改变。不论研究结果显示干预措施是否有效，其结论改变几乎都会应用于临床，然而有时对临床应用影响重大，如新纳入研究的数据解决了原系统评价尚不确定的问题。所有结论有重要改变的系统评价必须在摘要中报告。
2. 纠正错误（勘误）后结论更改
如因纠正严重错误后系统评价结论改变以至于读者需重新阅读该评价，也视为结论更改的新引文版本。此类改变需在传统的纸质出版杂志中发表勘误表。
3. 紧急纳入新的证据后结论改变
如因紧急纳入关于干预措施疗效的新研究证据后改变了原有结论以至于读者需重新阅读该评价，也视为结论更改的新引文版本。
系统评价新建引文版本的标准：结论未更改

4. 新的作者

如果系统评价增加了大量新信息，或者方法学上有重要变化，或者较大范围重写或复制，虽结论未更改，那么在CRG和作者的共同判断下，即使结论未改变，也可能视为新引文版本。如果作者发生重要变化（包括更改第一作者，但一般不包括重排作者顺序和删除作者）也是如此。符合第4章4.2.2节中列出的标准才能作为作者。

作者承诺维护系统评价可能需要大量工作来更新系统评价，更新有可能不会改变结论。如果系统评价由同一个工作组更新，且结论没有更改将不能有新引文版本。然而如果系统评价小组增加或替换了作者，则此类系统评价可能作为新引文版本以给予新作者适当肯定。

5. 积累性改变

如果系统评价发表超过5年，且现在的版本与最初版本已有很大不同，不论其结论和作者是否改变，经作者和CRG讨论后都可视为新的引文版本。积累性变化可能表现为：重写、增加研究数量、或方法学发生重大变化等随时间累积，从而使系统评价发生变化。

每个系统评价都应包含一个最后更新日期。因此该条判断系统评价为新引文版本的标准仅用于MEDLINE和SCI等引文数据库提示新的引文版本，而不用于决定修改或更改事件的日期。

6. 更正引文的严重错误（勘误）

如果引文版本中有严重错误需要更正，虽然结论未更改，但也归为新的引文版本，如作者名字拼写错误，需要传统的杂志中发表勘误。更新时未必要勘误，有影响结论的严重错误时参见上述第2条标准。

3.2.5.4 不被视为新发表的系统评价修改实例

除非符合表 3.2.c 中 6 条标准的一项或多项，以下一些情况不能将系统评价归为新引文版本。这些改变为更新或者修订，但依然保持原来的引文版本。

- 增加新的研究。
- 分析结果改变（点值估计或可信区间），结论未改变。
- 系统评价内容改变（如背景和讨论部分）。
- 方法发生变化。
- 修改作者排序（除第1作者改变），或者删除作者。
- 勘误。

3.3 与Cochrane系统评价相关的重要日期

3.3.1 引言

Cochrane 系统评价有几个重要的时期，有些是由 RevMan 自动生成的，有些需要作者输入。这些日期便于读者了解系统评价，也有助于系统评价出版管理。在更新和修订时输入相关的日期时应用这些定义是很必要的。

3.3.2 系统评价更新日期需要作者输入（仅在系统评价全文中，不包括计划书）

在发表时，将这个时期放在显著的位置，以告知读者系统评价被评定为最新的日期。评定标准见表 3.3.a。

尽管系统评价发表多年后只进行过很小的修改（如评价发表以来最新文献检索未检出新证据），也可认为是实时更新的。所有发表的系统评价都应该明确评定为最新的日期。这个日期应该由作者输入，并与作者提交准备在 CDSR 上发表的系统评价日期相符。系统评价被接受发表后可适当修改确定为最新的日期。

表3.3.a 评定系统评价为最新的指南

必须选择系统评价评定为最新的日期以便使系统评价（新的，更新的或者修订的）满足以下标准： 1. 干预措施疗效的证据是目前最新的 纳入研究需包括所有可获得的证据，最新检索的结果是在系统评价评定为最新的6个月之内。另外，下面的条件也是很重要的，但不是必须的： 2. 系统评价的方法是最新的 所有Cochrane系统评价要求的方法（据关于干预措施的Cochrane系统评价手册的目前版本所描述）都应包括在内。 3. 系统评价中描述的事实是正确的 描述的事实（如背景和讨论）不应过度陈旧。

3.3.3 检索的日期

这个日期由作者输入（仅在系统评价全文中，不包括计划书）。“检索”在这里是指检索所有系统评价中指定的数据库。如果不同的数据库是在不同时间检索的，就应该在系统评价全文中列出检索每个数据库最近的日期，然后在“data of search”这个地方填写最早检索数据库的时间。例如：某系统评价检索了以下数据库及日期：

MEDLIEN 2007 年 6 月 5 日

EMBASE 2007 年 6 月 12 日

Specialized Register 2007 年 6 月 26 日

CENTRAL 2007 年 6 月 28 日

此系统评价检索的时间为：2007 年 6 月 5 日。

3.3.4 预期进入下一个阶段的时间

由作者自行输入：

计划书：预期转为全文的时间；

全文：预期更新的时间。

3.3.5 最后一次编辑的时间

这个时间由 RevMan 根据对系统评价进行的修改记录自动生成，不发表。这个时间用于确定目前发表的系统评价终稿时间。

3.3.6 声明系统评价不再需要更新的时间

这个时间在少数系统评价中应用，应该咨询 CRG 后谨慎决定。系统评价在很长一段时间内（数年，而不是数月）都很有可能维护其相关性者才无需再更新。这类系统评价不受 Cochrane 协作网更新规则限制，必须与 CRG 协商，且要定期审查。此类系统评价包括以下两种情形：

- 干预措施是可取代的（记住Cochrane系统评价应该是国际性的）；
- 系统评价的结论很确切，即使增加新的信息也不会被更改，且干预措施没有可预见的不良反应；

只要在最近的“新内容”条目声明“不再更新”，那么该系统评价就保持“不再更新”。如果随后新增“新内容”条目，这个系统评价就如其他系统评价一样需要及时更新了。

3.4 更新系统评价需要考虑的方面

3.4.1 更新从何开始

很少有方法学的文章介绍应该何时、如何更新系统评价（Mother 2008），但是某个研究领域是在不断发展和改进的。本章指南根据方法学研究的新成果定期更新。更新系统评价一般每两年一次，每次更新都应检索新的研究。如果有新的研究检出，看其是否符合纳入标准，如果符合纳入标准就整合到原系统评价中。在更新系统评价时应考虑额外的问题，如：

- 是否需要对研究问题和纳入标准进行修改，如增加新的结局指标或对照组，随着疾病分类方法的完善增加新的亚组分析；
- 修改方法，如纳入研究偏倚风险评估或者增加结果总结表格

3.4.2 更新不修改研究问题的系统评价

3.4.2.1 重新检索

如果不修改研究问题和纳入标准，更新的第一步就是检索新的研究，并确定更新的步骤。因为 CRG 有丰富的资源，定期鉴别潜在的相关研究并向系统评价员传送引文是编辑小组（常为检索协调员）不间断的一个职责。其他情况下作者要自行检索。系统评价更新检索新研究的策略应至少包括从上次更新的“检索日期”开始重新检索（见第 6 章 6.4.12 节）。

如果在检索方法上有些改进或者作者认为可完善原有的检索策略，就需要从最近一次检索时间开始进行新检索，且检索策略中添加或修改的检索词直接覆盖原文中的检索词。

3.4.2.2 更新无新研究的系统评价

如未检索到相关研究或无符合纳入标准的研究，那么系统评价更新只需在相应部分进行记录即可。几个部分的内容可能需要修改：

1. 检索方法（确保“检索时间”记录正确）；

2. 结果部分对研究的描述（修改检出研究的数量，筛查、排除的研究）；

3. 结果（确保日期均合适）；

4. 作者的结论（特别是对进一步研究的需要）；

5. 摘要和通俗语言总结

除系统评价内容需要修改外，作者还应该确保相关的日期是否正确，是否反映更新状态（见 3.3 节），“新内容”表格是否完成（见 3.5）。

为了提示读者其阅读的是已更新的系统评价，在摘要背景部分可添加一句话说明这是更新的系统评价并引用之前的版本，以及附上之前版本的参考文献和发表时间。在该评价的背景部分，该句子也可包括对原系统评价结果的讨论。

最后，检查系统评价中没有过时的内容也是很重要的（如引用的其他 Cochrane 系统评价可能已更新，感兴趣疾病的患病率和发病率，“最近，1998 年显示……”，“下一年，2002 年，有……”的描述）。如果在“致谢”和“声明利益冲突”中有更改也应该进行修改。

3.4.2.3 更新有新研究的系统评价

如果检出新的潜在相关研究，就需要评估研究是否能纳入，其筛选的方法与原系统评价相同（筛选研究的信息见第 5 章）。

如果新的研究纳入更新的系统评价，其引文应该输入 RevMan 软件，并提取数据（见第 7 章）和评估偏倚风险（见第 8 章）。新检索并纳入的研究数据需要输入 RevMan，如有可能需重新进行 Meta 分析（见第 9 章）。更新系统评价采用的方法应与原系统评价相仿，除非明确地改动（例如系统评价方法改进如采用“偏倚风险评估”表格和“结果总结”表格）。更新系统评价与原系统方法不一致的地方及理由要在“系统评价和计划书的不同处”部分说明。

有新研究纳入的系统评价更新修改量取决于新数据对结果的影响。例如，纳入研究为小样本时对系统评价的结果和结论基本无影响（那么除在 3.4.2.2 描述之外对正文内容基本无需修改），只是增加先前研究的可信度；某些情况中新纳入研究可改变结论（需重写系统评价结果、讨论、结论、“结果总结”表，摘要和通俗语言总结）。提示读者他们阅读的是更新的系统评价的声明（见 3.4.2.2 节）也应包括在摘要和背景中。

作者还应该确保相关的日期是否正确，是否反映更新状态（见 3.3.2 节），“新内容”表格是否完成（见 3.5 节）。最后，检查系统评价中没有过时的内容也是很重要的（如引

用的其他 Cochrane 系统评价可能已更新，患病率和发病率可能描述为“最近，1998 年显示……”，“下一年，2002 年，有……”)。如果在“致谢”和“声明利益冲突”中有更改也应该进行修改。

3.4.3 修改研究问题和纳入标准

更新系统评价时除重新检索外，有时需要修改研究问题、研究纳入标准。例如随着技术的发展需要纳入新的对照，或者病人类型（如小孩和成人），或者重要的结局指标（如不良反应）可能没有在原系统评价中得到强调。更新系统评价与原系统评价不一致的地方要在“系统评价与计划书不同处”部分记录并说明理由，在全文中（背景、目的和方法部分）解释，在“新内容”部分标明。

除此而外，检索策略也可能需要修改，检索年限不但要包括原系统评价“检索时间”后的时间，还包括原系统评价的检索年限。有时采用原检索策略更新检索，用新的纳入标准筛选文献。

如果添加了新的对照和结局指标，这种情况下有必要检查原系统评价纳入的研究，看是否涉及新的对照和结局指标。原资料提取表也可能需要修改，并且要再次进行预试验。新的对照和结局指标也应纳入分析。

最后如增加了新的对照、受试者或者结局指标就需要对系统评价全文（背景和方法）进行修改。如果纳入了新的研究就需要对结果、结论、通俗语言总结和“结果总结”表格进行修改。

3.4.4 分割系统评价

有时系统评价太大，需要将其分割为两个或多个范围较窄研究文献可能较少的系统评价，这样可减轻更新的工作量。

分割一个系统评价产生至少一个新的引文版本，与原系统评价的正规链接可能会丢失。分割一个系统评价有时会撤销原系统评价。不能轻易决定分割系统评价，需咨询 CRG 编委会。

Cochrane 系统评价再评价（见第 22 章）可能会使分割系统评价变得容易些，可能将多个范围较窄的系统评价（如单个干预措施治疗某种疾病）在针对某种特定疾病的所有干预措施的再评价中综合在一起。

3.4.5 系统评价方法的修订

除检索新的研究、修改研究问题和纳入标准外，维护系统评价还包括对系统评价的方法进行修改（Shea 2006）。原系统评价发表后的方法学进展可能在更新时需要修改或扩展系统评价的方法。作者可在更新的系统评价中加入新的分析策略（如采用之前RevMan 中没有的统计方法）。第 8 章中介绍的“偏倚风险评估”表（第 8 章）和“结果总结”表（见第 11 章）都是 RevMan 5 新增的内容，更新时最好采用这些新方法，但不是强制性的。更新时加入“偏倚风险评估”表后，作者要决定是否采用新的方法重新评估之前纳入的研究，或者只用于评估新纳入研究的偏倚风险。在已发表的系统评价中“偏倚风险评估”表只纳入有数据的研究（即没有空白行）。

作为系统评价更新的一部分，作者可能希望囊括“结果总结”表格（见第 11 章）。“结果总结”表中要选择对卫生决策有重要意义的结局指标（常是系统评价的主要结局指标），而且应该在更新之前确定，以免受阳性结果影响而选择性报告结果，不依据重要性进行选择。

方法学修改可能牵涉到修改原有的计划书，这些修改和理由必须在“系统评价与计划书不同之处”部分和“新内容”表中标明。

3.4.6 系统评价其他更改

如果主要作者发生变化，有新的作者加入系统评价小组，或者由新的小组负责更新，则作者署名（作者列表）就需要修改。是否署名和作者排序应根据对更新系统评价的历史贡献及对最终更新文件的审批综合考虑。如果某个作者不再能够批准更新的系统评价，那么这个作者就不能署名，但可以在致谢中提到。所有作者对原系统评价和更新系统评价的贡献都应该在“作者贡献”部分描述。

修改系统评价的作者可能对给予系统评价新引文版本有影响（见 3.2.5.3）。

3.4.7 编辑过程

完成更新后，更新的全文应提交给编辑部进一步处理。更新系统评价在编辑中有多种情况。如果更新的系统评价没有进一步分析或没有改变结论就无需特殊处理，如果进行了新的分析或结论改变或方法改变就将重复和原系统评价一样的发表前评审过程。

很少有系统评价从 CDSR 上被撤销。这可能是暂时的（如系统评价严重过期或存在

重大错误），也可能是永久被撤销（如系统评价被分割为一些小的系统评价）。撤销的系统评价应该在“发表备注”部分标明，也会在每期的 CDSR 上发表。如果系统评价被暂时撤销，当其内容经系统评价作者和 CRG 判定为合适时可再次被认可。如果一个系统评价因与其他的系统评价合并而被撤销，需要在“发表备注”中说明撤销的原因。

3.5 “新内容”和历史事件表格

3.5.1 “新内容”事件

所有对计划书和全文进行的修改都应列在“what’s new”表格中，以便读者可以快速清楚了解哪些地方进行了修改。表格中的更新事件决定计划书或全文在 CDSR 中处于什么状态，包括采用旗帜或者其他装置标注及评估是否作为新的引文版本。

3.5.2 完成“新内容”表格

“新内容”或历史事件表格每一行包括：

- 事件进行或记录的时间；
- 事件的类型；
- 简述更新的内容。

表 3.5.a 和表 3.5.b 列出了计划书和全文可能的更改，作者应该参考相关章节以在“新内容”表中列出合适的更新事件。撤销应该与“修订”事件相关。

表3.5.a 计划书可能的更新事件

事件类型	定义或讨论	与原计划书的联系
修订	见3.2.2和3.2.4.1	无
反馈	见3.6	计划书标记为“评述”
新引文：无重大更改	见3.2.4.2	新引文
新引文：重大更改	见3.2.4.2	新引文，计划书标记为“重大更改”

表3.5.b　全文可能的更新事件

事件类型	定义或讨论	与原系统评价的联系
修订	见3.2.2和3.2.5.2	无
更新	见3.2.2和3.2.5.1	标注为“新研究”
反馈	见3.6	标注为“评述”
新引文：结论未更改	见3.2.3和3.2.5.3	新引文（如MEDLINE记录），重设影响因子
新引文：结论更改	见3.2.3和3.2.5.3	标注为“结论更改”，新引文（如MEDLINE记录），重设影响因子
停止更新	见3.3.6	无

当可输入多个事件到更新表格时，作者应确认此次改变是否不同于上次的版本。更为重要的是这个表格中只能有一个新引文和一次更新记录（以前的事件应该加入历史事件表格中）。

3.5.3　历史事件表格

非本次更新的事件都应移到历史事件表格。此外，历史事件表格还包括以下信息，由协作网信息管理系统自动完成：

- 计划书首次发表的年限
- 全文首次发布的年限
- 每个新引文发布的年限

3.6　Cochrane系统评价反馈的归纳综合与处理

Cochrane 图书馆有正式的机制管理系统评价用户的反馈信息。反馈原称评述和批评，设计为“……根据新证据修订评价……反映任何来源的新数据、有效反馈、请求或非请

求的出现”（Chalmers 1994）。

系统评价发表后任何时间都可以提出反馈意见，并提交给 CRG 的反馈编辑小组。编辑确认反馈信息和语言合适后传递给作者（通常要求 1 个月内）。作者回复时要求：

- 回复信息限制在反馈意见提出的相关问题；
- 回复每个重要的问题，清楚表明“同意”或“不同意”反馈意见，并提出相关证据支持；
- 描述系统评价根据反馈做了哪些修改；
- 采用清楚的通俗语言。

更新系统评价让反馈信息有机会整合到系统评价中，通过反馈机制可以有效地阐述关心的问题并增加符合纳入标准的研究。

3.7 本章信息

作者：Julian PT Higgins, Sally Green和Rob JPM Scholten.

本章引用格式：Julian PT Higgins, Sally Green和Rob JPM Scholten. Chapter 3: M aintaining reviews: updates, amendments and feedback. In: Higgins JPT, Green S (editors), Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.

致谢：Cochrane协作网更新工作小组（成员Mike Clarke, Mark Davies, Sally Henderson, Harriet MacLehose, Jessie McGowan, David Moher, Rob Scholten（会议召集人）和Phil Wiffen）对草稿提出的意见。

3.8 参考文献

Chalmers 1994

Chalmers I, Haynes B. Reporting, updating, and correcting systematic reviews of the effects of health care. BMJ 1994; 309: 862-865.

Chapman 2002

Chapman A, Middleton P, Maddern G. Early updates of systematic reviews - a waster of resources? Pushing the Boundaries: Fourth Symposium on Systematic Reviews, Oxford, 2002.

Moher 2006

Moher D, Tsertsvadze A. Systematic reviews: when is an update an update? The Lancet 2006; 367: 881-883.

Moher 2007

Moher D, Tsertsvadze A, Tricco AC, Eccles M, Grimshaw J, Sampson M, Barrowman N. A systematic review identified few methods and strategies describing when and how to update systematic reviews. Journal of Clinical Epidemiology 2007; 60: 1095-1104.

Moher 2008

Moher D, Tsertsvadze A, Tricco AC, Eccles M, Grimshaw J, Sampson M, Barrowman N. When and how to update systematic reviews. Cochrane Database of Systematic Reviews 2008, Issue 1. Art No: MR000023.

Shea 2006

Shea B, Boers M, Grimshaw JM, Hamel C, Bouter LM. Does updating improve the methodological and reporting quality of systematic reviews? BMC Medical Research Methodology 2006; 6: 27.

Shojania 2007a

Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Annals of Internal Medicine 2007; 147: 224-233.

Shojania 2007b

Shojania KG, Sampson M, Ansari MT, Ji J, Garritty C, Rader T, Moher D. Updating Systematic Reviews. Technical Review No 16 (Prepared by the University of Ottawa Evidence-based Practice Center under Contract No 290-02-0017). Rockville (MD): Agency for Healthcare Research and Quality, 2007.

（王凌译，岑啸、张龙浩初审）

第四章　Cochrane 计划书及系统评价内容指南

本文选自工作手册 5.0.1 版本。有关如何引用它的指南，见 4.13 节。这些材料还刊登于 Higgins JPT 和 Green S 编辑的《关于干预措施的 Cochrane 系统评价手册》（书号 978-0470057964）。该手册由 John Wiley＆Sons 出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- Cochrane系统评价有其特定格式要求，使用RevMan软件可以很方便地遵循这一格式。本章讲述了在一篇Cochrane计划书或系统评价的各个组成部分中，要求作者纳入的内容以及读者希望了解的内容。
- 本章同样也可以作为Cochrane手册其他部分的指南，它包含了其他章节的链接，我们可以通过这些链接找到关于方法学问题的深入讨论。
- “系统评价信息”（或“计划书信息”）部分包括作者详情和维护、更新系统评价的重要日期。
- 正文应该简洁易读，这样非该领域的专业人员也能读懂。计划书的正文应该以方法部分结束。

- “纳入研究和参考文献”部分应提供框架以便划分纳入研究、排除研究、在研的研究、信息不完善的研究和其他参考文献。
- 研究特征表系统呈现系统评价考虑的研究的关键特征。
- “数据和统计分析”部分具有层次结构，纳入研究的数据可进行不同的亚组分析，比如以特定结局为指标进行Meta分析，或以某特定的干预措施进行分析。任何Meta分析、森林图和漏斗图都能用RevMan软件生成。
- 其他的图表、附件可作为内置表格的补充。

4.1 引言

所有 Cochrane 干预性系统评价都有同样的格式，使用 Review Manager（RevMan）软件可更易于制作该格式的系统评价。本章介绍一篇完整的系统评价（或计划书）所包括的内容并明确各部分要点。Cochrane 手册中收录其他章节的大量参考文献，以便为各章节提供相应指导意见。RevMan 软件的使用方法可以在软件本身的帮助系统中获得。

4.2 标题与系统评价信息（或计划书信息）

4.2.1 标题

标题应简要陈述系统评价的干预措施以及该干预措施所要解决的问题。构建 Cochrane 系统评价题目的明确指南见表 4.2.a。

表4.2.a　Cochrane系统评价标题结构

类型（scenario）	结构	举例
基本结构	［干预措施］治疗［健康问题］	抗生素治疗急性支气管炎
两种有效干预措施对比	［干预措施A］与［干预措施B］对比治疗［健康问题］	及时治疗与延迟治疗子宫颈上皮内瘤变的对比研究
明确提到研究对象类型或干预的地点	［干预措施］治疗［研究对象/地点］的［健康问题］	一氧化氮吸入治疗早产儿呼吸衰竭
未指定某个具体的“健康问题”（如：家庭生产与医院生产）或者该干预措施会影响到各种问题（如：早产儿合成表面活性剂的预防性应用）	［干预措施］用于或治疗［研究对象/地点］	早产儿水摄入量限制与不限制的对比研究
有时需要说明干预措施是用于预防、治疗、或者防治某健康问题。如必要的话，在单词“for”后紧跟“preventing（预防）”、“treating（治疗）”、或者“preventing and treating（防治）”，这比“for the prevention of（为了预防……）”更好。		水池栅栏防止小孩溺水； 阿莫地喹治疗疟疾； 维生素C防治普通感冒

4.2.2　作者

所有科学论文（包括 Cochrane 计划书和系统评价）的著作权确立了问责制、责任制和信誉制（Rennie 1997，Flanagin 1998，Rennie 1998）。当我们在决定谁应该出现在 Cochrane 评价的作者署名时，应区分对该系统评价有实质贡献的个人（应列在作者栏）和在其他方面提供帮助的人（应在致谢部分标注），这非常重要。基于“提交给生物医学杂志的手稿统一要求”（医学杂志出版社国际委员会 2006），著作权的确立应具备下面所述的三个步骤。作者必须签署一份明确以下 3 项贡献的“发表许可”。

- 研究的构思和设计，或者数据的分析和解释
- 起草系统评价或批判性评论其学术内容
- 最终同意发表文稿

详细贡献应该列于“作者的贡献”部分（详见下）。作者列表可以是一个个体的名字，多个个体的名字，一个合作小组（如：晚期膀胱癌评价组）的名字或者一个或多个作者与合作组的组合。理论上，作者的排列顺序应与他们对于该系统评价的贡献大小相关。对系统评价贡献最大的人应列于第一位。

4.2.3 通讯作者

应列出负责系统评价的通讯作者详细联系方式，同时该作者要负责维护和更新系统评价。通常，通讯作者应（1）负责建立和组织系统评价小组；（2）与编辑部门进行交流；（3）确保系统评价在约定的时间内完成；（4）向编辑部门递交系统评价；（5）与联合作者交流反馈意见；（6）确保更新工作的准备。

通讯作者并非必须为第一作者，通讯作者的选择并不影响系统评价的引用。如果当前的通讯作者不想再负责已发表的系统评价而且该系统评价组的其他成员也不想为此负责，系统评价组协调人员的联系信息将会列出。系统评价的通讯作者并不一定作为作者列出。

4.2.4 日期

4.2.4.1 评定为更新

系统评价最后更新日期应和作者提交系统评价并考虑在 Cochrane 系统评价数据库发表的时间一致。

同样可见于：

- 系统评价定为最新的具体标准见第3章（3.2小节）

4.2.4.2 检索时间

这个时间有助于确认某个评价是否已经更新，并且告知该系统评价应该更新的时间。Cochrane 系统评价数据库不发表该信息。

同样可见于：

- 检索时间的具体标准见第3章（3.3.3小节）
- 检索方法的详细讨论见第6章（6.3小节）

4.2.4.3 预期的下个阶段

在内部使用的时间（不会发表在 Cochrane 系统评价数据库中），仅提示完成系统评价（计划书中）的时间或下一次更新系统评价（评价中）的时间。

同样可见于：

- 系统评价更新策略见第3章（3.1小节）

4.2.4.4 首次发表计划书

Cochrane 系统评价数据库将给出首次发表计划书的期号（如：2004 年第 2 期），该时间不能在 RevMan 软件中编辑。

4.2.4.5 系统评价首次发表

Cochrane 系统评价数据库将给出系统评价全文首次发表的期号（如：2005 年第 1 期），该时间不能在 RevMan 软件中编辑。

4.2.4.6 最新引文期号

Cochrane 系统评价数据库将列出系统评价的当前引文版式首次发表的期号（如：2007 年第 1 期），该时间不能在 RevMan 软件中编辑。

同样可见于：

- 引文版本的详细讨论见第3章（3.2小节）

4.2.5 新内容和历史

“新内容”部分应标出计划书或系统评价在 Cochrane 系统评价数据库发表至今的所有更改情况。在系统评价的每次更新和修改中，应至少记录一次“新内容”事件，包括事件类型，更改时间和描述更改内容。该描述可简要总结系统评价增加了多少新信息（如：研究的数量、参加者或额外的数据分析）和系统评价的结果、结论或方法有哪些重要改变。与当前系统评价引文版本不相关的“新内容”表格词条应列在“历史”表格中。

同样可见于：

- “新内容”表格事件的详细讨论见第3章（3.5小节）

4.3 摘要

一篇完整的系统评价必须包含一篇少于 400 字的摘要。摘要应简洁但不能遗漏重要内容。Cochrane 系统评价摘要发表在 Medline 和 SCI 上，可以从互联网免费获得。它可以作为独立文献阅读，因此非常重要。

同样可见于：

- 摘要内容指南见第11章（11.8小节）

4.4 通俗语言总结

通俗语言总结（以前称总结）旨在总结系统评价，使其以通俗易懂的方式为卫生保健用户所明白。通俗语言总结可在网上免费获得，所以常以独立文献阅读。通俗语言总结由两部分构成：通俗语言标题（即将系统评价的标题用简单术语重述）和少于 400 字的总结。通俗语言总结内容指南详见第 11 章（11.9 小节）。

4.5 正文

系统评价的正文应简明易读。尽管 Cochrane 系统评价没有正式的字数限制，但如果没有特殊原因要写一篇更长的系统评价，那么 10000 字绝对是上限了。实际上大部分系统评价都远少于这个字数。一篇系统评价的撰写应该能让非该领域专家的人读懂，我们可借鉴 Cochrane 手册讲述的策略声明（www.cochrane.org/admin/manual.htm）：

“Cochrane 系统评价的目标读者是医疗卫生决策者，这其中包括卫生保健专家、患者和对重大疾病或问题有基本理解的决策制定者。

Cochrane 协作网的基本原则和使命之一是让任何想要对医疗卫生做出某种决定的人能够更容易获得关于某种干预措施有效性的系统评价。但这并非指任何教育背景的人都必须理解 Cochrane 系统评价。这是不可能的，就像我们不能单用某一种语言来写

Cochrane 系统评价而期望全世界每个人都能看懂。

Cochrane 系统评价的撰写，应该让对该题目有基本认识但未必是该领域专家的人容易读懂。对一些术语和概念的解释很可能有用，甚至很重要。但是，太多的解释会减弱系统评价的可读性。简洁明了对于可读性是至关重要的。Cochrane 系统评价的可读性应与一篇普通医学杂志上的好文章相当。”

Cochrane 系统评价包括很多由 RevMan 软件设置的标题和子标题。作者可以在任何地方增加子标题。推荐所有作者使用部分特定子标题（RevMan 软件能启用或不启用）。但这并不是强制性的，而且如果它使个别部分内容出现不必要的过短时，我们应避免使用。另外，下面将讨论可能与个别系统评价有关或无关的可选择子标题。想要将推荐子标题与可选择子标题相结合的评价作者应该确保它们以适当的一致性风格呈现，这或许会要求你手工创建标题而不启用 RevMan 软件内置的所有推荐标题。

“方法”、“纳入研究标准”、“结果”和“作者结论”这些固定标题下带有多个固定子标题并且可以没有内容紧随其后。

背景 ［固定，一级标题］

阐述较好的系统评价，其问题产生于许多已形成的知识体系中。“背景”应体现出该内容，这有助于确立系统评价的合理性，解释该系统评价问题的重要性。背景应简洁（一般打印出来一页左右）并能被调查研究该干预措施的用户所理解。所有的信息来源应该标注出来。

问题描述 ［推荐，二级标题］

系统评价应以简要描述要解决的问题及其重要性开始，包括生物学、诊断、预后和公共卫生重要性（包括患病率和发病率）等方面的信息。

干预措施描述 ［推荐，二级标题］

对试验组干预措施的描述应放在任何标准或备选的干预措施背景中。应该明确对照组干预措施在规范的临床实践中的作用。对于药物，如可能，应该说明其临床药理学的基本信息。该信息可能包括剂量范围、代谢途径、选择性效应、半衰期、作用时间以及与其他药物的已知交互作用。对于更加复杂的干预措施，应提供其主要组成成分。

干预措施的作用机制 ［推荐，二级标题］

这部分将描述系统评价的干预措施对系统评价中的受试者产生影响的理论依据，例如药物的干预措施与一定情况下的生物学因素有关。作者可能会参考许多经验性证据，例如相似的干预措施对于该人群也有作用或者同一干预措施对于其他人群有影响。作者

同样也会参考很多证明其可能有效的文献。

该系统评价的重要性 ［推荐，二级标题］

背景部分应明确陈述该系统评价的合理性并解释提出问题重要的原因。该部分可能也会提到为什么要做该系统评价及与一个普遍问题的更宽泛的系统评价关联情况如何。如果该版系统评价是早期系统评价的更新，那么写上“这是一篇首次发表于哪年更新于哪年的 Cochrane 系统评价的更新”将非常有用。可以增补一个关于早期版本主要发现的简要描述和更新评价的特殊原因。

目的 ［固定，一级标题］

该部分应以精确陈述系统评价的主要目的开始，理论上用一句话表示。可用以下形式表示：评估［干预措施或对照措施］治疗［某类人群、疾病或问题和特定背景］的［健康问题］的效果。紧随其后为针对不同研究对象、不同干预措施的比较或不同结局测量指标的一系列具体目的。此部分没有必要陈述特定的假说。

方法 ［固定，一级标题］

撰写计划书中的方法部分时应使用将来时态。因为 Cochrane 系统评价会随着新证据的积累而更新，撰写计划书中的方法部分时通常应假设检出的研究数量能满足研究的目的（即使在写时知道这并不属实）。

撰写系统评价中的方法部分时应使用过去时态，并应表述为达到当前系统评价的结果和结论所做的事情。鼓励系统评价作者引用他们的计划书以证明其存在。通常因为证据不足，系统评价并不能采用计划书中的所有方法。在此情况下，建议在题为“系统评价与计划书的不同之处”部分（见下）概述系统评价中未实施的方法，作为将来更新系统评价的方案。

系统评价纳入研究的标准 ［固定，二级标题］

研究类型 ［固定，三级标题］

这部分应陈述合格的研究设计，以及基于研究实施或其偏倚风险确定的纳入标准阈值。例如“所有随机对照比较”或“所有对结局评估实施盲法的随机对照试验”。排除特定类型的随机研究（如交叉-对照研究）应有正当的理由。

同样可见于：

- 研究设计类型的合格标准讨论见第5章（5.5小节）。

研究对象类型 ［固定，三级标题］

该部分应表述感兴趣的疾病或情况，包括对诊断、年龄组和环境等的任何限制。亚

组分析不应该写在这里（见“方法”部分的“亚组分析和异质性研究”）。

同样可见于：

- 研究对象类型的合格标准讨论见第5章（5.2小节）

干预措施类型 ［固定，三级标题］

该部分应界定试验组干预措施和对照组干预措施，可适当使用单独的子标题。应清楚标明对哪种对照感兴趣。应陈述药物剂量限制、频率、强度和作用时间。亚组分析不应该写在这里（见“方法”部分的“亚组分析和异质性研究”）。

- 同样可见于：干预措施类型的合格标准讨论见于第5章（5.3小节）

结局指标类型 ［固定，三级标题］

注意：结局指标不一定构成纳入研究标准的一部分。如果没有，就应在该部分清楚写明。但关注的结局指标不论是否构成纳入标准的一部分都应在此部分列出。

同样可见于：

- 结果类型讨论见第5章（5.4小节）
- 病人相关性结果重要性的进一步讨论见第11章（11.5.2小节）；病人报告结果的扩展讨论见第17章。

主要结局 ［推荐，4 级标题］

系统评价的主要结局通常应反映至少一项潜在获益和至少一项潜在危害，并且应尽可能少。如能检出合格的研究，我们通常期望系统评价能分析这些结果，并且系统评价的结论很大部分基于干预措施在这些结果中的效应。

次要结局 ［推荐，4 级标题］

非主要结局应列在此处。总结局数应尽可能少。

以下可供选择的标题（4 级）或许会有帮助，作为以上标题的补充或替代：“结果总结”表中的主要结果

结局测量的时机

不良结果

经济学数据

鉴定研究的检索方法 ［固定，2 级标题］

应总结研究的检索方法。以下为推荐标题。在开始写这部分之前，作者应与相应的 Cochrane 系统评价小组联系以获得指导。

同样可见于：

- 检索方法详见第6章（6.3小节）

电子检索 ［推荐，3 级标题］

检索的文献数据库、检索时间和时间段以及任何限制词（如语言）都应该写明。每个数据库的所有搜索策略都应列于系统评价的附录中。如果 CRG 已生成专业数据库并为系统评价进行了检索，那么可以参考该注册库的标准描述，但需列出最近何时且如何为当前版本的系统评价检索该注册库的信息，以及所使用的检索词。

同样可见于：

- 检索策略详见第6章（6.4小节）

其他资源检索 ［推荐，三级标题］

列出灰色文献来源，例如内部报导和学术会议论文集。如果某系统评价手工检索了期刊，则应在本部分标注出来；但如果该作者进行手工检索是为了建立 CRG 的专业数据库则不用列出，因为这包涵在注册库的标准描述中。列出联系的人员（如试验者、课题专家）和相关组织。列出所有的其它来源，如参考文献列表、万维网和个人收集的文章。

以下为可能用到的可选标题，要么列在“其它资源检索”（此时作为 3 级标题）或作为子标题（4 级）：

灰色文献

手工检索

参考文献列表

信件

同样可见于：

- 其它检索来源相关讨论见第6章（6.2小节）

数据收集和分析 ［固定，二级标题］

该部分描述数据收集和分析的方法。

研究筛选 ［推荐，三级标题］

该部分描述应用筛选标准的方法。应写明筛选标准是否由多个作者独立应用，并写明如何解决不同意见。

同样见于：

- 研究筛选讨论见第7章（7.2小节）

数据提取和管理 ［推荐，三级标题］

该部分描述从已发表的报道或原始研究作者处提取或获得数据的方法（如使用数据收集表格）。应表明数据的提取是否由多个作者独立完成和如何处理不同意见。如果有关联，应描述准备分析时处理数据的方法。

同样可见于：

- 数据收集详见第7章，包括收集何种数据（7.3小节）、数据来源（7.4小节）、数据收集表格（7.5小节）以及从报告中提取数据（7.6小节）

纳入研究的偏倚风险评估 ［推荐，3 级标题］

该部分描述偏倚风险（或方法学质量）的评估方法。应表明是否由多个作者独立应用方法和如何处理不同意见。应描述所用工具或参考的工具，并指出评估结果如何融入结果解释中。

同样可见于：

- 评估偏倚风险的推荐工具见第8章（8.5小节）

治疗效果测量 ［推荐，三级标题］

该部分应描述所选择的效应指标。例如对于二分类数据应用比值比（OR）、相对危险度（RR）或危险度差（RD）；对于连续数据用均数差（MD）或标准化均数差（SMD）。以下为可能用到的可选用标题，要么放在“治疗效果测量”部分（作为 3 级标题）或作为子标题（4 级）：

二分类数据

连续型数据

时间-事件数据

同样可见于：

- 数据类型和效果测量讨论见第9章（9.2小节）

分析单位问题 ［推荐，3 级标题］

描述分析非标准设计研究（如交叉对照试验和群组随机对照试验）时的特殊问题。也可选择特别针对研究类型的可选标题（三级），如：

整群随机对照试验

交叉对照试验

多个治疗组的研究

同样可见于：

- 分析单位问题讨论见第9章（9.3小节）

- 非标准设计试验的讨论详见第16章，包括整群随机对照试验（16.3小节），交叉对照试验（16.4小节），多个治疗组的研究（16.5小节）。非随机研究讨论见第13章。

缺失数据处理 ［推荐，3 级标题］

该部分应描述处理缺失数据的方法。原则上应包括试验中途退出的受试者（并且说明是否进行意向性分析），和缺失的统计数据（如标准差或相关系数）。

同样可见于：

- 缺失数据的相关问题讨论详见第16章（16.1小节）和意向性分析见第16章（16.2小节）

异质性评估 ［推荐，3 级标题］

该部分描述评估临床异质性的方法，并应说明作者如何决定进行 Meta 分析是合理的。也应描述鉴别统计异质性的方法（如直观图示、使用异质性指标 I^2、使用 X^2 检验）。

同样可见于：

- 异质性评估讨论见第9章（9.5小节）。

发表偏倚评估 ［推荐，3 级标题］

该部分将描述如何处理发表偏倚和报告偏倚（如：漏斗图、统计学检验，赋值）。作者应记住，不对称的漏斗图未必由发表偏倚造成（并且发表偏倚也不一定会造成漏斗图的不对称）。

同样也见于：

- 报告偏倚讨论见第10章

数据合成 ［推荐，3 级标题］

应描述选择了何种 Meta 分析方法，包括是否使用固定效应模型或随机效应模型。如果不采用 Meta 分析，应描述综合多个研究结果的系统方法。

同样可见于：

- Meta分析和数据合成讨论见第9章（9.4小节）

亚组分析和异质性探讨 ［推荐，三级标题］

该部分列出所有计划进行的亚组分析（或 Meta 回归的自变量）。其他探讨异质性效应的方法也应描述。

同样可见于：

- 异质性探讨见第9章（9.6小节）

敏感性分析 ［推荐，三级标题］

描述针对系统评价过程中的决策变化，系统评价结论稳定性的分析方法。如从 Meta 分析中纳入或排除某个特殊研究、缺失数据赋值或选择结果分析方法。

同样见于：

- 敏感性分析见第9章（9.7小节）

以下为对“方法”部分可能有帮助的可选标题（3 级）：

经济学问题

进一步更新的方法

系统评价中作者若要包括干预措施的经济学方面，就需要从计划书形成的早期阶段考虑经济学问题。

同样可见于：

- 经济问题讨论见第15章
- 更新系统评价的问题讨论见第3章

结果 ［固定，1 级标题］

研究描述 ［固定，2 级标题］

检索结果 ［推荐，3 级标题］

这部分应以总结检索结果开始（如电子检索到的参考文献数量，以及经筛选后可能合格的数量）。

同样见于：

- 检索结果描述讨论见第6章（6.6小节）

纳入研究 ［推荐，3 级标题］

清楚描述纳入研究数量非常必要。该部分应简明概括“纳入研究特征”表中的信息。该表格中应包括每个纳入研究的参考文献。应描述纳入研究的重要特征，包括研究的受试者、研究地区（如国家）、研究环境（如果重要）、干预措施、对照、结局指标以及任何重要的研究间的不同之处。受试者的性别和年龄范围应清楚描述，除非他们的特征非常明显（如所有的参与者都是孕妇）。应提供干预措施的重要细节（如放射治疗应总结其总剂量、粒级数、使用射线种类；对于药物治疗，或许应总结药物的制备、用药途径、剂量和频率）。作者应注意到他们认为系统评价的读者应该知道的其它重要研究特征。以下为可能用到的可选子标题（4 级）：

设计方法

样本量

研究环境

受试者

干预措施

结果

同样可见于：

- “纳入研究特征”表详细讨论见4.6.1小节

排除研究 ［推荐，3 级标题］

这部分应参考“排除研究特征”表的信息。该表格应包括排除的每个研究的参考文献，并应提供一个这些研究被排除的简要总结。

同样可见于：

- “排除研究特征”表的详细讨论见4.6.3小节

以下为可能会在“研究描述”部分用到的可选标题（3 级）：

在研的研究

待分类研究

更新中发现的新研究

纳入研究的偏倚风险 ［固定，2 级标准］

该部分应总结纳入研究结果存在的总偏倚风险，不同研究间的变异性和单个研究存在的重要缺陷。评估偏倚风险的标准应在“方法”部分描述或参见“方法”部分，而非在此处描述。每个研究是否符合每一条标准的情况应在“偏倚风险”表中报告，而不用在文中详细描述，此处只需要简要概括。

同样可见于：

- “偏倚风险”评估表见第8章（8.6小节）

对于大型的系统评价，可在下列标题下针对主要结局指标总结偏倚风险评估情况：

分组 ［推荐，3 级标题］

该部分应简要总结如何产生分配序列及隐藏干预措施分配的方法。并判断所采用方法可能产生的偏倚风险。

盲法 ［推荐，3 级标题］

该部分应简要概括研究实施和分析中谁是不知情方。针对不同结果，对结果评估者

采用盲法的含义是不同的，因此这些需要分别陈述。并应总结与盲法相关的偏倚风险评价。

不完整数据 ［推荐，3 级标题］

针对每个主要结果的不完整数据都应简要归纳于此。应报道评价者对排除的研究对象和过多（或特异）的中途退出的研究对象的担心。

选择性报道 ［推荐，3 级标题］

该部分应简要总结数据的选择性利用问题，包括选择性报告结果、时间点、亚组或分析的证据。

其他潜在的偏倚来源 ［推荐，3 级标题］

该部分应总结其他任何应关注到的偏倚。

干预措施效应 ［固定，2 级标题］

该部分应总结系统评价中涉及的干预措施效应的主要发现。在这里应直接强调系统评价的目的而非依次列出纳入研究的发现。每个研究的结果及其统计总结都应放在“数据和分析”表中。结果一般应按照其在“结果测量类型”中的顺序列出。为使读者更容易理解，鼓励使用子标题（例如：对于每个不同参加组，评价中的多个研究结果比较或结果测量）并应报道所进行的敏感性分析。

作者应避免在本部分做出推论。在描述结果或做出推论时，常常需要避免的一个错误是不要混淆“没有证据证明其有效”和“有证据证明其无效”。当存在不确定证据时，声称干预措施没有作用或声称干预组和对照组没有区别都是不正确的。在这种情况下，比较安全的方式是报告数据和可信区间，这与结果减少或增加相一致。

同样可见于：

- 结果表述见第11章（11.7小节）
- 数值结果解释讨论见第12章（12.4，12.5，12.6小节）

讨论 ［固定，1 级标题］

结构式讨论有助于理解系统评价的含义（Docherty 1999）。

同样可见于：

- 结果详见第12章

主要结果总结 ［推荐，2 级标题］

该部分总结主要结果（不要重复“干预措施效果”部分）和显著的不确定性、平衡重要的利弊关系。明确参考“结果总结”表格。

总体完整性和证据的适用性 ［推荐，2 级标题］

该部分描述证据与系统评价问题的相关性，这有助于判断系统评价的外部真实性。筛选出的研究是否足以回答所有的系统评价目的？是否研究了所有相关的受试者、干预措施和结局指标类型？尽管作者应牢记全球的临床实践情况并不一致，但此处仍应包含系统评价的结果如何与当前实践背景相适应的评论。

证据质量 ［推荐，2 级标题］

针对系统评价的目的，能从鉴定出的证据中作出一个强有力的的结论吗？该部分应总结纳入证据的数量（包括研究和受试者数量），陈述研究主要的方法学局限性以及重申研究结果的一致性或差异性。这将有助于总体判断系统评价的内部真实性。

评价过程中的潜在偏倚 ［推荐，2 级标题］

本部分应阐述系统评价在防止偏倚方面的优势和局限性。这些因素或许在评价者的控制之内或之外。讨论可能包括鉴定出所有相关研究的可能性、是否获得所有相关的数据、所用的方法（如检索、研究筛选、数据收集和分析）是否会引入偏倚。

与其他研究或系统评价的异同 ［推荐，2 级标题］

该部分包括纳入研究如何与其他证据相适应的评论，并清楚陈述其他证据是否经过系统的评价。

作者结论 ［固定，1 级标题］

系统评价的主要目的是提供信息，而不是提供建议。作者的结论分为两部分：

对实践的意义 ［固定，2 级标题］

对实践的意义应尽可能实用和明确。它不应超出评价过的证据范围，并被系统评价呈现的数据证明是合理的。“没有有效的证据”不应与“有无效的证据”混淆。

对研究的意义 ［固定，2 级标题］

人们越来越多地用 Cochrane 系统评价帮助规划未来的研究。因此作者应尝试写出对此有用的东西。像“对实践的意义”部分一样，该部分内容也应基于可获得的证据，避免使用未在系统评价中纳入或讨论过信息。

作者在制作该部分时，或许可以以研究类型、受试对象、干预措施和结果作为构架，考虑研究的不同方面。作者应区分怎样做和报告研究与未来的研究该做什么的含义。例如：需要（Randomized control trial，RCT）而非其他研究类型、需要在某个特定主题的系统评价中更好的描述研究、或者需要常规收集某些特定的结局指标，应与当已有明确有效且恰当的有效治疗措施时而无需继续再选择安慰剂做对照、或需要比较特别指定的

干预措施、或者需要在特定类型的人群进行研究相区别。

该部分描写应尽可能清楚和明确，这非常重要。而包含很少或不包含具体信息的一般陈述如“今后的研究需更好的实施”或“需要更多的研究”对人们做决策用处甚微，应尽量避免。

同样可见于：

- 明确陈述结论的指南见第12章（12.7小节）

致谢 ［固定，1 级标题］

该部分用于致谢作者想要致谢的人或组织，包括未列在作者栏中的相关人员。还可能包括 Cochrane 系统评价以前的作者或资源支持，也可能包括 CRG 编辑部的贡献。致谢者需获得被致谢人许可。

作者贡献 ［固定，1 级标题］

此部分应描述计划书或系统评价当前合作者的贡献。应有一个作者作为该系统评价的保证人。在系统评价递交给 CDSR 发表前，所有作者应进行讨论并一致通过关于他们各自贡献的描述。当系统评价更新时，应检查该部分以确保其内容的准确和更新。

以下可能贡献改编自 Yank 等（Yank 1999）。这是推荐的方案，该部分应描述人们做了什么，而不是尝试将某人的贡献套入这些类别中。理论上，作者应将他们的贡献用自己的语言描述出来。

- 构思系统评价
- 设计系统评价
- 协调系统评价
- 收集系统评价的数据
  - 设计检索策略
  - 实施检索过程
  - 筛选检索结果
  - 组织获取全文
  - 根据合格标准筛选获取的全文评价文献质量
  - 提取文献数据
  - 联系文献作者获取额外信息
  - 提供文献的额外数据
  - 从未发表的论文中获取并筛选数据

- 系统评价的数据管理
  - 在 RevMan 软件中录入数据
- 数据分析
- 数据解释
  - 从方法学角度
  - 从临床角度
  - 从政策角度
  - 从用户角度
- 撰写系统评价（或计划书）
- 提供系统评价的综合建议
- 提供系统评价经费
- 开展当前系统评价的前期基础工作

声明利益冲突 ［固定，1 级标题］

作者应报告现在或过去与对系统评价结果有兴趣的任何组织或实体之间的关系，这些关系可能导致真实或可察觉的利益冲突。其他人认为可能会影响系统评价作者判断的境况包括个人冲突、政治冲突、学术冲突和其他可能的冲突，还有财政冲突。如果作者与系统评价中纳入的某个研究有关联，必须在此处说明。

同样可见于：

- 协作网关于利益冲突的政策总结见第2章（2.6小节）

财政冲突最该引起重视，应尽量避免，但如果确实存在，就必须报道出来。任何可能过度影响系统评价中判断（例如：关于研究的纳入和排除，评估纳入研究的有效性或结果的解释）的任何次要利益冲突（如个人冲突）均应报告出来。

如果不存在已知的利益冲突，也应明确陈述。例如表述“没有已知的利益冲突”。

计划书与系统评价的不同之处 ［固定，1 级标题］

有时系统评价中可能需要用到不同于原始计划书中描写的方法。这可能是因为：

- 处理特殊问题的方法未在计划书中详细说明；
- 计划书中的方法不能采用（如：由于应用该方法的数据不充足或确乏必要信息）；
- 需要改变以前计划书中的方法，因为发现了更可取的方法。

进行系统评价时改变计划书中的一些方法是可以接受的，但是应在该部分完整描述出来。该部分应总结随着时间改变，系统评价方法的主要有哪些改变。

- 指出在最初发表的计划书后决定的任何方法（如增加或改变结局指标；增加“偏倚风险”表格；增加“结果总结”表格）。
- 总结计划书中在当前系统评价中不能应用的方法（如因系统评价未鉴定出合格的研究，或因为没有研究符合研究前划分的特定亚组）。
- 说明从计划书到系统评价过程中任何改变的方法，陈述改变的时间并改变的合理性。这些改变不应受干预措施效果的发现的影响。应考虑任何方法的改变对于系统评价结论的潜在影响，并采用敏感性分析评估这些影响。

发表备注 ［固定，1 级标题］

发表备注信息会在 CDSR 的系统评价中列出，包括编辑注释和来自 Cochrane 系统评价小组的评价，例如编辑或评审人强调的问题被认为值得与系统评价同时发表。应详细说明这些评论的作者或来源（如是来自编辑还是评审人）。

必须完成所有撤销的计划书或系统评价的备注并且给出撤销的原因。撤销的计划书和系统评价只发表其基本引文信息、资助来源和发表的备注。

4.6 表格

4.6.1 纳入研究特征

在“纳入研究特征”表格中每个研究有 5 个条目：方法、受试者、干预措施、结果和备注。当不能被上述分类涵盖时，最多可附加 3 个特定条目，例如：提供随访期信息、基金来源，或者不太可能直接引起偏倚风险的研究质量指标（见 4.6.2 小节：关于偏倚风险的信息收录）。表格中可使用编码或英文缩写以便在一个条目中清楚简明地表达多个信息；例如作者可将国家、研究地点、年龄和性别归于受试者这个条目。脚注可用于解释编码或英文缩写（这些都将发表在 CDSR 中）。

同样可见于：

- “纳入研究特征”表详细指导见第11章（11.2小节）

4.6.2 偏倚风险

尽管强烈推荐，但“偏倚风险”表格是可选项，它可作为“纳入研究特征”表的扩展。标准的“偏倚风险”表包括分配序列产生、分配方案隐藏、盲法、结果数据不完整、

选择性报告结果和“其他问题”。对于每个条目，该表提供

了关于研究中发生情况报道的描述和对防止偏倚的主观判断的描述(“是”用来表示偏倚风险低，“否”表示偏倚风险高，其余用“不清楚”表示)。

同样可见于：

- “偏倚风险”表格的讨论见第8章（8.6小节）

4.6.3 排除研究特征

应列出某些看起来符合纳入标准但被排除的研究并附上排除原因（如：对照干预措施不恰当)。排除情况应简要记录，通常一个排除原因就足够了。

同样可见于：

- 排除研究的选择的讨论见第7章（7.2.5小节）

4.6.4 待分类研究特征

“待分类研究特征”表（以前为“待评估研究”）与“纳入研究特征”表具有相同的结构。可用于两类研究：

- 因为当前不能获得足够的信息，某些研究不能确定纳入或排除。在系统评价发表前，所有合理的获取信息的方式均尝试后仍不能进行文章分类，才能将该文章列于此部分。但是我们也不能为了获取此类信息而过度延迟系统评价的发表，尤其是在该研究的分类并不影响到系统评价的结论时。当有关条目的信息不能获得时，应将其内容表示为“未知”。
- 已经鉴定出的相关研究，但在等待系统评价更新。特别是对系统评价结论有着潜在影响，提及该研究是非常适当的。或者该研究在系统评价更新期间受到广泛关注。总结在表中的该类研究可以帮助修正系统评价。随着这些研究的加入，应尽快完成系统评价的全面更新。当表格中某个条目的信息不能获得时，应适当应用“尚未评估”或“未知”来表示。

4.6.5 在研研究的特征

“在研研究的特征”表格针对每个研究有 8 个条目：研究名称、方法、受试者、干预措施、结果、开始日期、联系信息和备注。这些条目内容可与“纳入研究特征”表格

的内容类似。脚注用来解释表中所用英文缩写（这些都将发表在 Cochrane 系统评价数据库中）。

4.6.6 结果的总结

尽管强烈推荐，“结果总结”表格是可选项，用于陈述最重要结果，无论这些结果是否有支持它们的证据。“结果总结”表格包括证据数量总结、接受试验和对照措施的特有绝对危险度、相对危险度估计（如：风险比、比值比）、证据质量描述、评论和脚注。证据的质量评估应符合 GREAD 分级标准，应综合考虑偏倚风险、证据直接性、异质性、精确性和发表偏倚。

同样可见于：

- “结果总结”表格的详细说明和讨论见第11章（11.5小节）
- GREAD分级系统见第12章（12.2小节）

4.6.7 附加表格

附加表格用于收录不方便放入文章中或固定表格中的信息。例如：

- 背景支持信息
- 研究特征总结（例如干预措施和结果的详细描述）

同样可见于：

- 附加表格见第11章（11.6小节）

4.7 研究和参考文献

4.7.1 研究的参考文献

研究由 4 个固定标题组织起来。每个标题可以包含多个研究或没有研究。一个研究由“研究 ID”来鉴别，通常由第一作者的姓和研究的主要参考文献的发表年份组成。年份可与每个研究明确相关（通常是完成年分或其主要参考文献的发表年份），因此可将其作为随机对照试验编码的国际标准（ISRCTN）。另外，每个研究都应指定一个以下的“数据资源”范畴：

- 仅发表的数据；
- 发表和未发表的数据；
- 仅未发表数据；
- 仅含发表数据（查找了未发表数据但未使用）。

每项研究可有多篇参考文献，给予每篇参考文献像 Medline ID 或 DOI 之类的标识符。每个研究有一个参考文献被视为“主要参考文献”。为准确起见作者应检查每篇参考文献。

4.7.1.1 纳入研究

符合纳入标准并纳入系统评价中的研究。

4.7.1.2 排除研究

不符合纳入标准并排除的研究。

4.7.1.3 待分类研究

已经过鉴定的相关研究，但是因缺乏更多的数据或信息，不能评估为纳入研究。

4.7.1.4 在研研究

达到（或接近）纳入标准的正在进行的研究。

4.7.2 其他参考文献

非上述研究的参考文献划分为两类。注意 RevMan 软件也包含“未分类”，这便于在系统评价准备阶段管理参考文献。所有参考文献在系统评价完成准备递交 Cochrane 系统评价数据库前应移出此分类，因为待分类中保存的参考文献不会被发表。

4.7.2.1 附加参考文献

该部分应列出文章中引用的其他参考文献，包括在背景和方法部分引用的参考文献。如果一个研究报告在文章中因为某些原因被引用了而不是参考该研究（例如，因为参考文献中的某些背景和方法学信息），该文献既要列在此处也要在相关研究部分列出。

4.7.2.2 该系统评价的其他发表版本

在期刊、教材、Cochrane 系统评价数据库或其他地方发表的系统评价其他版本的参考文献应列于此处。

作者应检查所有参考文献以确保其准确性。

4.8 数据和分析

系统评价中纳入研究的结果按照层次排列：研究嵌套在（可选择）亚组中，亚组嵌套在结果指标中，结果指标嵌套在比较组中（见图 4.8.a）。在分析过程中，一项研究会被多次纳入。

RevMan 软件会自动根据“数据和分析”板块输入的数据自动生成森林图说明数据、效应估计值和 Meta 分析结果（选择时）。作者可控制是否进行 Meta 分析和怎样进行 Meta 分析。

注意：“数据和分析”应作为补充信息，因为它不会以某种格式出现在某些发表的系统评价中。主要的森林图（包括每个研究的数据）可能作为图形（见 4.9 小节）而被一直保留在系统评价的全文中。但是 Cochrane 系统评价数据库发表的系统评价则以系列森林图或表格的方式包含了所有“数据和分析”部分。

作者应避免列出无数据的比较或结果（如没有研究的森林图）。但是，作者应在系统评价正文中记录这些比较没有数据。但是如果系统评价有“结果总结”表格，不管从纳入研究中是否可获取数据，其主要结果应列于这个表中。

同样可见：

- 分析见第9章，包括比较组讨论（9.1.6小节），结果数据的类型（9.2小节）和亚组（9.6小节）。将报告的数据转换为需要的形式见第7章（7.7小节）

4.8.1 比较

比较应与“目的”部分的问题和假设相一致。

4.8.2 结果

可能有 5 种类型的结果：二分类数据、连续型数据、“O – E”和“V”型统计资料、通用倒方差（估计值和标准误）以及其他数据（只含文本）。

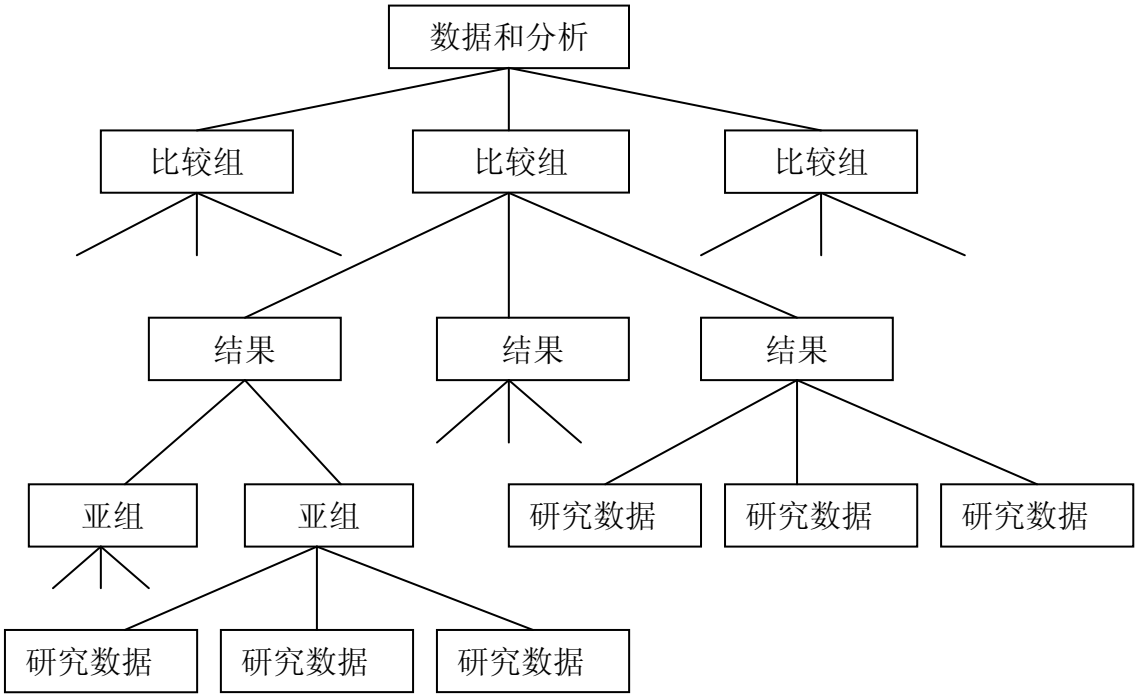
4.8.3 亚组

亚组与研究的亚类（如采不同疗程的物理疗法的试验）或结果的细分（例如短期、中期、长期）有关。

4.8.4 研究数据

由于结果数据的类型不同，每个研究的数据录入的格式也不同（如：连续性数据中每个组的样本量、均数和标准差）。

图形4.8.a “数据和分析”部分的层次图：



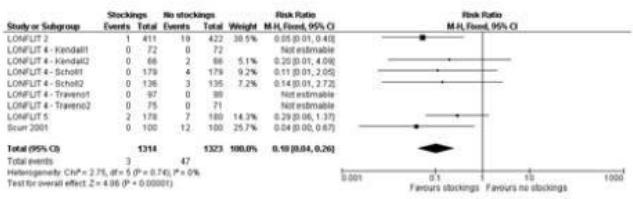
4.9 图形

系统评价正文中可能包含 5 种类型的图形（见表 4.9.a）。这些图形在系统评价的全文发表中会一直保留。每个图形必须有标题，提供简要图形描述（或解释），并且在系统评价的正文中必须提到图形（即建立图形与文字的相关链接）。

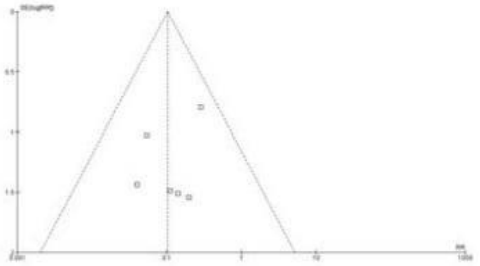
图形选择问题的讨论见第 11 章（11.4.2）章节：

表4.9.a　Cochran系统评价中所包含的图形类型

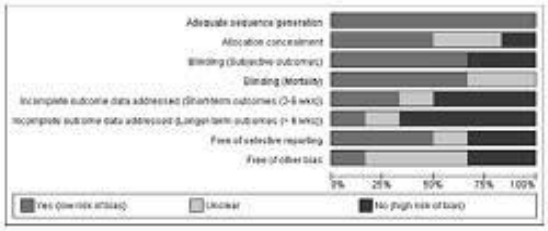
RevMan 森林图



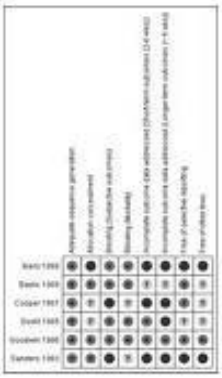
RevMan 漏斗图



RevMan “偏倚风险”图



RevMan “偏倚风险”总结



其他图表

4.9.1　RevMan软件的图和表格

在“数据和分析”部分的森林图和漏斗图可选为图形。关于偏倚风险的判断结果图示也可由 RevMan 软件自动生成并作为图形保存。

同样可见：

- 森林图的讨论见第11章（11.3.2小节）
- 漏斗图的讨论见第10章（10.4小节）
- “偏倚风险”图形和“偏倚风险”总结的讨论见第8章（8.6小节）

4.9.2 其他图形

非 RevMan 软件生成的图形和其他图像了可作为“图形”收录。但如果 RevMan 软件能以其它形式生成这些图形，则不能使用，例如作为森林图或附加表格。

作者负责获取系统评价中图像的使用许可权，并遵循指南确保图像符合发表要求。如果获得了有版权保护图像的许可，图像标题的最后必须附上：“版权 ©［年份］［版权所有者姓名或其他要求的词语］：允许复制。”

同样可见：

- 展示统计分析的图形应遵守统计方法学组的相关指南（补充信息可见于手册网页：www.cochrane.org/resources/handbook）.

4.10 支持系统评价的资源

作者应致谢对该系统评价的资助和其他形式的支持，如他们的大学或机构以工资形式的支持。支持资源分为“内部”（由生产系统评价的机构提供）和“外部”（由其他机构或基金委提供）。应提供每种支持来源的起源国家及其提供的支持方式。

4.11 反馈

系统评价的每个反馈由短标题和日期区分。“摘要”、“回复”和“撰稿人”为该部分的子标题。摘要由 Cochrane 系统评价小组的反馈编辑准备，如有必要，可咨询提交评论的人。系统评价作者应准备回复反馈。回复反馈意见人的名字应列于“撰稿人”下。

同样可见：

- 关于反馈的进一步信息见第3章（3.6小节）

4.12 附录

该部分提供补充信息，如：

- 详细的检索策略（推荐放置此处）；
- 非标准化统计方法的冗长细节；
- 数据收集表格；
- 结果详情（测量尺度）。

在某些发表的系统评价格式中不会出现附录。

4.13 本章信息

编辑：Julian PT Higgins and Sally Green.

本章引用格式：Higgins JPT, Green S (editors). Chapter 4: Guide to the contents of a Cochrane protocol and review. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Intervention. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.

致谢：本章建立在手册的早期版本上。想获取先前作者和编辑信息，请查看第一章（1.4小节）。列出的推荐标题由Julian PT Higgins与Mike Clark、Sally Hopewell、Jacqueline Birks、数个评价小组协调员、评估偏倚风险的工作组和手册顾问组成员讨论而成。最近更新中有贡献的作者包括Ginny Brunton, Mike Clarke, Mark Davies, Frances Fairman, Sally Green, Julian Higgins, Nicki Jackson, Harriet MacLehose, Sandy Oliver, Peter Tugwell and Janet Wale. 感谢Lisa Askie, Sonja Henderson, Monica Kjeldstrom, Carol Lefebvre, Philippa Middleton, Rasmus Moustgaard and Rebecca Smyth，他们提供了有益的评论。

4.14 参考文献

Docherty 1999

Docherty M, Smith R. The case for structuring the discussion of scientific papers. BMJ 1999; 318: 1224-1225.

Flanagin 1998

Flanagin A, Carey LA, Fontanarosa PB, Phillips SG, Pace BP, Lundberg GD, Rennie D. Prevalence of articles with honorary authors and ghost authors in peer-reviewed medical journals. JAMA 1998; 280: 222-224.

International Committee of Medical Journal Editors 2006

International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals: Writing and editing for biomedical publication ［Updated February 2006］. Available from: http://www.icmje.org （accessed 1 January 2008）.

Rennie 1997

Rennie D, Yank V, Emanuel L. When authorship fails. A proposal to make contributors accountable. JAMA 1997; 278: 579-585.

Rennie 1998

Rennie D, Yank V. If authors became contributors, everyone would gain, especially the reader. American Journal of Public Health 1998; 88: 828-830.

Yank 1999

Yank V, Rennie D. Disclosure of researcher contributions: a study of original research articles in The Lancet. Annals of Internal Medicine 1999; 130: 661-670.

（王凌译，岑啸、张龙浩初审）

第五章：立题与制定纳入研究标准

编辑：Denise O'Connor, Sally Green and Julian PT Higgins。

本文选自工作手册 5.0.1版本。有关如何引用它的指南，见5.8节。这些材料还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容摘要

- 一篇定义清楚、目的明确的系统评价始于精心构建的问题。在 Cochrane 系统评价中，问题被概括陈述为系统评价”目的”，并在“系统评价纳入研究的标准”中具体详述。
- 系统评价问题应指明人群（研究对象）类型，干预措施（和对照措施）类型，以及所关注的结局类型。为便于记忆简称 PICO，即研究对象（Participants），干预措施（Interventions），对照措施（Comparisons），和结局（Outcomes）。这些问题的要素，以及对纳入研究类型的规定构成了系统评价预先设定的纳入标准的基础。
- Cochrane 系统评价应包括所有可能有意义的结局，不包括不重要的结局。主要结局数目应很少且应包括有益及不利的结局。
- Cochrane 系统评价涉及的问题可以很宽泛，也可很局限，各有利弊。

5.1 问题与合格标准

5.1.1 精心构建问题的原理

与其他任何研究一样，制作系统评价首要和最重要的是确定研究的焦点，也就是清晰地勾勒出系统评价所要回答的问题。构思完善的问题将指导系统评价过程的多个方面，包括确定合格标准、检索研究、从纳入研究中收集数据和提交结果（Jackson 1980, Cooper 1984, Hedges 1994）。在Cochrane系统评价中，问题被概括陈述为系统评价的“目的”，并在“系统评价纳入研究的标准”中详细说明。这些部分除指导系统评价过程，也可作为读者初步判断系统评价是否与他们面临的问题直接相关。

陈述系统评价的目的应从明确陈述主要目的开始，最好用一句话完成。如果可能，可采用如下格式：“在（某类人群、疾病或问题，如指定，研究环境）中，评价（干预或比较措施）治疗（健康问题）的效果”，后面可以跟着一个或多个次要目的，如针对不同受试者、不同干预措施比较、或不同结局测量的问题。

详尽描述系统评价问题需要考虑几个关键要素（Richardson 1995，Counsell 1997）。“临床问题”应具体说明人群类型（研究对象）、干预措施类型（及对照措施类型）、感兴趣的结局类型，为便于记忆简称PICO（Participants, Interventions，Comparisons and Outcomes）。陈述时，并不需要对于每个PICO部分同样强调，例如，某系统评价关注治疗某期乳腺癌的干预措施的比较，则会明确定义疾病的分期和严重程度；反之若关注点是某特定药物治疗任何期的乳腺癌的效果，则需明确定义该治疗方案。

5.1.2 合格标准

系统评价区别于描述性综述的一个特点就是预先设定系统评价纳入和排除研究的标准（合格标准）。该标准由临床问题的各个方面与回答该临床问题的研究类型共同组成。临床问题中的受试者、干预措施和对照措施直接转化为系统评价的合格标准，而结局通常不作为纳入标准的内容。Cochrane系统评价通常搜寻针对特定人群比较特定干预措施的所有严谨的研究（如随机对照试验），而不在意其测量或报道的结局。然而，某些系统评价的确将合格标准合理限制为针对具体结局指标。例如，同一干预措施在同一人群中研究，但目的不同（如激素替代治疗或阿司匹林）；或某系统评价可能专门回答某干预措施治疗几种疾病时某的不良反应（参见第14章14.2.3节）。

在5.2-5.5节，我们将综述问题的要素和研究类型，用一些例子阐述在考虑各要素和制定合格标准以指导研究纳入时的问题。

5.2 确定受试者类型：哪些人和人群？

系统评价的人群纳入标准应足够广泛，以涵盖研究的多样性，同时又要足够局限，以保证纳入研究合并时能得到有意义的答案。常用以下两步去确定感兴趣的人群类型：首先，对感兴趣的疾病或临床情况应有明确的标准界定，以确定某个研究中是否有这些疾病或临床问题，但应避免采用某些标准强制性、不必要地排除研究，例如，新近建立的诊断标准可能被认为是当前诊断感兴趣临床问题的金标准，在以前的研究中没有采用；昂贵的或新的诊断试验在许多国家或地区都未采用。

其次，感兴趣的人群和临床问题应定义宽泛：这决定是否要研究某特殊人群，如年龄、性别、种族、教育程度、是否存在某一特殊症状如心绞痛或气促。还可针对病人就诊场所：如社区，医院，老人院，慢性病研究所或门诊。表5.2.a列出了制定“研究对象类型”标准时要考虑的因素。

感兴趣的研究对象类型通常直接决定纳入研究时与研究对象相关的纳入标准。但事先确定的标准在处理仅部分涉及感兴趣人群的研究时可能存在问题，例如，对于儿童来说，年龄切点设为16岁是合理的，但不能确定如何处理研究对象年龄为12-18岁的研究。如果不能得到研究中年龄的具体信息，而采用武断的规则（如“<16岁的研究对象超过80%”）将会不切实际，一句“多数研究对象年龄<16岁”可能就足够了。虽然存在系统评价作者偏倚影响事后纳入研究的风险，但符合系统评价研究目的的常识性规则可能比武断的规则更重要。入选标准的艰难决定应在系统评价中有记录，敏感性分析可评估这些决定对评价结果的影响（参见第9章9.7节）。

将研究限定于特定的人群或背景，应有合理的原理。保证Cochrane系统评价的全球相关性非常重要，因此，如果基于人群特征来排除研究，应在系统中阐述理由。例如，根据生物学依据、先前发表的系统评价与目前的争论，评估40-50岁妇女进行乳腺摄片筛查的效果是合理。另一方面，应避免毫无生物学或社会学根据而单凭个人兴趣按病人年龄、性别或种族将病人分类进行研究。如果对不同亚组病人，其疗效是否存在重要差异不肯定，最好将所有相关病人都进行研究，分析时再检测各亚组疗效是否存在重要的

貌似有理的差异（参见第9章9.6节）。这应在事先计划好，作为次要目的表述，并且不受可用数据的影响。

表5.2.a：制定“研究对象类型”标准时要考虑的因素

- 如何界定疾病/临床问题的？
- 什么是这些研究对象（受试者）的最重要特征？
- 是否有相关的人口学因素（如，年龄、性别、种族）？
- 什么环境（如，医院、社区等）？
- 应由谁作诊断？
- 有其它应该排除出系统评价的人群类型（因为他们可能对干预措施的反应不同）吗？
- 如何处理只包括相关受试者亚组的研究？

5.3 确定干预措施类型：与哪个对照？

第二个需精心构思的问题要素是：明确感兴趣的干预措施以及用来与之比较的干预措施（对照组），特别是，与干预措施比较的是无效对照措施（如安慰剂、不处理、标准治疗或等候名单中的对照）或有效干预措施（如相同干预措施的不同变化、不同药物、不同种类的治疗）？

如果是药物干预，那么制剂、给予途径、剂量、持续时间和给予频率等因素应加以考虑。对于较复杂的干预措施（如教育或行为干预），需明确规定这些措施共同的或本质的特点。通常需要仔细考虑给予什么样的干预措施？给予强度？多长时间给予一次？谁来执行？以及实施者是否需要培训等问题。评价作者还应考虑干预措施的可变因素（如基于剂量/强度、给予方式、频率、持续时间等）是否足以对研究对象、研究结果产生本质上不同的影响，以致需采取措施加以限制。

表5.3.a　制定“干预措施类型”标准时要考虑的因素

- 感兴趣的试验和对照（比较）措施是什么？
- 干预措施是否有变化（如剂量/强度、给予方式、实施者、频率、持续时间、干预时机）？
- 是否包括所有可变因素（如是否存在一个临界剂量，低于这个剂量干预措施可能不适于临床）？
- 如何处理只包括部分干预措施内容的试验？
- 如何处理感兴趣干预措施与其他干预措施联合（协同干预）的试验？

5.4　确定结局类型：哪个结局指标最重要？

5.4.1　列出相关结局

尽管结局报告很少决定研究入选系统评价的合格性，但良好构思问题的第三个重要部分是描述感兴趣的特殊结局指标。通常，Cochrane系统评价应包括所有对临床医生、患者（用户）、一般公众、管理者和政策制定者有意义的结局，但不必包括那些琐碎的或没有意义的结局。被认为有意义并因此要在系统评价中描述的结局未必在单个研究中出现，例如，对晚期癌症患者是否用化疗，生活质量是要考虑的一个重要并可能是最重要的方面，即使现有的研究仅有生存率报告（参见第17章）。在系统评价中涵盖所有重要结局将突出原始研究的不足，促使研究者在将来的研究中关注这些不足。

结局可包括生存率（死亡率）、临床事件（如卒中或心肌梗死）、患者报告的结局（如症状、生活质量）、不良事件、负担（如对护理的要求、测试频率、生活方式限制）和与经济有关的结局（如耗资与资源使用），关键是评估不良事件以及有益事件的结局指标均是系统评价该关注的（参见第14章）。如果考虑复合结局指标，需要特别说明。例如，如果一个研究没有区分非致死性和致死性卒中，而关注的问题与卒中死亡有关，这些研究的数据还应纳入Meta分析吗？

系统评价作者应从采用的量表和测量的时机两方面考虑结局是如何测量的，可采用客观指标（如：血压、中风次数）或由临床医生、患者或护理人员测量的主观指标，（如残疾评分量表）。阐述采用的测量量表是否已发表或者经过验证很重要。在定义结局测量的时间时，作者可考虑是否将所有测量时点或仅选择的时间点纳入系统评价中，策略

之一是将一组时间点划分为预先设定的时间段，表述为“短期”、“中期”和“长期”结局，并且对每个研究的任何一个结局采用其中之一。重要的是当结局的测量时间对系统评价的结果产生影响时，就应给予充分考虑（Gøtzsche 2007）。

系统评价再评价纳入了越来越多的Cochrane系统评价（参见第22章），对相关问题的系统评价采用一致的结局指标有利于此过程。作者在确定系统评价结果的测量方法和时点时，采用那些已在相关系统评价中采用的测量方法非常有帮助。另外，一些临床领域正在建立用于随机对照试验的统一核心结局指标测量方法，这对系统评价定义结局指标测量细节很有帮助。

列出相关结局指标的途径有多种，包括系统评价者的临床经验、来自用户和咨询小组的意见（参见第2章），以及文献中的证据（包括有研究对象认为重要的结局指标的定性研究）。更多有关使用定性研究来构思系统评价问题，包括结局指标的类型的信息见第20章。

Cochrane系统评价应涵盖所有重要的临床结局，但是琐碎的临床结局应除外。系统评价者应避免用几乎无意义或没有意义的数据淹没读者和误导读者。另外，间接或替代结局指标，如实验室结果或影像学结果（如以骨矿物质丢失做为激素替代治疗骨折的指标）有潜在的误导作用，应避免采用，解释时也应谨慎，因为这些替代结局可能无法正确预测重要的临床结局。替代结局指标可提供治疗措施如何起效的信息，但不能告之是否真正起效。许多干预措施降低替代结局的风险，但对临床相关结局没有任何效果甚至有害，而且有的干预措施对替代指标没有影响却可以改善临床结局。

5.4.2 优化结局指标：重要、主要和次要结局指标

5.4.2.1 重要结局指标

一旦列出系统评价关注的所有相关结局指标，作者应按重要性排序并选出与系统评价问题相关的重要结局指标。重要结局指标是对决策制定有本质作用的结局指标，并且为“研究结果总结”表的基础。“研究结果总结”表给出了重要的比较和结局指标的证据数量、证据的质量和效应的关键信息（参见第11章11.5节）。重要结果不要超过7个，通常不应包括替代或中期结局指标，这些指标的选择不应基于预期效应或观察到的效应、或因为这些结果会在将要评价的研究中涉及。

5.4.2.2　主要结局指标

系统评价的主要结局指标来自重要结局指标中，如果有相关研究，是预计在系统评价中要分析的结局指标，并且系统评价中干预措施效果的结论将主要基于这些结局指标 。主要结局通常不超过3个，应至少包括一个有利结局和一个不利结局（分别评价利和弊）。

5.2.2.3 次要结局指标

重要结局中未被选为主要结局的指标将被列为次要结局指标。另外，次要结局可能包括系统评价打算要讨论的数量有限的额外结果，这些结果可能针对系统评价的中仅一些比较。例如，实验室检查和其他替代指标不能作为重要结局，因为这些结局在指导决策方面的重要性不如临床终点指标，但有助于解释干预措施的效果或确定干预措施的完整性（参见第7章7.3.4节）。

表5.4.a总结了制定“结局指标类型”标准时需要考虑的主要因素

表5.4.a　制定“结局指标类型”标准时需要考虑的因素

- 纳入“研究结果总结”表中的重要结局是制定决策时必须的，通常应强调对病人重要的结局指标。
- 主要结局指标是重要结局指标中的两个或三个，如有足够的研究，系统评价可能会讨论这些结局，以得出干预措施效果（有利的和不利的）的结论。
- 次要结局指标包括剩下的重要结局指标（除了主要结局）加上额外有助于解释效果的结局指标。
- 确保结局指标包括潜在的和现有的不良结果。
- 考虑对所有决策制定者有影响的结局，包括经济学数据。
- 考虑结局指标的类型和测量时点

5.4.3　不良结果指标

Cochrane系统评价包括干预措施的不良和有益结局两方面的信息十分重要，系统评

价作者应认真考虑如何在系统评价中纳入不良结局的数据，并且至少应有一个不良结局被设定为主要结局指标。不良结局的评价将在第14章详细讨论。

5.4.4 经济学数据

决策者需要考虑一种干预措施经济学方面的问题，如这种方法的应用是否提高资源利用的效率，因此经济学方面的数据如资源利用、成本或成本效果（或这些数据的联合）应纳入作为系统评价的结局指标。针对不同类别或条目分解资源利用和成本指标非常有用。讨论成本问题应着眼于全球。经济方面问题的详细讨论在第15章中。

5.5 确定研究类型

在回答某个研究问题时，某些研究设计方案常优于其它设计方案。系统评价作者应预先考虑，什么样的研究设计能提供最可靠的数据回答系统评价的研究目的。

由于Cochrane系统评价是解决卫生保健效果的问题，因此主要关注随机临床试验。随机是防止不同干预组受试者的基线特征因已知及未知的（或无法测量）的混杂因素（参见第8章）引起的系统性差异的唯一方法。对于临床干预措施而言，有许多因素包括预后因素均可影响谁接受或者谁不接受干预措施。经验性证据提示，一般情况下，非随机研究得出的效应估计值往往高于随机对照试验。但是，偏倚的程度、甚至方向难以预测。这些问题在第13章详细讨论，并提供Cochrane系统评价何时纳入非随机研究合适的指南。

出于实际的考虑也促使许多Cochrane系统评价局限在随机对照试验。Cochrane协作网鉴定随机对照试验的工作量与鉴定其他类型研究的工作量无法相比，因此，系统评价若纳入除随机对照试验外的其它研究类型需要付出更多的精力去鉴定相关研究和保持系统评价更新，这样可能增加系统评价结果受发表偏倚影响的风险。确定纳入研究类型（例如：是否根据语种和发表状况限制研究合格标准）时需重点考虑的这个问题以及其他偏倚相关问题将在第10章详细讨论。

研究设计和实施的具体细节在设定入选标准时也应考虑，即使系统评价只纳入随机对照试验。例如，应确定是否纳入整群随机临对照试验（第16章16.3节）和交叉试验（第16章16.4节），同样也应根据诸如采用安慰剂对照、盲法评价结果、或最短随访期等方面确定研究入选的合格标准阈值。在严格的（这将导致入选的研究很少，但偏倚风险较小）

和宽泛的（这将导致入选的研究较多，但偏倚风险较大）研究设计标准之间总要进行权衡，然而过于宽泛的标准可能纳入误导性的证据。例如，如果重点关注一种治疗是否提高慢性疾病患者的生存率，那寻找较短周期的研究就不适合了，除非明确指出系统评价不能回答感兴趣的问题。

5.6 确定系统评价问题的范畴（广义与狭义）

从研究范畴来看，系统评价要解决的问题可以是广义或狭义的问题。例如评价者可提出一个广义问题：抗血小板制剂是否能有效防止人类的血栓形成事件。另一方面，评价者也可提出一个狭义的问题：抗血小板制剂阿斯匹林是否能有效降低既往有脑卒中病史的老年人发生某种血栓事件卒中的危险性。

系统评价的研究范畴由许多因素决定，包括研究问题的相关性和潜在影响力；理论上、生物学和流行病学的依据；回答该问题的潜在的普遍性和真实性；可利用的资源。

广义还是狭义的研究问题，有各自的优缺点，其中一些总结在表5.6.a。非常广义的系统评价的真实性可能因“苹果与橙子混为一团”受到批评，特别是生物学或社会学的依据提示不同治疗方案的作用不同，或者不同定义下的疾病干预措施的效果显著不同。

实际上，Cochrane系统评价可能以宽乏的范畴启动（或已启动），但随着证据的累积以及原系统评价变得难以处理时，可被分为狭义的系统评价。这可能出于实践和逻辑上的原因，例如，为了及时更新更容易也使读者跟上最新的研究结果更容易。系统评价者必须与他们的系统评价小组协商决定，将宽泛的系统评价分为重点突出的狭义系列评价是否合适，以及达到这一目的方法（见第3章，3.4.4节）。如果要进行重要的变更，例如将一个广义的系统评价分成一系列重点突出的狭义系统评价，每一个系统评价要发表新的计划书并需要清楚描述纳入标准。

Cochrane系统评价再评价的出现（第22章，第22.1.1节），由于可以将多个系统评价进行总结，可能影响系统评价范畴的确定。再评价可总结不同干预措施处理同一情况的Cochrane系统评价，或者同一干预措施处理不同类型的受试者的多个系统评价。人们越来越认识到，设计一系列范畴相对较窄的系统评价，再通过再评价总结他们的研究结果，可能更合适。

表5.6.a　广义与狭义系统评价问题的某些优缺点

	广义系统评价	狭义系统评价
受试者的选择 例如：糖皮质激素注射治疗肩关节肌腱炎（狭义）或者糖皮质激素注射治疗任何肌腱炎（广义）	优点： 全面总结证据 能评价结果在各类型受试者的普遍性 缺点： 可能制作系统评价再评价更合适（见第22章） 检索、数据收集、分析和撰写需要更多的资源 “苹果和橘子混为一谈”的风险（异质性）；解释可能困难	优点： 系统评价小组易处理；易于阅读 缺点： 证据可能稀少 结果可能对其他背景或人群不具有普遍性 可由系统评价作者确定范围以得到预期的结果
干预措施的定义 例如：监督下跑步治疗抑郁症（狭义）或者任何运动治疗抑郁症（广义）	优点： 全面总结证据 能评价干预措施结果在不同实施条件下的普遍性 缺点： 检索、数据收集、分析和撰写需要更多的资源 “苹果和橘子混为一谈”的风险（异质性）；解释可能困难	优点： 系统评价小组易处理；易于阅读 缺点： 证据可能稀少 结果可能不能推广到其他的干预措施方案 可由系统评价作者确定范围以得到预期的结果
干预及对照措施的选择 例如：闹钟预防尿床（狭义）或者预防尿床的干预措施（广义）	优点： 全面总结证据要 缺点： 可能难以处理，制作系统评价再评价更合适（见第22章） 检索、数据收集、分析和撰写需要更多的资源	优点： 对评价小组来说易处理 目的清楚，易于阅读 缺点： 如果未纳入再评价时，可能价值有限

5.7 研究问题的变更

研究问题应在系统评价启动之前的计划书中提出，但这些不能阻止对意想不到的问题的探索（Khan 2001）。系统评价是对已有的资料进行分析，受到所纳入研究的研究对象、场所、干预措施、结局指标和研究设计的限制。如果对与某研究问题相关的研究一无所知，是不可能提出可回答的问题的。很明显，随着系统评价的进行，资料的积累，有时需要对提出的问题做适当的更改。

随着对研究问题的深入了解，虽然可以对提出的研究问题进行修改，但重要的是要保证修改问题时不受偏倚的影响。数据导向的问题可基于假的结果产生错误的结论。任何对系统评价问题的修改均应在“计划书与系统评价的不同之处”中记录。应使用敏感性分析评估修改对系统评价结果的影响（见第9章9.7节）。当修改立题时，有必要回答下列问题：

- 修改立题的动机是什么？
- 是否受到某纳入研究结果的影响而改变立题？
- 改变立题后，检索方法是否仍恰当（尤其是已经完成的检索方法）？
- 资料的收集方法是否适合改变后的题目？

5.8 本章信息

编辑：Denise O’Connor, Sally Green and Julian PT Higgins.

本章引用格式：O’Connor D, Green S, Higgins JPT (editors). Chapter 5: Defining the review question and developing criteria for including studies. In: Higgins JPT, Green S (editors), Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.

致谢：本部分建立在早期手册版本基础上。有关先前作者和编者的详细信息见第1章（1.4节）。

5.9 参考文献

Cooper 1984

Cooper HM. The problem formulation stage. In: Cooper HM (editors). Integrating Research: a Guide for Literature Reviews. Newbury Park (CA): Sage Publications, 1984.

Counsell 1997

Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Annals of Internal Medicine 1997; 127: 380-387.

G.tzsche 2007

G.tzsche PC, Hróbjartsson A, Maric K, Tendal B. Data extraction errors in Meta-analyses that use standardized mean differences. JAMA 2007; 298: 430-437.

Hedges 1994

Hedges LV. Statistical considerations. In: Cooper H, Hedges LV (editors). The Handbook of Research Synthesis. New York (NY): Russell Sage Foundation, 1994.

Jackson 1980

Jackson GB. Methods for integrative reviews. Review of Educational Research 1980; 50: 438-460.

Khan 2001

Khan KS, ter Riet G, Glanville J, Sowden AJ, Kleijnen J (editors). Undertaking Systematic Reviews of Research on Effectiveness: CRD's Guidance for those Carrying Out or Commissioning Reviews (CRD Report Number 4) (2nd edition). York (UK): NHS Centre for Reviews and Dissemination, University of York, 2001.

Richardson 1995

Richardson WS, Wilson MS, Nishikawa J, Hayward RSA. The well-built clinical question: a key to evidence based decisions. ACP Journal Club 1995: A12-A13.

（何林译，岑啸、张龙浩初审）

第六章　文献检索

作者：代表 Cochrane 信息检索方法组的 Carol Lefebvre, Eric Manheimer 和 Julie Glanville。

本文选自工作手册 5.0.2版本。有关如何引用它的指南，见6.7节。这些材料还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 系统评价作者在系统评价初始就应当与他们的Cochrane系统评价小组（CRGs）的试验检索协调员（Trial searching coordinator，TSC）紧密合作。
- 纳入Cochrane系统评价的是研究（不是研究报告），但鉴定研究报告是目前尽可能多地检出研究、获取研究相关信息及结果的最方便方法。
- 临床试验注册库和试验结果注册库逐渐成为重要的信息源。
- 所有Cochrane系统评价应检索Cochrane中心对照试验注册库（CENTRAL）、MEDLINE和EMBASE（如果系统评价作者或TSC能进入上述数据库），这些检索或是直接检索，或是通过CRG的专业注册库。

- 检索应力求高敏感性，但同时可导致相对低的精确度。
- 应当避免太多的不同检索概念，但是各种各样的检索词应在同一概念之内用OR进行组配。
- 自由词和主题词都应该使用（例如医学主题词表（MeSH）和EMTREE）。
- 应用现有的高敏感性检索策略（过滤器）检索随机对照试验，例如用新修订的 Cochrane 协作网高敏感性检索策略检索 MEDLINE 中的随机试验（但不适用于 CENTRAL 数据库检索）。

6.1 引言

Cochrane系统评价小组（CRGs）负责为评价者提供系统评价相关的研究参考文献。大多数CRGs雇用一名专职的试验检索协调员来提供这项服务（参见章节6.1.1.1）。本章中的信息旨在帮助评价者开展补充检索和提供背景信息以便他们能更好地理解检索过程。在开始文献检索之前，评价者应与他们评价小组的试验检索协调员紧密联系，以了解他们的支持程度。

本章也对新上任的以及那些经验丰富的、查阅该章作为参考资源的试验检索协调员有用。

本章概述了文献检索的一般性问题；介绍了潜在研究的主要信息源；并探讨如何规划、设计以及执行检索策略，管理在检索中发现的文献，以及正确记录和报告检索过程。

本章主要探讨随机试验的检索。不过，讨论的许多检索原理也可用于其他研究设计。对于某些评价主题如复杂干预，可能需要采用其他方式纳入随机试验以外的研究。建议系统评价作者从系统评价小组寻求具体的指南，也可以参考本手册的相关章节，例如第13章介绍的非随机研究，第14章介绍的不良反应研究，第15章的经济学数据，第17章的患者报告结果，第20章的定性研究以及第21章讨论的健康促进和公共卫生。做Cochrane诊断性试验准确性系统评价时，系统评价作者应参考Cochrane诊断性试验准确性系统评价手册以检索纳入研究。

本章列出的众多网站在2008年6月已经核实过。

6.1.1 一般问题

6.1.1.1 试验检索协调员的任务

每一个系统评价小组的试验检索协调员都负责辅助系统评价作者检索纳入评价的研究文献，提供帮助的范围依据每个CRG的信息源有所不同，但大致提供以下部分或全部帮助：从CRG的专业注册库（详细参见章节6.3.2.4）中提供相关的研究，为主要的书目数据库设计检索策略，在CRG可得的数据库中执行这些检索策略，存储检索结果，并发给评价者，建议作者如何去检索其他数据库以及如何将检索结果导入文献管理软件（参见章节6.5）。开始文献检索之前请与你的试验检索协调员联系以确定他能提供的帮助程度。

如果一个CRG当前没有试验检索协调员，应当向当地检索经验丰富的医学图书管理员或信息专家寻求帮助，在他们的指导下进行系统评价的文献检索。

6.1.1.2 减少偏倚

干预措施的系统评价应全面、客观以及可重复的检索系列资源以找到尽可能多的相关研究（有限资源内）。这是区分系统评价与传统叙述性综述的关键因素，并有助于减少偏倚，从而达到干预措施效果的可靠评估。

仅仅检索MEDLINE数据库是不够的。有系统评价结果显示，仅有30%-80%已发表的随机试验能通过检索MEDLINE获得（根据领域或特定问题不同而有所不同）（Dickersin 1994）。即使在MEDLINE中有相关记录，要检索到它们也是很难的事情（Golder 2006，Whiting 2008）。非常有必要检索MEDLINE以外的数据库，不仅可以确保获得尽可能多的相关研究文献，同时也可以减少纳入文献时的选择性偏倚。完全依赖MEDLINE检索不能代表通过多种渠道全面检索所获得的所有文献。

由于受时间和经费的限制，要求评价者能够在检索的全面性与有效使用时间和资金之间达到平衡，而达到这一平衡的最佳方法是使用各种限定策略来检索文献，认识并尽量减少发表偏倚和语言偏倚（参见第10章的10.2节）。

6.1.1.3 研究与研究的报告

系统评价是将研究作为关注和分析的单位。然而，单个的研究包含一个以上的报告，并且每一个研究的报告对系统评价都是有用的信息（参见第7章的7.2节）。检索6.2节中

列举了大部分数据库资源获取的是单个研究报告，但是也有一些基于研究的数据源，例如试验注册库和试验结果数据库（参见章节6.2.3.1至章节6.2.3.4）。

6.1.1.4 版权和许可

Cochrane协作网的政策规定所有评价者和其它参与的合作者都应遵守版权法及数据库许可协议的条款。对于文献检索，尤其是在检索数据库和下载记录时应遵守数据库许可协议以及在拷贝文献时遵守版权法。评价者应通过试验检索协调员或者当地医学图书馆员获得指导，因为版权法和许可协议在各个地区和各个机构之间是有不同的。

6.1.2 要点总结

- 系统评价者在文献检索之前应从该评系统价小组（CRG）的试验检索协调员那里寻求建议。
- 如果CRG当前没有试验检索协调员，应向当地医学图书管理员或信息专家寻求帮助，在经验丰富的检索员指导下开展文献检索。
- 使用目录导航到本章的具体章节。
- 仅检索MEDLINE是不够的。
- Cochrane协作网的政策规定所有评价者和其它参与的合作者都应遵守版权法及数据库许可协议的条款。

6.2 检索信息源

6.2.1 书目数据库

6.2.1.1 书目数据库——概述

一般来说，检索卫生相关的书目数据库是获得一系列相关研究报告最容易且最省时的办法。有些书目数据库，如MEDLINE和EMBASE数据库，还包括多数最近年份文献的摘要。这些数据库的一个关键优势是既可以通过标题或摘要中的字词，也可通过分配给每个记录的标准化检索词和对照词汇进行电子检索。

Cochrane协作网已建立一个对照试验报告的数据库或登记库，即Cochrane中心对照

试验注册库（全称英文，CENTRAL）。这被认为是获得Cochrane系统评价合格试验的最好单一来源。普遍认为CENTRAL、MEDLINE和 EMBASE这三个数据库是检索试验报告最重要的信息源，有关的详细阐述将在随后的章节介绍。

个人可以通过付费、订购或在线付费的平台使用数据库。也能通过免费提供的国家节点、受许可的机构（如大学、医院）网站来使用、专业组织会员或者Internet上的免费资源使用数据库。

也有一些国际举措，提供免费或者低成本的数据库（与全文期刊）在线访问。卫生互联网访问研究倡议（HINARI）提供了多种数据库包括Cochrane图书馆和来自多种出版机构近4000种生物医学及相关的社会科学主要期刊，供超过100个低收入国家当地的卫生保健人员和非营利性机构使用。网址如下：

- www.who.int/hinari/en/

科学出版物国际网络（INASP）也提供包括Cochrane图书馆和期刊的多种数据库检索，不同国家可获得的期刊名称不同，详情见：

- www.inasp.info/file/68/about-inasp.html

图书馆电子信息联盟（eIFL）以图书馆联盟为基础，用于支持中欧、东欧以及东南欧转型期的50个低收入国家，还包括前苏联、非洲、中东以及东南亚的一些国家使用可负担得起的期刊，网址如下：

- www.eifl.net/cps/sections/about

有关如何检索这些以及其他数据库的详细信息请参见章节6.3.3和6.4.

6.2.1.2 Cochrane 对照试验中心注册库（CENTRAL）

Cochrane对照试验中心注册库（CENTRAL）是对照试验报告最全的数据库来源。CENTRAL是Cochrane图书馆的一部分且按季度更新。截至2008年1月（2008年第1期）CENTRAL中包含近53万条试验报告引文和其他一些可能符合纳入Cochrane系统评价的研究报告，其中31万条试验报告来自MEDLINE，5万条来自EMBASE，其余17万条来自手检文献和其他数据库。

CENTRAL中的许多研究报告来自系统检索MEDLINE和EMBASE，具体描述见章节6.3.2.1和6.3.2.2，但是，CENTRAL还包括未收录在MEDLINE和EMBASE或者其它书目数据库的对照试验报告；出版的引文有多种语言；只在会议论文集或其它难以访问的资

源中的研究报告引文信息（Dickersin 2002）。这也包括来自试验注册库和试验结果注册库的文献记录（参见章节6.2.3）.

Cochrane系统评价小组通过Cochrane图书馆免费检索CENTRAL。Cochrane图书馆网址是： http://www.thecochranelibrary.com. 许多卫生、学术机构和组织为他们的成员提供访问渠道，并且许多国家为全民开放（例如通过受资助国家的许可证或给低收入国家的特殊协议）。有关具体国家访问Cochrane图书馆的信息可以在Cochrane主页顶部的“访问Cochrane图书馆”项下找到。

6.2.1.3 MEDLINE 和 EMBASE

目前MEDLINE中收录了1950年以来的期刊文献超过1600万条。目前收录37种语言的5200种期刊，访问地址如下：

- www.nlm.nih.gov/pubs/factsheets/medline.html

PubMed提供了MEDLINE的免费检索版本，也包括没有被索引进MEDLINE的最新记录，地址如下：

- www.nlm.nih.gov/pubs/factsheets/pubmed.html

此外，PubMed包括来自未被MEDLINE收录的期刊记录，以及被MEDLINE部分收录的超范围期刊记录，有关MEDLINE和PubMed差异的更多信息参见以下网址：

- www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html

MEDLINE也可以从一些在线数据库厂商订购获得访问，例如OVID。访问通常是向订阅机构的成员免费开放（如医院和大学）。

美国国家医学图书馆（NLM）已经开发出NLM网关，它允许用户同时检索MEDLINE（或PubMed）以及NLM的其他资源，如卫生服务研究项目数据库（HSRProj），会议文摘和毒理学文献的毒理学在线数据库，地址见下：

- gateway.nlm.nih.gov/gw/Cmd

EMBASE目前收录1974年以来的1200多万条记录，目前收录30种语言的4800种期刊，网址见下：

- www.info.embase.com/embase_suite/about/brochures/embase_fs.pdf

EMBASE.com是Elsevier公司独家版权的EMBASE，除了的1974年以来的1200多万条EMBASE独家刊文献，也包含1966年以来MEDLINE独家收录的700多万条文献，因此

它可同时检索两个数据库，网址如下：

- www.info.embase.com/embase_com/about/index.shtml

2007年，Elsevier推出了EMBASE的经典库，提供EM从1947到1973年印刷版期刊（EMBASE创建的原始印刷刊）的电子检索，网址如下：

- www.info.embaseclassic.com/pdfs/factsheet.pdf

EMBASE数据库只有订阅后才能访问。作者应当核实一下自己的CRG是否可以访问，如果不行，是否可通过当地图书馆获得使用。

有关如何检索MEDLINE和EMBASE中试验报告的指导，分别参见章节6.3.3.2、6.4.11.1以及6.4.11.2.

数据库重叠

EMBASE中收录的4800种期刊，其中1800种不被MEDLINE收录。同时，MEDLINE收录的5200种期刊，1800种不被EMBASE收录，有关信息参见：

- www.info.embase.com/embase_suite/about/brochures/embase_fs.pdf

数据库中实际的重叠程度依据不同的主题而异，但是通过对数据库的比较研究，就某一主题通常认为需要对两个数据库进行综合检索（Suarez-Almazor 2000）。虽然MEDLINE和EMBASE检索不能得到同样的参考文献，但发现他们可以检索相似数量的相关文献。

6.2.1.4 国家和地区数据库

普遍认为MEDLINE和EMBASE是国际上重要的综合性医学文献资源数据库。此外，许多国家和地区生产的电子书目数据库集中了当地出版的文献，他通常包括了当地出版的期刊文献和其他文献，这其中有许多数据库可以在网上免费使用。其余的一些只有通过订阅或在线付费平台使用。索引的复杂性和连贯性以及检索界面的复杂性不同，但是他们可以提供MEDLINE和EMBASE等其他国际数据库未收录的重要期刊资源，一些例子包括在框6.2.a。

框6.2.a　地区电子书目数据库举例

Africa: African Index Medicus

- indexmedicus.afro.who.int/

Australia: Australasian Medical Index (fee-based)

- www.nla.gov.au/ami/

China: Chinese Biomedical Literature Database (CBM) (in Chinese)

- www.imicams.ac.cn/cbm/index.asp

Eastern Mediterranean: Index Medicus for the Eastern Mediterranean Region

- www.emro.who.int/his/vhsl/

Europe: PASCAL (fee-based)

- international.inist.fr/article21.html

India: IndMED

- indmed.nic.in/

Korea: KoreaMed

- www.koreamed.org/SearchBasic.php

Latin America and the Caribbean: LILACS

- bases.bireme.br/cgi-bin/wxislind.exe/iah/online/?IsisScript=iah/iah.xis&base=LILACS&lang=i

South-East Asia: Index Medicus for the South-East Asia Region (IMSEAR)

- library.searo.who.int/modules.php?op=modload&name=websis&file=imsear

Ukraine and the Russian Federation: Panteleimon

- www.panteleimon.org/maine.php3

Western Pacific: Western Pacific Region Index Medicus (WPRIM)

- wprim.wpro.who.int/SearchBasic.php

6.2.1.5　专题数据库

除了CENTRAL，MEDLINE和EMBASE外，还需检索哪些数据库受系统评价主题、专题数据库访问及经费预算的影响。大多数的专题数据库只能订阅或者在线付费平台使用。访问数据库因此受CRG编辑部的试验检索协调员可访问的数据库及系统评价者所在

研究机构拥有的数据库的限制。选择更可能通过机构订阅（并因此“使用免费节点”）或者通过Internet免费使用的专题数据库见框6.2.b以及进一步的网址信息。访问的细节，视不同机构而不同，评价者应当向当地医学图书馆员寻求他们机构的访问情况。

除专题数据库，通用搜索引擎包括：

- Google Scholar (free on the internet):
  scholar.google.com/advanced_scholar_search?hl=en&lr=
- Intute (free on the internet):
  www.intute.ac.uk/
- Turning Research into Practice (TRIP) database (evidence-based healthcare resource) (free on the internet):
  www.tripdatabase.com/

框6.2.b　专业电子数目数据库举例

Biology and pharmacology • Biological Abstracts / BIOSIS Previews: o biosis.org/ • Derwent Drug File: o scientific.thomson.com/support/products/drugfile/ • International Pharmaceutical Abstracts: o scientific.thomson.com/products/ipa/
Health promotion • BiblioMap - EPPI-Centre database of health promotion research (free on the internet): o eppi.ioe.ac.uk/webdatabases/Intro.aspx?ID=7 • Database of Promoting Health Effectiveness Reviews (DoPHER) (free on the internet): o eppi.ioe.ac.uk/webdatabases/Intro.aspx?ID=2
International health • Global Health: o www.cabi.org/datapage.asp?iDocID=169 • POPLINE (reproductive health) (free on the internet): o db.jhuccp.org/ics-wpd/popweb/
Nursing and allied health • Allied and Complementary Medicine (AMED):

- www.bl.uk/collections/health/amed.html

- British Nursing Index (BNI):
  - www.bniplus.co.uk/
- Cumulative Index to Nursing and Allied Health (CINAHL):
  - www.cinahl.com/
- EMCare:
  - www.elsevier.com/wps/find/bibliographicdatabasedescription.cws_home/708272/description
- MANTIS (osteopathy and chiropractic):
  - www.healthindex.com/
- OTseeker (systematic reviews and appraised randomized trials in occupational therapy) (free on the internet):
  - www.otseeker.com/
- Physiotherapy Evidence Database (PEDro) (systematic reviews and appraised randomized trials in physiotherapy) (free on the internet):
  - www.pedro.fhs.usyd.edu.au/

Social and community health and welfare

- AgeLine (free on the internet):
  - www.aarp.org/research/ageline/
- Childdata:
  - www.childdata.org.uk/
- CommunityWISE:
  - www.oxmill.com/communitywise/
- Social Care Online (free on the internet):
  - www.scie-socialcareonline.org.uk/
- Social Services Abstracts:
  - www.csa.com/factsheets/ssa-set-c.php

Social science, education, psychology and psychiatry

- Applied Social Sciences Index and Abstracts (ASSIA):
  - www.csa.com/factsheets/assia-set-c.php
- Campbell Collaboration's Social, Psychological, Educational and Criminological Trials Register (C2-SPECTR) (free on the internet):
  - geb9101.gse.upenn.edu/
- Education Resources Information Center (ERIC) (free on the internet)
  - www.eric.ed.gov/
- PsycINFO:
  - www.apa.org/psycinfo/
- Social Policy and Practice (evidence-based social science research):
  - www.ovid.com/site/catalog/DataBase/1859.pdf
- Sociological Abstracts:
  - www.csa.com/factsheets/socioabs-set-c.php

6.2.1.6 引文数据库

科学引文索引/科学引文索引扩展版数据库收录了大概6000种主要的科学、技术和医学期刊出版的文献并且链接到他们的引用文献上（特征是被引检索）。它可通过在线SciSearch和internet上的 Web of Science访问。Web of Science也整合入Web of Knowledge。它像MEDLINE数据库一样作为源文献数据库检索，也可通过检索已知的相关源文献和核查引用源文献的每一篇文献来为系统评价检索文献。这是一种从一篇重要文献发表开始而逐年追踪检索的方式。记录也包括源记录列出的参考文献，这又是相关试验报告的另一个可能来源。引文检索是数据库检索和手工检索的一个重要辅助（Greenhalgh 2005）。有关这些产品的信息见：

- scientific.thomson.com/products/sci/
- scientific.thomson.com/products/wos/
- isiwebofknowledge.com/

类似的社会科学数据库即社会科学引文索引：

- scientific.thomson.com/products/ssci/

2004年，Elsevier发行了文摘和引文数据库-Scopus，它涵盖了15000种期刊（其中超过1200种是开放获取的期刊）和500种会议论文集。Scopus包含了3300多万的摘要和近4亿科技网页数据结果：

- info.scopus.com/overview/what/

6.2.1.7 学位论文数据库

博士和硕士学位论文一般不被MEDLINE和EMBASE这样的书目数据库收录，但是也有例外，如CINAHL就收录了护理论文。为了检索相关发表在博士或硕士数据库中的研究报告，建议检索专题学位论文数据库：参见框6.2.c.

框6.2.c　学位论文数据库举例

ProQuest Dissertation & Theses Database: indexes more than 2 million doctoral dissertations and masters' theses:

- www.proquest.co.uk/en-UK/catalogs/databases/detail/pqdt.shtml

Index to Theses in Great Britain and Ireland: lists over 500,000 theses:

- www.theses.com/

DissOnline: indexes 50,000 German dissertations:

- www.dissonline.de/

6.2.1.8　灰色文献数据库

对灰色文献而言，存在多种定义，但是通常认为是未正式发表在图书或期刊数据库上的文献。已证明Cochrane系统评价中参考的文献大约有10%来自于会议摘要和其他灰色文献研究报告（Mallett 2002）。在最近更新的Cochrane方法学系统评价中，5个评价的研究都表明公开发表的试验报告的疗效都要优于灰色文献试验（Hopewell 2007b）。因此，未能检索会议论文集和其它灰色文献会影响系统评价的结果。

会议摘要是一个特别重要的灰色文献源，在章节6.2.2.4.中介绍。

EAGLE（欧洲灰色文献开发协会）已经关闭了SIGLE（灰色文献信息系统）数据库，后者是一个广泛使用的灰色文献数据库。法国的INIST（科学技术研究所）推出了OpenSIGLE数据库，他可提供所有之前SIGLE记录，EAGLE成员新增加的数据和Greynet提供的信息，地址：

- opensigle.inist.fr

医疗卫生管理信息协会（HMIC）数据库包含来自英国卫生部图书馆与信息服务部（the Library & Information Services department of the Department of Health全称DH）和国王基金信息与图书馆服务的记录。它包括通告和新闻在内的所有英国卫生部出版物。国王基金是一个独立的医疗慈善机构，致力于发展和改善医疗保健和社会卫生服务管理。该数据库被认为是下列主题的灰色文献良好来源，如保健和社区服务管理、组织发展、卫生资源的不均衡、用户参与以及种族和健康。地址如下：

- www.ovid.com/site/catalog/DataBase/99.jsp?top=2&mid=3&bottom=7&subsection=10

美国国家技术信息服务部（The National Technical Information Service， NTIS）提供美国和非美国政府资助研究的报告检索，且能提供大多数技术报告全文，从1964年起NTIS放在Internet上免费使用：

- www.ntis.gov/

PsycEXTRA是PsycINFO心理学、行为科学与健康的一个同伴数据库，它包括通讯、杂志、报纸、技术和年报以及政府报告和用户手册。PsycEXTRA和PsycINFO格式不同，因为它包括技术报告摘要、引文以及大多数记录的全文，它与PsycINFO无重叠覆盖：

- www.apa.org/psycextra/

6.2.2 期刊和其它非书目数据库源

6.2.2.1 手工检索

手工检索是指利用人工手动方式将每期杂志或者会议文献逐页进行检查，来寻找所有合格的试验报告。在杂志中，试验报告可能出现在期刊论文、摘要、消息栏、社论、来信或者其它文本中。至少有两个原因认为手工检索医疗卫生期刊和会议论文集是对电子数据库检索有用的辅助：1）并非所有的试验报告都收录在电子书目数据库中；2）即使收录了，在试验报告标题、摘要或者索引术语表中它们也未必包含相关的检索词（概念）能够让你轻松地把文献检索出来（Dickersin 1994）。每一年的期刊或者会议论文集不论题目都应该由受过良好训练的手工检索者进行全面细致手工检索，因此一旦手检过的期刊就不需要重复地检索了。一篇Cochrane方法学系统评价研究发现，手工检索联合电子检索共同检出所有相关发表在期刊中的试验报告是必要的，即使那些期刊中的试验文献已被收录进MEDLINE（Hopewell 2007a）。尤其是发表在1991年之前的期刊论文，那时MEDLINE中并没有随机试验的索引术语且部分期刊内容（如增刊和会议文摘）未被常规收录于MEDLINE中。

为方便查找所有已发表的试验报告，Cochrane协作网已通过Cochrane系统评价小组、专业领域和Cochrane中心组织了广泛的手工检索。美国Cochrane中心负责前瞻性注册所有的潜在手检，并保存手检活动文件包括手检的期刊总清单和会议论文集总清单（见apps1.jhsph.edu/cochrane/masterlist.asp）。超过3000种期刊已经或正在协作网手检。总清单能记录检索进程和监测检索进展，以防止同样的期刊或会议论文集被多个协作组或个人重复手检。

Cochrane实体和评价者可以根据他们预期能检出最多试验报告的领域优先开展手工检索。手工检索的优先顺序的形成可以根据检索某一主题领域内的CENTRAL, MEDLINE和EMBASE数据库后确定哪些期刊检出的引文最多而定。初步证据显示，收录试验报告多的期刊大部分都索引在MEDLINE中（Dickersin 2002），但这也反映一个事实，Cochrane作者早期的检索都集中在这些期刊上。因此，未被MEDLINE 或EMBASE收录的期刊也应考虑手检。

并不期望系统评价作者在制作系统评价时都进行手检，但是他们应该与评价小组的试验检索协调员讨论，针对他们的具体情况，手检某些期刊或会议论文集是否有益。希望手检期刊或会议论文体的评价者应当咨询该组的试验检索协调员以确定该期刊是否已经手检过，如果没有，他们可以在相关的总清单注册手检，并提供手工检索培训。培训材料可以从美国Cochrane网站上获得（apps1.jhsph.edu/cochrane/handsearcher_res.htm）。

所有关于期刊或会议论文集检索的启动、进程及状态信息应在CRG试验检索协调员和美国Cochrane中心员工之间协调。

6.2.2.2 全文电子期刊

越来越多的期刊可以通过订购或者Internet免费获得电子全文。除了提供方便方法去检索已收录的全文，全文期刊也能进行电子检索即依靠检索界面用书目数据库类似的检索方法在电子全文数据库中检索。

详细说明期刊全文是否已进行电子检索很重要，有些期刊在电子版中省略了印刷版的部分内容，例如信件，某些文章只在电子版中有。

大多数学术机构订阅的电子期刊比较广泛，并进而通过使用节点免费向其机构成员开放。系统评价作者应当向当地机构图书馆寻求访问电子期刊的建议。有些专业组织向其会员提供电子期刊访问。一些国家通过国家许可的协议向医疗卫生服务人员提供类似服务。也有一些国际行动，在互联网上提供免费或低收费的网上查阅电子全文期刊（和数据库），包括卫生互联网访问研究计划（HINARI）、科技出版物国际网络（INASP）以及图书馆电子信息网（eIFL）。有关更多的信息请参见章节6.2.1.1。

一些全世界无需订购可免费获得全文的期刊源例子见框6.2.d。

由于杂志的订阅可能不是永久性的，建议采用当地电子拷贝或纸质复印本，并将从订阅期刊中检索的可能相关文章存档。 该杂志可能停止出版或更换出版商和停止提供

以前的文章。这同样适用于互联网上免费提供的期刊，因为特定期刊的可获得情况也可能改变。

框6.2.d　全世界免费的全文期刊源举例

BioMed Central:
www.biomedcentral.com/browse/journals/
Public Library of Science （PLoS）:
www.plos.org/journals/
PubMed Central （PMC）:
www.pubmedcentral.nih.gov/
Web sites listing journals offering free full-text access include:
Free Medical Journals:
freemedicaljournals.com/
HighWire Press:
highwire.stanford.edu/lists/freeart.dtl

6.2.2.3　目录

许多期刊，即使那些仅通过订阅的期刊，都提供免费的目录（Table of Contents，TOC）服务，按时通过邮件提醒或者RSS反馈。此外，提供目录服务的一些组织见框6.2.e。

框6.2.e　提供OTC服务的组织举例

British Library Direct (free):
direct.bl.uk/bld/Home.do
British Library Direct Plus (subscription):
www.bl.uk/reshelp/atyourdesk/docsupply/productsservices/bldplus/
British Library Inside (to be replaced by British Library Direct Plus) (subscription):
www.bl.uk/inside
Current Contents Connect (subscription):
scientific.thomson.com/products/ccc/
Scientific Electronic Library Online (SciELO) – Brazil (free):
www.scielo.br/
Zetoc (Z39.50 Table Of Contents) (free as specified below) Zetoc provides access to the British Library's Electronic Table of Contents. It is free of charge for members of the Joint Information Systems Committee (JISC)-sponsored higher and further education institutions in the UK and all of NHS Scotland and Northern Ireland:
zetoc.mimas.ac.uk/

6.2.2.4 会议文摘或会议论文集

虽然会议论文集未被MEDLINE和一些专业数据库收录，但它们也会收录在BIOSIS（http://www.biosis.org/）数据库中。会议论文集中超过一半以上的试验报告都未发表全文，即使最终全文发表也与未发表全文的完全不同（Scherer 2007）。因此，重要的是设法通过专业数据库和手检鉴定可能相关的会议摘要或电子检索以印刷、光盘或互联网方式提供的会议摘要。许多会议论文集都以增刊的形式出版，专业会议摘要源列表见框Box 6.2.f。

许多会议摘要免费出版在互联网上，例如美国临床肿瘤学会（ASCO）：

- www.asco.org/ASCO/Meetings

框6.2.f 专业会议摘要源举例

- Biological Abstracts/RRM (Reports, Reviews, Meetings):
  - scientific.thomson.com/products/barrm/
- British Library Inside (to be replaced by British Library Direct Plus):
  - www.bl.uk/inside
- British Library Direct Plus:
  - www.bl.uk/reshelp/atyourdesk/docsupply/productsservices/bldplus
- ISI Proceedings:
  - scientific.thomson.com/products/proceedings/

6.2.2.5 其它系统评价、指南和参考文献目录作为研究来源

现有综述是获得潜在相关研究最方便和易见的参考文献源。获得感兴趣题目（或与其相关）先前发表的系统评价并检查其参考文献以纳入和排除的研究。Cochrane图书馆除包括Cochrane系统评价数据库（CDSR）外，还包括疗效评价文摘库（DARE）和卫生技术评估数据库（HTA ），后二者都是由英国约克大学的评价与传播中心（CRD）制作。两个数据库提供已出版的卫生保健疗效评价信息。Cochrane图书馆按季出版和更新，这些数据库更多最新的版本可以通过CRD的网站免费获得，网站更新频率更高。例如在2007年1月出版的Cochrane图书馆，DARE和HTA的记录是CRD中心的工作人员2006年11月提供的。2007年1月出版的Cochrane图书馆是直到2007年4月的当前主题，所以Cochrane

图书馆中DARE和HTA的记录比网站滞后2到5个月。

- www.crd.york.ac.uk/crdweb

CRD用于制作CRD正在进行的系统评价数据库，可通过UK国家研究注册库（NRR）检索，但只是存档至2007年9月，继续进行的系统评价记录已经转移到HTA数据库。

综述和指南也可以为其建立检索策略提供有用的信息：见框6.2.g特定循证检索服务如将研究转化为实践（Turning Research into Practice，TRIP）数据库可用来检索综述和指南。TRIP检索的系统评价源范围见：

- www.tripdatabase.com/Aboutus/Publications/index.html?catid=11
- www.guideline.gov

MEDLINE、EMBASE和其他书目数据库也可用于检索系统评价文章和指南。在MEDLINE中，最适合于系统评价文章的索引是1993年引入出版类型条目中的“Meta-analysis”，或者1966年引入的“Review”。指南应当用1991年引入出版类型条目中的“Practice Guideline”来检索。EMBASE也有同义词，“Systematic Review”在2003年引入，“Practice Guideline”在1994年引入。

在PubMed的临床提问检索链接中有一种叫做“系统评价”的检索策略或者过滤器：

- www.ncbi.nlm.nih.gov/entrez/query/static/clinical.shtml

在其广泛的范畴内检索的许多文献都不是系统评价。该策略描述如下：这一策略的目的是为检索系统评价、Meta-分析、临床试验系统评价、循证医学、会议共识、指南、以及对临床医生有价值的研究专业期刊引文。

- www.nlm.nih.gov/bsd/pubmed_subsets/sysreviews_strategy.html

已经建立的检索策略或过滤器被用于检索MEDLINE中的系统评价（White 2001，Montori 2005）和EMBASE（Wilczynski 2007）。在其他数据库中检索系统评价和指南的检索策略列在InterTASC信息专家小组检索过滤源网站上。

- www.york.ac.uk/inst/crd/intertasc/sr.htm

而且检索现有系统评价和Meta-分析、已检出研究的参考文献目录也可能用于进一步的研究检索（Greenhalgh 2005）。由于研究者可能选择性引用结果为阳性的研究，参考文献作为其他检索方法的补充应当谨慎使用（见第10章，10.2.2.3节）。

框6.2.g　循证指南范例

- Australian National Health and Medical Research Council: Clinical Practice Guidelines:
  - nhmrc.gov.au/publications/subjects/clinical.htm
- Canadian Medical Association – Infobase: Clinical Practice Guidelines:
  - mdm.ca/cpgsnew/cpgs/index.asp
- National Guideline Clearinghouse (US):
  - www.guideline.gov/
- National Library of Guidelines (UK):
  - www.library.nhs.uk/guidelinesFinder/
- New Zealand Guidelines Group:
  - www.nzgg.org.nz
- NICE Clinical Guidelines (UK):
  - www.nice.org.uk/aboutnice/whatwedo/aboutclinicalguidelines/about_clinical_guidelines.jsp

6.2.2.6　网络检索

使用一般的因特网搜索引擎如谷歌是否能检索可能相关的研究，几乎没有经验证据（Eysenbach 2001）。检索研究资助者和器械制造商的网站可能富有成效，检索制药企业网站可能是有用的，尤其是它们的试验注册库，见章节6.2.3.3。如果进行互联网搜索，建议评价者应该打印互联网上发现的任何相关研究文件的副本或者保存在本地文件夹，而不是简单地用“书籍标签”标记该网站，以防该试验记录被删除或后来变更。重要的是要记录访问网站的日期以作引用。

6.2.3　未发表和在研的研究

有些完成的研究从未发表。许多研究记载了有意义结果与发表的关系，总结在第10章（10.2）中。找出未发表的研究，并把合格和合适的研究纳入系统评价，这对减少偏倚很重要。尚无简单和可靠的方法获取已完成但从未发表的试验研究的信息。一系列的措施使这种情况正在好转。

启动了第一个在线服务的国际标准随机对照试验号注册计划，为全世界卫生保健各

个领域的随机对照试验提供唯一注册号，及随后的ClinicalTrials.gov （见章节6.2.3.1）；

研究者在试验之初注册试验的重要性越来越得到认可；

试验注册得到了世界顶级医学杂志出版商的支持，他们拒绝发表未正确注册的试验报告（De Angelis 2004，De Angelis 2005）；

美国国立卫生研究院（NIH）公共获取政策（见publicaccess.nih.gov/）在2007年12月之前都是自愿的，但现在要求“所有受NIH资助的研究人员，把他们最后经同行评价已接受出版的手稿电子版递交或已递交国立医学图书馆医学PubMed Central的时间不迟于正式出版日期后12个月。”

- publicaccess.nih.gov/policy.htm

同行可能是获取未发表研究信息的重要来源，有时非正式的沟通渠道是确定未发表数据的唯一方法。采用正式信件请求提供信息也能用来识别已完成但未曾发表的研究报告。这样做的一个办法是向纳入系统评价的研究报告的第一作者发送一个连同纳入标准的完整清单，询问他们是否知道任何相关的额外研究报告（发表或未发表的）。向这一领域感兴趣其他专家和制药公司或其他人发送相同的信件也可能是可行的。应该牢记的是，向研究者询问有关完成但未发表的研究报告信息并不总是富有成效的（Hetherington 1989，Horton 1997），尽管有些研究人员报告这是检索系统评价研究报告的一个重要方法（Royle 2003，Greenhalgh 2005）。有的组织建立系统评价项目网站，列出研究项目确定日期和要求提交未列出研究的信息。也有人建议立法，比如拥有信息自由法的英国和美国可能通过法律来获取有关未发表试验研究的信息（Bennett 2003，MacLean 2003）。

查找正在进行的研究同样重要，便于系统评价的更新时可能纳入。可能相关的正在进行研究的信息应当列在“正在进行的研究特征”表中（见第4章的4.6.5部分）。应意识到在研的相关研究也可能影响何时更新某个具体的系统评价。遗憾的是，尚不存在单个全面的、集中的在研试验注册系统（Manheimer 2002）。然而，一些组织包括代表制药公司和制药企业的组织，以国家或国际为基础致力于提供集中访问的在研试验以及某些已完成的试验结果。为改善这种情况，世界卫生组织（WHO）2007年5月推出国际临床试验注册平台检索入口以检索大范围试验注册库，与数年前由‘同期对照试验’（Current Controlled Trials） 发起的所谓Meta注册（ metaRegister）类似。目前（截止2008年6月）WHO入口仅提供检索三个主要的注册库（澳大利亚和新西兰的临床试验注册库，ClinicalTrials.gov和同期对照试验国际标准随机对照试验号注册库），但随着这一项目计划的进展预计将纳入其它的注册库。

6.2.3.1 国家或国际试验注册中心

框6.2.h列出国家和国际试验注册中心

此外，Drugs@FDA提供了有关美国自1939年以来批准的大部分药品的信息。对于最近批准（1998年以来）的药品，通常有一个“综述”，它包含了为新药品批准提供的科学分析基础。

- www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm

其他国家和地区的药品审批机构也可能是获取试验信息的有用资源。

框6.2.h 国家和国际试验注册中心举例

- The Association of the British Pharmaceutical Industry (ABPI) – Pharmaceutical Industry Clinical Trials database:
  - www.cmrinteract.com/clintrial/
- The Australian New Zealand Clinical Trials Registry:
  - www.anzctr.org.au/
- CenterWatch Clinical Trials Listing Service:
  - www.centerwatch.com/
- Chinese Clinical Trial Register:
  - www.chictr.org/Default.aspx
- ClinicalTrials.gov register:
  - clinicaltrials.gov/
- Community Research & Development Information Service (of the European Union) (trials and other research):
  - cordis.europa.eu/en/home.html
- Current Controlled Trials metaRegister of Controlled Trials (mRCT) – active registers:
  - www.controlled-trials.com/mrct/
- Current Controlled Trials metaRegister of Controlled Trials (mRCT) – archived registers:
  - www.controlled-trials.com/mrct/archived
- European Medicines Agency (EMEA):
  - www.emea.europa.eu/index/indexh1.htm
- German trials register – not yet launched. Final agreement reached 30 August 2007 – will be included under the WHO International Clinical Trials Registry Platform Search Portal – for further details as and when available see:
  - www.who.int/trialsearch

- Hong Kong clinical trials register - HKClinicalTrials.com:
  - www.hkclinicaltrials.com/
- Indian clinical trials registry - Clinical Trials Registry-India (CTRI):
  - www.ctri.in
- International Clinical Trials Registry Platform Search Portal:
  - www.who.int/trialsearch
- International Federation of Pharmaceutical Manufacturers and Associations (IFPMA) Clinical Trials Portal:
  - www.ifpma.org/clinicaltrials.html
- International Standard Randomised Controlled Trial Number Register:
  - www.controlled-trials.com/isrctn/
- Netherlands trial register (Nederlands Trialregister – in Dutch):
  - www.trialregister.nl/trialreg/index.asp
- South African National Clinical Trial Register:
  - www.sanctr.gov.za/
- UK Clinical Research Network Portfolio Database:
  - portal.nihr.ac.uk/Pages/Portfolio.aspx
- UK Clinical Trials Gateway:
  - www.controlled-trials.com/ukctr/
- UK National Research Register (NRR) (trials and other research – archived September 2007 – see UK Clinical Trials Gateway):
  - portal.nihr.ac.uk/Pages/NRRArchive.aspx
- University hospital Medical Information Network (UMIN) Clinical Trials Registry (for Japan)– UMIN CTR:
  - www.umin.ac.jp/ctr/

6.2.3.2 专题试验注册库

有很多专病试验注册库，尤其是在癌症领域—这不胜枚举。他们可以通过互联网搜索和检索上述资源发现，例如同期对照试验的对照试验Meta注册库（mRCT）。

6.2.3.3 制药企业试验注册库

通过一些制药公司的网站可获得他们的临床试验信息，替代或增加通过国家或国际网站（如上列）获得的信息。一些例子包括在框6.2.i.

框6.2.i　制药企业试验注册库举例

- AstraZeneca Clinical Trials web site:
  - www.astrazenecaclinicaltrials.com/
- Bristol-Myers Squibb Clinical Trial Registry:
  - ctr.bms.com/ctd/registry.do
- Eli Lilly and Company Clinical Trial Registry (also includes trial results)
  - www.lillytrials.com/
- GlaxoSmithKline clinical trial register:
  - ctr.gsk.co.uk/medicinelist.asp
- NovartisClinicalTrials.com:
  - www.novartisclinicaltrials.com/webapp/etrials/home.do
- Roche Clinical Trial Protocol Registry:
  - www.roche-trials.com/registry.html
- Wyeth Clinical Trial Listings:
  - www.wyeth.com/ClinicalTrialListings

6.2.3.4　试验结果注册库和其他资源

对已完成的试验结果进行注册是一个较新现象，下面仅从在研试验注册列举详细信息来说。他们具有特殊的价值，因为试验结果并不总是要发表，即使发表也并非全面公布。在最近的美国立法中，美国食品药品监督管理局（FDA）著名的修订法案2007的801条款（FDAAA 801），于2007年9月颁布，要求扩大ClinicalTrials.gov和增加一个临床试验结果数据库。试验结果注册的例子提供在框6.2.j.中

此外，临床试验结果（Clinical Trial Results）是一个寄存临床试验者报告临床试验结果的幻灯片的网站：

- www.clinicaltrialresults.org/

框6.2.j　试验结果注册库举例

International Federation of Pharmaceutical Manufacturers and Associations (IFPMA) Clinical Trials Portal:

www.ifpma.org/clinicaltrials.html

PhRMA Clinical Study Results Database:

www.clinicalstudyresults.org/about

Bristol-Myers Squibb Clinical Trial Results:

ctr.bms.com/ctd/results.do

Eli Lilly and Company Clinical Trial Registry:

www.lillytrials.com/

Roche Clinical Trials Results Database:

www.roche-trials.com/results.html

Wyeth Clinical Trial Results:

www.wyeth.com/ClinicalTrialResults

6.2.4　要点总结

Cochrane系统评价作者应当向其相关评价小组的试验检索协调员征求有关信息源检索的建议。

CENTRAL被认为是Cochrane系统评价纳入的试验报告的最佳来源。

CENTRAL，MEDLINE和EMBASE三个数据库通常被认为是检索纳入Cochrane系统评价的研究报告的重要书目数据库来源。

根据系统评价主题应选择性检索国家、地区和特定专题的数据库。

会议文摘和其他灰色文献也是系统评价纳入研究报告的重要来源。

其他系统评价、指南、已纳入（和排除）的研究和其他相关文献的参考文献目录也可作为进一步检索的来源。

应当努力去检索未发表的试验。

评价完成期间，应鉴定并追踪可能纳入的在研试验。

试验注册库和试验结果注册库是正在进行和未发表试验的重要来源。

6.3 规划检索过程

6.3.1 邀请试验检索协调员和卫生保健图书馆员参与检索过程

支持系统评价作者检索纳入研究报告是每个CRG小组的职责，大多数CRG小组聘请一位试验检索协调员来发挥这一作用（见章节6.1.1.1）。大多数CRG小组在系统评价作者选题规划初期就提供支持，直到完成系统评价并发表在CDSR中。这种支持可能包括设计检索策略或为设计提供咨询、运行检索策略，尤其是评价者所在其机构不能获得数据库，以及从CRG专业数据库和可能的其他数据库提供系统评价作者研究参考文献目录。根据现有可利用资源的不同各CRG小组提供的服务范围不尽相同。因此，鼓励系统评价作者在做系统评价的最早阶段积极与CRG小组的试验检索协调员联系以获得支持和建议。

如果系统评价作者自己检索时，他们应当向CRG小组的试验检索协调员咨询哪些数据库应当检索以及运行正确的检索策略。还应当牢记，自始至终需要足够详细地记录检索过程，确保在系统评价中正确地报告，最大程度使所有数据库检索具有可重复性。每个数据库的完整检索策略应包括在系统评价附录中。因此，重要的是：系统评价作者应当保存所有的检索策略，及时做笔记以确保检索部分在适当的时间完成。关于这部分的进一步指导，系统评价作者应与相关CRG组的试验检索协调员联系，和参见章节6.6。

如果当前CRG小组没有试验检索协调员，建议系统评价作者向当地卫生保健图书馆员或信息专家寻求指导，他们可能提供系统评价检索经验帮助。

6.3.2 协作网检索倡议

为了避免不必要的重复努力，在规划检索过程时应当考虑到哪些检索已经进行。例如，多年来相当大的努力都是在检索MEDLINE和EMBASE这两大主要国际数据库，以及将从中获得的试验报告输入Cochrane中心对照试验数据库（CENTRAL）。因此，必要的是任何一个特定的系统评价在做其他检索时都应考虑到以前都做过些什么。图6.3.a说明了CENTRAL的内容。

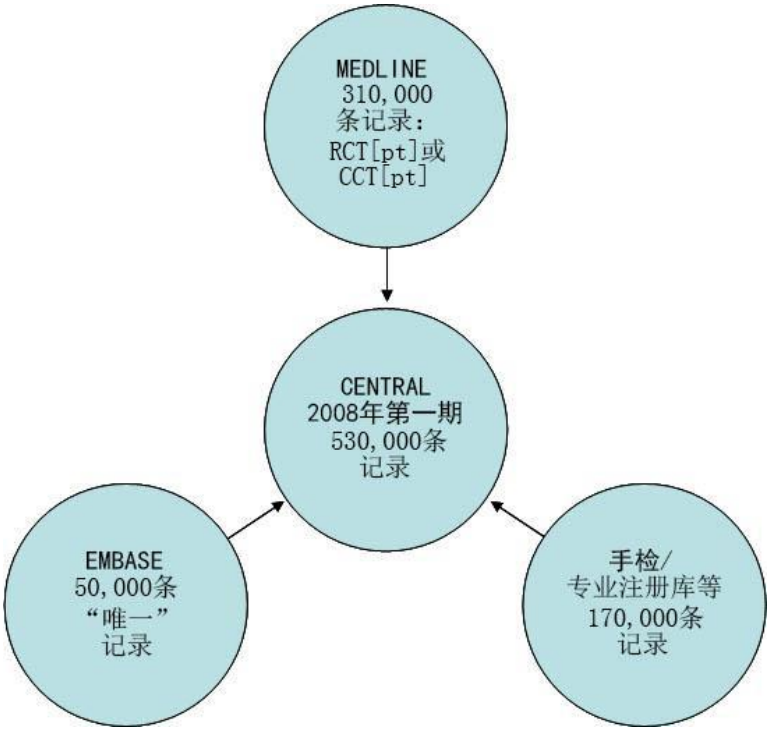


图6.3.a　图解CENTRAL

6.3.2.1　CENTRAL 中哪些来自 MEDLINE？

CENTRAL包含MEDLINE中所有出版类型以‘Randomized Controlled Trial’或‘Controlled Clinical Trial’索引并且是人类研究的文献记录。这些记录按季由Wiley-Blackwell从MEDLINE中下载，作为建立Cochrane图书馆出版内容的一部分，详细信息见：

- www3.interscience.wiley.com/cgi-bin/mrwhome/106568753/CENTRALHelpFile.html

MEDLINE中相当大比例的以随机对照试验（Randomized Controlled Trial）或临床对照试验（Controlled Clinical Trial）作为出版类型字段中的记录编码是Cochrane协作网编码工作的成果（Dickersin 2002）。Cochrane中心从MEDLINE收录的期刊中手检的结果已经送往美国国立医学图书馆（NLM），在那些记录的出版类型中重新标记为‘随机对照试验’（Randomized Controlled Trial）或‘临床对照试验’（Controlled Clinical Trial）。此外，美国Cochrane中心（前身为新英格兰Cochrane中心，Providence办公室和Baltimore Cochrane中心）和英国Cochrane中心已经对1996—2004年的MEDLINE进行了电子检索以检出可从文献的标题和/或摘要中确定的随机对照试验报告，但尚未在MEDLINE中索引为随机对照试验，使用的是1994年（Dickersin 1994）首次出版、后来更新并纳入Cochrane手册中的前两阶段Cochran高敏感检索策略。使用的自由词是：clinical trial；（singl$ OR doubl$ OR

trebl$ OR tripl$）AND（mask$ OR blind$）；placebo$；random$.其中$符号是截词符。以下主题词（MeSH）扩检：randomized controlled trials；random allocation；double-blind method；single-blind method；clinical trials；placebos。下面主题词（MeSH）未扩检：research design. 发表类型词为：randomized controlled trial；controlled clinical trial; clinical trial。

上列检索词一阶段和二阶段检索已进行，1994年Cochrane高敏感度检索策略词进行了一个三阶段测试，认为检索词的精确度太低，以致于不能保证这些检索词适合上述项目（Lefebvre 2001）。然而，也有人认为针对某些特定系统评价时，运用某些检索词与主题词组合进行检索可能是有用的（Eisinga 2007）。

上述文献检索限于人类。随后几年的检索由美国Cochrane中心（1966–1984；1998–2004）和英国Cochrane中心（1985–1997）完成。这些成果已提交给美国国立医学图书馆（NLM）重新标记在MEDLINE中，因而收录进CENTRAL。该项目目前被搁置。如果美国Cochrane中心能够为这个项目吸引到经费资助，他们将继续检索2005年及以后收录入MEDLINE的记录。这种情况的任何更新将被详细记录在Cochrane图书馆的CENTRAL创建详情文件中：

- www3.interscience.wiley.com/cgi-bin/mrwhome/106568753/CENTRALHelpFile.html

CENTRAL来自MEDLINE的试验报告不仅包括了满足Cochrane 严格定义的试验报告（Box6.3.a），而且也包括了满足初始美国国立医学图书馆（NLM）不太严格定义的试验报告（Box 6.3.b），其中包括历史性对照。无论是CENTRAL还是MEDLINE中，目前尚无更严格的区分方法区分哪些记录满足更严格的Cochrane试验报告的定义，因为它们都在出版类型中使用索引术语‘Controlled Clinical Trial’。

框6.3.a 随机对照试验（RCTs）及临床对照试验（CCTs）的Cochrane定义和标准

确定纳入的记录应当满足1992年12月制定并达成的纳入标准，其首次出版于1994年，在手册的第一个版本（见章节1，第1.4部分），根据这些合格标准基于可获得的最佳信息（一般来自一个及上出版报告）来判断其是否合格：

- 采用随机分配或一些半随机分配方法（如交替、出生日期、病例号）将试验中的个体（或其他单位）肯定或可能是前瞻性地分配到两种或以上卫生保健组中的一组。
- 通过读者对纳入试验合格者的分类来确定是否已进行随机分配形成对比的试验组，如果作者阐述明确（通常使用“随机”的一些同义词来描述分配过程）试验中的比较组是通过随机分配确定的，那么这个试验就分类为RCT（随机对照试验）。如果作者没有明确阐述试验是否随机，但是随机化又不能排除，那就分类为CCT（临床对照试验）。CCT的分类也适用于半随机研究，半随机研究的分配方法是明确的但并不认为是严格的随机，可能是半随机试验。半随机分配方法的例子包括交替、出生日期和病历号。
- RCT或CCT的分类仅仅是根据作者的描述，而不是读者的解释。因而，它并不是要反映一个真实性或分配过程的评价。例如，虽然‘双盲’几乎都是随机试验，许多试验报告没有明确提到随机分配，因此应当分类为CCT。
- 相关报告是指出版在任何一年、比较的至少是两种形式的卫生保健措施（卫生保健治疗及教育，诊断性试验或技术，预防性干预等等）的研究报告，且研究是在活人或身体的部分或者移植到活人身上的人类部分器官（例如，捐赠的肾脏）上进行的报告。尸体、拔除的牙、细胞系的研究等等是不相关的。检索者应确定所有对照试验符合这些标准，而不考虑他们与附属实体的相关性。

CENTRAL中可能包括了最高比例的卫生保健对照试验报告。因此，进行文献检索以鉴定试验时应报告出任何的疑问。系统评价作者将决定是否将某个报告纳入一个系统评价。

框6.3.b 美国国立医学图书馆2008年对出版类型条目为“随机对照试验”和“临床对照试验”的定义

随机对照试验

- 一个临床试验的组成至少包括一个试验组和一个对照组，两组人数同时入组并随访，且通过一个随机方案选择实施的治疗措施，例如使用随机数字表法。

临床对照试验

- 一个临床试验的组成包括一个或更多试验组、至少一个对照组和指定的结局指标以评价研究的干预措施，并采用无偏倚的方法分配病人到试验组。治疗措施可能是药物、器械、诊断性研究程序、治疗性或预防性效果。对照措施包括安慰剂、阳性药物、无治疗、剂量和方法、历史比较等等。当采用数学技术的随机化如随机数字表将患者分配入试验组或对照组，则该试验为“随机对照试验”。

6.3.2.2 Cochrane 对照试验中心注册库（CENTRAL）中哪些来自 EMBASE？

一个类似上述描述MEDLINE的研究表明，英国Cochrane中心已经完成了（Lefebvre 2008）检索EMBASE中没有索引进MEDLINE中的试验报告。（索引在MEDLINE中的试验报告已经纳入CENTRAL，有关阐述见章节6.3.2.1，因此，避免EMBASE重复记录作为检索过程的一部分。）下列术语目前用于该项目并已在1980年-2006年的检索中使用：自由词：random$；factorial$；crossover$；cross over$；cross-over$；placebo$；doubl$ adj blind$；singl$ adj blind$；assign$；allocat$；volunteer$；索引词，即EMTREE词：crossover-procedure；double-blind procedure；randomized controlled trial；single-blind procedure。下列自由词1974年至1979年的检索已完成：random$; factorial$; crossover$ and placebo$。这个$符号表示使用的是截词符。

这些检索已经获得8万条未能在MEDLINE中及时检索并索引的试验报告。在EMBASE出版商Elsevier和Cochrane协作网合约之下，目前这些试验报告均出版在CENTRAL中。8万条中的5万条是CENTRAL中独有的，即不包括在CENTRAL来源于MEDLINE中的记录。此检索按年更新，更新详细说明在Cochrane图书馆的CENTRAL创建详情文件中：

- www3.interscience.wiley.com/cgi-bin/mrwhome/106568753/CENTRALHelpFile.html

新内容章节在Cochrane图书馆主页：

- www3.interscience.wiley.com/cgi-bin/mrwhome/106568753/HOME

6.3.2.3 CENTRAL 中哪些来自其它数据库和手检？

其他如澳大利亚和中国出版的综合性卫生保健数据库已开展了类似的系统检索，为CENTRAL确定试验报告。澳大利亚Cochrane中心协调澳大利亚国立医学图书馆检索1966年以来的澳大利亚医学索引（McDonald 2002）。这个检索最近已经更新，并纳入记录到2007年。中国Cochrane中心在澳大利亚Cochrane中心的支持下，协调检索1999-2001年中国生物医学文献数据库。在英国Cochrane中心支持下，中国Cochrane中心正在进行一个项目，目的是检索大量中文数据源并将结果纳入CENTRAL中。同样，巴西Cochrane中心与巴西区域图书馆（BIblioteca REgional de MEdicina – BIREME）协作，正计划协调检索泛美卫生组织数据库拉美加勒比卫生科学文献（Latin American Caribbean Health Sciences Literature，LILACS）。

每一个Cochrane中心都有责任检索他们国家或地区的综合性医疗卫生文献。CRG小组和专业组有责任协调检索他们感兴趣的领域的专业卫生文献。超过3000期刊已经或正在进行手检。检出的试验报告若不符合CRG小组范围而不适合小组的专业数据库（见下），则作为手检结果提交给Wiley-Blackwell。手检记录可在CENTRAL检出，因为他们标记为HS-HANDSRCH或HS-PRECENTRL。

o www3.interscience.wiley.com/cgi-bin/mrwhome/106568753/CENTRALHelpFile.html

6.3.2.4 CENTRAL 中哪些是来自 Cochrane 系统评价小组和领域的专业注册库？

CRG小组的一个“基本核心功能”是“编辑小组应建立和维护一个包括所有他们感兴趣领域的相关研究的专业注册库，并将这些研究报告按季提交给CENTRAL”，作为“Cochrane系统评价小组的核心功能”在Cochrane手册的3.2.1.5节概述（www.cochrane.org/admin/manual.htm）。

专业注册库的作用是保证系统评价小组的评价作者容易和可靠的找到与他们评价主题相关的试验，通常通过试验检索协调员可以完成。CRG小组使用手册该章中描述的方法为他们的专业注册库检出试验。大多数CRG小组还建立了完备的系统以确保任何作者针对他们的系统评价所检出的额外合格报告能纳入CRG的专业注册库。反过来，该注册库按季提交并纳入CENTRAL。因此，每一个CRG包括在专业注册库的记录均可通过CRENTRAL被其他CRG检索。如上所述许多领域也建立自己的专题注册库并把他们提交给CENTRAL。想要从CENTRAL的一个专业注册库中找到相关记录可检索专业注册标记，例如SR-STROKE。所有的专业注册标记目录可以在Cochrane图书馆的“CENTRAL创建详细”帮助文件中的“附录：系统评价小组或专业/网络专业注册代码”中找到：

o www3.interscience.wiley.com/cgi-bin/mrwhome/106568753/CENTRALHelpFile.html

在CRG的专业注册库中的记录通常包含编码及不包括在CENTRAL中的其他信息，因此试验检索协调员常常能在他们的专业注册库中通过检索专业注册库中的编码找到不能在CENTRAL检出的额外记录。相反，用于管理专业注册库的书目检索或其他软件功能通常比Cochrane图书馆的检索功能简单些，因此通过检索CENTRAL来检索专业注册库记录不如检索专业注册库本身容易。因此建议分别检索CENTRAL和专业注册库来扩大检索。

6.3.3 CENTRAL，MEDLINE和MEDLINE检索：特殊问题

对所有Cochrane系统评价，如果可行的话：建议至少检索CENTRAL和MEDLINE以及EMBASE。

6.3.3.1 检索 CENTRAL：特殊问题

CENTRAL的记录来自多种数据源（见章节6.2.1.2和6.3.2以及亚章节），所以不存在记录格式或内容的一致性。

来自MEDLINE的310000条记录最好由主题词和自由词组合检索。其它记录，包括来自EMBASE的50000条记录，最好使用自由词全字段检索。

来自非MEDLINE及非EMBASE的大多数记录（在2008年第一期Cochrane图书馆中大约170000条）没有摘要及任何索引款目词。要检索这些仅有标题的记录，必须进行非常宽泛的自由词检索，对于检索整个CENTRAL来说过于宽泛。

通过检索CENTRAL中PubMed或EMBASE检索号识别来自MEDLINE和EMBASE的记录是可能的。然后从宽泛检索CENTRAL中将其排出也是可能的，如框6.3.c.所示

与检索CENTRAL相关的一般信息，见章节6.4.

框6.3.c 检索CENTRAL时排除MEDLINE和EMBASE记录举例

```
#1  ″accession number″ near pubmed
#2  ″accession number″ near2 embase
#3  #1 or #2
#4  tamoxifen
#5  (breast near cancer)
#6  #4 and #5
#7  #6 not #3
```

注：本例仅为说明的目的，针对这一主题的系统评价检索 CENTRAL 时还需要选择宽泛的他莫昔芬和乳腺癌的其他术语（款目词）。

6.3.3.2 MEDLINE 和 EMBASE 检索：特殊问题

尽管已经系统检索MEDLINE和EMBASE中的试验报告，且已经纳入CENTRAL是事实（章节6.3.2.1和6.3.2.2已述），仍建议补充检索MEDLINE和EMBASE。但任何这样的

检索应当考虑采用什么样的检索语言来避免重复劳动。

MEDLINE检索

索引在MEDLINE中的记录和作为试验报告索引在CENTRAL中大约有几个月的滞后，因为CENTRAL是按季更新。例如出版在2007年1月的Cochrane图书馆的内容，是由Wiley-Blackwell工作人员在2006年11月从MEDLINE下载的。2007年1月出版的Cochrane图书馆直到2007年4月才能更新，所以CENTRAL中的记录与MEDLINE时滞在2到5个月之间。因此对最近几个月的MEDLINE，至少应检索那些以‘随机对照试验’或‘临床对照试验’为出版类型索引的记录，以查找那些最近索引在MEDLINE中的RCTs或CCTs。

此外，在该项目的支持下，2004年是检索MEDLINE中的试验报告并发送到美国国立医学图书馆重新标记的最近年份，因此从2005年收入MEDLINE的记录的检索应当用章节6.4.11.1中检索策略之一。

最后，为了更加敏感的检索，或使用随机试验过滤器不合适时，系统评价作者应仅使用主题词检索所有年度MEDLINE。

应当记住在章节6.3.2.1描述的MEDLINE重标记项目仅基于标题和摘要评价被检出的记录是否是试验报告，因此任何补充检索MEDLINE是通过随后评价全文（最可能是通过方法学部分）而非仅通过标题或摘要辨别检出的额外试验报告。有关在MEDLINE索引版和包含“进行中”及其他未索引记录的MEDLINE版本中单独运行检索策略指南，请参考章节6.4.11.1。

系统评价作者检出的任何试验报告可发送给试验检索协调员以确保纳入CENTRAL。依据美国国立医学图书馆的定义（见章节6.3.2.1），MEDLINE索引为试验的记录若根据全文确定不是试验报告，任何这样的错误应向试验检索协调员报告，以便他们告知国立医学图书馆并且纠正。

有关检索MEDLINE的一般信息，见章节6.4。

EMBASE检索

检索EMBASE中的试验报告并纳入CENTRAL是以一年为基础进行，描述在章节6.3.2.2，因此出现在EMBASE与CENTRAL中的试验报告大约有一到二年的时差。因此，应当检索过去两年的EMBASE以覆盖正在进行的工作。一些建议的检索条目列在章节6.3.2.2中。一个由McMaster Hedges小组设计的检索过滤器也可用（Wong 2006）。

最后，为了更加敏感的检索，或使用随机试验过滤器不合适时，系统评价作者应当仅使用主题词检索EMBASE的所有年份，正如上述类似MEDLINE的情况。应当记住上

述描述的EMBASE项目，仅从标题和摘要基础上评价被检索的记录是否是试验报告，同上述MEDLINE项目描述的方式。因此任何补充检索EMBASE是通过随后评价全文（最可能是通过方法学部分）而非仅通过标题或摘要辨别检出的额外试验报告。

有关检索EMBASE的一般信息见章节6.4。

6.3.4 要点总结

Cochrane系统评价作者应当在整个检索过程中从试验检索协调员那里寻求建议。

作为最低要求，建议所有的Cochrane系统评价都应检索CENTRAL和MEDLINE，对CRG小组或系统评价作者而言，可行的话还应检索EMBASE。

每一个数据库的完整检索策略应当保存在系统评价的一个附录中，因此所有的检索策略应当保存，并且注释每一个数据库检索到的记录数。

CENTRAL包含超过350000条来自MEDLINE和EMBASE的记录，因此在检索MEDLINE和EMBASE时注意避免不必要的重复劳动。

应当从2005年开始使用修订和更新的Cochrane高敏感检索策略之一来检索MEDLINE中的随机对照试验，在章节6.4.11.1中描述。

应当检索最近两年的EMBASE。在章节6.4.11.2中描述。

检索CENTRAL已检索过的MEDLINE和EMBASE年份，通过获得全文并阅读，特别是方法学部分来获得额外的研究。

6.4 设计检索策略

6.4.1 设计检索策略-简介

本章重点介绍设计检索策略时需考虑的问题，但并不充分阐述这方面的许多复杂问题。特别是强烈建议有文献检索技巧的试验检索协调员或医疗卫生图书馆员的参与。许多下述突出的问题涉及检索方法（例如检索随机试验报告）和检索主题两方面，对于需要稳健检索者两方面都需要同等的重视，以确保不丢失相关记录。

系统评价中的研究纳入标准将告知如何进行检索（见第5章）。纳入标准中将阐明设计类型、受试者类型、干预类型（试验和比较）以及，在某些情况下，阐明结果类型。制定检索策略时需要考虑的问题如下：

- 系统评价是否仅限于随机试验或其他研究设计是否也将纳入（参见第13章）；
- 要求确定不良反应数据（参见第14章）；
- 待评估的干预措施的性质；
- 任何需考虑的地理因素，如中医药研究需检索中文文献；
- 评价这些干预措施已发生的时期
- 是否纳入未发表研究的数据。

6.4.2 检索策略架构

检索策略的架构应当基于系统评价确定的主要概念。对于Cochrane系统评价，标题应当提供这些概念和研究的纳入标准，将有助于选择合适的主题词和文本词以制定检索策略。

检索系统评价所涵盖的临床问题（通常指的是PICO——即患者（或受试者或人群），干预措施，比较措施，结果）的每一方面通常是不必要甚至是不需要的。虽然一个研究问题可能探讨特定人群、背景或结果，但这些概念在一篇文章的标题或摘要中难以很好地描述，往往没有很好的对照词汇术语索引。因此，将给完整检索带来困难。在综合性数据库中如MEDLINE，针对一篇系统评价纳入的研究报告的检索策略，典型的有三套术语：1）与感兴趣的健康状况相关的检索术语，例如人群；2）与待评价的干预评价相关的检索术语；3）与纳入的研究设计类型（通常是随机试验“过滤器”）相关的检索术语。然而，CENTRAL的目的是仅仅包含Cochrane系统评价要纳入的相关研究设计报告，因此检索CENTRAL不应使用试验过滤器。已经专为MEDLINE开发了检索随机试验和对照试验的过滤器，同时也提供了检索EMBASE的指南：见章节6.4.11和分节。对于复杂干预措施的系统评价，有必要采用一个不同的方法，如仅通过检索人群或干预措施（Khan 2001）。

6.4.3 服务提供商和检索界面

许多数据库服务提供商通过一系列检索界面提供MEDLINE和EMBASE；例如Dialog提供Dialog和DataStar。此外美国国立医学图书馆和Elsevier公司各自提供自己版本的MEDLINE和EMBASE：MEDLINE通过PubMed可以在互联网上免费检索，而EMBASE仅可通过订购在EMBASE.com利用。不同界面检索语法不同。例如，检索出版类型术语“Randomized Controlled Trial”通过不同的检索界面必须输入如下词：

randomized controlled trial.pt. (in Ovid)

randomized controlled trial [pt] (in PubMed)

randomized controlled trial in pt (in SilverPlatter)

许多服务提供商提供其他出版商网页的全文版链接，例如PubMed的‘Links / LinkOut’特征。

6.4.4 检索敏感度与精确性

系统评价检索的目标是尽可能宽泛全面，以便确保尽可能多地纳入重要和相关的研究报告。但是，在建立检索策略时必须尽力保持全面和相关的平衡。提高全面性（或敏感度）检索将降低精确性并检出更多非相关性文献。

敏感度是指检出的相关报告数量除以存在的相关报告总数。精确性是指检出的相关报告数量除以检出报告的总数。

检索策略的制定是在检索基础上反复修改检索术语（词）的一个过程。存在检索回报递减；一定阶段后，每一个额外投入检索的时间单元得到相关系统评价的参考文献越少。因而，存在一个不值得进一步检索额外的参考文献的点。在检索过程中决定投入多少时间应依据系统评价涉及的问题，CRG小组的专业数据库建立的程度，以及可利用的资源而定。但应当指出通过快速阅读检索得到的文献摘要可以确定潜在相关性。保守估计的阅读率是每分钟两篇摘要，因此数据库检索结果的浏览是每小时120篇（或大约1000篇需要超过8小时）。这样的话，与投入系统评价的总时间相比，系统评价检索的高产出和低精确性也就不会像乍一听起来那么令人生畏了。

6.4.5 受控词表和文本词

MEDLINE和EMBASE（和许多其他数据库）可以利用标准化学科术语分配的标准主题词进行检索。标准主题词（作为受控词汇或主题词表中的一部分）是有用的，因为他们提供了一个使用不同词汇描述同一概念的文献检索途径，同时他们能提供标题和摘要之外的文献信息。然而，当进行系统评价文献检索时，主题词应用的程度应当谨慎考虑。文献作者可能未很好地描述方法和目的，编索引的人员也并不总是他们所标引文献的主题领域或方法学方面的专家。此外，现有可利用的索引词可能不符合检索员要使用的词汇。

MEDLINE（MeSH）和EMBASE（EMTREE）的受控词汇检索不同，都不是索引的方法。例如，EMBASE制药学或药理学方面一条记录通常比同等的MEDLINE标引更深，并且近几年Elsevier公司增加了分配给每一条记录的索引词。因此，即使出现在两个数据库中的记录，检索EMBASE可能比检索MEDLINE获得更多的记录。需要为每个数据库定制检索策略。

开始为某一个特定数据库确定受控词汇的方法是从数据库中检索符合纳入系统评价标准的文章，并注意普通文本词和编索引的人员运用在文章中的主题词，然后可用于完整的检索上。确定一篇关键文章后，更多的相关文章可以通过使用Ovid的选项‘Find Similar’或PubMed的’Related Articles’找到。应当使用数据库提供的检索工具确定额外的受控词汇，例如Ovid的检索工具下的轮排主题索引和PubMed选项的MeSH数据库。

许多数据库叙词表提供主题词扩展检索以便在检索时自动包括更多的款目词。例如，使用MeSH词BRAIN INJURIES检索MEDLINE，如果扩检的话，不仅对BRAIN INJURIES，而且对更特异的词SHAKEN BABY SYNDROME将自动检索。MEDLINE中关于婴儿摇晃综合征的文章仅被标引成更特异的词SHAKEN BABY SYNDROME，而不是更为普通的主题词BRAIN INJURIES。因此，为了不漏掉相关的文献，在适当的地方扩检主题词是必要的。同样的原理也适用于EMTREE检索EMBASE以及其它数据库。有关这一主题的进一步指南，评价作者应当咨询它们的试验检索协调员或医学图书馆员。

特别重要的是MEDLINE中区别出版类型款目词和MeSH款目词。例如，一篇随机试验报告在MEDLINE的出版类型中将被标引成’Randomized Controlled Trial’，而一篇随机对照试验的文章在MeSH词索引中应标引RANDOMIZED CONTROLLED TRIALS AS TOPIC（注意后者是复数）。这同样适用于其它试验、综述（评价）和meta-分析的标引（索引）词。

系统评价作者应当假设早期的文章比最近的更难鉴定。例如，在1976年之前发表的文章MEDLINE中大多不包含摘要，因此，文本词检索仅适用于标题。此外，20世纪90年代以前MEDLINE很少有与研究设计相关的索引词，因此文本词检索早期记录是必要的。

为了尽可能多的检索到相关记录，应当从受控词汇表或叙词表中挑选主题词 （适当扩检）与宽泛的自由文本词组合检索。

6.4.6 同义词、相关词、不同拼写、截词和通配符

当设计一个检索策略时，应当尽可能全面，有必要为每一概念挑选宽泛的自由文本词。例如：

同义词：‘pressure sore’ OR ‘decubitus ulcer’，等

相关词：‘brain’ OR ‘head’，等

不同拼写：‘tumour’ OR ‘tumor’

服务提供商通过截词或通配符来捕获这些变化：

截词：random* （for random or randomised or randomized or randomly，等

通配符：wom?n （for woman or women）.

这些特征因提供商不同而不同。进一步的信息请参阅服务提供商的数据库帮助文件。

6.4.7 布尔运算符（与、或、非）

建立检索策略时应当包括每一概念的受控词汇、文本词、同义词和相关词，用布尔逻辑运算符OR将每一概念的每一术语组合起来：示范检索策略见框6.4.f。这意味着检出的文章至少包含这些检索词中的一个。建立术语集通常包括健康状况、干预措施和研究设计，这三套术语可以用运算符AND连接。最后一步用运算符AND连接这三套术语是：限制检出的文献为研究设计恰当且关注的是感兴趣的健康状况和待评价的干预措施。然而，有必要对种方法保持谨慎：如果一篇文章不包含来自这三套术语中的任何之一，将不会被检出。例如，如果一篇文献记录中没有加入干预措施索引词以及干预措施没有在标题和摘要中提及，该文献将漏检。一个可能的补救措施是省略三套术语之一，而基于检索数量和可用于检查的时间确定哪些记录需要检查。NOT运算符应避免无意中排除了相关记录。例如，在检索索引词female时，‘NOT male’将排除关于男女两性的记录。

Cochrane系统评价的检索式可能会非常长，通常超过100个检索表达式。检索设置中使用组合有时比较冗长，例如‘#1 OR #2 OR #3 OR #4 … OR #100’。有些服务供应商提供替代选择，例如Ovid提供使用语法‘or/1-100’。对于那些不可能提供的服务商，包括Cochrane图书馆检索CENTRAL，建议输入上述完整的字符串并保存为Word文档，根据需要再将必要数量的检索组合复制并黏贴来进行检索。如上用#符号录入字符串后，对不使用#符号的服务提供商，替补的方式是用查找替换方式删除全部的“#”符号，从而生成字符串‘1 OR 2 OR 3 OR 4 … OR 100’进行检索。

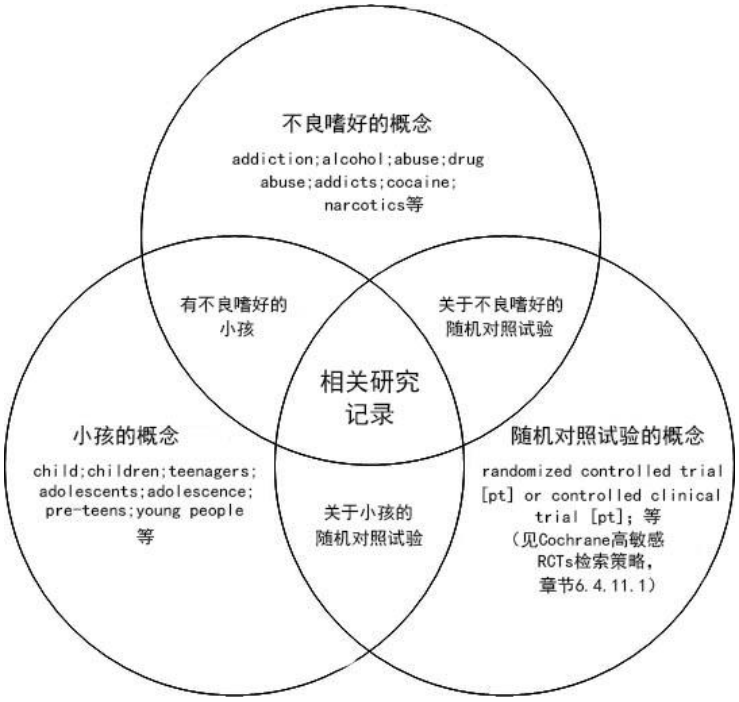


图6.4.a　检索设置时多概念结合

6.4.8　相邻运算符（NEAR，NEXT and ADJ）

在某些检索界面有必要指定，例如使用‘NEXT’或‘ADJ’运算符，两个检索词应当彼此相邻，犹如使用‘AND’运算符，以默认方式在文件中同时发现两个词。应当注意“NEXT”运算符在检索Cochrane图书馆时比使用引号搜索短语更敏感（如命中更多的记录），因为引号指的是确切词组，而‘NEXT’运算符包括自动多元和自动单元以及其他异性结尾词搜索。

此外，许多检索界面可能具体规定了检索词之间所能间隔的单词数量。例如，’NEAR’运算符在Cochrane图书馆中使用时能发现相邻6个词的检索词。这结果的敏感度比简单词组检索或使用’NEXT’运算符高，精确性优于‘AND’运算符。因此，如果可能或相关的话，最好使用这个运算符进行检索。

6.4.9　语言、日期和文献格式的限制

近来，与鉴定试验相关的研究主要关注比较Meta-分析中排除和纳入非英语语言发表的试验的效果。这个问题尤其重要，因为这些以非英语语言报告的试验的鉴定和翻译、

至少数据提取将大幅增加完成一个系统评价的费用和时间。对于这些问题的进一步讨论，请见第10章（章节10.2.2.4）。系统评价作者应尽可能检索和获取任何语言发表的、符合纳入标准的试验报告。检索策略不应当有语言限制。有关时间限制问题，除非知道某一研究仅仅在某个特定时间内才会涉及。例如，如果干预措施只是在某一时间点后才使用。不提倡限制格式，例如不提倡排除来信，因为来信可能包含重要的早期试验报告的相关信息或者尚未在任何地方报道的试验信息。

6.4.10 识别欺诈性研究、其它撤回发表物、勘误和意见

在考虑纳入Cochrane系统评价的合格研究时，重要的是必须意识到：已经发现有些研究自出版后因可能存在欺诈或其他原因已被撤回。在MEDLINE中有索引但已被撤回的研究报告（由于欺诈或其它原因）将在出版类型中增加‘Retracted Publication’标记。文章给予了撤销通知书，将在出版类型中标记为‘Retraction of Publication’。在决定收回一篇文章前，参考了一篇原始文章的任何文章都有可能发表，也会引起类似的关注。这类文章将被分类为评论。美国国立医学图书馆关于这类的政策是：“认为文章类型是评论的是：……报导科学性有问题或有学术不端行为的公告或通知（有时作为“表达关注”出版）”。

- www.nlm.nih.gov/pubs/factsheets/errata.html

此外，文章可能部分撤回，通过已发表勘误更正或已被纠正，再重新全文发表。当更新一个系统评价时，最重要的是要检索MEDLINE中纳入研究记录引文的最新版本。在某些MEDLINE版本的显示格式中有撤回出版、勘误和评论声明，并包含在引文数据中标题之后而引人瞩目。但并非总是这样，应注意通过下载合适的字段及引文数据确保在所有检索中可检索到这些信息（见章节6.5.2）。有关NLM政策的详细信息和这一领域的实践见：

- www.nlm.nih.gov/pubs/factsheets/errata.html

6.4.11 检索过滤器

检索过滤器是为检索特定记录设计的检索策略，例如特定的方法学设计。他们可能是主观来源的检索策略，例如为检索MEDLINE中随机试验的原Cochrane的高敏感检索策略（Dickersin 1994），或是客观通过词频分析和相关记录数据集测试来评价它们的敏

感度和精确性，如检索MEDLINE中随机试验检索策略（Glanville 2006）。最近英国InternerTASC信息专家小组（ISSG）建成了一个检索过滤器网站，该网站由支持英格兰和苏格兰研究组的信息专业人员构成，其为英国国家健康与临床优化研究所（NICE）提供技术评估（Glanville 2008）。该网站的目的是列出方法学检索过滤器，并对各种过滤器进行严格评价。此外，该网站包括了从一系列服务提供商的数据库中检索系统评价、随机和非随机研究以及定性研究的过滤器。

- www.york.ac.uk/inst/crd/intertasc/

应用搜索过滤器要谨慎。不仅要评估制作搜索过滤器的可靠性和报告的绩效，还要评估频繁的接口和影响数据库索引的条件下当前的准确性、相关性和效果。

ISSG提供一个搜索过滤器评价工具，有助于评价过滤器。相关范例见：

- www.york.ac.uk/inst/crd/intertasc/qualitat.htm

6.4.11.1 Cochrane 检索 MEDLINE 中随机试验的高敏感检索策略

Carol Lefebvre设计了第一个Cochrane检索MEDLINE随机对照试验的高敏感检索策略，并于1994年发表（Dickersin 1994）。这一策略随后出版在手册中，并随着时间的过去进行了必要的调整和更新。后续部分检索MEDLINE的Cochrane高敏感检索策略改编自2006年首次发表的策略（Glanville 2006），该策略是根据MEDLINE索引的随机对照试验报告标题和摘要中MeSH词和自由词出现的频率分析结果制订的，采用的是作者首次制订的检索MEDLINE中系统评价的检索策略设计方法（White 2001）。

提供两个检索策略：一个敏感度最大化版本和一个敏感度和精确性最大化版本。建议检索纳入Cochrane系统评价的试验报告时先用敏感度最大化版本和高敏感主题组合检索。如果该组合检索策略检出难以管理的参考文献数量，应当使用敏感度和精确性最大化版本策略。应当牢记MEDLINE摘要可以快速阅读，因为他们相对较短，保守估计30秒读一篇，1000篇摘要大约需要8小时.

重新分析获得这些策略的数据后，对检索策略进行了更新以反映自从最初的分析和检索语法改变后美国国立医学图书馆索引政策的变化。这些变化包括：

- 所有索引记录中出版类型标有‘Randomized Controlled Trial’或‘Controlled Clinical Trial’的不再在出版类型中标记为‘Clinical Trial’；
- 将主题词CLINICAL TRIALS改编为CLINICAL TRIALS AS TOPIC。

PubMed策略在框6.4.a 和框6.4.b中，Ovid检索策略在框6.4.c和框6.4.d中。

我们必须牢记，以下检索策略是以MEDLINE索引记录数据为基础，且用于MEDLINE检索。这些策略的设计并非检索“正在处理的”及其它未标引MeSH的记录。因此，建议这些策略在索引版MEDLINE中运行，并单独检索包含有“正在处理”的和非索引记录的数据库。例如以下Ovid检索策略应当在‘Ovid MEDLINE（R）1950 to Month Week X 200X’中运行和更新，而非索引记录应当在‘Ovid MEDLINE（R）In-Process & Other Non-Indexed Citations Month X，200X’中检索。为了检索非索引记录，需要一系列的自由词截词，如random，placebo，trial,等等，而且检索不能仅限于人类（因这些记录尚未添加人类索引）

如第6.3.2.1节讨论，已检索了MEDLINE1966-2004，包括使用Cochrane先前确定的随机对照试验版本的高敏感检索策略，而试验报告记录（仅在标题和摘要基础上）被重新在MEDLINE索引并纳入CENTRAL。请参阅章节6.3.2.1和6.3.3.2的进一步指导以便合理使用这些高敏感检索策略。

框6.4.a　Cochrane鉴定MEDLINE中随机对照试验的高敏感检索策略：敏感度最大化版（2008版）；PubMed格式

#1	randomized controlled trial [pt]
#2	controlled clinical trial [pt]
#3	randomized [tiab]
#4	placebo [tiab]
#5	drug therapy [sh]
#6	randomly [tiab]
#7	trial [tiab]
#8	groups [tiab]
#9	#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8
#10	animals [mh] NOT humans [mh]
#11	#9 NOT #10

PubMed检索语法

[pt] 表示发表类型；

[tiab] 表示题目或摘要中的字词；

[sh] 表示副主题词；

[mh] 表示医学主题词（MeSH）（“扩展”）；

[mesh: noexp] 表示医学主题词（MeSH）（未“扩展”）；

[ti] 表示题目中的字词。

框6.4.b Cochrane鉴定MEDLINE中随机对照试验的高敏感检索策略：敏感度和精确性最大化版（2008版）；PubMed格式

```
#1  randomized controlled trial [pt]
#2  controlled clinical trial [pt]
#3  randomized [tiab]
#4  placebo [tiab]
#5  clinical trials as topic [mesh: noexp]
#6  randomly [tiab]
#7  trial [ti]
#8  #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7
#9  animals [mh] NOT humans [mh]
#10 #8 NOT #9
```

PubMed检索语法

[pt] 表示发表类型；

[tiab] 表示题目或摘要中的字词；

[sh] 表示副主题词；

[mh] 表示医学主题词（MeSH）（“扩展”）；

[mesh: noexp] 表示医学主题词（MeSH）（未“扩展”）；

[ti] 表示题目中的字词。

框6.4.c　Cochrane鉴定MEDLINE中随机对照试验的高敏感检索策略：

敏感度最大化版（2008版）；Ovid格式

1	randomized controlled trial.pt.
2	controlled clinical trial.pt.
3	randomized.ab.
4	placebo.ab.
5	drug therapy.fs.
6	randomly.ab.
7	trial.ab.
8	groups.ab.
9	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8
10	exp animals/ not humans.sh.
11	9 not 10

Ovid检索语法：

.pt. 表示发表类型；

.ab. 表示摘要中的字词；

.fs. 表示漂浮主题词；

.sh. 表示医学主题词（MeSH）；

.ti. 表示题目中的字词。

框6.4.d　Cochrane鉴定MEDLINE中随机对照试验的高敏感检索策略：

敏感度和精确性最大化版（2008版）；Ovid格式

1	randomized controlled trial.pt.
2	controlled clinical trial.pt.
3	randomized.ab.
4	placebo.ab.
5	clinical trials as topic.sh.
6	randomly.ab.
7	trial.ti.
8	1 or 2 or 3 or 4 or 5 or 6 or 7
9	exp animals/ not humans.sh.
10	8 not 9

Ovid检索语法：

.pt. 表示发表类型；

.ab. 表示摘要中的字词；

.fs. 表示漂浮主题词；

.sh. 表示医学主题词（MeSH）；

.ti. 表示题目中的字词。

6.4.11.2 鉴定 EMBASE 中随机对照试验的检索过滤器

英国Cochrane中心正在设计鉴定EMBASE中随机对照试验报告的客观的高敏感检索策略，使用类似于在设计鉴定MEDLINE中随机对照试验的高敏感检索策略时采用的词频分析法，如章节6.4.11.1描述（Glanville 2006）。系统评价作者希望在EMBASE中运行自己的检索策略，同时希望考虑使用章节6.3.2.2所列的检索词，而目前英国Cochrane中心正使用这些检索词检索EMBASE随机试验报告以纳入CENTRAL（Lefebvre 2008）。另外，也可使用Wong及其同事设计的检索过滤器（Wong 2006），以检索EMBASE中他们认为“可靠的临床治疗性研究”。

如章节6.3.2.2讨论，已使用该章节中列出的词语检索了1980-2006的EMBASE，试验报告记录（仅在标题和摘要基础上）也已纳入CENTRAL。

6.4.12 检索更新

更新一个Cochrane系统评价时，不得不重新评价检索过程（例如，决定应当检索哪些年的哪些数据库和其他数据源）。先前检索过并认为与更新相关的数据库需要重新检索，先前的检索策略需要更新以反映各种问题。例如：索引变化，如增加或移除受控词汇（MeSH，EMTREE等等）；检索语法变化；对先前检索策略的评论或批评。更新时任何先前检索过的数据库若不予检索，应当解释和说明理由。还应考虑：新数据库或其它资源有可能产生，或成为系统评价员或试验检索协调员的可用资源。

当在PubMed或Ovid MEDLINE“In-Process & Other Non-Indexed Citations and Ovid MEDLINE 1950 to Present”文件等数据库同时检索MEDLINE索引记录和非索引记录时，应谨慎使用更新限制。如有可能，应单独选择文件并单独检索，例如Ovid MEDLINE ‘1950 to Month Week X 200X’，以及Ovid MEDLINE中‘In-Process & Other Non-Indexed

Citations Month X，200X’非索引记录文件。关于这一问题的进一步指导，联系试验检索协调员。

6.4.13 检索策略示范

框6.4.e提供了一个主题为“它莫西芬治疗乳腺癌”的CENTRAL检索策略演示。注意：它仅包括主题词（随机对照试验过滤器不适合CENTRAL）。没有限制于人类。该策略只用于演示目的：检索CENTRAL中研究以纳入系统评价时针对每一个概念需要更多的检索词汇。

框6.4.f提供一个主题为“它莫西芬治疗乳腺癌”的Ovid MEDLINE检索策略演示。注意MEDLINE使用了主题词和一个随机对照试验过滤器，检索仅限于人类。提供这一策略仅作为演示目的：检索MEDLINE中研究以纳入系统评价时针对每一概念需要更多的检索词汇。

框6.4.e 主题为“它莫西芬治疗乳腺癌”的CENTRAL检索策略示范

#1	MeSH descriptor Breast Neoplasms explode all trees
#2	breast near cancer*
#3	breast near neoplasm*
#4	breast near carcinoma*
#5	breast near tumour*
#6	breast near tumor*
#7	#1 OR #2 OR #3 OR #4 OR #5 OR #6
#8	MeSH descriptor Tamoxifen explode all trees
#9	tamoxifen
#10	#8 OR #9
#11	#7 AND #10

“near”运算符默认为在6个字内；

‘*’表示阶段符。

框6.4.f　主题为“它莫西芬治疗乳腺癌”的MEDLINE（Ovid格式）检索策略示范

1	randomized controlled trial.pt.
2	controlled clinical trial.pt.
3	randomized.ab.
4	placebo.ab.
5	drug therapy.fs.
6	randomly.ab.
7	trial.ab.
8	groups.ab.
9	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8
10	animals.sh. not (humans.sh. and animals.sh.)
11	9 not 10
12	exp Breast Neoplasms/
13	(breast adj6 cancer$).mp.
14	(breast adj6 neoplasm$).mp.
15	(breast adj6 carcinoma$).mp.
16	(breast adj6 tumour$).mp.
17	(breast adj6 tumor$).mp.
18	12 or 13 or 14 or 15 or 16 or 17
19	exp Tamoxifen/
20	tamoxifen.mp.
21	19 or 20
22	11 and 18 and 21

‘adj6’运算符表示在6个字内；

‘$’表示截断符；

.mp.表示检索标题、原标题、摘要、实义词及主题词。

6.4.14　要点总结

Cochrane系统评价作者在开始检索前应联系所在CRG小组的试验检索协调员；就大多数Cochrane系统评价而言，数据库检索结构包括：主题检索人群或状况、干预措施、研究设计的方法学过滤器如随机对照试验；

检索CENTRAL，不适合用随机对照试验过滤器和限制为人类；

避免过多不同检索概念，但使用宽泛的同义词和相关词（自由词和受控词汇），每一个概念内用OR组合；

用运算符AND组合不同的概念；

避免在组合检索设置中使用运算符NOT；

力求高敏度则应准备接受低精确性；

检索策略中不要应用语言限制；

为某一特定数据库和服务供应商设计的检索，应用于其他数据库或服务供应商时需要“翻译”；

确保能够意识到任何撤回出版物（例如，欺诈性出版物）、勘误和评论.

为了检索MEDLINE中的随机对照试验，首先使用Cochrane高敏度检索策略的敏感最大化版本检索。如果检索到的文献量过多，则需要使用敏感度和精确性最大化版本进行检索；

在任何可能的情况下，更新检索时应分开选择数据库文件和分开检索MEDLINE索引记录和正在处理的非索引记录.

6.5 参考文献管理

6.5.1 书目文献管理软件

专门设计的书目或参考文献管理软件，如EndNote、ProCite、Reference Manager和RefWorks是有用的，而且用于追溯研究的参考文献和报告时相对容易。系统评价作者机构的支持和可及性将影响选择使用哪款软件。如需上述产品的比较和与其它书目软件包的评价链接请见：

- www.burioni.it/forum/dellorso/bms-dasp/text/

上述列出的软件包中，普遍认为ProCite对识别重复参考文献非常有效，但提供商不再更新了。它除正确输入英文参考文献外不支持其它字符集，而EndNote可以。书目软件也方便保存检索方法和过程信息。例如，单独未使用的字段可以用来存储信息：1）用于鉴定试验报告的数据库名或其它来源详细信息；2）什么时间和从那里订购一篇文章以及文章收到日期；3）与文章相关的研究是否应纳入系统评价或排除，如果排除，理由是什么。

从Cochrane信息检索方法学组网站可获得将来自CENTRAL的参考文献导入书目文献管理软件的文件：

- www.cochrane.org/docs/import.htm

6.5.2 下载哪些字段

除了完整的引文记录，应当考虑从数据库下载一些关键字段。有关字段下载的指南已由试验检索协调员工作组编辑完成，而且可以从“TSC管理专题数据库和手检记录用户指南”文件中获得，地址：

- www.cochrane.org/resources/hsearch.htm

摘要：摘要可以用来消除明显无关的报告，无须获得全文或以后返回书目数据库。

收录号/唯一识别号：最好预留未使用的字段存储下载记录的唯一识别号/收录号，例如PubMed的PMID。以链接到完整的数据库记录，也有利于信息管理，如重复检测和删除。

联系/地址：可能包括机构联系地址和/或作者的e-mail地址。

文章识别码/数字对象标识码（DOI）：可用于引用和链接到完整的记录。

临床试验注册码：如果记录包含ClinicalTrials.gov或ISRCTN计划分配的一个临床试验注册号，或由试验赞助者分配的一个号，这些应下载并有助于将试验报告与原始研究链接的。例如：最近在EMBASE引入的临床试验注册号（CN）字段。

索引词/叙词/关键词：见章节6.4.5。如果标题和摘要缺乏详细信息，索引词/叙词/关键词有助于说明为什么要检索该记录。

语言：原始论文发表语言

评论、更正、勘误、撤回和更新：确保随后发表的评论、更正、勘误、撤回和更新相关的任何字段都包含在下载记录中非常重要，以便考虑这些出版物的任何后续影响。框6.5.a描述了需要考虑的最重要字段及他们在PubMed中对应的字段标签。

www.nlm.nih.gov/bsd/mms/medlineelements.html#cc

框6.5.a　PubMed中重要的字段标签

CIN: ‘Comment in’

CON: ‘Comment on’

CRI: ‘Corrected and republished in’

CRF: ‘’Corrected and republished from’

EIN: ‘Erratum in’

EFR: ‘Erratum for’

PRIN: ‘Partial retraction in’

PROF: ‘Partial retraction of’

RIN: ‘Retraction in’

ROF: ‘Retraction of’

RPI: ‘Republished in’

RPF: ‘’Republished from’

UIN: ‘Update in’

UOF: ‘Update of’

6.5.3　要点总结

- 使用书目文献管理软件管理参考文献。
- 确保下载所有必要的字段。

6.6　记录和报告检索过程

6.6.1　记录检索过程

整个检索过程的细节都需要记录，以确保在系统评价中正确地报告。在最大程度上确保所有数据库检索可重复。从开始应牢记每一个数据库全部检索策略都应输入系统评价的一个附录中。运行的检索策略需要正确复制和黏贴，且包括所有检索文献号和检索记录数。检索到的记录数需要记录在系统评价结果中“检索结果标题”部分下（见第4章，4.5节）。检索策略不应重新输入而产生误差。最近的一项研究显示对Cochrane系统评价手册中关于检索策略指南的描述缺乏依从性（Sampson 2006）。在大多数Cochrane系统评价小组中，要求试验检索协调员评论系统评价的检索策略，并作为系统评价在CDSR发表前签署同意过程之一。因此，建议系统评价作者应与试验检索协调员联系，尽早指导

检索记录过程，以便在系统评价中书写这部分内容。正如本章节所提及，重要的是以文件或打印副本方式保存在互联网上发现的任何信息，如在研试验信息。因为当撰写系统评价时，这些信息有可能不再检索。

根据2011年引入的新指南，强调鼓励系统评价作者采用PRISMA推荐的研究流程图（见第4章4.5节），而流程图应呈现的详细信息（见11章11.2.1节）如下：

- 检索鉴定出的记录数
- 初筛排除的记录数（如根据题目和摘要）
- 获取了全文的记录数
- 建议检索过程应做记录，以确保使用的流程图是完全正确的。

6.6.2 报告检索过程

6.6.2.1 在计划书中报告检索过程

- 列出待检索的数据库
- 规定在列出数据库中检索研究的最早日期
- 注明任何语言或发表状态的限制
- 列出待检索的灰色数据库资源
- 列出系统评价特别要手检的所有杂志和会议论文集
- 列出任何其它的资源（例如，参考文献目录、互联网）

Cochrane系统评价计划书中纳入详细的检索策略（一行一行的）是可选项。有必要在计划书阶段就记录已开展的检索方案，以便与计划书中的其它部分一样进行评价。有些系统评价小组认为直到计划书完成并发表时才应进行检索，因为已有研究的知识可能影响计划书的某些方面如纳入标准。

6.6.2.2 在系统评价中报告检索过程

在系统评价摘要中报告检索过程

- 列出所有检索过的数据库。
- 注明每一个数据库最后检索日期或检索期间。
- 注明所有语言或出版状态的限制（参见章节6.4.9）。
- 列出联系个人或组织。

有关这一信息的指南参见第11章（11.8部分）

在方法部分报告检索过程

在“研究的检索方法”部分：

- 列出所有检索过的数据库
- 注明每一个数据库最后检索日期及检索期间
- 注明所有语言和出版状态的限制
- 列出灰色文献资源
- 列出联系个人和机构
- 列出系统评价特别要手检的所有杂志和会议论文集
- 列出任何检索过的其它资源（例如，参考文献目录、互联网）

每一个数据库的完整检索策略应包括在系统评价的附录中，以免打断流畅的系统评价全文。每一个数据库的完整检索策略包括所有的检索序号均应按运行时正确复制和粘贴。不应因为重新输入而产生错误。有关详细的指南请联系试验检索协调员。

在结果部分报告检索过程

- 通过电子检索命中的记录数应当记录在结果部分。

报告检索日期

检索日期应当在“检索日期”字段标出，以表明最近什么时候开始完整地检索。欲了解更多有关指定日期的信息，见第3章（3.3.3节）。

6.6.3 要点总结

- 开始检索之前向试验检索协调员寻求记录检索过程的指导
- 每一个数据库每次的完整检索策略应当复制和粘贴在系统评价的附录中
- 每一个检索策略命中总数应记录在结果部分
- 在本地或打印文件副本保存从互联网发现的任何信息，如在研试验信息
- 参阅第4章（4.5节）和第11章（11.8节）了解更多关于如何在系统评价和摘要中报告信息

6.7 本章信息

作者：代表Cochrane信息检索方法组的Carol Lefebvre, Eric Manheimer 和Julie Glanville

本章引用格式：Lefebvre C, Manheimer E, Glanville J. Chapter 6: Searching for studies. In: Higgins JPT, Green S （editors）. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.2 [updated September 2009]. The Cochrane Collaboration, 2009. Available from www.cochrane-handbook.org.

致谢：本章基于1995年Kay Dickersin, Kristen Larson, Carol Lefebvre and Eric Manheimer合著手册版本部分。众所周知本章所列的许多资源经过了许多协作者多年的努力，对此，我们深表感谢。我们要感谢信息学家与我们分享这些信息和记录他们的研究过程。我们也要感谢Cochrane试验检索协调员、Cochrane信息检索方法学组成员（见表6.7.a）、卫生技术评估国际信息资源专题兴趣组和InternetTASC信息专家小组对本章早期草稿的评论，感谢Anne Eisinga校对检索策略，以及两个同行评审专家Steve McDonald 和 Ruth Mitchell的详细和建设性意见。

框6.7.a Cochrane信息检索方法学组

信息检索方法学组（IRMG）旨在提供建议和支持，开展研究和促进信息检索方法交流，以支持Cochrane协作网信息检索活动。该组在2004年11月向协作网正式注册。该组成员主要提供制定检索方法的实际支持，促进信息检索。该组通过以下活动实现目标：

- 提供有关信息检索政策和实践建议.
- 提供培训和支持.
- 开展信息检索方法实证研究（包括系统评价）.
- 帮助监督系统评价中检索技术的质量.
- 与Campbell协作网合作，避免对共同感兴趣领域的重复信息检索。
- 作为一个讨论论坛

网址: www.cochrane.org/docs/irmg.htm

6.8 参考文献

Bennett 2003

Bennett DA, Jull A. FDA: untapped source of unpublished trials. The Lancet 2003; 361: 1402-1403.

De Angelis 2004

De Angelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, Laine C, Marusic A, Overbeke AJ, Schroeder TV, Sox HC, Van der Weyden MB, International Committee of Medical Journal Editors. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. JAMA 2004; 292: 1363-1364.

De Angelis 2005

De Angelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, Laine C, Marusic A, Overbeke AJ, Schroeder TV, Sox HC, Van der Weyden MB, International Committee of Medical Journal Editors. Is this clinical trial fully registered? A statement from the International Committee of Medical Journal Editors. JAMA 2004; 293: 2927-2929.

Dickersin 1994

Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ 1994; 309: 1286-1291.

Dickersin 2002

Dickersin K, Manheimer E, Wieland S, Robinson KA, Lefebvre C, McDonald S, CENTRAL Development Group. Development of the Cochrane Collaboration's CENTRAL Register of controlled clinical trials. Evaluation and the Health Professions 2002; 25: 38-64.

Eisinga 2007

Eisinga A, Siegfried N, Clarke M. The sensitivity and precision of search terms in Phases I, II and III of the Cochrane Highly Sensitive Search Strategy for identifying reports of randomized trials in MEDLINE in a specific area of health care - HIV/AIDS prevention and treatment interventions. Health Information and Libraries Journal 2007; 24: 103-109.

Eysenbach 2001

Eysenbach G, Tuische J, Diepgen TL. Evaluation of the usefulness of Internet searches to identify unpublished clinical trials for systematic reviews. Medical Informatics and the Internet in Medicine 2001; 26: 203-218.

Glanville 2006

Glanville JM, Lefebvre C, Miles JN, Camosso-Stefinovic J. How to identify randomized controlled trials in MEDLINE: ten years on. Journal of the Medical Library Association 2006; 94: 130-136.

Glanville 2008

Glanville J, Bayliss S, Booth A, Dundar Y, Fleeman ND, Foster L, Fraser C, Fernandes H, Fry-Smith A, Golder S, Lefebvre C, Miller C, Paisley S, Payne L, Price AM, Welch K, InterTASC Information Specialists' Subgroup. So many filters, so little time: The development of a Search Filter Appraisal Checklist. Journal of the Medical Library Association （in press, 2008）.

Golder 2006

Golder S, McIntosh HM, Duffy S, Glanville J, Centre for Reviews and Dissemination and UK Cochrane Centre Search Filters Design Group. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. Health Information and Libraries Journal 2006; 23: 3-12.

Greenhalgh 2005

Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. BMJ 2005; 331: 1064-1065.

Hetherington 1989

Hetherington J, Dickersin K, Chalmers I, Meinert CL. Retrospective and prospective identification of unpublished controlled trials: lessons from a survey of obstetricians and pediatricians. Pediatrics 1989; 84: 374-380.

Hopewell 2007a

Hopewell S, Clarke M, Lefebvre C, Scherer R. Handsearching versus electronic searching to identify reports of randomized trials. Cochrane Database of Systematic Reviews 2007, Issue 2. Art No: MR000001.

Hopewell 2007b

Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in Meta-analyses of randomized trials of health care interventions. Cochrane Database of Systematic Reviews 2007, Issue 2. Art No: MR000010.

Horton 1997

Horton R. Medical editors trial amnesty. The Lancet 1997; 350: 756.

Khan 2001

Khan KS, ter Riet G, Glanville J, Sowden AJ, Kleijnen J (editors). Undertaking Systematic Reviews of Research on Effectiveness: CRD's Guidance for those Carrying Out or Commissioning Reviews (CRD Report Number 4) (2nd edition). York (UK): NHS Centre for Reviews and Dissemination, University of York, 2001.

Lefebvre 2001

Lefebvre C, Clarke M. Identifying randomised trials. In: Egger M, Davey Smith G, Altman DG (editors). Systematic Reviews in Health Care: Meta-analysis in Context (2nd edition). London (UK): BMJ Publication Group, 2001.

Lefebvre 2008

Lefebvre C, Eisinga A, McDonald S, Paul N. Enhancing access to reports of clinical trials published world-wide - the contribution of EMBASE records to the Cochrane Central Register of Controlled Trials (CENTRAL) in The Cochrane Library. Emerging Themes in Epidemiology (in press, 2008).

MacLean 2003

MacLean CH, Morton SC, Ofman JJ, Roth EA, Shekelle PG. How useful are unpublished data from the Food and Drug Administration in Meta-analysis? Journal of Clinical Epidemiology 2003; 56: 44-51.

Mallett 2002

Mallett S, Hopewell S, Clarke M. Grey literature in systematic reviews: The first 1000 Cochrane systematic reviews. Fourth Symposium on Systematic Reviews: Pushing the Boundaries, Oxford (UK), 2002.

Manheimer 2002

Manheimer E, Anderson D. Survey of public information about ongoing clinical trials funded by industry: evaluation of completeness and accessibility. BMJ 2002; 325: 528-531.

McDonald 2002

McDonald S. Improving access to the international coverage of reports of controlled trials in electronic databases: a search of the Australasian Medical Index. Health Information and Libraries Journal 2002; 19: 14-20.

Montori 2005

Montori VM, Wilczynski NL, Morgan D, Haynes RB. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. BMJ 2005; 330: 68.

Royle 2003

Royle P, Milne R. Literature searching for randomized controlled trials used in Cochrane reviews: rapid versus exhaustive searches. International Journal of Technology Assessment in Health Care 2003; 19: 591-603.

Sampson 2006

Sampson M, McGowan J. Errors in search strategies were identified by type and frequency. Journal of Clinical Epidemiology 2006; 59: 1057-1063.

Scherer 2007

Scherer RW, Langenberg P, von Elm E. Full publication of results initially presented in abstracts. Cochrane Database of Systematic Reviews 2007, Issue 2. Art No: MR000005.

Suarez-Almazor 2000

Suarez-Almazor ME, Belseck E, Homik J, Dorgan M, Ramos-Remus C. Identifying clinical trials in the medical literature with electronic databases: MEDLINE alone is not enough. Controlled Clinical Trials 2000; 21: 476-487.

White 2001

White VJ, Glanville JM, Lefebvre C, Sheldon TA. A statistical approach to designing search filters to find systematic reviews: objectivity enhances accuracy. Journal of Information Science 2001; 27: 357-370.

Whiting 2008

Whiting P, Westwood M, Burke M, Sterne J, Glanville J. Systematic reviews of test accuracy should search a range of databases to identify primary studies. Journal of Clinical Epidemiology 2008; 61: 357.e1-357.e10.

Wilczynski 2007

Wilczynski NL, Haynes RB, Hedges Team. EMBASE search strategies achieved high sensitivity and specificity for retrieving methodologically sound systematic reviews. Journal of Clinical Epidemiology 2007; 60: 29-33.

Wong 2006

Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically sound treatment studies in EMBASE. Journal of the Medical Library Association 2006; 94: 41-47.

（何林译，岑啸、张龙浩初审）

第七章　选择研究报告和收集数据

本文选自工作手册5.0.1版本。有关如何引用它的指南，见7.9节。这些材料还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 评价研究的合格性及从研究报告中提取数据，至少应由两人独立完成。
- Cochrane 干预措施系统评价以研究作为兴趣单位而非报告，因此需将相同研究的多种报告连接在一起。
- 数据收集表是非常重要的。他们应当针对系统评价的目的而设计，而且对每一个新的系统评价（或评价小组）有启示作用。
- 有些技巧，有助于数据收集表的设计和使用。
- 数据可能以不同的格式报告，但往往可以转化为一种合适做 Meta-分析的格式。

7.1 引言

一个系统评价的结果主要取决于纳入的相关研究，以及从这些研究中提取和分析的数据。如何选择研究及对研究数据进行提取和分析的方法必须透明，选择的方法应当使偏倚和人为误差缩小。以下将描述Cochrane系统评价如何选择研究、如何提取研究中哪些数据的方法。

7.2 选择研究

7.2.1 研究（非报告）作为兴趣单位

一个Cochrane系统评价是对多个研究的一个评价，而这些研究应符合预先制定的系统评价纳入标准。由于每个研究可能在几篇文章、摘要或其它报告中报道，因此，全面的文献检索可能从相关的研究中检出许多报告。为此需要两种不同的方法确定哪些研究能纳入系统评价。方法之一是将所有相同研究的不同报告连接在一起；其二是用各种报告中可得的信息确定哪些研究适合纳入。虽然有时一个研究仅一个研究报告，但绝不应假设这种情况。

7.2.2 识别同一研究的多个研究报告

如果无意中将研究纳入Meta-分析超过一次，重复发表将引起实质性偏倚（Tramèr 1997）。重复发表形式多样，从相同的手稿到报告不同参与者数和不同结局均可（Von Elm 2004）。有可能很难检测重复发表，需要系统评价的作者来做“检测工作”。

一些最有用的比较报告标准是：

- 作者姓名（大多数重复报告作者相同，但并不总是这样）；
- 地点和环境（尤其是被命名的机构，如医院）；
- 具体的干预措施（如剂量，频率）；
- 参与者数量和基线数据；以及
- 研究日期和持续时间（这也可以说明不同的样本大小，是否因不同时期新的研究对象的纳入）

考虑这些和其他因素后仍然不确定，则有必要联系该报告的作者。

7.2.3 选择研究的典型过程

选择纳入系统评价的研究的典型过程如下（这个过程应当在评价的计划书中详细描述）：

1. 使用参考文献管理软件合并检索结果，并删除相同报告的重复记录（见第6章6.5节）。
2. 检查标题和摘要，删除明显不相关的报告（在这个阶段作者一般包括过多）。
3. 获取潜在相关报告的全文。
4. 将同一研究的多个报告连接在一起（见章节7.2.2）。
5. 查阅报告全文，判断研究是否符合纳入标准。
6. 适当时，联系研究者以证明研究合格性（可同时要求更多的信息，如结果遗漏）。
7. 最后决定研究的纳入和着手数据收集。

7.2.4 选择过程的实施

决定哪一个研究应纳入系统评价是评价过程中最有影响的决策之一。然而决策基于判断。为确保这些判断是可重复的，最好是多个作者参与判断过程。在实践中，针对每一个系统评价的确切方法各不相同，它在某种程度上依赖于系统评价作者的经验和专业知识。

系统评价作者必须首先决定是否多一位作者来评估检出的标题和摘要（7.2.3节第2步）。至少由两个作者参与评估可能减少相关报告被排除的可能性（Edwards 2002）。最重要的是：研究是否纳入系统评价的最后选择应该至少由两人来完成（7.2.3节第5步）。

某一特定领域的专家通常有预先成形的观点，这将对文章的相关性和真实性评价产生偏倚（Cooper 1989，Oxman 1993）。因此 一方面应该至少有一位评价者是该评价领域的专家，而第二个作者最好不擅长该领域。有些评价者认为应由那些不知道文章信息（如发表的杂志、作者、机构、结果的大小和方向）的人员来评估文章的相关性。为此，他们需要编辑文章副本。但这需要花费更多时间，并且从防止偏倚而言，是否有效也难肯定，而且也不能保证所需要的资源（Berlin 1997）。

有关研究报告是否应被纳入的意见分歧常可通过讨论来解决。意见分歧的原因常常是由于评价者中某一人的疏忽所造成。当由于理解不一致产生分歧时，这可能需要第三方仲裁。有时对一篇研究是否应被纳入产生分歧，在不增加信息的情况下是难以解决的。

在这种情况下，评价者可将选择的研究报告归类为待评估报告直至从作者那里获得进一步信息再做决定。

总之，计划书和系统评价的方法部分应当详述：

- 是否由多个作者检查每一篇文献的标题和摘要以排除显而易见的不相关报告；
- 检查每一篇全文报告以决定是否独立地进行纳入工作（至少有两人完成）；
- 上述决定是否由相关专业专家或方法学家参与，或两者都参与决定；
- 当评估人员应用纳入标准评价研究报告的相关性时，是否了解文章作者姓名、单位、出版的期刊和结果；
- 如何处理分歧。

研究不满足众多纳入标准中的一个便足以排除该研究纳入系统评价。因此，在实践中，对每个研究的纳入标准应按照重要性顺序评估，以便第一个回答“不”时即可作为排除该研究的主要原因，其余标准就不需要评价。

就大多数评价而言，值得选择一个研究报告的样本（包括10-12篇文献，其中有合格的，有不合格的和合格性有争议的）作纳入标准的预试验。预试验可用于完善和明确纳入标准，培训使用该纳入标准的人员，使纳入标准经多人使用都能保持一致。

7.2.5 选择“排除研究”

一个Cochrane系统评价包括一个已排除研究的列表，详述读者可能期望纳入的研究。这包括所有表面上看符合纳入标准，但是细究却不是的研究，以及并不符合所有纳入标准，但众所周知且很可能被某些读者认为相关的研究。通过列出这些排除研究并给出排出的主要原因，表明评价者考虑了这些研究。排除研究目录应尽可能简单。不应列出通过全面检索得到的所有报告，以及显然不符合系统评价“研究类型”、“受试者类型”、“干预措施类型”纳入标准的研究，尤其是如果评价只包含随机对照试验，不应该列出明显是非随机的研究。

7.2.6 测量一致性

正式测量一致性可用于描述多个作者评价的一致程度（Orwin 1994）。我们在章节7.2.6.1描述如何用Kappa统计量计算两个作者在作简单的纳入/排除决策时的测量一致性。Kappa值在0.40-0.59之间认为一致性好，在0.60-0.74之间认为相当好，在0.75及以上认为

一致性非常好（Orwin 1994）。

不建议把Kappa统计量作为Cochrane系统评价的标准，尽管他们能揭示问题，尤其是在预试验的早期阶段。比较任意的切割点，Kappa值不太能揭示有关评价分歧的实质影响。例如，评估一个实施良好的大规模研究合格性时的分歧比一个小型有偏倚风险的研究的分歧对系统评价影响更大。任何分歧的原因应加以探讨。他们可能表明需要重新审查合格标准或数据收集编码方案，以及需要报告由此引发的任何变更。

样本Kappa统计量的计算

假定K研究是依据表7.2.a.中数字a到i分布，那么：

$$\text{kappa} = \frac{P_O - P_E}{1 - P_E},$$

这里

$$P_O = \frac{a + e + i}{K}$$

是一致性的研究比例，

$$P_E = \frac{I_1 \times I_2 + E_1 \times E_2 + U_1 \times U_2}{K^2}$$

由机遇单独引起的期望一致性的研究比例，以表7.2.b的数据为例.

$$P_O = \frac{5 + 7 + 3}{25} = 0.6,$$
$$P_E = \frac{12 \times 5 + 10 \times 10 + 3 \times 10}{25^2} = 0.304,$$

因此

$$\text{kappa} = \frac{0.6 - 0.304}{1 - 0.304} = 0.43.$$

表7.2.a　简单Kappa统计计算的数据

		评价作者2			
		纳入	排除	不确定	总计
评价作者1	纳入	a	b	c	I1
	排除	d	e	f	E1
	不确定	g	h	i	U1
	总计	I2	E2	U2	K

表7.2.b　样本Kappa统计量的数据举例

		评价作者2			
		纳入	排除	不确定	总计
评价作者1	纳入	5	3	4	12
	排除	0	7	3	10
	不确定	0	0	3	3
	总计	5	10	10	25

7.3　收集何种数据

7.3.1　什么是数据

根据本章的目的，我们定义“数据”是任何关于（或来源于）一个研究的信息，包括具体的方法、受试者、实施场地、背景、干预措施、结局、结果、出版物和研究者。评价者应预先计划什么样的数据符合其系统评价要求，并制定获取相关信息的策略。以下各章节回顾了应查找信息的类型，并总结在表7.3.a.。章节7.4回顾了主要数据源。

表7.3.a 数据收集或数据提取应考虑的项目清单

来源

- 研究ID（评价者创编）
- 报告ID（评价者创编）
- 评价者ID（评价者创编）
- 引文和联络方式

合格性

- 确认符合系统评价的合格标准
- 排除原因

方法

- 研究设计
- 研究期限
- 序列产生*
- 分配序列隐藏*
- 盲法*
- 其它有关偏倚问题*

参与者/受试者

- 总数
- 实施场地
- 诊断标准
- 年龄
- 性别
- 国家
- ［并存疾病］
- ［社会人口统计］
- ［种族］
- ［研究日期］

干预措施

- 干预组总数
- 对于每个关注的干预组和对照组
- 具体干预措施
- 干预措施细节（如果可行，足以重复）
- ［干预措施完整性］

结果

- 结果和时间点：（i）收集结果和时间点；（ii）报告结果和时间点*

对每一个感兴趣的结果：

- 结果定义（诊断标准，如相关性）
- 测量单位（如相关）
- 对于量表：上、下限，以及是否高或低的分数是好的

结论

- 分配到每一个干预组的受试者数量
- 对每一个感兴趣的结果：
- 样本量
- 失访受试者*
- 每个干预组的汇总数据（例如二分类数据的四格表；连续数据的均值和SDs标准差）
- ［效应估计值和可信区间；P值］
- ［亚组分析］

其他

- 资金来源
- 纳入研究作者的重要结论
- 纳入研究作者的其它评论
- 其它相关研究的参考文献
- 要求的通信信息
- 系统评价者的其他评论

*需全面描述在‘偏倚风险’工具的标准条目（见第 8 章，第 8.5 节） 无括号的项目通常应在所有的系统评价中收集；方括号内的项目只在一些相关的系统评价中收集。

7.3.2 方法和潜在的偏倚源

不同的研究方法对结果产生不同的偏倚从而影响研究结局。研究设计的基本特征应收集在“纳入研究特征”表中，包括研究是否随机，研究是否为整群设计或交叉设计以及研究期限。如果系统评价包括非随机研究，应当描述研究的相应特征（见第13章，13.4节）。

还应当使用第8章（8.5节）中描述的工具来收集信息，以便评价每一个纳入研究的偏倚风险。该工具包括序列生成、分配序列隐藏、盲法、结局数据不完整和选择性报告结局。工具中的每个条目，需要描述研究中发生了什么，其中可能包括从研究报告中逐字引用。从结局和结果中收集评价结局数据不完整和选择性报告结局的信息可能最方便。第8章（8.3.4节）将讨论如何收集评估偏倚风险的信息。

7.3.3 受试者和实施场地

受试者和实施场地的详细资料主要收集在“纳入研究特征”表中。有些Cochrane系统评价小组制定了哪些特征应予以收集的标准。通常情况下，应当收集那些可能（或被认为）会影响干预措施效果的有无或大小的信息，及有助于用户评价适用性的信息。例如，如果评价者怀疑在不同的社会经济群体中（这是罕见例子）干预效果存在很大差异的时候，应当收集这种信息。如果认为干预效果在这些群体中很稳定，并且这些信息对结果的应用无益，就不必收集。

有助于评估适用性的受试者特征包括年龄和性别，如果从上下文看不是很明显的话，概要信息中应进行收集。这些信息可能会以不同形式呈现（例如，年龄以均数或中位数表示，包括标准差或四分位间距；性别为百分比或计数；要么以整个研究描述结果，要么以每一干预组分别描述结果）。评价者应尽可能寻找一致的量化指标，并决定是以整个研究还是分不同干预组总结特征更恰当。有时其他特征也很重要，包括种族、社会人口统计资料（例如，教育程度）和并存疾病的存在。

如果研究的实施场地可能影响干预效果或适用性，那么应当收集这些相关信息。卫生保健干预研究的典型场所包括急救医院、急救设施、全科病房、附加医疗保健设施（例如老人院、办公室、学校以及社区）。有时在文化差异显著的不同地区进行的研究将影响干预的实施和效果。研究时间可能与重要技术差异或时间趋势相关。如果这些信息对系统评价解释很重要，也应该收集。

用于定义健康状况的诊断标准是不同研究出现差异的特别重要的来源，应当收集此信息。例如，在评价药物治疗充血性心力衰竭时，知道各研究如何定义心衰及严重程度（例如，收缩或舒张功能不全，严重收缩功能不全伴射血分数小于20%）很重要。同样，在评价抗高血压治疗时，描述受试对象的基础血压值也很重要。

7.3.4 干预措施

所有与系统评价相关的试验和对照措施都应当收集在“纳入研究特征表”中。同样，影响效果的存在和大小或有助于用户评价适用性的详细资料也要收集。如果可行，应当收集（并在系统评价中展现）足以重复研究中的干预措施的信息包括作为研究一部分的任何合并使用的干预措施。

对于许多非复杂干预的临床试验如药物或物理干预措施，给药途径（如口服或静脉给药，使用外科技术）、剂量（如总量或每一治疗强度，给药频率）、时间（如确诊后24小时内）和疗程也可能相关。对于复杂的干预措施，如心理治疗、行为和教育措施或医疗保健实施方案等，在评价时要注意收集干预措施的信息，如谁实施干预措施、其内容、形式及时间等。

干预措施完整性

是否按计划实施规定的干预措施程序或内容会对研究结果产生重要影响。我们将此描述为干预措施完整性；相关术语包括依从性（compliance）和真实性（fidelity）。在预防措施和复杂干预措施系统评价中，确认干预措施完整性可能尤为重要，因为常常是在存在诸多障碍的环境下实施的（Dane 1998）。信息完整性有助于确定不理想的结果是否由于不够理想的干预或对规定内容的不完全实施所致。评估干预措施的实施也揭示了在真实环境下实施一项干预措施的可行性重要信息，特别是干预措施能够并将会按计划执行的可能性有多大。如果在实践中很难做到全面实施，该方案可行性将很低（Dusenbury 2003）。

以下是Dane 和Schneider描述的预防性研究方案完整性的五个方面（Dane 1998）：

1. 指定的干预措施按规定实施的程度（依从性）；
2. 干预措施内容实施的次数、疗程和频率（暴露情况）；
3. 不直接与规定内容实施相关的干预实施质量方面，如执行者热情/对执行者的培训/疗程效果的总体评估以及领导人对干预的态度（实施质量）；

4. 受试者对干预措施的反应性测量，可能包括参与水平和热情等指标（受试者的反应）；
5. 防止干预措施扩散的保障措施，也就是确保试验组的每一个对象只接收计划的干预措施（计划变异）。

一项干预措施的完整性可在研究过程中使用过程测量进行管理，而从评价中得到的反馈信息将有助于干预措施的改进。过程评价研究的特点是用灵活的方法收集数据、用多种方法生成一系列不同类型的数据，包括定量和定性方法。干预措施的过程评价和结果评估可分开发表。当认为重要时，评价者应当强调该项试验是否解释或测量了关键的过程因素、而完全处理好完整性问题的试验是否具有更大的影响。过程评价可能是一个潜在影响干预效果的因素来源。然而要注意，测量盲法的成功（例如在安慰剂对照的药物试验中）可能没有价值（见第8章第8.11.1节）。

评价干预措施完整性的Cochrane系统评价例子是孕妇戒烟的系统评价（Lumley 2004）。评价者发现，干预措施的过程评价仅发生在一些试验中，而在其他试验的应用中相当不理想（包括一些大规模试验）。该系统评价显示，当干预措施从一个实施场地转移到另一个实施场地时，因干预措施的要素发生改变或文化方面存在差异，可能会降低其有效性。

7.3.5 结局测量

评价者应当预先决定他们是否收集研究中所测量的所有结局信息，或者只收集那些在系统评价中（预先设定的）感兴趣的结局。因为我们在7.3.6节建议只收集预先设定的结果，同时我们建议只有在计划书中列出的结局才详细描述。然而，所有测量结局的一个完整清单允许更详细评估由于选择性结局报告导致的偏倚风险（见第8章8.13节）需要。

可能重要结局的信息包括：

- 定义（诊断方法/量表名称/阈值定义/行为类型）；
- 时机；
- 测量单位（如果相关）；以及
- 量表：上下限，及高分还是低分数有利。

收集与量表相关的引用报告细节可能是有益的，因为这些可能包含上下限、效应值的方向、典型平均值和标准差、最小有重要意义的效应尺度及量表验证的进一步信息。

对经济学结局的进一步讨论在第15章（15.4.2部分），对病人报告的结局的讨论在17章。

不良结局

收集不良反应结局可能特别困难，具体在第14章讨论。归属于术语“不良反应”、“不良药物反应”、“副作用”、“毒副作用”、“不良事件”和“并发症”下的信息被视为评估一个干预措施不良影响时最合适提取的数据。此外，一个结局是否应当归类为不良结局可能尚不明确（而且同样的结局可能在有的研究中认为是不良结果，有的研究却不是）。未提及副作用并不一定意味着没有不良反应发生。通常最安全的假设是不良反应不确定或未被记录。生活质量指标是总体的测量指标，通常不包含干预措施的具体不良影响。虽然生活质量量表可以用来衡量整体水平，但它们不应该被视为详细评价安全性和耐受性的替代品。

应当准确记录不良反应结局的定义和强度，因为它们在不同研究间可能不相同。例如，在阿司匹林与胃肠道出血的一个系统评价中，有些试验仅报到了胃肠道出血，而另一些报到了具体出血类型，如呕血、便血和直肠出血（Derry 2000）。出血严重程度的定义和报告（例如重大、严重、需要入院）各试验也不尽相同（Zanchetti 1999）。此外，一个特定的不良反应可能在研究中用不同的方法描述或测量。例如，术语“疲劳”、“疲乏”或“昏睡”可能都用在副反应报告中。研究者也可能使用不同的阈值表达“异常”结果（例如，在诊断低血钾采用血钾浓度3.0 mmol/l或3.5 mmol/l）。

7.3.6 结果

应只收集计划书中指定的感兴趣结局的结果。除非方案修改增加了其它结局的结果，否则不应提取，而且这种修改应在系统评价中报告。然而，评价者应注意可能重要、意想不到的结果，特别是严重的不良反应。

研究报告往往包括同一结局指标的几种结果。例如，可能使用不同的测量尺度、结果可能会分别出现在不同的亚组，而且结局可能在不同时间点测量。选择的数据不同，结果的差异可能会非常大（Gotzsche 2007）。研究计划书应尽可能明确测量何种结局、时间点和合并的统计量（例如，终值和与基线差值相比较）。可能需要改进研究计划书以帮助决定哪些结果应提取。

7.7节描述了做Meta-分析所需的数据。如果不明显的话，分析单位（如受试对象、群组、身体部分、治疗期）应做记录（见第9章9.3节）。结局数据类型决定了将要寻找的每个结局数据的属性。例如，对一个二分类（“是”或“否”）结局，将找出受试对象人数和每一组经历结局的人数。重要的是要收集与每个结果相关的样本量，虽然这并不总是显而易见的。如果其不能在发表的报告中获得，制作CONSORT声明（Moher 2001）推荐的流程图有助于确定一项研究中的受试对象流程（可从这里下载www.consort-statement.org）。Metal-分析所要求的数据并非总是能够获得，有时需要收集其它统计数据并转化为要求的格式。例如，对于一个连续性结局，通常最方便的是找到每组的受试对象数、均值和标准差。这些通常不是直接可得，尤其是标准差，用替代统计数据可计算或估计丢失的标准差（如标准误、可信区间、检验统计量（如t检验或F检验）或P值）。详细情况见第7.7节。处理缺失数据的进一步探讨见第16章（16.1节）。

7.3.7 收集其它信息

每一个研究报告所需要提取的其他信息，包括引文、研究者详细联系资料、以及其它信息源的细节（例如能在一个临床试验注册库中找到该研究的标识符/注册号）。在许多领域特别重要的是研究资金来源，或研究者潜在的利益冲突。有些系统评价者希望收集研究特征信息，该信息会影响研究实施的质量但不太可能直接导致偏倚风险，例如是否得到伦理许可和是否进行了样本量计算。

我们建议评价者收集纳入研究的作者报告的重要结论。在系统评价中未必要报告这些结论，但是应当用来验证评价者的分析结果，特别是有关效果的方向。研究者的更多评论，例如对发现的任何非预期结果的解释都需要注意。研究报告中引用的其他研究参考文献可能是有用的，但评价者应当注意引文偏倚的可能性（见第10章10.2.2.3节）。

7.4 数据来源

7.4.1 报告

多数Cochrane系统评价获取的大部分数据来自研究报告。研究报告包括期刊论文、图书、学位论文、会议摘要和网站。但是要注意，这些资料的可信度以及详细级别不同。例如，会议摘要提供初步发现但是可能需要证实最后的结果。强烈建议建立数据提取表

格来收集研究报告数据（见7.6节）。

7.4.2 联系研究者

评价者往往发现他们无法从现有报告中提取所有寻求的信息，包括研究的详细资料和数值结果。在这种情况下，建议评价者联系原始研究作者。评价者需要考虑是否以开放式的要求联系研究作者，寻找具体信息，包括一个数据收集表（无论是未完成或部分完成），或寻求单个受试对象层面的数据。如果不能从研究报告中获得研究者的详细联系资料，往往可以从最近的一个其它出版物、或通过大学职员名单或综合检索互联网获得。

7.4.3 单个患者数据

原始研究数据可以直接从负责每项研究的研究者那里获得，而不是从研究出版物中提取数据。单个患者数据（Individual patient data，IPD）的系统评价是指从每一个研究中获取每一个患者的研究数据，是数据可用性的金标准。IPD可以集中再分析，如果合适，可在Metal分析中合并。IPD系统评价的详细信息在第18章介绍。

7.5 数据提取表

7.5.1 数据提取表的基本原理

数据收集表是连接原始研究者（例如，在杂志文章、摘要、个人通讯）与系统评价作者最终报告内容的桥梁。数据提取表具有几个重要功能（Meade 1997）。首先，表格是直接与系统评价问题及评价研究合格性标准相关，并且为鉴别研究报告提供了一个清晰的总结。其次，数据提取表是整个评价过程中大量决策的历史记录（决策的改变）。第三，表格是纳入分析的数据来源。

鉴于数据提取表的重要功能，在设计时应给予充裕的时间和思考。因为每个系统评价不同，其数据提取表也不同。但是，在重要信息类型方面有许多相似之处，因此可以将一个系统评价的表格改编后用于另一个系统评价。虽然我们使用的术语“数据提取表”为单数，但在实践中它可能是用于不同目的的系列表格：例如，在评价中，为便于快速确定应排出的研究，将制定评估纳入研究合格性的单独表格。

7.5.2 电子与纸质数据提取表

决定使用电子还是纸质数据提取表格很大程度上取决于评价者的偏好，纸质表格可能的优点如下：

- 方便或偏好；
- 数据提取几乎可以在任何地方进行；
- 易创建和实施（不需要计算机程序和专业软件）；
- 提供所有操作和修改的永久记录（如果没有擦去这些操作和修改）；
- 可简单比较不同评价者完成的表格。

电子表格可能的优势包括：

- 方便或偏好；
- 数据提取和输入一步完成；
- 表格可以编程（例如，使用Microsoft Access）来引导作者完成数据提取，例如，依据前一问题的答案提出问题；
- 纳入大量研究的系统评价其数据更易储存、分类和检索；
- 数据提取同时允许简单数据转化（例如，标准误转为标准差；磅转为公斤）；
- 快速比较不同评价者完成的表格；
- 环境因素。

已建立了满足两种方法多数优点的电子系统（包括商业SRS软件：见www.trialstat.com）。如果评价者计划使用电子制表软件或数据库软件建立自己的电子表格，我们建议（i）首先建立一个纸质表格，并由多个评价者在几个研究报告中试用；（ii）数据输入的结构要有逻辑性，回应编码尽可能简单和一致；（iii）检查结果输出与RevMan的兼容性；（iv）考虑数据记录、评价和纠正输入错误的机制。

7.5.3 数据提取表设计

当改编或设计一个数据提取表时，评价者首先应考虑应当收集多少信息。收集过多的信息会导致表格比原始研究报告更长，而且非常浪费时间。收集太少信息，或遗漏关键数据，将导致系统评价过程中需要重新返回研究报告。

以下是数据提取表设计的一些技巧，它基于非正式整理许多评价者的经验而成。在表7.3a中也已经列出咨询清单。

- 包括系统评价标题或一个唯一识别号。数据提取表适合多个系统评价，某些评价者参与了多个系统评价。
- 包括修改日期或数据提取表的版本号。表格有时需要修改，这样可以减少错误使用过期表格的机会。
- 记录完成表格的人名（或ID号）。
- 在表头附近留下空间以便注释，避免将注释、问题或提醒放在最不易被察觉的表格最后一页。重要提示可输入RevMan的“纳入研究特征”表的“备注”栏，或输入系统评价正文。
- 包括一个唯一的研究ID和报告ID。这为相同研究的多个报告间提供了一个链接。每个纳入研究必须有一个专一标识符，用于RevMan中（通常包括第一作者的姓和研究最初参考文献的出版年）。
- 表格开始处包括系统评价纳入研究的合格性评价（或核实）。那么表格的开头部分可用于评估合格性过程。这样就很容易从评估中推断出研究排出的原因。例如，如果仅仅真正的随机对照试验才合格，数据提取表的问题可能是：“随机？是、否、不清楚”。如果研究采用交替分配，问题的答案是“否”，这个信息可能被输入“排除研究特征”表，以此作为排除原因。
- 记录收集的每一条重要信息的来源，包括在报告的哪里发现 （可通过纸质版上突出显示来完成）或信息是否来自非出版源或个人通讯。未发表信息编码应与发表信息编码一致。
- 使用选项框或编码可节省时间。
- “是”或“否”回答旁边包括‘未报告’或‘不清楚’选项。
- 应考虑收集结果的格式与RevMan数据表匹配。数据提取表应有足够的灵活性，以满足数据报告的变化。强烈建议提取结果数据时应当采用报告中的格式（然后在随后的步骤中转化格式）。
- 除了收集初始（例如，随机的）数量外，收集结果数据时总是需要收集样本量。有可能由于失访或排除，不同的结局有不同的样本量。
- 留下足够的注释空间。

7.5.4 编码和解释

为所有使用数据提取表的作者提供详细的说明至关重要（Stock 1994）。说明可插入在数据表中相邻或相近的数据领域，或直接包含在数据单元格中（例如，作为Microsoft Excel中的一个评论）。如果该说明过长，可以单独提供在另一页上。使用编码方案效率高且有利于在系统评价中全面呈现研究特征。为使数据收集不易混淆或分类不会出错，应准确编码，且编码不应太复杂。应该对不同评价者使用的编码方案的一致性进行检查。

7.6 从研究报告中提取数据

7.6.1 引言

在大多数Cochrane系统评价中，每一个研究的主要信息来源通常是以期刊论文形式出版的研究报告。一个系统评价中最重要且最耗时的一项就是从研究报告中提取数据。数据提取表的设计通常会考虑到数据提取的问题。

电子全文检索有助于定位研究报告中的信息，例如PDF浏览器、internet浏览器和文字处理软件检索设施的使用。然而，由于信息可以使用不同的术语来呈现，因此全文检索不应视为可替代阅读报告。

7.6.2 谁应提取数据

为缩小误差和减少潜在偏倚，强烈建议评价者应一人以上提取每个研究报告数据。至少当涉及包括主观解释和对结果阐述重要的信息（例如，结局数据）时应当至少两人独立提取。按照遴选程序执行（7.2.4节），数据提取者最好来自互补（交叉）学科。例如，一个方法学家和一个主题领域的专家。重要的是，所有参与数据提取的人都应练习过使用这个表格，如果提取表由别人设计，数据提取者则应接受适当的培训。

支持双重数据提取的证据间接源于几方面因素。一项研究发现：两个作者独立提取数据所犯的差错少于一个作者提取数据并由另一人核实（Buscemi 2006）。已经观察到数据提取的错误发生率高（34个系统评价中的20个有错误）（Jones 2005）。一项关于计算标准化均差的数据提取研究发现：27篇系统评价中至少有7篇有重大错误（Gotzsche 2007）。

7.6.3 数据提取的准备

所有表格应当使用具有代表性的评价研究报告进行预测试。这种测试可确定数据提取表中遗漏或多余的内容。使用这些测试报告来起草“纳入研究特征”表（第11章11.2节）和“偏倚风险”表的条目是明智的（第8章8.5节）。表格用户可提供某些编码指示不清或不完整的反馈建议（如，一个选项列表可能不包括所有情况）。修改表格之前，如果有可能，评价者之间最好达成共识以避免任何误解或之后的分歧。在首次测试后如果要大修改，可能需要用一组新的报告重复预测试。

有时数据提取表格问题在预测试完成后出现，这时即使已经开始提取数据，可能仍需要修改表格。事实上，在预测试完成后需要修改数据提取表很常见。当表格或编码说明改变时，有必要返回已经完成数据提取的报告。在某些情况下，可能仅仅需要阐明编码说明而不需要修改实际数据提取表。

有人提出，研究报告中的有些信息如作者，在数据提取和偏倚风险评估之前应对系统评价作者采用盲法（Jadad 1996）；另见第9章（8.3.4节）。然而，不让系统评价作者知道研究报告的某些方面并不是Cochrane系统评价的常规推荐（Berlin 1997）。

7.6.4 从同一研究的多个报告中提取数据

研究常常不止在一个出版物中报告（Tramèr 1997，von Elm 2004）。然而，Cochrane干预性系统评价中关注的单元是研究而非研究报告。因此，需要核对多个报告的信息。放弃任何一个纳入研究的报告都不恰当，因为某些有价值的信息可能并未包括在主要研究报告中。评价者需要在这两个策略之间决策：

- 分别从每一个报告中提取数据，然后把多个数据提取表信息合并。
- 从所有的报告中提取数据并直接将信息放入一个提取表。

选择使用哪个策略将取决于报告属性，不同研究及不同报告间不同。例如，如果有一个完整的期刊文章和多个会议论文，可能大多数信息将从期刊文章获得，而从每一篇会议文摘中完成一个新的数据收集表可能会浪费时间。相反，如果有两个或更多详细的期刊文章，它们或许有不同的随访时期，那么对这些文章进行单独数据提取，而后再从数据提取表中整理信息可能会更容易。

如CONSORT声明中推荐的（Moher 2001），绘制研究的受试者流程图对核查来自多个报告的信息特别有帮助。

7.6.5 可靠性和达成共识

不止一人从同一研究报告中提取数据时，可能会有意见分歧。应当在方案中确定辨别和解决分歧的明确程序或决策规则。大多数情况下，分歧是由一个数据提取者造成的错误产生的，这种情况容易解决。因此，首先在系统评价作者之间讨论解决是很明智的。少数情况下，发生的分歧可能也需要第三方来仲裁。任何不能解决的分歧应当联系研究作者；如果不成功，则应在系统评价中报告。

应认真记录分歧原因和解决过程。保留一份“原提取”（除了共识数据之外）的数据，以评价编码的可靠性。可以实现的方法包括：

- 采用一位评价者的数据提取表并用不同颜色笔记录统一认识后的变更。
- 使用一个单独表格来记录共识数据。
- 将达成共识的数据输入一个电子表格中。

虽然Cochrane系统评价未常规使用，但编码项目的一致性可以量化，例如使用Kappa统计量（Orwin 1994）。简单计算两个作者一致性的描述见7.2.6节。如果要做一致性评估，那么仅针对最重要的数据（例如，重要偏倚风险评估，或可获得的关键结局）即可。

应在整个评价过程中考虑数据提取的可靠性。例如，若开始几个研究达成共识后，系统评价作者发现某个数据频繁出现分歧，那么可能需要修改编码说明。此外，由于遗忘了编码规则，作者的编码方式可能随时间而改变，表示需重新进行培训，有些可能需重新编码。

7.6.6 总结

- 总之，方案和系统评价的方法部分应当详述：
- 收集的数据类型；
- 如何验证从每一个报告中提取的数据（例如，两个评价者独立提取的数据）；
- 数据是由主题专家、还是方法学家提取，还是两者都参与；
- 应对数据提取表进行预测试和培训，并应具备数据提取表的编码说明；
- 如何从同一研究的多份报告中提取数据；
- 如果一个以上作者从报告中提取数据，如何处理分歧。

7.7 提取研究结果和转化成需要的格式

7.7.1 引言

现在我们要概述分析二分类结局、连续性结局和其它类型结局数据时需要从每一个报告中提取的数据。这些数据类型讨论见第9章（9.2节）。常常是分别收集每一个干预组的汇总数据并输入RevMan，然后计算效应估计值。有时只能间接获得所需要的数据，而且相关结果并不明显。本章提供一些有用的方法和技巧，来处理这类情况中的一些问题。如果不能从每一个干预组获得汇总数据，可能会直接报告效应估计值。在7.7.7节我们描述了如何从可信区间和P值中获得效应估计值的标准误。

7.7.2 二分类结局的数据提取

二分类数据在第9章9.2.2节描述，有关其Meta分析在第9章9.4.4节描述。二分类结局需要的数据仅是每一个干预组中两类结局的分别数量(要求填写在四格表SE, FE, SC, FC的数字见第9章表9.2a)。这些数字可作为两组发生结局的人数和总样本量输入RevMan中。收集二分类结局数据最可靠的方式是收集每组明确经历和未经历结局的数量。虽然在理论上这相当于收集的总数和发生结局的数量，但并不总是很清楚报告的总人数是否为测量结局的人数。偶尔，经历了事件的人数需要从百分比中计算（尽管不总是很清楚使用哪个作分母，但四舍五入后的整数百分比不止与一个分子兼容)。

有时候受试者人数和事件数量不详，但可能会报告诸如比值比或风险比这样的效果估计，例如在会议文摘中。这种数据可采用反方差法纳入Meta分析，但只有同时报告了不确定性测量指标如标准误、95%可信区间或确切P值等指标才能使用：见7.7.7节。

7.7.3 连续性结局数据提取

连续性数据在第9章9.2.3节描述，其Meta分析在第9章9.4.5节讨论。使用均数差什值（均差）或标准化均数差值（标化均差）做连续性数据的Meta分析的评价者需寻求：

- 各干预组结局指标的平均值（ME and MC)；
- 各干预组结局指标的标准差（SDE and SDC)；
- 各干预组测量结局的参与者人数（NE and NC)。

由于报告质量差及差异大，可能难以或无法从数据汇总中获得必要的信息。不同研

究采用不同统计量总结平均数（有时用中位数而不是均数）和变异情况（有时使用标准误、可信区间、四分位距和极差/全距而不是标准差）。也有选择不同的刻度值来分析数据（例如，干预后测量值与基线改变值；原始数据值与对数值）。

把标准误当作标准差是经常发生的一种曲解。遗憾的是，并不总是很清楚报道的是什么，而且可能需要推理以及与其它研究比较。标准差和标准误在研究报告中有时混淆，而且使用术语不一致。

必要时，应向作者索取遗漏信息和核实报告的统计量。但是，一些变异性测量指标与标准差有近似或直接代数关系，所以即使未发表在论文中也可能获得所需要的统计数据，相关信息见第7.7.3.2节到7.7.3.7节。更多详情和例子可在其它地方获得（Deeks 1997a, Deeks 1997b）。第16章（16.1.3节）讨论了如果在试图获得标准差后仍然缺失的选择。

有时候参与者人数、均数和标准差不详，但是报告了效果估计值均差或标准化均差，例如会议文摘中。在这种情况下可采用逆差法将这些数据纳入Meta分析，但只有同时报告了不确定性测量指标如标准误、95%可信区间或确切P值才能这么做。从可信区间获得均数差值的合适标准误，可采用7.7.3.3节中描述的初期步骤。关于标准化均差见7.7.7节。

7.7.3.1 干预后值与基线改变值

连续数据的一个共同特征是：评估每一个参与者结局的指标，在基线时也会测量，即实施干预前要测量。为此，有可能用基线改变值（也称改变评分）作为主要结局。建议评价者不要关注基线改变值，除非这种分析方法应用在某些研究报告中。

当处理基线改变值时，为每个参与者创建了一个单一的测量值，由基线测量结果减去最后测量结果获得还是最后测量结果减去基线测量获得都可以。此时，如同任何其它类型的连续结局变量一样，采用改变值进行分析，而不是用最后测量结果进行分析。

通常，纳入系统评价的研究可能混合使用基线改变值和终值。有些研究报告二者；另一些则仅仅报告改变值或终值。如第9章（9.4.5.2节）解释，终值和改变值有时会合并在同一个分析，所以这并不一定是难题。如果需要的均数和标准差可获得，评价者可能希望同时提取基线改变值和终值。选择哪一个数据分析的关键问题是：是否存在选择性报告一种夸大结果的可能性，而评价者应寻找存在这种可能性的证据（见第8章8.13节）。

最后一个提取基线改变值信息的问题是：通常由于失访和退出研究，基线测量值和最终测量值报告的参与者人数不同。可能难以确定同时报告了基线测量值和最终测量值

的参与者人数并计算他们的差值。

7.7.3.2 从标准误和可信区间获得每组均数的标准差

标准差可以通过均数的标准误乘以样本量的平方根获得：$SD = SE \times \sqrt{N}$

当进行这一转化时，标准误必须是从一个干预组计算得到的均数标准误，而不是计算不同干预组的均数差值的标准误。

也可用均数的可信区间计算标准差。同样，以下适用于从一个干预组计算得到的均值的可信区间而非干预组之间差值的估计（这些见7.7.3.3节）。大多数可信区间是95%的可信区间。如果样本量很大（即每一组大于100），95%可信区间是3.92标准误范围（3.92=2x1.96）。每一组的标准差等于可信区间长度$SD = \sqrt{N} \times (\text{upper limit} - \text{lower limit})/3.92$

对于90%可信区间应当用3.29代替3.92，而对于99%可信区间应当用5.15代替。

如果样本量小（即每组少于60），那么可信区间应当使用一个t分布值来计算。这些数字3.92、3.29和5.15需要用更大一点的t分布值代替，t分布值可以用自由度（等于组样本量减去1）从t分布表得到。有关t分布的详细资料可从许多统计学书附录或标准的计算机电子表格软件包获得。例如样本量为25的95%可信区间的t值，可通过键入=tinv（1-0.95,25-1）在Microsoft Excel电子表格的单元格里而获得（结果是2.0639）。上述公式中的除数3.92将改为2×2.0639=4.128。

中等规模样本量（即介于60与100之间），可使用t分布或标准正态分布。评价者应当寻找使用何种的证据，如果疑惑的话可使用t分布。

以下为例，考虑如何展示如下数据：

组别	样本量	均数	95%CI
试验组	25	32.1	（30.0, 34.2）
对照组	22	28.3	（26.5, 30.1）

可信区间应当基于自由度分别是24和21的t分布。从上面可见，试验干预组的除数是4.128，该组的标准差是$\sqrt{25} \times$（34.2 – 30.0）/4.128 = 5.09。用相似的方法可计算对照组的标准差。

重要的是检查可信区间是以均值对称（上下限与均值的距离应当是同样的）。如果不是这样，可能需要计算转化值的可信区间（见7.7.3.4节）。

7.7.3.3 从标准误、可信区间、t 值和 P 值获得均数差的标准差

标准差可从两组间均数差的标准误、可信区间、t值或P值获得。均数差本身（MD）要求从t值或P值计算得到。所有情况下的假设是：两组结局测量值的标准差应相同，那么该标准差可用于两个干预组。首先，我们描述如何从P值得到t值，然后是如何从t值或可信区间获得标准误，最后是如何从标准误获得标准差。系统评价者在这个过程中根据他们可获得什么样的结果选择合适的步骤。相关方法可用于从确定的F统计量计算标准差，因为计算F值的平方根可产生同样的t值。常常需要谨慎确保使用一个合适的F值，并建议听取资深统计学家的意见。

从P值到t值

实际P值可通过t检验获得，相应t值可从一个t分布表获得。自由度为NE + NC – 2，NE 和NC 分别是试验组和对照组样本量。我们将举例阐述。假设比较试验干预组（NE = 25）与对照组（NC = 22），组间均数差是MD=3.8。据此比较P值是P=0.008，使用一个两样本的t检验获得。

与P=0.008和自由度是25+22-2=45对应的 t值为t=2.78。这可以通过一个自由度45的t分布表或一个计算机（例如，通过输入=tinv（0.008,45）到Microsoft Excel电子表格任意单元格）获得。

当只报告显著水平（如P<0.05或通常P=NS表示P>0.05），而不是精确的P值时就会遇到困难。一个保守的做法是采取上限P值（例如P<0.05取P=0.05，若P<0.01取P=0.01，若P<0.001取P=0.001）。不过，这不是报告P=NS时的一个解决方案：见7.7.3.7节。

从t值到标准误

t值是均数差与均数差的标准误的比值。因此均数差的标准误等于均数差除以t值：

$$SE = \frac{MD}{t}$$

在该例中，均数差的标准误是3.8除以2.78得到1.37.

从可信区间到标准误

如果均数差有95%的可信区间，那么同样的标准误就可计算为：

$$SE = (\text{upper limit} - \text{lower limit})/3.92$$

只要试验样本很大，对于90%的可信区间3.92应改为3.29，对99%的可信区间而言应当改为5.15。如果样本量很小，那么可信区间用t分布来计算。这些数字3.92，3.29和5.15需要更换为t分布和样本量所对应的更大的数字，可以用自由度（等于NE + NC – 2，NE和NC分别是两组的样本量）从t分布表计算。相关t分布的详细信息可在许多统计学书的附录或计算机电子表格软件包中获取。例如，样本量为25和22的比较对应的95%可信区间的t值，可在Microsoft Excel电子表格的单元格中输入=tinv（1-0.95,25+22-2）得到。

从标准误到标准差

组内标准差可用均数差的标准误通过以下公式计算得到：

$$SD = \frac{SE}{\sqrt{\frac{1}{N_E} + \frac{1}{N_C}}}$$

在这个例子中，

$$SD = \frac{1.37}{\sqrt{\frac{1}{25} + \frac{1}{22}}} = 4.69$$

请注意，这里标准差是试验组和对照组平均标准差，并应输入RevMan两次（每个干预组一次）。

7.7.3.4 数据转换和偏态数据

报告的汇总统计量可能是原始数据转换之后的。例如，可得到对数值的均数和标准差（或，等同的，其几何均数和可信区间）。这样的结果应当收集，因为它们可能纳入Meta分析，或在某种假设下，可能转换回原始数值。

例如，某试验报告了接种C型脑膜炎疫苗和对照疫苗12个月后脑膜炎双球菌抗体情况（MacLennan 2000），几何平均滴度分别为24和4.2，95%可信区间分别为17-34和3.9-4.6。这些总结通过找到抗体滴度自然对数值的均数和可信区间获得（疫苗组3.18：95%CI（2.83-3.53），对照组1.44（1.36-1.53）），并取它们的指数（反对数）。基于这些自然对数的抗体反应值就可进行Meta分析。而对数转换后数据的标准差可用7.7.3.2描述的方法从后面的一对可信区间获得。有关偏态数据Meta分析的更多讨论见第9章（9.4.5.3节）。

7.7.3.5 中位数和四分位间距

当数据分布对称时，中位数与均数很类似，因此中位数有时直接用于Meta分析。然

而，如果数据为偏态分布，均数和中位数可能有很大的不同，经常因数据偏态而报道中位数（见第9章9.4.5.3节）。

四分位间距描述的是中心50%研究对象结局分布的位置。当样本量大和结局分布类似正态分布时，四分位间距宽度大约是标准差的1.35倍。在其它情况下，特别是当结局分布为偏态时，不能从四分位间距来估计标准差。注意使用四分位间距而不是标准差时，通常可看作是结局为偏态分布的一个指标。

7.7.3.6 极差

和其它变量指标相比，极差是非常不稳定的，且随样本量增加而增大。它描述的是观察结局的极限值而不是平均变量。极差不应当用来估计标准差。一个常见的方法就是：数据如为正态分布数据，其95%的值将位于均数两侧的2×SD范围内。因此，SD估计约为典型数据值范围的四分之一。这种方法不成熟，我们建议不要使用。

7.7.3.7 无变异性数据信息

如果上述方法均不能从试验报告中（而且信息不能从试验员处得到）计算出标准差，那么，为了作Meta分析，作者可能被迫为丢失数据赋值（填写值）或从Meta分析中排除该研究：见第16章（16.1.3节）。也可使用描述方法合成信息。即使不能纳入进行正式的Meta分析，将纳入系统评价的所有研究可获得的结果列表展示也是有价值的。

7.7.3.8 合并组

有时希望将报告的两个亚组合成一个组。例如，如果一个研究分别报告了每一个干预组中男女的样本量、均数和标准差，就可以如此做。可用表7.7.a中的公式将每一个干预组的数据合并为一个样本量、均数和标准差（即，该例中合并男性与女性）。看起来相当复杂的SD公式产生的结局指标的SD，犹如该合并组从未被分成过二个组。该标准差的近似值是采用常规的合并标准差获得，它提供了一个比期望标准差略微低的估计值。

这些公式也适用于多个干预组比较的研究，可将两个干预组合并成一个干预组（见第16章16.5节）。例如，第1组和第2组可能是一个干预措施的两种变异形式，而参与者被随机分配入其中一个变异组。

如果有2个以上组合并，最简单的方法是依次应用上述计算公式（即合并组1和组2

变成组“1+2”，然后合并组“1+2”和组3变成组“1+2+3”，依此类推）。

表7.7.a　合并组的计算公式

	组 1 （如男性）	组 2 （如女性）	合并组
样本量	N1	N2	N1 + N2
均数	M1	M2	$\frac{N_1M_1 + N_2M_2}{N_1 + N_2}$
标准差	SD1	SD2	$\sqrt{\frac{(N_1 - 1)SD_1^2 + (N_2 - 1)SD_2^2 + \frac{N_1N_2}{N_1 + N_2}(M_1^2 + M_2^2 - 2M_1M_2)}{N_1 + N_2 - 1}}$

7.7.4　等级结局数据的提取

结局被分为几种有序类别时，称为等级数据，其相关内容在第9章9.2.4节描述，其Meta分析在第9章9.4.7节讨论。等级结局需要提取的数据取决于是否将等级数值转化为二分类进行分析（见7.7.2节、是否以连续性结局变量处理（见7.7.3节）或直接作为等级数据分析。这一决定也将影响研究者采用哪一种方法分析数据。因此，可能无法预先指定数据提取是否会涉及根据某个确定的阈值计算高于和低于该阈值的参与者人数，或均值和标准差。实际上，在不清楚哪一种是最常用数据的情况下，明智的做法是提取文中报告的所有数据形式，直到系统评价完成为止。在某些情况下，一个系统评价有理由纳入一种以上的分析方式。

等级数据转化为二分类数据时，有几种选择切点（或任意选择切点）的方式，从一开始就计划用敏感性分析研究切点选择的影响是明智的（见第9章9.7节）。要做到这一点，有必要收集用于每一种二分法的数据。因此，更可取的是记录每类等级数据的数量，以避免从文章中多次提取数据。当研究使用略有不同的短序等级时，这种记录所有类别的方法也是明智的，目前还不清楚是否有一个所有研究共同的用于二分类法的切点。

如果使用比例比值比法，也有必要记录每一个干预组每类等级数据的数量（见第9章9.2.4节）。

7.7.5 计数数据的提取

计数资料在第9章9.2.5节描述，其Meta分析在第9章9.4.8节讨论。本身就是计数的数据可使用好几种方法分析。至关重要的决定在于：是否将感兴趣的结局分为二分类、连续、时间事件或率。一个常见的错误是直接把计数资料当作二分类数据，将受试者总数或随访的总人年数作为样本量大小。这些方法都不适合每一个受试者发生不止一次事件的情况。这可能导致事件总数超过样本量，出现无意义的结果。虽然最好提前决定怎样分析计数数据，但选择通常由现有数据的格式决定，因而在大多数研究评价之前是不能决定的。因此，评价者一般都以研究报道的格式提取计数数据。

有时没有提供事件数和暴露人群人年数的详细数据，但可获得从它们计算出的结果。例如，在会议文摘中可能报告了比值比或率差的估计。只有同时报告了这些数据的不确定指标如95%可信区间时才可能纳入Meta分析：见7.7.7节。为此可以计算出标准误和使用反方差法做Meta分析。

7.7.5.1 计数资料作为二分类数据提取

如果认为结局是一个二分类资料结局，作者必须确定每一个干预组的参与者人数、以及每个干预组至少经历一个事件的人数（或其它将所有参与者划分到两个可能组之一的合适标准）。尽管可能创建一系列的二分类结果数据，但通过这种方法会导致数据的时间元素丢失，例如，随访的第一年至少有一个中风事件发生，随访的前两年至少有一个中风事件发生，等等。可能难于从发表的报告中获得这些数据。

7.7.5.2 计数资料作为连续性数据提取

当作连续性资料提取计数资料（即每名患者发生事件的平均数）时，应当遵循章节7.7.3中的指导，应特别注意这些数据可能明显呈偏态分布。

7.7.5.3 计数资料作为时间事件数据提取

对于可能发生一次以上的罕见事件，作者可能要面对以首次事件发生的时间处理数据的研究。以时间事件提取计数资料时，应当遵循章节7.7.6的指导。

7.7.5.4 计数资料作为率数据提取

如果能够提取每一组中事件总数，以及每一组暴露人年数的总数，那么计数数据可作为率数据来分析（见第9章9.4.8节）。注意对于率数据分析不要求受试者总数，但是作为描述研究的一部分应当记录。

7.7.6 时间事件结局数据的提取

有关时间事件结局在第9章9.2.6节描述，其Meta分析在第9章9.4.9节讨论。时间事件数据的Meta分析通常包括从原始研究者那里获得单个患者数据，再分析数据获得对数风险比的估计值及其标准误，然后进行Meta分析（见第18章）。使用已发表的论文或试验报告的汇总信息进行Meta分析通常是有瑕疵的，因为通常未呈现最合适的汇总统计量。无论是使用单个患者数据还是汇总数据，有两种方法可用来获取对数风险比的估计值及其标准误，在Meta分析中使用倒方差法（逆差法）；对于实际指导，评价者应当参考Tierney等（Tierney 2007）。

第一种方法可从对数秩分析时计算的统计量中获得对数风险比的估计值。如果采用这一方法，建议与资深统计学家协作。对数风险比（试验组与对照组比）通过（O – E）/V公式估算，标准误是1/√V，O是试验组观察到的事件数，E是试验组的对数秩期望事件发生数，O – E是对数秩统计量，V是对数秩统计量的方差。因此有必要获取每一研究的O-E值和V值。

如果能获得单个患者的数据，很容易计算这些统计量，有时可从引用统计数据和生存曲线中提取（Parmar 1998，Williamson 2002）。或者，有时候可使用每一个试验的每一干预组的汇总资料。例如，假设数据包括第一年、第二年等发生事件的受试者人数，及每年末未发生事件以及仍在随访的人数。虽然通过对数秩处理这些数据可以得到O-E值和V值，但是需要仔细考虑截尾时间的处理。由于粗分组的对数风险比仅仅是粗略估计，在有的系统评价中被称作对数比值比（早期乳腺癌试验协作组 1990）。如果时间间隔大，更合适的方法是基于区间截尾生存分析方法（Collett 1994）。

如果试验者采用Cox风险比例模型来分析数据，或如果Cox模型适合单个患者数据，则采用第二种方法。Cox模型可直接推导出对数风险比估计值和标准误（因此可做逆差法Meta分析）。如果报告中引用了风险比例和可信期间或P值，如在章节7.7.7中描述可获得标准误估计。

7.7.7 效应估计值数据的提取

7.7.7.1 效应估计值和逆差 Meta 分析法

在有些系统评价中，总体效应估计值来自每一个研究而非每一个干预组的汇总数据。例如，对非随机研究、交叉试验、整群随机对照试验、或具有事件结果的研究可能是如此。如果其标准误可获得的话，在RevMan中使用逆差法对此类效应估计值进行Meta分析（见第9章9.4.3节）。当从非随机对照研究，以及一些随机对照研究中提取数据时，可能获得校正后的效应估计值（例如，从Logistic回归分析获得的校正的比值比，或从Poisson回归分析获得的校正的率比）。尽管应记录被校正的变量（见第13章13.6.2节），但数据提取和使用逆差法进行分析的过程与未校正的效应估计值是一样的。

有时，可能要寻找每个干预组的汇总数据（例如，事件和受试者数，或均数和标准差），但不能被提取。在这种情况下，使用逆差法仍然可将研究纳入Meta分析。这一方法的局限是，即使提供了干预组的汇总数据，也必须计算同一Meta分析中所有纳入研究的同一效应指标的效应估计值和标准误。例如，如果已知有些研究中干预组每个结果类型的数据，但其它研究仅仅比值比（ORs）可用，那么需要计算第一组研究的OR，然后输入RevMan用逆差法与第二组研究一起进行Meta分析。RevMan可以用来计算这些ORs（它们作为二分类数据输入），而且RevMan计算出的可信区间使用下面方法可转化为标准误。

关注的效应指标估计值可能与可信区间或P值一起报告。通常希望通过这些数字计算出标准误，以便在RevMan中用逆差法进行Meta分析。获得标准误的程序有赖于效应指标是一个绝对测量值（例如，均数差、标准化均数差、危险差）还是一个比值测量值（例如，比值比、危险比、风险比、率比）。我们分别在7.7.7.2节和7.7.7.3节描述这些过程。但是，对于连续结果测量，从一个干预组提取均数结果，以及提取两均数差的结果的特殊情况，在7.7.3节讨论。

7.7.7.2 从可信区间和 P 值获得标准误：绝对（差异）测量

如果一个干预措施效应的绝对测量的95%可信区间可用（例如，SMD、危险差、率差），那么标准误可用如下方法计算

SE = （上限 – 下限） / 3.92.

90%可信区间则除以3.29而不是3.92；99%可信区间则除以5.15

当同时报告了精确P值和干预效应估计值时，可用来估计标准误。虽然统计学意义的所有检验产生了P值，但不同的检验使用不同的数学方法获得P值。此处的方法假定P值已通过特别简单的方法，即用效应估计值除以标准误差获得，然后用这一结果（用Z表示）与标准正态分布比较（统计学家通常称这为Wald检验）。若显著检验采用的是其它数学方法，则获得的估计标准误与真实的标准误并不完全一致。

第一步是要从一个标准正态分布表中获得与报告的P值对应的Z值。那么标准误的计算如下：

SE = 干预效应估计值 / Z.

作为一个例子，假定一个会议文摘提供一个0.03的危险度差估计值（P=0.008）。P值为0.008对应的Z值是Z=2.652。这可以从一个标准正态分布表或计算机（例如，通过输入=abs（normsinv（0.008/2）进MicroSoft Excel电子表格的任何单元格）获得。危险度差的标准误是危险度差（0.03）除以Z值（2.65）得0.011。

7.7.7.3 从可信区间和 P 值计算标准误：率测量

获得率测量值的标准误的过程类似于获取绝对测量值的标准误，但是有额外的第一步。利用自然对数表做率测量分析（见第9章9.2.7节）。对于一个率的测量，如风险比、比值比或危险比（我们在这里统称为RR），先计算

下限=In（RR的可信下限）

上限=In（RR的可信上限）

干预效果估计值=InRR

然后可使用7.7.7.2节中的公式。注意标准误是指率测量的对数，当在RevMan中使用逆差法时，数据应按自然对数表输入，也就是正如此处计算的InRR和InRR的标准误（见第9章9.4.3节）。

7.8 数据管理

可以在纸质数据提取表上提取数据，再直接输入RevMan。但经常需要在输入RevMan之前，在中间的计算机软件中处理数据。多种软件和数据管理程序可能对此有帮助，包括电子表格软件（例如，MicroSoft Excel）和数据库程序（如，MicroSoft Access）。

例如，在电子表格中整理提取的研究信息并做成表格有助于对研究分类比较和分亚组。此外，统计转换，例如标准误转换成标准差，理想的方法是使用计算机而不是手持计算器，因为它能够对原始数据、计算的数据以及所用的计算方法永久记录和保存。

7.9 本章信息

编辑：Julian PT Higgins and Jonathan J Deeks.

本章引用格式：Higgins JPT, Deeks JJ (editors). Chapter 7: Selecting studies and collecting data. In: Higgins JPT, Green S (editors), Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.

致谢：本章是早期手册版本的改版。有关先前手册作者和编者详细信息见第1章（1.4节）。文本由Andrew Herxheimer, Nicki Jackson, Yoon Loke, Deirdre Price and Helen Thomas所写。Stephanie Taylor和Sonja Hood提出了设计数据提取表的建议。我们感谢Judith Anzures, Mike Clarke, Miranda Cumpston 和 Peter Gotzsche的有益建议。

7.10 参考文献

Berlin 1997

Berlin JA. Does blinding of readers affect the results of Meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. The Lancet 1997; 350: 185-186.

Buscemi 2006

Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. Journal of Clinical Epidemiology 2006; 59: 697-703.

Collett 1994

Collett D. Modelling Survival Data in Medical Research. London (UK): Chapman & Hall, 1994.

Cooper 1989

Cooper H, Ribble RG. Influences on the outcome of literature searches for integrative research reviews. Knowledge 1989; 10: 179-201.

Dane 1998

Dane AV, Schneider BH. Program integrity in primary and early secondary prevention: are implementation effects out of control? Clinical Psychology Review 1998; 18: 23-45.

Deeks 1997a

Deeks J. Are you sure that's a standard deviation? (part 1). Cochrane News 1997; Issue No. 10: 11-12.（Available from www.cochrane.org/newslett/ccnewsbi.htm）.

Deeks 1997b

Deeks J. Are you sure that's a standard deviation? (part 2). Cochrane News 1997; Issue No. 11: 11-12. (Available from www.cochrane.org/newslett/ccnewsbi.htm).

Derry 2000

Derry S, Loke YK. Risk of gastrointestinal haemorrhage with long term use of aspirin: Meta-analysis. BMJ 2000; 321: 1183-1187.

Dusenbury 2003

Dusenbury L, Brannigan R, Falco M, Hansen WB. A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. Health Education Research 2003; 18: 237- 256.

Early Breast Cancer Trialists' Collaborative Group 1990

Early Breast Cancer Trialists' Collaborative Group. Treatment of Early Breast Cancer. Volume 1: Worldwide Evidence 1985-1990. Oxford (UK): Oxford University Press, 1990. (Available from www.ctsu.ox.ac.uk).

Edwards 2002

Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. Statistics in Medicine 2002; 21: 1635-1640.

Gotzsche 2007

Gotzsche PC, Hróbjartsson A, Maric K, Tendal B. Data extraction errors in Meta-analyses that use standardized mean differences. JAMA 2007; 298: 430-437.

Jadad 1996

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, McQuay H. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Controlled Clinical Trials 1996; 17: 1-12.

Jones 2005

Jones AP, Remmington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. Journal of Clinical Epidemiology 2005; 58: 741-742.

Lumley 2004

Lumley J, Oliver SS, Chamberlain C, Oakley L. Interventions for promoting smoking cessation during pregnancy. Cochrane Database of Systematic Reviews 2004, Issue 4. Art No: CD001055.

MacLennan 2000

MacLennan JM, Shackley F, Heath PT, Deeks JJ, Flamank C, Herbert M, Griffiths H, Hatzmann E, Goilav C, Moxon ER. Safety, immunogenicity, and induction of immunologic memory by a serogroup C meningococcal conjugate vaccine in infants: A randomized controlled trial. JAMA 2000; 283: 2795-2801.

Meade 1997

Meade MO, Richardson WS. Selecting and appraising studies for a systematic review. Annals of Internal Medicine 1997; 127: 531-537.

Moher 2001

Moher D, Schulz KF, Altman DG. The CONSORT Statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. The Lancet 2001; 357: 1191-1194. (Available from www.consort-statement.org).

Orwin 1994

Orwin RG. Evaluating coding decisions. In: Cooper H, Hedges LV (editors). The Handbook of Research Synthesis. New York (NY): Russell Sage Foundation, 1994.

Oxman 1993

Oxman AD, Guyatt GH. The science of reviewing research. Annals of the New York Academy of Sciences 1993; 703: 125-133.

Parmar 1998

Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform Meta-analyses of the published literature for survival endpoints. Statistics in Medicine 1998; 17: 2815-2834.

Stock 1994

Stock WA. Systematic coding for research synthesis. In: Cooper H, Hedges LV (editors). The Handbook of Research Synthesis. New York (NY): Russell Sage Foundation, 1994.

Tierney 2007

Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into Meta-analysis. Trials 2007: 16.

Tramèr 1997

Tramèr MR, Reynolds DJ, Moore RA, McQuay HJ. Impact of covert duplicate publication on Meta-analysis: a case study. BMJ 1997; 315: 635-640.

von Elm 2004

von Elm E, Poglia G, Walder B, Tramèr MR. Different patterns of duplicate publication: an analysis of articles used in systematic reviews. JAMA 2004; 291: 974-980.

Williamson 2002

Williamson PR, Smith CT, Hutton JL, Marson AG. Aggregate data Meta-analysis with time-to-event outcomes. Statistics in Medicine 2002; 21: 3337-3351.

Zanchetti 1999

Zanchetti A, Hansson L. Risk of major gastrointestinal bleeding with aspirin （Authors' reply）. The Lancet 1999; 353: 149-150.

（何林译，岑啸、张龙浩初审）

第八章　纳入研究的偏倚风险评价

编辑：Julian PT Higgins, Douglas G Altman and Jonathan AC Sterne 代表 Cochrane 统计方法学组和 Cochrane 偏倚方法学组。

本文选自工作手册5.1.0版本。有关如何引用它的指南，见8.16节。该手册的早些版本（5.02）由John Wiley＆Sons有限公司以“Cochrane系列丛书”印记出版，Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。电子邮件（供订购及客户服务查询）地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 单个卫生保健干预研究的设计和执行问题使研究结果的真实性受到质疑；实证证据支持该顾虑。
- 对 Cochrane 评价纳入研究的真实性进行评估，应强调其结果的偏倚风险，即高估或低估干预措施真实效应的风险。
- 许多工具可用于评估临床试验方法学质量，但我们不推荐使用量表产生总分。
- Cochrane 协作网推荐一种用于评估纳入研究偏倚风险的工具，由对“偏倚风险”表中每个条目的判断及判断的依据组成，每个条目都代表纳入研究某方面的特征。每个条目的判断涉及评估偏倚风险为“低风险”、“高风险”和“不清楚”，后者表示信息缺乏或对潜在偏倚不确定。

- 偏倚风险评估图可通过 RevMan 软件绘制。
- 临床试验中，偏倚的类型有：选择性偏倚、实施偏倚、测量偏倚、随访偏倚、报告偏倚和其他未能分类的偏倚。
- 对平行对照试验而言，Cochrane 系统评价标准的“偏倚风险”评估表条目有：随机序列生成（选择性偏倚）、分配序列隐藏（选择性偏倚）、对受试者和研究人员实施盲法（实施偏倚）、对结果测量者实施盲法（测量性偏倚）、结果数据不完整（随访偏倚）、选择性报告（报告偏倚）以及其他潜在来源的偏倚。
- 本章将提供评估这些条目的详细注意事项。

8.1 引言

Cochrane系统评价得到的某个干预措施效果的结论，其可信度取决于纳入研究数据和结果是否真实。尤其是Meta分析纳入不真实的研究可能会导致误导的结果，产生一个围绕错误的干预效果估计值的狭窄可信区间。因此，纳入研究的真实性评价是Cochrane系统评价的重要部分，它影响到系统评价的数据分析、结果解释和结论。

研究的真实性可从两个方面考虑。第一，是否提出了正确的研究问题，常称之为“外部真实性”，对它的评价是基于研究的目的。外部真实性与研究结果的推广和应用有密切联系，我们将在12章中具体介绍。

第二，是否正确地回答了所提出的问题，即是否通过各种方法减少了偏倚，常称作“内部真实性”，也是本章所描述的真实性。由于大部分Cochrane系统评价关注随机对照试验，我们将重点叙述如何评价随机对照试验的真实性。第13章介绍非随机研究的评价，14章涉及不良反应研究的评价。内部真实性评价通常描述为“方法学质量评价”或者“质量评价”。然而，我们将会避免使用这些术语，原因将在文中介绍。下面我们将定义“偏倚”并且将其与随机误差和质量的相关概念区别。

8.2 什么是偏倚？

8.2.1 偏倚和偏倚风险

偏倚是研究结果或统计推断中的一种系统误差，或与真实值的偏差。偏倚有一定方

向性：不同偏倚可导致低估或高估干预措施的真实效应。偏倚也有量的变化：有些偏倚较小（与观察到的效应相比可忽略不计），有些偏倚较大（明显的效应完全是由偏倚造成的）。即使相同来源的偏倚也会有不同方向：某种设计缺陷产生的偏倚（例如，无分配方案隐藏）在一个研究中可能低估其疗效，而在另一个研究中却可能高估疗效。虽然实证证据告诉我们，随机临床试验在设计、实施和分析中存在的缺陷会导致偏倚，但是要知道这些偏倚对结果到底造成多大程度的影响几乎是不可能的（见8.2.3节）。尽管研究存在方法学缺陷，但事实上结果也可能没有偏倚，所以考虑为“偏倚风险”更恰当。偏倚风险的不同有助于我们解释系统评价中纳入研究结果的差异（如解释结果的异质性）。严格设计和实施的研究能得到更接近真实的结果。如果纳入不够严格的研究倾向于高估干预措施的效果，那么这些不真实结果的Meta分析会错误地认为某干预措施是有效的；如果研究倾向于低估干预措施的效果，那么也可能错误地认为干预措施是无效的。（Detsky 1992）

无论纳入研究的差异是来自研究结果或真实性，评估系统评价中所有纳入研究的偏倚风险都是很重要的。例如，纳入研究的结果趋于一致，但所有研究都可能存在缺陷。在这种情况下，系统评价的结论不可能与纳入研究的设计和实施严格且结果趋于一致的系统评价所得出的结论有同样的说服力。在Cochrane系统评价中，其评价过程被描述为“纳入研究偏倚风险评估”。RevMan软件已可完成纳入研究偏倚风险评估，将在8.5节中介绍。本章余下部分将介绍该软件的原理以及解释如何总结偏倚风险评估并整合到最终的结果分析中（8.6至8.8节）。8.9至8.14节为我们介绍了一些背景，有助于评价者使用这个工具。

偏倚不能与不精确相混淆。偏倚指的是系统误差，也就是说多次重复同一个研究仍会得到错误的平均效应结果。不精确指的是随机误差，即多次重复同一个研究，尽管因为抽样误差而得到不同效应估计值，但其平均效应是真实的。小样本研究更会受到抽样误差的影响，因此精确度较差。各个研究干预效果估计的可信区间和Meta分析中每个研究的权重反应不精确度。更精确的结果会被赋予更大的权重。

8.2.2 偏倚风险和质量

偏倚应与质量区分开来。“方法学质量评价”一词已经被广泛用于系统评价方法中对纳入研究进行严格评估。这个术语揭示研究者以可能的最高标准实施研究的程度，本

手册描述了方法学质量评价和偏倚风险评估的区别，并且对后者给予详细介绍。区别原因如下：

1、Cochrane系统评价主要考虑纳入研究结果的真实程度。偏倚风险评估与该问题直接相关。

2、尽可能以最高标准实施的一项研究仍可能存在重要的偏倚风险。例如，在很多情况下，对受试者或研究人员实施盲法是不切实际或不可能的。将这些研究全部描述为“低质量”是不恰当的，但不能说研究者知道了干预情况而不会产生偏倚。

3、获得伦理批准、样本量计算和按照CONSORT声明报告研究（Moher 2001c)等医学研究的质量指标与偏倚风险无直接关联。

4、强调偏倚风险能克服研究报告质量和研究质量之间的模糊含义（虽然偏倚风险评估依赖研究报告的问题依然存在）。

尽管对术语“质量”有担忧，但术语“证据质量”在Cochrane系统评价“结果总结”表格中仍用来描述纳入研究的效应估计值接近真实效应的可信度，我们将在11章（11.5节）和12章（12.2节）中介绍。如本文所定义的那样，当判断大量证据的质量时，每个研究结果的偏倚风险对效应估计值的影响是许多需要考虑的因素之一。

8.2.3 评估偏倚的实证证据

运用“Meta-流行病学”评估与个别研究特征相关的偏倚(Naylor 1997, Sterne 2002)。Meta 流行病学研究是对多个meta-分析的分析，其中每个meta-分析纳入的研究按照研究水平特征进行分层。早期的例子是一个结果为二分类的临床试验的Meta分析，来自Cochrane Pregnancy and Childbirth数据库（Schulz 1995b）。该研究表明与报告分配隐藏完善的研究相比，分配隐藏不完善或报告不充分的研究干预措施效应会被夸大，而且没有描述双盲的研究也有相似（但小些）的关系。

一个简单的Meta-流行病学分析是计算每个Meta-分析的“比值比（OR）的比”（例如，分配隐藏不完善或不清楚的研究的比值比除以分配隐藏完善的研究的比值比）。这些比值比的比又通过一个新的Meta分析综合。因此，这样的分析过程也被称为“Meta分析的Meta分析”。本章后面的部分中，来自Meta-流行病方法学研究的偏倚实证证据被引用作为评估每种潜在偏倚的部分理论支持。

8.3 研究质量和偏倚风险评估工具

8.3.1 工具类型

现有许多工具用于评估系统评价中纳入研究的质量。其中大部分是量表，对涉及研究质量的各条目进行评分，最后算出总分；或者是条目清单，由具体的问题构成。（Jüni 2001）

1995年，Moher及其同事分析了评估随机试验真实性或研究“质量”的25个量表和9个清单(Moher 1995，Moher 1996)。这些量表和清单包含3~57个条目不等，并且评估一个研究需花10~45分钟。几乎所有的条目都基于临床试验教科书建议或“普遍认可”的标准。许多也包括与内部真实性不直接相关的条目，如是否计算了检验效能（与结果精确度相关）或者是否清楚描述了纳入和排除标准（针对于实用性而非真实性）。量表比清单更可能包括与内部真实性没有直接联系的评估标准。

Cochrane协作网推荐的偏倚风险评估工具既不是量表也不是清单。而是基于“维度评估”，即对研究质量的不同方面进行严格独立评估，将在8.5节中详述。由方法学家、编辑和评价者组成的工作小组在2005~2007年制定。因为在一个既定的研究中要得到偏倚程度（或者真实的偏倚风险）是不大可能的，所以评估工具的验证也受到限制。再实际地评估一个研究的真实性也可能带有主观性：例如，评估未对病人实施盲法是否真的影响到癌症这类严重疾病的复发。

8.3.2 报告与实施

在偏倚风险或质量评估中，难点是研究报告信息不完整所造成的障碍。虽然强调偏倚风险评估应该基于研究实际的设计和实施，但仍依赖于研究报告的准确性。Moher等人评价的工具中很多对此混为一谈（Moher 1995）。此外，量表评分也常基于某些条目是否被报告（比如报告受试者如何被分配），而非研究实施是否恰当。

8.3.3 质量量表和Cochrane系统评价

Cochrane系统评价中，不提倡使用量表评估偏倚风险或研究质量。虽然该方法简单易行，但不被实证证据支持（Emerson 1990，Schulz 1995b）。计算总分必然要对量表中不同条目赋予权重，而且权重分配的正当性难以证明。此外，量表评估的真实性并不可

靠（Jüni 1999），而且对系统评价证据使用者而言也不透明。使用简单并可以详尽报告的方法评估真实性更易于接受（即每个试验在每个标准上是如何判断的）。

由Jadad及其同事为疼痛研究中的随机试验制定的评估量表较为常用（Jadad 1996）。但不推荐使用这个量表，因为和其他量表一样，他强调研究报告而非实施过程，而且没有包括随机试验中一个重要的潜在偏倚——分配隐藏（见8.10.1节）。

8.3.4 收集偏倚风险评价的信息

尽管报告有局限性，但关于研究设计和实施过程的信息通常能通过发表的研究报告，如期刊论文、图书章节、学位论文、会议摘要和网页（包括试验注册库）而获得。研究计划书是一个很有价值的信息源，从中提取信息的讨论详见第7章。资料提取表应有足够的空间提取研究细节，以完成协作网的“偏倚风险”工具的评估（见本章8.5节）。提取此信息时，应详实记录每部分信息的来源（包括在文献中的准确位置）。它对评价小组在预试验中检查数据收集表格和偏倚风险评估有很大帮助，可确保标准一致地贯彻应用，并可以达成共识。如果可能，3到6篇偏倚风险高低不等的文献可能是比较合适的预试验样本。

评估文章的方法时，系统评价员须决定是否对评估偏倚风险的人实施盲法，即不让他们知道纳入研究的作者、单位、杂志和研究结果。一项研究提示，对研究报告进行盲法评估可能比开放性（未采用盲法）评估得到更低、更一致的研究质量分级（Jadad1996），而其他研究显示盲法评估没有益处（Berlin 1997，Kjaergard 2001）。盲法评估非常费时，当评估者对纳入研究很熟悉，盲法也不太可能实现，并且不是所有偏倚风险都能独立于结局数据而进行评价。此外，若知道谁承担的此项研究有时也有助于评估者对该项研究的实施情况做出合理的假设（评价者须报告此假设）。研究者在决定是否采用盲法评估时，必须权衡潜在的利弊

研究者因不同层次的方法学培训和经验可能识别出不同的信息，从而造成偏倚风险评估的差异。虽然专业领域的专家预先已有的观点可能影响他们的判断（Oxman 1993），但是他们对研究的真实性评估可能比其他人更一致（Jadad 1996）。专业领域的专家可以对偏倚风险的大小提出有价值的意见，有经验的方法学家对不明显的潜在偏倚也有很深的见地。每个系统评价小组应该包括此专业领域的专家和方法学家，并确保他们充分理解所有相关的方法学问题。

倚风险评估常受制于未完整报告研究实施中的具体细节。联系原始研究作者是获取缺失信息的途径之一。但是，联系研究的作者可能导致过度阳性的答复。在一项对104名试验人员的调查中，采用直接提问的方式调查对试验人员实施盲法的情况，43%的人反映他们的双盲试验中对数据分析者实施了盲法，19%反映对论文撰写者实施了盲法（Haahr 2006）。这不太可能是真实的，因为发表的相应研究中报告此过程的分别只有3%和0%，而且在其他的试验报告中也很少有描述。

为降低过度阳性答复的风险，当向试验者咨询关于研究设计和实施的信息时，评价员应采用开放性式问题。例如，要获取关于盲法的信息，以下问题形式比较恰当：“是否有确保受试者和主要的试验人员不知道受试者接受了哪种干预措施的方法，若有，请描述。”要获得随机过程的信息，以下问题形式比较适用：“您如何决定下一个受试者接受何种治疗？”可询问更多密切相关的问题以澄清剩余的不确定处。

8.4 临床试验偏倚来源介绍

随机试验结果的真实性取决于避免潜在偏倚的程度。评估每个纳入研究结果的偏倚风险是系统评价的重要部分。一种有用的偏倚分类即分为选择偏倚、实施偏倚、测量偏倚、随访偏倚及报告偏倚。本小节将描述这些偏倚并在8.5节中对协作网“偏倚分析”工具中评估的7个相应维度进行介绍。在8.9节至8.15节，对每个问题提供更详细的阐述。

8.4.1 选择性偏倚

选择偏倚是指比较组的基线特征之间的系统差异。随机化的独特优势是，如果设计成功，它可以在分配患者干预措施时防止选择性偏倚。正确的随机化取决于几个相关的环节是否完善。分配干预措施的规则以机遇（随机）为基础，须详细说明，称为“随机序列生成”。此外，必须采取措施防止随机分配的预知，以确保严格执行随机分配。这个过程通常被称为分配隐藏，但更准确地说应描述为分配序列的隐藏。因此，一个恰当的分配是使用一个简单、不可预知的随机序列，并向选择受试对象入组的人员隐藏下一个受试者的分配情况。

8.4.2 实施偏倚

实施偏倚是指除感兴趣的干预措施外，组间护理、暴露因素等存在的系统差异。进入研究后，对受试者和研究人员采用盲法，可降低由于他们知晓接受的是何种干预措施而影响结局的风险，而不是干预措施本身对结局的影响。有效的盲法也可确保比较组间所受到的护理、辅助治疗、诊断方式相似。然而，盲法并不总是可行的。例如，对是否实施了外科大手术来说盲法是不可行的。

8.4.3 测量偏倚

测量偏倚是指测量组间结局存在的系统差异。对结果评估者采用盲法，可降低由于他们知晓接受的是何种干预措施而影响结果测量的风险，而不是干预措施本身对结局的影响。对结局评估者实施盲法在评估主观性结局指标时十分重要，如术后疼痛程度。

8.4.4 随访偏倚

随访偏倚是指组间研究病例退出导致结局数据不完整造成的系统差异。临床试验中病例退出或数据不完整有两个原因：排除指一些病例资料的数据试验人员可得到但在分析报告中删去的情况，失访指结果数据无法得到的情况。

8.4.5 报告偏倚

报告偏倚是指报告和未报告结果之间的系统差异。发表的研究报告中，干预组之间结果有统计学差异比无统计学差异被报告的可能性更大。这种“研究内发表偏倚”通常被称为结局报告偏倚或选择性报告偏倚，并可能是影响单个研究结果最重要的偏倚之一（Chan2005）。

8.4.6 其他偏倚

此外，还有其他来源的偏倚，是在某些特定的情况下出现的。有些主要在特定的试验设计中出现（如交叉试验中的后遗效应、整群随机试验招募偏倚）；有些可在广泛的试验中找到，但只针对某些特殊情况（如试验组和控制组干预措施混用造成的沾染，例如受试者共用他们的药物）；有些偏倚来源于特定的临床环境。

对于所有偏倚的潜在来源，重要的是要考虑到偏倚可能的大小和方向。例如，如果研究所有的方法学局限性预示偏倚使干预措施趋于无效，而研究结果表明干预是有效的。那么即使存在潜在偏倚，也可以得出干预措施有效的结论。

表8.4.a　常见偏倚分类

偏倚类型	描述	协作网“偏倚风险”工具的相关领域
选择性偏倚	比较组的基线特征之间存在的系统差异	• 序列生成 • 分配隐藏
实施偏倚	除感兴趣的干预措施外，组间护理、暴露因素等存在的系统差异	• 对受试者、研究人员实施盲法 • 其他对真实性的潜在威胁
测量偏倚	测量组间结局存在的系统差异	• 对结局测量者实施盲法 • 其他对真实性的潜在威胁
随访偏倚	组间病例退出造成的系统差异	• 结局数据不完整
报告偏倚	报告与未报告结果间存在的系统差异	• 选择性报告结果（见第 10 章）

8.5　Cochrane 协作网偏倚风险评估工具

8.5.1　概述

本节描述了推荐的评价Cochrane系统评价纳入研究偏倚风险的方法。它由两部分组成，七个重要的偏倚来源将在8.9至8.15节中分别讨论（即随机序列生成、分配隐藏、受试者及研究人员的盲法、结局评估者的盲法、结果数据不完整、选择性报告结果及其他问题）。表8.5.a归纳了该工具。2010年末，经过一项评估项目后对该工具进行了修订，具体修改总结见表8.5.b。

工具的每一维度在“偏倚风险”表中至少包含了一个条目。每一个条目中，工具的第一部分是研究报告中该条目的具体描述，以支持偏倚风险的评价；工具的第二部分是对该条目偏倚风险的判断，通过判断为“低风险”、“高风险”、“风险不清楚”完成。

随机序列生成、分配隐藏和选择性报告结果这三个维度在工具中应该是每个研究一

个条目。对受试者和研究人员实施盲法、结果评估者实施盲法和结局数据不完整，可能有两个或多个条目，因为需分别评估不同结局指标或同一结局指标不同时点的情况。系统评价员应该试着将不同结局指标分组以减少条目数，例如，对结果测量实施盲法评估可分为“客观”或“主观”两类结局指标；对结果数据不完整的评估可分为“6个月随访”或“12个月随访”两类。同样的结局指标分组可用于系统评价的每个纳入研究。偏倚来源的最后一个方面“偏倚的其他来源”对于所有纳入研究可作为一个条目（在RevMan软件中为默认）。尽管如此，我们仍建议采用多个预先设定的条目用于评估其他来源的偏倚。这些评价员自拟的条目可将所有纳入研究视为一个整体评估，也可以对单个研究或分组的结局指标进行评估。

表8.5.a　Cochrane协作网偏倚风险评估工具

领域	判断依据	评估者的判断
选择性偏倚		
随机序列生成	详细描述随机分配序列产生的方法，以便评估不同分配组是否具有可比性	由于产生随机分配方案的方法不正确导致的选择性偏倚（干预措施分配偏倚）
分配隐藏	详细描述隐藏随机分配方案的方法，确定干预措施的分配方法在分组前、期间是否被预知	由于随机分配方案隐藏不完善导致的选择性偏倚（干预措施分配偏倚）
实施偏倚		
对受试者、试验人员实施盲法（需对各项主要结局或结局的种类分别评估）	描述所有对受试者和试验人员施盲的方法。提供所有与盲法是否有效相关的信息	由于研究中干预措施的分配情况被受试者及试验人员知晓导致的实施偏倚
测量偏倚		
对结局评估员施盲（需对各项主要结局或结局的种类分别评估）	描述所有对结局评估员施盲的方法。提供所有与盲法是否有效相关的信息	由于干预措施的分配情况被结局评估员知晓导致的测量偏倚
随访偏倚		
结果数据不完整（需对各项主要结局或结局的种类分别评估）	描述每个主要结局指标结果数据的完整性，包括失访、排除分析的数据。明确是否报告失访和排除分析数据的情况，每个干预组的人数（与分配入组时的人数比较），是否报告失访与排除的原因，以及系统评价员再纳入分析的数据	由于不完整结果数据的数量、种类及处理导致的随访偏倚

报告偏倚		
选择性报告结果	阐明系统评价员如何检查可能发生的选择性结果报告，发现了什么	由于选择性报告结果导致的报告偏倚
其他偏倚		
偏倚的其他来源	工具中没提到的与偏倚有重要关联的情况 如果系统评价的计划书中有预先设定的问题或条目，需一一回答	其他引起偏倚风险的因素

表8.5.b　5.0.1/5.0.2版手册“偏倚风险”工具与5.1.0版（本版）手册修订后“偏倚风险”工具的不同处

盲法的分离	在先前的版本中，对受试者、试验人员、结局评估员施盲的偏倚虽然是对不同结局间分别评估，但属于同一维度。修订后，对受试者、试验人员施盲的偏倚与对结局评估员施盲的偏倚分属于不同维度
判断的种类	现在用“低偏倚”、“高偏倚”、“风险不清楚”表达判断。删除了问题及回答（“是”表示低偏倚，“否”表示高偏倚）
轻微的修改	RevMan 重新命名了条目，删去了基于提问的判断： 序列生成正确吗？改为 随机序列生成 分配隐藏了吗？改为 分配隐藏 盲法了吗？改为 对受试者、试验人员施盲及对结局评估员施盲 结局数据不完整处理了吗？改为 结局数据不完整 没有选择性报告吗？改为 选择性报告 没有其他偏倚吗？改为其他偏倚
插入偏倚分类	修订后的工具阐明了各种偏倚所属维度：选择性偏倚（随机序列生成、分配隐藏）、实施偏倚（对受试者、研究人员施盲）、测量偏倚（对结局评估施盲）、随访偏倚（结局数据不完整）、报告偏倚（选择性报告）及其他偏倚
重新审查包括提前结束试验等“其他偏倚”问题的合格性	对“其他偏倚”维度的指南进行了编辑，以保证附加条目仅在特殊情况下使用且这些条目可能直接引起偏倚。尤其是删除了提前结束试验，因为(i)证据表明在 meta-分析中纳入提前结束试验不会引起实质性偏倚，且(ii)排除提前结束试验可能导致 meta-分析趋于零及精度损失。

8.5.2　判断依据

判断依据简明总结了偏倚风险作出判断的依据，旨在使判断的过程透明化。对于某个研究来说，判断依据的信息通常来自一个发表的研究报告，但也可能通过研究报告、

计划书、发表的评论和与研究人员联系来获得。适当情况下，判断依据应逐字引用报告或信件。除此之外，还可以包括已知事实的总结，或系统评价员的评论。特别是它还应包括任何可以影响判断的其他信息（如同一研究者参与的其他研究所提供的信息）。一个对补充引用报告模糊的有用解释，是表示可能做或可能没有这样做，为判断提供依据。当没有可用信息时，应该明确说明。建议的格式见表8.5.c。

表8.5.c　随机序列生成条目判断依据的例子

随机序列生成	注释：未提供信息
随机序列生成	引用：“受试者随机分配”
随机序列生成	引用：“受试者随机分配” 注释：可能是，因为相同研究者的早期研究报告清楚描述了随机序列的使用（Cartwright 1980）
随机序列生成	引用：“受试者随机分配” 注释：可能不是，因为一个由这些试验人员参与的相似研究也有相同的措辞但采用的是交替分配（Winrow 1983）
随机序列生成	引用（报告中）：“受试者随机分配” 引用（来自信函）：“按治疗日期随机分配” 注释：非随机

8.5.3　判断

评价者的判断应分为偏倚风险低、偏倚风险高及风险不清楚。评价时应考虑重要的偏倚而不是任何偏倚。我们将“重要的偏倚”定义为对试验结果或结论有显著影响的偏倚，并认识到任何判断均有主观性。表8.5.d提供工具中七个方面的偏倚风险判断标准。如果研究中没有充分报告相关细节，判定通常是偏倚风险“不确定”。如果研究报告了相关细节，但其偏倚风险是未知的，也应该做“不确定”的判断；或者条目与该研究无关（当研究中没有测量用于评价结果的条目时，特别是对盲法以及不完整数据的评价）。

表8.5.d　偏倚风险评估工具的偏倚风险评价标准

<table>
<tr><td colspan="2">随机序列生成
生成随机序列的方法不恰当导致的选择性偏倚（干预措施分配偏倚）</td></tr>
<tr><td>低偏倚风险的判断标准</td><td>研究者在序列产生过程中描述了随机方法如：
• 随机数字表
• 计算机产生随机数字
• 抛硬币法
• 洗牌或信封
• 掷骰子
• 抽签法
• 最小化法*
*最小化(Minimization)可以不按随机方法实施，但等同于随机。</td></tr>
<tr><td>高偏倚风险的判断标准</td><td>研究者在序列产生过程中描述了非随机的方法。通常，该描述包括一些系统的、非随机的方法，如：
• 根据生日的奇数或偶数产生分配序列由入院日期（或天数）产生
• 由住院或就诊号码产生
其他非随机方法较以上这几种系统方法较少见，他们通常包括主观判断或其他一些非随机分组方法，如：
• 根据临床医师的判断分配
• 根据病人意愿分配
• 基于实验室结果或一系列检查结果分配
• 根据干预措施的有效性分配</td></tr>
<tr><td>偏倚风险不确定的判断标准</td><td>序列产生的信息不详，难以判断是“低风险”还是“高风险”</td></tr>
<tr><td colspan="2">分配隐藏
由于随访分配方案隐藏不完善导致的选择性偏倚（干预措施分配偏倚）</td></tr>
<tr><td>低偏倚风险的判断标准</td><td>受试者及招募受试者的研究人员不能预知分配情况，因为采用以下原因或者等效的方法来隐藏随机分配方案：
• 中心分配（包括电话、网站和药房控制随机）
• 外形相同且有序的药物容器
• 有序的、不透光的密封信封</td></tr>
<tr><td>高偏倚风险的判断标准</td><td>受试者或招募受试者的研究人员可能会预知分配情况而导致选择性偏倚，如以下的分配方法：
• 运用开放性随机分配表（如随机数字表）
• 信封缺乏恰当的保护（即信封不是密封的，或不是有序的，或是透明的，）
• 交替或轮流分配
• 出生日期
• 病例号
• 其他明确不能隐藏的方法</td></tr>
</table>

偏倚风险不确定的判断标准	无充分信息判断“低风险”或“高风险”。通常是隐藏的方法没描述或者没充分的描述而不能给出明确的判断，例如，描述了使用信封分配，但不确定是否按顺序编号，是否透明，是否密封等。
对病人、试验人员实施盲法 研究中干预措施的分配情况被受试者及试验人员知晓导致的实施偏倚	
低偏倚风险的判断标准	存在以下任一项： • 无盲法或盲法不完善，但系统评价员判断结局不会受到未施盲法的影响 • 对受试者和主要研究人员实施盲法，且盲法不会被破坏
高偏倚风险的判断标准	存在以下任一项： • 未采用盲法或盲法不完善，结果判断或测量会受到影响 • 对受试者和主要研究人员实施盲法，但该盲法可能被破坏
偏倚风险不确定的判断标准	存在以下任一项： • 无充分信息判断为“是”或“否” • 研究中没有报告该结局指标
对结局评估者实施盲法 因结局评价者知道干预措施分组情况导致的实施偏倚	
低偏倚风险的判断标准	存在以下任一项： • 未实施盲法，但系统评价员判断结局测量不会受到未施盲法的影响 • 对结局测量者实施盲法，且盲法不会被破坏
高偏倚风险的判断标准	存在以下任一项： • 未采用盲法或盲法不完善，结果判断或测量会受到影响 • 对结局测量者实施盲法，但该盲法可能被破坏
偏倚风险不确定的判断标准	存在以下任一项： • 无充分信息判断为“是”或“否” • 研究中没有报告该结局指标
结果数据不完整 由于不完整结果数据的数量、种类及处理导致的随访偏倚	
低偏倚风险的判断标准	存在以下任一项： • 无缺失数据 • 缺失数据不影响结果分析（如生存分析中的缺失值） • 组间缺失的人数和原因相似 • 对二分类数据，缺失数据的比例与观察到的事件相比，不足以严重影响干预措施效应值 • 对于连续性变量数据，缺失数据的效应值（均数差值或标准化均数差值）不足以严重影响观察到的效应值 • 采用恰当的方法处理了缺失数据

高偏倚风险的判断标准	存在以下任一项： • 组间缺失的人数和原因不平衡 • 对于二分类数据，缺失数据的比例与观察到的事件相比，不足以严重影响干预措施效应值 • 对于连续性变量数据，缺失数据的效应值（均数差值或标准化均数差值）不足以严重影响观察到的效应值 • 采用“as-treated”分析，但改变随机分配的干预措施的人数较多 • 不恰当的方法处理缺失数据
偏倚风险不确定的判断标准	存在以下任一项： • 信息不全，难以判断数据是否完整（如缺失人数或原因未报告） • 研究未提及完整性的问题
选择性报告 由于选择性报告结果导致的报告偏倚	
低偏倚风险的判断标准	存在以下任一项： • 有研究计划书，且系统评价均按预定的方式报告了所有预定的结局指标（主要和次要结局） • 无研究计划书，但发表的研究报告中所有期望的结局（包括了预定的结局）均已报告，包括那些预先设定的（有说服力的文本较少见）
高偏倚风险的判断标准	存在以下任一项： • 未报告所有预先指定的主要结局指标 • 报告的一个或多个主要结局指标采用预先未指定的测量、数据分析方法或数据子集（如子量表） • 报告的一个或多个主要结局指标未预先设定（除非证实报告它们是必须的，如没有预料到的不良反应） • 系统评价关心的一个或多个结局指标报告不完善，以致不能纳入行 meta 分析 • 未报告重要的结局指标
偏倚风险不确定的判断标准	信息不全，难以判断是否存在选择性报告结果的风险。可能大部分的研究会判断为这种类别
其他偏倚 该表格其他地方未包含的偏倚	
低偏倚风险的判断标准	研究无其他偏倚来源
高偏倚风险的判断标准	至少有一个重要的偏倚风险。如：该研究 • 有与特殊研究设计有关的潜在偏倚； • 声明有造假行为 • 一些其他问题
偏倚风险不确定的判断标准	可能存在偏倚风险，也可能是其他 • 没有充分的信息判断是否存在重要偏倚风险 • 无充分理由或证据证明这个问题可以导致偏倚

8.6 风险评估描述

“偏倚风险”表在RevMan软件中作为Cochrane系统评价“纳入研究特征”表的一部分。对于每个条目，判断（低偏倚风险；高偏倚风险；或者偏倚风险不确定）之后附着一个文本框，用于描述判断依据相关的研究设计、实施、或观察的信息。图8.6.a提供了图样。如果这个文本框是空的，且判断是“风险不确定”，那么在CDSR发表时，这个条目将从“偏倚风险”表中省去。

系统评价文章中偏倚风险评价的描述在第4章讨论（4.5节）（属于结果的子标题“纳入研究偏倚风险”和讨论的子标题“证据质量”）。系统评价中纳入研究的偏倚风险可用Revman软件产生两个图。一个是“偏倚风险图”，描述工具中每个条目每种判断（低风险、高风险、风险不确定）的研究比例（见图8.6.b）。另一个“偏倚风险总结图”表示各个研究的每个条目的判断结果交叉表（见图8.6.c）。

第一个图（“偏倚风险图”）的另一个可选择的、可能更佳的版本可显示低风险、风险不明确及高风险的信息而不是研究的比例，将注意力放在特别重要的meta-分析中的研究。Meta-分析中，可根据信息的比例赋予每个研究的权重。但是，RevMan尚不能制作该图。

图8.6.a 单个研究的“偏倚风险”表（虚构）

条目	判断	判断依据
随机分配序列的生成（选择偏倚）	低偏倚	引用：“病人随机分配” 评论：很可能是，因为相同研究者的早期研究报告描述了随机序列产生的正确方法（Cartwright 1980）
分配隐藏（选择偏倚）	高偏倚	引用：“采用随机数字表” 评论：很可能未隐藏随机分配方案
对受试者、研究人员施盲（实施偏倚）	低偏倚	引用：“双盲双模拟”；“高、低剂量的片剂或胶囊从外观上不能区分。对每种药物，都用了不能区分的相似安慰剂（通过检查分发之前的药物来评价盲法是否成功）。” 评论：很可能是
对结局评估员施盲（测量偏倚）	低偏倚	引用：“双盲”。 评论：很可能是

结果数据不完整（随访偏倚）（短期[2-6 周]）	高偏倚	4 周：干预组 110 例中 17 例缺失（9 例因为“缺乏效果”）；对照组 113 例有 7 例缺失（2 例因为“缺乏效果”）
结果数据不完整（随访偏倚）（长期结局＞6 周）	高偏倚	12 周：干预组 110 例中 31 例缺失；对照组 113 例中 18 例缺失。每组缺失原因不同
选择性报告（报告偏倚）	高偏倚	方法中提到三个认知量表，但只报告了其中一个（结果有统计学意义）

图8.6.b “偏倚风险图”图形例子

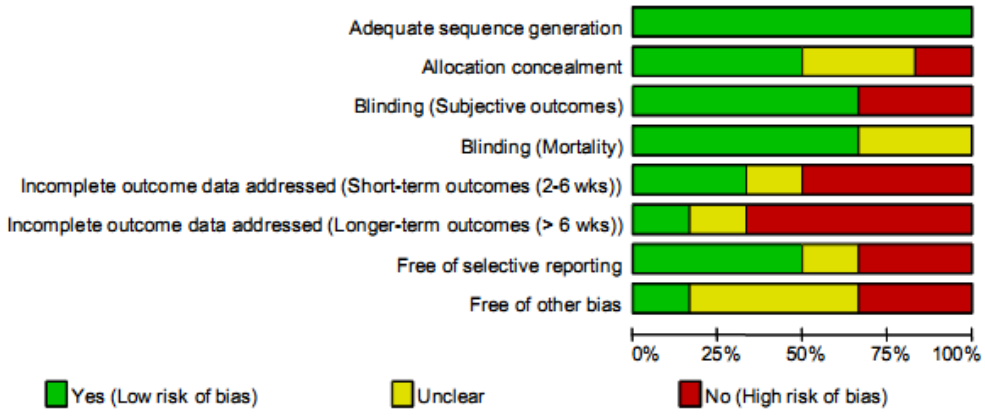
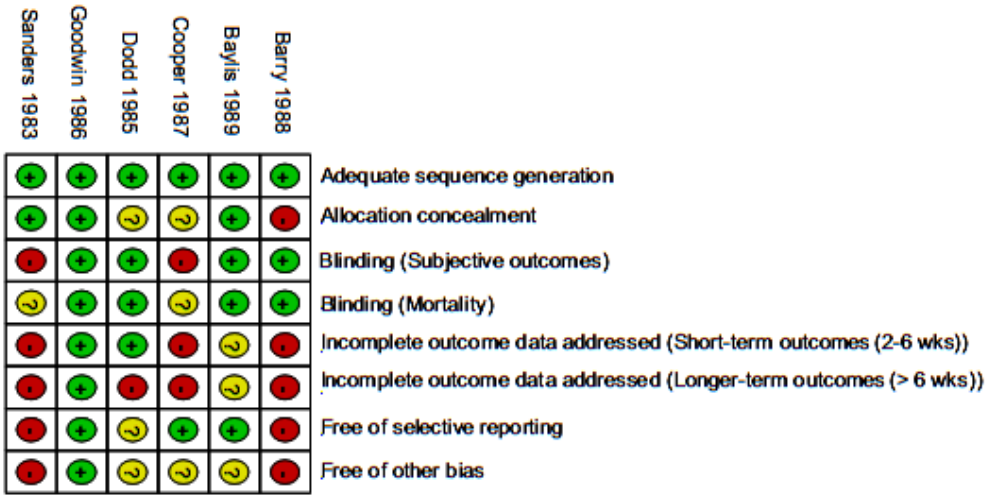


图8.6.c 偏倚风险总结图”举例



8.7 偏倚风险评估小结

协作网推荐的纳入研究偏倚风险评估工具从每个偏倚来源方面均进行了评估和描述，如分配隐藏和盲法。为得到某个结局指标总的偏倚风险就需进行总结。一般不提倡使用量表评估（通过多个条目的得分相加产生一个总分），原因见8.3.1节。尽管如此，任何总偏倚风险的评估都应该考虑不同偏倚来源的重要性。系统评价员必须判断当前系统评价中哪种偏倚来源最重要。例如，对于高度主观性结局如疼痛，评价员可能认为对受试者采用盲法至关重要。做出如此判断的依据需明确，如：

- 偏倚的实证证据：8.5节到8.15节总结了各来源偏倚的实证证据，如分配隐藏、盲法与疗效估计值大小的关系。尽管如此，证据仍然不完善。
- 偏倚的可能方向：现有的实证证据认为不符合重要的标准如分配隐藏完善与过高估计效应相关。如果某个方面的偏倚使得疗效被低估（偏向无效），而系统评价认为该干预措施是有效的，那么无须太关注该方面的偏倚。
- 偏倚的可能强度：不同来源的偏倚强度可能是不一样的。例如对受试者采用的盲法不完善所产生的偏倚可能对主观性结局指标影响更大。一些经验性证据提示了可能的偏倚强度（见上文），但仍缺乏充分信息清楚说明在某些特殊情况下偏倚的大小。尽管如此，考虑偏倚强度与其影响大小的关系仍是可行的。例如，随机分配方案隐藏不完善加上一个较小的效应估计值，那么这个结果的可信度就会大大降低；反之，如果不完整数据的处理存在较小的缺陷，那么对一个较大效应估计值的可信度就不会产生实质性的影响。

总结偏倚风险可从4个层面考虑：

- 针对单个研究的所有结局指标总结偏倚风险：某些偏倚会影响一个研究的所有结局指标的偏倚风险：如随机序列产生和分配隐藏。其他偏倚如盲法和结果数据不完整，对单个研究不同结局可能产生不同偏倚风险。因此，系统评价员不应该假设一个研究的所有结局的偏倚风险都是相同的。而且，针对同一研究的所有结果总结偏倚风险，意义不大。
- 针对一个研究内的某一结果总结偏倚风险（涉及所有偏倚来源）：这是推荐的总结研究偏倚风险的方式，因为不同结局其偏倚风险可能不同。总结某个结局偏倚风险应该包括所有的与该结局相关的条目：即研究水平的条目（如分配序列隐藏）和结局水平特有的条目（如盲法）。

- 针对所有研究的同一结果总结偏倚风险（如用于meta分析）：这些是系统评价员主要采用的偏倚风险总结方法并纳入“结果总结”表中的“证据质量”判断，见第11章11.5节。如上解释，Meta-分析时纳入高偏倚风险研究结果比排除这些研究时证据质量更低。
- 系统评价的整体偏倚风险总结（针对所有研究的所有结果）：应避免总结某一个系统评价整体的偏倚风险，有两个原因：第一，需要判断哪些结局指标对决策很重要。但系统评价纳入的研究常常缺乏一些重要结局指标的数据，如不良反应。在同一系统评价中，这些结局的偏倚风险也很少相同。第二，因为价值观或其他的因素存在差异，不同环境中对影响决策的重要结局的判断也是不同的，如基础危险度。因此，针对所有研究的所有结局总结总体偏倚风险的方式只能在特定情况下采用，如临床实践指南，而不应在为不同地方提供决策依据的系统评价中采用。系统评价员应该针对单个研究和所有研究中重要的结局指标作出明确的判定。这需要鉴定出最重要的偏倚来源（“关键来源”），以总结偏倚风险评估。表8.7.a为在一个研究内或所有研究中总结重要结局指标的偏倚风险提供了一个可行的途径。

表8.7.a　在单个研究内或者所有研究中概括每个重要结局偏倚风险的方法

偏倚风险	解释	单个研究内	所有研究中
低偏倚风险	存在的偏倚不可能严重影响研究结果	所有条目的评估结果均为低偏倚风险	大部分信息来源于低偏倚风险的研究
不确定	存在的偏倚引起对研究结果的怀疑	一个或或多个条目的评估结果为偏倚风险不确定	大部分信息来源于低偏倚风险或偏倚风险不确定的研究
高偏倚风险	存在的偏倚严重减弱研究结果的可信度	一个或多个条目的评估结果为高偏倚风险	来源于高偏倚风险研究的信息的比例足够影响结果的解释

8.8 将评估结果纳入分析

8.8.1 引言

统计学需要在偏倚与精确度之间进行权衡。纳入所有研究的Meta分析可能得到一个高精度（可信区间窄）的结果，但也可能因为一些研究的实施中的缺陷而产生严重偏倚。另一方面，只纳入各方面偏倚风险均低的研究也可能得出无偏倚但不精确的结果（如果只有少数高质量的研究）。在进行和报告Meta分析时，系统评价者必须说明纳入研究结果的偏倚风险。忽略纳入研究在偏倚风险评估中发现的缺陷，而基于所有研究进行分析和解释是不恰当的。高偏倚风险的研究所占比例越高，在结果的分析解释中越需谨慎，证据质量级别也越低。

8.8.2 偏倚风险影响

8.8.2.1 基于偏倚风险的图表结果

下面将讨论单个偏倚来源及总结研究的偏倚风险（见本章8.7）

根据偏倚风险分层的干预效果估计值的图表（如森林图）可能是开始衡量潜在偏倚对Meta分析结果影响的一个有效途径。Revman5软件可做出根据每个“偏倚风险”条目判断的森林图。该图形能直观地呈现低偏倚风险，偏倚风险不清楚或者高偏倚风险的研究的相对贡献程度以及这些研究之间干预效应估计值的差异程度。通常，比较明智做法是将这些图表限制在几个关键的偏倚条目上（见本章8.7）。

8.8.2.2 偏倚风险不确定的研究

下面几种情况中研究可判断为偏倚风险不确定：无足够的信息判断为“高”或“低”偏倚风险；虽然有关于研究实施的足够的信息但偏倚风险未知；或是某个条目与研究不相关（如研究所关注的结局不在条目所适用的结局里面）。当以第一种情况为主时，认为这些研究所得结果的平均偏倚比高偏倚风险研究所得的结果低是合理的，因为在偏倚风险不确定的研究中，有的研究实际上可能避免了偏倚。由分别考察偏倚风险“高”和“不确定”研究的实证研究得出的有限证据认为：例如，Schulz等研究发现对于分配隐藏不恰当（高偏倚风险）的研究，干预效应的比值比（OR）夸大41%，而分配隐藏不明确（不确定偏倚风险）的研究则夸大30%（Schulz 1995b）。但是，在与“低”偏倚风险

的研究相比时，大部分实证研究都先将偏倚风险“高”和“不确定”的研究进行合并。在数据分析时最好不要将低偏倚风险和不确定偏倚风险的研究合并，除非有明确的理由认为这些研究很可能以能够避免偏倚的方法实施。在本节其余部分，我们将把偏倚风险低的研究单独作为一类。

8.8.2.3 Meta 回归和亚组间比较

根据偏倚风险进行干预效果间的正式比较可以通过Meta回归实现（详见第九章9.6.4节）。对于结果为二分类数据的研究，Meta回归分析结果通常用比值比（或危险比）来表示高、不确定偏倚风险的研究与低偏倚风险研究的效应差异。

比值比之比（Ratio of odds ratios, ROR）=高偏倚风险或不确定偏倚风险研究干预措施效应比值比/低偏倚风险的干预措施效应比值比

此外，也可以进行比较高偏倚风险、不确定偏倚风险与低偏倚风险研究的分别比较。对于连续性变量（如血压）资料研究，干预措施效应以组间均差表示，则Meta回归分析结果就对应的是不同偏倚风险研究的均差之差。

如果干预措施的效应估计值在不同偏倚风险的研究中一样，那么比值比（危险比）之比就等于1，而均差之间的差异则为0。正如8.2.3节的解释，来自Meta流行病学研究的多个meta分析结果的经验证据表明：通常情况下，较之于低风险偏倚的研究，干预措施的效应估计值在高、不确定偏倚风险的研究中会被夸大。当一个Meta分析纳入许多研究时，Meta回归分析可以包括多个偏倚来源（如分配隐藏和盲法）。

Meta回归分析的结果包括比值比之比的可信区间，还有零假设（即高、不确定偏倚风险的研究和低偏倚风险的研究的结果没有差异）的P值。因为Meta分析纳入研究数量通常较少，比值比之比的估计可能会不准确。因此很重要的一点是，在P值无统计学差异时，不要做出高、不确定偏倚风险与低偏倚风险研究的结果之间无差异因而偏倚对结果无影响的结论。可信区间可以显示高、不确定偏倚风险与低偏倚风险研究间的差异与无偏倚和偏倚的实质影响是一致的。亚组间差异的统计学检验也可以对单个条目进行比较（如比较分配隐藏适当和不适当的研究）。固定效应Meta分析模型，可用Revman5软件实现。如果没有相应的可信区间，得到的P值的价值也有限，并且无论是亚组间或亚组内，只要在异质性存在的情况下，所得的P值都将非常小。

8.8.3 在分析中纳入对偏倚风险的评估

一般来说，较之低偏倚风险的研究，高偏倚风险或不确切偏倚的研究应在Meta分析中给予较小的权重（Spiegelhalter 2003）。但是，目前Cochrane系统评价推荐使用的综合高偏倚风险和低偏倚风险研究结果的统计方法还不完善（见8.8.4.2），因此，Cochrane系统评价中整合偏倚风险的主要方法是将Meta分析限制在低偏倚风险（或较低偏倚）的研究或根据偏倚风险对研究进行分层。

8.8.3.1 可能的分析策略

当Meta分析纳入研究偏倚风险存在差异时，主要有三种策略可用于选择报道哪一结果作为系统评价的主要结果（例如，决定某个结果是否包括在摘要中）。计划使用的策略应在系统评价计划书中描述。

1. 只分析低（或低和不明确）偏倚风险的研究（来自5.02）

第一种方法即根据主要偏倚来源（见8.7节）确定纳入主要分析研究的阈值，即只有符合特定标准的研究才被纳入分析。可根据系统评价纳入标准或合理的参数（可能根据Meta流行病学研究关于偏倚的经验证据得出）来制定阈值。在极个别研究中，Meta分析内部对于高偏倚风险和低偏倚风险研究的比较可能得出干预效应估计值有差异的证据，就使得把分析限制在低偏倚风险的研究有了正当理由（见8.8.2.3节）。如果主要分析包括了偏倚风险不明确的研究，作者则需给出正当理由。理想情况下，计划书中应明确阈值或其确定的方法。作者应记住任何阈值都是主观的，理论上，研究可能处在“没有偏倚”到“确认有偏倚”变化的任何位置。阈值越高，研究的偏倚风险越相似，但数量越少。如果系统评价使用这种方法确定纳入分析的研究，那么最好做敏感性分析以确定纳入高偏倚风险的研究对结论的影响。

2. 进行多元（分层）分析

根据总的偏倚风险分层可能产生干预措施效果的至少3种估计值：来自高偏倚风险研究、低偏倚风险研究以及所有研究。两个或多个这样的估计值可能同等重要，例如，一个是纳入所有的研究，一个是仅纳入低偏倚风险的研究。这样的做法避免了在不同结果间做出艰难的抉择，但是对于读者来说可能十分困惑，尤其是那些通常需要基于效应的单个估计值作出决策的人来说。同时，“结果总结”表常常只需给出每种结局的单个结果。另外，分层的森林图也清楚的展示所有信息。

选择策略1或2应根据系统评价的具体情况，并在排除高偏倚风险或不明确偏倚风险的研究的情况下，权衡潜在偏倚及精确度的损失。正如在8.8.2.3节解释的，因为meta回归分析的检验效能低，不应将高偏倚风险与低偏倚风险的研究结果间差异没有统计学显著性解释为研究不存在偏倚。

3. 分析所有研究并对所有研究的偏倚风险作叙述性讨论

将偏倚风险评估整合到最终结果中最简单的方法是给出基于所有纳入研究的干预措施效应的估计值，并描述所有研究在每个方面的偏倚风险，或者总的偏倚风险。仅当所有纳入的研究均为高偏倚风险、不确定或低偏倚风险时，这是可行的方法。但当纳入的研究间偏倚风险不同时，我们不推荐使用这种方法，原因有二：首先，偏倚风险评估一般在结果部分详细描述，并在讨论中进行谨慎的解释，但在结论、摘要和“研究结果概要”中经常不会提及，所以最终的解释会忽略偏倚风险，而后续的决策就可能至少部分是基于有缺陷的证据；其次，它不能降低高偏倚风险研究的权重，并将得出一个非常精确同时存在潜在偏倚的总干预效应。

在对所有研究进行分析时，对偏倚风险的整体评估必须纳入对于每个重要结局证据质量的明确指标，例如使用GRADE系统（Guyatt 2008）。这样可以明确在解释系统评价结果时，已适当地考虑了对偏倚风险的判断及不准确性、异质性、发表偏倚等影响证据质量的因素。

8.8.4 其他解决偏倚风险的方法

8.8.4.1 直接加权法

已有根据真实性（效度）或偏倚风险对Meta分析纳入研究进行加权的方法报告（Detsky 1992）。在合并多个研究时，常用的统计方法中是根据各研究贡献的信息量（具体就是效应估计值方差的倒数）赋予权重。这会使结果越精确（可信区间窄）的研究的权重越大。此外，根据真实性（效度）赋予权重也是可行的，使真实性好的研究对合并的结果有更多的影响。也可以联合使用倒方差和真实性评估进行加权，这种方法的主要缺点是需要每个研究真实性的具体数值，并且目前还没有可以决定对不同偏倚来源赋予多大权重的经验证据。而且，如果一些研究存在偏倚，结果的平均权重也会产生偏倚。应避免采用由真实性或偏倚风险评估对效应估计直接加权（Greenland 2001）。

8.8.4.2 Bayesian 方法

Bayesian方法允许加入关于偏倚性质的外部信息或意见（见第16章，16.8节）（Turner 2008）。干预效果评估中具体偏倚的先验分布可基于偏倚的经验证据、专家意见或其他合理的观点。用于偏倚调整Meta分析的Bayesian方法目前正在研究探索中，其尚未发展完善而不能广泛使用。

8.9 随机序列生成

8.9.1 偏倚相关理论

根据Cochrane协作网偏倚风险评估工具中的随机序列生成，确定研究是否采用随机分配。包括两方面的内容，一是处理分配的过程，第二个是隐藏随机序列（分配隐藏）。下面我们将说明两者的区别。无偏倚的干预性研究首先是确保同类受试者接受各自的干预措施,一些相关的过程需要考虑：首先，如果实施完善，对于必须使用随机分配序列，应能平衡各干预组间的预后因素。随机化在这里起着根本性作用。其他一些有争议的分配规则，如交替（两个干预交替）或循环（两个以上干预间的循环）分配也可以同样实现这个过程（Hill 1990）。然而，理论上无偏倚的规则还不足以防止实践中的偏倚。如果分配情况可以被预知，不管是通过预测还是其他办法知道了分配序列，都会因在将要进行的干预分配中对受试者的选择性招募和不招募而出现选择性偏倚。

分配方案能被预知有以下几个原因：(i)知道分配的规则，如交替、出生日期或入院日期；(ii)无论是随机或非随机，知晓分配序列（例如，如果随机分配序列被贴在墙上）；(iii)根据之前的分配，能成功预测今后分配（有时候可能是随机方法为了确保不同干预间准确的分配比例）。干预性研究中，干预分配在理论与实践方面复杂的相互关系使得选择性偏倚的评估充满挑战。随机分配序列的隐藏也许是在实践方面最为重要的部分，即防止下一个分配序列被预知。经验证明，这在Cochrane系统评价中已经经过评估了。在偏倚风险评价工具中，我们将分配序列隐藏作为一个单独的偏倚来源（见8.10节）。

随机化要求序列不可预测。不可预测的随机序列和分配隐藏相结合应能防止选择性偏倚。然而，如果进行了随机化但随机分配方案未隐藏，选择偏倚也可能发生，同时，即使分配方案隐藏，如果潜在的序列是非随机的，也可能会出现选择性偏倚（至少在理论上是）。我们知道即使分配隐藏方法完善，一个随机序列也并不总是完全不可预测。

有时这可能就是问题所在，例如，如果采用区组随机，一旦进入试验，所有的分配情况是已知的，不过，我们不在序列生成或分配隐藏中考虑这种特殊的情况，而在8.15.1.3节中单独讨论。

方法学研究已评估了序列生成的重要性。至少4个研究避免了因疾病或干预造成的混杂，这对评估是至关重要的（Schulz 1995b，Moher 1998，Kjaergard 2001，Siersma 2007）。随机分配序列生成方法不正确与研究间干预措施效应的偏倚有一定关系（Als Nielsen 2004）。在一项将分析限制在已报告分配隐藏充分的79个试验的研究中，与序列生成恰当的研究相比，序列生成不恰当的的研究平均来说会夸大干预措施的效应估计值（OR=0.75，95％可信区间0.55-1.02，P=0.07）。这些结果表明，如果分配序列是非随机的，即使分配隐藏方法完善，但分配方案仍可被预知（Schulz 1995b）。

8.9.2 有关随机序列生成是否恰当的偏倚风险评估

在随机对照试验的设计和实施阶段，常常没有对随机序列的生成做适当的描述，并在发表的报告中也时常被忽略，这些均是偏倚风险评估中面临的主要问题。当使用Cochrane协作网的工具时，下面的建议可帮助评估者确定随机序列生成是否正确（见8.5节）。

8.9.2.1 序列生成的正确方法

用于序列生成的随机成分应足够。

对于生成的分配序列没有限制的随机称为简单随机或非限制随机。原则上以下方法也可实现对干预措施的分配，如抛硬币法、扔骰子或扑克牌法（Schulz 2002c，Schulz 2006）。更常用的是查随机数字表或者计算机随机。大样本量的试验中（每个随机组至少100例，Schulz 2002c，Schulz 2002d，Schulz 2006），简单随机法可以得到例数相对接近的比较组。但在小样本量的试验中，简单随机法有时会导致组间存在差异，在一些偶然的情况下，组间的例数或者预后因素（即“混合病例“变异）可能相差很大（Altman 1999）。

例子（低偏倚风险）：通过随机数字表进行简单随机化分配，得到相同分配比的两个比较组。

有时，限制性随机用来产生确保干预组特定分配比例的序列（如1:1）。区组随机（随机排列区组）是一种常见的限制性随机形式（Schulz 2002c，Schulz 2006）。区组能使分配到各对组受试者的受试者在区组间达到平衡，例如，每10个连续进入的受试者中5个

进入一个组，另外5个进入另一组。为减少干预分配被预测的可能性，区组的大小可以随机进行变化。

例子（低偏倚风险）：

我们用区组随机产生了两个干预组的分配名单，用计算机随机数字生成器选择了区组大小8且相等的分配比例的随机排列区组。

分层随机也比较常用，它是在每一层内单独进行限制性随机分配。这种分配方法是在根据重要的预后因素确定的多个亚组内产生各自的随机方案，如疾病的严重程度和研究中心。如果每一层内采用简单（非限制性）随机，那么分层可能无效，但随机仍然有效。无论试验是否说明进行了分层，均使用同样方式判断偏倚风险。

另一种整合了分层随机和限制性随机一般概念的办法——最小化法，它可用于产生某些特征相似的小样本干预组。使用最小化法不应该被机械地认为增加了研究的偏倚风险。然而，一些方法学者对最小化法的可接受性仍持谨慎态度，尤其当它是在没有任何随机过程被使用时，而有些人则认为这种方法非常有吸引力（Brown 2005）。

随机化的其他方法有硬币法或抽签法、替代随机、混合随机和最大随机化法（Schulz 2002c，Schulz 2002d，Berger 2003）。碰到这些或其他方法时，可能有必要咨询统计学家。

8.9.2.2 序列生成的不正确方法

交替分配、按出生日期、病例记录号和记录日期分配等系统方法，有时被称为“半随机（假随机）”。交替（或循环，对于两个以上干预组时）原则上可形成相似的干预组，但其他序列生成的系统方法则可能不能，例如病人入院的时间，就不仅是机遇的问题。

所有系统方法主要的缺点是分配序列隐藏通常不太可能，这会使研究中招募的受试者的分配情况可被预知，从而分配存在偏倚（见8.10节）。

例子（高偏倚风险）：基于某月的第几周将患者分配到各干预组。

例子（高偏倚风险）：将出生天数为偶数的患者分配到干预组A，出生天数为奇数的患者被分配到干预组B。

8.9.2.3 偏倚风险不确定的序列生成方法

“我们进行了随机分配”或“采用随机化设计”这类简单的表述往往不能充分证明分配序列是否真的随机。即使不合理但很多作者仍使用“随机化”这个术语：许多系统化的分配方法也被研究作者描述为随机。如果存在疑问，那么序列生成的正确性应被视

为不清楚。

有时试验者提供了一些信息，但也不能完全确定所使用的方法以及实施过程中是随机的。例如，作者指出采用了区组随机，但选择区组的过程，如随机数字表或计算机随机数字生成器，却未详细说明。这时序列产生的正确性应为不清楚。

8.10 分配序列隐藏

8.10.1 偏倚相关理论

随机生成分配序列是必要的，但并不是防止干预措施分配中的偏倚的保障。如果这些序列在招募和分配受试者时不能充分隐藏，所产生的不可预测和无偏倚的序列就可能是无效的。

预知分配序列（如公开张贴在布告栏的随机数字表）可能导致基于预后因素而选择性招募受试者。将被分配到“不适合”干预组的患者可能会拒绝。通过延迟分配直到出现合适的受试者，其他受试者也可能被有意地进入“合适”的干预组。即使试图隐藏分配序列，但破译的情况时有发生。例如未密封的分配信封可能被打开；半透明的信封内容可以在亮光下被看到（Schulz 1995a，Schulz 1995b，Jüni 2001）。个人报道显示因为隐藏方法不完善，许多分配方案被研究者破译（Schulz 1995a）。

防止分配序列被预知可避免上述选择性偏倚。应在不知道将要接受的干预措施情况下，决定受试者是否入组和是否给予他们知情同意。完善的分配序列隐藏方法可防止试验人员预知接下来的分配情况。

若干方法学研究着眼于在避免了疾病或干预等混杂因素后，分配序列隐藏是否与有对照的临床试验的效果估计值的大小存在关联。综合分析7个方法学研究发现，分配隐藏不完善或不清楚的研究，其“有利”效果估计比分配隐藏完善的研究平均高出18％（95%可信区间5%-29%）（Pildal 2007）。最近对于3组这类合并数据的（来自146个Meta分析的1346个试验）的详细分析着重于这些研究的异质性。在一些主观性结局指标的试验中，当其分配隐藏不恰当时，干预效果的估计会被夸大，而客观指标的偏倚还无从证实（Wood 2008）。

8.10.2 分配隐藏是否完善的偏倚风险评价

在使用Cochrane协作网的评估时，下列因素有助于系统评价员评估分配隐藏方法是否足以防止偏倚（见8.5）。完善的分配序列隐藏方法保证分配序列是在未预知干预分配的前提下严格执行的。分配隐藏方法是指按序列实行分配的技术，而不是产生序列的方法（Schulz 1995b）。然而大多数分配序列产生的方法不正确，如根据入院天数或病例记录号的分配，不能充分隐藏，并因此在两方面均失败。不正确的序列也可能被充分隐藏（负责招募和分配干预措施的人员不知道所实施序列的是不恰当的），这只是理论上可能，实际不大可能。然而，正确（即随机）的序列生成方法也可能隐藏得不完善，例如将分配序列张贴在墙上。

一些系统评价员混淆了分配隐藏和干预措施分配的盲法。分配隐藏的目的是在干预分配中通过保护分配之前和分配入组时的随机序列来防止选择性偏倚，并且不管任何研究都能成功实施（Schulz 1995b，Jüni 2001）。相比之下，盲法旨在通过保护分配后的序列来防止实施偏倚和测量偏倚（Jüni 2001，Schulz 2002a），并且不一定都能实施，如比较药物与手术的研究。因此分配隐藏的影响到干预措施分配为止，而盲法则影响干预分配之后，其解决不同来源的偏倚并且可行性不同。

分配隐藏的重要性可能取决于研究潜在的受试者有不同预后的程度，研究者和受试者是否对干预措施利弊抱有很强的信念，以及干预措施疗效的不确定性是否被研究涉及的所有人接受（Schulz 1995a）。在不同的分配隐藏方法中，由第三方实施的中心随机是最可取的。而采用信封法比其他方法更易于操作（Schulz 1995b）。如果研究者使用信封，他们应该制定并监测分配的过程以保持隐藏。除了使用顺序编号、不透明、密封的信封，还应确保按顺序打开信封，并且不可逆的用于受试者。

8.10.2.1 分配序列隐藏的正确方法

表8.10.a规定了判断分配隐藏正确的最低标准（左）和确保分配隐藏确为正确的扩展标准（右）。

例子（低偏倚风险）[发表的判断隐藏过程为正确的描述，由Schulz和Grimes编制（Schulz 2002b）]：“……药物分配结合编码数字。每个区组十个数字，由中心分配给每个中心负责随机分配的人员。这些人（药剂师或不照顾试验患者的护士和不在研究地点的研究者）负责分配、制备和记录试验输液。这项试验的输液在单独场所制备，然后每

隔24小时由护士带到床边。护士以适当的速度给病人给药。这样的随机分配序列对所有护理人员、病房医生及其他研究人员隐藏。”（Bellomo 2000）。“……以有序编号、密封、不透明的信封隐藏，并由两个中心的药剂师保存。”（Smilde 2001）。“治疗措施通过电话核实纳入标准后由研究中心分配……”（de Gaetano 2001）。“Glenfield医院药学部门做了随机，分送研究试剂，并保存试验代码，这些都在研究完成后公布。”（Brightling 2000）。

表8.10.a　判断分配隐藏充分（低偏倚风险）的最低和扩展标准

判断分配序列隐藏充分的最低标准	判断分配序列隐藏充分的扩展标准
中心随机	中心随机处远离患者招募中心。受试者的详情通过电话、传真或电子邮件提供给随机中心，且分配序列隐藏到受试者确定入组后。
有序编号的药物容器	药物容器由独立的药房准备，按顺序编号，并依次打开。容器的外观相同、防篡改、重量相等。
顺序编号、不透明、密封的信封	受试者的详情写在信封上后，信封按顺序编号，并依次打开。信封中的感光纸或复写纸会传递受试者的详情。信封内用厚纸板或铝箔不会透过强光。用防干扰的扎带将信封密封。

8.11　对受试者和工作人员的盲法

8.11.1　偏倚相关理论

临床试验中可对多种人员施盲：见框8.11.a。工具中的前两项明确说明了针对受试者和工作人员（医药卫生提供者）的盲法。缺乏对受试者或工作人员的盲法可影响受试者的真实结局而导致偏倚。这可能是因为对对照组缺乏期望，或干预组间的不同行为（如脱落不同、与替代措施交叉不同，或合并干预措施的不同）。

目前尚没有因缺乏对受试者和工作人员的盲法导致偏倚的实证证据。但有证据表明，描述为“盲法”或“双盲”的研究常常是对一组或两组人员进行盲法。实证研究表明，缺乏盲法的随机试验其干预效应比值比（OR）估计值平均夸大9%（Pildal 2007）。这些研究处理了多种结局，其中一些为客观结局。一般来说，越主观的结局，观察到的偏倚

越多（Wood 2008）。不论结局指标的类型，如果干预组在新增试验人员或合并干预方面不同，也可因缺乏盲法而导致偏倚。

对于有些人来说实施盲法几乎是不可能的（如接受手术治疗的病人）。但这种研究可采取其他措施减少偏倚，如治疗采用严格的诊疗计划，以减少病人和医护人员不同的行为。对受试者和试验人员实施盲法，并不能确保盲法成功的实施。对于大多数干预盲法也是会打折扣的。对于许多采用盲法的药物试验，药物的副作用可能提示某些受试者接受的是何种干预措施，除非该研究比较的是两个类似的药物，如副作用相似的药物，或使用活跃的安慰剂（Boutron 2006）。

在采用盲法的研究中，特别是安慰剂对照的研究，应考虑是否真的对受试者施盲（有时也会考虑干预提供者）。一些研究小组建议，在试验结束时让受试者猜测自己接受的治疗措施（Fergusson 2004，Rees 2005），并且这些报告的一些综述已经发表(Fergusson 2004，Hróbjartsson 2007）。超过50％的受试者猜测正确就表明盲法可能被破坏，但在实际中可以简单地反映试验中受试者的体验：一个好的结局或者明显的副作用往往归因于阳性治疗，差的结局则归因于安慰剂（Sackett 2007）。我们认为，若盲法一直保持着，当干预措施疗效或者出现不良反应的有差异时，可以认为有些猜测是正确的，但当疗十分效相似时，便不会有正确的猜测。因此，系统评价员应慎重考虑这样的研究结果。

框8.11.a　临床试验中的盲法

一般来说，盲法是指研究的受试者和试验人员，包括结果评估者，在受试者进入试验后不知道干预措施的分配情况。盲法可以减少因干预措施被知晓后，非干预措施因素影响结局和结局评估的风险。

临床试验中可对受试者和工作人员实施盲法的类型有以下几种（Gøtzsche 1996，Haahr 2006 年）：

1、受试者（如病人或健康人）；

2、医疗服务提供者（如医生或负责护理的护士）；

3、结果评估者，包括原始数据收集人员（如负责测量和收集结局数据的随访人员）和任何辅助评估者（如外部结局评审委员）；

4、数据分析者（如统计人员）；

5、撰稿者。

在条目“对受试者和研究人员的盲法”描述了前两类人群。条目“结局评估中的盲法”表述了第三类人群。最后两类没有做明确的表述。

8.11.2 对受试者及研究人员的盲法是否完善的偏倚风险评价

研究报告描述的盲法常比较宽泛，如“双盲”。这样就不可能知道被实施盲法的对象是谁（Schulz 2002a）。这些术语使用时也非常不一致（Devereaux 2001, Boutron 2005, Haahr 2006）。尽管CONSORT声明里有明确的建议（Moher 2001b），但即使发表在高端医学期刊上的试验，明确报告对受试者和工作人员施盲的研究仍然很少（Montori 2002）。一篇关于盲法使用的综述强调实践中实施盲法的多样性（Boutron 2006）。使用Cochrae协作网的工具时，下列因素可能有助于系统评价员评估研究中盲法的使用是否足以防止偏倚（8.5节）。

在考虑对受试者和研究人员的盲法不完善导致的偏倚风险时，重点考虑下面几点：

1、谁被施盲，谁未被施盲；

2、研究中因盲法不完善产生偏倚的结局指标（如因为共同干预或行为差异）；

即使研究中同一类型人员未被施盲，不同结局指标的偏倚也可能不同。例如知道分配的干预措施可能会影响行为结果（如就诊次数），但不会影响生理结果或死亡结局。即使结果评估者知道干预措施的分配，许多情况下总死亡率评估可能被认为是无偏倚的。因此，评估因盲法不完善引起的偏倚风险的时需要对不同结局指标分开处理。

通常比较简便的做法是对偏倚风险相似的结局指标分组评估，而非单独评估每个结局指标（见8.5节）。例如，将所有主观结局指标与客观结局指标分开，并各自进行一个总体评估。

8.12 对结局盲法的评估

8.12.1 偏倚相关的理论

在一个临床试验中，可以对几种人实施盲法：见8.11.a表。表中第三、四条关注的是对结果的评估者进行盲法。如果结果评估者清楚干预措施的分配情况，那这些评估的结果中可能有偏倚存在。结果评估者可以是受试者自己，也可以是医务人员或者独立的结果评估者。

实证研究表明，未实施盲法的随机对照试验干预效果的OR值平均夸大9%（Pildal 2007）。这些研究涉及到多种结局，有些是客观的。通常如果试验的结局越主观，则产生的偏倚越大（Wood 2008）。

所有的结果评估均可能受到未实施盲法的影响，尤其是主观性结果（比如疼痛，感冒的天数等）。因此在考虑盲法时，应考虑结局的主观性或者客观性如何。在同一个试验中不同结局，盲法的重要性及可能性是不同的。表面上看起来是客观性的评估也可能存在主观性，如医生评估心理或生理伤残的程度（Noseworthy 1994）。

结局评估的盲法有时是不可能的（比如病人做了大手术）。然而，这并不意味着潜在的偏倚能被忽略，并且系统评价必须评价整篇系统评价纳入的所有研究由于结果评估缺少盲法而产生的偏倚。

8.12.2 结局评估中是否正确实施盲法的偏倚风险评估

研究报告盲法时常常采用较宽泛的术语，如“双盲”。这样就不知道到底是谁被盲（Schulz 2002a）。这些术语使用时也非常不一致（Devereaux 2001, Boutron 2005, Haahr 2006），并且一些顶尖杂志中的试验对于研究的参与者与实施人员的盲法情况的报告率很低（Montori 2002），尽管CONSORT中的推荐是明确的（Moher 2001）。一篇关于盲法使用的方法综述强调实践中实施盲法的多样性（Boutron 2006）。当使用协作网的工具时，以下的推荐将帮助系统评价作者评估研究中盲法的采用是否足以使研究免受偏倚（8.5节）。

在考虑结局评估中因缺少盲法所致的偏倚风险时，重点考虑以下几个方面：

1、由谁来评估；

2、评估中存在的偏倚风险（考虑结局的主观性或客观性）。

对于一些结局的评价者可能被施盲，而其它的评价者则没有。例如，在一项病人知道自身干预措施的有关手术的试验中，病人报告的结局（如生活质量）即明显是在知晓所受干预的情况下搜集的，而其它由独立的医师来测量的结局（如体能），则可能是符合盲法的。即使研究中同一类型人员未被施盲，不同结局指标的偏倚也可能不同。比如，知道所分配的干预措施，可能会影响病人报告的结局（如疼痛的水平），而死亡率等结局则不受影响。在大多数情况下，认为总死亡率的评价不存在偏倚，即使结局的评价者知道干预措施。因此对于未实施盲法而产生的偏倚风险的评估要根据不同结局分别考虑。

通常比较简便的做法是对偏倚风险相似的结局指标分组评估，而非单独评估每个结局指标（见8.5节）。例如，将所有主观结局指标与客观结局指标分开，并各自进行一个总体评估。

8.13 不完整结果数据

8.13.1 偏倚相关原理

研究中失访或分析时被排除导致结局数据缺失，进而增加了所观测的效应估计值偏倚的可能性。使用术语“不完整结果数据（incomplete outcome data）”来表示失访和排除。当单个受试者的结果无法得到时，我们将称其为缺失。失访发生的可能原因：

- 受试者从研究中主动或被动退出。
- 受试者不参加本来应该参加的结果测量。
- 受试者参加了测量，但没有提供相关数据。
- 受试者未完成记录或问卷。
- 不能找到受试者（失访）。
- 因研究者不适当的决定中止了随访。
- 数据或记录丢失，或者因其他原因无法获得。

此外，一些受试者可能在时被排除，原因如下：

- 入组后发现受试者不合格。
- 进行了“实际治疗”（依从方案）分析（只有接受了计划书中的预期的干预措施的受试者才被纳入；见第8.13.2）。
- 受试者因其他原因被排除分析。

在排除某些受试者不会导致结局数据缺失时，可能是符合要求的情况（Fergusson 2002）。例如：发现随机分配后的受试者不合格则应排除，只要不合格的受试者未受到随机化的干预措施的影响，且基于盲法分配所做的决定。排除这些受试者的意图应在观察结局出现之前就已详细说明。

意向性分析（intention-to-treat，ITT）常被看作随机试验中评估干预效果偏倚最小的方法（Newell 1992）：见第16章16.2节。ITT分析的原则如下：

1、确保受试者在原随机分配的干预组，不管他们实际上是否接受了该干预措施；

2、测量所有受试者的结局数据；

3、分析中纳入所有随机分配的受试者。

第一条原则总是使用。但第二个原则常因试验者不能控制失访而不能实施，因此，第三个原则仅在对缺失数据做出估计的前提下进行（见下文）。因此如果没有进行数据

填补，几乎没有研究能真正做到意向性分析，尤其在随访时间很长时。实际中，即使某些结局数据丢失，研究者也可能将其描述为ITT分析。“ITT”一词并没有一个明确、一致的定义，并且在研究报告中的使用也不一致（Hollis 1999）。只有在遵循所有上述的三个原则时，系统评价者才能使用该术语，并对使用该术语但没有说明的研究谨慎解释。

系统评价者也可能会遇到描述为“调整的意向性分析（modified intention-to-treat）”的分析，这通常是指如果受试者没有接受最少数量的指定干预措施将会被排除。这个术语在多种方法中使用，所以系统评价者应收集其所包含的准确信息。

如果排除的理由不恰当，并且评价作者可获得相应的数据，那么可能考虑将研究作者排除的病人再次纳入分析（重新纳入）。在可能和适当的情况下，我们支持系统评价者这样做。

结局数据不完整导致的偏倚主要是从理论上考虑的。一些实证研究着眼于各方面丢失数据是否与效果估计值的大小有关。大多数研究没有发现存在偏倚的明显证据（Schulz 1995b，Kjaergard 2001，Balk 2002，Siersma 2007）。Tierney等观察到，与纳入所有受试者的分析相比，排除受试者数据后的分析结果偏向于支持试验组的干预措施（Tierney 2005）。但是，有一些“依从方案”分析所致偏倚的典型例子（Melander 2003），并且有一篇系统评价已经发现，对于同一个试验，与“ITT”分析相比，“依从方案”分析会更加夸大效应估计值（Porta 2007）。因为对排除情况很少报告，特别是在1996年CONSORT出现之前（Moher 2001），就使得对着这些实证研究结果的解释十分困难。例如：Schulz观察到“明显”缺乏排除数据的研究，会得出更“有利”的效应估计，同时分配隐藏充分的可能性减小（Schulz 1996）。因此Schulz的研究中没有报告排除数据可能是因为试验实施较差，而不是真的没有被排除的数据。

实证研究还调查了试验报告中对不完整结局数据的处理是否恰当。一项根据来自4本医学杂志的71个试验报告的研究表明，数据缺失是常见的，而且往往在统计学分析中没有进行充分的处理（Wood 2004）。

8.13.2 结局数据不完整所致偏倚风险的评估

结果数据不完整引起的偏倚取决于几个因素：缺失数据在不同干预组的数量和分布、结果缺失的原因、有或无结局数据的受试者间可能的差异，以及研究者在分析报告和临床条件下针对此问题所采取的措施及临床意义。因此，以一个简单的规则来判断偏倚风

险的高低不太可能。在使用Cochrane协作网的工具时，下列因素可能有助于系统评价员评估不完整的结局数据是否以无偏倚的方式进行了处理（8.5节）。

通常常认为缺失数据的比例大，或者不同干预组缺失数据比例差异较大是引起偏倚的主要原因。然而，这些特征本身不足以产生偏倚。这里，我们将详细说明某个分析判断为低偏倚风险或高偏倚风险的情况。同时，必须考虑数据缺失的数量和缺失的原因。

8.13.2.1 结局数据不完整所致的低偏倚风险

要得出无缺失结局数据的结论，评价员者应确信分析中纳入分析的受试者人数与随机分配入组的人数一致。如果随机分配到各个干预组的人数没有明确报告，则偏倚风险为不清楚。如上所述，随机分配的受试者后来发现不合格的情况，并不总是被视为结局数据缺失。

例子（低偏倚风险）：“所有患者完成了研究，没有失访，治疗中没有退出，试验组没有变化以及无严重不良反应事件”。

缺失数据可接受的理由：

一个健康的人决定搬离临床试验所在的地理位置，并不太可能得到他们后续的结局。对于随访期较长的研究，这样的退出是不可避免的。

对报告时间-事件（time-to-event data）数据的研究而言，所有在最后一次随访中都没有出现结局事件的受试者则被认为是“截尾”数据（我们不知道随访结束后结局事件是否会发生）。对于这种类型的分析，其考虑重点是：是否可以将这样的截尾认为是无偏倚的，即在定的随访结束前，干预措施在截尾的个体与其他个体中的效果（如危险比作效应指标）相同。换言之，若截尾与预后无关则没有偏倚。

如果两干预组均存在缺失数据，但报告了缺失原因且组间原因相似，那么，除非这些原因在对比较组有不同的影响，否则不会产生重要偏倚。例如，“拒绝参与”可能意味着在锻炼组不愿锻炼，然而，拒绝也可能意味着不满意另一组不锻炼的建议。在实际中，对缺失原因报告不完整可能会妨碍系统评价员作出评估。

缺失数据对效果估计的潜在影响

二分类的结果数据缺失的潜在影响依赖于结局出现的频率（或风险）。例如，如果10%的受试者缺失数据，那么对事件风险概率为10%的潜在影响比5%的要大得多。下表说明了所观察到的风险的潜在影响。假设A和B分别代表两个试验，1000名受试者中90%的个体观察到了结局，并且900名受试者中的相对危险度（RR）为1。此外，在这两个

试验中，我们假设干预组失访病人具有事件发生的高风险（80%），对照组则比较低（20%）。A与B唯一的区别是所观察的受试者的风险不同。A试验的风险为50%，根据已有的观察，缺失数据的影响较低。B试验的风险为10%，根据已有的观察，缺失数据的影响较大。一般来说，发生事件的受试者缺失数据的比例越高，偏倚的可能性越大。A试验缺失的比例是100/450（0.2），而B试验是100/90（1.1）。

	随机分配人数	观察对象的风险	观察数据	缺失受试者的假设风险	缺失数据	完整数据	所有受试者的比值比
A 试验							
干预	500	50%	225/450	80%	40/50	265/500	1.13
对照	500	50%	225/450	20%	10/50	235/500	
B 试验							
干预	500	10%	45/450	80%	40/50	85/500	1.55
对照	500	10%	45/450	20%	10/50	55/500	

连续性变量结果数据缺失的潜在影响因缺失数据的受试者的比例增大而增加。还必须考虑在缺失结局的受试者中合理的干预效果。下表显示了不同比例缺失数据的影响。A和B代表两个假设的试验，各有1000个受试者，试验组和对照组干预效应的均差为0。此外，我们假设，干预组失访的病人均值较大，而对照组较低。试验A与B的唯一区别是失访人数的不同。在试验A中，90%的受试观察到结局，10%失访，缺失数据对均差的影响较小。在试验B中，一半的受试者失访，缺失数据对均差影响很大。

	随机分配人数	观察人数	观察均值	失访人数	失访病人的假定均值	整体均数（加权平均法）	所有受试者的平均差
A 试验							
干预	500	450	10	50	15	10.5	1
对照	500	450	10	50	5	9.5	
B 试验							
干预	500	250	10	250	15	12.5	5
对照	500	250	10	250	5	7.5	

8.13.2.2 不完整结果数据所致的高偏倚风险

缺失数据不能接受的理由

如果结果数据的可用性由受试者的真实结局决定，组间不完整数据比例的差异则应该被考虑。例如临床结局差的受试者更可能因为不良反应而退出并且这主要发生在试验组，那么对干预措施的效应估计将会发生偏倚，而倾向于支持试验组的干预。如果因“无效”或“无好转”而排除受试者人数在各组间不平衡，则也会产生偏倚。需要注意的是即使对缺失数据进行差异性检验的结果无统计差异，也并不能确定避免了偏倚，特别在小样本研究中。

例子（高偏倚风险）：“在西布曲明与安慰剂对照治疗肥胖的试验中，西布曲明组35例有13例退出，其中7例因为无效退出。安慰剂组34例有25例退出，17例因为无效退出。一个只纳入了剩余的受试者的“意向性分析”"（Cuellar 2000）（即安慰剂组只剩9例）。

即使不完整结果数据的人数在各组间平衡，如果缺失的原因不同也可引起偏倚。例如，在一个研究戒烟的干预试验中，可能出现：对照组中一定比例的受试者因对不新颖的事物缺乏热情而退出研究（继续吸烟）的比例，而试验组中一定比例的受试者则会因戒烟成功而退出试验。

戒烟研究中，处理缺失数据常见方法（假设每个退出试验的人都继续吸烟）并不总能避免偏倚。这个例子强调了评估偏倚风险时考虑结局不完整数据原因的重要性。实践中，知道每个受试者数据缺失的原因不太可能，尽管实证研究证据显示在63个有缺失数据的研究中有38个提供了数据缺失的原因（Wood 2004），当然这也可通过使用CONSORT声明来改善（Schulz 2010）。

“实际治疗”（依从方案）分析应根据受试者随机分配入组时的情况进行分析，而不管他们是否真的接受该干预措施。因此，在一个比较前列腺癌手术与放射治疗的研究中，随机分组后拒绝手术而选择放射治疗的患者应纳入手术组进行分析。这是因为受试者改变干预措施的倾向可能与预后相关，并可能导致选择性偏倚。除非改变的人数不足以对干预措施的效应估计值产生重要的影响，否则报告“实际治疗”分析应的研究被认为是因结局数据不完整导致的高偏倚风险，虽然严格地讲，这是分析方法不当而不是结局数据不完整。

另一个类似的不恰当的研究分析方法是只关注按计划书接收干预的受试者。一个典型的例子：安妥明用于降低血脂的药物试验（Coronary Drug Project Research Group 1980）。安妥明组1103人的5年死亡率为20.0%，而安慰剂组2789人的死亡率为20.9%（P=0.55）。

安妥明组依从性好的受试者5年死亡率（15.0%）比依从性差的死亡率（24.6%）低。同样依从性“好”与“差”之间的差别也在安慰剂组观察到（15.1% vs 28.3%）。因此，依从性与预后高度相关，而不是安妥明的疗效。这些结果表明，基于病人对干预措施的反应所决定的亚组中来评估干预措施的效果是非常困难的。因为未接受干预措施比无法得到结果的数据提供的信息多，尽管不完整的数据比例较小，将分析限制在依从的受试者也会出现高偏倚风险。

8.13.2.3 处理报告中的缺失数据：填补（imputation）

处理缺失数据的一种常见但存在潜在风险的方法是填补结果，将其当做好像真正测量过一样（见第16章16节）。例如，有缺失结果数据的个体可指定为该干预组的平均效应，或者治疗成功或失败。但这样处理会导致严重偏倚和过窄的可信区间。另一种有些不同的，且有效性更难评价的方法是末次观测值结转（last observation carried forward，LOCF）。在这里，我们将最后一次测量的结果假定为后续研究时点所有的测量结果（Lachin 2000，Unnebrink 2001）。LOCF也会导致严重的偏倚，例如，在一个治疗退行疾病药物的试验中，如阿尔茨海默病，失访可能与该药物的不良反应有关。因为结局往往随时间恶化， LOCF将会使效应结果向支持该药物的方向偏倚。另一方面如果大多数人结转的结果与真实的测量值接近，那么使用LOCF是比较适当的。

有大量以有效方式的统计方法处理缺失数据的文献，见第16章（16.1）。但这些方法在实际的临床试验报告中应用相对较少（Wood 2004）。如果系统评价员遇到此类方法，最好寻求统计学家的意见。如果要深入了解可链接www.missingdata.org.uk。

8.14 选择性的结果报告

8.14.1 偏倚相关原理

选择性的结果报告被定义为：根据研究结果对所纳入文章中记录的原始变量进行选择；见第10章（见10.2.2.5）。特别关注的是，无统计学差异的结果可能被选择性地拒绝发表。目前所发表的关于选择性结果报告的证据是有限的。最初仅有一些个案研究。一个通过了当地研究伦理委员会批准的的小样本研究，完成队列随访后发现15个发表的研究中只有6个在计划书中报告了主要结局指标，8个提到了预期的分析策略，但其中7个

发表研究都没有按计划进行数据分析（Hahn 2002）。在一项对于Cochrane 系统评价数据库中发表的5篇Meta分析所纳入的试验进行的综述中，研究内有明显或可疑的选择性报告（Williamson 2005a）。

研究存在选择性报告偏倚令人信服的直接证据来自最新的三个研究。第一个研究（Chan 2004a）中，有122篇研究报告发表和3736个结局指标的102个试验。总体而言，每个平行分组试验有（中位值）38%的疗效指标和50%的安全性指标没有完全报告，即无可以被纳入Meta分析的足够信息。与无统计学差异的结果相比，有统计学差异的结果在全文中报告的比例较高，对于疗效指标（合并的比值比为2.4；95%可信区间1.4-4.0）和伤害指标为（4.7；1.8-12）均是如此。此外，在比较发表的研究报告与计划书时，62%的试验至少有一个主要结局指标被更改、添加或删除。第二个研究纳入了由加拿大卫生研究院资助的48个试验，得到了大致相同的结果（Chan 2004b）。第三个研究涉及对519个试验的回顾性分析和作者的随访调查，比较了同一篇文章中报道的结果指标与在方法部分所提到的结局指标（Chan 2005）。平行分组试验中平均有超过20%的结局指标报告不完全。与完全报道的结局指标的研究相比，试验中这样的结局指标有很大的风险是无统计学差异的（疗效指标的比值比为2.0（1.6-2.7）；有害的结局指标为1.9（1.1-3.5））。这三个研究表明与选择性的结果报告相关的比值比（OR）约为2.4，例如，与72%有统计学差异的结果发表相比，大约只有50%无统计学差异的结果被发表。

在所有的3项研究中，作者均被问及是否有未发表的结果，这些结果是否有统计学差异，以及这些结果为什么没有发表。对于未发表的结果，最常见的原因是“缺乏临床意义”或“缺乏统计学意义”。因此，未包括没有发表的结局的Meta分析，可能高估干预疗效。此外，即使在计划书和研究报告中提及的结局，作者一般也不提及这些结局的存在。

最近的研究也得出了类似的结果（Ghersi 2006, von Elm 2006）。在不同类型的研究中，当可用于Meta分析数据的研究较少时，其综合效应会偏大（Furukawa 2007）。这表明有些结果可能会被试验者基于效果值大小而有选择性地不报告。

对于具有相同特征的不同测量指标的选择性报告有关的偏倚也相似。在治疗精神分裂症的试验中，当采用量表进行评价时，未发表的研究比发表的研究更易观察到相似的干预措施效果（Marshall 2000）。作者假设来自未发表量表的数据在其无统计学意义时不太可能被发表，或者在后续的数据分析中，会将不利的条目也被踢出以得出明显有利的效果。

在许多系统评价中，针对某一结局指标，只有少数研究可以纳入Meta分析，因为许多研究并没有提供必要的数据。尽管在一些研究中没有对结局指标进行评价，但对于一些研究来说几乎总是存在一种报告偏倚的风险。系统评价员需考虑结果是收集了但未报告还是根本未收集。

选择性结果报告可以多种形式出现，可作为整体影响研究（下面第一点），而有的则影响特定的结局指标（2-6点）：

1、结局报告中选择性的省略：发表的报告只纳入了部分分析结果。如果选择是基于这些研究结果，特别是有统计学意义的结果，那么相应Meta分析的估计值将可能是有偏倚的。

2、结局数据的选择性抽取：对于某个具体的结局指标，可能有不同的测量时点，或者在同一时点所用的测量工具可能不同（如不同的量表或不同的评价者）。例如，在一个骨质疏松试验的报告中，有可供选择的12个不同数据集来估计骨中矿物质含量。这12项数据的标准化均数差在-0.02和1.42之间（Gøtzsche 2007）。如果研究者根据这些结果做出选择，那么Meta分析的估计值将是有偏倚的。

3、选择性报告采用同一数据的分析结果：结果分析通常有若干不同的方法。例如，连续性结局如血压降低值，既可作为连续性也可作为二分类变量进行分析，还可进一步选择多个分界点进行分析。另一种常见的分析可选择终末得分与较之基线的变化（Williamson 2005b）。常将预期的最终值比较转化为基线变化的比较，是因为基线间的不平衡会引入偏倚，而非去掉它（研究者可能这样猜想）。（Senn 1991，Vickers 2001）。

4、选择性报告亚组数据：如果结局数据可再细分，那么就会出现选择性报告的情况，如选择一个完全测量表的子量表或结局事件的一个亚组。例如，真菌感染可能在基线或随机分配后一两天内检测到，或者在随机后的好几天内检测到所谓的“突破”性真菌感染，如果仅选择这些感染的亚组就可能导致报告偏倚（Jørgensen 2006，Jørgensen 2007）。

5、选择性漏报数据：一些研究虽然报告了某些结局，但无可以纳入Meta分析数据的详细信息。有时这与结果明确相关，如仅报告描述“无显著性”或“P>0.05”。

选择性报告的其他形式在这里不做详细介绍，如亚组分析的选择性报告、校正分析的选择性报告以及交叉试验中第一期结果的选择性报告（Williamson 2005a）。另外，“主要”、“次要”等结局描述可能知道结果后有所改变（Chan 2004a，Chan 2004b）。只要它不影响所发表结果，系统评价者（不留意各个研究中比较突出的结果的评价者）一般不关注这个问题。

8.14.2 选择性报告结果所致偏倚风险的评估

虽然研究间的发表偏倚只有通过考虑所有研究的情况下才能估算出概率值（见第10章），但研究内选择性报告结局的概率可通过系统评价中纳入的每一个研究进行检查估计。在使用Cochrane工具时，下列因素可能有助于系统评价者评估结局报告是否足够完整、透明以避免偏倚（8.5节）。

检测研究内选择性报告结局的统计方法尚未制定出来。但是，存在检测这类偏倚的其他方法，尽管彻底的评估有可能很费力。如果可获得计划书，就可以进行计划书中的结局和发表报告的比较。如果没有，那么可进行文章方法部分的所列结局与结果部分报告的结果间比较。如果提到了无显著性的结果但没有充分报告，Meta分析可能会出现偏倚。更多信息也可通过该研究报告作者的得到，但应该意识到这种信息可能是不可靠的（Chan 2004a）。

计划书和发表物间的差别可能是计划书合理的改变。虽然这种变化应在发表中报道，但是Chan等人两个样本的150个研究中都没有这样做（Chan 2004a，Chan 2004b）。

系统评价者应通过在某个领域中针对某个问题常规地测量了少数的关键结局指标的研究者来努力收集证据。系统评价者应该考虑在Meta分析中可能缺失数据的原因（Williamson 2005b）。寻求这些证据方法还未建立，但我们描述了一些可能的策略。

比较有用的第一步是建立一个矩阵，说明有哪些研究记录了哪些结果，如以行表示不同的研究，列表示不同的结局指标。完整和不完整的报告也须注明。这个矩阵将显示某些研究并没有报告的结局被大多数其它研究如何报告。

撰写研究计划书应检索Pubmed、其他主要的引文数据库以及互联网；少数的情况下，在研究报告中会给出网址。或者，不久的将来随着试验强制性登记变得越来越普遍，可试验登记库的获得研究的详细描述。关于研究的描述性摘要可能包含在后续的发表物中未提到的信息。此外，系统评价者应该仔细检查发表文章的方法部分，以获得被评价结局的详情。

应特别重视的是那些理应记录的缺失信息。例如，有些指标是同时出现的，如收缩压和舒张压，所以如果只报告其中一个时我们应该探究其原因。另一个例子是：报告连续性变量的改变超出某些阈值的受试者的比例的研究；调查人员必须有原始数据的使用权，因此才能以变化量的的均值和标准差显示结果。Williamson等人举了一些例子，包括一个Cochrane系统评价，其中有9个试验报告了治疗失败的结局但只有5个试验报告了

死亡率。然而，死亡率是治疗失败的的一部分，因此这四个缺少死亡率分析的试验的数据一定也被收集了。单独报告或未报告死亡率的研究间在治疗失败结果方面的显著性差异能提示偏倚（Williamson 2005a）。

当怀疑或有选择性报告结果的直接证据时应询问研究作者更多的相关信息。例如，可要求作者提供研究计划书和报告不完整结局的所有信息。此外，文章或计划书中提及但未报告的结局，可以要求他们阐明这些结局测量指标是否真的进行了分析，如果有则提供有关资料。

在主要Meta分析中一般不建议进行报告偏倚“校正”。敏感性分析是调查选择性的报告结局的可能影响的一种较好的方法（Hutton 2000，Williamson 2005a）。

对于选择性报告结果引起的偏倚风险的评价应将研究作为一个整体，而不是单个结局。尽管可能清楚的是对于某些特定的研究，某些结果易于受到选择性报告的影响，而其它的不会，但我们建议采用研究水平方法，因为在“偏倚风险”表里列出所有完全报告的结果不太实际。工具的“判断依据”（见8.5.2节）应用来详细描述选择性（或不完整）报告的证据。该研究水平的判断评价了研究对于选择性报告偏倚的整体敏感度。

8.15 对有效性的其他潜在威胁

8.15.1 偏倚相关理论

以下方面（序列生成、分配隐藏、盲法、不完整结局数据和选择性报告结局）与所有医疗保健领域的临床试验偏倚的重要潜在来源相关。除了这些具体的条目，系统评价者应该注意可能导致偏倚的更多问题。“偏倚风险”评价工具中的第七个条目是其他的偏倚来源的“收集箱”（catch-all）。对一些主题领域的系统评价来说，存在对所有研究都应该问的更多的问题。特别是一些研究设计应特殊考虑时。如果某研究设计是预期的（如交叉试验、非随机的研究类型），这类型的研究将涉及更多有关偏倚风险的问题。非随机研究偏倚风险的评估见第13章，整群随机试验、交叉试验以及多个干预组试验的风险评价见第16章。此外，一些主要的、意料之外的、特定研究的问题在系统评价或Meta分析过程中可能会被发现。例如，受试者的基线特征相当不平衡。本章将在下面讨论几个例子。

8.15.1.1 具体设计的偏倚风险

非随机化研究中首要考虑的偏倚风险是因试验组与对照组受试者类型不同的而存在的选择性偏倚。系统评价者应参考第13章（13.5节）对此详细讨论的部分。整群随机试验主要关注的偏倚风险的问题是：(i)招募偏倚（不同干预组间招募了不同的群体）；(ii)基线不平衡；(iii)群体的丢失；(iv)分析不当；(v)与单个随机试验的可比性。交叉试验主要关注的偏倚风险问题是：(i)交叉设计是否适当；(ii)是否有携带效应；(iii)是否只有一期数据可用；(iv)分析不当；(v)与平行分组试验结果的可比性。这些将在第16章中具体讨论（16.3节和16.4节）。两个及其以上干预组的研究的偏倚风险也在第16章中讨论（16.5节）。

8.15.1.2 基线不平衡

与结局测量密切相关因素的基线不平衡可导致干预措施效应估计的偏倚。这可能仅仅因为机遇而产生，但干预措施的非随机（未隐藏）分配也会导致不平衡增加。有时试验作者可能排除了一些随机分配的个体，从而引起受试者特征在不同干预组间的不平衡。序列生成，无分配隐藏或排除受试者，都应在工具中用相应条目进行分别处理。如果观察到更多令人费解的基线不平衡并足以导致效果估计值的严重夸大增大，那么应该进行备注。对于真正的随机试验检验其基线不平衡毫无意义，但非常小的P值提示在干预的分配中存在偏倚。

例子（高偏倚风险）：在卡托普利vs常规抗高血压药物试验中，组间的身高、体重、收缩压和舒张压均存在很小但非常显著不平衡：P= 10-4-10-18（Hansson 1999）。这种组间不平衡表明在一些中心随机化（密封信封）失败（Peto 1999）。

8.15.1.3 非盲法试验的区组随机

一些序列生成方法、分配隐藏方法和盲法的组合共同产生了在干预措施分配中的选择性偏倚风险。一个特别的组合方法是在非盲的试验或盲法被破坏（如因为副作用的特点）的盲法试验中使用区组随机化。采用随机区组时，且分配被后续招募进入实验的受试者知道时，那么有时可以预测到未来的分配，这在区组大小固定且没有被分到多个招募中心时，尤其是个问题。即使按表8.5.d给出的标准进行充分的分配隐藏，这种预测未来分配的可能性也会存在（Berger 2005）。

8.15.1.4 不同的诊断活跃性

尽管盲法有效，结果评估也可能出现偏倚。尤其是诊断活动增加可能导致事实上存在却无伤害的疾病的诊断增加。例如，许多胃溃疡患者无症状也没有临床相关性，但这些病例在服用药物产生胃部不适而进行更多的胃镜检查时，则会检测出更多的胃溃疡。同样，如果药物引起腹泻，这可能引起更多的直肠镜检查，因而，也会检查出更多前列腺癌的无害病例。显然，有益效果的评估也可因这样的原因产生偏倚。干预措施也可能导致不同的诊断活跃性，例如，如果试验干预是护士在病人家里进行访视，而对照措施是不随访的。

8.15.1.5 潜在偏倚的更多例子

如下临床研究的其他潜在的偏倚来源的列表可能有助于发现进一步的问题。

- 这项研究的实施因中期结果受到影响（如从得出更有利结果的亚组中招募额外受试者）。
- 研究计划书未反映临床实践的偏差（如事后逐步增加剂量至极大的水平）。
- 干预措施在随机化之前的实施能加强或减小亚组、随机、干预的效果。
- 某项干预措施分配不恰当（或者共同干预）。
- 沾染（如受试者混合了药物）。
- 干预措施实施不充分或过度扩大纳入标准引起的“零偏倚”事件（Woods 1995）。
- 用于测量结果的工具不敏感（这可能导致对有利和有害效果的低估）。
- 选择性报告亚组。
- 欺骗。

资助者（或者一般来说和结果有利益相关者）的不恰当的影响通常被认为是一个重要偏倚。例如在一个实证研究中，在超过一半由工厂发起的试验的计划书中提到，试验发起者拥有数据或需要批准手稿，或两者都有；这些局限都未在任何试验的发表物中说明（Gøtzsche 2006）。收集利益相关者并呈现这些信息很重要。因此系统评价者必须在“纳入研究特征”表提供这个信息（见11.2.2）。“偏倚风险”表应在具体的方法学方面评价受利益相关者影响以及由此可能直接导致的偏倚。应备注一些可能受到利益相关者影响的决定，比如选择极低剂量的对照药物，应表述“异质性来源”而不是通过“偏倚风险”工具，因为它们不会直接影响到结果的内在效度。

8.15.2 其它来源偏倚风险的评估

确定合适的主题作为“其他来源”的一般准则将在下面介绍。特别是，适当的主题应构成潜在的偏倚来源而非不准确的来源，多样性的来源（异质性）或与偏倚无关的研究质量指标。工具在该条目下的主题包括之前在8.15.1介绍的例子。然而，除了这些具体的问题，系统评价者应注意到可能导致偏倚的研究具体的问题，并应制订本领域的工具对此进行判断。使用Cochrane协作网的工具时，下列因素可能有助于系统评价者评价研究是否存在其他来源的偏倚风险（8.5节）。

只要有可能，系统评价计划书应预先明确任何需要解决的问题，这会导致“偏倚风险”表中的条目分离。例如，如果交叉试验是系统评价所处理的问题的常用研究设计，那么与交叉试验偏倚相关的具体问题应提前明确。

由偏倚风险评价工具引起的问题一定是偏倚的一个潜在来源，而不仅仅是不精确的原因（见8.2节），这适用于“其他来源偏倚”的评价。某个潜在的偏倚来源必定能改变效应估计值的大小，而不准确的来源只影响估计值的不确定性（即可信区间）。影响评估精确性的的潜在因素包括技术上的变异性（如测量误差）和观察员的变异性。

由于该工具只涉及内部偏倚，在这个方面的任何问题都应是内部偏倚的潜在来源，而非多样性的来源。多样性的原因包括药物剂量、随访时间和受试者特征（如年龄、疾病阶段）的差异。研究可能选择利于试验药物与对照药物比较的剂量。例如，旧的药物往往过量（Safer 2002）或在并未反映临床实践的次优的情况下给出（Johansen 2000，Jørgensen 2007）。此外，受试者可能基于以前证明对试验组干预有反应的情况下被选择性的纳入研究。Cochrane系统评价中对这样的偏倚性选择进行处理很重要。虽然未在本章“偏倚风险”工具中描述，但是有时会在分析时进行处理（例如通过亚组分析和Meta回归），并应在“结果总结”表的证据分级和解释中予以考虑（见11和12章）。

对于临床试验的设计和实施可作出许多判断，但并非所有的都与偏倚有关。“质量”指标常常与引起偏倚的方面有很强相关性。然而，系统评价者应关注引起偏倚的原理，而不是对反映“质量”的研究的描述。不应在该部分进行评估的“质量”指标包括标准的适用性、“普遍性”或“外部真实性”（包括如上述的），标准相关的精确性（如样本量或样本量（或效能）计算）、报告标准、伦理准则（如研究是否得到伦理批准或受试者是否知情同意）。这些因素可能很重要，但应在“纳入研究特征”表或附表中进行描述（见第11章）。

最后，为避免重复计算，如果他们在工具中更适宜由排在前面的条目报告，潜在的偏见来源不应该被列为“其他来源的偏倚”。例如，阿尔茨海默病，患者在试验期间随时间推移显著恶化。一般来说，治疗的效果较小，且治疗措施有明显的毒性。想要处理好受试者的失访是非常困难的。治疗组的受试者可能因为副作用或死亡而提早脱落，因此对这些人的测量是在在研究的初期，这将支持干预措施。一般很难得到受试者的连续监测以便进行对随机分配的所有受试者进行分析。这个问题尽管第一眼会觉得是一个特定主题的偏倚，但更为合适是作为不完整结局数据。

8.16 本章信息

编辑：Julian PT Higgins, Douglas G Altman and Jonathan AC Sterne on behalf of the Cochrane Statistical Methods Group and the Cochrane Bias Methods Group.

本章引用格式：Higgins JPT, Altman DG, Sterne, JAC (editors). Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

特约作者：Doug Altman, Gerd Antes, Peter Gøtzsche, Julian Higgins, Peter Jüni, Steff Lewis, David Moher, Andy Oxman, Ken Schulz, Jonathan Sterne and Simon Thompson.

致谢：感谢 Hilda Bastian, Rachelle Buchbinder, Iain Chalmers, Miranda Cumpston, Sally Green, Peter Herbison, Victor Montori, Hannah Rothstein, Georgia Salanti, Guido Schwarzer, Ian Shrier, Jayne Tierney, Ian White和 Paula Williamson的有益点评. Cochrane统计方法学组的详情见第9章框9.8.a，Cochrane偏倚方法学组的详情见第10章框10.5.a。

8.17 参考文献

Als-Nielsen 2004

Als-Nielsen B, Gluud LL, Gluud C. Methodological quality and treatment effects in randomized trials:a review of six empirical studies. 12th Cochrane Colloquium, Ottawa (Canada), 2004.

Altman 1999

Altman DG, Bland JM. How to randomize. BMJ 1999; 319: 703-704.

Balk 2002

Balk EM, Bonis PAL, Moskowitz H, Schmid CH, Ioannidis JPA, Wang C, Lau J. Correlation of quality measures with estimates of treatment effect in Meta-analyses of randomized controlled trials.JAMA 2002; 287: 2973-2982.

Bassler 2007

Bassler D, Ferreira-Gonzalez I, Briel M, Cook DJ, Devereaux PJ, Heels-Ansdell D, Kirpalani H, Meade MO, Montori VM, Rozenberg A, Schünemann HJ, Guyatt GH. Systematic reviewers neglect bias that results from trials stopped early for benefit. Journal of Clinical Epidemiology 2007; 60: 869-873.

Bellomo 2000

Bellomo R, Chapman M, Finfer S, Hickling K, Myburgh J. Low-dose dopamine in patients with early renal dysfunction: a placebo-controlled randomised trial. Australian and New Zealand Intensive Care Society (ANZICS) Clinical Trials Group. The Lancet 2000; 356: 2139-2143.

Berger 2003

Berger VW, Ivanova A, Knoll MD. Minimizing predictability while retaining balance through the use of less restrictive randomization procedures. Statistics in Medicine 2003; 22: 3017-3028.

Berger 2005

Berger VW. Quantifying the magnitude of baseline covariate imbalances resulting from selection bias in randomized clinical trials. Biometrical Journal 2005; 47: 119-127.

Berlin 1997

Berlin JA. Does blinding of readers affect the results of Meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. The Lancet 1997; 350: 185-186.

Boutron 2005

Boutron I, Estellat C, Ravaud P. A review of blinding in randomized controlled trials found results inconsistent and questionable. Journal of Clinical Epidemiology 2005; 58: 1220-1226.

Boutron 2006

Boutron I, Estellat C, Guittet L, Dechartres A, Sackett DL, Hróbjartsson A, Ravaud P. Methods of blinding in reports of randomized controlled trials assessing pharmacologic treatments: a systematic review. PLoS Medicine 2006; 3: 1931-1939.

Brightling 2000

Brightling CE, Monteiro W, Ward R, Parker D, Morgan MD, Wardlaw AJ, Pavord ID. Sputum eosinophilia and short-term response to prednisolone in chronic obstructive pulmonary disease: a randomised controlled trial. The Lancet 2000; 356: 1480-1485.

Brown 2005

Brown S, Thorpe H, Hawkins K, Brown J. Minimization: reducing predictability for multi-centre trials whilst retaining balance within centre. Statistics in Medicine 2005; 24: 3715-3727.

Chan 2004a

Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004; 291: 2457-2465.

Chan 2004b

Chan AW, Krleža-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. Canadian Medical Association Journal 2004; 171: 735-740.

Chan 2005

Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. BMJ 2005; 330: 753.

Coronary Drug Project Research Group 1980

Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. New England Journal of Medicine 1980; 303:1038-1041.

Cuellar 2000

Cuellar GEM, Ruiz AM, Monsalve MCR, Berber A. Six-month treatment of obesity with sibutramine 15 mg; a double-blind, placebo-controlled monocenter clinical trial in a Hispanic population. Obesity Research 2000; 8: 71-82.

de Gaetano 2001

de Gaetano G. Low-dose aspirin and vitamin E in people at cardiovascular risk: a randomised trial in general practice. Collaborative Group of the Primary Prevention Project. The Lancet 2001; 357: 89-95.

Detsky 1992

Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbé KA. Incorporating variations in the quality of individual randomized trials into Meta-analysis. Journal of Clinical Epidemiology 1992; 45: 255-265.

Devereaux 2001

Devereaux PJ, Manns BJ, Ghali WA, Quan H, Lacchetti C, Montori VM, Bhandari M, Guyatt GH. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA 2001; 285: 2000-2003.

Emerson 1990

Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Controlled Clinical Trials 1990; 11: 339-352.

Fergusson 2002

Fergusson D, Aaron SD, Guyatt G, Hébert P. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. BMJ 2002; 325: 652-654.

Fergusson 2004

Fergusson D, Glass KC, Waring D, Shapiro S. Turning a blind eye: the success of blinding reported in a random sample of randomised, placebo controlled trials. BMJ 2004; 328: 432.

Furukawa 2007

Furukawa TA, Watanabe N, Omori IM, Montori VM, Guyatt GH. Association between unreported outcomes and effect size estimates in Cochrane Meta-analyses. JAMA 2007; 297: 468-470.

Ghersi 2006

Ghersi D, Clarke M, Simes J. Selective reporting of the primary outcomes of clinical trials: a follow-up study. 14th Cochrane Colloquium, Dublin (Ireland), 2006.

Gøtzsche 1996

Gøtzsche PC. Blinding during data analysis and writing of manuscripts. Controlled Clinical Trials 1996; 17: 285-290.

Gøtzsche 2006

Gøtzsche PC, Hróbjartsson A, Johansen HK, Haahr MT, Altman DG, Chan AW. Constraints on publication rights in industry-initiated clinical trials. JAMA 2006; 295: 1645-1646.

Gøtzsche 2007

Gøtzsche PC, Hróbjartsson A, Maric K, Tendal B. Data extraction errors in Meta-analyses that use standardized mean differences. JAMA 2007; 298: 430-437.

Greenland 2001

Greenland S, O'Rourke K. On the bias produced by quality scores in Meta-analysis, and a hierarchical view of proposed solutions. Biostatistics 2001; 2: 463-471.

Haahr 2006

Haahr MT, Hróbjartsson A. Who is blinded in randomised clinical trials? A study of 200 trials and a survey of authors. Clinical Trials 2006; 3: 360-365.

Hahn 2002

Hahn S, Williamson PR, Hutton JL. Investigation of within-study selective reporting in clinical research: follow-up of applications submitted to a local research ethics committee. Journal of Evaluation in Clinical Practice 2002; 8: 353-359.

Hansson 1999

Hansson L, Lindholm LH, Niskanen L, Lanke J, Hedner T, Niklason A, Luomanmaki K, Dahlof B, de Faire U, Morlin C, Karlberg BE, Wester PO, Bjorck JE. Effect of angiotensin-converting-enzyme inhibition compared with conventional therapy on cardiovascular morbidity and mortality in hypertension: the Captopril Prevention Project (CAPPP) randomised trial. The Lancet 1999; 353: 611-616.

Hill 1990

Hill AB. Memories of the British streptomycin trial in tuberculosis: the first randomized clinical trial. Controlled Clinical Trials 1990; 11: 77-79.

Hollis 1999

Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 1999; 319: 670-674.

Hróbjartsson 2007

Hróbjartsson A, Forfang E, Haahr MT, ls-Nielsen B, Brorson S. Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding. International Journal of Epidemiology 2007; 36: 654-663.

Hutton 2000

Hutton JL, Williamson PR. Bias in Meta-analysis due to outcome variable selection within studies. Journal of the Royal Statistical Society Series C 2000; 49: 359-370.

Jadad 1996

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, McQuay H. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Controlled Clinical Trials 1996; 17: 1-12.

Johansen 2000

Johansen HK, Gøtzsche PC. Amphotericin B lipid soluble formulations versus amphotericin B in cancer patients with neutropenia. Cochrane Database of Systematic Reviews 2000, Issue 3. Art No: CD000969.

Jørgensen 2006

Jørgensen KJ, Johansen HK, Gøtzsche PC. Voriconazole versus amphotericin B in cancer patients with neutropenia. Cochrane Database of Systematic Reviews 2006, Issue 1. Art No: CD004707.

Jørgensen 2007

Jørgensen KJ, Johansen HK, Gøtzsche PC. Flaws in design, analysis and interpretation of Pfizer's antifungal trials of voriconazole and uncritical subsequent quotations. Trials 2007; 7: 3.

Jüni 1999

Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for Meta-analysis. JAMA 1999; 282: 1054-1060.

Jüni 2001

Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ 2001; 323: 42-46.

Kjaergard 2001

Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in Meta-analyses. Annals of Internal Medicine 2001; 135: 982-989.

Lachin 2000

Lachin JM. Statistical considerations in the intent-to-treat principle. Controlled Clinical Trials 2000; 21: 167-189.

Marshall 2000

Marshall M, Lockwood A, Bradley C, Adams C, Joy C, Fenton M. Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. British Journal of Psychiatry 2000; 176: 249-252.

Melander 2003

Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(i)ased medicine - selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. BMJ 2003; 326: 1171-1173.

Moher 1995

Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Controlled Clinical Trials 1995; 16: 62-73.

Moher 1996

Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: Current issues and future directions. International Journal of Technology Assessment in Health Care 1996; 12: 195-208.

Moher 1998

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in Meta-analyses? The Lancet 1998; 352: 609-613.

Moher 2001a

Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. The Lancet 2001; 357: 1191-1194.

Moher 2001b

Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. The Lancet 2001; 357: 1191-1194.

Moher 2001c

Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. The Lancet 2001; 357: 1191-1194.

Moher 2001d

Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. The Lancet 2001; 357: 1191-1194.

Montori 2002

Montori VM, Bhandari M, Devereaux PJ, Manns BJ, Ghali WA, Guyatt GH. In the dark: the reporting of blinding status in randomized controlled trials. Journal of Clinical Epidemiology 2002; 55: 787-790.

Montori 2005

Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, Lacchetti C, Leung TW, Darling E, Bryant DM, Bucher HC, Schünemann HJ, Meade MO, Cook DJ, Erwin PJ, Sood A, Sood R, Lo B, Thompson CA, Zhou Q, Mills E, Guyatt GH. Randomized trials stopped early for benefit: a systematic review. JAMA 2005; 294: 2203-2209.

Naylor 1997

Naylor CD. Meta-analysis and the Meta-epidemiology of clinical research. BMJ 1997; 315: 617-619.

Newell 1992

Newell DJ. Intention-to-treat analysis: implications for quantitative and qualitative research. International Journal of Epidemiology 1992; 21: 837-841.

Noseworthy 1994

Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. Neurology 1994; 44: 16-20.

Oxman 1993

Oxman AD, Guyatt GH. The science of reviewing research. Annals of the New York Academy of Sciences 1993; 703: 125-133.

Peto 1999

Peto R. Failure of randomisation by "sealed" envelope. The Lancet 1999; 354: 73.

Pildal 2007

Pildal J, Hróbjartsson A, Jørgensen KJ, Hilden J, Altman DG, Gøtzsche PC. Impact of allocation concealment on conclusions drawn from Meta-analyses of randomized trials. International Journal of Epidemiology 2007; 36: 847-857.

Porta 2007

Porta N, Bonet C, Cobo E. Discordance between reported intention-to-treat and per protocol analyses. Journal of Clinical Epidemiology 2007; 60: 663-669.

Rees 2005

Rees JR, Wade TJ, Levy DA, Colford JM, Jr., Hilton JF. Changes in beliefs identify unblinding in randomized controlled trials: a method to meet CONSORT guidelines. Contemporary Clinical Trials 2005; 26: 25-37.

Sackett 2007

Sackett DL. Commentary: Measuring the success of blinding in RCTs: don't, must, can't or needn't? International Journal of Epidemiology 2007; 36: 664-665.

Safer 2002

Safer DJ. Design and reporting modifications in industry-sponsored comparative psychopharmacology trials. Journal of Nervous and Mental Disease 2002; 190: 583-592.

Schulz 1995a

Schulz KF. Subverting randomization in controlled trials. JAMA 1995; 274: 1456-1458.

Schulz 1995b

Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995; 273: 408-412.

Schulz 1996

Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. BMJ 1996; 312: 742-744.

Schulz 2002a

Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomized trials. Annals of Internal Medicine 2002; 136: 254-259.

Schulz 2002b

Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. The Lancet 2002; 359: 614-618.

Schulz 2002c

Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. The Lancet 2002; 359: 515-519.

Schulz 2002d

Schulz KF, Grimes DA. Unequal group sizes in randomised trials: guarding against guessing. The Lancet 2002; 359: 966-970.

Schulz 2006

Schulz KF, Grimes DA. The Lancet Handbook of Essential Concepts in Clinical Research. Edinburgh (UK): Elsevier, 2006.

Senn 1991

Senn S. Baseline comparisons in randomized clinical trials. Statistics in Medicine 1991; 10: 1157-1159.

Siersma 2007

Siersma V, ls-Nielsen B, Chen W, Hilden J, Gluud LL, Gluud C. Multivariable modelling for Meta-epidemiological assessment of the association between trial quality and treatment effects estimated in randomized clinical trials. Statistics in Medicine 2007; 26: 2745-2758.

Smilde 2001

Smilde TJ, van Wissen S, Wollersheim H, Trip MD, Kastelein JJ, Stalenhoef AF. Effect of aggressive versus conventional lipid lowering on atherosclerosis progression in familial hypercholesterolaemia (ASAP): a prospective, randomised, double-blind trial. The Lancet 2001; 357: 577-581.

Spiegelhalter 2003

Spiegelhalter DJ, Best NG. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. Statistics in Medicine 2003; 22: 3687-3709.

Sterne 2002

Sterne JA, Jüni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'Meta-epidemiological' research. Statistics in Medicine 2002; 21: 1513-1524

Tierney 2005

Tierney JF, Stewart LA. Investigating patient exclusion bias in Meta-analysis. International Journal of Epidemiology 2005; 34: 79-87.

Turner 2008

Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. Journal of the Royal Statistical Society Series A (in press, 2008).

Unnebrink 2001

Unnebrink K, Windeler J. Intention-to-treat: methods for dealing with missing values in clinical trials of progressively deteriorating diseases. Statistics in Medicine 2001; 20: 3931-3946.

Vickers 2001

Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Medical Research Methodology 2001; 1: 6.

von Elm 2006

von Elm E, Röllin A, Blümle A, Senessie C, Low N, Egger M. Selective reporting of outcomes of drug trials. Comparison of study protocols and pulbished articles. 14th Cochrane Colloquium, Dublin (Ireland), 2006.

Williamson 2005a

Williamson PR, Gamble C. Identification and impact of outcome selection bias in Meta-analysis. Statistics in Medicine 2005; 24: 1547-1561.

Williamson 2005b

Williamson PR, Gamble C, Altman DG, Hutton JL. Outcome selection bias in Meta-analysis. Statistical Methods in Medical Research 2005; 14: 515-524.

Wood 2004

Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clinical Trials 2004; 1: 368-376.

Wood 2008

Wood L, Egger M, Gluud LL, Schulz K, Jüni P, Altman DG, Gluud C, Martin RM, Wood AJG, Sterne JAC. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. BMJ 2008; 336: 601-605.

Woods 1995

Woods KL. Mega-trials and management of acute myocardial infarction. The Lancet 1995; 346: 611-614.

（李雨璘、崇乐、李江、齐国卿、柯法勇、林晓亭、柳文杰译, 马彬、岑啸、 贾鹏丽、秦天强初审）

第九章　数据分析和 Meta 分析

编者：代表 Cochrane 统计方法学组的 Jonathan J Deeks, Julian PT Higgins 和 Douglas G Altman。

本文选自工作手册5.1.0版本。有关如何引用它的指南，见9.8节。这些材料还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- Meta 分析是对两个或多个独立研究结果的统计学合并。
- Meta 分析的潜在优势包括增加检验效能、提高准确性、回答单个研究无法回答的问题和解决相互矛盾的观点引发的争论。然而，Meta-分析也会带来潜在的严重误导，尤其是未仔细考虑特定的研究设计、研究本身的偏倚、研究间的变异和报告偏倚情况下。
- 熟悉测量单个研究结局指标的数据类型（如分类变量、连续性变量），并选择恰当的效应指标比较干预组间差异非常重要。
- 多数 Meta 分析方法因不同研究间的加权平均效应值而变化。
- 研究间的变异（异质性）必须加以考虑，虽然多数 Cochrane 系统评价并未纳入足够

的研究以分析产生异质性的确切原因。随机效应模型 Meta 分析假设潜在效应值服从正态分布而容许异质性的存在。

- 在制作 Cochrane 系统评价或 Meta 分析的过程中需进行许多判断。敏感性分析用于检测汇总结果对潜在影响的决策是否仍稳定。

9.1 引言

9.1.1 请勿从此处开始

当进行系统评价时，容易过早地进入统计分析。对许多作者而言，得到森林图底部的菱形结果是一个令人兴奋的时刻，但如果对提出系统评价问题，确定纳入标准，查找、筛选和严格评价研究，收集恰当的数据，并决定什么分析有意义，未给予适当的关注，则Meta分析结果很可能会产生误导。系统评价者应在进行Meta分析前先学习相关章节。

9.1.2 制定分析计划

在原始研究中，研究者从个体患者筛选和收集数据；在系统评价中，研究者则是从原始研究中筛选和收集数据。原始研究是对受试者的分析，而系统评价是对原始研究的分析。分析可能为描述性的（如对研究特征和结果的结构式小结和讨论），也可能为定量的（包含统计分析）。Meta分析——对来自两个或多个独立研究结果的统计学合并——是最常使用的统计学方法。Cochrane系统评价写作软件（RevMan）可以进行多种Meta分析，但必须强调的是，Meta分析并非对所有Cochrane系统评价都恰当。本节和9.1.4节将讨论决定系统评价中Meta分析是否恰当的问题。

比较卫生保健干预措施的研究，尤其是随机对照试验，是采用受试者结局以比较不同干预措施的效果。Meta分析是用于两组干预措施的比较，如试验组与对照组干预措施相比，或两个试验措施间的比较。此处所用的术语（试验组和对照组干预措施比较）是指前者，虽然这一方法同样适用于后者。

给予不同处理的两组的结局差异被作为“效果”、“治疗效果”或“干预效果”。对纳入研究的分析是定性的还是定量的，综合研究的一般性框架可考虑以下四个问题：

1. 效应的方向是什么？
2. 效应的大小是什么？

3. 研究间效应一致吗？

4. 效应证据强度是什么？

Meta分析对问题1和3提供了统计分析。问题4的评价还有赖于对研究设计和偏倚风险及对不确定性的统计测量的判断。

对于Meta分析既不可行也不合理的系统评价，定性合成使用主观的（而非统计学）方法以回答问题1-4。在定性合成中，每一阶段使用的方法应该预先确定，并合理和系统地实施。如果过分强调某一个研究的结果，则可能带来偏倚。

分析计划遵从系统评价的科学目的。系统评价有不同类型的目的，可能因此需要不同的分析方法。

1. 最简单的Cochrane系统评价收集比较两种治疗措施的研究，如比较卡法根提取物与安慰剂治疗焦虑（Pittler 2003）。如有一致的结局指标，则Meta分析和相关技术可用于：
   - 建立是否有效的证据；
   - 评估效应量及其不确定性；
   - 分析研究间效应是否一致
2. 一些系统评价关注的内容可能更广，而不仅是单一比较。其目的首先是确定和比较针对同一疾病和状况的多种干预。一个例子是足部皮肤和指甲真菌感染的局部治疗的系统评价，其纳入了任何局部治疗的研究（Crawford 2007）。其次，相关目的是找到“最佳”干预。一篇紧急避孕措施的系统评价寻找哪种干预措施是最有效的（同时也考虑潜在的不良反应）。这样的系统评价可能包括所有可能的成对治疗措施间的多个比较和Meta分析，需要在设计统计分析时多加注意（见9.1.6节和16章16.6节）。
3. 有时候，系统评价的范围特别宽泛，使Meta分析的应用存在问题。如，在工作场所干预的禁烟方法的系统评价涵盖了各种类型的干预措施（Moher 2005）。当系统评价包含了非常广泛的研究时，Meta分析可用于回答是否有证据显示，如基于工作场所的干预有效（见9.1.4节）。但如果实施环境差异过大以致效应估计值在任何特定的环境下都不能被解释，则采用Meta分析描述效应量可能就没有意义。
4. 某些系统评价的目的之一是了解效应量大小与研究的某些特征间的关系。在Cochrane系统评价中，很少将其作为主要目的，但可作为次要目的。如，一篇比较倍氯米松与安慰剂治疗慢性哮喘的系统评价，其研究兴趣在于倍氯米松的

给药剂量是否会影响其效果（Adams 2005）。这样的异质性分析需要谨慎进行（见9.6节）。

9.1.3 为什么在系统评价中做Meta分析？

系统评价中Meta分析的增值作用要看其在什么情况下使用，见9.1.2节。在系统评价中考虑采用Meta分析的原因如下：

1. 增加检验效能。检验效能是检出真实存在的效应有统计学意义的概率。很多单个研究因样本量太小而不能检出很小的效应，但当几个研究合并后则有较大机会检出这种效应。
2. 提高精确性。当对一种干预效果的估计基于更多信息时，其估计精确性可得到提高。
3. 回答单个研究不能回答的问题。原始研究常常纳入特定类型的患者和明确定义的干预措施。选择上述特征不同的研究可以评估效应的一致性，如果必要，还可评估效应差异的原因。
4. 解决明显矛盾的研究引起的争论或产生新的假设。结果的统计分析允许正式评估矛盾的程度，并且对结果不一致的原因进行探索和定量。

当然，统计方法的使用并不能保证系统评价结果的真实性，其所能起到的作用不可能比在原始研究中更多。而且，同任何工具一样，统计方法可能被滥用。

9.1.4 何时在系统评价中不使用Meta分析

如果使用恰当，Meta分析对于从数据中得出有意义的结论是一个强大的工具，并且有助于在解释中避免错误。然而，有一些情况下Meta分析可能弊大于利。

- 对Meta分析通常的批评是他们“将苹果和橘子合并在一起”。如果各研究间有临床差异，Meta分析则可能无意义，且可能掩盖真正的效果差异。特别重要的一种差异就是原始研究中进行的干预措施比较。通常在单一Meta分析中合并所有纳入研究并无意义：有时是不同治疗措施与对照措施间的混合比较，需要分别考虑每一个比较的合并。而且，很重要的是不要合并差异太大的结局。对什么应该合并和什么不应该合并的决策不可避免地带有主观性，不能依从于统计学方法而需要讨论和临床判断。有些情况很难达成一致性。

- 对存在偏倚风险的研究进行Meta分析可能产生严重误导。如果每一个（或某些）研究存在偏倚，Meta分析就会简单地综合这些错误，并产生“错误”结果，但可能解释成更可靠。
- 最后，如果存在严重发表和/或报告偏倚，Meta分析可能产生不恰当的合成。

9.1.5 系统评价承担了什么？

虽然系统评价中使用统计方法非常有帮助，但分析的最基本要素是一个深思熟虑的过程，包括定性和定量的要素。这包括对以下问题的考虑：

1. 应做哪些比较？
2. 在每个比较中应使用哪些研究结果？
3. 对每个比较最佳效应量是什么？
4. 在每个比较中研究结果相似吗？
5. 效应量的可靠度如何？

解决这些问题的第一步是决定要做哪些比较（见9.1.6节）和对于感兴趣的结局哪类数据是恰当的（见9.2节）。下一步是采用表格形式总结纳入每个比较的研究特征和结果（提取数据并转换成目标格式见第7章7.7节）。然后才可能采用系统的方式合成研究间的效应值（9.4节），以测量和分析研究间的差异（9.5和9.6节）和解释发现并对其可信度做出结论（见12章）。

9.1.6 应做哪些比较

制定分析计划的第一步也是最重要的一步是确定将要做的配对比较。在规划系统评价时，系统评价中所确定的比较应与提出的研究问题或假设明确且直接相关（见第5章）。在完成系统评价计划书时就应确定要进行的主要比较。然而，根据数据收集的情况调整和新增一些比较常常很有必要。如，干预措施的重要变更可能仅在数据收集后才能被发现。

确定哪些研究有足够的相似性因而其结果可以合并在一起，需要考虑系统评价所要解决的问题，以及作者和用户的判断。有关系统评价问题的构建已在第5章讨论。同样需要考虑的是决定做哪些比较、合并哪些结局和分析效应值变异（异质性）时考虑哪些关键特征（研究设计、受试者、干预措施和结局）。当在RevMan中建立“数据和分析”表以及在决定什么信息需要放进“纳入研究特征”表时，必须要考虑上述问题。

9.1.7 撰写计划书的分析部分

Cochrane系统评价计划书的分析部分比其他部分更易改变（如研究纳入标准和如何评价方法学质量）。要事先估计到所有可能出现的统计学问题几乎不可能，如，发现结局彼此间相似但不一样；结局在多个或不同的时点测量；联合治疗的使用。

然而，研究方案应明确表示作者如何进行研究结果的统计学评估。但在撰写计划书时，系统评价小组至少有一名成员应熟悉本章的主要内容。作为指南，我们推荐强调如下内容：

1. 确保分析策略明确地解决系统评价确定的目的（见9.1.2节）。
2. 思考哪类研究设计适合于系统评价。常规的是平行组试验，但其他随机设计也可能适合该系统评价（如交叉试验、整群随机试验、析因试验）。决定如何在分析中处理这类研究见9.3节。
3. 确定是否设计Meta分析并考虑如何决定Meta分析是否恰当（见9.1.3节和9.1.4节）。
4. 确定结局指标的可能类型（如分类变量、连续变量等）（见9.2节）。
5. 考虑是否可能事先确定要采用的干预措施效应指标（如二分类变量：危险比、比值比和危险差（risk difference，RD）；连续性变量：均差或标准差）（见9.4.4.4和9.4.5.1）
6. 确定如何鉴定或量化统计学异质性（见9.5.2节）。
7. 确定随机效应模型Meta分析、固定效应模型Meta分析或两种方法是否都用于每一计划进行的Meta分析（见9.5.4节）。
8. 思考怎样评估临床和方法学差异（异质性）和是否（和怎样）将其整合进分析策略（见9.5和9.6节）
9. 确定怎样评估纳入研究的偏倚风险并在分析中强调（见第8章）。
10. 事先确定可能引起异质性的研究特征（见9.6.5节）。
11. 思考如何处理缺失数据（如意向性分析赋值）（见第16章16.1和16.2节）。
12. 决定是否（和怎样）寻找可能存在发表偏倚和/或报告偏倚的证据（见第10章）。

显然，在撰写计划书时可能需要其他专业人员；如果是这样的话，系统评价小组中应有统计专家的参与。

9.2 数据类型和效应值测量

9.2.1 数据类型

所有有效性研究的Meta分析首先要明确结局指标的数据类型。本章中，结局数据包括5种：

1. 二分类（或二元）数据，每一个体的结局为两种可能答案中的一种；
2. 连续性数据，每一个体结局为一个数量测量值；
3. 有序数据（包括量表），其结局为几个有序分类中的一个，或通过打分和合计分类回答而产生；
4. 计数和率，其通过计算每一个体发生事件数获得；
5. 时间-事件（以生存为代表）数据，其分析事件发生的时间，但并非研究中的所有个体均会出现事件（截尾数据）

干预措施效果的测量方式有赖于所收集数据的特征。本节我们简要分析了有关临床试验的系统评价中可能遇到的结局指标类型，和干预措施效应指标的标准测量方法的系统评价定义、性质和解释。在9.4.4.4和9.4.5.1节，我们将讨论选择这些指标用于特定Meta分析的问题。

9.2.2 二分类结局指标的效应值

二分类结局数据是指每个受试者的结局为两个可能性中一个，如死亡或生存，或有临床改善和无临床改善。本节考虑当分析的结局为二分类形式时可能的汇总统计量。临床试验中最常遇到的二分类数据效应指标为：

- 风险比（RR）（也称相对危险度）；
- 比值比（OR）
- 危险度差值（RD）（也称绝对危险度降低率）；
- 需治疗的例数（NNT）

前三类指标的计算详见框9.2.a。对NNT的讨论详见第12章（12.5节）。

另：因为事件有时是所期望的而非不期望的，因而倾向于使用更中性的术语而非风险（如概率），但由于约定俗成的缘故，我们在此处使用术语风险比和风险差。考虑到与其他术语间的一致性，我们也使用术语“风险比”来代替“相对危险度”。两者可互换，

并都可缩写为“RR”。也应注意：在使用单词“危险”和“率”时我们应当心。这两个词常常被当作同一意思。然而，对数据类型“计数和率”，我们试图保留使用单词“率”，其描述了在所测时间范围内的事件频率。

框9.2.a　采用2×2四格表计算风险比（RR）、比值比（OR）和危险差值（RD）

临床试验的结果可用 2×2 四格表显示：			
	有事件 （“成功”）	无事件 （“失败”）	合计
试验组干预措施	S_E	F_E	N_E
对照组干预措施	S_C	F_C	N_C

此处，S_E、S_C、F_E 和 F_C 为每一组（“E”或“C”）每一结局（“S”或“F”）的受试者数，可用于计算如下合并统计量：

RR=试验组事件风险/对照组事件风险= $(S_E/N_E)/(S_C/N_C)$

OR=试验组事件比值/对照组事件比值= $(S_E/F_E)/(S_C/F_C)=S_EF_C/F_ES_C$

RD=试验组事件风险-对照组事件风险= $S_E/N_E-S_C/N_C$

9.2.1.1　风险和比值

通常，在描述同样的数量时，术语“风险”和“比值”互换使用（这与术语“机遇”、“概率”和“似然”类似）。然而，在统计学上，风险和比值有特定意义，并按不同方式计算。如果忽略他们之间的差异，系统评价结果可能被错误解释。

风险是患者和卫生专业人员更熟悉的概念，风险描述了卫生结局（通常为不良事件）发生的概率。研究中，风险常表述为0至1之间的某一数值，但偶尔被转换成百分率。在Cochrane系统评价的“结果汇总表”中，其常被表述为在每1000个体中的数量（见11章11.5节）。掌握风险和可能事件发生之间的相关性很简单：在100人的样本中所观察到的事件数为平均风险×100。如，当风险为0.1时，表示每100人中将有约10人发生事件；当风险为0.5时，表示每100人中将有约50人发生事件。在1000人的样本中，这些数字将分别为100和500。

比值是赌博者更熟悉的概念。比值是某一特定事件发生概率和不发生概率的比值，并且可能是0和无穷大之间的任一数值。在赌博中，比值描述了潜在赢得赌金的可能性大小之比；在卫生保健中，其为发生某事件和未发生某事件人数之比。其通常表述为两

个整数之比。如，比值为0.01常被写为1:100，0.33的比值为1:3，3的比值为3:1。使用如下公式，比值可转换为风险，风险可转换为比值：

风险=比值/（1+比值）

比值=风险/（1-风险）

比值的解释比风险更复杂。确保解释正确的最简单方法为首先将比值转换成风险。如，当比值为1:10或0.1时，与每10个未发生事件的人相对都有1人将发生事件，公式显示事件风险为0.1（1+0.1）=0.091。在100人的样本中，有9人将发生事件，91人将不会发生事件。当比值等于1时，与每个未发生事件的人相对都有1人发生事件，因而在100人的样本中，有100 ×1/(1+1) = 50人将会发生事件，50人将不会发生事件。

对罕见事件，比值和风险间的差异很小（如对上述第1个例子的解释一样，0.091的风险与0.1的比值差不多）。对常见事件，如在临床试验中的常见情况，比值和危险差异巨大。例如，0.5的风险与1的比值相等；0.95的风险与19的比值相等。

临床试验二分类结局效应指标包括对两干预组风险或比值的比较。在对这些指标进行比较时，我们可以看他们的比值（风险比或比值比）或其风险的差（危险差）。

9.2.2.2 相对效应指标：风险比和比值比

相对效应指标表述了某一组相对与另一组的结局。风险比（或相对风险）是在两组中某一事件风险的比，而比值比是某一事件比值的比（见框9.2.a）。对该两个指标，值为1表明估计效应在两干预间相同。

如一个研究的对照组无事件发生，则不能计算RR和OR。这是因为，如框9.2.a中的公式所见，我们将会除以0。如果干预组每一个体都发生了事件，也不能计算OR。在这些情况以及不能计算标准误的情况下，通常对2×2表格中的每一格都增加1/2（在必要时RevMan会自动做出校正）。在两组无事件（或全事件）被观察到的情况下，研究不会提供关于事件相对概率的信息，并将从Meta分析中自动忽略。这是完全恰当的。当所关注的事件为罕见事件时，0尤其容易出现——如事件为通常不希望的不良结局。对这类稀疏数据（常有许多0）效应指标选择的进一步讨论见16章（16.9节）。

风险比描述了随试验干预使用而产生的风险增加。例如，治疗的风险比为3，意为治疗带来的事件是未治疗事件的3倍。或者说，治疗增加了100× (RR – 1)% = 200%的事件风险。相似地，风险比为0.25解释为治疗带来的事件概率是未治疗的1/4。其也可以表述为治疗减少了100 ×(1 –RR)% = 75%的事件风险。这被称之为相对危险减少（也见12

章12.5.1节）。在不知道无治疗事件风险的情况下，不能做出对所给风险比的临床重要性的解释：风险比为0.75可能反映了从80%到60%的有临床重要意义的事件减少，或4%到3%的有小的、较少临床重要意义的事件减少。

所观察到风险比的数值一定是介于0至1/CGR之间，CGR（control group risk，对照组风险，有时指对照事件率）是对照组观察到的事件风险（表述为0至1之间的一个数）。这意味着，对通常的事件，不可能得到较大的风险比值。例如，当对照观察到的事件风险为0.66（或66%）时，所观察到的风险比不可能超过1.5。这一问题仅适于风险的增加，和仅当结果被外推至超过在研究中所观察到的风险时会带来问题。

和比值一样，比值比更难于解释（Sinclair 1994，Sackett 1996）。比值比描述了随干预使用产生结局比值的增加。为理解从事件数改变来看比值比意味着什么，最简单的是首先将其转换为风险比，然后按上述基于对照组风险解释风险比。将比值比转换为风险比的公式见12章（12.5.4.4节）。有时，与对照组风险相比超过1时，计算RR更恰当。

9.2.2.3 注意：OR 和 RR 不一样

因为对常见事件而言，风险和比值是不同的，因而风险比和比值比也不同。风险比和比值比的不等价并非是说哪一个不对：他们都是描述干预效果的完全正确的方式。然而，如果将比值比错误地解释为风险比，问题可能随之出现。对增加事件机会的干预而言，比值比将比风险比更大，因此，错误解释将导致过高估计干预效应，尤其是对（事件风险超过20%的）常见事件。对减少事件机会的干预而言，比值比将小于风险比，因而错误解释将高估干预效应。不幸的是，这一错误解释在发表的单个研究和系统评价报告中非常普遍。

9.2.2.4 绝对效应的测量：危险度差值

风险差值是两组间观察到的风险（发生所观察结局的个体的构成比）的差（见框9.2.a）。任何研究甚至当各组均无事件发生时，都可计算风险差。风险差是直接解释：其描述了试验组和对照组干预措施间所观察到事件风险的确切差异。对个体而言，其描述了可能发生事件的估计差异。然而，风险差值的临床意义可能有赖于事件潜在的风险。例如，风险差值为0.02（2%）可表示小的无临床意义的58%到60%风险改变；或相对更大的1%到3%有潜在意义的改变。虽然风险差值比相对指标提供了更直接的相关信息（Laupacis 1988，Sackett 1997），当解释风险差值时，了解事件的潜在风险和事件结果

仍然很重要。绝对指标如风险差，对于权衡一种干预可能带来的获益和危害上尤其有用。

和风险比一样，风险差值有其天生的不足，当将结果应用于其他患者组和环境时，其可能带来困难。例如，如果一个研究或Meta分析估计风险差值为-0.1（或-10%），对一个起始风险为7%的研究组而言，其结局将得到不可能的-3%的负概率。风险增加时也会出现类似的情况。仅当结果被用于在研究中观察到的不同风险的患者时，这些问题才会出现。

NTT从风险差值获得。虽然NNTs常用于总结临床试验的结果，但其不能在Meta分析中进行合并（见9.4.4.4节）。然而，比值比、风险比和风险差可有效地转换成NNTs并在12章（12.5节）所讨论的解释Meta分析结果时使用。

9.2.2.5 什么是事件？

在二分类结局的情况下，卫生保健干预措施要么减少不良结局产生的风险，要么增加有利结局出现的机会。9.2.2节描述的所有效应指标均可应用于这两种情况。

在许多情况下，我们很自然地将一个结局状况作为一个事件。例如，当受试者在研究开始时有特定症状时，关注的事件通常是恢复或治愈。如果受试者在研究开始时是健康的或有某些不良结局的风险，则事件是疾病发生或不良结局出现。因为关注点通常放在试验干预组上，对于试验干预减少不良结局出现的研究将得到小于1的比值比或风险比，以及负的风险差。试验干预增加有利结局出现的研究将得到大于1的比值比或风险比，以及正的风险差（见框9.2.a）。

然而，可以用发生事件和未发生事件来取代未恢复患者和未发生事件患者的比例。对使用风险差或比值比的Meta分析，该转换的影响并不会带来大的后果：转换只是改变了风险差的正负号，同时对比值比而言，新的比值比是原比值比的倒数（1/x）。

相反，对结果的转换可能导致风险比的明显差异，影响效应估计、其显著性和干预间研究效应的一致性。这是因为在低风险和高风险状况下，风险比估计的准确性显著不同。在Meta分析中，很难预测这种颠倒的效应。因此在数据分析前，对哪个风险比可能是最相关的统计量的判断很重要，并将在9.4.4.4节进一步讨论。

9.2.3 连续结局效应指标

通常，术语“连续”在统计学中是指在特定范围内可取任何值的数据。当处理数值

资料时，这意味着任何数值都可以被测量和报告为任意小数位。真实的连续性资料的例子包括重量、面积和体积。实践中，在Cochrane系统评价中对其他类型的数据、最常见的测量量表和大量事件的计数，我们能使用同样的统计方法（见9.2.4节）。

常用于连续性资料Meta分析的合并统计量有两个：均数差（mean difference，MD）和标准化均数差（standardized mean difference，SMD）。无论来自每个个体的数据是单一评估或相对基线测量值的改变，均可被计算。通过获得均数比、或通过比较除均数外的其他统计量（如中位数）来对效应进行测量也是可能的。但在此处不讨论这些方法。

9.2.3.1 均数差（或均数的差）

均数差（更准确的应该是“均数的差”）是测量临床试验中两组间均值绝对差值的统计量。其评估了试验干预对结局的平均改变相对于对照的数量。当所有研究的结局测量值是基于同样的度量单位得到时，其可用作为Meta分析的汇总统计量。

附注：基于该效应指标的分析过去在Cochrane系统评价数据库中被称为加权均数差（weighted mean difference，WMD）分析。该名称可能引起混淆：虽然Meta分析计算了这些均数差的权重平均值，但单个研究的统计汇总计算中并未引入加权。而且，所有Meta分析都包含了对估计值的加权合并，而对于其他方法我们并未使用“加权”一词。

9.2.3.2 标准化均数差

在Meta分析中，当所有研究都评估了同样的结局但按不同的方法进行测量（如，所有研究都测量了抑郁，但使用了不同的心理测验量表）时，使用标准均数差作为汇总统计量。这种情况下，在研究合并之前，对研究结果进行标化以达到统一度量单位（scale）很有必要。标准均数差表达了每个研究中与观察到的变异相关的干预效应的大小。（此外，真实情况下，干预效应是均数的差而非差的均数。）：

SMD=组间结局的均数差/受试者结局的标准差

因此，无论用于测量的实际度量单位是多少，均数差与标准差比例相同的研究将有同样的标准化均数差。

然而，该方法假设研究间标准差的差异反应了测量尺度的差异而非研究人群变异的真实差异。当我们希望了解不同研究受试者间变异的真实差异时，该假设可能就会存在问题。例如，在系统评价中合并实效试验和探索性试验时，实效试验可能纳入更大范围的受试者并且可能得到更高的标准差。在系统评价中，因为总干预效应是以标准差的单

位进行报告，而不是以所用任何测量尺度的单位进行报告，因此总干预也很难解释，但有些情况下，可将效应转换回特定研究所用的单位（见12章12.6节）。

术语“效应量”常用于社会科学，尤其在Meta分析中。效应量（虽然并不总是）是标准化均数差的另一典型说法。推荐在Cochrane系统评价中使用术语“标准化均数差”代替“效应量”以避免出现后者作为“干预效应”或“效应估计”的同义词所具备的更广泛的医学用途而带来的混淆。Cochrane系统评价所用标准化均数差的特定定义是在社会科学被称为Hedges’校正的g效应量。

应该注意到，SMD方法对于度量单位方向并不正确。如果一些度量单位随严重程度增加而另一些减少，那么需要对一系列研究乘以-1（或可从度量单位最大可能值减去均数）以确保所有度量单位点都在同一的方向。任何这样的调整都应在系统评价的统计方法部分进行描述。标准差无需修正。

9.2.4 有序结局和量表的效应指标

当每个受试者被分入一类且分类有自然顺序时，就会出现有序结局。例如，按分类排序的“三分类”结局（如疾病严重程度分为“轻”、“中”和“重”）就是有序类型。随分类数的增加，有序结局会表现出与连续性结局相似的特征，并且在临床试验中可能会按连续性结局进行分析。

量表是一种特殊类型的有序结局，常用于测量难于定量的情况，如行为、抑郁和认知能力。典型的量表包括一系列问题或任务，对每一项进行打分，然后求和得到总“积分”。如果认为条目间重要性不同，可使用加权求和。

明确量表是否有效很重要：也就是说，证明他们测量了他们要测量的情况。当在临床试验使用量表评价结局时，为理解研究的目的、目标人群和评估问卷，应对量表所引参考文献进行研究。研究者常常通过增添、修改或去掉某些问题来改编量表以满足他们自己的目的，系统评价作者应检查使用的是原始还是改编的量表。这对于Meta分析结局合并时尤其重要。临床试验可能使用了看似相同的评分量表，但仔细检查时可能发现他们之间存在必须考虑的差异。为了显示出从试验干预中获益的部分，研究者可能根据研究结果对量表进行修改。

对于以成比例比值比（proportional odds ratio）来表示其效应的有序结局资料，可用高级方法进行分析，但RevMan未提供这些方法，同时，当分类数很大时，这些方法则

变得用处不大（或不必要）。事实上，较长的有序量表在Meta分析中常被作为连续性变量分析，而较短的有序量表常通过将相邻分类合并在一起而转换成二分类资料。如果有明确合理的分界点，后者尤其适用。不恰当地选择分界点可能导致偏倚，尤其是选择的分界点使临床试验中两干预组的差异最大化时。

当使用处理二分类资料的方法来总结有序量表时，所分两类中的一类被定义为事件，且干预效应使用风险比、比值比或风险差进行描述（见9.2.2节）。当使用处理连续性资料的方法来总结有序量表时，干预效应被表述为均数差或标准化均数差（见9.2.3节）。如果原始研究已经使用了中位数来总结其结果，则可能遇到困难（见第7章，7.7.3.5节）。

除非可获得单个患者的资料，否则临床试验中研究者所报告的分析方法决定了Meta分析中所使用的方法。

9.2.5 计数和率的效应指标

某些类型的事件在一个人身上可能发生超过一次，例如，心肌梗死、骨折、不良反应或住院。明确这些事件发生的次数而不仅仅是每个人经历了任何事件（换句话说，而不是以二分类资料来处理他们）可能有必要或更好。我们把这种类型的资料归为计数资料。根据应用目的，计数资料可分为罕见事件计数和常见事件计数。

罕见事件计数在统计学中常被归为“Poisson资料”。罕见事件的分析常关注率。率表明了事件数与其发生的时间时长的关系。例如，临床试验中一组结果可能是在随访的314人年期间在该组所有患者发生了18次心肌梗死。事件发生率是0.057/人年或5.7/100人年。在Meta分析中通常使用的统计量是率比（rate ratio，RR），其通过一组除以另一组来比较两组中事件的发生率。也可使用率差作为统计量，但很少使用。

更常见事件的计数，如龋坏的、缺失的或充填的牙的计数，常以与连续性结局资料同样的方式进行处理。所用干预效应为均数差，其比较干预组受试者与对照组受试者发生事件平均数（可能被标化到某一单位事件范围）的差异。

9.2.5.1 注意：计数事件还是计数受试者？

一个常见错误是将计数资料作为二分类资料进行处理。如在刚才所举的例子中，314人年来自对157例患者平均2年的观察。一个可能得到的结果是18/157。如果总数18中包括来自同一患者的多次心肌梗死（比如说，如果18来自12例发生1次心肌梗死的患者和3

例发生2次心肌梗死的患者），这就不恰当了。事件总数在理论上可能超过患者数，使结果不可理解。例如，在过去的1年期间，1个研究中35例癫痫患者可能有63次发作。

9.2.6 时间-事件（生存）结局的效应指标

当关注点落在某一事件发生前所经历的时间时，就会出现时间-事件资料。因为关注的事件常常是死亡，尤其在癌症和心脏疾病，因此在统计学中他们一般被称为生存资料。时间-事件资料包括对每一个体两方面的观察：（1）未观察到事件的时间长度；（2）发生事件或观察期结束的终点指针。在观察期结束也未发生事件的受试者称为“截尾值（censored）”。他们的无事件时间也有意义，并应纳入Meta分析中。时间-事件资料也可基于除死亡外的事件，如疾病复发（如，无癫痫发作时间）或出院。

时间-事件资料有时可被作为二分类资料分析。这需要知道在一研究中一个固定时点所有受试者的状况。例如，如果所有患者被随访至少12个月，且知道两组12个月前发生事件的比例，则可构建2×2表格（见框9.2.a），并以风险比、比值比或风险差来表述干预效应。

使用处理连续性结局的方法来分析时间-事件资料并不恰当（如使用平均时间-事件），因为只能知道发生了事件的受试者亚组的相关时间。使用这种方法必须排除未发生事件的受试者，这样会引起偏倚。

综合时间-事件资料最恰当的方法是使用生存分析的方法，并以危险比（hazard ratio）来表述干预效应。危险在概念上与风险相似，但也有细微差别，其测量的是瞬时风险，并会不断变化（如，当您穿越一条繁忙道路时，你的死亡危险会改变）。危险比解释方式与风险比相似，因为其描述了一个受试者接受试验干预比对照干预更可能或更不可能在某一特定时点上发生事件的倍数。当在一个研究或Meta分析中对干预措施进行比较时，常简单假设危险比是贯穿整个随访期的常数，即使危险自身可能不断变化。这被称为比例风险假设（proportional hazards assumption）。

9.2.7 以对数形式表述干预效应

干预效应的比值（如比值比、风险比、率比和危险比）在分析前常进行对数变换，并且他们可能偶尔以其对数变换值的方式被提及。有代表性的是使用自然对数变换（对数底为e，写成“ln”）。

比值的汇总统计量共有的特征是其可取的最小值是0，1相当于无干预效应，比值比的最大值可取无穷大。这一记数度量单位并不均称。例如，比值比为0.5（减半）和比值比为2（加倍）正好相对，就是说他们平均后应该是无效应，但0.5和2的平均并不是比值比为1而是1.25。对数转换的度量单位均称：0取对数为负无穷，1取对数为0，正无穷取对数为正无穷。例如，0.5的OR取对数为-0.69，2的OR取对数为0.69。-0.69和0.69的均数是OR为1的对数转换值0，正确地显示了无平均干预效应。

按比例度量单位所做Meta分析的图形显示通常使用对数度量单位。因为同样的原因，其有使可信区间看起来对称的作用。

9.3 研究设计和确定分析单元

9.3.1 分析单元问题

临床试验的重要原则是分析必须考虑随机实施的水平。大多数情况下，分析中的观察数应与随机的单元数相匹配。在一个简单平行组设计的临床试验中，受试者被独立地随机分配到两个干预组中，并且收集和分析每个受试者的每个结局的测量值。然而这一设计上可有许多变动，作者应该考虑是否在每个研究中：

- 将成组个体一起随机分配到同一干预（如整群随机试验）；
- 个体接受不止一种干预(如在交叉试验中,或每个个体同时进行的多部位治疗)；
- 对同一结局的多重观察（如，重复测量、复发事件、不同身体部分的测量）。

下面将详细介绍单元分析问题常见情况及其讨论。

9.3.2 整群随机试验

在整群随机试验中，成组的受试者被随机分配到不同的干预措施。例如，组可以是学校、村庄、医疗单位、由一个医生看的患者、家庭。见16章（16.3节）。

9.3.3 交叉试验

在交叉试验中，所有受试者顺次接受所有干预：他们被随机分配到一个干预序列，并且所有受试者作他们自己的对照。见16章（16.4节）。

9.3.4 受试者的重复观察

在长疗程研究中，结果可能按几个随访期呈现（如，6个月、1年和2年）。在不产生单元分析误差的情况下，来自每个研究不止一个时点的结果不能以标准Meta分析的形式进行合并。其可选择以下一些方式进行处理。

- 获得个体患者数据，并用每个患者的整个随访期进行分析（如时间-事件分析）。或者，对每个个体受试者结合所有时点计算一个效应测量值，如总事件数、总体均数或随时间变化的趋势。偶尔在发表的报告中可见到这种分析。
- 基于不同的随访时段定义几个不同的结局，并分别进行分析。例如，不同时限可用于反映短期、中期和长期随访。
- 选择一个时点并分析研究中该时点的数据。理想情况下，其应为临床重要时点。有时其可能是从可得数据中选择的最大值，即使作者知道存在报告偏倚的可能性。
- 从每个研究选择最长随访时间。这可能导致研究间缺乏一致性，增加异质性。

9.3.5 可重复发生的事件

如果关注的结局是可能发生不止一次的事件，则必须注意避免单元分析误差。计数资料不应按二分类资料处理。见9.2.5节。

9.3.6 多次治疗

相似地，每个受试者接受多次治疗可引起分析单元误差。必须注意确保可信区间的计算使用随机分配的受试者数，而不是治疗尝试数。例如，有关低生育力的研究中，妇女可能接受多周期治疗，作者可能错误地使用治疗周期而不是妇女作为分母。这与整群试验（除每个受试者是一个“群”外）中的情况相似。见16章描述的方法（16.3节）。

9.3.7 身体多个部分（一）：身体多个部分接受相同的干预

受试者被随机分配，但身体的多个部分（或部位）接受相同的干预，对每个身体部分独立进行结局判断，且身体部分数在分析中被作为分母。例如，可能未在左、右眼之间进行非独立性调整而错误地将眼作为分母。这与整群试验（除外将多个受试者作为

“群”）中的情况相似。见16章描述的方法（16.3节）。

9.3.8 身体多个部分（二）：身体多个部分接受不同的干预

不同情况是将身体的不同部分随机分配到不同的干预。口腔卫生保健中的“分口”设计属于这种情况，在该设计下不同的口腔区域被分配接受不同的干预。这类试验与交叉试验相似：不同的是在交叉试验中个体在不同的时间接受不同的治疗，而在这类试验中个体则是在不同的部位接受不同的治疗。见16章描述的方法（16.4节）。将这些研究与那些受试者在多个部位接受同样干预的研究区分开来非常重要（9.3.7节）。

9.3.9 多个干预组

对超过两个干预组进行比较的研究需小心处理。这类研究往往会被纳入Meta分析，因为它们将干预组所有可能的组合进行了多个成对比较。如果同样一组受试者在同一个Meta分析中被纳入两次，会出现一系列的分析单元问题（例如，如果“剂量1 vs 安慰剂”和“剂量2 vs 安慰剂”都被纳入同一个Meta分析，则对安慰剂治疗患者作了两次比较）。见16章（16.5节）。

9.4 汇总研究效应

9.4.1 Meta分析

系统评价的重要步骤是仔细思考所有（或部分）研究的数值结果的合并是否恰当。这样一个Meta分析会得到一个总体统计量（及其可信区间），其汇总了试验干预与对照干预相比的效果（见9.1.2节）。本节描述了对可能遇到的主要数据类型进行Meta分析的原理和方法。

补充文件提供了所有所述方法的公式，RevMan 5（在手册网站可获得）的统计算法和本节中讨论问题的更详细的讨论见Deeks等的研究（Deeks 2001）。

9.4.2 Meta分析的原则

所有Meta分析常用方法遵循如下基本原则。

1. Meta分析典型的过程分两步。第一步，计算每个研究的汇总统计量，以描述所观察的干预效应。例如，如果是二分类资料，汇总统计量可为风险比；或如果是连续性资料，汇总统计量可为均数差。

2. 第二步，以个体研究中估计的干预效应的权重均数计算（合并的）汇总干预效应估计值。权重均数定义为：

$$\text{加权平均值} = \frac{\text{（估计值 x 权值）的和}}{\text{权值和}} = \frac{\sum Y_i W_i}{\sum W_i}$$

此处，Yi是第i个研究的估计干预效应，Wi是对第i个研究所给权重，和是所有研究相加。注意如果所有权重相同，则权重均数等于平均干预效应。如果对第i个研究所赋权重较大，则其权重均数的贡献就更多。因而权重被用于反映每个研究所包含的信息量。对比值指标（OR、RR等），Yi是测量值的自然对数。

3. 研究间干预效应估计值的合并可能随机地引入假设，即研究并不全是估计同样的干预效应，而是估计服从研究间分布的干预效应。这是随机效应Meta分析的基础（见9.5.4节）。或者，如果假设每个研究准确地估计同样的量，则应进行固定效应Meta分析。

4. 汇总（合并）干预效应的标准误可用于推导可信区间（其传达了汇总估计的准确度（或不确定性）），和推导P值（其传达了相对于无干预效应的无效假设的证据强度）。

5. 除了得到合并效应的汇总量之外，所有Meta分析方法都包含评估是否不同研究结果间的变异可归为随机变异，或是否这种变异足够显示出研究间干预效应的不一致性（见9.5节）。

9.4.3 Meta分析的倒方差法

Meta分析程序的常见和简单版本通常是指倒方差法。该方法在RevMan中以其最基本的形式进行，并且用于二分类和连续性资料的Meta分析。

倒方差法得名是因为每个研究所赋权重来自效应估计方差的倒数（如，标准误平方分之一）。因此，较大样本的研究（其标准误较小）比较小样本的研究（其标准误较大）所赋权重大。这一权重选择在最大程度上减小了合并效应估计的不精确性（不确定性）。

固定效应Meta分析使用倒方差法计算加权均数为：

$$\text{方差倒数的加权平均值} = \frac{\sum Y_i \left(1/SE_i^2\right)}{\sum \left(1/SE_i^2\right)},$$

此处，Yi为第i个研究的估计干预效应，SEi为估计值的标准误，和包含所有研究。因此需要纳入分析的基本数据是每个研究干预效应及其标准误的估计值。

9.4.3.1 Meta 分析的随机效应（DerSimonian 和 Larid）方法

倒方差法的变异引入一个假设，即不同的研究评估了不同但相关的干预效应。这产生了随机效应Meta分析，DerSimonian和Laird法（DerSimonian 1986）是已知最简单的版本。随机效应Meta分析在9.5.4节讨论。为进行随机效应Meta分析，对研究特定估计值的标准误（上述SEi）进行调整，以引入在不同研究中观察到干预效应的差异程度的指标或异质性（该差异通常是指τ^2）。差异的量和因此而进行的调整可以从Meta分析纳入研究的干预效应及其标准误进行估计。

9.4.3.2 RevMan 中倒方差结局类型

估计值及其标准误可以直接输入RevMan“倒方差”结局之下。软件将进行固定效应Meta分析和随机效应（DerSimonian和Laird）Meta分析，并进行异质性评估。对于比值一类的干预效应指标，数据应按自然对数输入（如按log OR和log OR的标准误）。然而，其直接通过软件以原始（如OR）度量单位的形式显示结果。不是分别显示各治疗组的汇总数据，森林图除了研究的标识符外，将以他们进入的形式显示估计值及其标准误。可给出一列来提供两组的样本量来补充或替换他。

注意：RevMan中直接输入估计值及其标准误的功能在Meta分析中创造了高度的灵活性。例如，其促进了交叉试验、整群随机试验和非随机研究，以及有序、时间-事件或率等结局类型的分析。然而，在连续性和二分类结局资料分析最常见的情况中，更好的是将更多细节资料输入RevMan（如明确作为每组二分类或连续性资料的简单汇总）。这避免了作者对计算效应估计值的需要，并允许特定针对不同资料类型的方法的使用（见9.4.4节和9.4.5节）。并且，这有助于系统评价读者去看每个研究每个干预组的汇总统计量。

9.4.4 二分类结局的Meta分析

有四种广泛使用的二分类变量Meta分析方法，三种为固定效应方法（Mantel-Haenszel，Peto和倒方差），一种为随机效应方法（DerSimonian和Laird）。在

RevMan软件中，所有这些方法都可以作为一种分析选择获得。Peto法仅能合并OR，而其他三种方法能合并OR、RR和RD。所有Meta分析方法公式由Deeks等（Deeks 2001）提出。

注意：发生率为零（如一组无事件）在一些方法估计值及其标准误的计算中会带来问题。对于这类研究RevMan软件自动在2×2表格的每一格中增加0.5。

9.4.4.1 Mantel-Haenszel 法

Mantel-Haenszel法是编程在RevMan中默认的Meta分析固定效应方法。不管是事件发生率低还是样本量小所致，当数据稀少时，倒方差所用的效应估计值的标准误可能欠佳。Mantel-Haenszel法根据使用的效应指标（如RR、OR、RD），采用了不同的加权方式。当为小概率事件时，其表现出较好的统计性能。由于在Cochrane系统评价中这是常见情况，因此Mantel-Haenszel法一般优于倒方差法。其他情况下，两种方法结果相似。

9.4.4.2 Peto OR 法

Peto法（Yusuf 1985）仅用于合并OR。其利用倒方差法近似估计OR的对数值，且使用不同的权重。另一种观点认为Peto法是“O-E”统计量的总和。此处，O是观察的事件数，E是每个研究中试验组期望的事件数。

当干预效应很小（OR接近1）、事件并不特别常见以及试验组和对照组事件数相似时，使用近似法来计算log OR的方法效果较好。其余情况下，其可能给出有偏倚的结果。因为这些标准并不总是能满足，因此并不推荐将Peto法作为Meta分析的默认方法。

当使用Peto法时，不必对四格表中为零的格子进行校正。或许因为该原因，当事件非常罕见时，该法很有用（Bradburn 2007）（见16章16.9节）。而且，Peto法可对二分类结局数据的研究和使用进行了对数秩检验的时间-事件分析的研究进行合并（见9.4.9节）。

9.4.4.3 随机效应法

随机效应法（DerSimonian 1986）引入了一个假设，即不同的研究估计了不同但相关的干预效应。如在9.4.3.1节描述的一样，该法基于倒方差法，根据不同干预效应间差异的程度或异质性对研究权重进行调整。当研究间无异质性时，随机效应法和固定效应法结果相同。有异质性时，如使用随机效应法，平均干预效应的可信区间比使用固定效应法宽，并且对统计学意义的相应要求更保守。如果观察的干预效应与样本量之间存在

相关性，干预效应的估计中值也可能改变。对这些问题的进一步讨论见9.5.4节。

对二分类变量，RevMan可完成两种随机效应方法：Mantel-Haenszel法和倒方差法。两种方法间的差异很小：前者通过对每个研究的结果和Mantel-Haenszel法固定效应Meta分析结果进行比较来估计研究间差异的量，而后者过对每个研究的结果和倒方差固定效应Meta分析结果进行比较来估计研究间差异的量。实际上，区别可能是细微的。RevMan5中增加了倒方差法。

9.4.4.4 二分类结局包含哪些指标？

9.2.2节介绍了二分类资料的汇总统计量。干预效应可用相对或绝对效应来表达。RR和OR是相对指标，而RD和NNT是绝对指标。更为复杂的是事实上有两个风险比。我们可以计算一个事件发生的风险比或一个事件未发生的风险比。Meta分析中他们有不同的合并结果，有时差异较为显著。

Meta分析中汇总统计量的选择需平衡三个标准（Deeks 2002）。首先，我们需要在Meta分析中对所有研究都给出相似值的一个汇总统计量，和干预将被应用的人群分类。汇总统计量越一致，对表达干预效应作为一个单一汇总数的论证就越强。其次，汇总统计量必须具备实施有效Meta分析所有的数学特性。第三，汇总统计量应易于被系统评价用户理解和使用。应以一种有助于读者适当解释和应用结果的方式来呈现干预效应的汇总结果。在二分类资料的效应指标中，没有哪一个指标是绝对最好的，因此，对指标的选择不可避免地要进行折中。

一致性：实证证据表明，一般情况下相对效应指标比绝对效应指标一致性好（Engels 2000，Deeks 2002）。因此，避免做风险差的Meta分析是明智的，除非有清楚的理由表明在特定的临床情况下，风险差是一致的。总体上讲，OR和RR在一致性上差异很小（Deeks 2002）。当研究旨在降低不良结局的发生率时（见9.2.2.5节），有实证证据表明，不良结局的RR比无事件的RR更一致（Deeks 2002）。在特定的情况下选择一致性最好的效应指标并不是一个推荐的策略，因为其可能导致最大化Meta分析精确度的假象。

数学特性：最重要的数学标准是可以进行可靠的差异评估。NTT并无一个简单的差异评估法，Meta分析时不能直接使用，尽管其可以通过其他汇总统计量进行计算（见12章12.5节）。另两种经常引用的数学特性的重要性未达成一致共识：比值比和风险差的大小不依赖于这两个结局编码的事件，以及比值比是唯一的无限统计量。

可解释性：在实践中，OR是最难理解和应用的汇总统计量，许多临床医生在使用

它们时遇到困难。有许多发表的例子显示，作者从Meta分析中错误地按RR解释了OR。我们必须注意到，常规以OR来表述系统评价结果通常导致当其应用于临床实践时，治疗的获益和危害被过高估计。绝对效应指标比相对效应指标更易被临床医生理解（Sinclair 1994），并在干预措施可能的获益和可能的危害间做出权衡取舍。然而，绝对效应指标却难以推广。

实证证据表明，在汇总统计量不太可能对干预效应（风险差）给出一致估计时，避免其使用似乎很重要，并且，在不能进行Meta分析时，使用统计量NNT很重要。因此，通常推荐分析从使用RR（注意哪些结局归类为事件）或OR开始。进行敏感性分析以了解汇总统计量的选择（和事件分类的选择）对于Meta分析的结论是否关键，是明智的（见9.7节）。

在Meta分析中使用一个统计量，并在结果重新表述时使用另一个更易解释的统计量，通常是明智的。例如，通常Meta分析可能最好使用相对效应指标（RR或OR）并在结果再表述时使用绝对效应指标（RD或NNT——第12章，12.5节）。这是在Cochrane系统评价中给出“结果汇总”表的关键原因之一：见11章（11.5节）。如果OR被用于Meta分析，他们也能被再表述为RR（见12章，12.5.4节）。在所有情况下，上下置信限的转换可使用同样的公式。然而，重要的是要注意，所有这些转换需要设定基线风险值（其显示了将试验干预应用到“对照”人群可能的结局风险）。在该假设的对照风险所选值接近研究间典型的所观察的对照组的风险时，则可获得相似的绝对效应估计，而不论是OR还是RR被用于Meta分析。在假设的对照风险与典型的观察的对照组的风险不一致时，绝对获益的预测将根据Meta分析所用统计量而不同。

9.4.5 连续性结局的Meta分析

RevMan中对连续性结局的Meta分析有两种分析方法：倒方差固定效应法和倒方差随机效应法。当无异质性时，该两种方法结果相同。有异质性时，采用随机效应法比固定效应法所得平均干预效应的可信区间宽，且相应的P值显著性更低。如果所观察的干预效应与样本量之间存在相关性，则干预效应的中央估计值将改变。对这些问题的进一步讨论见9.5.4节。

作者应清楚，连续性资料Meta分析方法的假设前提是每个研究每个干预的结局服从正态分布。这一假设可能并不总是满足，尽管它对大样本研究意义不大。这有助于思考

偏态分布数据的可能性（见9.4.5.3节）。

9.4.5.1 连续性结局有哪些指标？

有两个汇总统计量用于连续性结局的Meta分析，均数差（MD）和标准化均数差（SMD）（见9.2.3节）。连续性数据汇总统计量的选择主要取决于研究在结局报告上是否均使用了同样的度量单位（这时可用均数差）或使用了不同的度量单位（这时必须使用标准化均数差）所决定。

应该理解在两种方法中，观察结局的标准差所起的不同作用。

- 对于均数差法，标准差与样本量被共同用于计算每个研究的权重。标准差较小的研究被赋予相对大的权重，而标准差较大的研究被赋予相对小的权重。如果研究间标准差的差异能反应结局指标可靠性的差异，则是恰当的；但如果标准差的差异反应了研究人群结局差异性的真实差异，则可能并不恰当。
- 对于标准化均数差法，标准差被用于将均数差标化为一个度量单位（见9.2.3.2节），并用于研究权重的计算。其假设研究间标准差的差异仅能反应测量度量单位的差异，而非结局指标可靠性或研究人群间变异的差异。

观察到研究间标准差意想不到的变异时，这些方法的局限性应牢记在心。

9.4.5.2 积分改变的 Meta 分析

在一些情况下，基于基线改变的分析将比最终值的比较更有效，且功能强大，因为其从分析中排除了受试者之间的变异。然而，积分改变的计算需要对结局进行两次测量，并且在实践中，对不稳定或难于准确测量的结局，其效率可能不高，因其测量误差可能比受试者间真实的基线变异更大。如果基线结局改变量比最终测量结局分布更正态，其也可能被优先考虑。虽然有时对于随机化不理想的情况进行校正，但并不推荐这样做。

计算基线结局测量结果优先考虑的统计方法是在回归模型或协方差分析（ANCOVA）中将基线结局测量结果作为协变量纳入。这些分析会得到治疗效应及其标准误的“调整的”估计值。这些分析使用频率最少，但因为他们给出了治疗效应最准确和偏倚最小的估计值，因此当他们可以使用时，分析中应该使用这些方法。然而，每个干预组的均数和标准差不可得，因此他们仅能用于使用普通倒方差法的Meta分析。

在实践中，作者可能发现，纳入系统评价的研究可能既包括基线改变值，也可能包括最终值。然而，当谈到均数差的Meta分析时，混合的结局指标并不是一个问题。当在

RevMan中使用（非标准化的）均数差法时，为何在Meta分析中以基线改变值为结局的研究不应与以最终测量值为结局的研究进行合并，并无统计学的理由。在随机试验中，在基线改变值的基础上获得的均数差通常被假设与在最终测量值基础上进行的分析可以得到同样准确的干预效应。也就是说，大体上，平均最终值的差异与平均积分改变相当。如果积分改变的使用不能增加准确性，在最终值被使用的情况下，呈现积分改变的研究将在分析中恰当地被赋予比他们已有的权重更大的权重，因为他们将有更小的标准差。

在进行数据合并时，作者必须小心对每个研究使用恰当的均数和标准差（最终测量结果的或基线改变的）。因为对两种类型的结局，均值和标准差可能略有不同，因此，建议将其分为两个亚组以避免对作者引起混淆，但可以将亚组结果合并在一起。

然而，最终值和积分改变不能作为标准化均数差合并在一起，因为标准差的差异并不能反映测量度量单位的差异，而是反映测量可靠性的差异。

与纳入基线改变量指标相关的常见实践问题是改变量的标准差并未报告。对缺失标准差的插补问题的讨论见第16章（16.1.3节）。

9.4.5.3 偏态资料的 Meta 分析

对于近似正态分布的数据和来自大样本试验的数据，用均数来进行分析是恰当的。如果结局真实的分布是不对称的，则是说数据是偏态的。偏态有时可从结局的均数和标准差进行判断。可以进行粗略地检查，但其仅在一个已知结局的最小或最大可能值存在的情况下才可用。因此，这种检查可用于像体重、体积和血液浓度一类的结局（其最小可能值为0），或有最小或最大积分的量表类结局，但其对基线改变量指标可能并不恰当。这种检查包括计算所观察的均数减最小可能值（或最大可能值减所观察的均数），并用标准差来除该差值。其比值小于2表明呈偏态分布（Altman 1996）。如果比值小于1，则表明有很强的证据表明数据呈偏态分布。

对原始结局数据进行转换可在很大程度上降低偏度。试验报告可以转换度量单位，通常是对数度量单位，来呈现结果。由试验专家来对恰当的数据进行收集、归纳，或获取个体患者的数据，是当前可选择的方法。对个体患者数据的恰当数据归纳和分析策略视情况而定。建议咨询学识渊博的统计学家。

在数据以对数度量单位进行分析之处，结果通常以几何均数和几何均数比值的形式表示。Meta分析于是可以在对数转换数据度量单位的基础上进行；所有计算均数和标准

差的例子见第7章（7.7.3.4节）。该方法实施的前提是能够获得所有研究转换的数据；已有将一种度量单位转换为另一种度量单位的方法（Higgins 2008a）。在一个Meta分析中，经对数转换和未经对数转换的数据不能合并。

9.4.6 合并二分类和连续性结局

有时作者会碰到这样一种情况，对于同样的结局，在一些研究中数据以二分类资料表示，而在另一些研究中数据以连续性资料表示。例如，抑郁量表积分可以干预后的均数或达到某些点而被诊断为抑郁患者的百分比（如积分在特定截值之上）的形式报告。当将其分为两类后，这类信息通常更易理解和更有用。然而，截点的选取可能带有随意性，并且当连续性数据转换成二分类数据后，信息会损失。

有几种方法可以处理二分类和连续性数据的合并问题。通常，以一种相似的方式从所有相关的、有效的研究中总结结果是理想的，但这往往不可能。为了达到这一目的，可能的方式是从研究者那里收集缺失数据。如果做不到，以下三种方式有助于总结数据：作为连续性结局记录均数和标准差、作为二分类结局记录计数和作为“其他数据”结局以文本的形式记录所有的数据。

有统计学方法可将OR转换成标准化均数差（反之亦然），将二分类和连续性数据合并在一起。基于每个干预组的连续性指标服从逻辑分布（这是一种对称分布，在形状上与正态分布相似，但数据更多分布在尾部），且结局变异在治疗和对照受试者中相同这样假设，根据下述简化公式可将OR转换为SMD（Chinn 2000）。

$$SMD = \frac{\sqrt{3}}{\pi} \ln OR$$

log OR的标准误可通过乘以同样的常数（$\sqrt{3}/\pi = 0.5513$）而转换为SMD的标准误。或者，SMD可通过乘以$\pi/\sqrt{3} = 1.814$而转换成log OR。一旦在Meta分析中计算出所有研究的SMD（或log OR）及其标准误，则在RevMan中可使用倒方差法对其进行合并。通过在RevMan中输入二分类和连续性结局类型的数据，可计算出所有研究的标准误，并可视情况将log OR和SMD的可信区间转换成标准误（见第7章，7.7.7.2节）。

9.4.7 有序结局和测量量表的Meta分析

有序和测量量表结局根据研究作者实施原始分析的方式最常以二分类数据（见9.4.4节）或连续性数据（见9.4.5节）进行Meta分析。

偶尔，在有序度量单位类别数少，每个干预组每个类别数量可知，且所有研究都使用相同有序度量单位的情况下，可使用比例优势模型（proportional odds model）进行数据分析。该方法可比二分类法更有效地使用可获数据，但需获得统计软件的使用权限，并且要从汇总统计量的结果中找到临床意义是个挑战。

比例优势模型使用了优势比值比作为干预效应指标（Agresti 1996）。假如有三个类别，其根据愿望（如1是最好，3是最差）进行排序。数据可按两种方式分成两类。即，将分类1作为成功，将分类2-3作为失败；或将分类1-2作为成功，将分类3作为失败。优势比例模型假设两种数据的分类方法OR相等。因此，通过优势比例模型计算的OR可解释试验干预相对于对照成功的比值，不考虑这种有序的分类如何被分成成功或失败。已有方法（尤其是多分类logistic回归模型）可用于计算log OR及其标准误的研究估计值和在高级软件包中进行Meta分析（Whitehead 1994）。

通过比例优势模型获得的log OR及其标准误的估计值可在RevMan中使用普通倒方差法进行Meta分析（见9.4.3.2节）。随机效应和固定效应分析方法均可用。如果所有的研究使用了相同的有序度量单位，但在一些报告中以二分类结局的形式进行表述，则在Meta分析中仍可纳入所有研究。在3分类模型的情况下，其可能意味着在一些研究中将分类1作为成功，而其他研究中将分类1和2作为成功。已有处理这一情况的方法，同时也有合并度量单位相关但其分类定义不同的数据的方法（Whitehead 1994）。

9.4.8 频数和率的Meta分析

当每个受试者可能发生一次和不止一次事件时，结果可按计数资料表达（见9.2.5节）。例如，“卒中数”或“医院就诊数”为计数资料。这些事件可能完全未发生，但如果发生了，对个体没有理论最大发生数。

如第7章（7.7.5节）所描述的，计数资料可以使用分析二分类（见9.4.4节）、连续性（见9.4.5节）和时间-事件资料（见9.4.9节）的方法进行分析，同时也可作为比率资料进行分析。

如果将对每个受试者的计数与观察时间一起测量，则会出现比率数据。当被计数的事件为罕见事件时尤其适用。例如，一个妇女在两年的随访期内经历了两次卒中。其卒中率为每随访年1次（或等价于每随访月0.083次）。率通常在小组水平进行汇总。例如，临床试验对照组的受试者在总共2836个随访人年中可发生85次卒中。与率的使用相关的

假设前提是一个事件的风险在受试者间和整个观察时间内是不变的。对每个情况，都应仔细斟酌该假设。例如，在避孕研究中，率被用于描述每100个随访妇女-年的妊娠次数（称为Pearl指数）。但因为夫妇有不同的受孕风险，且每个妇女的风险在观察期内会发生变化，则现在这被认为并不恰当。目前对妊娠更常使用生命表或时间-事件方法（其能够研究在首次妊娠前经过的时间）进行分析。

将计数资料按率来分析并非总是最恰当的方法，且在实践中并不常用。这是因为：

1 潜在风险保持不变的假设并不恰当；

2 统计方法并不如其他数据类型的统计方法完善。

研究结果可以用率比表达，其是试验干预组的率相对于对照组率的比。假设在试验干预组中在TE个随访受试者年中发生EE次事件，在对照组中在TC个随访受试者-年中发生EC次事件。则率比为：

$$\text{比率} = \frac{E_E/T_E}{E_C/T_C} = \frac{E_E T_C}{E_C T_E}$$

研究间率比的（自然）对数可使用倒方差法进行合并（见9.4.3.2节）。率比对数的近似标准误计算公式为：

$$\ln(\text{比率})\text{的标准误} = \sqrt{\frac{1}{E_E} + \frac{1}{E_C}}$$

在0事件的情况下，可在每个计数上加0.5进行校正。注意：事件单位的选择（如患者-月、妇女-年等）是无关的，因为其在计算率比时被抵消了，并且不会在标准误中出现。但当呈现研究结果时，单位仍应显示出来。可以通过Whitehead的方法估算率比（Whitehead 1991）。

在随机试验中，率比可能常常与将受试者分为两类后获得的相对危险度非常相似，因为所有干预组的平均随访期应该是相似的。然而，如果干预影响了某些受试者发生多次事件的可能性，则率比和相对危险度将会有所差异。

也可能关注率差，

$$\text{率差} = \frac{E_E}{T_E} - \frac{E_C}{T_C}$$

率差的近似标准误为

$$\text{率差的标准误} = \sqrt{\frac{E_E}{T_E^2} + \frac{E_C}{T_C^2}}$$

在RevMan中，其分析再一次需要使用倒方差法。对率的Meta分析的仅有的讨论，仍然十分短，见Hasselblad和McCrory的研究（Hasselblad 1995）。

9.4.9 时间-事件结局的Meta分析

在RevMan中可使用两种方法进行时间-事件Meta分析。采用哪一种取决于从原始研究中提取或对单个患者数据再分析的数据类型。

如果通过对个体患者数据的再分析，或对研究报告中呈现的统计量进行整合而获得了“O-E”和“V”统计量，则可使用“O-E和方差”结局类型将这些统计量直接输入RevMan中。有几种计算“O-E”和“V”统计量的方法。应用于二分类资料的Peto法可得到OR；对数秩法可得到危害比，分析时间-事件资料的改良Peto法可得到他们之间一些东西。在RevMan中应确定恰当的效应指标。在RevMan中，对于“O-E和方差”结局，只能使用固定效应Meta分析法。

或者，如果已从Cox比例风险回归模型的结果中获得了危害比的对数值及其标准误，则研究结果可使用倒方差法进行合并（见9.4.3.2节）。固定和随机效应分析均可使用。

如果从研究中同时获得了对数秩和Cox模型估计值，则所有结果可使用倒方差法进行合并，因为对数秩估计值可使用第7章（7.6.6节）所给的公式转换成危害比的对数值及其标准误。

9.4.10 RevMan中可用的Meta分析方法小结

表9.4.a列出了RevMan中可用的统计方法选择。RevMan要求作者对每个结局都选择一个最佳方法。如果没有进行选择，则软件默认对二分类变量为固定效应Mantel-Haenszel OR，对连续性结局为固定效应均数差，对倒方差结局为固定效应模型。重要的是作者要清楚当结果在系统评价中呈现的时候，他们使用了那种方法，因为其不能确保显示给用户的Meta分析就是经选择的最佳方法。

表9.4.a　RevMan中可用的Meta分析方法小结

资料类型	效应指标	固定效应方法	随机效应方法
二分类	比值比（OR）	Mantel-Haenszel（M-H） 倒方差（IV）	Mantel-Haenszel（M-H） 倒方差（IV）
		Peto	
	风险比（RR）	Mantel-Haenszel（M-H） 倒方差（IV）	Mantel-Haenszel（M-H） 倒方差（IV）
	风险比差（RD）	Mantel-Haenszel（M-H） 倒方差（IV）	Mantel-Haenszel（M-H） 倒方差（IV）
连续性	均差（MD）	倒方差（IV）	倒方差（IV）
	标准化均数差（SMD）	倒方差（IV）	倒方差（IV）
O – E和方差	用户自定（默认为“Peto比值比”）	Peto	None
普通倒方差	用户自定	倒方差（IV）	倒方差（IV）
其他资料	用户自定	None	None

9.4.11　Meta分析投票计数的使用

偶尔，Meta分析使用“投票计数”来比较阳性研究数和阴性研究数。投票计数只限于回答“有任何效应证据吗”这样简单的问题。投票计数可能出现两个问题，建议对其尽可能避免。第一，如果主观决定或统计学意义被用于定义“阳性”和“阴性”研究，会出现问题（Cooper 1980, Antman 1992）。为恰当地进行投票计数，应对显示危害的研究数和显示获益的研究数进行比较，而不是统计学意义或其结果的大小。可用符号检验评价存在正负效应的证据的意义（如果无效，则研究将均匀地分布于无差异的零假设周围）。第二，投票计数未考虑对每个研究赋予不同的权重。在标准Meta分析方法不能用的情况下（如无一致的结局指标），投票计数被认为是最后的办法。

9.5 异质性

9.5.1 什么是异质性？

一个系统评价纳入的研究难免会有差异。系统评价中研究间任何种类的变异都被称为异质性。区分不同类型的异质性可能有用。所研究的受试者、干预措施和结局的变异可被称为临床多样性（有时称为临床异质性），而研究设计中的多样性和偏倚风险可称为方法学多样性（有时称为方法学异质性）。在不同研究中所评估的干预效应中的多样性被称为统计学异质性，并且它是研究间临床和/或方法学多样性的结果。统计学异质性表明所观察干预效应间的差异比所期望的仅由随机误差（机遇）所致的差异大。我们将按照惯例，将“统计学异质性”简称为“异质性”。

如果干预效应受研究间因素改变的影响，则这种临床变异将导致异质性。最明显的是特定的干预措施或患者特征。换句话，真实的干预效应在不同研究间将不同。

研究间在方法学因素上的差异，如盲法和分配隐藏的使用，或如果研究间在定义结局和测量的方法上存在差异，则可能导致在所观察的干预效应间存在差异。来自方法学多样性或结局评估中的差异的统计意义异质性表明，研究并非都有同样的估计效应，但并不一定表明真实的干预效应存在差异。特别地，仅与方法学多样性相关的异质性表明，研究受到不同程度偏倚的影响。实证证据表明，设计的某些方面可能影响临床试验结果，虽然情况并非总是如此。进一步的讨论见第8章。

系统评价的范围将在很大程度上决定系统评价纳入研究差异的大小。有时，一篇系统评价将纳入针对不同问题的研究，例如，当关注于针对同一疾病的几种不同干预时（也见第5章，5.6节）。每种干预纳入研究都应分开分析和呈现。当一组研究在受试者、干预措施和结局方面对于提供一个有意义的汇总结果有足够的同质性时，才考虑进行Meta分析。系统评价通常比单个临床试验关注的范围更广。一个常见类比是系统评价将苹果和橘子放在一起，将他们合并在一起可能产生毫无意义的结果。如果关注苹果和橘子自身，则这是正确的；但如果他们被用于回答更宽泛的关于水果的问题，则不正确。例如，一个系统评价可能通过合并评估了属于同一类别不同药物效果的试验结果合理地评估了一类药物的平均效果。

系统评价可能特别关注于分析研究的临床和方法方面如何与其结果相关。这些分析应尽可能事先确定，如在系统评价计划书中。对于一个系统评价而言，应该关注于分析

研究的某些临床特征和干预效应大小之间的相关性，而不是获得一系列研究的汇总效应估计值。对于该目的，Meta回归是最佳方法，虽然其在RevMan中不能完成（见9.6.4节）。

9.5.2 识别和测量异质性

重要的是思考在什么范围下研究结果是一致的。如果个体研究结果的可信区间（通常用横线进行图示）重叠较少，这通常表明存在统计学异质性。更正规的是对异质性进行统计学检验。Cochrane系统评价的森林图中包括了检验异质性的卡方（χ^2或Chi^2）检验。其评价了所观察结果间的差异是否仅由机遇所致。小的P值（或与其自由度相关的大的卡方统计量）提供了干预效应存在异质性的证据（效应估计值的变异超出了机遇）。

在对卡方检验进行解释时必须小心，因为当研究样本量较小或数量较少时，在Meta分析的（通常）情况下，其检验效能较低。这意味着有统计学意义的结果可能表明存在异质性问题，无显著性结果不能视为无异质性的证据。这也就是为何有时将0.01的P值，而不是常规0.05的水平，用于确定统计学意义。该检验的另一个问题（其在Cochrane系统评价中很少出现）是当Meta分析纳入很多研究时，有较高的检验效能可能检出并无临床意义的少量异质性。

有一些争论认为，在Meta分析中总会产生临床和方法学多样性，因此统计学异质性不可避免（Higgins 2003）。因此异质性检验与分析的选择不相关；异质性将总是存在，而不论我们是否能使用统计学检验碰巧将其检测出来。已有方法用于对研究间的不一致性进行量化，其将关注点从检测是否存在异质性转向评估其对Meta分析的影响。对不一致性进行量化的有用统计量为：

$$I^2 = \left(\frac{Q - df}{Q}\right) \times 100\%$$

此处，Q是卡方统计量，df是自由度（Higgins 2002，Higgins 2003）。该公式描述了由异质性而不是抽样误差（机遇）所致的干预估计值变异的百分比。

I^2解释的阈值可能会产生误导，因为不一致的重要性由几个因素决定。以下给出对解释的初略指导：

- 0%至40%：可能不重要；
- 30%至60%: 可能存在中度异质性*；
- 50%至90%: 可能存在实质性异质性*；
- 75%至100%：存在较大异质性*。

* 所观察I^2值的重要性依赖于：（i）效应的大小和方向；（ii）异质性证据的强度（如从卡方检验获得的P值，或I^2的可信区间）。

9.5.3 解决异质性的策略

如果在一组研究间（从其他方面看适合进行Meta分析）找出了（统计学）异质性，许多方法可以使用。

1. 再次检查数据是否正确

严重异质性可能表明数据被错误地提取或输入了RevMan。例如，如果对于连续性结局标准误被误输为标准差，这可能表现为可信区间非常窄，重叠非常少，并因此产生实质性异质性。分析单位误差也可能是异质性的原因（见9.3节）。

2. 不做Meta分析

系统评价不一定包含Meta分析（O'Rourke 1989）。如果结果存在很大差异，尤其如果效应方向存在不一致，则使用干预效应的平均值可能会产生误导。

3. 探索异质性

令人感兴趣的无疑是确定导致研究结果间存在异质性的原因。异质性可通过进行亚组分析（见9.6.3节）或Meta回归（见9.6.4节）来探索，虽然后一种方法在RevMan中不能进行。理想地，对可能与异质性相关的纳入研究特征的分析应该在系统评价计划书中事先说明（见9.1.7节）。可靠的结论仅能从在看到纳入研究的结果前真正事先确定的分析中得到，即便如此，这些结论也应被谨慎解释。在实践中，当撰写研究计划书时，作者常常会去熟悉一些研究结果，因此真正事前确定并不可能。在找到异质性后才计划进行异质性的探索在最乐观的情况下也可能导致假设的产生。他们应该被更加谨慎地解释，并通常不应列于系统评价的结论中。而且，当研究很少时，异质性分析的价值值得怀疑。

4. 忽略异质性

固定效应Meta分析就忽略了异质性。通常，固定效应Meta分析得到的合并效应估计值被认为是干预效应的最佳估计值。然而，异质性的存在表明，干预效应不是单一值，而是一个分布。因此，固定效应合并估计值可能是并非在任何人群都存在的干预效应，因此该估计值可能存在一个太窄而等同于无意义的可信区间（见9.5.4节）。然而从固定效应Meta分析获得的P值的确提供了对无效假设（每个研究都无效）有意义的检验。

5. 进行随机效应Meta分析

随机效应Meta分析可用于综合研究间的异质性。这并不能取代对异质性的彻底分析。其主要用于不能解释的异质性。对该内容的更多讨论见9.5.4节。

6. 改变效应指标

异质性可能是由于对效应指标的不恰当选择而人为造成。例如，研究选择使用不同量表或不同单位的连续性结局时，当使用均数差时可能出现显著异质性，而当使用更恰当的标准化均数差时，则不会出现显著异质性。而且，对二分类变量效应指标（OR、RR、或RD）的选择可能影响研究结果异质性的程度。尤其，当对照组风险不同时，同质的OR或RR将肯定得到有异质性的RD，反之亦然。然而，在一个特定Meta分析中干预效应的异质性是否是用于选择这些指标的合适标准，仍然不清楚（也见9.4.4.4节）。

7. 排除研究

异质性可能是由于一两个与其他研究结果相冲突的结果不一致的研究的存在。通常情况下，基于研究的结果而将部分研究从Meta分析中排除掉是不明智的，因为这可能带来偏倚。然而，如果造成结果不一致的原因明确，则大可将其排除掉。因为通常在任何Meta分析中对任何研究都能找到至少一个与其他研究不同的特征，因为这太容易了，所以这一标准并不可靠。建议对纳入或不纳入结果不一致的研究进行敏感性分析。如果可能，应在计划书中详细说明可能导致临床多样性的潜在原因。

9.5.4 通过随机效应模型综合异质性

固定效应Meta分析提供了作为一种“典型的干预效应”（来自Meta分析纳入研究）可见的结果。为了针对固定效应Meta分析计算可信区间，前提假设是在每个研究中真实的干预效应（包括大小和方向）具有相同的值（即在研究间是固定的）。这一假设意味着，所观察到的研究结果间的差异仅由机遇所致，即没有统计学异质性。

当异质性难以解释时，一种分析方法是将其综合进随机效应模型。随机效应Meta分析模型的假设是，虽然不同研究间估计的效应不同，但服从某种分布。该模型说明我们缺乏关于为何真实的、或外在的干预效应存在差异的知识（我们认为这种差异是随机的）。该分布的中心为平均效应，其宽度则为异质性的程度。分布的常规选择为正态分布。任何分布假设的效度都难以确定，这通常是对随机效应Meta分析的批评。对这一分布特定假设形状的重要性并不清楚。

注意：随机效应模型不能“计算”异质性，就意义而言它不再是一个问题。建议对异质性的可能原因进行探索，虽然很少有研究充分地这样做（见9.6节）。

在RevMan中，对随机效应分析，合并估计值及其可信区间是指干预效应分布的中心，而不是描述分布的宽度。通常，合并估计值及其可信区间被作为固定效应Meta分析中评估的定量估计值而分别引用，这并不恰当。来自随机效应Meta分析的可信区间描述了在不同研究中不同效应平均值位置的不确定性。其通常并未描述研究中可以被相信的异质性程度。例如，当Meta分析中有许多研究时，即使存在较大异质性时，围绕平均效应的随机效应的估计值，可得到较窄的可信区间。

与其他Meta分析软件相同，RevMan在随机效应Meta分析中给出了研究间变异的估计值（被称为τ^2）。该数值的平方根（即τ）是研究间潜在效应的估计标准差。对于绝对效应指标（如RD、MD、SMD），可通过创造一个从随机效应合并估计值之下$2\times\tau$，到其之上$2\times\tau$的区间来获得潜在效应的近似95%范围。对相对指标（如OR、RR），区间以合并估计值的自然对数为中心，反对数（对数还原）界限以获得基于比率度量单位的区间。对于在一个新研究中被预测的效应，替代区间已被提出（Higgins 2008b）。研究中所观察的干预效应的范围可能被认为给出了真实干预效应的大概理想的分布宽度，但事实上，其有点太宽了，因为其也描述了观察的效应估计值的随机误差。

如果效应变异（统计学异质性）被认为是由临床多样性所致，对随机效应合并估计值与固定效应估计值的解释应不同。随机效应估计值及其可信区间解决“什么是平均干预效应”的问题，而固定效应估计值及其可信区间解决“什么是干预效应最佳估计值”的问题。当无异质性存在，或当干预效应的分布基本对称时，对这些问题的回答是一致的。当回答不一致时，随机效应估计值在任何被研究的特定的人群中都不可能反映真实的效应。

方法学多样性通过影响不同研究结果的偏倚来产生异质性。如果导致干预效应过高和过低估计值的偏倚均衡分布（不太可能是这种情况），则随机效应合并估计值仅能估计平均治疗效果。在实践中，对于区分异质性结果到底是来自临床还是方法学多样性非常困难，在大多数情况下，两种情况都有，因此难于对这些区别给予解释。

对于任何存在异质性的一组特定研究，随机效应合并估计值的可信区间比固定效应合并估计值的可信区间宽。如果I2统计量大于零，将出现该情况，即使未通过卡方检验检测到异质性（Higgins 2003）（见9.5.2节）。在固定效应和随机效应Meta分析之间进行选择，不应以异质性的统计学检验为基础。

在有异质性的一组研究中，随机效应Meta分析将比固定效应Meta分析赋予较小样本的研究更大的权重。这是因为小样本研究能提供更多的信息，来了解研究间效应的分布，而不是把它假定为一个普通的干预效应。必须注意，仅当确证干预效应存在“随机分布”时，才能使用随机效应分析。尤其，如果小样本研究结果与大样本研究结果相比存在系统差异，小样本研究可能存在发表偏倚或研究内偏倚（Egger 1997，Poole 1999，Kjaergard 2001），而随机效应Meta分析将进一步加重这种偏倚效应（也见第10章，10.4.4.1节）。固定效应分析受影响更少，虽然严格意义上讲其也不恰当。在这种情况下，明智的是既不做Meta分析，也不排除小样本研究进行敏感性分析。

相似地，当信息很少时，或是因为研究很少，或是如果研究样本很小，随机效应分析将得到分布较宽的较差干预效应估计。

RevMan所采用的随机效应Meta分析版本是DerSimonian和Laird所描述的（DerSimonian 1986），该方法的优点是计算是直接的，但其理论缺点在于可信区间稍微有点窄而不能涵盖足够的由所估计的异质性程度而导致的不确定性。虽有能够涵盖更多不确定性的替代方法，但他们需要更高级的统计软件（也见16章，16.8节）。在实践中，结果的差异可能很小，除非研究较少。对于二分类变量，RevMan可完成DerSimonian和Laird随机效应模型的两个版本（见9.4.4.3）。

9.6 研究异质性

9.6.1 交互作用和效应修正

干预效应会随着不同的人群或干预特征（如剂量或疗程）而变化吗？这种变异被统计学家称为交互作用，被流行病学家称为效应修正。用于找出这种交互作用的方法包括亚组分析和Meta回归。所有方法都存在较大缺陷。

9.6.2 什么是亚组分析

亚组分析包括将所有受试者数据分到不同亚组中，以致在各亚组间能进行比较。亚组分析可以对不同受试者（如男性和女性）或不同的研究（如在不同地点实施）进行。亚组分析也可作为分析异质性结果的方法而进行，或用于回答有关特定患者、干预类型或研究类型的问题。

在对研究文献进行的系统评价中，针对研究内不同受试者的亚组分析很少见，因为需提取的关于不同受试者资料的足够信息很少在研究报告中发表。相反，当收集到个体患者数据时，则容易对不同受试者进行分析。我们在9.6.3节描述的方法是针对试验亚组。

来自多个亚组分析的结果可能产生误导。亚组分析是通过特征来观察的，而不是基于随机的比较。亚组分析进行越多，出现假阴性和假阳性显著性检验的可能性迅速增加。如果其结果作为最终结论被呈现，则无疑会出现患者拒绝有效干预并接受无效（或甚至有害）干预的风险。亚组分析也会对未来的研究方向产生误导性推荐，如果发生，其将对稀缺资源造成浪费。

区分“定性交互作用”和“定量交互作用”的概念是有用的（Yusuf 1991）。如果效应方向相反，即如果干预在一个亚组有益但在另一个亚组有害，则会出现定性交互作用。定性交互作用非常少。这可能是对Meta分析的最适结果是所有亚组的总效应争论。当效应量而不是效应方向不同时，即干预的益亚组间程度不同，处在则会出现定量交互作用。

作者可在Oxman和Guyatt（Oxman 1992）以及Yusuf等（Yusuf 1991）有关亚组分析的文章中找到有用的建议。也见9.6.6节。

9.6.3 进行亚组分析

在RevMan中可进行亚组分析。亚组内的Meta分析和合并几个亚组的Meta分析都是可以的。分别思考每个亚组的Meta分析结果以比较不同亚组的效应估计值是令人感兴趣的。这仅应通过比较效应大小而非正式的完成。注意在一个亚组内效应或异质性检验有统计学意义，同时在另一个亚组无统计学意义，并不表明亚组因素解释了异质性。因为不同的亚组可能所含的信息量不同，因此对于效应检测有不同的能力，仅仅比较结果的统计学意义会带来很大的误导。

9.6.4 Meta回归

如果把研究分入不同亚组（见9.6.2节），作为一种分析方法可以看到在Meta分析中分类研究特征与干预效应的相关程度。例如，分配序列隐藏充分的研究与那些隐藏不充分的研究相比，可能得出不同的结果。此处，分配序列隐藏，包括充分或不充分，是研究水平的分类特征。Meta回归是亚组分析的扩展，其可以对连续和分类特征的效应进行分析，并原则上可以同时对多个因素的效应进行分析（虽然由于研究数量不够，这几乎

不可能）（Thompson 2002）。在Meta分析中当少于10个研究时，一般不考虑Meta回归。

Meta回归在本质上与简单回归相似，在简单回归中，结局变量可根据一个或更多解释变量的值进行预测。在Meta回归中，结局变量为效应估计值（如MD、RD、log OR或log RR）。解释变量是可能影响干预效应大小的研究特征。其通常被称为“潜在效应修饰因子”或协变量。Meta回归通常与简单回归在两方面存在差异。首先，因为研究通过其各自效应估计值的精确度被赋予权重，因此较大样本的研究对相关性的影响比较小样本的研究大。其次，明智的是保留未由解释变量得到干预效应间的残差异质性。这提出了术语“随机效应Meta回归”，因为附加变异未以与随机效应Meta分析中同样的方式被处理（Thompson 1999）。

从Meta回归分析获得的回归系数将描述结局变量（干预效应）如何随解释变量（潜在效应修饰因子）单位的增加而改变。回归系数的统计学显著性是对干预效应和解释变量间是否有线性关系的检验。如果干预效应为比值指标，在回归模型中应总是使用其对数转换值（见9.2.7节），回归系数的指数将给出随解释变量增加一个单位，干预效应相关改变的估计值。

Meta回归也能用于分析亚组分析中所引入的分类解释变量的差异。如果有J个亚组，则这J个亚组的从属关系在Meta回归模型中（例如在标准线性回归模型）可以使用J-1个哑变量（只能取0和1）来显示。回归系数将估计每个亚组的干预效应如何与指定的参考亚组存在差异。每个回归系数的P值将显示是否该差异有统计学意义。

Meta回归可以使用Stata统计包中的“metareg”宏来完成。

9.6.5 用于亚组分析和Meta回归的研究特征的选择

对于进行亚组分析和对所做分析进行解释，作者需要小心。此处将概述在筛选可能影响干预效应量的特征（也称为解释变量、潜在的效应修饰因子或协变量）时需考虑的一些内容。这些需考虑的内容既适用于亚组分析也适用于Meta回归。更多细节可从Oxman和Guyatt （Oxman 1992）以及Berlin和Antman（Berlin 1994）的文章获得。

9.6.5.1 确保有足够的研究用于亚组分析和 Meta 回归

除非有足够数量的研究，否则异质性分析很难得出有用的结果。对于进行简单回归分析，值得注意以下典型建议：对于每一个特征变量，都至少应获得10个观察结果（如

一个Meta分析中的10个研究）。然而，当协变量分布不均衡时，10个研究甚至都不够。

9.6.5.2 事先确定研究特征

作者应尽可能在计划书中事先确定随后将用于亚组分析或Meta回归的研究特征。事先确定研究特征可减少出现假结果的可能性（首先通过限制分析的亚组数，其次通过防止知道研究结果而影响要分析的亚组）。在系统评价中，真正事先确定很难，因为一些相关研究的结果常常在起草计划书时就知道了。如果在计划书中忽略了一个被外部证据所证实的重要研究特征，于是作者难免会去分析它。然而，像这样的事后分析应该交代。

9.6.5.3 选择少量的研究特征

亚组分析和Meta回归中假阳性结果的可能性会随被分析研究特征数的增加而增加。要对最多可关注多少个研究特征给出建议很困难，尤其事先并不知道可获得的研究数时。如果超过1或2个研究特征被分析，则需要调整显著性水平以对多重比较进行解释。建议寻求统计学家的帮助（见16章，16.7节）。

9.6.5.4 确保对分析的每一个特征都有科学的理论

研究特征的选择应该由生物学和临床假设所启发，理想情况下应有来自纳入研究之外的研究证据所支持。使用不可信或临床不相关的特征的亚组分析毫无用处并应该避免。例如，干预效应与发表年限的关系本身没有临床意义，如果有统计学意义，就会面临这样的危险，即从事后数据挖掘出的因素实际是由于时间的变化而变化。

预后因素是预测疾病或病情结局的因素，而效应修饰因素是影响干预措施对结局影响程度的因素。在计划进行亚组分析时，尤其在计划书阶段，预后因素和效应修饰因素间常常容易混淆。对于亚组分析而言，预后因素并不是理想的选择，除非他们被认为可以调整干预效应。例如，称为吸烟者可能是未来10年内一个强的病死率指标，但其并不是影响病死率药物治疗效果的原因（Deeks 1998）。潜在的效应调整因素可能包括明确的干预措施（活性治疗剂量、比较治疗的选择）、研究如何实施（随访时长）或方法学（设计和质量）。

9.6.5.5 认识到并非总能找出研究特征的效应

许多可能对干预措施的作用程度有重要效应的研究特征不能通过亚组分析或Meta

回归进行分析。虽然这些受试者特征在研究内可能变化很大，但其仅能在研究的水平上进行汇总。一个例子是年龄。考虑收集包括18-60岁的成年人的临床试验。在每个研究内年龄和干预效应之间可能有强相关性。然而，如果试验平均年龄相似，通过观察试验平均年龄和试验水平的效应估计值将无相关性。这是将个体结果整合到一起带来的问题之一，并被称为聚集偏倚、生态学偏倚或生态谬误（Morgenstern 1982，Greenland 1987，Berlin 2002）。

9.6.5.6 思考是否研究特征是否与另一个研究特征密切相关（混杂）

混杂问题是亚组分析和Meta回归的解释错综复杂并可能导致不正确的结论。如果两个特征对干预效应的影响不能被分开，那么他们将被混杂在一起。例如，如果那些实施了加强转胎位术治疗的研究恰好是纳入更严重疾病患者的研究，那么则不能判断哪方面是这些研究和其他研究间效应估计值的任何差异的原因。在Meta回归中，在潜在效应修饰因素间的共线性可导致与Berlin和Antman的文章中讨论的相似的困难。计算研究特征间的相关性将给出关于哪些研究特征可能彼此混淆的一些信息。

9.6.6 亚组分析和Meta回归的解释

亚组分析和Meta回归的恰当解释需要小心。更详细的讨论见Oxman 和Guyatt的文章（Oxman 1992）。

- 亚组比较是观察性的

必须记住亚组分析和Meta回归在本质上完全是观察性的。这些分析调查了研究间的差异。即使个体在一个临床试验内可以被随机分入一组或其他组，但他们不能被随机分入一个试验或另一个试验。因此，亚组分析存在任何观察性研究都有的局限性，包括其他研究水平特征的混杂而带来的偏倚。而且，甚至亚组间的真实差异并不必然归于亚组分类。例如，骨髓移植治疗白血病的亚组分析可能显示同胞供者的年龄与移植成功之间存在强相关性。然而，该可能性并不意味着供者的年龄很重要。事实上，受者的年龄可能是一个关键因素，且亚组结果只是简单地归于受者年龄与其同胞年龄间存在强相关性。

- 分析是事先确定的还是事后确定的？

作者应说明是否亚组分析是事先确定的或在知晓研究结果后才实施（事后分析）。如果一个亚组分析是事先确定的少量分析之一，则其可靠性更高。进行大量的事后亚组

分析以解释异质性是数据捕捞。数据捕捞应避免，因为通过分析大量不同的特征通常可能找到明显但却是错误的异质性解释。

- 有间接证据支持结果吗？

如果亚组间的差异要具有说服力的话，那么这种差异应具有临床合理性并被其他外部或间接证据支持。

- 差异的大小实际重要吗？

如果亚组间的差异大小不能导致对不同的亚组作出不同的推荐意见，那么仅呈现总分析结果可能更好。

- 亚组间有统计学差异吗？

为证实在不同的情况下，是否一种干预措施有不同的效应，应该对不同亚组效应的大小直接进行比较。特别地，不应对不同亚组分析内的结果的统计学显著性进行比较。见9.6.3.1节。

- 亚组分析是针对研究内或研究间的相关性吗？

对患者和干预特征，在研究内观察到的亚组差异比研究层面的亚组分析更可靠。如果这种研究内的关系在研究间得到重复，则其可增加结果的可靠性。

9.6.7 分析基线风险效应

一系列研究间的异质性潜在重要来源之一是当研究间结局时间的潜在平均风险不同时。特定事件的基线风险可以被视作为病例组合因素的综合性指标，如年龄或疾病严重程度。其通常作为每个研究对照组所观察的事件风险（对照组风险（CGR））被测量。在其与临床实践的相关性上，这一概念存在争议，因为基线风险反应了已知和未知风险因素的总和。因为基线风险视随访期长短（其常在研究间不同）而定，问题也随之出现。然而，基线风险已在Meta分析中受到了特别的关注，因为一旦二分类资料被准备用于Meta分析，则信息容易获得。Sharp对该主题作了全面讨论（Sharp 2000）。

直观的看，根据受试者的风险状况，受试者或多或少地可能从一个有效的干预中获益。然而，基线风险与干预效应间的相关性是一个复杂问题。例如，从对所有患者一种干预措施能将事件（如卒中）风险降低到基线风险的80%这个意义上讲，这种干预的获益是相同的。那么从绝对风险差来看，其获益则不同，因为其可将50%的卒中率降低10个百分点到40%（NNT=10），也可以将20%的卒中率降低4个百分点到16%（NNT=25）。

不同汇总统计量（RR、OR、和RD）的使用将证明与基线风险的不同相关性。显示与基线风险几乎没有相关性的汇总统计量通常被首选用于Meta分析（见9.4.4.4节）。

对效应估计值与对照组风险相关性的分析也会被一种被称为向均数回归的技术现象所复杂化。其提出是因为对照组风险是效应估计值不可缺少的一部分。完全因为机遇而观察到的对照组高风险将得到平均比预期更高的效应估计值，反之亦然。这一现象将导致效应估计值和对照组风险间出现假相关性。有方法（要求更高级的软件）可以对向均数回归进行校正（McIntosh 1996，Thompson 1997）。这些方法应用于这类分析，并应考虑统计专家的意见。

9.6.8 剂量-效应分析

Meta回归的原理可用于研究干预效应和剂量（通常被称为剂量-效应）、治疗强度或治疗疗程间的关系（Greenland 1992，Berlin 1993）。如果在一个研究内受试者被随机分到不同剂量组，而得出的由于剂量（或相似因素）差异所致的效应间差异的结论，且在相似研究间也发现了一致的关系的话，则这一结论的强度最高。同时作者应对这些效应加以考虑，尤其作为一种可能的异质性解释，他们在研究间差异的基础上下结论应该小心。作者对于声称不存在剂量-效应关系应该尤其小心，

9.7 敏感性分析

进行系统评价的过程包括一系列决策。同时，许多这些决策都目的清楚并没有争议，而有一些则是随意和不清楚的。例如，如果纳入标准包括数值，值的选择通常是随意的：例如，定义老年人组，可能60、65、70或75岁，或其之间的任何值都是合理的。因为一个研究报告未能纳入所需信息，一些决策可能不清楚。一些决策因为纳入研究自身未获得所需信息而不清楚：例如，那些不幸失访者的结局。更多的决策因为不能确定用于特定问题的最佳统计方法而不清楚。

我们希望证明系统评价的结果并不是依赖于这种随意或不清楚的决策。敏感性分析是对最初分析或Meta分析再次分析，来取代随意或不清楚的备选决策或决策的价值范围。例如，如果Meta分析中一些研究的合格性因为其并未包含足够的细节而值得怀疑，敏感性分析可以是进行两次Meta分析：首先，纳入所有研究；其次，仅纳入那些明确被认为

合格的研究。敏感性分析会问：结果对获取他们的过程中所做的决策足够强吗？

在系统评价过程中有许多可能需要进行敏感性分析的决策节点。例子包括：

研究检索

- 摘要（其结果不能在随后发表的全文中得以证实）应被纳入系统评价吗？

纳入标准

- 受试者特征：在一个研究中大部分而不是所有人都满足年龄范围，该研究应被纳入吗？
- 干预措施的特征：Meta分析中应纳入什么剂量范围？
- 对照特征：对于定义用于对照组的常规治疗，需要什么标准？
- 结局特征：什么时点或时点范围适合纳入？
- 研究设计：应包括盲法和非盲法的结局评估吗，或研究纳入应该被方法学标准的其他方面严格限制吗？

应分析什么数据？

- 时间-事件数据：如何假设截尾数据的分布？
- 连续性数据：在标准差缺失之处，他们应在何时和怎样被处理？
- 有序度量单位：用于将短的有序度量单位分入两组的截点是什么？
- 整群随机试验：当试验分析为进行群集性调整时，应使用什么组内相关系数值。
- 交叉试验：当在原始报告内不能得到受试者内相关系数时，其取值应为多少？
- 所有分析：为进行ITT分析，关于缺失结局应做什么假设？应该调整还是不调整所用治疗效应的估计值？

分析方法：

- 应使用固定效应还是随机效应方法进行分析？
- 对于二分类结局，应使用OR、RR还是RD？
- 对于连续性结局，在几个度量单位都评价了同一内容时，其结果应以涵盖所有度量单位的标准化均数差进行分析还是对每个度量单位以均数差单独进行分析。

一些敏感性分析可在研究计划书中事先确定，但许多适合进行敏感性分析的问题仅能在系统评价过程中（确定所分析研究的个体特点时）被确定。当敏感性分析显示总体结果和结论不受可能在系统评价过程中所做的不同决策影响时，系统评价结果可被认为有较高程度的肯定的。当敏感性分析找到了能够在很大程度上影响系统评价结果的特定决策或缺失信息时，更多的资源可被用于尝试并解决不确定性，并获得额外信息（可能

通过联系试验作者和获得个体患者数据）。如果做不到，结果必须谨慎恰当解释。这样的结果对进一步的分析和未来的研究给出建议。

系统评价中敏感性分析的最佳报告方式为制作一张汇总表。对于所做的每个敏感性分析都制作一个森林图几乎没有必要。

敏感性分析优势会和亚组分析相混淆。虽然一些敏感性分析包括只针对一个亚组的全部研究进行分析，但两种方法在两方面存在差异。首先，敏感性分析不会对从分析中排除研究的干预效进行评估，而在亚组分析中，对每个亚组都要进行评估。其次，在敏感性分析中，是在估计同一事情的不同方法间进行非正式比较；而在亚组分析中，是在所有亚组间进行正式的统计比较。

9.8 本章信息

编辑：Jonathan J Deeks、Julian PT Higgins和Douglas G Altman，代表Cochrane统计方法组。

本章引用方式：Deeks JJ, Higgins JPT, Altman DG (editors). Chapter 9: Analysing data and undertaking Meta-analyses. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

做出贡献的作者：Doug Altman、Deborah Ashby、Jacqueline Birks、Michael Borenstein、Marion Campbell、Jon Deeks、Matthias Egger、Julian Higgins、Joseph Lau、Keith O'Rourke、Rob Scholten、Jonathan Sterne、Simon Thompson和Anne Whitehead。

致谢：我们感谢以下对我们的初稿给予有益批评的人：Bodil Als-Nielsen、Doug Altman、Deborah Ashby、Jesse Berlin、Joseph Beyene、Jacqueline Birks、Michael Bracken、Marion Campbell、Chris Cates、Wendong Chen、Mike Clarke、Albert Cobos、Esther Coren、Francois Curtin、Roberto D'Amico、Keith Dear、Jon Deeks、Heather Dickinson、Diana Elbourne、Simon Gates、Paul Glasziou、Christian Gluud、Peter Herbison、Julian Higgins、Sally Hollis、David Jones、Steff Lewis、Philippa Middleton、Nathan Pace、Craig Ramsey、Keith O'Rourke、Rob Scholten、Guido Schwarzer、Jack Sinclair、Jonathan Sterne、Simon Thompson、Andy Vail、Clarine van Oel、Paula Williamson和Fred Wolf。

框9.8.a　Cochrane统计方法组

统计学问题是Cochrane协作网多数工作的核心内容。统计学方法组（SMG）是与Cochrane协作网的工作相关的统计学问题讨论的论坛。其涵盖领域广泛，包括与统计方法、培训、软件和研究相关的问题。其也致力于确保各系统评价小组能得到充分的统计学和技术支持。

SMG的历史可追溯到1993年。现在要成为SMG的成员需先成为工作组Email讨论区的成员。该讨论区被用于讨论所有工作组的重要问题，不论研究、培训、软件或管理。工作组有来自20多个国家的超过130位成员。极力鼓励所有与各Cochrane系统评价工作组（CRGs）一起工作的统计学家都加入SMG。

特别地，该工作组旨在：

1. 为协作网就与卫生保健干预系统评价相关的所有统计学问题制定普通政策建议。
2. 为本手册与统计学相关的章节负责。
3. 为CRG在工作中提供统计学支持。
4. 就新出现的必要的主题举办培训班和研讨会。
5. 撰写和审阅协作网内提供的培训材料的统计内容。
6. 开发和验证协作网内使用的统计软件。
7. 生产和更新统计学方法组的清单，细化其感兴趣和专业的领域，并以讨论相关方法学问题论坛的形式建设一个Email讨论区。
8. 针对协作网目前和将来功能的重要问题制定研究计划，并鼓励有意解决这些问题的研究。

网址：www.cochrane-smg.org

9.9　参考文献

Adams 2005

Adams NP, Bestall JB, Malouf R, Lasserson TJ, Jones PW. Beclomethasone versus placebo for chronic asthma. Cochrane Database of Systematic Reviews 2005, Issue 1. Art No: CD002738.

Agresti 1996

Agresti A. An introduction to categorical data analysis. New York (NY): John Wiley & Sons, 1996.

Altman 1996

Altman DG, Bland JM. Detecting skewness from summary information. BMJ 1996; 313: 1200-1200.

Antman 1992

Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of Meta-analyses of randomized control trials and recommendations of clinical experts: Treatments for myocardial infarction. JAMA 1992; 268: 240-248.

Berlin 1993

Berlin JA, Longnecker MP, Greenland S. Meta-analysis of epidemiologic dose-response data. Epidemiology 1993; 4: 218-228.

Berlin 1994

Berlin JA, Antman EM. Advantages and limitations of metaanalytic regressions of clinical trials data. Online Journal of Current Clinical Trials 1994; Doc No 134.

Berlin 2002

Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman KA, Anti-Lymphocyte Antibody Induction Therapy Study Group. Individual patient- versus group-level data Meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Statistics in Medicine 2002; 21: 371-387.

Borenstein 2008

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-analysis. Chichester (UK): John Wiley & Sons, 2008.

Bradburn 2007

Bradburn MJ, Deeks JJ, Berlin JA, Russell LA. Much ado about nothing: a comparison of the performance of Meta-analytical methods with rare events. Statistics in Medicine 2007; 26: 53-77.

Chinn 2000

Chinn S. A simple method for converting an odds ratio to effect size for use in Meta-analysis. Statistics in Medicine 2000; 19: 3127-3131.

Cooper 1980

Cooper HM, Rosenthal R. Statistical versus traditional procedures for summarizing research findings. Psychological Bulletin 1980; 87: 442-449.

Crawford 2007

Crawford F, Hollis S. Topical treatments for fungal infections of the skin and nails of the feet. Cochrane Database of Systematic Reviews 2007, Issue 3. Art No: CD001434.

Deeks 1998

Deeks JJ. Systematic reviews of published evidence: Miracles or minefields? Annals of Oncology 1998; 9: 703-709.

Deeks 2001

Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in Meta-analysis. In: Egger M, Davey Smith G, Altman DG (editors). Systematic Reviews in Health Care: Meta-analysis in Context (2nd edition). London (UK): BMJ Publication Group, 2001.

Deeks 2002

Deeks JJ. Issues in the selection of a summary statistic for Meta-analysis of clinical trials with binary outcomes. Statistics in Medicine 2002; 21: 1575-1600.

DerSimonian 1986

DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986; 7: 177-188.

Egger 1997

Egger M, Smith GD, Schneider M, Minder C. Bias in Meta-analysis detected by a simple, graphical test. BMJ 1997; 315: 629-634.

Engels 2000

Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in Meta-analysis: an empirical study of 125 Meta-analyses. Statistics in Medicine 2000; 19: 1707-1728.

Greenland 1985

Greenland S, Robins JM. Estimation of a common effect parameter from sparse follow-up data. Biometrics 1985; 41: 55-68.

Greenland 1987

Greenland S. Quantitative methods in the review of epidemiologic literature. Epidemiologic Reviews 1987; 9: 1-30.

Greenland 1992

Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to Meta-analysis. American Journal of Epidemiology 1992; 135: 1301-1309.

Hasselblad 1995

Hasselblad VIC, McCrory DC. Meta-analytic tools for medical decision making: A practical guide. Medical Decision Making 1995; 15: 81-96.

Higgins 2002

Higgins JPT, Thompson SG. Quantifying heterogeneity in a Meta-analysis. Statistics in Medicine 2002; 21: 1539-1558.

Higgins 2003

Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in Meta-analyses. BMJ 2003; 327: 557-560.

Higgins 2004

Higgins JPT, Thompson SG. Controlling the risk of spurious findings from Meta-regression. Statistics in Medicine 2004; 23: 1663-1682.

Higgins 2008a

Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects Meta-analysis. Journal of the Royal Statistical Society Series A (in press, 2008).

Higgins 2008b

Higgins JPT, White IR, Anzures-Cabrera J. Meta-analysis of skewed data: combining results reported on log-transformed or raw scales. Statistics in Medicine (in press, 2008).

Kjaergard 2001

Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in Meta-analyses. Annals of Internal Medicine 2001; 135: 982-989.

Laupacis 1988

Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. New England Journal of Medicine 1988; 318: 1728-1733.

Mantel 1959

Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute 1959; 22: 719-748.

McIntosh 1996

McIntosh MW. The population risk as an explanatory variable in research synthesis of clinical trials. Statistics in Medicine 1996; 15: 1713-1728.

Moher 2005

Moher M, Hey K, Lancaster T. Workplace interventions for smoking cessation. Cochrane Database of Systematic Reviews 2005, Issue 2. Art No: CD003440.

Morgenstern 1982

Morgenstern H. Uses of ecologic analysis in epidemiologic research. American Journal of Public Health 1982; 72: 1336-1344.

O'Rourke 1989

O'Rourke K, Detsky AS. Meta-analysis in medical research: strong encouragement for higher quality in individual research efforts. Journal of Clinical Epidemiology 1989; 42: 1021-1026.

Oxman 1992

Oxman AD, Guyatt GH. A consumers guide to subgroup analyses. Annals of Internal Medicine 1992; 116: 78-84.

Pittler 2003

Pittler MH, Ernst E. Kava extract versus placebo for treating anxiety. Cochrane Database of Systematic Reviews 2003, Issue 1. Art No: CD003383.

Poole 1999

Poole C, Greenland S. Random-effects Meta-analyses are not always conservative. American Journal of Epidemiology 1999; 150: 469-475.

Sackett 1996

Sackett DL, Deeks JJ, Altman DG. Down with odds ratios! Evidence Based Medicine 1996; 1: 164-166.

Sackett 1997

Sackett DL, Richardson WS, Rosenberg W, Haynes BR. Evidence-Based Medicine: How to Practice and Teach EBM. Edinburgh (UK): Churchill Livingstone, 1997.

Sharp 2000

Sharp SJ. Analysing the relationship between treatment benefit and underlying risk: precautions and practical recommendations. In: Egger M, Davey Smith G, Altman DG (editors). Systematic Reviews in Health Care: Meta-analysis in Context (2nd edition). London (UK): BMJ Publication Group, 2000.

Sinclair 1994

Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. Journal of Clinical Epidemiology 1994; 47: 881-889.

Thompson 1997

Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in Meta-analysis. Statistics in Medicine 1997; 16: 2741-2758.

Thompson 1999

Thompson SG, Sharp SJ. Explaining heterogeneity in Meta-analysis: a comparison of methods. Statistics in Medicine 1999; 18: 2693-2708.

Thompson 2002

Thompson SG, Higgins JPT. How should Meta-regression analyses be undertaken and interpreted? Statistics in Medicine 2002; 21: 1559-1574.

Whitehead 1991

Whitehead A, Whitehead J. A general parametric approach to the Meta-analysis of randomised clinical trials. Statistics in Medicine 1991; 10: 1665-1677.

Whitehead 1994

Whitehead A, Jones NMB. A Meta-analysis of clinical trials involving different classifications of response into ordered categories. Statistics in Medicine 1994; 13: 2503-2515.

Yusuf 1985

Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomised trials. Progress in Cardiovascular Diseases 1985; 27: 335-371.

Yusuf 1991

Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991; 266: 93-98.

（杜亮译，岑啸、贾鹏丽、王霁初审）

第十章　论述报告偏倚

编辑：Cochrane 偏倚方法学组 Jonathan AC Sterne、Matthias Egger、David Moher。

本文选自工作手册5.0.1版本。有关如何引用它的指南，见10.5节。这些材料还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 仅部分研究方案将会发表在系统评价作者能够容易查到的出版物上。一旦研究结果的传播受其结果属性和方向的影响，就会产生报告偏倚。
- 结果无统计学意义的原始研究对系统评价证据的贡献与结果有统计学意义的原始研究一样重要。
- 有说服力的存在多种报告偏倚（本章概述）的证据表明，确有必要全面查找符合 Cochrane 系统评价纳入标准的原始研究。
- 前瞻性试验注册现已成为众多杂志发表文章的一项要求，这条规定可能会从根本上减少发表偏倚的影响。

- 漏斗图可用于纳入足够数量原始研究的系统评价，但如漏斗图不对称，并不一定等同于存在发表偏倚。
- 有些方法可检验漏斗图是否不对称，本章将提供参考意见，帮助选择合适的检验方法。

10.1 引言

研究结果的传播，并非按照是否公开发表来划分；相反，研究结论公布形式广泛，包括与同事分享论文草稿、在会议上的介绍、或公开发表的摘要、或由各大书目数据库收录的杂志发表的论文（Smith，1999）。众所周知，长期以来仅有部分研究项目最终能在收入数据库的某期刊上公开发表，进而容易被系统评价查到。

一旦研究结论的传播受种种结果自身特性及其指向的影响，就会产生报告偏倚。那些表明某种干预措施有效的有统计学意义的“阳性”结果，更可能被发表，发表起来也更便捷、更有可能以英语发表、更有可能不止一次被发表、更有可能在高影响力的杂志上发表进而更可能被他人引用。无显著性结果的研究对系统评价证据完整性的贡献其实与有显著性结果的研究所作贡献一样重要。

表10.1.a总结了一些不同类型的报告偏倚，10.2节将予以详细介绍，届时我们会专门提供证据，证明这些报告偏倚确实存在。10.3部分，我们会探讨种种在Cochrane系统评价里规避报告偏倚的办法；10.4部分将介绍漏斗图及其它查明潜在偏倚的统计学方法。尽管为了探讨这些偏倚，我们有时称有统计学意义（P＜0.05）的结果为“阳性”结果、称统计学差异不明显或无效结果为“阴性”结果，但Cochrane系统评价作者不宜采用这类说法。

表10.1.a　部分报告偏倚的定义

报告偏倚类型	定义
发表偏倚	由研究结果自身特性及其指向影响的已发表/未发表的研究结论
时滞偏倚	受研究结果自身特性及其指向影响，导致研究结论发表过快或延期发表
重复发表（二次发表）偏倚	受研究结果自身特性及其指向影响，研究结论重复发表或单次发表
检索偏倚	受研究结果自身特性及其指向影响，发表于标准数据库收入杂志上的研究结论，这些数据库检索权限、编录等级各不相同
引用偏倚	受研究结果自身特性及其指向影响，经引用/未经引用的研究结论
语言偏倚	受研究结果自身特性及其指向影响，用某种语言发表的研究结论
结局报告偏倚	受研究结果自身特性及其指向影响，选择报告部分结局、而不报告其它结局

10.2　报告偏倚的类型及支持证据

10.2.1　发表偏倚

在一篇1979年的文章里“‘文件抽屉问题（即因未收集到研究领域内相当数量未发表的、通常为阴性结果的研究而造成的偏倚）’及对无效结果的容忍”，Rosenthal介绍了令人沮丧的现状，即“各种学术期刊上有5%的研究存在Ⅰ型错误；另一方面，实验室的文件抽屉里有95%的研究结果无统计学意义（如P＞0.05）”（Rosenthal，1979年）。此前，在社会科学领域早就对文件抽屉问题有过质疑，一篇对几本心理学杂志的综述发现，20世纪50年代发表的294项研究中，有97.3%在5%的检验水平（P＜0.05）拒绝零假设（Sterling，1959）。加入3本杂志（《新英格兰医学杂志》、《美国流行病学杂志》及《美国公共卫生杂志》）后，该研究进行了进一步更新（Sterling，1995）。这些年来，心理学杂志发表情况几乎毫无变化（95.6%的研究报道有统计学意义的结果），全科医学、公共卫生类杂志中有统计学意义的报告比例也相当之高（占85.4%），其它众多医学领域报告的结果也与之相似，如急诊医学（(Moscati，1994）、补充和替代医学（Vickers，1998，Pittler，2000）和急性卒中试验（Liebeskind，2006）。

提示某种干预措施确有疗效、或认为其疗效更佳的研究能被发表，而结论相反的类似资料则未能发表，这是有可能的。这种情况下，仅纳入已发表研究的系统评价，可能

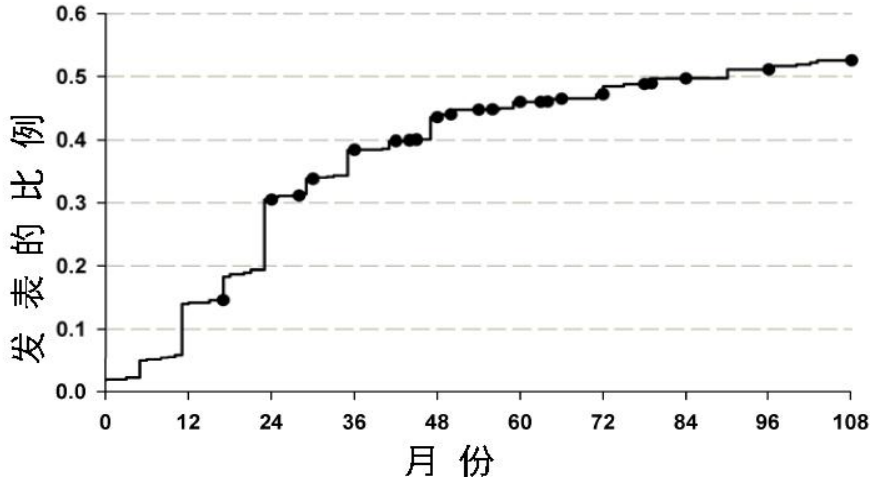
会得出某种干预假的有益效果，或漏掉干预措施某些重要的不良反应。以心血管类药物为例，1980年，有研究人员发现，急性心肌梗死患者经一类抗心律失常药治疗，死亡率增加，结果他们将其视为偶然发现而漏掉了，当时并未将其试验结果发表出来（Cowley，1993）。他们结论本来能够有助于及时发现后来被证实的与使用一类的抗心律失常药相关的死亡率增加的问题（Teo，1993；CLASP Collaborative Group，1994）。（CLASP Collaborative Group全称Collaborative Low-dose Aspirin Study in Pregnancy，即低剂量阿司匹林对怀孕研究联合工作组）

从经验上讲，用于检查发表偏倚的研究一般可分为两类：间接证据、直接证据。因为不知道在所有假设检验中，零假设的确是假的比例有多少，故如上所述的对已发表结果进行的调查仅能提供关于发表偏倚的间接证据。同样存在发表偏倚的大量直接证据，Roberta Scherer及其同事最近更新了一篇系统评价，纳入了79个研究，这些研究介绍了后来完整发表、而开始仅以摘要形式、或短篇报道形式发表的一些研究（Scherer，2007）。图10.2.a总结45个研究的数据，这些研究交代了发表时间的数据。会议提交的摘要中仅半数后来得以全文发表（而随机试验为63%），而这些后续发表多涉及阳性结果（Scherer，2007）。

另一方面，直接证据可从大量提交给伦理委员会和机构审查委员会的队列研究标书的数量（Easterbrook，1991；Dickersin，1992；Stern，1997；Decullier，2005；Decullier，2007）、向注册单位申请的试验（Bardy，1998）、对试验注册中心进行的分析（Simes，1987）、或由专门资助机构资助的前瞻性研究（Dickersin，1993）等资料中得到。就上述各类研究，数年后会联系主要研究人员，以了解各项已完成试验的发表情况。这些研究中，更有可能发表干预措施疗效佳、且有统计学意义的研究。

Hopewell等近来完成一篇关于上述研究的方法学综述，仅限于被认为是临床试验的那些研究（Hopewell，2008）。该系统评价里所纳入的5个研究，以全文形式发表的比例从36%到94%不等（表10.2.a）。较之阴性结果，阳性结果向来更有可能被发表。如图10.2.b示，若结果有统计学意义，则能发表的可能性约大4倍（OR=3.90，95%置信区间（2.68，5.68））。其它因素如研究规模、资助来源、学科排名、主要研究人员的性别并不一定与能否发表有关，也不太可能单独用来评价临床试验（Hopewell，2008）。

图10.2.a　45个研究中以会议摘要提交、后经全文发表的时间-累计全文发表率



月份	0	12	24	36	48	60	72	84	89	108
#已经全文发表数	362	2460	3,348	15,19	800	280	282	84	27	10
#未经全文发表数	20227	19091	16313	10758	9032	6518	4030	1803	1352	246

N（**发表总量**）=20,227个摘要

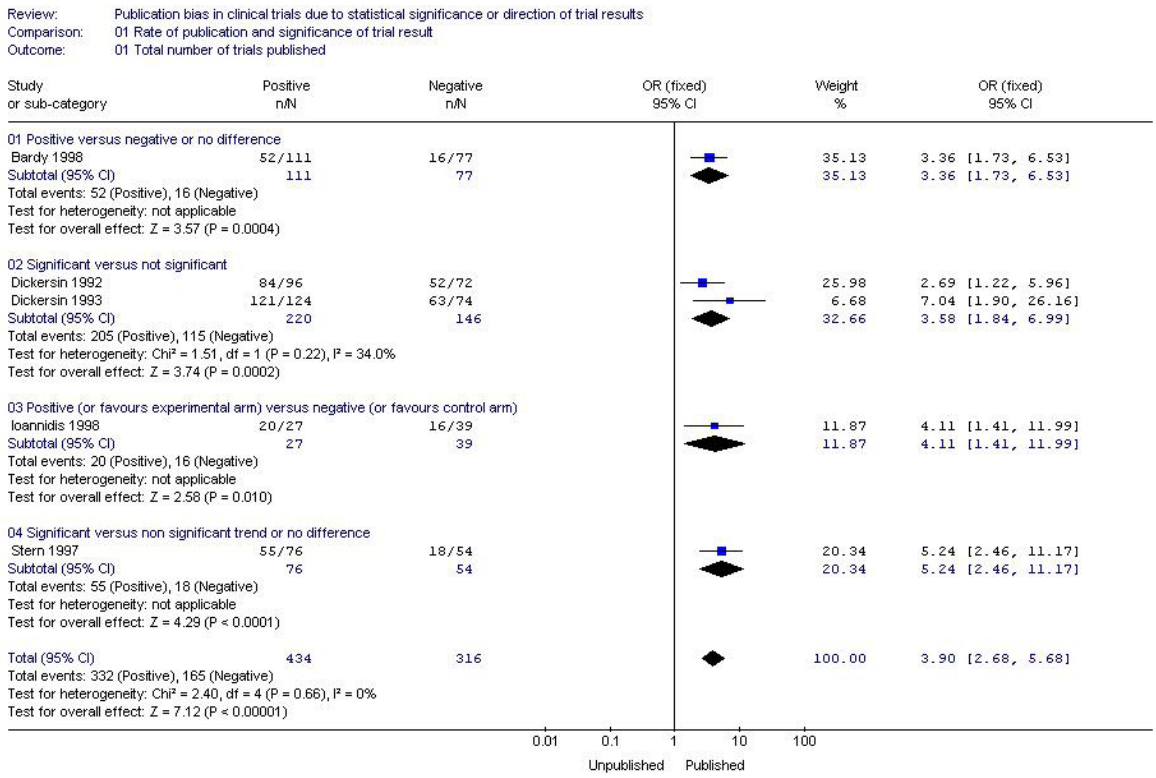
经圆圈标注点表示因报道终止随访而截尾的数据

表10.2.a　通过伦理审查委员会/机构审查委员会审查的5个队列研究标书，这些标书随访期内已完成计划及分析。（摘自Hopewell等（Hopewell，2008））

	巴的摩尔Johns Hopkins大学	美国国家卫生研究院	悉尼皇家Alfred王子医院	芬兰国家药事局	美国国家卫生研究院HIV/AIDS多中心试验
参考	Dickersin，1992年	Dickersin，1993年	Stern，1997年	Bardy，1998年	Ioannidis，1998年
审核通过时间	1980年	1979年	1979-1980年	1987年	1986-1996年
随访年份	1988年	1988年	1992年	1995年	1996年
审核通过标书数量	168	198	130	188	66
发表数	136（81%）	184（93%）	73（56%）	68（36%）	36（54%）
阳性报告*	84/96（87%）	121/124（98%）	55/76（72%）	52/111（47%）	20/27（75%）
阴性报告*	52/72（72%）	63/74（85%）	3/15（20%）	5/44（11%）	16/39（41%）
未包括/空白报告（如系独立评价）	未予评价	未予评价	15/39（38%）	11/33（33%）	未予评价

*不同研究定义不同

图 1 0.2.b　由试验结果统计学意义或试验结果的导向在临床试验中产生的发表偏倚（摘自Hopewell，等（Hopewell，2008年））



10.2.1.1　时滞偏倚

待伦理委员会通过立项多年后，研究才陆续地付梓发表。Hopewell及其同事对调查临床试验得出结果到文章发表的时间的研究进行了综述，（Hopewell，2007a）。该系统评价中的两篇研究（Stern，1997；Ioannidis，1998）发现，约半数试验以阳性结果发表，且这些阳性结果试验往往比无效结果、阴性结果试验早两三年发表。

在向澳大利亚悉尼皇家Alfred王子医院伦理委员会提交申请的标书中，约85%阳性结果研究10年内得以发表，同期仅65%的无效结果研究最终发表（Stern，1997）。阳性结果研究发表的中位时间为4.7年，阴性结果/无效结果则为8.0年。与此相似的，在美国，HIV感染领域由多中心试验组进行的试验，如果试验结果有统计学意义，则从纳入患者算起，平均经过4.3年即可发表；若试验结果呈阴性，则需等上6.5年才能发表（Ioannidis，1998）。最近研究也得到类似结果（Decullier，2005）。实际上相当多研究即使已经完成试验、也进行了分析，但10年后还是未发表，这种情况比较麻烦，因为对系统评价者、系统评价读者来说，可能重要的信息依然不得而知。

Ioannidis和同事还发现，在完成随访耗时上，阳性结果试验和阴性结果试验差之甚少（Ioannidis，1998）。相反，时滞是由从试验完成到发表之间的用时决定的（Ioannidis，1998）。上述结论表明，即使绝大多数、甚或全部研究终以发表，系统评价可能还是会带入时滞偏倚。阳性结果的研究将在文献中占主要地位，且在若干年里引入偏倚，直到阴性（但和阳性结果一样重要的）结果最终发表为止。此外，较之短期的有利效应，罕见的不良事件更可能在研究的过程中被发现。

10.2.1.2 谁来为发表偏倚负责？

阴性结果研究仍未发表，可能因为作者没有起草投稿，或者这类研究不太受同行评审的青睐，还可能系期刊编辑不想发表阴性结果的研究。同行评审有时并不可信，易受主观性、偏倚及利益冲突的影响（Peters，1982年；Godlee，1999）。那些检验提交给同行评审或杂志的手稿的实验性研究表明，同行评审可能更倾向于评判那些和他们自己观点一致的结果（Mahoney，1977；Epstein 1990；Ernst，1994）。例如，让选出来的一群作者同行评审一篇关于经皮电神经刺激（TENS）的虚构文章，他们会受自己的发现及个人知识的影响。其它研究发现，投稿发表和研究结果间没有联系（Abbot，1998；Olson，2002），这一结论表明，尽管同行评审他们可能持有能够影响他们评价的强大的个人信念，但是不存在支持或反对阳性结果的总体偏倚。许多研究曾直接问过作者，为什么他们没有将研究结果发表。最常见的回答是，他们没足够的兴趣认为这值得发表（比如杂志不太可能接受这类投稿）（Easterbrook，1991；Dickersin，1992；Stern，1997；Weber，1998；Decullier，2005），或者调查者没足够的时间撰写文稿（Weber，1998；Hartling，2004）。很少会以杂志拒稿作为研究结果未发表理由。发表偏倚主要取决于作者选择性投稿，而非同行评审选择性荐稿、或编辑选择性录用。另外，Dickersin等研究了从投稿（JAMA杂志）到全文发表间的时间，发现这段时间和研究中任何研究特点、包括研究结果的统计学显著性没有任何联系（Dickersin，2002）。所以，时滞偏倚也可能是作者延期投稿的结果，而非杂志延期发表的结果。

10.2.1.3 外部资助及商业利益的影响

现已发现，外界资助和有统计学显著性结果的研究的发表存在独立的关联（Dickersin，1997）。在提交给伦理委员会的3个队列研究的标书中，政府机构的资助明显和发表有关（Easterbrook，1991；Dickersin，1992 ；Stern 1997），而在另两个研究中

发现，受药厂资助的研究能发表的可能性则不大（Easterbrook，1991；Dickersin，1992）。实际上，大部分由制药公司向注册机构提交的临床试验尚未发表（Hemminki，1980；Bardy，1998）。

在一篇系统评价中，Lexchin等纳入了1966-2002年间发表的30个研究，这些研究调查了由药厂资助的药物研究是否和其结果对赞助方有利有关。他们发现，较之由其它方资助的研究，制药公司资助的研究更不易于发表，且制药公司资助的研究结果比其它方资助的研究更有利于资助者（Lexchin，2003）。其它研究也发现了这些联系，得出类似结果（Bhandari，2004年；Heres，2006）。在一个抗精神病类药物头对头比较的研究中，Heres等发现，在90%已评价的研究中，试验的总体结果都支持由资助方生产的药品（尽管一些类似研究得出相反结论），这些研究中每个都支持试验赞助方的产品（Heres，2006）。

上述结论暗示，药业方会倾向于阻止受其资助的阴性结果研究发表。比如，某文报道一项比较通用的左旋甲状腺素与品牌左旋甲状腺素两者生物等效性的试验，该试验未能取得试验资助方Boots药业公司想要的结果，于是在Boots药业向大学校方、研究人员起诉后中止了。Boots药业这一做法换来的是该研究的论文（Dong，1997）延期了约7年才发表，对此JAMA杂志编辑Drummond Rennie有过详细介绍（Rennie，1997）。在美国一个针对生命科学教职员工的全国调查中，20%受访人员称，他们曾延期超过6个月才发表其工作，他们解释为何没有发表的理由中包括“延期发布不想要的结果”（Blumenthal，1997）。论文延期发表和商业化、学术-产业研究关系的介入有关，还和男性调查人员及调查人员的学术排名有关（Blumenthal，1997）。

10.2.2 其它报告偏倚

尽管很早就认识到发表偏倚存在且充分讨论，而其它因素也会影响系统评价，使其引入有偏倚的研究。事实上，在已发表的研究中，找出与Meta分析相关的研究的可能性还受这些研究结果影响。对这些偏倚的考虑远少于发表偏倚，然他们所致的后果很可能同等重要。

10.2.2.1 重复（多次）发表偏倚

1989年，Gøtzsche发现，在244个比较非甾体类抗炎药治疗类风湿性关节炎的试验的

报告中，44个（18%）系冗余的重复发表，和之前发表的文章基本重复。20个试验重复发表两次，10个试验重复3次，还有1个试验达4次（Gøtzsche，1989）。由同一研究得到的多个重复发表会从若干方面带来偏倚（Huston，1996）。重要的是，阳性结果研究更有可能造成多次发表和报告（Easterbrook，1991），这就使得研究更有可能被找到并纳入Meta分析。来自同一研究的重复发表，以及一群研究受试对象在同一分析中纳入两次，这种情况并非总是显而易见。纳入重复发表的数据可能会高估干预措施疗效，研究昂丹司琼预防术后恶心、呕吐疗效的研究就是例证（Tramèr，1997）。

当多中心试验在多个刊物上发表时，有作者描述了因研究的冗员和“分散”所带来的（对于重复发表判断上）困难和挫折（Huston，1996；Johansen，1999）。重复发表彼此间往往不互相援引（Bailey，2002；Barden，2003）；而两文报道同一试验、但作者却无一相同，这样的例子也不是没有（Gøtzsche，1989；Tramèr，1997）。这样的话，如果不联系原文作者，系统评价作者将很难、甚至无法确定两篇重复发表论文是来自同一研究、还是两个不同的研究，从而对使用该数据进行的Meta分析造成偏倚。

10.2.2.2 检索偏倚

有研究认为，与研究结果可获得性有关的各种因素与各个试验效应量的大小有联系。比方说，在一组补充替代医学领域的试验中，Pittlter和同事研究了试验结果、方法学质量及样本量三者与发表这些试验的杂志的特点之间的关系（Pittlter，2000）。他们发现，在低影响力或无影响力杂志上发表的试验比那些在具有高影响力的主流医学杂志上发表的试验更有可能报告阳性结果，同时试验的质量和刊发杂志也有关系。与此类似，一些研究认为，英文杂志发表的试验比非英文杂志发表的试验更有可能给出强阳性的结果（Egger，1997b），但是这一结论并非一直如此（Moher，2000；Jüni，2002；Pham，2005）。

“检索偏倚”用来表示基于电子数据库不同的标引方式下，研究的可获得性。从临床问题出发，选择检索哪些数据库，这可能对Meta分析效果估计造成偏倚。例如，有研究发现，未经MEDLINE（即美国国立医学文献数据库）收录的杂志发表的试验可能比MEDLINE收录杂志刊载的试验显示的疗效可能更好（Egger，2003）。对另一含61个Meta分析的研究发现，一般来讲，由EMBASE（即荷兰医学文摘数据库）收录、但MEDLINE未收录的杂志所发表的试验较之MEDLINE收录杂志发表的试验，其疗效的估计值偏小；但考虑到仅EMBASE里收录的试验并不普遍，这一偏倚的风险可能很小（Sampson，2003）。如上所述，这些结论可能随这正在研究的临床问题而彻底改变。

检索偏倚的一种最终形式即是地区偏倚或称为发达国家偏倚。支持这种偏倚存在证据的研究表明，在某些国家发表的研究可能比其它国家发表的研究更易得出干预措施有明显疗效。Vickers及同事证实了这种偏倚有可能存在（Vickers，1998）。

10.2.2.3 引文偏倚

细读文章参考文献目录是一种广泛使用的用来发现更多相关研究的方法，尽管还没有证据来支持这种方法。这种方法存在的问题是，援引既往文章缺乏客观性，并且通过浏览参考文献列表检索文献可能得到一个有偏倚的样本。引用一篇文章可能出于很多目的。Brooks曾采访来自Iowa大学众多教职员工中的学术作者，问他们在自己最近一篇文章引用每个参考文献的理由（Brooks，1985）。说服力即让同行信服、证明他们自己的观点，成为引用其他文献一个最重要的原因。Brooks认为，作者为宣扬其个人主张，并用其它文献来证明自己的观点，“这些作者可以视为知识界他们个人主张的信徒，总是搜罗文献来佐证。”（Brooks，1985）

Gøtzsche对非甾体类抗炎药治疗类风湿性关节炎试验的分析证明，新药疗效更优的试验比阴性结果的试验更容易被引用（Gøtzsche，1987）。另一篇对肝胆疾病随机试验的分析也得出类似结果（Kjaergard，2002）。类似的，如果降低胆固醇以预防冠心病的试验结果支持降低胆固醇，则其受引用的概率往往约大6倍（Ravnskov，1992）。另外也会发生过度引用不支持的研究。Hutchison等调查了肺炎球菌疫苗效果的系统评价，发现不支持其有效的试验比表明疫苗有效的试验更有可能被引用（Hutchison，1995）。

引文偏倚可能会影响“二次”文献。例如，ACP杂志俱乐部旨在总结原始文献和系统评价，以便会员医师与最新证据保持同步。但是，Carter等发现，在控制了其它的选择原因后，阳性结局试验更易于被总结（Carter，2006）。若阳性结果试验更容易被引用，则它们更有可能被检索到继而更有可能被纳入系统评价，最后使系统评价的结果产生偏倚。

10.2.2.4 语言偏倚

系统评价常常只基于英文发表的研究。比如，1991-1993年间在主要英文的全科医学杂志上报道的36个Meta分析中，26个（72%）将其检索文献限于以英文报道的研究（Grégoire，1995）。随着最近一个针对300篇系统评价再评价的发表，这种趋势将有所改变；这篇再评价发现：约16%的系统评价仅纳入英文发表的试验；纸质版杂志发表的

系统评价比Cochrane系统评价更有可能将检索限制在以英文发表的试验（Moher，2007）。此外，对于关注治疗的系统评价，Cochrane系统评价比非Cochrane系统评价更有可能报告无语言的限制（两者比例依次为62%、26%）（Moher，2007）。

非英语国家的研究人员会在本国杂志上发表其部分研究（Dickersin，1994）。可以想象，如果研究结果呈阳性，作者更愿意在国际性英文杂志上报道；反之，结果呈阴性，则更有可能在本国杂志上发表。德语文献即是例证（Egger，1997b）。

仅限于英文报道的系统评价可能引入偏倚（Grégoire，1995；Moher，1996）。但针对这一问题相关研究的结果存在冲突。在一个包括50篇采用全面检索策略的系统评价的研究中，包括英文试验与非英文试验，Jüni等报道，非英文试验更有可能得到阳性结果（P<0.05），且非英文试验干预措施更有益且效果估计值比英文试验平均多16%，置信区间95% CI（3%，26%）（Jüni，2002）。相反，Moher与同事在两个针对Meta分析的研究中调查了纳入和排除英文试验的影响，他们发现，总体上，排出非英文报道的试验对Meta分析结果的影响不显著（Moher，2003）。仅分析传统的医学试验的系统评价也得到类似结果。然而，在针对补充替代医学试验单独进行Meta分析时，则在排出非英文的报道后，Meta分析效应量明显降低（Moher，2003）。

由于近来英文发表的研究有些变化，语言偏倚的程度、影响可能有所减少。2006年，Galandi等报道，在德语医疗保健杂志上发表的随机试验数量明显减少：1999年以后，每本杂志上发表的随机试验少于两个（Galandi，2006）。尽管Meta分析中非英文发表研究的潜在影响可能很小，但还是很难预测，在何种情况下，这种排出将给系统评价造成偏倚。系统评价作者可能想要进行不限制语言的检索，并且在可能需要在个案的基础上纳入做出纳入非英文报告的研究。

10.2.2.5 结局报道偏倚

在众多研究中，虽然相当多结局值都记录下来，但并非对所有记录的结局都进行报道（Pocock，1987；Tannock，1996）。选择报道的结局会受结果的影响，可能使发表的结果具有误导性。例如，对于一个评价阿莫西林对儿童化脓性中耳炎疗效的随机双盲对照试验，的两次独立的分析（Mandel，1987；Cantekin，1991）得到的结论相反，主要是因为研究中对于所评价的结局指标的权重不同。伴随着对支持阿莫西林有效的研究组行为不当的指控，分歧被公诸于众。该研究组的领导人接受了阿莫西林制造商提供的大量的财政扶持，既有研究经费，又有私人酬金（Rennie，1991）。这是个不错的例子，

证明了依赖调查人员选择性报道的数据会怎样的歪曲真相（匿名，1991）。这种“结局报道偏倚”对不良反应可能尤为重要。Hemminki调查了芬兰、瑞典两国由制药公司申请新药注册机构提交的的临床试验报告，并发现未发表的试验能比发表的试验提供更多有关不良反应的信息（Hemminki，1980）。其后另一些研究也表明，临床试验对不良反应及安全性结局的报道往往不够，且系存在选择性（Ioannidis，2001；Melander，2003；Heres，2006）。最近由加拿大、丹麦、英国联合组建的一个研究组开创了对于研究结果选择性报告的实证研究（Chan，2004a；Chan，2004b；Chan，2005）。第8章对这些研究有过介绍（见8.13节），并对结局报道偏倚有更详细的讨论。

10.3 避免报告偏倚

10.3.1 报告偏倚证据的影响

10.2节介绍过，能表明报告偏倚存在的、有说服力的证据证明，有必要为Cochrane系统评价全面检索符合纳入标准的所有研究。系统评价作者应确保采取多渠道检索，比如，只检索MEDLINE可能不够。第6章已详细介绍了检索途径、检索方法。但是，全面检索不可能消除偏倚。系统评价作者应牢记，比方说，研究报告可能有选择地报告结果、参考文献目录会选择性的援引资源、重复发表结果很难找出，等等。还有，研究信息的可获得性可能受时滞偏倚的影响，特别是在那些进展神速的研究领域。我们这里探讨两种可进一步减少、或者说有可能避免报告偏倚的方法：一是纳入未发表研究，一是利用试验注册库。

10.3.2 系统评价中纳入未发表研究

发表偏倚对任何系统评价的有效性都是一种主要威胁，尤其对非系统、描述性的系统评价而言更是如此。找到并从未发表的试验提取数据似乎是避免该问题的明显方法。Hopewell及同事在随机试验的Meta分析中对比较纳入、排出“灰色”文献（这里指各级政府、学术团体、商务界、工业界以印刷版及电子格式做的报告，其中未由商业发行人经管）的效应做过系统评价（Hopewell，2007b）。他们纳入了5个研究（Fergusson，2000；McAuley，2000；Burdett，2003；Hopewell，2004），这些研究都表明，已发表试验的干预措施疗效总的来说大于灰色试验。一篇包含其中3个试验的Meta分析提示，平均来说，

已发表试验给出的干预措施疗效比灰色试验大9%（Hopewell，2007b）。

从未发表的试验中纳入数据本身也会产生偏倚。检索到的研究可能是所有未发表研究的非代表性样本。未发表研究的方法学质量可能不如已发表研究：一个针对60篇纳入已发表及未发表试验Meta分析的研究发现，未发表试验不太可能充分隐藏干预的分配，也不太可能对结局评估施盲（Egger，2003）。相反，Hopewell及同事发现这类信息的报告质量并无差异（Hopewell，2004）。

进一步的问题是关于调查人员检索未发表研究以提供研究数据的意愿。这也取决于研究结论，更有利的结果自然更易于提供。但会给系统评价的结论造成偏倚。有趣的是，致力于获得围产医学未发表试验的信息，Hetherington等曾接触了18个国家的42，000名产科医师、儿科医师，他们仅发现18个已经完成两年以上的未发表试验（Hetherington，1989）。

曾有调查问卷评估对纳入未发表数据的态度，问卷发给150个Meta分析的作者及发表这些Meta分析的杂志的编辑（Cook，1993）。研究人员和编辑对Meta分析纳入未发表数据的观点不一样。大多数（78%）Meta分析作者赞成引用未发表材料，而杂志编辑则相对较少（47%）（Cook，1993）。这项研究最近又重复做了一遍，为的是了解系统评价中灰色文献的纳入情况，发现赞成纳入灰色文献的受访者数增加了，尽管两组受访者态度仍存在种种差异（系统评价作者：86%，编辑：69%），但比起Cook等给的数据，这些差异还是减少了（Tetzlaff，2006）。不愿纳入灰色文献，原因包括未发表文献没有进行同行评审。但要切记的是，评审过程也不可能确保发表结果是真实的（Godlee，1999）。准备Cochrane系统评价的小组中应至少包括一位类似于杂志同行评审级别的专家，来评价未发表研究。另一方面，针对从感兴趣文献来源未发表数据的Meta分析显然值得关注。

10.3.3 试验注册与发表偏倚

2004年9月，很多隶属国际医学杂志编辑委员会（ICMJE）的主流医学杂志宣布，他们将不再发表一开始即未注册的试验（Abbasi，2004）。2005年9月后开始招募受试者的所有试验，如考虑在这些杂志上发表，需在招募受试者时或之前就在公共试验注册机构注册。ICMJE是这么描述“认可的”注册机构的：可经电子检索的、公众能免费查及的、向所有注册人开放的、由非营利组织经管的。类似的，ICMJE还要求临床试验申请人恪守世界卫生组织提出的最小数据集。

如果这种期待已久的倡议获得成功，有可能从根本上减少发表偏倚的影响。但是，这取决于系统评价作者通过检索在线的试验注册库找出全部相关的试验、并且还取决于系统评价作者经可查及的注册库找出的未发表试验的结果。管理注册试验的结果这一倡议目前尚处在初期，但是发展迅猛，而且将对这些数据的可获得性产生影响。不断有证据表明，尽管注册机构要求的某些数据字段不完整（Zarin，2005），这可能随着时间不短改善。试验注册机构能多大程度上促进Cochrane系统评价作者的工作，目前还不清楚。关于检索试验注册机构的建议，参见第6章（6.2.3部分）。

10.4 发现报告偏倚

10.4.1 漏斗图

漏斗图是一个简单的散点图，反映研究在一定样本量或精确性下单个研究的干预效应估计值。和森林图一样，漏斗图最常见的是在横轴为各研究效应估计值，纵轴为研究样本量。这和传统散点图的图示相反，传统散点图中，结局（如干预措施疗效）标于纵轴，协变量（如研究的样本量）标于横轴。

“漏斗图”的称法是源于随着研究样本量增加，干预措施疗效估计值的精确度增加。因此，小样本研究的疗效估计值在漏斗图底部更分散，而较大样本的研究则分布得较窄。在没有偏倚的情况下，图像中的点应聚集成一个大致对称的（倒置的）漏斗。图10.4.a：图A阐明了此种情况，图中大样本研究的疗效估计接近于干预措施真实的比值比（OR，Odds Ratio）0.4。

若存在偏倚，例如由于疗效无统计学意义的小样本研究尚未发表（即图10.4.a：图A空心圈所示），将使漏斗图外观不对称，图形底角有空白（图B）。这种情况下，Meta分析计算出的效果可能会高估干预措施疗效（Egger，1997a；Villar，1997）。不对称越明显，越有可能存在实质的偏倚。

漏斗图最初用于教育研究和心理学领域，绘制对应于不同总样本量的效应估计值（Light，1984）。现常建议纵轴用干预措施疗效估计值的标准误，而非样本总量（Sterne，2001）。这是因为除样本量之外，试验的统计效能还由多个因素决定，诸如发生二分类结局事件的受试者人数，以及连续性变量结果的标准差。例如，与一个有1000个受试者、发生事件数100例的研究相比，一个有100000个受试者，发生事件10例的研究不太可能

得出一个有统计学意义的干预措施效果。而标准误涵盖了上述的这些其它因素。在倒置的刻度上绘制标准误，可将较大的、或者说更有效的研究置于靠近的位置。使用标准误另一个好处，即可大致勾画一个简单的三角形区域，希望区域内95%的研究既无偏倚、也无异质性。图10.4.a就包含这些区域。用RevMan软件，可绘制在不同的标准误（在倒置的刻度上）下效应估计值的漏斗图。图中可引入基于固定效应模型Meta分析的三角形95%置信区，并且不同的绘图标记能使不同亚组的研究易于区分。

发表偏倚不一定引起漏斗图不对称。如果干预措施无效，仅根据P值的选择性发表将使漏斗图不对称，因处在两端的研究比处在中间的研究更有可能被发表。这会给所估计的研究间方差不齐造成偏倚。

干预措施疗效的比率指标（如比值比、风险比）要在对数尺度上绘制，这能使同样大小、但方向相反的疗效值（如比值比0.5、比值比2）与1.0等距。对以连续性（数值型）尺度表示的结局（如血压、抑郁评分），应以均数差或标准化均数差衡量干预措施疗效，这些统计指标可作为漏斗图的横轴。据我们了解，目前还没有实证研究选择将连续性的结局用作漏斗图的坐标轴。就均数差而言，标准误与受试者人数平方根的倒数大致成正比，故标准误可作为纵轴毫无争议的选择。

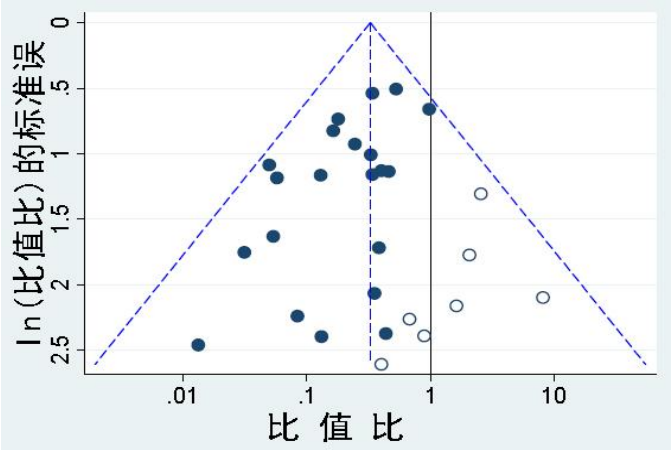
有些作者认为，对漏斗图的目测解释过于主观而用处不大。尤其是，Terrin等发现，研究人员只有非常有限的能力，可以正确无误找出受发表偏倚影响的Meta分析的漏斗图（Terrin，2005）。

漏斗图还有个重要问题，就是有些疗效估计值（如比值比、标准均数差）本来就与其标准误相关，在漏斗图中可引起虚假的不对称。我们将在10.4.3部分对这一问题详细讨论。

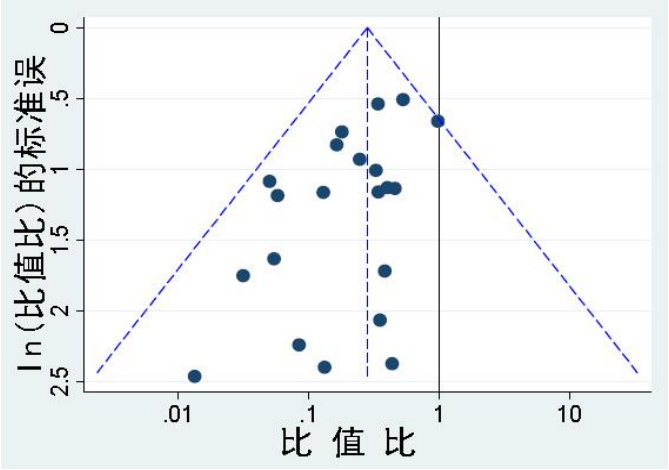
图10.4.a　假想的漏斗图

图A：无偏倚的对称漏斗图。图B：存在报告偏倚的不对称漏斗图。图C：方法学质量不高、进而夸大干预措施疗效估计值的小样本研究（空心圈所示）所致偏倚的不对称漏斗图。

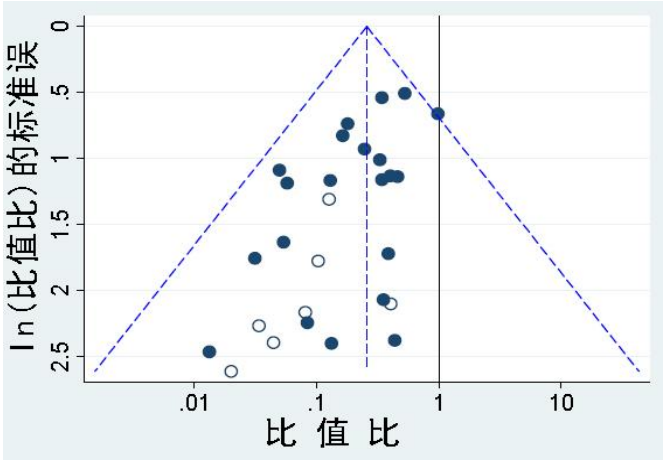
图A



图B



图C



10.4.2 造成漏斗图不对称的不同原因

尽管早就将漏斗图不对称与发表偏倚同等看待（Light 1984；Begg 1988），但漏斗图应视为表示小样本研究效应（估计的干预效果在小样本研究与大样本研究中存在不同的一种趋势）的通用方法（Sterne 2000）。小样本研究效应可能取决于发表偏倚之外的其它因素（Egger 1997a；Sterne 2000）。其中部分因素见表10.4.a。

方法学质量的差异是漏斗图不对称一个重要的潜在原因。与大样本研究相比，小样本研究在实施和分析的方法学上可能不严谨（Egger 2003）。低质量的试验还同样可能得出较大的干预措施疗效（Schulz 1995）。因此那些本来是“阴性”的试验，如果实施和分析得当，可能变为“阳性”（图10.4.a：图C）。

干预措施疗效真实的异质性也会使漏斗图不对称。比如，仅在就干预措施影响的结局而言处于高风险的患者中，才能看出干预的实质获益；而早期阶段的小样本研究更有可能纳入这些高风险患者（Davey Smith 1994；Glasziou 1995）。此外，小样本试验往往在大样本试验确立前就已经实施，在干预疗程期间内标准治疗可能已经得到改进（使大样本试验中干预措施的疗效偏小）。而且，有些干预措施在大样本试验里可能实施得不彻底，这样也可能会使干预措施的疗效估计值偏小（Stuck 1998）。最后，当然有可能仅仅是机遇的原因使漏斗图不对成得到。Terrin等认为漏斗图不适用于存在异质性的Meta分析，因为发起漏斗图的前提条件是所有研究来自潜在的同一总体（Light 1984；Terrin2003）。

有个可强化漏斗图的提议（Peters 2008），即引入等高线，这些等高线相当于所谓有统计学意义（P=0.01、0.05、0.1等等）的“里程碑”。这样做能够兼顾研究效应估计值的统计学显著性，以及被视为缺失的研究。这种“经等高线强化的”漏斗图可帮助系统评价作者鉴别因发表偏倚造成的不对称及其它因素所致的不对称。比如，如果研究看似缺失无统计学显著性的那部分（示例参见图10.4.a：图A），那么这就是不对称性肯很可能是由发表偏倚所致的凭证。反之，如果假设缺失的研究来自统计学显著性较高的那部分（示例参见图10.4.a：图B），就意味着不对称原因更有可能取决于其它因素、而非发表偏倚（参见图10.4.a）。如果有统计学显著性的研究不存在，发表偏倚可能就不是漏斗图不对称的一个合理解释（Ioannidis 2007b）。

在解释漏斗图时，系统评价作者要能区分表10.4.a列举的造成漏斗图不对称的各个可能原因。对特定干预措施及其在不同研究中实施的环境的了解，有助于找出导致漏斗

图不对称的实际存在的异质性。值得留意的是，目测解释漏斗图本身就有主观性。所以，我们这里将讨论对漏斗图不对称进行统计检验，并探讨统计检验多大程度上能有助于客观解释漏斗图。如果系统评价作者担心小样本研究效应影响Meta分析结果，他们可能想进行敏感性分析，以进一步探索Meta分析对于漏斗图不对称原因的不同假设所得结论的稳定性，这些在10.4.4部分予以讨论。

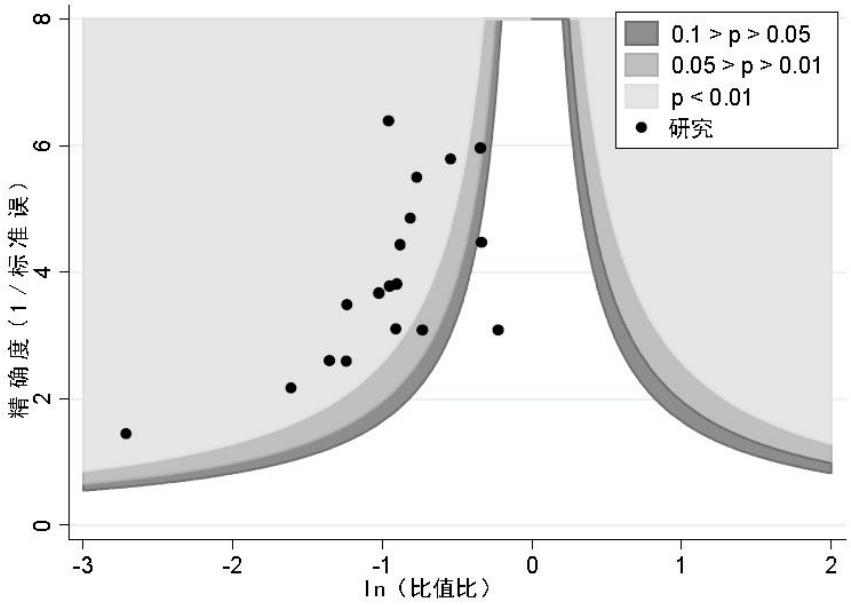
表10.4.a　可能造成漏斗图不对称的原因——改编自Egger等（Egger 1997a）

1.选择偏倚
1.1发表偏倚
1.1.1延期发表偏倚（亦称为“时滞偏倚”、“管道偏倚”）
1.1.2检索偏倚
1.1.2.1语言偏倚
1.1.2.2引文偏倚
1.1.2.3重复发表偏倚
1.2选择性的结果报道
2.方法学质量过低，造成小样本研究得出虚假的夸大疗效
2.1方法学设计不良
2.2分析不完整
2.3（学术）欺诈
3.真实的异质性
效应量不同由研究的样本大小造成（如由干预措施强度或不同样本量的研究之间潜在差异风险造成的）
4.人为因素
有些情况下（参见*10.4.3部分*），抽样变异会使干预措施疗效与疗效标准误之间有联系。
5.机遇

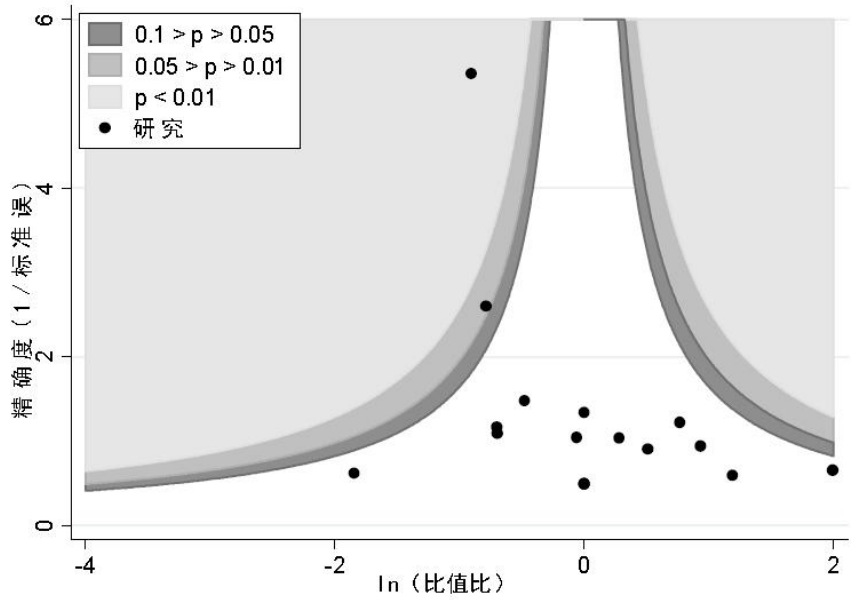
图10.4.b　等高线强化的漏斗图

图A：提示缺失的研究处在漏斗图的右侧，大致分布在无统计学意义的那些区域（图中P>0.1的白色部分），这部分可用发表偏倚合理解释。图B：提示缺失的研究处在漏斗图底部的左侧，由于部分主要包括有显著统计学意义区域（图中以深色阴影标示），减少了发表偏倚系漏斗图不对称的潜在原因的合理性。

图A



图B



10.4.3 对于漏斗图不对称的检验方法

漏斗图不对称（小样本研究效应）的检验方法检验估计的干预疗效和研究样本量的测量值（如干预措施疗效的标准误）间的联系是否大于机遇产生的联系。采取连续性（数值型）尺度测量结局合理而直接。用Egger等推荐的方法（Egger 1997a），我们可用干预措施疗效估计值的标准误对其进行线性回归，权重为干预措施疗效估计值的方差的倒数，以寻找干预措施疗效及其标准误间的直线关系。如果无效假设是没有小样本研究效应（如见图10.4.a：图A），该直线将垂直于横轴。如果干预措施疗效及标准误间的联系愈大（如见图10.4.a：图B），漏斗斜线将越偏离中垂线。需要注意，权重对于确保回归估计值不受小样本研究的主导很重要。

如果结局指标属于二分类，干预措施疗效以比值比表示，则Egger等（Egger 1997a）推荐的方法相当于对数比值比及对数比值比标准误间的线性回归，权重为对数比值比方差的倒数（Sterne 2000）。迄今为止，本法系漏斗图不对称最常用的检验法。遗憾的是，这种方法还是存在统计学问题，因为即使没有小样本研究效应，对数比值比的标准误在数学上依然和比值比的大小有关（Irwig 1998）（参见Deeks等对本现象的代数解释（Deek 2005））。这会使以对数比值比绘制的漏斗图不对称，意味着用Egger等使用的检验法求得的P值过小，从而得到假阳性的检验结果。如果干预措施疗效很大、存在明显的研究间异质性、或各研究发生的事件数很少、或所有研究样本量相似，则这些问题更有可能出现。

众多作者因此提出其它检验漏斗图不对称的方法，表10.4.b总结了这些方法。正因为发表偏倚的准确成因不得而知，才要求在根据很多发表偏倚成因假设的前提下、用模拟试验（用计算机产生的海量的数据集来评估检验方法）评估这些检验法的特点（Sterne 2000；Macaskill 2001；Harbord 2006；Peters 2006；Schwarzer 2007）。Rücker等报道了最为全面的研究（在检验的场景、实施的模拟、参照的各种检验等方面）（Rücker 2008）。这一研究及其它已发表的模拟研究提供了对于漏斗图不对称检验的如下建议。尽管模拟研究能提供了十分有用的深入见解，但它们评价的环境不可避免的不同于某个特定的Meta分析的具体环境，因此在解释模拟研究的结果时务须慎重。

大部分的这类方法学工作主要集中于用比值比表示的干预措施疗效。对于以危险度或标准化均数差表示的干预措施疗效，预期将出现相同的问题尽管看似合理，但需要对这种情形进一步的调查。，

对于模拟试验用的参数值的代表性，以及没有明确的合理性但经常用于模拟发表偏倚和小样本研究效应的机制，目前仍有争议。不同检验法一些可能有效的变更仍未经检查验证。因此在选择漏斗图不对称性的检验方法时，不可能给出明确建议。尽管如此，对于想要检验漏斗图不对称的系统评价作者，我们仍能够找到3种值得考虑的方法。

RevMan软件未使用这里任何一种检验法，具体使用时宜咨询专业统计人员。

表10.4.b　推荐的用于漏斗图不对称的检验法

N_{tot}表示样本总数，N_E、N_C分别表示试验组、对照组例数，S代表两组事件发生的总例数，且F= N_{tot}－S。

注意：仅前3种检验法（Begg，1994；Egger，1997a；Tang，2000）可用于连续性结局指标。

参考文献	检验的条件
Begg 1994	标准化的干预措施效应值及其标准误间存在秩相关
Egger 1997a	干预措施疗效的估计值对应于其标准误呈线性回归，权重为干预措施疗效估计值方差倒数
Tang 2000	干预措施疗效的估计值对于1/$\sqrt{N_{tot}}$ 呈线性回归，权重为Ntot
Macaskil 2001*	干预措施疗效的估计值对于Ntot呈线性回归，权重为S×F/Ntot
Deeks 2005*	比值比的对数对于1/$\sqrt{ESS}$ 呈线性回归，权重为ESS（有效样本量）且ESS=4NE×NC/Ntot
Harbord 2006*	Egger等据比值比对数的（O-E）“得分”及 “方差得分”（V）提出的校正检验法
Peters 2006*	干预措施疗效估计值对于1/Ntot呈线性回归，权重为S×F/Ntot
Schwarzer 2007*	使用非中心超几何分布的均数和方差进行秩相关检验
Rücker 2008	对研究间的异质性明确建模，检验基于所观测风险的余弦转换

*用比值比表示的检验法，但可能适于干预效应的其它指标。

10.4.3.1 关于漏斗图不对称检验的建议

对于所有类型的结局指标：

- 根据经验，仅当Meta分析纳入至少10个研究时方可使用漏斗图不对称检验，因

为如果纳入研究过少，检验效能将过低，将无法区别机遇和真正的不对称。

- 如果所有研究的样本量近似（干预措施疗效估计值的标准误相似），不宜采用漏斗图不对称检验。然而我们并不知道来自模拟实验的证据，可以指导何时将研究样本量视为“过于相似”。
- 应通过目测漏斗图来解释漏斗图不对称检验的结果。比如，小样本研究真的使得干预效应的估计值更加有利或不太有利？在Meta分析中是否存在干预效果估计值明显不同或者具有高度影响的研究？小的P值是否仅由一个研究造成？通过观察等高线强化的漏斗图，正如10.4.1部分所述，可有助于进一步解释检验结果。
- 如果有证据支持小样本效应，发表偏倚应视为众多可能的解释之一（参见表10.4.a）。尽管漏斗图（及各种漏斗图不对称检验法）可提醒系统评价作者注意要考虑的问题，但它们没有给出解决问题的办法。
- 最后，系统评价作者应牢记，由于这些检验法的检验效能往往相对较低，甚至在某种检验法不能给出漏斗图不对称的证据时，仍然无法排除偏倚（包括发表偏倚）。

对于以均数差表示干预措施疗效的连续性结局指标：

- 可以用Egger提出的的方法检验漏斗图不对称（Egger 1997a）。目前，没有任何支持后来提出的用于这种情形下的检验方法的理由，尽管还没有正式考察这些检验法的优缺点。我们知道，对于连续性情况，尚未有专门针对这种方法检验效能的研究，但常识告诉我们检验效能会比二分类结局的大一些，但将该法用于远远少于10个研究的情况就不明智了。

对于以比值比表示干预措施疗效的二分类结局指标：

- Harbord等（Harbord 2006）、Peters等（Peters 2006）提出的检验方法避免了Egger等提出的检验法中当明显的干预效应存在时，对数比值比与其标准误间的数学联系（并因此得到假阳性检验结果），同时保留了较之其它的检验法相当的检验效能。但是，当存在明显的研究间异质性时，还是会出现假阳性结果。
- 当干预措施确有疗效、同时存在明显研究间异质性的情况下，Rücker等（Rücker，2008）提出的检验法可避免假阳性结果。一般来说，如果估得的对数比值比研

究间异质性方差（τ方）大于0.1，只有包含随机效应（即Rücker等表示的“AS+RE”）在内的反正弦检验有效。但若无异质性，这种检验就有些保守；因其建立在反正弦转换基础上，故其结果解释不太为人所熟知。（注意：尽管本方法是是以其它影响方法检验效能的因素（包括不同研究的样本量及其分布）τ方的大小为基础。我们现在无法在这些方法中中加入这些因素。）

- 如果异质性方差τ方小于0.1，可使用Harbord 2006、Peters 2006或Rücker2008提出的检验方法（检验效能一般随τ方增加而减弱）。
- 只要有条件，系统评价作者应事先确定其检验策略（应注意所选检验法取决于观察到的异质性程度）。他们应从上面推荐列表里只选一种检验法，并只报告由所选检验方法得到的结果，以和特定Meta分析的上下文相适应。同时运用两种或多种检验方法并不可取，因为由一组检验方法得到的极端P值（最大值或最小值）不能被很好的解释。

对于以危险度或风险差表示的干预措施疗效的二分类结局指标和以标准化均数差表示的连续性结局指标：

- 漏斗图中对这些效应量可能存在问题的研究不如对比值比的研究深入，并且现在对此尚无可靠的指导。
- 一般认为用风险差的Meta不如用比率做效应量指标的Meta分析合适（参见第9章9.4.4.4部分）。也因为类似原因，很少会对采用风险差的漏斗图感兴趣。如果各研究间的危险度（或比值比）固定不变，且小样本研究的基线风险值过高（或过低），使用风险差的漏斗图将不对称。

据对Cochrane系统评价数据库所发表Meta分析的调查，显示应该用于漏斗图不对称检验方法的标准仅被用于少部分的Meta分析（Ioannidis 2007b）。

对推荐使用的证据不足的检验方法：

下述建议适于所有的干预措施指标。Begg与Mazumdar（Begg 1994）提出的检验法存在同样的统计学问题，但检验效能低于Egger等的方法，故不推荐使用。唐金陵与Joseph刘（Tang 2000）提出的检验方法尚未在模拟试验里进行过评价，同时Macaskill等（Macaskill 2001）提出的检验法比最近提出的其它方法检验效能低。Schwarzer等（Schwarzer 2007）提出的方法虽避免了对数比值比及其标准误间的数学联系，但检验

效能相对低于上面讨论的检验方法。

在本章讨论的有关干预措施研究的Meta分析的内容中，Deeks等（Deeks 2005）提出的方法的检验效能似乎低于最近提出的其它方法。该方法并不是为了检验随机试验系统评价中的发表偏倚，而是在于诊断性试验准确性研究的meta分析，这类研究中比值比极大、各研究之间也极不均衡，会给其它的检验方法带来麻烦。

10.4.4 敏感性分析

如果系统评价作者发现小样本研究效应，他们应考虑进行敏感性分析，根据与这类效应产生原因相关的不同假设，调查系统评价结果将如何变化。我们强调这种分析的探索性质，是因为发表偏倚的调整与生俱来就有困难，同时缺少对这种基于10.4.3部分所述发表偏倚检验方法结果的检验效能的研究。这一领域还相对欠完善，故建议使用以下方法。

10.4.4.1 比较固定效应估计值和随机效应估计值

存在异质性的情况下，随机效应Meta分析赋予各研究的权重较固定效应分析相对更为均衡。由此可见，对于图10.2.a展示的那些小样本研究效应（即：小样本研究中干预措施的疗效更佳），干预效应的随机效应估计值将比固定效应估计值更加有利。Poole与Greenland（Poole，1999年）对此有所总结，他们认为“随机效应的Meta分析并不总是保守的”。第9章也讨论过这一问题（9.5.4部分）。

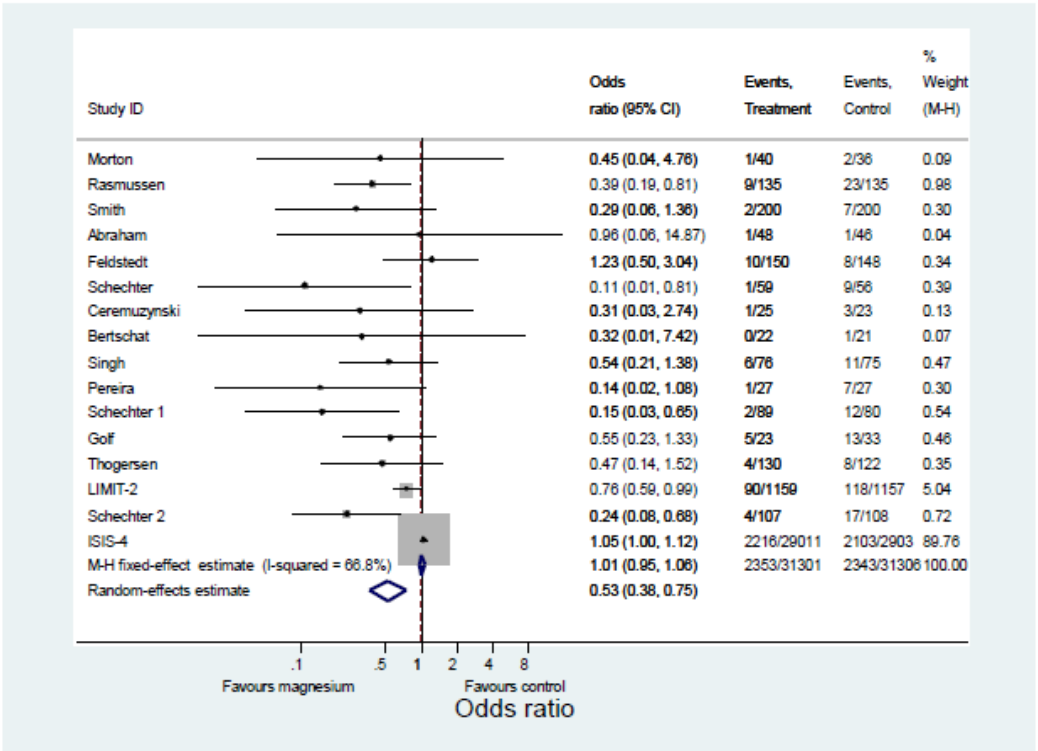
如存在小样本研究效应，会出现固定效应分析、随机效应分析间的差异。一个典型的例子如图10.4.c所示，图中给出了用固定效应、随机效应估计静脉补镁对心肌梗死病死率影响的结果。在这个经典案例中，小样本研究的Meta分析发现干预措施确有疗效，但与后来进行的大规模ISIS-4试验（即国际心肌梗死存活第4试验，由牛津大学主持、遍布30个国家，旨在比较镁、卡托普利及硝酸甘油对心脏病死亡事件的预防作用）结论相反，后者并没有发现镁对心肌梗死死亡率有影响的证据

正因为研究间存在显著的异质性，随机效应分析赋予各研究的权重比固定效应分析均衡得多。如用固定效应分析，ISIS-4试验所占权重将为90%，并得出没有证据支持干预措施有效的结论。如小样本研究占主体时采用随机效应分析，则貌似有明确证据支持干预措施有效。欲解释不断汇集的证据，需要判断小样本研究合并证据的可能真实性，

并与ISIS-4试验所得证据的真实性进行比较。

我们建议，如果系统评价作者注意到小样本研究效应对研究间有异质性证据（I^2>0）的Meta分析有影响时，他们应比较对干预措施疗效的固定效应和随机效应估计。如果两者估计相近，则任何小样本研究效应对干预措施疗效的估计都影响甚小。若随机效应估计更有效，系统评价作者就应考虑，小样本研究中干预措施更有效的结论是否合理。如果大样本研究所用的方法学更严格，或干预措施实施的环境更接近实际，系统评价作者就应考虑，报告仅限于设计严格的大型研究的Meta分析的结果。在模拟的研究中对这样的策略进行正式的评估将是可取的。需要注意的是：进行干预效果固定效应和随机效应估计值的比较是不可能的；在没有异质性证据时，即使干预措施疗效的固定效应估计值及其随机效应估计值相等，小样本研究效应还是可能对Meta分析结果造成偏倚。

图10.4.c　静脉补镁对心肌梗死死亡率效果的固定效应模型Meta分析与随机效应模型Meta分析的比较



10.4.4.2 剪补法

“剪补”法旨在查实及纠正因发表偏倚所致的漏斗图不对称（Taylor 1998；Duval 2000）。这种方法基础是：（1）“修剪”（去掉）引起漏斗图不对称的小样本研究，（2）用修剪过的漏斗图估计漏斗的真实“中心”，接着（3）在重估的中心周围替换省去的研究和他们对应的缺失研究（填补）。除了给出缺失研究数量的估计值，还要用纳入这些填充的研究进行Meta分析求得调整后的干预措施疗效。

剪补法无需对导致发表偏倚的原因进行假设，但要提供缺失研究数量的估计值以及因发表偏倚（基于填补的研究）而“调整”的干预措施疗效的估计值。但是，这是建立在本应对称的漏斗图、且不能保证经调整的干预措施疗效与无发表偏倚时观察到的情况相配的强假设基础之上，因为我们无法知道发表偏倚的真正成因。同样重要的是，剪补法并未考虑发表偏倚以外的其它造成漏斗图不对称的原因。基于此，在解释这种方法得到的“校正的”干预措施疗效的估计值时务须慎重。在已知研究间异质性很大时，这种方法的效能很差（Terrin 2003；Peters 2007）。此外，估计、推论都是基于包含填补的干预效应估计值的数据集。有人会认为，这些估计值会不恰当地提供减少总的干预效应中不确定性的信息。

10.4.4.3 失安全数

Rosenthal建议，可通过计算“失安全数”评估发表偏倚对Meta分析结果的可能影响，即要使Meta分析的P值增至0.05以上，所需的额外“阴性”研究（即干预措施的疗效等于零的研究）的数量（Rosenthal 1979）。但是，失安全数的估计高度取决于给予未发表的研究假设的平均干预措施疗效（Iyengar 1988），并且可用的方法得到的额外研究的数量差异极大（Becker 2005）。尽管也给出了针对效应大小的相关方法，这种方法有违一般医学研究，尤其是系统评价的准则，应该关注的是估计的干预效应的大小及其可信区间，而不是关注P值是否达到人为设定的某个特定的临界值，（Orwin 1983）。故不建议在Cochrane系统评价中使用本方法或相关的其它方法。

10.4.4.4 其它可选模型

（Dear 1992；Hedges 1992）。这些方法还能扩展用于估计干预措施的疗效，校正估计的发表偏倚（Vevea 1995）。但是，它需要相当多的研究，以至于能够包括足够范围的研究P值。模拟未观察研究的数量和结局的贝叶斯法也被提出作为校正发表偏倚的干预

措施疗效估计值的一种方法（Givens，1997）。近期的研究调查了根据多种权重函数评估稳健性的可能性，进而可避免对大量研究的需要（Vevea 2005）。统计方法的复杂性及需要大量的研究或能解释为何可选模型未在实际广泛应用。

10.4.4.5 基于可选模型的敏感性分析

Copas设计出一种模型，模型中某研究被Meta分析纳入的可能性取决于其标准误。因无法精确地估计所有的模型参数，他提倡进行敏感性分析，根据选择偏倚严重程度的种种假设计算出干预措施疗效的估计值（Copas 1999）。本方法并不是一个发表偏倚的干预效应估计值的校正，读者可以看到，随着估计的选择偏倚数量递增，估计的疗效值（及置信区间）是如何变化的。本法用于环境性吸烟（environmental tobacco smoke，ETS）与肺癌的流行病学研究，认为发表偏倚可以解释部分在这类研究的Meta分析中所观察到联系（Copas 2000）。

10.4.4.6 对额外的有意义结果研究的检验

Ioannidis与Trikalinos提出一种简单的检验方法，希望能估计形式上有统计学意义结果的研究是否过量（Ioannidis，2007a）。该检验方法比较在形式上有统计学意义结果的研究数量和期望在基于效应量大小的不同假设下有统计学意义结果的研究的数量。最简单的假设即效应量等于Meta分析中所观察到的总效应（但这会产生一个循环成分）。另外也可采用其它的潜在效应量值及不同的具有统计学意义的标准。因此，像在10.4.1部分所介绍的等高线漏斗图（但不同于回归检验），本方法法考虑了结果有意义的研究的分布情况。但是，和回归检验及等高线漏斗图都不同的是，这种检验法不对小样本研究效应做任何假设。额外的有意义结果要么反映全部研究的抑制情况、要么反映相关的选择性/操作性分析和可能引起类似超额的报告行为。

和其它大多数检验方法一样，如果研究过少和有统计学意义结果的研究过少，这种方法的检验效能将很有限。由于本方法在和其它方法比较时没有通过模拟进行严格的评价并在不同的场景下进行，故目前不建议将本方法作为10.4.3部分介绍的那些检验方法的替代方法。

本方法的一个新特点在于可用于同一研究领域的大量Meta分析中，以调查某个临床研究领域发表偏倚及选择报告偏倚的程度。此外，欢迎对本法进行深入评估。

10.4.5 小结

尽管确有证据证明，发表偏倚及其它报告偏倚会使得干预效应的估计过于乐观，但要解决发表偏倚、查找发表偏倚及校正发表偏倚却存在问题。进行全面的检索非常重要，特别对于像随机试验那样有明确定义的研究更是如此。但是，全面的检索也尚不足以防止有些实质上潜在的偏倚。

发表偏倚应视为可能造成“小样本研究效应”的众多原因之一，这种效应会存在小样本研究中干预措施疗效的估计值更优的趋势。漏斗图允许系统评价作者进行目测评估，判断一篇Meta分析是否存在小样本研究效应。对于以均数差表示干预措施疗效的连续性（数值型）结局指标，漏斗图及漏斗图不对称的统计检验是有效的。但对于以比值比表示干预措施疗效的二分类结局指标，即使不存在小样本研究效应，对数比值比的标准误在数学上还是和比值比的大小有关联。这会使对数比值比（或用对数尺度表示比值比）绘制的漏斗图不对称，并意味着使用Egger等提出的检验方法法得到的P值过小。对与其它的效应指标尚无可靠的指导提供。Cochrane系统评价中，在有至少10个研究纳入的条件下，推荐使用3种小样本研究效应的统计检验方法。但是，RevMan软件未采用上述任何的一种方法，因此常需来自统计学的支持。当研究间的异质性方差超过0.1时，仅有一种检验被证明有用。宜谨慎解释漏斗图不对称的检验结果。当存在小样本研究效应时，发表偏倚应仅视为众多可能的解释之一。这些情况下，系统评价作者需试着理解小样本研究效应的来源，并且考虑它们对于敏感性分析的影响。

10.5 本章信息

编辑：Cochrane偏倚方法学组Jonathan AC Sterne、Matthias Egger、David Moher

本章引用格式：Sterne JAC, Egger M, Moher D (editors). Chapter 10: Addressing reporting biases. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.

有贡献的其他作者：James Carpenter、Matthias Egger、Roger Harbord、Julian Higgins、David Jones、David Moher、Jonathan Sterne、Alex Sutton、Jennifer Tetzlaff

致谢：我们向Doug Altman、Jon Deeks、John Ioannidis、Jaime Peters、Gerta Rücker表示感谢，

感谢他们提供有益的意见。

利益声明：James Carpenter、Jon Deeks、Matthias Egger、Roger Harbord、David Jones、Jaime Peters、Gerta Rücker、Jonathan Sterne、Alex Sutton均系提出漏斗图不对称检验方法的论文作者。

图10.5.a　Cochrane偏倚方法学组

偏倚方法学组（下简称BMG），即以前报告偏倚方法学组，2000年作为方法学组正式注册。BMG着眼于各种不同形式偏倚，如发表偏倚、语言偏倚、选择性报告结局偏倚及其它设计研究、实施研究时产生的偏倚。本组主要目的（在和统计方法学组合作的条件下）在于提供新指导，以评估Cochrane系统评价纳入研究的发表偏倚。

BMG成员学术活动含：

- 进行专业研究，调查哪些情况下，各种偏倚可能对系统评价、包括对Cochrane方法学系统评价产生实质上的影响。
- 进行方法学研究，调查怎样明确、怎样报告系统评价、Meta分析中的潜在偏倚。
- 帮助完成、协调进行本组研究范围相关的方法学系统评价。
- 为Cochrane组织提供建议。
- 通过正式/非正式途径为Cochrane系统评价作者、非Cochrane系统评价作者进行培训。

BMG成员电子邮件列表可当做讨论、传播信息的论坛。Cochrane方法学组通讯、Cochrane新闻、CCInfo（Cochrane协作网资讯——译者注）等Cochrane通讯、电子邮件分布列表也可用于传播本组活动。

Funding: The BMG receives infrastructure funding as part of a commitment by the Canadian Institutes of Health Research (CIHR) and the Canadian Agency for Drugs and Technologies in Health (CADTH) to fund Canadian-based Cochrane entities. This supports dissemination activities, web hosting, travel, training, workshops and a full time Co-ordinator position.

资助：BMG所获基础建设资助，系加拿大卫生研究院（CIHR）、加拿大药品及卫生技术局（CADTH）对加拿大国内Cochrane组织承诺资助的一部分。该资助可供传播事宜、经营网站、出行培训、举办研讨会及设立一全职协调员岗位之需。

网址: www.chalmersresearch.com\bmg

10.6　参考文献

Abbasi 2004

Abbasi K. Compulsory registration of clinical trials. BMJ 2004; 329: 637-638.

Abbot 1998

Abbot NC, Ernst E. Publication bias: direction of outcome is less important than scientific quality. Perfusion 1998; 11: 182-182.

Anonymous 1991

Anonymous. Subjectivity in data analysis. The Lancet 1991; 337: 401-402.

Bailey 2002

Bailey BJ. Duplicate publication in the field of otolaryngology-head and neck surgery. Otolaryngology and Head and Neck Surgery 2002; 126: 211-216.

Barden 2003

Barden J, Edwards JE, McQuay HJ, Moore RA. Oral valdecoxib and injected parecoxib for acute postoperative pain: a quantitative systematic review. BMC Anesthesiology 2003; 3: 1.

Bardy 1998

Bardy AH. Bias in reporting clinical trials. British Journal of Clinical Pharmacology 1998; 46: 147-150.

Becker 2005

Becker BJ. Failsafe N or file-drawer number. In: Rothstein HR, Sutton AJ, Borenstein M (editors). Publication Bias in Meta-Analysis. Chichester (UK): John Wiley & Sons, 2005.

Begg 1988

Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. Journal of the Royal Statistical Society Series A 1988; 151: 419-463.

Begg 1994

Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994; 50: 1088-1101.

Bhandari 2004

Bhandari M, Busse JW, Jackowski D, Montori VM, Schünemann H, Sprague S, Mears D, Schemitsch EH, Heels-Ansdell D, Devereaux PJ. Association between industry funding and statistically significant pro-industry findings in medical and surgical randomized trials. Canadian Medical Association Journal 2004; 170: 477-480.

Blumenthal 1997

Blumenthal D, Campbell EG, Anderson MS, Causino N, Louis KS. Withholding research results in academic life science. Evidence from a national survey of faculty. JAMA 1997; 277: 1224-1228.

Brooks 1985

Brooks TA. Private acts and public objects: an investigation of citer motivations. Journal of the American Society for Information Science 1985; 36: 223-229.

Burdett 2003

Burdett S, Stewart LA, Tierney JF. Publication bias and Meta-analyses: a practical example. International Journal of Technology Assessment in Health Care 2003; 19: 129-134.

Cantekin 1991

Cantekin EI, McGuire TW, Griffith TL. Antimicrobial therapy for otitits media with effusion ('secretory' otitits media). JAMA 1991; 266: 3309-3317.

Carter 2006

Carter AO, Griffin GH, Carter TP. A survey identified publication bias in the secondary literature. Journal of Clinical Epidemiology 2006; 59: 241-245.

Chan 2004a

Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;291: 2457-2465.

Chan 2004b

Chan AW, Krleža-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. Canadian Medical Association Journal 2004; 171: 735-840

Chan 2005

Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. BMJ 2005; 330: 753.

CLASP Collaborative Group 1994

CLASP Collaborative Group. CLASP: a randomized trial of low-dose aspirin for the prevention and treatment of pre-eclampsia among 9364 pregnant women. The Lancet 1994; 343: 619-629.

Cook 1993

Cook DJ, Guyatt GH, Ryan G, Clifton J, Buckingham L, Willan A, McIlroy W, Oxman AD. Should unpublished data be included in Meta-analyses? Current convictions and controversies. JAMA 1993; 269: 2749-2753.

Copas 1999

Copas J. What works?: selectivity models and Meta-analysis. Journal of the Royal Statistical Society Series A 1999; 162: 95-109.

Copas 2000

Copas JB, Shi JQ. Reanalysis of epidemiological evidence on lung cancer and passive smoking. BMJ 2000; 320: 417-418.

Cowley 1993

Cowley AJ, Skene A, Stainer K, Hampton JR. The effect of lorcainide on arrhythmias and survival in patients with acute myocardial infarction: an example of publication bias. International Journal of Cardiology 1993; 40: 161-166.

Davey Smith 1994

Davey Smith G, Egger M. Who benefits from medical interventions? Treating low risk patients can be a high risk strategy. BMJ 1994; 308: 72-74.

Dear 1992

Dear KBG, Begg CB. An approach to assessing publication bias prior to performing a Meta-analysis. Statistical Science 1992; 7: 237-245.

Decullier 2005

Decullier E, Lheritier V, Chapuis F. Fate of biomedical research protocols and publication bias in France: retrospective cohort study. BMJ 2005; 331: 19.

Decullier 2007

Decullier E, Chapuis F. Oral presentation bias: a retrospective cohort study. Journal of Epidemiology and Community Health 2007; 61: 190-193.

Deeks 2005

Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. Journal of Clinical Epidemiology 2005; 58: 882-893.

Dickersin 1992

Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. JAMA 1992; 263: 374-378.

Dickersin 1993

Dickersin K, Min YI. NIH clinical trials and publication bias. Online Journal of Current Clinical Trials 1993; Doc No 50.

Dickersin 1994

Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ 1994; 309: 1286-1291.

Dickersin 1997

Dickersin K. How important is publication bias? A synthesis of available data. AIDS Educucation and Prevention 1997; 9: 15-21.

Dickersin 2002

Dickersin K, Olson CM, Rennie D, Cook D, Flanagin A, Zhu Q, Reiling J, Pace B. Association between time interval to publication and statistical significance. JAMA 2002; 287: 2829-2831.

Dong 1997

Dong BJ, Hauck WW, Gambertoglio JG, Gee L, White JR, Bubp JL, Greenspan FS. Bioequivalence of generic and brand-name levothyroxine products in the treatment of hypothyroidism [see comments]. JAMA 1997; 277: 1205-1213.

Duval 2000

Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in Meta-analysis. Biometrics 2000; 56: 455-463.

Easterbrook 1991

Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. The Lancet 1991; 337: 867-872.

Egger 1997a

Egger M, Smith GD, Schneider M, Minder C. Bias in Meta-analysis detected by a simple, graphical test. BMJ 1997; 315: 629-634.

Egger 1997b

Egger M, Zellweger Z, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomized controlled trials published in English and German. The Lancet 1997; 350: 326-329.

Egger 2003

Egger M, Jüni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technology Assessment 2003; 7: 1.

Epstein 1990

Epstein WM. Confirmational response bias among social work journals. Science, Technology and Human Values 1990; 15: 9-37.

Ernst 1994

Ernst E, Resch KL. Reviewer bias: A blinded experimental study. Journal of Laboratory and Clinical Medicine 1994; 124: 178-182.

Fergusson 2000

Fergusson D, Laupacis A, Salmi LR, McAlister FA, Huet C. What should be included in metaanalyses? An exploration of methodological issues using the ISPOT Meta-analyses. International Journal of Technology Assessment in Health Care 2000; 16: 1109-1119.

Galandi 2006

Galandi D, Schwarzer G, Antes G. The demise of the randomised controlled trial: bibliometric study of the German-language health care literature, 1948 to 2004. BMC Medical Research Methodology 2006; 6: 30.

Givens 1997

Givens GH, Smith DD, Tweedie RL. Publication bias in Meta-analysis:a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. Statistical Science 1997; 12: 221-250.

Glasziou 1995

Glasziou PP, Iriwg LM. An evidence based approach to individualising treatment. BMJ 1995; 311: 1356-1359.

Godlee 1999

Godlee F, Dickersin K. Bias, subjectivity, chance, and conflict of interest in editorial decisions. In: Godlee F, Jefferson T (editors). Peer Review in Health Sciences. London (UK): BMJ Books, 1999.

Gøtzsche 1987

Gøtzsche PC. Reference bias in reports of drug trials. British Medical Journal (Clinical Research Edition) 1987; 295: 654-656.

Gøtzsche 1989

Gøtzsche PC. Multiple publication of reports of drug trials. European Journal of Clinical Pharmacology 1989; 36: 429-432.

Grégoire 1995

Grégoire G, Derderian F, LeLorier J. Selecting the language of the publications included in a metaanalysis: is there a Tower of Babel bias? Journal of Clinical Epidemiology 1995; 48: 159-163.

Harbord 2006

Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in Meta-analyses of controlled trials with binary endpoints. Statistics in Medicine 2006; 25: 3443-3457.

Hartling 2004

Hartling L, Craig WR, Russell K, Stevens K, Klassen TP. Factors influencing the publication of randomized controlled trials in child health research. Archives of Pediatrics and Adolescent Medicine 2004; 158: 983-987.

Hedges 1992

Hedges LV. Modeling publication selection effects in Meta-analysis. Statistical Science 1992; 7: 246-255.

Hemminki 1980

Hemminki E. Study of information submitted by drug companies to licensing authorities. British Medical Journal 1980; 280: 833-836.

Heres 2006

Heres S, Davis J, Maino K, Jetzinger E, Kissling W, Leucht S. Why olanzapine beats risperidone, risperidone beats quetiapine, and quetiapine beats olanzapine: an exploratory analysis of head-to-head comparison studies of second-generation antipsychotics. American Journal of Psychiatry 2006; 163: 185-194.

Hetherington 1989

Hetherington J, Dickersin K, Chalmers I, Meinert CL. Retrospective and prospective identification of unpublished controlled trials: lessons from a survey of obstetricians and pediatricians. Pediatrics 1989; 84: 374-380.

Hopewell 2004

Hopewell S. Impact of grey literature on systematic reviews of randomized trials (PhD thesis). University of Oxford, 2004.

Hopewell 2007a

Hopewell S, Clarke M, Stewart L, Tierney J. Time to publication for results of clinical trials. Cochrane Database of Systematic Reviews 2007, Issue 2. Art No: MR000011.

Hopewell 2007b

Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in Meta-analyses of randomized trials of health care interventions. Cochrane Database of Systematic Reviews 2007, Issue 2. Art No: MR000010.

Hopewell 2008

Hopewell S, Louden K, Clarke M, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database of SystematicReviews (to appear).

Huston 1996

Huston P, Moher D. Redundancy, disaggregation, and the integrity of medical research. The Lancet 1996; 347: 1024-1026.

Hutchison 1995

Hutchison BG, Oxman AD, Lloyd S. Comprehensiveness and bias in reporting clinical trials.

Canadian Family Physician 1995; 41: 1356-1360.

Ioannidis 1998

Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA 1998; 279: 281-286.

Ioannidis 2001

Ioannidis JP, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. JAMA 2001; 285: 437-443.

Ioannidis 2007a

Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. Clinical Trials 2007; 4: 245-253.

Ioannidis 2007b

Ioannidis JP, Trikalinos TA. The appropriateness of asymmetry tests for publication bias in metaanalyses: a large survey. Canadian Medical Association Journal 2007; 176: 1091-1096.

Irwig 1998

Irwig L, Macaskill P, Berry G, Glasziou P. Bias in Meta-analysis detected by a simple, graphical test. Graphical test is itself biased. BMJ 1998; 316: 470-471.

Iyengar 1988

Iyengar S, Greenhouse JB. Selection problems and the file drawer problem. Statistical Science 1988: 109-135.

Johansen 1999

Johansen HK, Gøtzsche PC. Problems in the design and reporting of trials of antifungal agents encountered during Meta-analysis [see comments]. JAMA 1999; 282: 1752-1759.

Jüni 2002

Jüni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in metaanalyses of controlled trials: empirical study. International Journal of Epidemiology 2002; 31: 115-123.

Kjaergard 2002

Kjaergard LL, Gluud C. Citation bias of hepato-biliary randomized clinical trials. Journal of Clinical Epidemiology 2002; 55: 407-410.

Lexchin 2003

Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. BMJ 2003; 326: 1167-1170.

Liebeskind 2006

Liebeskind DS, Kidwell CS, Sayre JW, Saver JL. Evidence of publication bias in reporting acute stroke clinical trials. Neurology 2006; 67: 973-979.

Light 1984

Light RJ, Pillemer DB. Summing up. The science of reviewing research (1). Cambridge (MA): Harvard University Press, 1984.

Macaskill 2001

Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in Meta-analysis. Statistics in Medicine 2001; 20: 641-654.

Mahoney 1977

Mahoney MJ. Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive Therapy and Research 1977; 1: 161-175.

Mandel 1987

Mandel EH, Rockette HE, Bluestone CD, Paradise JL, Nozza RJ. Efficacy of amoxicillin with and without decongestant-antihistamine for otitis media with effusion in children. New England Journal of Medicine 1987; 316: 432-437.

McAuley 2000

McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in Meta-analyses? The Lancet 2000; 356: 1228-1231.

Melander 2003

Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(i)ased medicine – selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. BMJ 2003; 326: 1171-1173.

Moher 1996

Moher D, Fortin P, Jadad AR, Jüni P, Klassen T, Le Lorier J, Liberati A, Linde K, Penna A. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. The Lancet 1996; 347: 363-366.

Moher 2000

Moher D, Pham B, Klassen TP, Schulz KF, Berlin JA, Jadad AR, Liberati A. What contributions do languages other than English make on the results of Meta-analyses? Journal of Clinical Epidemiology 2000; 53: 964-972.

Moher 2003

Moher D, Pham B, Lawson ML, Klassen TP. The inclusion of reports of randomised trials published in languages other than English in systematic reviews. Health Technology Assessment 2003; 7: 1-90.

Moher 2007

Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Medicine 2007; 4: e78.

Moscati 1994

Moscati R, Jehle D, Ellis D, Fiorello A, Landi M. Positive-outcome bias: comparison of emergency medicine and general medicine literatures. Academic Emergency Medicine 1994; 1: 267-271.

Olson 2002

Olson CM, Rennie D, Cook D, Dickersin K, Flanagin A, Hogan JW, Zhu Q, Reiling J, Pace B. Publication bias in editorial decision making. JAMA 2002; 287: 2825-2828.

Orwin 1983

Orwin RG. A fail-safe N for effect size in Meta-analysis. Journal of Educational Statistics 1983; 8: 157-159.

Peters 1982

Peters DP, Ceci SJ. Peer review practices of psychology journals: The fate of published articles, submitted again. The Behavioral and Brain Sciences 1982; 5: 187-255.

Peters 2006

Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in Meta-analysis. JAMA 2006; 295: 676-680.

Peters 2007

Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. Statistics in Medicine 2007; 26: 4544-4562.

Peters 2008

Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. The contour enhanced funnel plot: an aid to interpreting funnel asymmetry. Journal of Clinical Epidemiology 2008; 61: 991-996.

Pham 2005

Pham B, Klassen TP, Lawson ML, Moher D. Language of publication restrictions in systematic reviews gave different results depending on whether the intervention was conventional or complementary. Journal of Clinical Epidemiology 2005; 58: 769-776.

Pittler 2000

Pittler MH, Abbot NC, Harkness EF, Ernst E. Location bias in controlled clinical trials of complementary/alternative therapies. Journal of Clinical Epidemiology 2000; 53: 485-489.

Pocock 1987

Pocock S, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. New England Journal of Medicine 1987; 317: 426-432.

Poole 1999

Poole C, Greenland S. Random-effects Meta-analyses are not always conservative. American Journal of Epidemiology 1999; 150: 469-475.

Ravnskov 1992

Ravnskov U. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. BMJ 1992; 305: 15-19.

Rennie 1991

Rennie D. The Cantekin affair. JAMA 1991; 266: 3333-3337.

Rennie 1997

Rennie D. Thyroid Storms. JAMA 1997; 277: 1238-1243.

Rosenthal 1979

Rosenthal R. The 'file drawer problem' and tolerance for null results. Psychological Bulletin 1979; 86: 638-641.

Rücker 2008

Rücker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in Meta-analyses with binary outcomes. Statistics in Medicine 2008; 27: 746-763.

Sampson 2003

Sampson M, Barrowman NJ, Moher D, Klassen TP, Pham B, Platt R, St John PD, Viola R, Raina P. Should Meta-analysts search Embase in addition to Medline? Journal of Clinical Epidemiology 2003; 56: 943-955.

Scherer 2007

Scherer RW, Langenberg P, von Elm E. Full publication of results initially presented in abstracts. Cochrane Database of Systematic Reviews 2007, Issue 2. Art No: MR000005.

Schulz 1995

Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995; 273: 408-412.

Schwarzer 2007

Schwarzer G, Antes G, Schumacher M. A test for publication bias in Meta-analysis with sparse binary data. Statistics in Medicine 2007; 26: 721-733.

Simes 1987

Simes RJ. Confronting publication bias: a cohort design for Meta-analysis. Statistics in Medicine 1987; 6: 11-29.

Smith 1999

Smith R. What is publication? A continuum. BMJ 1999; 318: 142.

Sterling 1959

Sterling TD. Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. Journal of the American Statistical Association 1959; 54: 30-34.

Sterling 1995

Sterling TD, Rosenbaum WL, Weinkam JJ. Publication decisions revisted: The effect of the outcome of statistical tests on the decision to publish and vice versa. American Statatistician 1995; 49: 108-112.

Stern 1997

Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ 1997; 315: 640-645.

Sterne 2000

Sterne JAC, Gavaghan D, Egger M. Publication and related bias in Meta-analysis: Power of statistical tests and prevalence in the literature. Journal of Clinical Epidemiology 2000; 53: 1119-1129.

Sterne 2001

Sterne JAC, Egger M. Funnel plots for detecting bias in Meta-analysis: Guidelines on choice of axis. Journal of Clinical Epidemiology 2001; 54: 1046-1055.

Stuck 1998

Stuck AE, Rubenstein LZ, Wieland D. Bias in Meta-analysis detected by a simple, graphical test. Asymmetry detected in funnel plot was probably due to true heterogeneity. Letter. BMJ 1998; 316: 469-471.

Tang 2000

Tang JL, Liu JL. Misleading funnel plot for detection of bias in Meta-analysis. Journal of Clinical Epidemiology 2000; 53: 477-484.

Tannock 1996

Tannock IF. False-positive results in clinical trials: multiple significance tests and the problem of unreported comparisons. Journal of the National Cancer Institute 1996; 88: 206-207.

Taylor 1998

Taylor SJ, Tweedie RL. Practical estimates of the effect of publication bias in Meta-analysis. Australian Epidemiologist 1998; 5: 14-17.

Teo 1993

Teo KK, Yusuf S, Furberg CD. Effects of prophylactic antiarrhythmic drug therapy in acute myocardial infarction. An overview of results from randomized controlled trials [see comments]. JAMA 1993; 270: 1589-1595.

Terrin 2003

Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. Statistics in Medicine 2003; 22: 2113-2126.

Terrin 2005

Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. Journal of Clinical Epidemiology 2005; 58: 894-901.

Tetzlaff 2006

Tetzlaff J, Moher D, Pham B, Altman D. Survey of views on including grey literature in systematic reviews. 14th Cochrane Colloquium, Dublin (Ireland), 2006.

Tramèr 1997

Tramèr MR, Reynolds DJ, Moore RA, McQuay HJ. Impact of covert duplicate publication on metaanalysis: a case study. BMJ 1997; 315: 635-640.

Vevea 1995

Vevea JL, Hedges LV. A general linear model for estimating effect size in the presence of publication bias. Psychometrika 1995; 60: 419-435.

Vevea 2005

Vevea JL, Woods CM. Publication bias in research synthesis: sensitivity analysis using a priori weight functions. Psychological Methods 2005; 10: 428-443.

Vickers 1998

Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. Controlled Clinical Trials 1998; 19: 159-166.

Villar 1997

Villar J, Piaggio G, Carroli G, Donner A. Factors affecting the comparability of Meta-analyses and largest trials results in perinatology. Journal of Clinical Epidemiology 1997; 50: 997-1002.

Weber 1998

Weber EJ, Callaham ML, Wears RL, Barton C, Young G. Unpublished research from a medical specialty meeting: why investigators fail to publish. JAMA 1998; 280: 257-259.

Zarin 2005

Zarin DA, Tse T, Ide NC. Trial Registration at ClinicalTrials.gov between May and October 2005. New England Journal of Medicine 2005; 353: 2779-2787.

（钟大可译，岑啸、贾鹏丽、秦天强初审）

第十一章　结果报告和“结果汇总”表（SoFs 表）

作者：代表 Cochrane 应用与推荐方法学组和 Cochrane 统计方法学组的 Holger J Schünemann, Andrew D Oxman, Julian PT Higgins, Gunn E Vist, Paul Glasziou 和 Gordon H Guyatt。

本文选自工作手册5.0.2版本。有关如何引用它的指南，见11.10节。这些材料还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 图表有助于系统、清晰地呈现纳入的研究及其结果。
- 森林图是展示单个研究及 Meta 分析结果的标准方式。其可通过 RevMan 软件生成，在 Cochrane 系统评价正文中亦可选入部分森林图。
- “结果总结”表可提供有关以下内容的关键信息：证据质量、所检测干预措施的效应大小、给定比较所有重要结果的可得数据汇总。
- Cochrane 系统评价摘要的首要目标人群为卫生保健决策者（包括临床医生，信息来源广的用户和政策制定者）；此外，“简明概要”以一种直接明了的方式表述结果，该方式易为卫生保健用户所理解。

11.1 引言

系统评价的结果部分应以清晰、合理的顺序总结结果，且应明确针对该系统评价的目的。系统评价作者可以通过使用多种图表以更方便地呈现信息：

- “纳入研究特征”表（包括“偏倚风险”表）
- “数据和分析”（所有的数据表和森林图）
- 图形（文献筛选流程图、森林图、漏斗图、“偏倚风险”图和其它图形）
- “结果总结”表
- 附加表格

“纳入研究特征”表展示了单个研究的信息；“数据和分析”表和森林图展示了从单个研究中得出的结果且有可能还包括Meta分析；“结果汇总”表提供的是最重要结果的合成信息、数据及证据质量。此外，还应将系统评价的结果总结成一个摘要和一个简明概要。

“结果汇总”表是以上所有方式中最关键的，因此在本章中有一个大篇幅的部分专门对其进行阐述。我们讨论了可能与用户考虑干预措施时相关的重要结局的详述，该步骤我们认为Cochrane系统评价中常被忽略。我们还提供了“结果汇总”表的实例，并解释了所给表格的内容。第12章将会讨论结果的解读问题。

11.2 研究的检索和筛选结果

11.2.1 研究流程图

研究流程图用于说明研究结果和系统评价纳入研究的筛选过程。图11.2.a为研究流程图的示例，其采用了PRISMA声明(Liberati 2009)中的模板。使用PRISMA声明模板的流程图可用RevMan生成，且RevMan可生成结构灵活流程图。

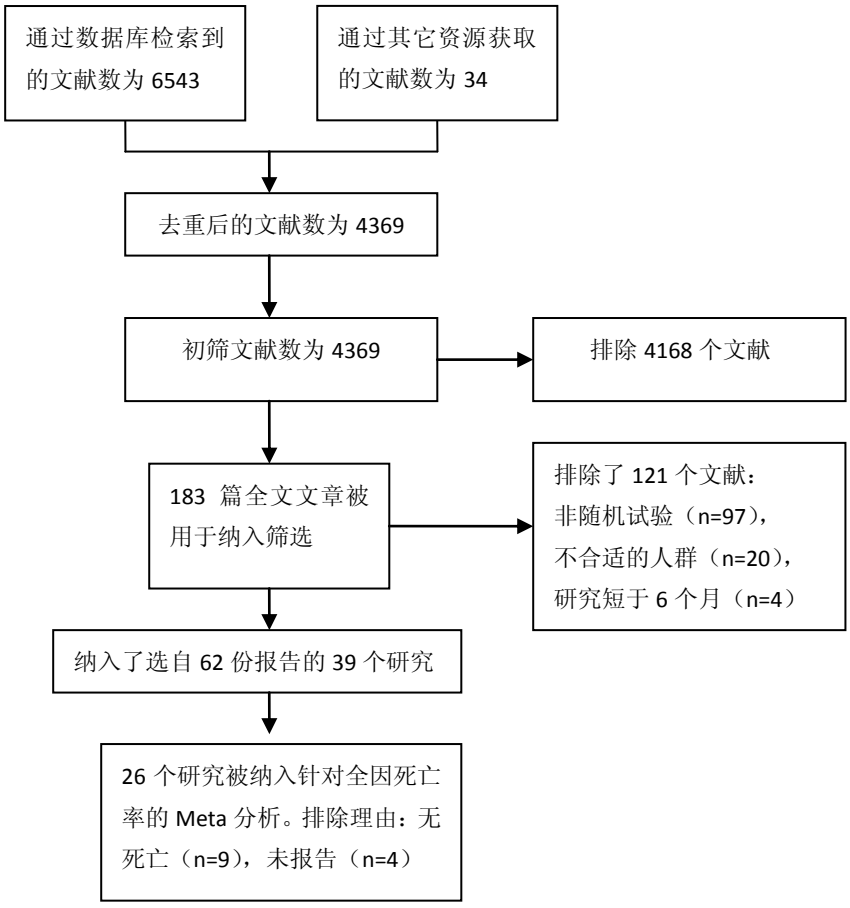
流程图的文字部分对研究和记录明确地区分开十分重要。文献是某研究的信息来源，如期刊文章、书本、网页或其它文献。研究则是指那些原始研究，典型的就是Cochrane系统评价中的随机试验。通常，流程图的开始部分为检索到的文献数量（大多数文献来自书目数据库）。剔重后，这些文献被反映为特定的研究（见第7章，7.2节），流程图转而强调研究数以体现该点。

流程图应该展示的内容有：

- 研究检出的文献数；
- 初筛（如通过标题和摘要）排除的文献数；
- 以全文形式获取的文献数；
- 查看全文后排除的文献和研究数，并提供简明原因；
- 符合系统评价纳入标准的研究数（并因此而用于定性合成）
- 用于生成主要结局的研究数。

关于制作研究流程图的进一步指导请参见PRISMA声明的解读文献(Liberati 2009)和一篇关于发表流程图的综述（Hind 2007）。

图11.2.a PRISMA研究流程图举例



11.2.2 “纳入研究特征”表

评价者必须明确所纳入研究的哪些特征可能与读者（使用者）相关。大多数特征很

可能在系统评价计划书中就已列出。系统评价作者应该至少在“纳入研究特征”表中体现以下信息：

研究方法：研究设计（说明该研究是否随机），必要的话，也应明确指出该研究的研究设计是否不同于平行随机设计（如：交叉或整群随机设计）；纳入研究的试验周期（如果没有在干预措施部分中说明）。注意：本部分不应包括偏倚风险评估，偏倚风险评估应该列在“偏倚风险”表中（见第8章，8.5节）。

受试人群：受试者入组时所处的环境（如急诊、门诊、住院等）；受试者健康状况的相关说明；年龄；性别；国籍。应提供足够的受试者信息，以方便系统评价的使用者判断该研究是否适用于其所面临的人群以及明确各个不同研究的受试者是否不同。

干预措施：研究中干预组的简明列表。如果可行的话，应提供每个干预的充分信息以便在临床实践中重复这种干预措施；对药物干预来说，则应提供药物的详细名称、剂量、服药次数、给药方式（如果不显而易见）、持续时间（如果没有在方法部分说明）；对非药物干预来说，则应需提供其相应的信息和说明。

结局：为以下两种内容中一种的简明列表：（ｉ）系统评价计划观察的结局及结局测量时间点；（ii）纳入研究所测量（或报告）的结局和时间点。研究结果不应出现在此处（也不能出现在“纳入研究特征”表的其它部位）。

备注：进一步说明系统评价作者对各研究中未被上述分类涵盖的内容。需要注意的是，对偏倚风险的评价应该出现在“偏倚风险”表中。

此外，还可以在“纳入研究特征”表中添加最多三个附加版块。如果合适，建议系统评价作者利用一个附加版块来呈现各研究的基金来源。

11.3 数据和分析

11.3.1 系统评价的“数据和分析”部分

Cochrane系统评价的“数据和分析”部分是结果的详细阐述。其中包括结果资料（数字或文字），森林图和Meta分析结果。“数据和分析”的基本内容是比较、结果和亚组（任选）的表格。表格中列出的分析通常为一个结果表（“其它数据”表），更为常见的是附有森林图的数据表。“数据和分析”表包含在Cochrane系统评价的全文中。然而，有些形式的已发表系统评价可能会省去森林图和“其它数据”表（出现在附件中），所以它

们通常被当做补充材料，因而关键结果应呈现在系统评价正文的“结果”部分。最终发表的系统评价中通常有一个包含所有分析的总结表（其中包括每个比较中每个结局下亚组的研究数和Meta分析结果）。系统评价应以图的形式呈现从“数据和分析”资源中选入的最重要的森林图，且应该在“结果”部分进行说明（见11.4.2节）。

11.3.2 森林图

森林图展示了单个研究和Meta分析的效应估计值及可信区间（Lewis 2001）。每个研究都由位于干预效果点估计值位置的方块来代表，同时一条横线分别向该方块的两边延伸出去。方块的面积代表在Meta分析中该研究被赋予的权重，而横线代表可信区间（通常为95%可信区间）。方块面积和可信区间传达的信息是相似的，但在森林图中两者的作用却不同。可信区间描述的是与研究结果相符的干预效果的范围，且能表示每个研究是否有统计学意义。较大的方块意味着较大权重的研究（通常为可信区间较窄的研究），这些研究也决定了最终合并的结果。

11.3.2.1 RevMan 中的森林图

RevMan为Cochrane系统评价的“数据和分析”部分制作森林图提供了一个灵活的框架。Cochrane森林图的的构成见框11.3.a，且图11.3.a给出了一个RevMan中的例子，所使用的是压力袜预防航空乘客深静脉血栓的系统评价中的结果（Clarke 2006）。RevMan使用教程可从RevMan中获得（可登录www.cc-ims.net获取）

RevMan提供了多种选择以改变分析方法（如：固定效应和随机效应Meta分析，或使用不同的效应指标；见第9章，9.4节）和图表制作（如坐标轴的尺度和研究的排序）。输入RevMan的每组数据对应的森林图可被自动整合到正式发表的Cochrane系统评价中。除非进行设置，否则将按默认设定显示。RevMan的默认设定如下：二分类变量默认的是Mantel-Haenszel比值比（OR）；连续性变量默认的是均数差固定效应模型Meta分析；“O–E和方差”结果默认的是Peto比值比（OR）；而倒方差结果默认的是固定效应模型Meta分析（见第9章，9.4节）。当在RevMan中设置或编辑结果时，作者应跳过和正文中报告的结果不相关的默认设置。这样才能保证显示的结果和文中描述的结果是一致的。此外，应选择坐标轴尺度以使点估计值能在森林图中显示出来（即使不能显示出整个可信区间，也应尽可能多的显示）。

按照Cochrane系统评价数据库(CDSR)的既往惯例，二分类变量关注的是不利结果指标，故RR和OR小于1（RD小于0）就意味着试验组干预优于对照组干预。这将导致当效应估计值落在森林图垂线的左边时就认为试验组干预有益。因其缺乏普适性，因此不再鼓励沿袭这一惯例。更好的方法是，在森林图的坐标轴上进行方向标记，以此清晰地说明在横线的哪边表示哪种干预有益。RevMan允许作者为每个结局对应的“试验”和“对照”组分别制定标签。而其定义的标签将被用于Cochrane系统评价数据库。因此，有必要弄清楚图形是如何制作的，以及如何对其进行解释。这对于测量量表数据尤其重要，因为对读者而言，他们并不总是清楚哪个方向意味着会损害健康。

当没有研究纳入时，不应制作森林图；当针对某特定结局只纳入了一个研究时，也不主张制作森林图。为显示仅在单个研究中被调查的结果，作者可使用为每个结果设置了亚组的森林图（确保未启用数据合并项）。此外，单个研究的结果在附加的表格（见11.6节）中列出会更为方便。

框11.3.a　Cochrane森林图所提供的详细信息

二分类变量和“O–E 和方差”结果的森林图默认显示如下：

1.对应每个研究的原始数据（2×2 的表格）；

2.所选效应测量的点估计值及可信区间，以方块、横线和文字表示；

3.采用指定效应测量和方法（固定或随机效应）得到的各亚组 Meta 分析，以菱形和文字的表示；

4.试验干预和对照干预组的受试者总数和事件发生总数；

5.异质性检验得到的统计量（随机效应 Meta 分析时的研究间变异（Tau^2 或 τ^2）、卡方检验、I^2 统计量，如果有亚组则还有亚组间的差异）；

6.对总体效应的检验结果（随机效应 Meta 分析的总平均效应）；

7.各研究被赋予的百分比权重。

注：以上 3-7 项只有在合并数据时才会显示。此外，亚组间差异检验在 Mantel-Haenszel 分析时不会显示。而对于“O–E 和方差”结局，亦可显示 O–E 值和 V 值。

连续性变量结局的森林图默认显示如下：

1.每个研究各组结果的原始数据（均数、标准差和样本量）；

2.所选效应测量的点估计值及可信区间，以方块、横线和文字表示；

3.采用指定效应测量和方法（固定或随机效应）得到的各亚组 Meta 分析，以菱形和文字的表示；

4.试验干预和对照干预组的受试者总数和事件发生总数；

5.异质性检验得到的统计量（随机效应 Meta 分析时的研究间变异（Tau^2 或 τ^2）、卡方检验、I^2 统计量，如果有亚组则还有亚组间的差异）；

6.对总体效应的检验结果（随机效应 Meta 分析的总平均效应）；

7.各研究被赋予的百分比权重。

注：以上 3-7 项只有在合并数据时才会显示。

倒方差法的森林图默认显示如下：

1.作者录入的各研究的数据汇总（对于比例的测量，结果会描绘在自然对数坐标上）；

2.所选效应测量的点估计值及可信区间，以方块和横线及以文字的形式（对于比例的测量，图形将被描绘在自然坐标而非对数坐标上）；

3.采用指定方法（固定或随机效应）的亚组 Meta 分析，以菱形和文字表示；

4.异质性检验得到的统计量（随机效应 Meta 分析时的研究间变异（Tau^2 或 τ^2）、卡方检验、I^2 统计量，如果有亚组则还有亚组间的差异）；

5.对总体效应的检验（随机效应 Meta 分析的总平均效应）；

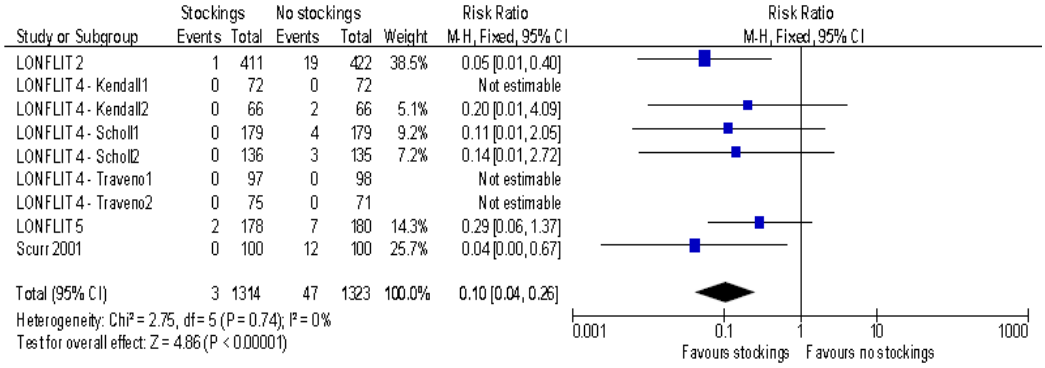
6.各研究被赋予的百分比权重。

注：以上 3-6 项只有在合并数据时才会显示。可补充输入试验组和对照组的样本量。其录入应视研究设计而定。样本量不会包含在分析中，但如果录入则显示为以下两项：

7.各研究试验组和对照组的受试者人数；

8.试验组和对照组的受试者总数。

图11.3.a　RevMan森林图实例



11.3.3 其它数据表

“数据和分析”部分可有“其它数据”这一结局类型。单个试验结果可以在此处以纯文本的形式录入。这一选项非常适合录入非标准的汇总统计量，如中位值，或在倒方差法结局类型下录入的估计值及标准误对应的基础数据（如，交叉试验的均数及标准差）。

11.4 图

11.4.1 图的类型

在Cochrane系统评价的正文部分，可包含以下三种类型的图。

1. 所有从RevMan的“数据和分析”中得到的森林图（见11.3.2节）。
2. 所有从RevMan的“数据和分析”中得到的漏斗图（见第10章，10.4.1节）。
3. 附加图。

由于在Cochrane系统评价的一些发表形式中，可能不包含“数据和分析”部分，因此作者应将最重要的森林图放入系统评价的正文部分，并在正文的相关部分提及这些图。然而需要注意的是，所有“数据和分析”森林图所得的Meta分析和亚组分析结果在Cochrane系统评价的所有发表形式中都将以表格的形式呈现。

一般说来，图形能够以清晰、系统地方式将单个研究和Meta分析的结果呈现出来。然而，附有大量图形的系统评价却常常很难理解，尤其是当各图形仅含极少信息时。许多科学期刊限制一篇论文中的图表数量在6个左右，同样地，大多数Cochrane评价组也考虑了以上问题。

所有图的重要结果应在系统评价正文的结果部分进行概述。无论在系统评价正文中何处取用图中的数值结果，作者都应清楚地解释其意义及其含义，并说明该数值是引自哪张图。

11.4.2 以图片形式选择RevMan分析

来自“数据和分析”的森林图和漏斗图可被选入并以图的形式纳入已发表的Cochrane系统评价。对于主要结局，包含所有研究和研究数据详细信息的森林图通常以图的形式纳入。如果纳入了足够多的研究，对一个或多个主要结局的漏斗图可作为这些森林图的重要补充（见第10章，10.4.1节）。

11.4.3 附加图

虽然RevMan可以制作森林图和漏斗图，但也可以在系统评价中纳入其它类型的图。例如：

1. “系统评价再评价”森林图，该类型森林图中每条线代表一个Meta分析而不是一个研究（例如，表示多个亚组分析或敏感性分析）；

2. Meta回归分析图；

3. L’Abbé图。

这些图可通过非RevMan的其它软件制作并作为“附加”图纳入。图片和图表可以同样的方式在Cochrane系统评价的其它部分纳入使用。

附加图很少是必需的，且不应额外绘制可用RevMan绘制的森林图。若需附图，则应使用能绘制达到出版质量的图片的统计软件包来绘制，如Stata、SAS、SPSS、S-Plus或专门的Meta分析软件。一般用途的电子表格程序既无适当的灵活性，也不能进行高质量的输出。

另有一篇文献——《对Cochrane系统评价中图表的思考和建议：统计图》提供了针对各种数据类型的附加图表的全面指导（可从www.cochrane.org/resources/handbook获得）。该文献含有对上述所列图表和几种其它插图的描述和建议。作者在提交一篇包含附图的系统评价前，应参考该文献。在一篇Cochrane系统评价提交给CDSR之前，其所有附图都应由统计学编辑或顾问进行评价。作者应该明白，附图通常很大并占用Cochrane图书馆宝贵的存储空间。关于附图技术方面的指导可从RevMan文件中获得（http://www.cc-ims.net）。

虽然在RevMan中插入附图的功能在技术上允许作者将表格作为图片格式插入。但是由于图片文件会占据很大的存储空间，因此不鼓励作者采用这种方式。相应地，为达到以上目的，可建议作者改用附加表格功能。

11.5 “结果总结”表（SoFs表）

11.5.1 “结果总结”表简介

“结果总结”表以简单易懂的表格形式呈现了一篇系统评价的主要结果。具体而言，即它们提供了关于证据质量、所检测干预措施的效应大小并汇总可得到的主要结局数据等关键信息。对多数系统评价来说，最好能有一个简单的“结果总结”表。某些系统评价可能包括不止一个SoFs表，如：涉及一个以上的主要比较或针对的群体存在显著差异的系统评价。在CDSR中，最重要的“结果总结”表将先于背景部分，位于系统评价的开头。其它“结果总结”表则将会在结果和讨论部分之间呈现。

在系统评价之初选择纳入（i）系统评价和（ii）“结果总结”表的结局指标时，就应

计划好“结果总结”表。因为这是至关重要的一个步骤，也是在经典Cochrane系统评价中很典型但非正式要求的，因此接下来我们将对结局指标筛选的问题进行综述。

11.5.2 “结果总结”表的结局指标筛选

Cochrane系统评价首先需要拟定一个系统评价问题并列出对患者和其他决策者而言重要的主要结局，以确保产生的是最有用的信息。对系统评价计划书的评审和反馈有助于加强以上过程。

重要结局可能包括那些众所周知的事件，如：死亡率和重要的发病率（例如，卒中和心肌梗死）。然而，它们也可能包括常见的轻微副作用和罕见的严重副作用、症状、生活质量、治疗负担及资源问题（费用）。负担包括依从某种干预时所需的而患者或其护理者（如，家人）可能不喜欢的要求，例如不得不接受更频繁的检查或某些干预措施所要求的生活方式限制。

通常，当确定涵盖所有决策所需的患者重要结局的问题时，系统评价作者将会面临随机试验并未报告以上所有结局的困境。对于不良结局尤其如此。例如，随机试验可能提供了关于预期效应及常见的较轻副作用的数据，但不会针对像自杀倾向这样的罕见不良结局的相对危险度（RR）。第14章将讨论合理处理不良反应的策略。为获取所有重要结局的数据，可能有必要查阅观察性研究的结果。

如果一篇系统评价仅纳入了随机试验，则在系统评价的限制下要针对所有重要结局不太可能。系统评价作者应了解这些局限性，并告知读者。

承担为所有相关结局收集和汇总最佳证据这一任务的系统评价者可能会面对一系列挑战。例如进行危害分析的研究，其受试者可能不同于进行疗效分析研究的受试者。因此，系统评价者可能需要考虑观察性研究的受试者与随机试验的受试者有多大差异。这将会影响证据的质量，因为其涉及到直接性问题（见第12章，12.2节）。如系统评价者在系统评价中未包括这些重要结局的信息时，应进行说明。第13章有对这些问题的进一步讨论。

11.5.3 “结果总结”表的通用模板

尽管对于某些系统评价来说，有充分的理由去修改“结果总结”表的格式，但是为了确保系统评价的一致性和易用性、对决策者所需的最重要信息的纳入及这些信息的最

佳呈现，已为“结果总结”表制定出了一个标准格式。故标准Cochrane“结果总结”表采用了固定的格式，并包含以下六个要素（见图11.5.a）。

1. 所有利弊方面重要结局的列表；
2. 对这些结局一般情况下的负担测量（如，对照组的危险度或均值）
3. 绝对效应和相对效应的大小（如果均适用）。
4. 针对这些结局的受试人数和研究数。
5. 对每个结局证据群的总体质量分级（其可因结局而异）。
6. 备注栏。

作为衡量效应大小的方式，对二分类变量，表格通常同时提供相对指标（例如RR值或OR值）和绝对危险度指标。对于其它类型的资料，或仅提供绝对指标（例如连续性变量的均数差），或仅提供相对指标（例如时间-事件资料的风险比）。然而，如果可能，相对和绝对指标都应提供。含有一个以上主要比较的系统评价，要求将每个比较的“结果总结”表分开。图11.5.a为“结果总结”表的示例。

对“结果总结”表内容的详细描述见11.5.6节。

图11.5.a “结果总结”表举例

结果总结：

长途飞行者穿压力袜与不穿压力袜的比较						
患者或人群：任何长途（>6 小时）飞行者 背景：国际航空旅行 干预：压力袜 [1] 对照：无压力袜						
结局指标	危险估计值* (95% CI)		相对效应 (95% CI)	受试者人数（研究数）	证据质量 (GRADE)	备注
	对照危险 无压力袜	干预危险 压力袜				
有症状深静脉血栓 (DVT)	见备注	见备注	不可估	2821(9)	见备注	这些研究中无受试者发展成有症状深静脉血栓
无症状深静脉血栓	低风险人群 [2]		RR 0.10 (0.04，0.26)	2637(9)	⊕⊕⊕⊕ 高	
	10/1000	1/1000(0-3)				
	高风险人群 [2]					
	30/1000	3/1000(1-8)				
浅静脉血栓	13/1000	6/1000(2-15)	RR 0.45 (0.18，1.13)	1804(8)	⊕⊕⊕○ 中 [3]	
水肿 飞行后测量值，评分等级为：0 分（无水肿）~10 分（极度水肿）	对照组平均水肿评分为 6 -9	干预组平均水肿评分降低 4.7 分 [95% CI（–4.9，–4.5）]		1246(6)	⊕⊕○○ 低 [4]	
肺栓塞	见备注	见备注	不可估	2821(9)	见备注	这些研究中无受试者发展为肺栓塞。[5]
死亡	见备注	见备注	不可估	2821(9)	见备注	这些研究中无受试者死亡。
不良反应	见备注	见备注	不可估	1182(4)	见备注	其中的 4 个研究对压力袜耐受性的描述为非常好，即未出现副作用 [6]
*对照危险的基本描述见脚注。干预组的风险（及其 95%CI）是基于干预组的对照风险和干预的相对效应（及其 95%CI）。 CI：可信区间；RR：相对危险度；GRADE: GRADE 证据分级工作组（见注释）。						

[1] 该系统评价纳入的 9 个试验中所有压力袜均为膝下压力袜。4 个试验中，压力袜在踝部的压力值为 20-30 mmHg。在另外 4 个试验中则为 10-20 mmHg。其压力袜的尺寸各异。如果压力袜将膝盖包裹太紧，就会妨碍自发的静脉回流从而使血液淤积在膝部。因此压力袜应该合适。此外，在长途飞行中，太紧的压力袜可能会勒破皮肤甚至可能导致溃疡和增加深静脉血栓的风险。有的压力袜可能比一般的袜子厚一些，可能对较小的鞋有一定限制。上飞机前在家进行试穿以确保压力袜穿着舒适不失为一个好主意。多数试验中，受试者在飞行前 2-3 小时就穿上了压力袜。最后，压力袜的实用性和费用也可不同。

[2] 两个试验招募了高风险人群（其定义为之前发生过深静脉血栓、凝血功能障碍、重度肥胖的人及由于骨关节问题而活动受限、前两年内患有肿瘤性疾病和大量静脉曲张的人，或在其中一个研究中，身高超过 190 cm 和体重超过 90 kg 的受试者也定义为高风险人群）。7 个排除了高风险受试者的试验发生率为 1.45%，2 个纳入高风险受试者（至少有一个危险因素）的发生率为 2.43%。我们对其四舍五入后其分别为 10/1000 和 30/1000。

[3] 可信区间跨越无效线不多。

[4] 水肿的测量未经验证且未对干预采用盲法。所有研究由同样的研究者实施。

[5] 如果事件数很少或无事件且受试者数很大，则对证据质量的判断（尤其是对精确性的判断）可基于绝对效应。如果恰当地评估了结局且在 2821 例研究的受试者中事件未发生，该处质量分级可判定为“高”。

[6] 除 1 个试验中有 4 例膝部静脉曲张处的浅静脉血栓（由压力袜上口压力造成），其它试验均未报告不良反应。

11.5.4 制作“结果总结”表

附加软件GRADEprofiler (GRADEpro)，可协助系统评价作者准备“结果总结”表。GRADEpro可从RevMan导入资料，并可将其与用户录入的对照组风险整合，从而得出干预措施相关的相对效应和绝对效应。此外，该软件还可以协助用户参与GRADE分级的全过程（详见GRADEpro软件中的具体帮助文件），且能生成一个易于以“结果总结”表形式导入RevMan的表格。该表以特殊表格（见11.6节）形式导入，不能在RevMan中对其进行修改。系统评价制作者可在RevMan中选择性创建他们自己的表格。

11.5.5 “结果总结”表中的统计学因素

以下我们描述的是如何获得二分类变量的绝对效应和相对效应指标。RR值、OR值和RD值是比较两组二分类结局资料的不同方式（见第9章，9.2.2节）。而且，根据分析的关注点代表的事件（如“是”或“否”）的不同，有两种不同的RR值（见9.2.2.5）。在存在非零干预效果的情况下，如果研究间对照组风险存在差异，就不可能有一个以上的上述指标在每个研究中完全一样。流行病学理论一直推测道，在不同的情况下应用，相对效应指标较绝对效应指标一致性好。目前已有实证证据支持这一推测（Engels 2000, Deeks 2001）。基于上述原因，Meta分析通常应使用RR值或OR值作为效应指标（见第9章，9.4.4.4节）。相应地，单个相对效应估计值可能比单个绝对效应估计值更适用于总结。如果相对效应在研究间的确具有一致性，那么不同的对照组风险将说明了不同的绝对获益。例如，如果RR值一致为0.75，那么就意味着在干预组中治疗措施可将对照组中为80%的危险度降低到60%（绝对危险降低率为20%），也可以意味着在干预组中可将对照组为20%的危险度减少到15%（绝对危险降低率为5%）。

制定“结果总结”表是采取基于一致的相对效应这样一个假设的。因此考虑到该效应值对不同对照组风险对应的不同含义十分重要。对所有对照组风险，都可以从Meta分析的RR值或OR值估计一个相应的干预组风险。注意：在“干预风险”一栏所提供的数值是和其旁的“对照风险”栏对应的。

根据Meta分析的相对危险度（RR）和假设对照组危险度（assumed control risk，ACR），可算得相应的干预组危险度为：

相应的干预组危险（每千人）=1000×ACR×RR。

举例如下：在图11.3.a中，Meta分析的危险比为RR=0.10（95%CI　0.04，0.26）。假

设对照组危险为ACR=10/1000=0.01，因此我们可以得到：

相应的干预组危险（每千人）=1000×0.01×0.10=1，结果与图11.5中提供的一致。

根据Meta分析的比值比（OR）和假设对照组危险度（ACR），可算得相应的干预组危险度为：

$$\text{相应干预组危险度（每千人）} = 1000 \times \left(\frac{\text{OR}\times\text{ACR}}{1-\text{ACR}+(\text{OR}\times\text{ACR})}\right)$$

用RR值或OR值可信区间的上下限替换RR值或OR值分别代入上述公式就可得到相应干预危险的可信区间上下限（例如，在上述例子中，分别用0.04替换0.1，用0.26替换0.1）。该可信区间未体现假设对照危险中的不确定性。

在处理相对危险度（RR）时，使“结果总结”表中使用的“事件”定义与Meta分析中所用的“事件”定义一致很关键。例如，如果Meta分析将“存活”而不是“死亡”作为事件，那么“结果总结”表中的假设危险和相应危险也必须指的是“存活”。

在（少有的）有明确理由假设Meta分析中存在一致危险差的情况下，原则上就能可呈现这一值为相关的“假设危险”及其相应危险，且对于每个假设危险可呈现相应的（不同的）相对效应。

11.5.6 “结果总结”表的详细内容

11.5.6.1 表格的标题和表头

每个“结果总结”表的标题都应指明其临床问题，标题应在说明人群的基础上，清楚、准确地说明所实施的干预及其对照。在表11.5.a中，人群为长途飞行者，干预为穿压力袜，对照是不穿压力袜。

每个“结果总结”表的首行都应提供以下“表头”信息：

患者或人群：这一部分进一步阐明目标人群（可能为亚组人群），并在理论上可阐明在某种治疗下最重要的不利结局的危险度大小。例如，长途飞行患者可能有不同程度的深静脉血栓风险；使用SSRIs的患者可能有不同的副作用风险；房颤患者可能有低度（<1%）、中度（1%-4%）或高度（>4%）的卒中年危险度。

背景：该部分应阐明可能会限制结果总结应用于其它环境的研究开展背景的特征；如，在欧洲和北美的基层医疗。

干预：试验干预措施。

对照：对照措施（包括未采取特定的治疗）。

11.5.6.2 结局

“结果总结”表的各行应包括所有对决策至关重要的利弊结局指标（按重要性依次列出），最多可列7个结局指标。如果系统评价中结局指标过多，作者需去掉较不重要的结局。还应提供量表和时间范围的详细信息。在计划书制定阶段和进行系统评价前，作者就应确定哪些结局对“结果总结”表重要。然而，系统评价制作者留意这样一种可能性，即某结局（例如，严重不良反应）的重要性可能仅在计划书完成或进行了Meta分析之后才知道，为将其纳入“结果总结”表应该做适当的处理。注意：无论相关数据是否可得，作者都应在表中将这些重要结局列出来。

严重不良事件应被纳入表格，但也可结合小的不良事件，然后在脚注中进行描述（注意：把不良事件放到一起并不恰当，除非已知他们各不相关）。多个时间点是一个特殊问题。通常，为保持表格的简易，仅有对决策重要的结局才应该在多个时间点呈现。其它结局均应在一个同样的时间点呈现。

连续性结局指标可在“结果总结”表中显示；系统评价制作者应尽量将其向目标读者解释清楚（见第12章，12.6节）。这要求其单位都是清晰并易于解释的，例如，疼痛天数，或头痛的频率。然而，许多测量工具对非专科临床医师或患者来说并不容易解释，例如，白氏抑郁症量表的得分或生活质量得分。对于以上情况，更容易解释的方式可能是将连续性结局转换为二分类结局，如>50%的提高（见第12章，12.6节）。

11.5.6.3 描述性比较风险（illustrative comparative risks）1：假定风险（对照干预）

作者应提供受试者接受对照干预的三个典型的危险度。推荐其以每1000人中（常用频率）发生事件的人数的形式呈现。超过1000的选择可能适用于罕见事件，100则可能适用于发生率更高的事件。假设对照危险可能基于对不同患者人群或不同随访时间的典型危险度的评估。理想情况下，危险度反映的是那些临床医生根据显示特点就能轻松区分的组别。脚注应阐明每个对照组风险的来源和原理，包括相应的时间范围。在表11.5.a中，临床医生很容易就能区分有和无深静脉血栓危险因素的个体。如果基线危险差异很小，系统评价制作者可使用各研究对照组的中位危险度。

11.5.6.4 描述性比较风险 2：相应风险（试验干预）

对二分类结局，应在假设对照危险栏后提供各假设对照危险相应的绝对危险度及其可信区间。该试验干预的绝对危险度通常由在相对效应一栏的Meta分析结果推导而来

（见11.5.6.5节）。公式见11.5.5节。系统评价作者应以对照干预假设危险同样的格式呈现绝对效应（见11.5.6.3节），如：每1000人中发生该事件的人数。

对连续性结局，均数差或标准化均数差应与其可信区间一起呈现。这些通常直接从Meta分析中获得。应如表11.5.a所示，采用解释性文字阐明其意义。

11.5.6.5 相对效应（95%CI）

相对效应通常为危险比（RR）或比值比（OR）（或偶尔为风险比）及其95%可信区间，它们均来自基于同一效应指标而作的Meta分析。当对照干预危险度较低且效应较小时，RR值或OR值是相似的，但随着危险度和效应的增加，两者的差异将增大。Meta分析可包含对固定或随机效应的假设，具体视系统评价作者认为哪种效应恰当而定。

11.5.6.6 受试者人数（研究数）

该栏应包括每个结局所纳入的研究所评价的受试者人数，以及提供这些受试者的相应研究的数量。

11.5.6.7 证据质量（GRADE）

作者将以“高”、“中”、“低”或“极低”来评价证据本身的质量。这是一个基于判断力的问题，但该判断过程是根据一个透明的框架进行的，详细介绍见第12章（12.2节）。例如，如果是对数个低偏倚风险的随机试验的总结，则其质量将为“高”，但如果涉及设计或实施、不精确性、不一致性、间接性或报告偏倚，则质量将会降级。作者应采用由GRADE工作组制定的特定证据分级系统（GRADE工作组2004），其详细介绍见第12章（12.2节）。应采用“结果总结”表的脚注或备注栏使除“高”质量以外的其它判断透明化（见表11.5.a）。

11.5.6.8 备注

备注栏旨在提供附加注释以帮助解释该行所列出的信息或数据。例如，其可能有关结局指标的可靠性或对于效应大小相关的变量的呈现。有关结果的重要说明应置于此处。不是每一行都需要备注，因此当该行没有内容需要附以备注时，最好保留空白。

11.6 附加表格

附加表格的特性为表格的创建提供了灵活的方式，其可展示试验、Meta分析结果及其它Meta分析性调查（如Meta回归分析）的结果。所有附加表格的重要结果应在系统评价正文的结果部分进行总结。

11.7 在正文中呈现结果

11.7.1 Meta分析结果

结果部分应按照系统评价方案中所规定的比较和结局顺序来安排，以明确阐述系统评价的目的。正文应以富有逻辑且系统的方式呈现全部结果：其没必要过分依赖表和图，或不断地提及图表以获得系统评价结果的清晰概况。更恰当的方式是，表格应作为提供更详细内容的附加资源使用。然而，应避免在正文部分过度重复已在表和图中提供的数据。

对于事后比较分析（post hoc analyses）和相对次要的问题，就算有足够的数据支撑，也不应过分强调。事后比较分析应一直遵循该原则。虽然分析方法本身应在方法部分予以描述，作者同时还应在结果部分再次说明每个引证结果采用的分析方法（特别是效应指标的选择，有利效应的方向和所用的Meta分析模型）。结果总应包括对不确定性的估计，如95%可信区间。摘要应只总结最重要的比较和结局，而非选择性报告那些差异最显著的结果。这也有助于显示分析所基于的信息量（研究数和受试者人数）。

每个图和附加表格都应在正文中明确提及。当提及图、表或“数据和分析”森林图（未被选为图）中的结果时，应在正文中提及相应的图、表或分析。

作者应考虑以易解释的格式呈现结果。例如，比值比和标准化均数差自身并不能直接用于临床实践，但却能以更易理解的方式进行转述。见第12章（12.5和12.6节）。

11.7.2 无Meta分析的结果

Meta分析方法可以量化效应方向、效应大小和效应的一致性（见第9章，9.1节）。如果不能获得适当的数值资料用于Meta分析，或不宜采用Meta分析时，还是应该对这些资料进行检查从而对所有可得数据进行综合的评价。

对证据的描述性评估可能富有挑战，特别是当系统评价纳入了大量研究时；当研究本身就检测了复杂的干预和结局时；当干预效应有很大差异时。效应类型和研究间的相似性或差异可能因此而不那么显而易见。采用系统的方法来呈现结果对于使系统评价结果能被理解很重要。如对每个研究的结果都给出一段描述，则描述应保持一致，包括每个研究同样的信息元素，按同样的顺序进行描述。如果系统评价纳入了大量研究（大于20），则提倡对研究进行分组或分类（如：根据干预类型、人群、背景等），这能使结果的叙述性描述过程更易控制。这也能帮助确定在组内和组间产生的结果的类型。

11.8 撰写摘要

所有完整的系统评价都必须包括一个不超过400字的摘要。摘要应在不遗漏重要内容的前提下尽可能保持简明。Cochrane系统评价的摘要被MEDLINE和SCI所收录，且通过网络可免费获得。因此它们可以被以单个完整文献的形式阅读，这是很关键的。

摘要应总结系统评价的关键方法、结果和结论，且不应包含系统评价中没有的信息。摘要不应包含对系统评价其它部分（如参考文献、研究、表格和图）的链接。一个假定的摘要举例见框11.8.a。

摘要应主要针对卫生保健决策者（临床医生、知证用户和政策制定者），而不应只针对研究者。术语应易于被普通而非专业卫生保健读者理解。除了那些广为人知的缩写（如HIV）外，应避免缩写。在有必要使用缩写时，在首次使用时应将全称拼写出来（括号内给出缩写）。药名和干预措施名称应尽可能使用国际广泛认可的。不应使用商品名。

每个摘要的标题下面应包括如下内容：

背景：应该为一两个解释背景或详细阐述系统评价的目的及原理的句子。如果该系统评价是对已有系统评价的更新，附一句诸如“这是对XX年首次发表的Cochrane系统评价的更新，且其已于XX年进行了更新”这样的话很有用。

目的：这部分应最好以一句话来明确阐述系统评价的主要目的，并紧扣系统评价正文主要内容的目的。其格式可能为以下形式：“评价在［某人群、疾病或问题和特定环境］中［某干预或比较］对［某卫生问题］的效果”。

检索方法：这部分应列出资源和最近检索日期，对每个资源，使用主动语态，即“我们检索了……”或，如果仅有一个作者，可使用被动语态，例如，“数据库X、Y、Z被

检索”。检索词不应在此处列出。如果使用了CRG专题注册库，则应该以“Cochrane X组专题注册库”的形式首先列出。其它数据库的罗列顺序应该是Cochrane对照试验注册中心，MEDLINE、EMBASE和其它数据库。还应给出每个数据库检索的日期范围。对于Cochrane对照试验注册中心，其形式应为“Cochrane对照试验注册中心（Cochrane图书馆，2007年第1期）”。对于绝大多数其它的数据库，如MEDLINE，其形式应为“MEDLINE（1966年1月至2006年12月）”。对相关引文进行书目文献检索可写成通用的“参考文献列表”。如果存在对语言或发表状态的限制，应将这些限制列出来。如果为获取研究而联系了个人或组织，也应予以说明，最好使用“我们联系了制药公司”而不是给出所有联系了的制药公司的列表。如果系统评价进行了手工检索，应加以说明，但为帮助建立CRG专题注册库所进行的手检不应列出。

筛选标准：该部分应写为“对［某疾病、问题或人群类型］［某干预或比较类型］的［研究类型］”。如果系统评价对结局进行了限定，则结局应在此处列出。

数据提取和分析：该部分仅限于说明数据是如何提取和评价的，不应包括何种数据被提取的详细信息。该部分应交代数据提取和偏倚风险评估是否不止由一人完成。如果系统评价作者联系了原始研究者以获取缺失的信息，应予以说明。如果存在，应提供确定不良反应的步骤。

主要结果：该部分首先应交代系统评价纳入的总研究数和总受试者人数，以及与结果解释相关的简要细节（如，研究的整体偏倚风险，或对研究可比性的评价，如果恰当的话）。该部分应基于主要目的，并限制在主要的定量和定性结果（通常包括的重要结果不超过6个）。纳入的结局的筛选应基于以下条件：其最有可能帮助他人对是否使用特定干预的决策。如果系统评价包括了对不良反应的评价，则应予以纳入。如果必要，应注明各结局对应的研究和受试者数，同时应涉及针对这些结局的证据的质量。如果数值结果不清楚或不直观（如从标准化均数差分析获得的结果），则在定量描述的同时应进行定性描述。摘要中的汇总统计量应与系统评价默认选择的统计量相同，且应以标准的方式进行呈现，如：“OR 2.31（95%CI 1.13-3.45）”。理想情况下，事件的危险度（百分比）或平均值（对连续性结局）在两个比较组都应报告。如果系统评价中并未计算总体结果，则可对结果的范围和模式进行定性评估或描述。然而，应避免“投票计数”（报告“阳性”和“阴性”研究数）。

作者的结论：系统评价的主要目的应是呈现信息，而不是提出建议或给出推荐。作者的结论应简洁并直接根据系统评价结果得出，以使其能直接、明显地反映主要的结果。

通常不应对实践场景、价值、偏好、利弊权衡作出假设；且通常不应给出建议或推荐。而应注明数据和分析的所有重要局限性。该部分还应包括对研究有启示意义的重要结论（如果这些意义并不明显的话）。

框11.8.a　假定的摘要举例

（由 Peach A、Apricot D 和 Plum P 完成的系统评价：“A 与 B 比较治疗成人流感”）

背景

A 和 B 均具有抗病毒特性，但由于对其性质的了解不全和对其可能的不良反应的考虑，其并未被广泛使用。本系统评价是对 1999 年首次发表的 Cochrane 系统评价的更新，前一次更新是在 2006 年。

目的

评价 A 和 B 治疗成人流感的效果。

检索方法

我们检索了 Cochrane 急性呼吸道感染组专题注册库（2007 年 2 月 15 日）、Cochrane 对照试验注册中心（Cochrane 图书馆，2007 年第 1 期）、MEDLINE（1966 年 1 月-2007 年 1 月）、EMBASE（1985 年 1 月-2006 年 12 月）和文后所附参考文献。我们也联系了相关的厂商和研究人员。

筛选标准

A 和/或 B 与安慰剂进行比较的随机和半随机对照试验，或是对 A 和/或 B 治疗成人流感不同剂量或疗程的比较。

数据收集

两位作者独立评价试验质量并提取资料。我们联系了原始研究作者以获取补充信息。我们从试验中收集了不良反应信息。

主要结果

共纳入 17 个试验包括 689 例患者。5 个试验包括 234 例患者比较了 A 与安慰剂。与安慰剂比较，A 将发热时间明显缩短了 23%（1 天，95%CI 0.73-1.29）。6 个试验包括 256 例患者比较了 B 和安慰剂。与安慰剂比较，B 将发烧时间明显缩短了 33%（1.27 天，95%CI 0.77-1.77）。直接比较 A 和 B 可得的少量信息（2 个试验包括 53 例患者）并未显示出两种药物的疗效存在差异，尽管可信区间非常宽。基于包括 73 例患者的 4 个试验中，A 的中枢神经系统效应明显比 B 更多见（RR 2.58，95%CI 1.54-4.33）。

作者的结论

A 和 B 对治疗流感均有效。尚无充分证据以确定是否一种比另一种更有效。两种药物都显示出相对较好的耐受性，尽管 B 可能更安全。

11.9 撰写通俗语言摘要

11.9.1 关于通俗语言摘要

通俗语言摘要旨在以一种易为卫生保健用户理解的简单形式对系统评价进行概述。通俗语言摘要可通过因特网免费获取，因此常被作为一种独立文件阅读。通俗语言摘要包括两部分：标题和正文。

通俗语言摘要的初稿通常应由系统评价作者撰写并与系统评价一起投给相关的CRG。该草稿可能需要修改，作者应对一次或多次的退修有心理准备。许多CRG的编辑组都有通俗语言摘要的撰写技巧。如该技巧不可得，中央支持服务部可以协助CRG进行撰写和编辑。虽然该服务由Cochrane用户网运作，但若系统评价作者需要撰写通俗语言摘要方面的帮助却应联系其CRG。

关于完成通俗语言摘要过程的更多信息可从Cochrane手册中获取（该手册可www.cochrane.org/admin/manual.htm获得）。

11.9.2 通俗语言摘要的标题

通俗语言摘要的第一部分为使用简单明了的语言对系统评价标题复述。标题应包括受试者和干预措施（当系统评价标题包含结局时则也应包含结局）。例如，一篇题为“抗胆碱药与其它药物比较治疗成人膀胱过度活动综合征”的系统评价，其通俗语言摘要的标题可以为“药物治疗膀胱过度活动综合征”。若系统评价标题本就容易理解，可以直接将其作为通俗语言摘要的标题，例如“减少长期吸烟危害的干预”。

通俗语言摘要标题不应为声明性的（即不应反映出系统评价结论）。且应该采用命题式的写法(即标题首个单词的首字母和人名大写，其余均小写；具体见前面的例子)，标题的长度不应超过256个字符，且末尾不加句号。

11.9.3 摘要正文

通俗语言摘要的第二部分或称主体，其篇幅不应超过400个单词，且应包括：

- 关于系统评价重要性的陈述：例如卫生保健问题的定义和背景、体征和症状、患病率、对干预措施的描述及其作用原理。
- 系统评价的主要结果：若系统评价是以数值形式报告结果的，则该部分可包含

数值型总结，但应以通俗易懂的形式呈现。通俗语言摘要中的结果应与系统评价中的结果保持一致（即，在摘要中无新结果出现）。如果可能，该部分还应指出结果是基于多少个试验和多少受试者得出的。

- 对不良反应的说明。
- 对系统评价局限性的简要说明（例如，是在特定人群中开展的试验、所纳入试验的方法学质量低）。

在通俗语言摘要末尾，作者应提供部分网页链接（例如，获取CRG网站中其它信息或决策辅助的链接，当然这应以遵照Cochrane协作网关于网络链接的政策为前提。）通俗语言摘要不应包含照片和图片。和Cochrane系统评价的其它组成部分一样，通俗语言摘要也应遵从Cochrane格式指南（该指南可从http：//www.cochrane.org/traning/authorsmes/cochrane-style-guide/cochrane-style-guide获得）。

11.10 本章信息

作者：代表Cochrane应用与推荐方法学组和统计方法学组的Holger J Schünemann, Andrew D Oxman, Julian PT Higgins, Gunn E Vist, Paul Glasziou和Gordon H Guyatt。

本章引用格式：Schünemann HJ, Oxman AD, Higgins JPT, Vist GE, Glasziou P, Guyatt GH. Chapter 11: Presenting results and 'Summary of findings' tables. In: Higgins JPT, Green S (editors), Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

致谢：Penny Hawe教授协助了不良反应部分。Sally Green, Janet Wale和Gill Gyte制作了通俗语言摘要的指南，我们也利用了Rebecca Ryan及用户和交流评价小组叙述合成的指南。撰写摘要的材料基于手册的早期版本。可参考1.4节获得早期手册的作者和编辑的详情，第12章框12.8.a可获得Cochrane应用与推荐方法学组的详情，第9章框9.8.a可获得统计方法学组的详情。

利益冲突：“结果总结”表的很多观点来自GRADE工作组，Holger Schünemann, Andrew Oxman, Gunn Vist, Paul Glasziou和Gordon Guyatt在其中担任不等的领导角色。

11.11 参考文献

Clarke 2006

Clarke M, Hopewell S, Juszczak E, Eisinga A, Kjeldstrøm M. Compression stockings for preventing deep vein thrombosis in airline passengers. Cochrane Database of Systematic Reviews 2006, Issue 2. Art No: CD004002.

Deeks 2001

Deeks JJ, Altman DG. Effect measures for Meta-analysis of trials with binary outcomes. In: Egger M, Davey Smith G, Altman DG (editors). Systematic Reviews in Health Care: Meta-analysis in Context (2nd edition). London (UK): BMJ Publication Group, 2001.

Engels 2000

Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in Meta-analysis: an empirical study of 125 Meta-analyses. Statistics in Medicine 2000; 19: 1707-1728.

GRADE Working Group 2004

GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004; 328: 1490-1494.

Hind 2007

Hind D, Booth A. Do health technology assessments comply with QUOROM diagram guidance? An empirical study. BMC Med Res Methodol 2007; 7: 49.

Lewis 2001

Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. BMJ 2001; 322: 1479-1480.

Liberati 2009

Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and Meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Medicine 2009; 6: e1000100.

（杜亮、肖晓娟、汪泽皓译，陈耀龙、岑啸、贾鹏丽、王霁初审）

第十二章　解释结果和得出结论

作者：代表 Cochrane 应用和推荐方法学组的 Holger J Schünemann, Andrew D Oxman, Gunn E Vist, Julian PT Higgins, Jonathan J Deeks, Paul Glasziou 和 Gordon H Guyatt。

本文选自工作手册5.1.0版本。有关如何引用它的指南，见12.8节。

该手册的早期版本（5.0.2版）还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- Cochrane 协作网所采用的 GRADE 法明确提出了质量有四个等级（高、中、低和极低），其最高质量等级是来自于以随机试验为基础的证据。5 个因素可能会降低随机试验的证据评级，3 个因素可能会升高观察性研究的证据质量。
- 应针对每个结局分别进行证据质量评级。
- 本章介绍了计算、显示和解释二分类结局指标的相对和绝对效应（包括 NNT）的方法。
- 对连续性结局指标，如果系统评价作者使用的是相同单位的研究，可显示提供其合并结果；而对于使用了相同构架但测量单位不同的研究，可提供其标准化均数差和效应量，以及标准化均数差转化后的比值比。

- 系统评价作者不应该将结果描述为“无统计学意义”或“无意义”，而应报告可信区间及其准确的 P 值。
- 系统评价作者不应作出推荐意见，但他们可以——在描述了证据质量和平衡利弊后——强调与特定价值观和意愿相一致的不同选择。

12.1 引言

Cochrane系统评价旨在促进患者和普通公众、临床医生、管理者和政策制定者作出卫生保健决策。清晰地陈述结果、有深度的讨论和清晰地陈述作者的结论是系统评价的重要部分。以下的一些问题尤其有助于人们更好地进行知证决策和增加Cochrane系统评价的使用价值：

- 包括不良结局在内的所有重要结局的信息；
- 每个结局的证据质量，当其应用于特定的人群、和特定的干预措施；
- 阐明特定的价值观和偏好以何种方式影响干预措施的获益、危害、负担和费用的平衡。

第11章（11.5节）介绍了“结果汇总”表，该表以快速可接受的方式提供了重要信息。鼓励作者在Cochrane系统评价中采用这样的表格，以及确保充分描述了纳入研究和Meta分析以支持其内容。文章的讨论部分应提供补充说明。作者应使用5个副标题以确保讨论部分涵盖了适当的内容，且保证将系统评价置于恰当的背景之下。其中包括“主要结果小结（利弊）”、“证据总的完整性和适用性”、“证据质量”、“系统评价过程中潜在偏倚”和“与其他研究或系统评价相比的异同点”。作者的结论分为“对实践的意义”和“对研究的意义”。因为Cochrane系统评价面向世界范围内的读者，讨论和结论部分应尽可能提供较广阔的国际视角，并且就如何在不同环境中应用结果提供指导，而不是将结果局限于特定的国家或局部地区。文化和经济差异可能在确定最佳方案过程中发挥重要作用。而且，社会中的不同个体对于健康以及如何使用社会资源以达到具体的健康状态，都有极为不同的价值观和偏好。甚至在价值观和偏好相同的条件下，人们也可能对同样的研究证据做出不同解释。基于上述这些原因，不同的人对于同样的证据也常常作出不同的决策。

因此，系统评价的目的是呈现信息和帮助解释，而不是提供推荐意见。讨论和结论

应帮助人们理解与实践决策相关的证据的意义，以及将结果应用于特定的环境。系统评价作者应避免对可获得资源和价值观进行假设而给出推荐意见。然而，作者可以通过设置不同的场景来描述特定价值结构以帮助决策。

本章我们描述结果解释主要方面中最重要的一方面，也就是完成“结果汇总表”的基本原则：与每个结局相关的证据质量。然后我们围绕适用性和数值结果的解释提供更详细的思考，并且为作者结论的呈现提供建议。

12.2 评价一组证据的质量

12.2.1 GRADE法

推荐、评估、制定与评价分级工作组（GRADE工作组）已制定出证据质量的评级系统（GRADE工作组2004，Schünemann 2006b，Guyatt 2008a，Guyatt 2008b）。超过20个组织包括世界卫生组织（WHO）、美国内科医师学会、美国胸内科医师学会（ACCP）、美国内分泌学会、美国胸科学会（ATS）、加拿大药品和卫生技术机构（CADTH）、BMJ临床证据、英国国家卫生与临床优化研究所（NICE）和UpToDate®已使用了原版或做了微小修改（Schünemann 2006b，Guyatt 2006a，Guyatt 2006b）。BMJ鼓励临床指南的作者使用GRADE系统（www.bmj.com/advice/sections.shtml）。Cochrane协作网已使用GRADE系统的原则来评价系统评价中所报告结局的证据质量。该评估与“结果汇总”表的介绍放在一起（见第11章，11.5节）。

基于系统评价的目的，GRADE将证据质量定义为对某一效应值或相关强度接近真实值的程度。多种因素可能影响一组证据质量，如12.2.2节所描述的，包括研究内偏倚的风险（方法学质量）、证据的方向、异质性、效应估计值的准确性和发表偏倚的风险。GRADE系统的作用就是对每一结局的一组证据质量进行评价。

GRADE法分为四个质量水平（表12.2.a）。最高质量的评级是针对随机试验的证据。然而，系统评价作者可根据表12.2.b所述的5个因素将随机试验证据降级为中、低、甚至极低质量证据。通常，一个影响因素可使证据质量降低一级，所有因素都存在时可使证据质量最大降三级。如果任何一个因素存在非常严重的问题（如当评价设计和实施的局限性时，所有研究均未采取分配隐藏和盲法，且患者失访超过50%），则该随机试验的证据质量级别可因该影响因素降低两级。

系统评价作者通常把来自可靠的观察性研究的证据评级为最低质量。然而，如果这样的研究得出了较大的效应且无明显的偏倚解释这种效应，那么系统评价作者可将证据质量评级提升为中级或——如果效应足够大——甚至可被评级为高质量（表12.2.c）。极低质量水平包括，但不限于，存在严重问题和非系统性的临床观察研究（如病例系列或病例报告）。

表12.2.a　在GRADE方法中一组证据的质量水平

基本的方法学	质量评级
随机试验；或升高两级的观察性研究	高
降级的随机试验；或升级的观察性研究	中
降两级的随机试验；或观察性研究	低
降三级的随机试验；或降级的观察性研究；或病例系列/病例报告	极低

表12.2.b　可能降低一组证据质量水平的因素

1. 可获得研究的设计和实施存在局限性，表明存在高度偏倚的可能性。
2. 间接证据（间接的人群、干预措施、对照、结局）。
3. 不能解释的异质性或不一致的结果（包括亚组分析问题）。
4. 结果不精确（可信区间较宽）。
5. 存在发表偏倚的高度可能性。

表12.2.c　可能升高一组证据质量水平的因素

1. 效应较大
2. 在结果显示无效时，所有可能的混杂因素都将降低所证实的效应或提示为假效应
3. 存在剂量-效应梯度

12.2.2　降低一组证据质量的因素

现在更详细地描述对于一个特定结局会降低一组证据质量等级的5个因素（表12.2.b）。在每一种情况中，如果发现一个因素降低了证据等级，其应该被区分为“严重”（降低质量评级一级）或“非常严重”（降低质量评级两级）。

1. 设计和实施的局限性：如果研究存在大的可能导致干预措施评价产生偏倚的限制因素，那么我们对于估计效应的置信度就将降低。对于随机试验，其方法学局限性包括缺乏分配隐藏、缺乏盲法（尤其对有偏倚的评估高度敏感的主观结局），高失访率，因为获益而较早终止的随机试验，或选择性报告结局。第8章详细讨论了研究偏倚风险的分级评估，并提出了研究结果偏倚风险的评估方法，分为“低偏倚风险”、“偏倚风险不清楚”和“高偏倚风险”（第8章，8.7节）。这些评估应直接针对该因素。尤其，“低偏倚风险”表示“无局限性”，“偏倚风险不清楚”表明既可能为“无局限性”也可能“存在严重局限性”；“高偏倚风险”表明存在“严重局限性”或“极严重的局限性”。作者必须根据潜在偏倚的可能大小，在可选类别间进行判别决策。每个研究在得到一个具体结局指标的结果时都会存在不同程度的偏倚风险。系统评价作者必须就是否一个结局的证据质量因研究局限性而降级做出整体判断。对研究局限性的评估应针对“结果汇总”表中对结果有贡献的研究，而不是针对被纳入分析的所有研究。我们在第8章曾讨论过，主要分析应限于存在低度（或低度和不清楚）偏倚风险的研究。

表12.2.d显示从偏倚风险评估到每个结局的研究局限性都必须做出判断。仅当绝大多数证据来自低偏倚风险研究时，才能得到高质量评级的证据。例如，在22个关于β阻滞剂对心衰患者病死率影响的研究中，大多数可能或肯定使用了分配隐藏，所有研究都至少对某些重要小组实施了盲法，且随机患者的随访几乎完成（Brophy 2001）。当绝大多数证据来自于在对某一标准上有严重局限性的研究，或在多个标准上有一些局限性的研究时，证据质量可能降低一级。例如，我们不能确信，在恶性疟疾患者中，阿莫地喹和磺胺多辛 乙胺嘧啶联用与单用磺胺多辛 乙胺嘧啶相比，可降低治疗失败率，因为失访者中事件发生率会影响磺胺多辛 乙胺嘧啶的明显获益（3个研究中有两个失访率>20%）（McIntosh 2005）。在腰椎间盘突出症患者中比较手术与保守治疗的证据是一个存在极严重局限性的例子，肯定可以降低证据质量两级（Gibson 2007）。我们不确定在一年或更长时间之后手术在改善症状方面的获益，因为纳入分析的一个试验分配隐藏不充分，且外科医生在未施盲的情况下使用了较粗略的评级方法评价结局。

2. 证据的间接性。证据间接性有两种类型。一种类型是，评价者欲比较可相互替代的两种干预措施（如A和B）的效果，虽然检索到了随机对照试验的证据，但要么是将A与安慰剂比较，要么是将B与安慰剂比较。因此，评价者只能间接比较A与B的疗效。另一种类型是，评价者检索到了符合标准的随机试验，但是这些试验在人群、干预措施、对照或结局等方面限制了主要问题的评价。例如，假设系统评价是关于冠心病的二级预

防的干预措施，发现纳入的的大部分研究中的冠心病患者恰好也患有糖尿病。那么，因为其人群局限在同时患有糖尿病的冠心病患者，而非是所有冠心病患者，因此这样的证据只能被视为间接证据。相反的情况也可能出现：针对糖尿病患者中冠心病预防策略效果的系统评价，发现纳入研究中的受试者有非糖尿病患者，这是提供的也是间接证据。如果针对目标人群（如，糖尿病人群），研究者实施的随机试验很少，这尤其可能。其他间接性来源可能来自所研究的干预措施（如，在所有纳入研究中，一种技术含量很高的干预措施由专业中心的专家和接受过很好培训的专科医师来实施，那么来自这些中心之外干预措施效果的证据可能是间接性的）、所使用的对照（如，对照组接受了在大多数情况下效果比标准治疗差的干预）和所评价的结局（如，当有关患者重要结局的数据不能获得时，或当研究者寻找有关生活质量的数据而仅有症状被报告时，选用替代结局而产生的间接性）。当系统评价作者认为基于主要小组预期效应的差异而使证据需要降级时，其判断必须透明化。

3. 不能解释的异质性或不一致的结果：当研究得出差异很大的效应估计值时（结果的异质性或变异），研究者应对异质性给出合理的解释。例如，药物在患病更重的人群或当给予更大剂量时，可能有更大的相对效应。对异质性及其研究的详细讨论见第9章（第9.5和9.6节）。如果存在重要的因素，有较强的证据表明在不同亚组（在理想情况下需事先确定）间重要结局不同，那么对于不同的人群可考虑分别制作“结果汇总”表。例如，对有症状、高度狭窄的患者，由高水平外科医生行颈动脉内膜剥离术，可使之获益（Cina 2000），而对于无症状的中度狭窄患者（如果他们认为值得做手术），手术不能使之获益（Chambers 2005）。当存在异质性并影响结果解释，但作者又未能找到合理的解释时，证据质量会降低。

4. 结果的不精确性：当研究纳入的受试者较少且事件发生率较低，并因而得到的可信区间较宽时，其证据质量评级会降低。结果汇总表中提供的可信区间，允许系统评价的使用者对证据质量级别作出自己的判断。

5. 发表偏倚的高度可能性：如果研究者由于结果的原因未将研究（特别是那些显示无效的研究：发表偏倚）或结局（特别是那些观察到可能有害或无效的结局：选择性报告偏倚）报告出来，那么证据质量可能会降低。作为偏倚风险评估的一部分（见第8章，第8.13节），选择性报告结局是在研究水平进行评价，因此，对“结果汇总”表中的结局有贡献的研究，其通过上述因素1予以解决（设计和实施的局限性）。如果系统评价纳入的大多数研究对所评价的结局没有贡献，或存在发表偏倚的证据，那么证据质量可能降

级。第10章对发表偏倚进行了详细讨论，包括发表偏倚，以及其在Cochrane系统评价中如何处理。非常典型的情况是，当发表的证据纳入大量小样本试验，且多数是由企业资助时，则可能怀疑存在发表偏倚（Bhandari 2004）。例如，14个类黄酮用于痔疮患者的试验表现出明显的巨大获益，但总共仅纳入1432例患者（也就是说，每个试验纳入了相对较少的患者）（Alonso-Coello 2006）。在绝大多数这些试验中，赞助者的深度介入会引发这样的疑问：是否未发表的试验显示的是未见获益。

一个特定的研究证据可能面临上述五个因素中不止一个因素，且遇到的问题越多，证据质量评级越低。可能虽然获得了随机试验，但所有或几乎所有这些局限性都存在，而且很严重。其证据质量评级应为极低。

表12.2.d　在GRADE评估中对因素1的更多指导： 从偏倚风险的评估到对主要结局的研究局限性的判断

偏倚风险	所涉及的研究	解释	考虑	研究局限性的GRADE 评估
低偏倚风险	多数信息来自低偏倚风险的研究	可能的偏倚不可能严重改变结果	无明显的局限性	无严重局限性，不降级
偏倚风险不清楚	多数信息来自低偏倚风险或偏倚风险不清楚的研究	可能的偏倚可能引发对结果的部分怀疑	潜在的局限性不可能降低对效应估计的可信度	无严重局限性，不降级
			潜在的局限性可能降低对效应估计的可信度	存在严重局限性，降低一级
高偏倚风险	来自高偏倚风险研究的信息所占比例足以影响对结果的解释	可能的偏倚会严重削弱证据的可信度	对一个标准存在严重局限性，或对多个标准存在一些局限性，足以降低对效应估计的可信度	存在严重局限性，降低一级
			对一个或多个标准存在严重局限性，足以在实质上降低对效应估计的可信度	存在非常严重局限性，降低两级

12.2.3 增加证据质量水平的因素

虽然观察性研究和降级的随机试验通常会得出证据质量低的评级，但在少数情况下，证据也可能会被“调高”到中甚至高质量（Table 12.2.c）。

1. 很少见的情况下，当方法学实施很好的观察性研究得出一种干预效应较大、一致和精确的估计值时，其结果可能尤为可靠。在没有可疑混杂因素存在时的大效应（如RR > 2或RR < 0.5）、或在可靠性上未受严重威胁的研究的非常大效应（如RR > 5或RR < 0.2），可能满足这一条件。在这些情况下，虽然观察性研究可能高估了真实效应，但较差的研究设计是不可能解释所观察到的明显获益的。因此，尽管对观察性研究的研究设计有所保留，但作者还是有信心认为效应是存在的。这些研究的效应大小可能将证据质量从低调至中等（如果在没有其他方法学问题时效应较大）。例如，观察性研究的Meta分析显示，自行车头盔在很大程度上能减少骑行者头部损伤的风险（OR=0.31，95%CI 0.26-0.37）（Thompson 2000）。这一较大效应，在没有与此相关的明显偏倚存在时，建议证据质量评级为中等。

2. 有时，来自观察性或随机研究的潜在偏倚可能会导致低估明显的干预效应。例如，如果仅有比较严重的患者接受了试验干预或暴露，但他们仍进展得更顺利，则真实的干预或暴露效应比数据所显示的效应更大。例如，一个严格实施的观察性研究的系统评价，共纳入3800万患者，显示在私立营利性医院比私立非营利医院有更高的死亡率（Devereaux 2004）。一个可能的偏倚与这两种类型的医院中疾病的严重程度不同有关。因为在非营利性医院的患者很可能比在营利性医院的患者更严重。因此，残余的混杂在某种程度上的存在可能会对非营利性医院的结果产生偏倚。第二个可能的偏倚是，收治的优良私人保险覆盖的患者越多，越可能为医院带来更多资源，并且这种溢出效应可以使未参保者获益。因为营利性医院比非营利性医院可能接受更多有很好保险的患者，偏倚再一次不利于非营利性医院。因为可能的偏倚将减小所显示的干预效应，一种可能的考虑是，将来自这些观察性研究的证据作为中等而非低质量。同时存在的类似情况是，观察性研究未能证明相关性，但所有可能的偏倚将增加干预效应。这种情况通常出现在探索明显的损害效应时。例如，因为降糖药苯乙双胍会引起乳酸中毒，相关药物二甲双胍也被怀疑具有同样的毒性。然而，很大样本的观察性研究未能证明该相关性（Salpeter 2007）。如果临床医生在使用该药时更警惕于乳酸中毒，并对其发生有过度报告，可能会认为这种中等甚或高质量证据对二甲双胍的典型治疗剂量与乳酸中毒间的因果关系

给予了反驳。

3. 量效关系的存在也可能增加我们对观察性研究结果的可信度，并因而提高证据质量级别。例如，国际标化比（international normalized ratio，INR）越高，出血风险越大，这种存在的量效关系会增加观察性研究结果（其显示有抗凝水平的患者出血风险增加）的可信度（Levine 2004）。

12.3 适用性问题

12.3.1 系统评价作者的角色

“当将研究结果应用到广大人群或某一特定个体时，都需要医学观念的跳跃性转变”。“在观念转变的过程中，我们总是需要在合理地扩大适用性和个体结论的保守性之间取得平衡”（Friedman 1985）。

为了充分说明系统评价与提出的研究目的相关程度，有些事是系统评价作者必须做的，而有些事是用户必须做的。在此我们讨论系统评价作者可以为用户做什么。Cochrane系统评价作者必须非常清楚他们所计划针对的人群、干预措施和结局。第11章（11.5.2节）强调了一个并未常规作为Cochrane系统评价组成部分的重要步骤：在作比较时与干预策略相关的所有患者重要结局的说明。

就受试者与干预因素而言，系统评价作者需对可能影响效果的因素作出先验假设，然后验证这些假设。如果评价者发现了明显的亚组效应，必须对该疗效的可信度作出评判（Oxman 2002）。需要谨慎解释亚组间的差异，尤其是对于研究间的差异。亚组间会不可避免地存在一些机会性误差，因此除非有很强的证据证明交互作用的存在，否则作者不应假设这些亚组效应存在。如果，尽管谨慎，系统评价作者判断亚组效应确实存在，他们也应对相关亚组分别作Meta分析，并为这些亚组制作单独的“结果汇总表”。

系统评价的使用者将受到结果“个体化”的挑战。例如，即使相对效应在亚组间相似，绝对效应也将因基线风险而有不同。系统评价作者可通过在“结果汇总”表中显示可识别的有不同风险的人群亚组来提供这一信息，如在第11章（11.5.5节）所讨论的。用户可在将患者归于某一特定风险亚组之前对其进行识别，并相应地评估其可能的利弊大小。

用户必须做出的另一个决策是他们所面对的患者是否与系统评价纳入的患者足够不同，以致他们完全不能使用系统评价和Meta分析的结果。系统评价作者在解释结果时

最好是给出充分的理由来说明为什么该证据不适合某类具体患者，而不是刻板强调研究的纳入与排除标准（Guyatt 1994）。作者有时能通过找出重要的变异（这种变异可能限制结果的适用性）而帮助临床决策者（Schünemann 2006a），包括：生物学差异、文化差异，对干预措施依从性的差异等。

在解决这些问题的过程中，作者不可能知道或解决整个世界环境中难以计数的差异。然而，他们能解决许多已知的重要差异，更重要地是，他们应避免在讨论结果和作出结论时假设其他人所处的环境与他们自身所处的环境一致。

12.3.2 生物学差异

作者应考虑的生物学差异问题包括病理生理学方面的差异（如女性和男性间的生物学差异可能影响对同一治疗措施的反应）和病原体方面的差异（如感染性疾病疟疾的病原体多种多样）。

12.3.3 文化背景差异

一些干预措施，尤其是非药物干预措施，可能在某些文化背景下起有效，但在另一些文化背景下无效；这种状况称为文化背景差异作用（Hawe 2004）。文化背景因素可能涉及干预措施的主要提供机构，如只有这样的主要机构才具有实施该干预措施的专业知识、经验和具备资质的工作人员，具有该竞争领域的领先优势，具有与该项目配套的服务和设施，并给该项目极大的重视和地位。文化背景因素还涉及目标人群的基本特征（如文化和语言差异、社会经济地位的差异、城乡差异等），这些基本特征就形成了不同文化背景下医患之间特殊的医患服务模式，而这种模式可能与所要实施的干预措施的价值和技术不匹配。多年来，当决策者认为其他国家的证据并不能应用于自己的国家时，文化背景差异在其中的主要影响已被普遍认可（但未清楚地说明）。

同时，某些项目/干预措施已被从一个文化背景转换到其他文化背景，并已观察到获益，其他则未见获益（Resnicow 1993，Lumley 2004）。系统评价作者在从一个文化背景向另一个文化背景进行推广时，应当谨慎。如果文化背景相关信息可获得的话，系统评价作者应在干预性研究中报告（Hawe 2004）。

12.3.4 依从性差异

患者和干预措施提供者的依从性差异会限制结果的适用性。可预见的依从性差异可能是由于经济状况或态度的不同使得某些干预措施在某些环境不可得或不可行，如在发展中国家（Dans 2007）。不能认为在严密监测的随机试验中的高度依从性在常规实践中也能有相似的依从性。

12.3.5 价值和偏好的差异

管理决策包含权衡所计划管理策略的利弊。对于有不同价值观和偏好的人而言，其正确选择可能不同，临床医生需要确保所作出的决策与患者的价值观和偏好一致。我们在12.7节介绍系统评价作者如何促进这一过程。

12.4 解释统计分析的结果

12.4.1 可信区间

单个研究和Meta分析的结果均以点估计值及其可信区间的形式报告。例如，“OR为0.75，其相应的95%可信区间为0.70-0.80”。点估计值（0.75）是对试验干预效应相比于对照干预效应大小和方向的最佳推测。可信区间描述了这一估计固有的不确定性和我们能确保真实效应位于其间的数值范围。如果可信区间相对较窄（如0.70-0.80），效应量则较精确。如果可信区间较宽（如0.60-0.93），虽然仍有足够的精确性作出关于干预措施使用的决策，不确定性则增加。区间非常宽（如0.50-1.10）则表明我们对效应知之甚少，需要进一步的信息。

95%的可信区间常被解释为显示一个范围，真实效应有95%的可能性位于该范围内。该说法是一种宽泛的解释，但作为一个粗略的理解是实用的。可信区间精确的解释是基于这一假设的概念：如果研究被重复多次，将获得考虑到的结果。如果一个研究被重复无数次，每次重复都计算一个95%的可信区间，则95%的这些区间将包含真实效应。

单个研究可信区间的宽度将依赖于样本量的大小。大样本的研究比小样本的研究倾向于得出更精确的效应估计值（因而有更窄的可信区间）。对连续性变量而言，精确性也依赖于结局指标的变异性（单个测量值间的标准差）；对二分类结局而言，精确性依

赖于事件风险；对于时间-事件结局，精确性依赖于所观察到的事件数。所有这些值都被用于计算效应估计值标准误，从而推导出可信区间。

Meta分析可信区间的宽度依赖于单个研究估计值的精确性和合并的研究数。而且，对于随机效应模型，随着异质性的增加精确性将降低，可信区间也将相应变宽（见第9章，9.5.4节）。当更多研究加入Meta分析时，可信区间的宽度通常会变窄。然而，如果增加的研究增加了Meta分析的异质性而使用了随机效应模型，则可信区间的宽度可能将增大。

可信区间和点估计值在固定效应和随机效应模型有不同的解释。固定效应估计值及其可信区间解决问题“最佳（单一）效应估计值是什么？”随机效应估计值假设效应有一个分布，估计值及其可信区间解决问题“最佳平均效应估计值是什么？”

可信区间可以任何可信度进行报告（虽然常取95%，有时为90%或99%），例如报告0.80的OR值，其80%的可信区间为0.73-0.88，90%的可信区间为0.72-0.89，95%的可信区间为0.70-0.92。随着可信度的增加，可信区间变宽。

可信区间和P值间有逻辑对应关系（见12.4.2节）。如果显著性检验得出了一个小于0.05的P值时，对于一个效应，95%的可信区间将不包含无效值（如OR=1或RD=0）。如果P值正好为0.05，那么95%可信区间的上限或下限将为无效值。相似地，当且仅当显著性检验得出了一个小于0.01的P值时，99%的可信区间将排除无效值。

点估计值和可信区间将共同提供信息以评估干预措施的临床有效性。例如，我们正在评估一种降低事件风险的治疗，且我们决定仅当其将一个事件发生风险从30%至少降低5%即降低到25%时，才认定这项措施有效（这些值依赖于特定的临床环境和结局）。如果Meta分析得出的合并结果显示时间风险降低了10%，且95%CI较窄（如7%-13%），则我们可以做出结论：治疗是有效的。因为点估计值和可信区间的都超过了预先设定的5%的临床有效性标准。反之，如果Meta分析报告了相同的降低10%的事件风险，但其可信区间较宽(如2%-18%)，尽管我们仍然能做出结论：对治疗效果的最佳估计是有效，但却没有太大的把握，因为没有排除效果可能是2%-5%的可能性。如果可信区间再宽一点，包括了差值0%，我们将不能排除治疗效果出现任何情况的可能性，在下结论时需更加谨慎。

不同可信度的可信区间可以表明对于不同程度的获益或危害有不同的证据。例如，报告同样的分析结果：（1）干预措施不引起危害时用95%的可信区间；（2）有一些效果时用90%的可信区间；（3）有患者重要的获益时用80%的可信区间。这些情况都可能表

明干预措施的有效性，且需要进一步研究。

系统评价作者可能使用同样的一般性方法做出干预措施无效的结论。如上例，如果设定最小患者重要差异为5%的风险差，2%的效应估计值，可信区间为1%-4%表明干预措施无效。

12.4.2 P值和统计学意义

P值是在无效假设成立的前提下，获得观察效应（或更大效应）的概率，其在Cochrane系统评价中可假设为“干预措施无效”或“研究间干预措施效果无差异”（无异质性）。因此，P值很小表明观察效应由偶然所致的可能性很小，并据此提供证据来推翻无效假设。通常情况是通过检测P值是否比特定的阈值小来解释。特别是P值小于0.05通常被报告为“有统计学意义”，并解释为偶然所致的可能性足够小以拒绝无效假设。然而，0.05的阈值是主观设定的，常用于医学和心理学研究的原因是在这些领域中通过比较检验统计量和统计分布特定百分点下的面积确定P值。RevMan，像其他统计软件包一样，报告精确的P值。如果系统评价作者决定在Meta分析结果中给出P值，应报告精确的P值及其95%的可信区间。

在RevMan中，提供两种P值。一种与Meta分析的汇总效应相关，来自于无效假设的前提下，Z检验得到干预措施无效（或随机效应Meta分析中平均无效）。另一种与研究间的异质性相关，来自于无效假设的前提下，卡方检验得到没有异质性（见第9章，9.5.2节）

对于汇总效应的检验，P值的计算涉及效应估计值和样本量（更确切地说是效应估计的准确性）。随着样本量的增加，由偶然因素导致干预措施有效的概率就降低。相应地，一个特定强度效应的统计学意义在大样本研究中比在小样本研究中更大（P值将更小）。

P值常常以两种方式被错误理解，一种情况是，将中等或大的P值（如大于0.05）错误地理解为“干预措施无效”。正确的解释是“证据尚不足以证明干预措施有效”。这两种解释所蕴含的意义明显不同。为避免这种错误的解释，系统评价作者对结果进行假设检验时应同时报告效果大小、95%可信区间及P值大小。在小样本研究或小样本Meta分析中，常常出现的情况是可信区间包括的效应范围既包含了无干预效果又包含了有干预效果。因此，建议系统评价的作者不要将结果描述为“无统计学意义”或“无意义”。

第二种错误的解释是认为合并结果的P值越小，则干预措施的效果就越明显。这种错误的解释最常见于大样本研究中，如纳入数十个研究和数千名受试者的Meta分析。P

值解决的是干预效果是否确实没有的问题；其不能检测干预措施是否对其潜在受者的效果足够大。在大样本研究中，即使干预措施效果很小，假设检验时也可以得出很小的P值。因此，对统计结果的正确解释要结合点估计值和可信区间（见12.4.1节）。

12.5 二分类结局结果解释（包括NNT）

12.5.1 相对和绝对风险降低

临床医生可能更偏好于开出一种能降低死亡风险25%的干预措施，而不是死亡风险降低1个百分点的干预措施，虽然这两种措施可能显示了相同的获益（如风险从4%降低至3%）。前者是指风险的相对降低，而后者是指风险的绝对降低。如第9章（9.2.2节）所述，有几个指标可用于两组的二分类结局比较。Meta分析通常采用风险比（RR）、比值比（OR）或风险差（RD）完成，但结果的表达有几种方式可供选择。

相对风险降低率（relative risk reduction：RRR）是以百分数降低的形式对RR进行再表达的一种便捷方式：RRR=100%×（1-RR）。在上述例子中，0.75的RR转换为25%的RRR。

风险差常常是指绝对风险降低（absolute risk reduction，ARR），其可以表述为百分数（如，1%）、小数（如，0.01）或计数（如，1000例中有10例）的形式。众所周知，RR的简单转换形式NNT是表述同样信息常用的替代方式。我们在12.5.2节讨论NNT，并在12.5.3节讨论在表述绝对效应时不同的选择。我们接下来描述如何从单个研究和Meta分析结果中计算得到这些数据。

12.5.2 关于NNT的更多内容

NNT（number needed to treat）被定义为：在给定的时间范围内，接受试验干预比接受对照干预为多增加一例发生或避免一次事件，预期需治疗的人数。例如，NNT为10可解释为“在给定的时间范围内，每10个受试者接受试验干预比接受对照干预预期可增加（或减少）一例发生一次事件”。需要明确的是：

1. 既然NNT是从RR得到，那么它仍是一个比较效应指标（试验 vs. 某一特定对照），且不具有单一干预措施的普遍特性；

2. NNT给出了一个“预期值”。例如，NNT=10并不是指在每10人和每10人所构成的小组中会发生一例额外事件。

NNT可计算有利和有害事件以及同时引起有利和有害结局的干预措施。在所有实例中，用正整数表述NNT，去掉所有小数。当一个干预导致结局恶化而非改善时，一些作者使用术语“用某种干预引起1例某种不良事件所需要的人数”（number needed to harm，NNH）。然而，该短语并不令人满意，可能带来误导即容易被解释为如果给予干预，将经历有害结局的人数。应当尽量避免使用术语“用某种干预引起1例某种不良事件所需要的人数”和“NNT”来表示直接效应。更好地是使用“对增加一例有利结局需干预的人数”（NNTB）和“为增加一例有害结局需干预的人数”这样的术语来显示直接效应。

因为NNT适用于事件，所以当二分类结局是基于量表结局的二分类时，其解释需仔细斟酌。例如，如果结局是以“无、轻度、中度或重度”量表来测量疼痛，则其可按照“无或轻度”与“中度或重度”进行二分类。对从这些数据得到的NNT（疼痛的NNT）则是不恰当的。其实质是“中或重度疼痛的NNT”。

12.5.3 绝对风险降低的表达

系统评价用户易受证据所选择的统计学表现形式的影响。Hoffrage等认为，临床医生处理“自然频数”——包括治疗和未治疗总例数（如治疗结果使妇女患乳腺癌的风险从20/1000降低至10/1000），比处理以百分数呈现的效应（如乳腺癌风险绝对降低1%）时，他们对统计结局的推断更恰当（Hoffrage 2000）。概率可能比频数更难于理解，尤其当事件罕见时。在改进研究证据（和参与卫生保健决策）的表述中标准化可能很重要。当前的证据表明，对于表达绝对风险的差异，自然频数的表达形式最易被卫生保健信息的用户所理解。这一证据为我们提供了在“结果总结表”中呈现绝对风险的合理形式为每1000例接受干预的受试者发生事件的例数。

风险比和相对风险降低仍然很重要，因为相对效应在不同的风险组间实质上比绝对获益更趋于稳定。系统评价作者可通过自己的数据来研究这种一致性（Cates 1999，Smeeth 1999）。风险差在基线事件率上很少一致；因此在系统评价中计算NNT一般不恰当。如果相对效应指标（OR或RR）被选择用于Meta分析，则在ARR或NNT计算时，需要设定对照组风险。对每个临床可辨别的风险组，阐明其绝对获益以及观察的时限很重要。包含不同疾病严重程度患者的研究，或不同随访时间的研究将几乎肯定有不同的对照组风险。在这些情况下，不同的对照组风险会得出不同的ARR和NNT（除非干预措施无效）。推荐的方法是在所假设的对照风险（ACRs）范围内将OR或RR用一系列的NNT

来表示（McQuay 1997，Smeeth 1999，Sackett 2000）。系统评价作者在制作“结果总结表”时和在系统评价的正文中均应加以考虑。

例如，口服抗凝剂预防卒中的系统评价通过描述不同基线风险下的绝对获益来将信息呈现给用户（Aguilar 2005）。他们呈现其主要结果为“在心房颤动患者中使用口服抗凝药的决策中应考虑卒中固有的风险，选择对这一治疗有最大获益的患者”（Aguilar 2005）。在每1000例有先兆卒中或一过性脑缺血发作（其每年卒中发生率约12%（120/1000））的高风险心房颤动患者中，华法林约可预防70例卒中。这一表述有助于用户理解典型的基线风险在其可预期的绝对获益上的重要影响。

12.5.4 计算

绝对风险降低（ARR）或需要治疗患者数（NNT）的直接计算依赖于从单个研究或Meta分析中可获得的汇总统计量（OR、RR或RD）。当表述Meta分析结果时，作者在计算中应使用他们认为最恰当的合并统计量（见第9章，9.4.4.4节）。此处，我们描述每1000受试者中降低的数量来获得ARR的结果。例如，-0.133的风险差相当于每1000例中，133例会免于事件。

ARRs和NNTs不应通过各试验间的受试者和事件累积总数进行计算。这样做会忽视研究内的随机化，如果任何研究的随机化不均衡，都可能对结果产生严重误导。

当计算NNTs时，所获得的值常规都上取整数。

12.5.4.1 通过风险差（RD）计算 NNT

NNTs可按如下方式对单个研究进行计算。注意：该方法虽然可用，但很少被用于风险差Meta分析的结果，因为Meta分析通常都是用相对效应指标进行（RR或OR）。

NNT可从风险差计算得到：

NNT=1/风险差的绝对值=1/|RD|，

此处分母的垂线（绝对值）表明对负号予以忽略，其通常将NNT向上舍入最近的整数。例如，如果RD是-0.12，NNT是9；如果NNT是-0.22，NNT是5。

12.5.4.2 通过风险比（RR）计算绝对风险降低或 NNT

为便于解释，系统评价作者愿意从风险比的Meta分析结果计算绝对风险降低或NNT。

为达此目的，需要假设对照风险（ACR）。对于一个范围内的不同ACR，这通常是恰当的。计算过程如下：

每1000例减少的数量=1000×ACR×（1-RR）

NNT=|1/ACR×（1-RR）|

例如，假设风险比是RR=0.92，并假设对照风险ACR=0.3（300/1000）。则在风险上的效应是每1000例减少24例：

每1000例减少的例数=1000×0.3×（1-0.92）=24

NNT是42：

NNT=|1/0.3×（1-0.92）|=|1/0.3×0.08|=41.67

12.5.4.3 通过 OR 计算绝对风险减低或 NNT

系统评价作者可能希望从OR的Meta分析结果计算绝对风险减低或NNT。为达成该目的，需有假设的对照风险（ACR）。对于一个范围内的不同ACR，这通常是恰当的。计算过程如下：

$$\text{number fewer per 1000} = 1000 \times \left(\text{ACR} - \frac{\text{OR} \times \text{ACR}}{1 - \text{ACR} + \text{OR} \times \text{ACR}} \right)$$

$$\text{NNT} = \frac{1}{\left| \text{ACR} - \frac{\text{OR} \times \text{ACR}}{1 - \text{ACR} + \text{OR} \times \text{ACR}} \right|}$$

例如，假设比值比为OR=0.73，并假设对照风险ACR=0.3。则在风险上的效应是每1000例减少62例：

$$\text{number fewer per 1000} = 1000 \times \left(0.3 - \frac{0.73 \times 0.3}{1 - 0.3 + 0.73 \times 0.3} \right)$$

$$= 1000 \times \left(0.3 - \frac{0.219}{1 - 0.3 + 0.219} \right) = 1000 \times (0.3 - 0.238) = 61.7$$

NNT为17：

$$\text{NNT} = \frac{1}{\left| \left(0.3 - \frac{0.73 \times 0.3}{1 - 0.3 + 0.73 \times 0.3} \right) \right|} = \frac{1}{\left| 0.3 - \frac{0.219}{1 - 0.3 + 0.219} \right|} = \frac{1}{|0.3 - 0.238|} = 16.2 .$$

12.5.4.4 通过比值比（OR）计算风险比

尽管风险比（RR）较比值比（OR）容易解释，但OR数学特征更好。系统评价作者可能决定基于OR实施Meta分析，但却以汇总风险比（或相对风险降低）的形式解释结果。这需有假设的对照风险（ACR）。则：

$$RR = \frac{OR}{1 - ACR \times (1 - OR)}$$

使用从Meta分析纳入研究得到的中位对照风险来进行这一转化通常是合理的。

12.5.4.5 计算可信限

通过应用上述对于汇总统计量（RD、RR或OR）的上下可信限的公式计算ARRs和NNTs可信限（Altman 1998）。注意：该可信区间并不包含对照组风险（CGR）的不确定性。

至于通常认为无统计学显著性的结果（如，OR或RR的95%可信区间包含1），可信限的一端表明获益，另一端则表明危害。因此，当以事件的形式呈现结果时，对每个界限都需要恰当使用单词“更少”和“更多”。对NNTs，两个可信限应标记为NNTB和NNTH以显示每种情况下的效应方向。NNT的可信区间存在“不连续性”：区间包含从无限大的NNTB到无穷大的NNTH。

12.6 通过连续性结局解释结果（包括标准化均数差）

12.6.1 连续性结局的Meta分析

当结局为连续性时，系统评价作者在表述合并结果时有多种选择。如果所有研究都使用了同样的单位，Meta分析可以平均反应差的形式，以相同的单位得出合并估计值（例如，见第11章，图11.5.a对于水肿的行汇总结果）。这样的结果单位可能难于解释，尤其当结果与评级量表相关时。“结果汇总”表应包括测量量表的最小和最大值及其方向（见第11章，图11.5.a，水肿列）。对患者感知的最小评分改变的认识很重要——最小重要差异——且能在很大程度上促进结果解释。知道最小重要差异可让作者和用户将结果置于研究背景中，且如果作者知道最小重要差异，应在结果汇总表的评论栏加以说明。

当研究使用了不同的工具来测量相同的内容时，Meta分析中可用标准化均数差（SMD）对连续性资料进行合并（见第9章，9.2.3.2节）。对临床解释而言，这种分析可能不如二分类反应和报告获益患者比例等方法好。已有方法根据报告的均数和标准差将数据转化为二分类数据，但这些方法需要的假设可能不满足（Suissa 1999，Walter 2001）。

SMD以标化的单位而非原始的测量单位来表达干预效应。SMD通过患者结局的合并标准差以区分试验和对照组平均效应的差异（见第9章，9.2.3.2节）。因此，SMD值取

决于效应量（均数差）和结局的标准差（患者间的固有变异）。

在没有指南的情况下，临床医生和患者可能对于如何解释以SMD呈现的结果束手无策。如下几种方式可对这种结果以更有用的方式进行表述。

12.6.2 使用效应量经验法则对SMD进行表述

已有用于解释SMD（或效应量）的经验法则，主要由社会科学领域的研究者提出。如，0.2代表小的效应，0.5代表中度效应，0.8代表大的效应（Cohen 1988）。也有不同的划分（如，<0.41=小，0.41-0.70=中度，>0.71=大）。系统评价作者可以考虑在结果汇总表的评论栏写入经验法则。然而，一些方法学家认为，这样解释存在问题，因为一个结果的患者重要性取决于环境，而并非具有普遍意义。

12.6.3 通过转化为OR对SMD进行表述

基于对两干预组内连续变量具有相等标准差逻辑分布的假设，可将SMD转化为OR或log OR（Furukawa 1999, Chinn 2000）。该假设不可能准确地予以把握，结果只能取近似值。log OR被估计为

$$\ln OR = \frac{\pi}{\sqrt{3}} SMD,$$

（或近似1.81×SMD）。所获得的OR则可与预设的对照组风险结合，以获得12.5.4.3节所讲的绝对风险降低。这些对照组风险是指在连续性结局中有一定（不明确的）程度改善的人群所占的比例（应答者）。表12.6.a显示了一些来自该方法的阐述结果。这些NNT可使用公式1000/NNT而转化为每千人中的人数。

表12.6.a NNT等于对于对照组中不同给定的“改善比例”的特定SMD。

对照组改善比例	10%	20%	30%	40%	50%	60%	70%	80%	90%
SMD=0.1	57	33	26	23	23	24	28	37	66
SMD=0.2	27	16	13	12	12	13	15	20	36
SMD=0.5	9	6	5	5	5	6	7	10	18
SMD=0.8	5	4	3	3	4	4	5	7	14
SMD=1.0	4	3	3	3	3	4	5	7	13

12.6.4 使用常用工具对SMD进行表述

对解释SMD最后的可能性是以一个或更多特定测量工具的单位对其进行表达。用对特定尺度的典型的不同个体间的标准差乘以SMD会得到在该尺度上结局评分均值之差（试验 vs. 对照）的估计值。标准差可通过研究之一基线评分的合并标准差获得。在实践中，为更好地反映不同个体间的变异，最好使用来自有代表的观察性研究的标准差。合并效因此应以特定工具的原始单位表述，且可解释干预效应的临床相关性和影响。然而，作者应该注意到如果被应用于个体研究而非汇总的效应指标，这种效应量的逆转换可能带来误导。（Scholten 1999）。考虑两个研究使用同样的工具，观察同样的效应，但观察了不同的参与者间的变异（或许由于不同的纳入标准），那么使用来自这些研究的不同标准差进行逆转换对同样的尺度和的效应将得到不同的效应大小。

12.7 得出结论

12.7.1 Cochrane系统评价的结论部分

Cochrane系统评价的结论可分为对实践的意义和对研究的意义。要决定这些意义是什么，应考虑四个因素：证据质量、利弊平衡、价值观和偏好以及资源利用（Eddy 1990）。考虑这些涉及到判断和效应的因素超出了绝大多数系统评价作者的努力。

12.7.2 对实践的意义

做出关于干预措施实践有用性的结论，必须要综合考虑利益、风险和费用，权衡取舍。之后，可以对今后的临床实践提出推荐意见。但是，提出推荐意见已经超出了系统评价的范围，推荐意见的提出还需要综合考虑许多系统评价之外的其他因素，而这正是临床实践指南制定者的工作。因此，系统评价作者不应提出推荐意见。

如果作者认为不得不对某项干预措施提出意见时，他们应该首先清楚地列出证据质量和风险效益的平衡，然后突出强调不同的价值观和医疗模式下对该干预措施所应采取的不同态度。同时，应强调其他可以影响临床决策的因素，包括任何已知的可以影响干预措施效果的因素、患者的基础状态和危险因素、费用和这些费用的支付者以及资源的可得性等。作者应确保考虑了患者所有的重要结局，包括可得的数据有限的结局。该过

程表明对与不同结局对应的价值观或意愿的判断要很明确。最明确地是采用敏感性分析（对价值观和意愿进行不同的假设）来进行正式的经济学分析。这超越了绝大多数Cochrane系统评价的范围（虽然他们很可能被用于这种的分析）（Mugford 1989，Mugford 1991）；对该问题的讨论见第15章。

一篇有关在癌症患者中使用抗凝剂以增加存活率的系统评价（Akl 2007）提供了一个范例（对干预措施满意和不满意的效应存在重要权衡取舍的情况下给出临床意义）：“对癌症患者开始肝素治疗以提高生存率的决策应权衡利弊和结合患者的价值观和意愿（Haynes 2002）。对延长生存时间有强烈意愿（即使延长的时间可能有限），且不害怕出血、不将肝素治疗作为一种负担的患者，可能选择使用肝素；而那些害怕出血和将肝素治疗视作负担的患者可能不会选择使用肝素治疗。”

12.7.3 对研究的意义

系统评价结论应有助于人们对未来的卫生保健研究做出知情决策。结论中“对研究的意义”应对未来的研究需要，尤其是解决相关临床问题最需要的研究作出评论。建议采取下列格式（EPICOT）来报告对未来研究方向的推荐意见（Brown 2006）：

E（Evidence，证据）：当前的证据是什么？

P（Population，人群）：诊断、疾病分期、合并症、危险因素、性别、年龄、种族、特定的纳入和排除标准、临床环境。

I（Intervention，干预措施）：类型、频次、剂量、疗程、预后因素；

C（Comparison，比较）：安慰剂、常规治疗、替代治疗/管理。

O（Outcome，结局）：研究者需要去测量、改进、影响或达到哪些临床或患者相关的结局？应使用哪些测量方法？

T（Time stamp，时间标记）：文献检索或推荐的时间。

在推荐时可能考虑的其他因素包括要解决情况的疾病负担、时限（如随访长度、干预疗程），以及最适合于后续研究的研究类型（Brown 2006）。

Cochrane系统评价作者应确保其包括了这一格式的PICO方面。这有助于注明解决研究问题最好的研究类型以及任何特定的设计特征。

在航空旅客中使用弹力袜预防深静脉血栓的系统评价举了一个对干预措施获益有确切证据的例子：“系统评价显示，在这些试验所研究的患者类型中，穿与不穿弹力袜

对无症状深静脉血栓效果的问题已得到回答。进一步的研究可调查不同强度的长筒袜或长筒袜与其他预防措施比较的相对效果。进一步的随机试验应针对穿或不穿弹力袜在效果仍不确定的结局（如死亡、肺栓塞或有症状深静脉血栓）。”（Clarke 2006）

焦虑障碍抚触治疗的系统评价提供了一个未发现合适研究时对研究意义的例子：“该系统评价强调，需要评价抚触治疗对于降低诊断为焦虑障碍患者焦虑症状有效性的随机对照试验。未来的试验需要在设计和实施上更加严格，之后的报告应对方法学的各方面有高质量描述，以便于对结果进行评价和解释。”（Robinson 2007）

12.7.4 得出结论中的常见错误

一个常见的错误是，在证据不足以得出确定结论时，将“没有某一疗效的证据”与“证据证明无效”混淆；并声称干预措施“无效”或与对照干预措施相比“无差异”。更可靠的做法是报告结局指标增加或减少量的数据及可信区间。当干预措施的疗效显示“有效”但没有统计学意义时，作者通常会解释成这是一种有希望的干预措施，而当干预措施的疗效显示“有害”且没有统计学意义时，作者会解释成这是对该干预措施的警告信号；这些描述性语言都没有正确解释结果，应避免在结论中出现。

另一种错误是用期望性词语来下结论。例如，当纳入研究显示在病死率上存在降低甚或是增加但未达到通常的统计学意义时，作者会写“纳入研究太少而不足以发现病死率的降低”。避免出现这种情况之一的方式是考虑结果盲法，也就是不报道研究结果有益或有害的倾向，而是分别从有益和有害两个方向作出假设。例如，如果对于某一干预措施的疗效估计的可信区间包含无效值，则该干预措施既可能是确实有益的措施，也可能是确实有害的措施。如果在结论中提及了其中一种可能性，则其他可能性也应提及。

另一种常见错误是超越证据得出结论。在没有参考得出系统评价对实践意义结论的附加信息或判断时常常不易察觉。即使该额外信息支持系统评价对实践意义的结论，系统评价作者也很少对此信息进行系统评价。而且，对实践的意义常常由特定的环境因素和价值决定，必须对此加以考虑。正如我们已经注意到的，当做出有关实践意义的结论时，作者应始终保持谨慎，且不应给出推荐意见。

12.8 本章信息

作者：代表Cochrane应用和推荐方法组的Holger J Schünemann, Andrew D Oxman, Gunn E Vist, Julian PT Higgins, Jonathan J Deeks, Paul Glasziou和Gordon H Guyatt。

本章引用格式：Schünemann HJ, Oxman AD, Vist GE, Higgins JPT, Deeks JJ, Glasziou P, Guyatt GH. Chapter 12: Interpreting results and drawing conclusions. In: Higgins JPT, Green S (editors), Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

致谢：Jonathan Sterne, Michael Borenstein和Rob JM Scholten执笔撰文。

利益声明：Holger Schünemann, Andrew Oxman, Gunn Vist, Paul Glasziou和Gordon Guyatt不同程度地在GRADE工作组担任领导角色，本章的许多思想来自该工作组。

框12.8.a Cochrane应用和推荐方法组

我们期望本章所描述的方法将继续改进。应用和推荐方法学（ARMG）和 GRADE 工作组是举行相关讨论的场所。两个讨论工作组都欢迎热心于更多学习和参与证据质量评级制定，以及探索 Cochrane 系统评价应用问题的新成员。

应用和推荐方法学组由对系统评价结果向个人和小组解释、应用以及转化感兴趣和有专长的人组成。ARMG 的目的是探索从证据向卫生保健推荐转化的过程。其最终目标是使这一过程尽可能严谨。

当前考虑重要的领域包括：

- 评价证据质量（www.gradeworkinggroup.org）；
- 效应基线风险的变化；
- 患者预期事件率或严重性获益预测；
- 思考证据强度大小以及临床和科研意义所产生效应的大小和精确性；
- 思考当基于个体的临床特征对利弊进行加权时，人们的价值观对临床和科研意义有何影响。

12.9 参考文献

Aguilar 2005

Aguilar MI, Hart R. Oral anticoagulants for preventing stroke in patients with non-valvular atrial fibrillation and no previous history of stroke or transient ischemic attacks. Cochrane Database of Systematic Reviews 2005, Issue 3. Art No: CD001927.

Akl 2007

Akl EA, Kamath G, Kim SY, Yosuico V, Barba M, Terrenato I, Sperati F, Schünemann HJ. Oral anticoagulation for prolonging survival in patients with cancer. Cochrane Database of Systematic Reviews 2007, Issue 2. Art No: CD006466.

Alonso-Coello 2006

Alonso-Coello P, Zhou Q, Martinez-Zapata MJ, Mills E, Heels-Ansdell D, Johanson JF, Guyatt G. Meta-analysis of flavonoids for the treatment of haemorrhoids. British Journal of Surgery 2006; 93: 909-920.

Altman 1998

Altman DG. Confidence intervals for the number needed to treat. BMJ 1998; 317: 1309-1312.

Bhandari 2004

Bhandari M, Busse JW, Jackowski D, Montori VM, Schünemann H, Sprague S, Mears D, Schemitsch EH, Heels-Ansdell D, Devereaux PJ. Association between industry funding and statistically significant pro-industry findings in medical and surgical randomized trials. Canadian Medical Association Journal 2004; 170: 477-480.

Brophy 2001

Brophy JM, Joseph L, Rouleau JL. Beta-blockers in congestive heart failure. A Bayesian Meta-analysis. Annals of Internal Medicine 2001; 134: 550-560.

Brown 2006

Brown P, Brunnhuber K, Chalkidou K, Chalmers I, Clarke M, Fenton M, Forbes C, Glanville J, Hicks NJ, Moody J, Twaddle S, Timimi H, Young P. How to formulate research recommendations. BMJ 2006; 333: 804-806.

Cates 1999

Cates C. Confidence intervals for the number needed to treat: Pooling numbers needed to treat may not be reliable. BMJ 1999; 318: 1764-1765.

Chambers 2005

Chambers BR, Donnan GA. Carotid endarterectomy for asymptomatic carotid stenosis. Cochrane Database of Systematic Reviews 2005, Issue 4. Art No: CD001923.

Chinn 2000

Chinn S. A simple method for converting an odds ratio to effect size for use in Meta-analysis. Statistics in Medicine 2000; 19: 3127-3131.

Cina 2000

Cina CS, Clase CM, Haynes RB. Carotid endarterectomy for symptomatic carotid stenosis. Cochrane Database of Systematic Reviews 2000, Issue 2. Art No: CD001081.

Clarke 2006

Clarke M, Hopewell S, Juszczak E, Eisinga A, Kjeldstrøm M. Compression stockings for preventing deep vein thrombosis in airline passengers. Cochrane Database of Systematic Reviews 2006, Issue 2. Art No: CD004002.

Cohen 1988

Cohen J. Statistical Power Analysis in the Behavioral Sciences (2nd edition). Hillsdale (NJ): Lawrence Erlbaum Associates, Inc., 1988.

Dans 2007

Dans AM, Dans L, Oxman AD, Robinson V, Acuin J, Tugwell P, Dennis R, Kang D. Assessing equity in clinical practice guidelines. Journal of Clinical Epidemiology 2007; 60: 540-546.

Devereaux 2004

Devereaux PJ, Choi PT, El-Dika S, Bhandari M, Montori VM, Schünemann HJ, Garg AX, Busse JW, Heels-Ansdell D, Ghali WA, Manns BJ, Guyatt GH. An observational study found that authors of randomized controlled trials frequently use concealment of randomization and blinding, despite the failure to report these methods. Journal of Clinical Epidemiology 2004; 57: 1232-1236.

Eddy 1990

Eddy DM. Clinical decision making: from theory to practice. Anatomy of a decision. JAMA 1990; 263: 441-443.

Friedman 1985

Friedman LM, Furberg CD, DeMets DL. Fundamentals of Clinical Trials (2nd edition). Littleton (MA): John Wright PSG, Inc., 1985.

Furukawa 1999

Furukawa TA. From effect size into number needed to treat. The Lancet 1999; 353: 1680.

Gibson 2007

Gibson JN, Waddell G. Surgical interventions for lumbar disc prolapse. Cochrane Database of Systematic Reviews 2007, Issue 2. Art No: CD001350.

GRADE Working Group 2004

GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004; 328: 1490-1494.

Guyatt 1994

Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. JAMA 1994; 271: 59-63.

Guyatt 2006a

Guyatt G, Gutterman D, Baumann MH, Addrizzo-Harris D, Hylek EM, Phillips B, Raskob G, Lewis SZ, Schünemann H. Grading strength of recommendations and quality of evidence in clinical guidelines: report from an American College of Chest Physicians Task Force. Chest 2006; 129: 174-181.

Guyatt 2006b

Guyatt G, Vist G, Falck-Ytter Y, Kunz R, Magrini N, Schünemann H. An emerging consensus on grading recommendations? ACP Journal Club 2006; 144: A8-A9.

Guyatt 2008a

Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ. What is 'quality of evidence' and why is it important to clinicians? BMJ 2008; 336: 995-998.

Guyatt 2008b

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008; 336: 924-926.

Hawe 2004

Hawe P, Shiell A, Riley T, Gold L. Methods for exploring implementation variation and local context within a cluster randomised community intervention trial. Journal of Epidemiology and Community Health 2004; 58: 788-793.

Haynes 2002

Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. ACP Journal Club 2002; 136: A11-A14.

Hoffrage 2000

Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G. Medicine. Communicating statistical information. Science 2000; 290: 2261-2262.

Levine 2004

Levine MN, Raskob G, Beyth RJ, Kearon C, Schulman S. Hemorrhagic complications of anticoagulant treatment: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. Chest 2004; 126: 287S-310S.

Lumley 2004

Lumley J, Oliver SS, Chamberlain C, Oakley L. Interventions for promoting smoking cessation during pregnancy. Cochrane Database of Systematic Reviews 2004, Issue 4. Art No: CD001055.

McIntosh 2005

McIntosh HM, Jones KL. Chloroquine or amodiaquine combined with sulfadoxine-pyrimethamine for treating uncomplicated malaria. Cochrane Database of Systematic Reviews 2005, Issue 4. Art No: CD000386.

McQuay 1997

McQuay HJ, Moore A. Using numerical results from systematic reviews in clinical practice. Annals of Internal Medicine 1997; 126: 712-720.

Mugford 1989

Mugford M, Kingston J, Chalmers I. Reducing the incidence of infection after caesarean section: implications of prophylaxis with antibiotics for hospital resources. BMJ 1989; 299: 1003-1006.

Mugford 1991

Mugford M, Piercy J, Chalmers I. Cost implications of different approaches to the prevention of respiratory distress syndrome. Archives of Disease in Childhood 1991; 66: 757-764.

Oxman 2002

Oxman A, Guyatt G. When to believe a subgroup analysis. In: Guyatt G, Rennie D (editors). Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice. Chicago (IL): AMA Press, 2002.

Resnicow 1993

Resnicow K, Cross D, Wynder E. The Know Your Body program: a review of evaluation studies. Bulletin of the New York Academy of Medicine 1993; 70: 188-207.

Robinson 2007

Robinson J, Biley FC, Dolk H. Therapeutic touch for anxiety disorders. Cochrane Database of Systematic Reviews 2007, Issue 3. Art No: CD006240.

Sackett 2000

Sackett DL, Richardson WS, Rosenberg W, Haynes BR. Evidence-Based Medicine: How to Practice and Teach EBM (2nd edition). Edinburgh (UK): Churchill Livingstone, 2000.

Salpeter 2007

Salpeter S, Greyber E, Pasternak G, Salpeter E. Risk of fatal and nonfatal lactic acidosis with metformin use in type 2 diabetes mellitus. Cochrane Database of Systematic Reviews 2007, Issue 4. Art No: CD002967.

Scholten 1999

Scholten RJPM. From effect size into number needed to treat [letter]. The Lancet 1999; 453: 598.

Schünemann 2006a

Schünemann HJ, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development: 13. Applicability, transferability and adaptation. Health Research Policy and Systems 2006; 4: 25.

Schünemann 2006b

Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, Fahy BF, Gould MK, Horan KL, Krishnan JA, Manthous CA, Maurer JR, McNicholas WT, Oxman AD, Rubenfeld G, Turino GM, Guyatt G. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. American Journal of Respiratory and Critical Care Medicine 2006; 174: 605-614.

Smeeth 1999

Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from Meta-analyses - sometimes informative, usually misleading. BMJ 1999; 318: 1548-1551.

Suissa 1991

Suissa S. Binary methods for continuous outcomes: a parametric alternative. Journal of Clinical Epidemiology 1991; 44: 241-248.

Thompson 2000

Thompson DC, Rivara FP, Thompson R. Helmets for preventing head and facial injuries in bicyclists. Cochrane Database of Systematic Reviews 2000, Issue 2. Art No: CD001855.

Walter 2001

Walter SD. Number needed to treat (NNT): estimation of a measure of clinical benefit. Statistics in Medicine 2001; 20: 3947-3962.

（杜亮译，吴红梅、岑啸、贾鹏丽、王霁初审）

第十三章　纳入非随机研究

作者：代表 Cochrane 非随机研究方法学组的 Barnaby C Reeves, Jonathan J Deeks, Julian PT Higgins and George A Wells。

本文选自工作手册 5.0.2版本。有关如何引用它的指南，见13.8节。这些材料还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（ISBN 978-0470057964），由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 有些Cochrane系统评价所研究的问题并不能通过随机试验解决，此时作者可以纳入非随机研究。
- 因为非随机研究相比随机试验的潜在偏倚可能更大，所以将其结果纳入系统评价和Meta分析时对结果的解读应谨慎。尤其应注意不同干预组间受试者的差异（选择性偏倚）和那些未明确报道其有计划书的研究（报告偏倚）。
- 我们建议，对纳入研究的纳入标准、数据收集、严格评价应强调研究设计的具体特

点（例如，研究的哪些部分是前瞻性的设计），而不仅仅是简单地提供研究设计的名称（例如，病例对照研究vs.队列研究）。

- 尽管相比随机试验必须更加关注选择性偏倚的可能性，仍可以采用与随机试验相似的方式评估非随机研究偏倚风险。
- 非随机研究的Meta分析必须考虑如何解决潜在的混杂因素，以及因残余混杂因素和其他因研究而异的偏倚而导致异质性增大的可能。

13.1 引言

13.1.1 本章主要内容

本章内容由Cochrane协作网非随机研究方法学小组（Non-Randomised Studies Methods Group，NRSMG）（见框13.8.a）完成。目的在于支持那些考虑在Cochrane系统评价中纳入非随机研究的系统评价作者。本手册将非随机研究（Non-Randomised Studies，NRS）定义为：对干预效果（危害或收益）评估，研究单位未采用随机方法进行对比组间分配的的定量研究。其包含那些依据日常治疗决策和患者选择而进行分配的研究，即通常所谓的观察性研究。非随机干预研究包括很多类型：队列研究、病例对照研究、前后对照研究、断点时间序列（Interrupted-time-series）序列研究和采用不恰当随机策略的对照研究（有时被称为半随机研究）。框13.1.a总结了非随机研究的常用研究设计分类。13.5.1节将解释我们为何并不一定建议在Cochrane系统评价中使用这些分类。

本章旨在描述那些在Cochrane系统评价纳入NRS时会出现的特殊问题，其内容基于理论或流行病学知识、实证研究和NRSMG成员的讨论结果。本章就如何处理某个事项做出推荐，其前提是能够基于已有证据和理论来支持该项推荐。若尚不能作出确定推荐，本章则力求陈述供选择方案的利弊并确定有待进一步进行方法学研究的问题。

考虑在Cochrane系统评价中纳入NRS的系统评价作者们，开始学习本章内容应在已熟悉一篇随机试验系统评价的准备过程后。无论Cochrane系统评价是仅纳入随机试验还是纳入了NRS，其格式和基本步骤是相同的。对这些步骤的详细阐述，读者可参见该手册的第一部分。若在系统评价中纳入了NRS，则完成该评价的每一步都会更加困难，并且作者应力求在其系统评价团队中包含流行病学家和方法学家。有一个这样的例子（Siegfried 2003），该系统评价有9个作者，其中5个都为方法学家。

以下对设计类型的区分均基于常用的分类标识和描述，而这些描述并非特定的，因为对标识的解读方式可视情况而异。NRSMG并不提倡使用这些分类，原因请见13.5.1节。

框13.1.a　一些用于评价干预效果的非随机研究设计类型

非随机试验	采用非随机方法将人群分配到不同干预组的实验性研究。
前后对照研究	对采用了试验干预措施的干预组和未采用干预的对照组进行干预前后的观察的研究。
断点时间序列研究	对干预前后多个时间点（即所谓“断点”）的观察值进行研究。该设计旨在探究某干预是否随时间推移具有比基础趋势更显著的效果。
历史性对照研究	将现时某干预的结果与既往的未给予该干预的相似病人的结果比较的研究。
队列研究	追踪一组既定人群（即队列）的结局以判定不同干预措施与其结局间的联系的研究。其中，前瞻性队列研究招募的是在研究开始前还未采用任何干预的受试者并随后对他们进行随访观察；回顾性队列研究纳入的对象则来自病历记录表明已采用了干预的个体并从这些病历记录的时间点开始对受试者进行随访。
病例对照研究	一种通过比较某群体内具有某结局的人群（病例组）与同一群体中不具该结局有的人群（对照组），以探究该结局和暴露因素（例如：进行某项干预）的关联程度的研究类型。这是一种尤其适用于罕见结局的研究设计。
横断面研究	一种通过收集人群在特定时间点的干预（过去或现在）信息和当前卫生结局，即限于健康状况，以评估干预的暴露史与结局的关联程度的研究类型。
病例系列（Case series ）（非对照性纵向研究）	在该研究中，通常是对接受同一干预的一系列个体进行干预前后观察，但是没有对照组。

13.1.2　为什么考虑纳入非随机研究？

相比于其他研究设计，随机试验更可能提供关于不同卫生保健方案的不同效果的无偏倚信息，因此Cochrane协作网重点关注的是随机试验的系统评价。只有当某个研究问题不能用随机试验的系统评价解决时，我们才可能进行非随机研究的系统评价。NRSMG认为系统评价作者纳入对偏倚的易感度适中的NRS是合理的。总体来说，NRSMG认为在Cochrane系统评价中纳入非随机研究有以下三个主要原因：

a) 通过提供对现有非随机研究缺点的详细评估以检验进行随机试验的案例。非随机研究的系统评价结果也可能对后续随机试验的设计很有用，如通过区别相关亚组。
b) 为不能够被随机化或极不可能用随机试验来研究的干预效果(有利或有害）提供证据。在这些情况下，一个系统报告了可得非随机研究结果及其局限性的公正的（没有偏倚的）系统评价将会很有用。
c) 为不能够用随机试验研究透彻的干预效果（有利或有害）提供证据，比如长远结局和罕见结局，或在现有主要随机试验实施时并不了解其重要性的结局。

另三个常被用来支持非随机研究系统评价但不充分的理由：

d) 研究未被纳入随机试验的人群（例如，儿童，孕妇和老年人）的干预效果。尽管考虑试验结果能否被推广应用于非试验人群的群体很重要，但目前尚不清楚非随机研究能否达此目的。暂不考虑NRS的估算结果与随机试验的结果是否一致，NRS的结果通常会存在潜在的偏倚进而导致作出不正确的结论。
e) 补充现有随机试验证据。在随机试验证据中引入非随机研究证据，可能使不精确但无偏倚的估算结果成为精确但有偏倚的，即将不确定性转变为误差。
f) 当干预的效应确实很大时可以纳入NRS。由于需要实施相关系统评价（或者其他的证据合成）来观察可能的效应大小，所以这是一个以结果为导向的或事后分析的判断理由。同时，在大家较易认为大效应比小效应更难通过偏倚来完全解释（Glasziou 2007）的情况下，对于卫生保健实践，获取效应作用强度的无偏倚估计以助临床和经济决策仍然是很重要的（Reeves 2006）。因此，对于效应大（并且它们并不需要太大，当效应当真很大时）的研究仍需要进行随机试验。对于NRS的系统评价结果已显示可能有很大效益的干预，其随机试验就可能会遭到伦理方面的质疑，从而使受试者难以随机化；此外，对于被普遍认为有很大效应的干预，由于其他原因也很难做到随机（例如，手术vs. 非手术）。然而，以上情况下选择纳入NRS的系统评价的理由应属于上述的原因（b）而非原因（f），即不大可能随机分配的干预。

13.1.3 将非随机研究纳入 Cochrane 系统评价存在的关键问题

随机试验是研究卫生保健干预效果的首选研究设计类型，因为在多数情况下，随机试验是最少出现偏倚的。任何Cochrane系统评价都必须考虑个体原始研究的偏倚风险，包括偏倚可能的方向和大小（见第八章）。同样地，纳入NRS的系统评价也要求其作者遵循这些，且关于偏倚风险评价的原则完全一致。但是，相对于随机试验，非随机对照研究的潜在偏倚可能更大。系统评价作者需要考虑到：（a）所用设计类型的不足（如：注意它们确定因果关系的能力），（b）通过对偏倚风险的细致评估了解该项研究的具体实施情况，尤其是（c）出现所有NRS都可能存在的选择性偏倚和混杂因素的可能，以及（d）报告偏倚的可能，包括选择性报告结果。

选择性偏倚的易感度（本手册中是指：不同干预组中个体特征的基线差异，而非样本是否能代表总体）被广泛地视为随机试验和NRS的首要差异。有充分的分配序列隐藏的随机，可减少随机试验的系统选择性偏倚，因此随机试验中组间特征的差异可视为偶然。而在NRS中，组间的分配取决于其他因素，且这些因素常常是未知的。当选择性偏倚导致干预组和对照组（或病例对照研究中的病例组和对照组）的不平衡增加时，就会出现预后因素的混杂因素，即与结局相关的因素存在组间分布差异。这种混杂对Meta分析有两种影响：（a）改变干预效果的估算结果（系统性偏倚）；（b）增大观察结果的变异度，造成各项研究间过大的异质性（Deeks 2003）。考虑以上两种可能的影响（见第13.6.1节）极其重要。本章的13.5节提供了关于NRS的偏倚易感度更详细的讨论。

13.1.4 计划书对于纳入了 NRS 的 Cochrane 系统评价的重要性

第2章明确了在实施Cochrane系统评价之前撰写计划书的重要性。因为在实施NRS的系统评价过程中，方法学的选择很复杂，并可能影响到系统评价结果，所以对于纳入了NRS的系统评价，其计划书更为重要。此外，开展纳入NRS的系统评价理由（见第13.1.2节）应在计划书中进行说明。相比随机试验的系统评价，NRS的系统评价计划书应该包括更为详细的信息、提前制定的关于使用的方法及分析计划的方法学策略。该计划书需要把与随机试验无关的细节列入说明（如：计划用于确定潜在混杂因素和评估原始研究对混杂因素易感度的方法），也需要对系统评价实施过程中那些在纳入了NRS后更复杂的标准步骤进行说明（如：说明筛选合格研究的纳入标准和检索策略）。

NRSMG指出，在方案中提前对所有关于系统评价要应用的方法进行决策也许不太可能。但不管怎样，系统评价作者应力求在不提及原始研究结果的情况下决定该系统评价将要采用的所有方法，并报告收集完研究结果的数据后必须重新制定或修订的方法学策略。

13.1.5 本章的章节内容安排

本章的各节应用相同结构依次阐述了系统评价过程的不同步骤。首先特别指出，我们概括了在Cochrane系统评价中纳入NRS（与随机试验比较）的不同之处及其适用情况，并描述了需要考虑的概念性问题。第一部分会提供一些现有的相关证据。其次，我们总结了一些指导意见，并在可能的情况下描述了现有的可获取资源以帮助系统评价作者。

13.2 制定纳入非随机研究的标准

13.2.1 纳入非随机研究有何不同？

13.2.1.1 同时纳入随机与非随机研究

由于随机试验的数量很少或因其目前已知的随机试验局限性，系统评价作者可能希望在系统评价中纳入NRS。本章中，我们强烈建议系统评价作者不要试图去合并随机试验和非随机研究的证据。该建议具体是指：若想评价某一干预对特定结局指标的干预效果时，纳入研究设计的标准一般都应明确是随机试验还是非随机研究。（然而，单一系统评价也可能由包含针对不同结局指标而采取不同研究设计的子系统评价组成，例如，评价收益的随机试验及评价危害性的非随机研究；详见第14章。）另外，在急需随机试验证据却无法获得的情况下，就可在纳入标准中合理说明：仅在无法获取随机试验时纳入NRS。随着该类系统评价的持续更新，当可以获取随机试验时，则应及时将NRS替换掉。当某一干预的随机试验与NRS同时存在,且鉴于13.1.2节中提到的一个或多个原因，两者都被纳入了系统评价时，则应分别呈现随机试验与NRS。此外，当有足够多的随机试验时，相关非随机研究的评论也可被纳入系统评价的讨论部分，但这鲜有用处。

13.2.1.2 评价收益与危害

Cochrane 系统评价旨在量化卫生服务干预有益与有害的、预期与非预期的效果。多数系统评价是采用随机试验来评估的干预措施的预期收益。随机试验可能会报告某项干预的某些危害，这些危害要么是预期的、该试验试图评估的，要么是非预期的、在试验中作为标准安全监控部分而被收集到的。然而，由于某项干预的许多严重危害很罕见或在随机试验的随访期间未出现，从而导致了这些危害不被报道。因此，非随机研究的系统评价的一个最重要的作用就是评估干预的潜在非预期危害或罕见危害（见13.1.1节中的原因（c））。设定用于评价罕见、长远副作用或不良效果的重要相关研究的筛选标准较难。尽管对于有益结局，不同研究设计的相对优缺点相同，但研究设计的选择还是要取决于该结局出现的频繁程度及其重要性。例如，对于一些罕见的不利结局，则只能获得病例系列或病例对照研究。当缺少更好的证据评估某严重事件时，易受偏倚影响的研究设计也是可以接受的。

与研究干预的预期效应相比，在研究其罕见不良作用或非预期效应时，混杂对系统评价真实性的威胁可能更小，因为临床医生首要关注的是结局，而“适应症混杂”主要影响治疗决策。但是此情况下，混杂仍无法排除，因为那些相同特征的预期效应混杂因素也可能是非预期效应的直接混杂因素，或者与混杂因素的特征相关。

与此相关的一个问题是需要区别量化干预效果与检测干预效果。量化干预的预期效益（最大化估计的精确度并最小化偏倚的易感度）对权衡不同干预适用于同种情况的优缺点是很关键的。同时一篇系统评价还应该量化干预的危害，并尽可能使对偏倚的易感度最小化。但是，如果一篇系统评价无可置疑的确定某干预会造成特别的伤害，那么估计效应量的精确度和偏倚的易感度可能就不那么重要了。也就是说，某干预所致伤害的严重性可能超过其任何效益。以上情形通常更可能出现于同一情况多个干预并存的时候。

13.2.1.3 决定纳入哪种类型的非随机研究

随机试验是一种前瞻性的实验性研究设计，其特点在于将受试者随机分配到干预组。尽管随机试验设计也有很多种（包括个体、整群或身体部位的随机分配；多臂试验，析因试验，交叉试验），但是它们仍属于同一个特殊的研究类别。相比之下，非随机研究则涵盖一些完全不同的设计，其中有些设计最初源于病因学的流行病学领域。其中部分设计类型已总结在框13.1.a中，虽然其并非详细清单，并且有许多研究会综合来自不同基础设计的理念。如13.2.2中所讨论的，表中的分类并不总是适用。NRS设计的多样性

引出了两个相关问题。第一，关于某特定有效性问题的所有非随机研究设计都应被纳入其系统评价中吗？第二，如果系统评价作者不纳入所有的非随机研究设计，那么应该采用什么标准来决定纳入或排除哪些研究设计？

大家普遍认为应当设立标准以限制系统评价纳入证据的种类。最主要的原因是不同研究间的偏倚风险不同。正因为如此，许多Cochrane系统评价仅仅纳入随机试验（当可以获得时）。鉴于同样的原因，有人认为系统评价作者应只纳入那些出现偏倚的可能性最小的NRS。当原始研究的结果可能存在偏倚时，尽管此时没有其他更好的证据，在系统评价中纳入这些原始研究也是毫无用处的。因为一个误导性的效应估计可能比不评估对患者危害更大，尤其是当使用这些证据进行临床决策的人们未意识到证据的局限性时（Doll 1993，Peto 1995）。

关于用来限制在Cochrane系统评价中纳入NRS的研究设计标准尚未达成一致。一种策略是只纳入那些能提供合理、有效的效应估计的研究设计。另一种策略是纳入那些已被用来回答某特定问题的最佳可得研究设计。第一种策略意味着系统评价间具有一致性并纳入相同类型的NRS，但一些系统评价未纳入任何NRS。第二种方法则表示不同的系统评价根据可利用资源纳入不同的研究。例如，与干预的收益相比，评价干预的危害时，也许完全可以使用不同的纳入标准。这种策略在Cochrane系统评价数据库（Cochrane Database of Systematic Reviews，CDSR）中已很明显，有些Cochrane系统评价小组（Cochrane Review Groups，CRG）的编辑将系统评价纳入的研究限定为随机试验，同时有些CRG编辑允许纳入特定类型非随机研究（在随机试验较少的领域如卫生保健领域中尤为典型）的情况下。

无论采纳以上哪种策略，纳入标准的选取都必须考虑原始研究设计的分级，即根据研究设计特点的偏倚风险进行排序。通过把病因学研究的证据分级运用到干预的有效性的研究上，目前似乎出现了大量关于有效性研究的“证据分级”（Eccles 1996，National Health and Medical Research Council 1999，Oxford Centre for Evidence-based Medicine 2001）。例如，我们通常认为与病例对照研究相比，队列研究能提供更好的证据。鉴于病因分级更着重于确立因果关系（比如，结局前的暴露因素剂量-反应关系）而非对效应值的有效量化，我们并不能确定这种分级总是适用。此外，用于研究干预效果的研究设计可能会更多样化和复杂化（Shadish 2002），或许也很难列入目前的证据等级（例子可见框13.1.a中的研究设计表）。不同的设计会对不同偏倚的易感，并且通常我们不能确定哪种偏倚的影响最大及它们如何随临床情况变化。

13.2.1.4 病因学和有效性研究问题的区别

原则上，在Cochrane 评价中纳入NRS时允许纳入在普通卫生保健或日常生活过程中进行干预的真实性观察研究。对于那些不局限于医疗场所的干预措施，则可能是研究对象自己的选择，比如：非处方药。在系统评价中纳入的观察性研究也允许研究的暴露因素是不明显的“干预措施”，比如：膳食选择，及其他影响健康的行为。这就在有效性和病因学的证据中产生了一个“灰色地带”。因此，仔细区别对于特定暴露的病因学和有效性的研究很重要。例如：营养学家感兴趣的可能是包含每天五份最小量的水果或蔬菜（“五份每天”）的饮食方式的健康效应，即病因学问题。而另一方面，公共卫生专业人员则可能对能促进患者改用包含“五份每天”的饮食方式的干预所带来的健康效应感兴趣，即有效性问题。在未认识或意识到以上两个研究解决的是不同的研究问题时，由于涉及这两类问题的研究的其他差别（比如，随访时间和调查的结果），我们通常会认为研究前一个问题类型（即病因学问题）的研究“更好”或“更相关”。某些情况下，NRS所评价的健康干预措施实施的目的并非改善健康状况。例如，一项关于包皮环切术以防止艾滋病病毒传播的系统评价中就纳入了这样的NRS，这些NRS中行包皮环切术是出于文化或宗教的原因（Siegfried 2003），并且我们也不确定以健康为目的来实施这项干预是否有相同的结果。

13.2.2 用于支持系统评价作者的指导意见和可用资源

系统评价作者应该首先向准备注册系统评价方案所属的CRG编辑们查明其是否存在关于系统评价纳入NRS的CRG-特殊政策。系统评价作者还应与编辑讨论CRG能提供的方法学建议的范围，因为相比只纳入随机试验，纳入NRS的系统评价作者可能需要更多的帮助，并且试着招募博学的方法学家到他们的系统评价小组。遗憾的是，NRSMG目前尚不能就特定的系统评价与系统评价作者进行合作交流，不过我们鼓励在系统评价中纳入了NRS的作者们将其经验反馈给NRSMG，尤其是那些能支持或反驳文章所阐述内容的经验。

若系统评价作者想要评价干预措施的不良反应（危害）则应阅读第14章的内容，其内容由不良反应方法学小组（adverse Effects Methods Group）制定。

我们建议系统评价作者在决定系统评价纳入哪种类型的NRS时，参考明确的研究设计特征（注意：不是研究设计分类）。出于以上目的，NRSMG的成员已制作出两个说

明研究设计特征的表格，但是其使用该表的经验还很有限。表13.2.a和表13.2.b 分别就个体分配研究和整群分配研究的研究设计特点进行了描述。表中的16个（或15个）条目被分到如下四组中：

1. 是否有对照？
2. 各研究组是如何设立的？
3. 研究的哪些部分为前瞻性的？
4. 不同干预组间哪些变量具有可比性？

这些条目用于描述研究的主要特点，是基于NRSMG 成员的经验及“基本原则”（而不是证据）设计的，它们也许能够确定主要研究设计类别或可能与偏倚易感度相关。该表格通过比框13.1.a中的分类更加详细的分类说明了不同非随机研究设计各自的特征，。对这些（栏目）分类的使用并没有完全达成一致。这些不一致并不代表横排条目不合适或者描述不充分；这两个表的价值取决于归类原始研究时系统评价作者们的共识。此外，我们建议将这两个表用作数据收集的清单或者研究严格评价的一部分(见第13.4.2节和第13.5.2节)。将这些条目作为清单的使用说明，框13.4.a提供了条目的深入解释。

当没有或者随机试验非常少时，许多机构会选择做NRS的系统评价。系统评价通常服务于给卫生保健专业人员发布政策或指引的机构，例如，英国国家卫生和临床技术优化研究所（National Institute for Health and Clinical Excellence，NICE)及加拿大药品和卫生技术署（Canadian Agency for Drugs and Technologies in Health，CADTH)，其实施则由大学相关卫生科学部门的系统评价团队完成。总体来看，这些团队中的系统评价员们已经在探寻将针对随机试验系统评价方法运用于NRS的系统评价，这些团队包括：

- Cochrane有效实践与医疗保健组（Effective Practice and Organisation of Care，EPOC）（www.epoc.cochrane.org）
- 系统评价与传播中心（The Centre for Reviews and Dissemination，CRD）（www.york.ac.uk/inst/crd）
- 伦敦大学教育学院EPPI中心（eppi.ioe.ac.uk）
- 有效公共卫生实践项目（The Effective Public Health Practice Project，EPHPP），加拿大卫生部，汉密尔顿市与长期保健，公共卫生服务（链接至EPHPP系统评价：old.hamilton.ca/phcs/ephpp）

考虑到研究设计或方法学的质量，CRG及Cochrane系统评价作者曾考虑限制NRS的纳入，他们认为NRS设计影响了偏倚易感度。例如，EPOC的CRG就只接受纳入了断点

时间序列研究和前后对照研究的系统评价方案，而拒绝纳入了其他类型非随机研究设计的方案。其他系统评价则限制为只纳入“方法学质量合格”的研究（Taggart 2001）。

13.2.3 总结

- 系统评价作者应在其系统评价中仔细阐述纳入NRS的依据。
- 系统评价作者应就关于纳入NRS的相关编辑政策向提起注册所在的CRG进行咨询。系统评价作者应考虑到CRG所能提供的方法学建议范围及该团队能够提供的方法学支持。
- 系统评价作者应根据研究者的的具体实施过程（即研究设计的重要方面）详细阐述纳入标准及与所研究问题相关的因素（即干预，人群，卫生问题）以避免不确定性。在该过程中，我们建议作者使用NRSMG清单及其类似清单中的条目。
- 系统评价作者也需要了解原始研究中研究者们的具体实施情况以归类纳入的研究。为此，我们建议作者使用NRSMG研究设计特征表或相似的工具，并记录研究设计中那些不清楚或者没有报道的重要方面。
- 若系统评价作者所研究的问题是关于干预措施的副作用（危害）的，则应阅读第14章的内容。

表13.2.a　研究设计特征表（以个体分配干预措施的研究）

	RCT	Q-RCT	NRCT	CBA	PCS	RCS	HCT	NCC	CC	XS	BA	CR/CS
是否有对照：												
接受不同干预的两个或多个受试组间的比较？	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
同一受试组内不同时间段的比较？	P	P	N	Y	N	N	N	N	N	N	Y	N
受试者分配到各组的依据：												
隐蔽随机分组？	Y	N	N	N	N	N	N	N	N	N	na	na
半随机？	N	Y	N	N	N	N	N	N	N	N	na	na
研究人员的其他方式？	N	N	Y	P	N	N	N	N	N	N	na	na
时间的不同？	N	N	N	N	N	N	Y	N	N	N	na	na
地点的不同？	N	N	P	P	P	P	P	na	na	na	na	na
治疗决策？	N	N	N	P	P	P	N	N	N	P	na	na
参与者的偏好？	N	N	N	P	P	P	N	N	N	P	na	na
基于结局？	N	N	N	N	N	N	N	Y	Y	P	na	na
其他过程？（请详细说明）												
研究的哪些部分为前瞻性的：												
受试者的确定？	Y	Y	Y	P	Y	N	P*	Y	N	N	P	P
基线的评估及干预的分配？	Y	Y	Y	P	Y	N	P*	Y	N	N	na	na
结果的评价？	Y	Y	Y	P	Y	P	P	Y	N	N	P	P
假设的提出？	Y	Y	Y	Y	Y	Y	Y	Y	P	P	P	na
各组间哪些变量具有可比性：												
可能的混杂因素？	P	P	P	P	P	P	P	P	P	P	N	na
结局变量的基线评估？	P	P	P	Y	P	P	P	N	N	N	N	na

Y=是；P=可能；P*=只在单组时可能；N=否；na=不适用。注意：此表中填“可能”表示在该单元格中“是”或“否”都有可能。在应用该清单时不能将其作为回答选项，若不确定应回答“无法判断”(见框 13.4.a)。RCT=随机对照试验；Q-RCT=半随机对照试验；NRCT=非随机对照试验；CBA=前后对照研究；PCS=前瞻性队列研究；RCS=回顾性队列研究；HCT=历史性对照研究；NCC=巢式病例对照研究；CC=病例对照研究；XS=横断面研究； BA=前后对照研究；CR/CS=病例报告/病例系列。

表13.2.b　研究设计特征表（以群体分配干预措施的研究）

	ClRCT	ClQ-RCT	ClNRT	CITS	CChBA	ITS	ChBA	EcoXS
是否有对照：								
接受不同干预的两个或多个群体的组间比较？	Y	Y	Y	Y	Y	N	N	Y
同一群组不同时间段的比较？	P	P	N	Y	N	Y	Y	N
群体分配到各组的依据：								
隐蔽随机分组？	Y	N	N	N	N	N	N	N
半随机？	N	Y	N	N	N	N	N	N
研究人员的其他方式？	N	N	Y	P	P	N	N	N
时间的不同？	N	N	N	Y	Y	Y	Y	N
地点的不同？	N	N	P	P	P	N	N	P
政策/公共卫生决策？	Na	na	P	P	P	P	na	na
群体偏好？	Na	na	P	P	P	P	na	na
其他过程？（请详细说明）								
研究的哪些部分为前瞻性的：								
受试群体的确定？	Y	Y	Y	P	P	P	P	N
基线的评估及干预的分配？	Y	Y	Y	P	P	P	P	N
结果的评价？	Y	Y	Y	P	P	P	P	N
假设的提出？	Y	Y	Y	Y	Y	Y	Y	P
各组间哪些变量具有可比性：								
可能的混杂因素？	P	P	P	P	P	P	P	P
结局变量的基线评估？	P	P	P	Y	Y	Y	Y	N

注意：“群体”指的是实体（比如一个机构），而不一定指一组观察对象；“组”指一个或多个群体；见框 13.4.a。

注意：此表中填“可能”表示在该单元格中“是”或“否”都有可能。在应用该清单时不能将其作为回答选项，若不确定应回答“无法判断”（见框 13.4.a）。

Y=是；P=可能；P*=只在单组时可能；N=否；NR=不作要求。ClRCT=整群随机对照试验；ClQ-RCT=整群半随机对照试验；ClNRT=整群非随机对照试验；CITS=中断时间序列对照研究（Shadish 2002）；CChBA=前后对照队列研究（Shadish 2002）；ITS=中断时间序列研究；ChBA=前后队列研究（Shadish 2002）；EcoXS=生态学横断面研究。

13.3 非随机试验检索

13.3.1 纳入非随机试验有什么不同

13.3.1.1 全面、系统的检索策略

如果一个系统评价仅仅只纳入随机试验，检索符合条件的研究的一个重要原则是评价者应该尽可能找到关于待系统评价的临床问题的已经发表的和已经开始实施所有随机试验。因此，建议系统评价撰写者应该检索注册的试验，会议摘要，灰色文献等，以及重要的书目数据库，如 MEDLINE、PUBMED、EMBASE（见第6章）。有一种观点认为系统评价为了避免发表偏倚，需要全面检索。同样有人认为纳入非随机试验的系统评价撰写者也应该全面检索（Petticrew 2001）。然而，重要的是建立一些能够支撑全面检索的基本原理的前提并仔细思考这些前提是否适用于非随机试验的系统评价。这些前提有：

a) 界定所评价问题的随机试验涉及的特定研究人群；
b) 通过全面、系统的检索，找到所有关于该人群的随机试验，由于随机试验相对容易识别，可以获得注册号，而且在没有基金支持和伦理审批下的试验实施是很困难的，这又提供了一条线索（Chan 2004）；
c) 涉及该人群的所有随机试验，如果实施得当，可提供有价值的资料信息；
d) 获取随机试验的信息的满意程度与他们的发现有关，所以最易检出的研究往往是具有偏倚的子集，这就是发表偏倚（即具有统计学意义和有利结果的研究更容易发表）（见第10章10.2）。因为小规模研究产生此类结果的可能较小，并且不能检出所有研究可能会导致漏斗图不对称。理论上讲，可通过检出所有随机研究得出一个没有偏倚的答案，即是通过全面检索，找到小规模研究，无统计学意义或不利结局的研究。小规模研究同样可能受其它偏倚不同程度的影响，使得由于其它原因导致漏斗图不对称发生的机会增加。但是这些偏倚风险能够被人们合理理解和评估（见第10章10.4）。

目前，尚不清楚这些理论是否适用于非随机试验。

13.2.1.3节指出非随机试验包括多种设计类型，且将它们分类存在困难。即使系统评价撰写者针对评估可能纳入的潜在非随机试验，确定具体的研究设计类型纳入标准，但很多潜在合格的非随机研究因没有报告足够的信息而无法确定是否纳入。

非随机研究在什么情况下能够被认可，这是需要确切定义的深层次问题。如一个队列研究，收集了干预措施和结局数据，但是还没有评估他们的联系，该研究是否是合格的非随机试验？用OR值分析相关关系，能否判断非随机试验是否合格？因此，很难为一个特定评价问题确定“非随机研究的限定群体”。一些已经完成的非随机试验根本无法追踪，即使是众所周知的问题也无法追踪。

尽管存在定义合格的非随机研究的组成要素的难题，但这也给NRS的实际检出提出了重大的挑战。这不仅仅是因为缺乏报告，而且还在于：

- 非随机研究不需要注册；
- 重要的研究设计特征缺乏索引，等等；
- 非随机研究往往不需要伦理批准（至少过去是）；
- 非随机研究不一定得到赞助或基金；并且
- 非随机研究不一定按照预定方案执行。

目前还没有证据可以说明报告偏倚对随机试验和非随机试验的影响有什么不同。但是，很难让人相信报告偏倚对非随机试验的影响小于随机试验，即使通过增加与随机试验实施和报告相关特征的报道来防止报告偏倚发生，而非随机试验通常没有报告这些指标（预先制定的计划书，包括过程和最终报告的伦理批准，CONSORT声明（Moher 2001），试验注册编号和在书目数据库中出版物类型索引等）。与随机试验不同，非随机试验发表偏倚的可能强度和的决定因素都尚不清楚。

全面检索对非随机研究的益处仍不清楚，对这个问题仍需进一步研究。如果研究是由于低质量设计和小样本量而难以查找到，那么最难发现的文献（研究）发生偏倚的风险也是最大。随机试验系统评价中，全面检索可以防止发生偏倚，因为符合研究所需的研究人群已存在，所以，没有显著性结果的小样本量的研究也能被最终检出。而在非随机试验系统评价中，即使在理论上，符合研究所需的研究人群被限定，人们也没有同样的信心确定，没有显著性结果的研究也可以被检出。

13.3.1.2 检索非随机研究

根据研究人群和疾病特征、干预措施及比较措施编辑检索式，很容易就可以制定出能检索到所有关于某种干预措施所有证据的检索策略。当系统评价仅仅纳入随机试验时，可以用很多种方法限定检索策略，使其针对随机试验（见第6章）：

a) 检索关于该系统评价问题已发表的系统评价。

b) 使用随机试验很丰富的资源，如CENTRAL或特定CRG的注册信息。

c) 使用过滤器和索引字段来限定检索可能是随机试验的研究，如MEDLINE中的出版类型。

d) 检索试验注册者

限定检索目标为特定的非随机研究设计会更加困难。在上述方法中，只有（a）和（b）可能有帮助。系统评价撰写者应当检索特定CRG中的注册信息，以获取潜在的非随机研究。在有些CRG（如EPOC组）的注册中包含一些特定类型非随机试验（若有需要，请咨询该小组）。在CENTRAL中纳入随机研究的过程意味着部分非随机试验也被纳入但是并非全部，因此，即使是对特定设计类型的研究，只检索这个数据库是不全面的。目前，还没有和CENTRAL一样的非随机研究数据库。

就像在13.2.3.1中讨论的一样，研究设计分类在作者的使用中并非一致，并且也不总是是被用作书目数据库的索引。策略（c）并没有太大的用处，因为研究设计分类除了随机试验，在书目数据库中进行标引是不确定的，而且研究设计分类也常常和原始研究作者所使用的不一致。有些系统评价撰写者已经尝试开发和建立有效的非随机试验检索策略（Wieland 2005，Furlan 2006，Fraser 2006），也试图优化不良反应研究的检索策略（见第14章，第14.5小节）（Golder 2006b，Golder 2006c）。由于做非随机试验的系统评价的耗时性，试图完善非随机试验检索策略的行动并没有关于此项研究的大量系统评价。因此，系统评价撰写者应该谨慎处理这个假设：以前的检索策略可用于现在的新课题。

13.3.1.3 引文和摘要评价

特别是实施报告标准后，根据题目、摘要很容易将随机试验从检索结果中区分出来。不幸的是，那些所需的符合评估标准的非随机试验的设计细节通常在题目、摘要中未描述，需要阅读全文才能获得。

13.3.2 支持系统评价作者的指南和可利用资源

NRSMG没有推荐与研究设计相关的索引词来限定检索策略。然而，系统评价撰写者可能希望与曾经报道过成功制定一些有效非随机试验检索策略的研究者取得联系（见13.3.1）以及其他做过与自己相似内容的非随机试验的Cochrane系统评价（或其他系统

评价）撰写者联系。

检索非随机试验的时，建议系统评价作者应该检索研究某干预措施的所有研究，而不是限定检索策略来检索特定结果（见第6章）。当检索干预措施的罕见或长期的结果时（通常是不良反应或意外结果），在检索策略中包括某特定结果的自由词和MeSH 主题词也可能是合理的。不良反应方法组的成员已经开展了这方面的研究（见第14章，14.5）。

系统评价作者应该咨询CRG编辑：具体的CRG注册中是否包括具有某些研究设计特点的研究以及寻求CRG和方法学组有关信息检索专家的建议（见第6章，6.7a）。

13.3.3 总结

- 为检出干预措施期望有利的研究，检索策略应包括针对干预措施、研究对象和健康问题的检索式。目前，没有任何推荐的方法是通过研究设计来限定检索策略。
- 系统评价作者在检索与“可疑”不良反应相关的证据时，可能想要检索自己感兴趣的特定结局（如不良反应）。显然，这种方法不能用于干预措施所有可能的不良反应的全面检索（见第14章，14.5）。
- 在随机试验中推荐的穷尽检索，但这对于非随机试验系统评价可能不可取。然而，在这个重要问题上，现在还没有能够指导研究者的研究。

13.4 选择研究和收集数据

13.4.1 当包括非随机研究时，检索策略会有何不同？

检索结果包括大量的不相关的引文和摘要，在有关非随机试验设计的细节方面，通常未能提供详细、足够的信息，而这些是判断合格性需要的，因此获取和阅读大量的全文是为了选择合格的研究，这不同于评价随机试验。

系统评价作者需要收集随机试验系统评价要求的所有数据(见第7章，7.3)，以及描述这些数据（a）原始研究设计的特点（见13.2.2），（b）可能的混杂因素极其控制方法（见13.1.3），（c）影响非随机试验具体的偏倚风险（见13.5.1），（d）结果（见13.6.1）

通常随机试验系统评价作者收集所有结果的原始信息，如对于二分类结局，需要收集研究对象总例数和每组出现结局的例数。如果研究对象被随机分组，可以认为原始数

据是无偏的。对于非随机试验，同样的原始数据是“未经校正”易于受到混杂因素的影响。作者通常还报告通过回归模型“校正”的估计结果，但这不能通过同样的方法进行总结。系统评价撰写者既应该报告纳入各组的样本量，实际分析例数和某事件发生人数，也需要报告任何调整的效应估计值及其标准误或可信区间。这些数据能够用于森林图展现调整的效应估计值和精确度，如果合适，可以进行不同研究数据的合并分析。

有趣的是：系统评价作者的经验是因为非随机试验报告质量差，以至难于查找所需的信息，同时不同的作者可能从同一篇文章中提取不同的信息。数据提取表可能需要根据所研究的问题来制定。由于潜在符合纳入的研究的多样性以及各自不同的报告方式，制定资料提取表时需要选择部分原始研究反复进行。不可能在此之前提前完成资料提取表。

根据报告方式和分析方法不同，采用对效果和不确定性或统计学差异的不同测量指标来呈现非随机试验的结果。统计学专家的建议可能帮助系统评价作者把文章中提供的信息转化成或者“还原”，以从不同的研究中获得一致的效应测量指标。资料提取表要能够处理作者可能遇到的各种研究结果的信息。

13.4.2 支持系统评价作者的指南和可获取的资源

与为确定合格性提供信息一样表13.2a、表13.2b中的问题代表了从非随机实验中提取研究设计特征相关数据的一个方便的清单。在使用这个清单来提取研究信息以及确定合格性时，目的中应该写清楚：在原始研究中研究者做了什么，而不是研究者声称他们做了什么或者认为他们做了什么。每个条目应该用“是”“否”“不清楚”来回答。13.4a为如何使用资料提取表提供了指导。

在NRSMG中，资料提取表已经制定出来并用于非随机试验提取数据。包括：研究设计清单、混杂因素信息提取模版、基线可比性、校正混杂的方法以及效果估计。这些信息（可以从www.cochrane.org/resources/handbook获得）可以用来指导系统评价作者制定需要的资料提取表的类型。然而系统评价作者在制定资料提取表时还需要根据评价的问题仔细考虑。

框13.4.a　使用表13.2.a或13.2.b的清单进行数据收集/研究评估的用户指南

注释：

使用者需要清楚知道“组”和“整群”在表中的使用方式。表13.2.a仅指组，是指传统意义上的多个独立个体。除了在结果基础上的分配，“组”就是指“干预组”。13.2.b包括组和整群。在这个表中，整群通常是指一个组织实体，例如家庭健康实践组织或一个行政区，不是指一个个体。就像在13.2a表中，组与干预组同义，组也用于描述一个分配单元的集合，但在13.2.b中，这些单元就是整群，而不是多个单一个体。此外，尽管个体处在群体之中，但是整群不是一定要代表一些特定个体的集合。如在整群分配的研究中，整群通常是研究在两个或多个时点（段），来各个时点收集数据的不同的个体集合。

是否只有一个对照？

通常，研究者比较接受不同干预措施的两个或多个组；对这些组的研究可能在同一个时间段，也可能在不同时间段（见下文）。有时候研究者比较同一组在两个不同时间点的结果，也有可能研究者两种情况都研究，即研究两个或多个组并在多个时间点测量结果。

研究对象（受试者/整群）通过什么方式分组？

该条目目的是描述如何分组。如果研究没有对研究对象进行两个或多个组的比较，这些条目就不适用。分组信息经常没有报道或很难在文章中找到。这些条目包括分组所要用到的主要方法。尽管有些选项是互相矛盾的（如研究是随机或非随机），但是一个研究仍可以选用不止一个选项。

随机分配：通过真正的随机序列分配研究对象。这类研究在本手册的其他部分均有标准的指导。仔细检查分配方案隐藏是否持续至研究对象招募终止。

半随机：使用假随机序列进行分配，如随意的或者甚至是入院顺序或出生日期，交叉入组。注意：当使用这种假随机方法时，问题在于很少能够做到分配方案隐藏。在只纳入随机试验的系统评价中经常纳入该类研究，可通过便宜风险评估将其与随机试验区别开来。

研究者的其他行为：包括多种情况，如果作者报道了内容应该详细记录下来。作者根据一些决定或系统结果来分配。如来自特定“单元”（如：病房号，全科）的研究对象被选择接受某干预措施，其他人接受对照措施。

时间不同：纳入两组研究对象的时间不同。如在历史对照研究中，通常对照组的研究对象比干预组招募的早；干预措施介入较晚，干预组研究对象招募的也晚。通常两组的研究对象在同一个环境下招募。如果设计不在研究者的控制范围内，这两种选择以及“研究者的其他行为”必须记录在一个研究中。如果通过引进一种新的干预措施，这个设计可以实施，这两种选择以及治疗决策必须在同一个研究中记录。

地点不同：不同地域的两组或多组比较不同组接受的干预和对照干预不是随机产生的。因此这两种选择以及“研究者的其他行动”必须记录在一个研究中。

治疗决策：干预组和对照组是根据治疗决策的自然变化分配，这个选项反映治疗决策主要是由医生负责的；以下的选项主要反映治疗决策是由研究对象的意愿决定的。如果来自同一个特定“单元”的研究对象的意愿不一致或随时间变化，这两个选择已经分配和时间不同应该记录下来。

病人意愿：干预组和对照组根据病人自己的意愿分配，这个选项反映治疗决策主要是由研究对象意愿决定的；前一个选项是要反映治疗决策主要是由医生决定。

基于结局：一组有了特定结果的人与没有这种结果的人比较。即病例对照研究。注意：这个选项应该记录为大样本量的特定结果的多风险因素分析，即是在这种研究中研究对象被分别分入有这种结果和没有这种结果的组。尽管是对连续性病人的前瞻性的收集数据，这种研究比队列研究更接近巢式病例对照研究。

对于整群分配研究的其他选项：

地域不同：见上

政策或公共卫生决策：干预组和对照组是根据公共卫生或服务提供政策执行责任人的决定进行分配。这些决定在整群是一致的或这些人就是研究者本人，这些条目与“研究者的其他行为”和“整群意愿”重叠。

整群意愿：干预组和对照组是按整群自然的意愿进行分配的，如意愿来自整群的个体的一致意见

研究的哪部分是前瞻性的？

这些条目的目的是描述研究的哪部分是前瞻性的。在随机对照试验中，这四条都是前瞻性的。对于非随机试验，尽管没有详细的内容显示出来，尤其是研究假设的形成，这四个条目也可能是前瞻性的，在一些队列研究中，研究对象可被识别且干预组的分配是回顾性的，但其结果是前瞻性的。

评估组间哪些变量具有可比性？

在前后比较的研究中应该区分这些问题。当结果是连续性变量的时候，评估结果变量的基线尤其重要，如：健康状况和生活质量。

回答选项

尽可能只用“是”“否”“不清楚”来回答这些选项。如果研究没有报道两组比较用N/a表示。

13.4.3 总结

- 通过检索来评估检出的引言和摘要是非常费时，首先引言数量大，其次需要用来判断合格性的信息可能没有在标题或摘要中。
- 随机试验的数据收集（即 研究细节，研究对象，纳入的样本量，实际分析的样本大小）
- 收集研究者所研究的数据（NRAMG清单或类似的）
- 收集需要考虑的混杂因素
- 收集影响组间可比性的混杂因素
- 收集用于控制混杂的方法
- 收集关于多重效应估计的数据（如果可以校正前后数据均收集）

13.5 非随机研究的偏倚风险评估

13.5.1 纳入非随机研究时会有什么不同？

13.5.1.1 非随机研究偏倚来源

和设计或完成质量低下的随机试验一样，非随机试验研究结果的偏倚可能来自多个方面（见第8章）。例如，通常非随机试验中排除人数不清楚，干预和结果评估常常不是按照预先设计计划书进行，结果评估可能未采用盲法。由于这些原因导致的偏倚与随机试验中由这些同样原因导致的偏倚相同，系统评价作者应该熟悉第八章中对这些问题的描述。在设计较好的前瞻性非随机试验中要解决这些问题中的任何一个，难度都不比在随机试验中小。

在非随机试验中，使用无随机隐藏的分配机制意味着组间不太可能可比。这些潜在不同干预组间研究对象特征的系统性差别可能是在大部分非随机试验中应该重点考虑的问题，这称为选择性偏倚。如果选择性偏倚导致与研究结局相关的预后因素在组间分布不平衡，此时就会出现混杂。统计学方法有时能通过校正干预效果的估计值来处理产生的混杂偏倚，并且部分研究质量评估可能涉及对分析方法的适用性以及研究设计和执行做出判断。

因为非随机试验包含多种研究设计类型，并且它们对于不同偏倚的易感性不同，这就使得制定一个稳健的通用的用于评价偏倚风险的工具困难重重。在纳入不同研究设计类型的非随机试验系统评价中，需要制定几种工具来评估偏倚风险。系统评价小组中加入方法学专家对辨识纳入的研究设计的重要薄弱点是必不可少的。

在随机试验中，评估偏倚风险主要集中在系统偏倚，这通常被认为是“乐观”的方向。不管是有意识的还是下意识的，研究者设计、实施、分析和报告原始研究以给出期望结果的趋势也可能用于研究者能够控制关键决策（如干预措施分配，中心选择）的非随机试验。真正的观察性非随机试验中，由于适应症混杂导致的偏倚可能不一致；即便根据当前病情严重程度和并发症情况，卫生保健专业人员对于病人改变干预措施的恰当性也可能有不同意见。不同地点间病例组合比较的差别可能是随意的。因此，进行非随机试验系统评价时，偏倚的变异度及由此带来的组间异质性至少与系统偏倚是一样重要的。

13.5.1.2 非随机研究偏倚风险的证据

通过比较低偏倚风险和高偏倚风险的随机试验可以获得一些对非随机试验偏倚风险的深刻理解。半随机分配研究对象或在招募研究对象期间分配方案隐藏失败的对照试验存在发生选择偏倚的风险，就像是一个前瞻性的公开的非随机试验或队列研究。第八章中评估了随机试验偏倚风险的几个方面的证据，并且指出随机试验中那些可能夸大干预效果有效性的方法学限制。

研究者比较了对于同一个研究问题单独纳入随机试验和非随机试验的Meta分析结果，认为这种系统评价的方法提供了一种研究非随机试验偏倚风险的方式。这类系统评价还报告了由研究设计产生的差异，但是很难做到客观的比较。原因有二：

- 针对恰好相同的问题的随机试验和非随机试验很少见；例如，使用不同设计方法研究同一干预措施的研究间，通常在研究人群，干预或结果方面根本不同，。
- 随机试验和非随机试验在与偏倚风险相关的方面可能根本不同（报告偏倚，选择性偏倚，适应症偏倚，测量和分配偏倚）以及非随机试验往往质量相对较差。

这些原因可以解释针对同一个研究问题的随机试验和非随机试验的方法学系统评价的结论的不一致性。Deek等人(Deeks 2003)评估了8个这样的的系统评价发现：

- 5个结论显示：随机试验和非随机试验对于很多但不是全部干预措施的评估效果不同，无一致性。
- 1个结论显示：在所有的干预研究中，非随机试验过高估计了干预的有效性（有益结果）
- 2个结论显示：随机试验和非随机试验估计效果惊人的相似。

还有一个采用类似方法学的系统评价比较了随机试验和病人意愿研究的结果（King 2005）。系统评价结果显示：无证据显示病人意愿“明显影响有效性”，病人意愿没有显示出对干预措施效应的混杂作用。

在解释这种实证研究数据时要谨慎。首先，发表的原始研究和系统评价作者选择的原始研究时可能均有偏倚存在。评价者对系统评价结果进行分类时也可能就存在偏倚。Deek等人发现，有时候同1个比较项目在2个系统评价中的分类不同，在前者可能是矛盾的，后者则可能是可比的，这就增加了定义“差异”的难度。

其次，对这些并不总是乐观的差异的观察仍是一个重要问题，并且是与较之按机遇的预期，非随机试验的效应估计值具有更大的异质性这一原则是一致的。（Greenland 2004）。一些创新性的模拟研究证明了该问题（Deeks 2003）。Deek等人指出非随机试

验中的偏倚更易变，而且可能最容易被认为是在结果中引入了额外的不稳定因素，而不是可估计的系统偏倚。这种包含在可信区间中的不确定性，在大量的研究中可能很容易使得95%区间的范围增大至5-10倍。

最后，方法学系统评价陷入了一个循环，它们需要假设非随机试验是有效地，并因此随机试验和非随机试验间效应估计值的不同也是有效的，并且可归结于外部因素，或者假设非随机试验有偏倚，因此随机试验和非随机试验间效应估计值的不同可以解释为不同的偏倚风险。真相可能介于这两个极端之间，但事实是方法学系统评价也不可能明确划分不同来源的差异。此外，如果多因素能区分随机试验和非随机试验并且能影响效应尺度，那么观察到随机试验和非随机试验效果大小无差异时就可以解释为多因素在不同方面对干预效果的影响；假设结果没有差异意味着非随机试验有效，而结果有差异则意味着非随机试验无效，这是不符合逻辑的。

13.5.2 支持系统评价作者的指南和可用的资源

13.5.2.1 评估非随机研究偏倚风险的整体思考

随机试验的报道相对来说比较简单和常见，有CONSORT声明指导（Moher 2001）。虽然是最近提出的，STROBE为报道观察性流行病学研究提出了一个类似的共识声明，（Vandenbroucke 2007）。因此，报告评估偏倚风险所需信息的质量要求，对于非随机试验可能不太适合。这有可能阻碍任何偏倚风险的评估。

计划书是防止发生偏倚的一种工具，在研究开始前进行注册，说明研究设计和分析要在招募研究对象前考虑，同时数据定义和标准化数据收集方法也应该确定。由于研究需要伦理学认可，所有的随机试验必须要有计划书，即便它们的质量和具体条目不同。许多随机试验，尤其是受企业支助的，也要有详细的研究手册。因此由计划书提供的保护在非随机试验中往往没有。还没有对于无计划书所带来影响的研究。然而，这意味着，举例来说，对研究者“择优选择”结果和亚组及分析方法来报告的趋势没有限制，即使在有计划书的随机试验中，这些情况发生机会可能更大或更小（Chan 2004）。

与随机试验一样，偏倚评估的维度包括选择性偏倚（涉及组间可比性，混杂和校正），实施偏倚（涉及干预精确性、研究队象接受何种干预的信息质量，包括对研究对象和医务人员施盲），测量偏倚（涉及无偏倚和正确的评估研究结果，包括对评估人员实盲），失访偏倚（涉及样本完整性，随访数据）以及报告偏倚（涉及发表偏倚和选择性报告结

果）。在随机试验中，通过识别用于阻止以上各种偏倚的研究设计特征和标记每个研究是否满足要求，偏倚风险的评估方法已经得到了提高。评估非随机试验的偏倚风险应该采用同样的方法，事先确定计划书中待评估的特征，记录研究过程中发生了什么，以及判断这些用来避免特定偏倚的方法是否足够或不清楚。确定这些特征时可能需要咨询流行病学专家，并结合具体的临床问题。尤其要注意混杂偏倚的评估。（见13.5.2.2）。

需要特别注意原始研究设计特点（如研究对象如何分组，哪些部分是前瞻性的等）而不是设计类别（如队列或横断面研究）是因为假设偏倚风险受研究的特定特征影响而不受所用方法的大体分类的影响。此外，像“队列”和“横断面”研究这些名词定义模糊不清，并涵盖了多种具体研究设计。尽管根据病因学研究和随机试验中偏倚风险的证据和理论可以构建一个名单，但没有描述与偏倚风险相关的研究设计特定的经验衍生列表（见13.2.2和13.4.2）。

由于非随机试验的多样性，不同设计特点需要不同的评估方法。根据结局（如病例对照研究）与更多基于干预措施分组的研究间存在重要区别。在前1种类型的研究中，最容易发生偏倚的是感兴趣的暴露因素而不是结果，系统评价撰写者应该询问研究者暴露的评估是否是在不知道研究对象出现结局与否（即病例还是对照）的情况下进行。病例对照研究非常适合研究罕见结果和多种暴露之间的关系，因此可能在生产关干预措施潜在副作用和意外的有益结局的证据上有重要作用。它们也被用于评估大规模公共卫生干预（进行随机试验评估很困难或耗资巨大），如事故预防和筛查(MacLehose 2000)。然而，系统评价撰写者应该熟悉尤其是应用此类研究的流行病学方面的考虑(Rothman 1986)。注意，一些病案分析与病例对照研究（例如，假定整个数据被分成发生特定结果和未发生特定结果两组，观察与结果相关的暴露因素。）相似。与随机试验相比，系统评价撰写者评估非随机试验的偏倚风险时需要更深入的流行病学相关知识。

13.5.2.2 混杂与校正

针对混杂因素研究者并不总是做出同样的处理方式，因此，用来控制混杂的方法是研究间异质性的重要来源。所考虑的混杂因素、控制混杂的方法以及数据分析中用于测量混杂因素的准确方法可能有不同。很多（并非全部）非随机试验描述了可能的混杂因素以及在研究设计和分析时是否考虑了混杂因素；大部分也报告了对比组间的基线特征。然而，评价研究者实际做过什么来控制混杂因素可能很难；很少有研究具体描述如何测量混杂因素或进行协变量模型分析（如连续变量，有序分类变量或无序分类变量等）。

评估选择性偏倚的一些具体建议有：

- 在撰写计划书阶段，列出潜在的混杂因素
- 找出研究者考虑到的和那些忽略的混杂因素。注意测量混杂的方法（混杂因素的控制能力取决于对因素的准确测量）。
- 评估对比组间主要预后因素或混杂因素基线的均衡性。

找出研究者控制混杂偏倚方法，即，出于此目的的研究设计特征（如对特定亚组的限制和匹配）和数据分析方法（如分层分析或带有倾向性评分和协变量的回归模型）。

没有用于识别一组预先指定的重要混杂因素的现成的方法。或许有人会说，应该“独立地”“系统地”列出潜在的混杂因素。不应该只是列出纳入系统评价的原始研究关注的混杂因素（至少一些独立的验证形式），因为潜在混杂因素的数量可能会随时间增加（因此，较老的研究可能过时），并且研究者自身可能也只是简单地选择测量一些先前的研究所关注的混杂因素（因此，这样的列表就有选择性），（但研究病因关系的研究者往往没有对他们选择的混杂因素做出解释（Pocock 2004）。）当然，这个列表应以证据为基础（尽管做系统评价要确定多种潜在预后因素是极端困难的），兼顾系统评价小组成员和顾问专家的意见。

在Cochrane系统评价中呈现混杂因素评估的结果最好是创建一个表格，以先前提到的混杂因素作为列，具体研究作为行，指明每个研究是否：①限制性地选择研究对象，以至所有组具有相同的混杂因素（如限定纳入对象都是男性）；②说明组间的混杂因素分布均衡；③匹配混杂因素；或④为量化效应的大小对统计分析中的混杂因素进行校正。

13.5.2.3 非随机试验中评估方法学质量或偏倚风险的工具

第八章（8.5）中描述了系统评价撰写者期望用来评估随机试验偏倚风险的工具。有6个特征需要考虑：① 随机序列产生；② 分配序列的隐藏；③ 盲法；④ 不完整数据；⑤ 选择性报告结果；⑥ “其它”可能的潜在偏倚。评估条目：① 描述研究中发生了什么；② 提供对于研究项目是否足够的判断。判断是通过回答预先制定的问题，如回答“是“表示偏倚风险低，“否”表示偏倚风险高，“不清楚”表示偏倚风险不清楚或未知。制定这个工具时未将非随机试验考虑在内，而且这6个方面未必适合非随机试验。但这个工具的大体框架和评估表有助于制定非随机试验偏倚风险的评估工具。

对于实验性对照研究和前瞻性队列研究（见表13.1a，13.2.2）标准的偏倚风险工具中的6个方面可以有效地进行评估，不论分组是否随机。这是系统评价撰写者应该做到

的最低要求，通常要求提供更多细节。需要额外评估由于混在所导致的偏倚风险。评估的深度可能要依赖于各研究间的异质性以及系统评价撰写者是否打算定量合成（见13.6章）。如果各研究具有异质性且没有定量合成的打算，那么不太详尽的评估仍然可以说明异质性及解释系统评价的结果。

许多评估非随机干预性研究的方法学质量的工具已被制作已创建，并且Deek等人进行了系统评价（Deeks 2003），在该系统评价中，找出了182种工具，最终减少到了14种并选出了其中认为对系统评价可能有用的6种，因为它们“迫使系统评价撰写者进行系统的研究评估并以可能最客观的方式做出质量判断”。然而，所有6个方面都需要一定程度的调整，因为他们忽视了关于研究对象如何分组的详细信息，从选择性偏倚方面来说，这是很关键的。不是所有6个工具都适合不同研究设计，与其它用来评估随机试验质量的一些工具一样，有的工具没有区分研究质量条目和报告质量条目。在这个系统评价中发现的两个最有用的工具是Downs and Black 工具and the Newcastle-Ottawa 量表（Downs 1998，Wells 2008）。

Downs and Black工具已经被修改并用于一个方法学系统评价（MacLehose 2000），系统评价撰写者发现29个条目中有些条目不适合病例对照研究，这个工具需要大量流行病学专家，并且使用起来费时。The Newcastle-Ottawa 量表被NRSMG工作组用于回答从原始的非随机试验中提取数据的问题，包括8个条目，使用起来简单（Wells 2008）。然而，这些条目仍然需要根据评估的具体的问题进行修订。系统评价撰写者也需要意识到不同国家的流行病学术语不同，如Newcastle-Ottawa量表使用“选择性偏倚”，其他可能用“适用性”或“普遍性”。承认区别“研究者做了什么”和“研究者报告了什么”的重要性，系统评价撰写者将发现它有助于考虑纳入随机试验（Moher 2001）和观察性流行病学研究（Vandenbroucke 2007）报告报表的条目，以突出非随机试验（Reeves 2004，Reeves 2007）在报告（和执行）方面的差距。

13.5.2.4 非随机研究偏倚风险评估中的困难

纳入非随机试验的2个系统评价研究指出只有少数的系统评价评估了纳入研究的方法学质量（Audige 2004，Golder 2006a）。NRSMG成员已经积累了试着评估非随机试验偏倚风险的经验。有趣的是，系统评价作者已经报道非随机试验普遍方法学质量低或报告质量差，以至于通过原始研究评估方法学质量和偏倚风险比较难或者说是不可能的（Kwan 2004）。甚至有报道说Newcastle-Ottawa量表不太好用，因此，想要在系统评价

撰写者间达成一致的可能性较小。方法学部分的信息很难从文章中找到，使得评价过程令人沮丧，尤其是使用一些更为详细的工具时；系统评价作者可能要花很长时间查找研究者做了哪些具体工作，最后发现这些文章中没有报告的信息。然而，收集一些确切信息（如混杂因素及研究者如何处理混杂）仍是有用的，因为这些信息说明了研究间异质性的程度。

13.5.3 总结

- 在撰写计划书阶段，将可能存在的潜在混杂因素及选择的理由汇总成表；
- 在撰写计划书阶段，决定怎样评估原始研究的偏倚风险，包括对混杂的控制程度；
- 根据协作网推荐的用于随机试验的方法，进行非随机试验完全前瞻性的研究；
- 没有单独的推荐工具，因此系统评价作者可能需要补充偏倚风险的工具或条目；
- 像混杂这类问题不可能简单通过新的偏倚风险工具来处理，需要建立其他报告评估表格；
- 收集一些确切信息（如混杂因素及研究者对偏倚做了什么处理）是有帮助的，因为这些信息说明了研究间异质性的程度；
- 在Cochrane系统评价中选择纳入病例对照研究的系统评价作者应该熟悉一般常见的可能影响这些研究的潜在的因素，以及使用以此目的研究设计工具评估偏倚的敏感性；
- 系统评价作者可能认为收集大量关于混杂风险以及其他偏倚的详细信息是不合理的，但如果使用这种方法，系统评价撰写者一定能够知道研究间由于潜在的残余混杂或其它偏倚导致的异质性的程度，以及表明在他们在解释纳入的原始非随机试验的系统评价的结果时他们已经考虑了这些异质性来源。

13.6 从非随机研究中整合数据

13.6.1 当纳入非随机研究时会有什么不同？

与随机试验系统评价相比，在非随机试验中系统评价撰写者应该预料到更大的异质性。因为非随机研究系统评价增加了潜在的方法学多样性，因为原始研究间选择性偏倚

风险的差异，在数据分析中混杂的处理方式的差异以及由于低质量研究设计和实施导致的其它更大的偏倚。目前，没有方法在原始分析中可以控制这些偏倚，也没有确定的方法评估这些偏倚是如何影响或者影响程度有多大（见第8章）。

有一种观点认为当非随机试验有明显效果时，对结果进行整合是合适的，但是这种观点的逻辑受到质疑，有明显效果的非随机试验与效果微弱的非随机试验相比，可能（或极有可能）存在偏倚和异质性。判断偏倚风险和异质性应该基于对纳入研究的方法和特征严格评估，而不是基于他们的结果。

系统评价作者在进行Meta分析前评估研究的相似性时，应该记住研究的一些特征，例如，结果评估是在知道干预措施分配的情况下完成，虽然这让非随机研究间的可能相对同质，但也让所有研究均存在偏倚风险。如果作者认为纳入的非随机研究在这个方面不易发生偏倚且相对同质，他们可能希望将研究中的数据进行Meta分析（Taggart 2001）。非随机试验不同于随机试验，它通常需要进行分析校正效应估计值，而不是非校正，即是试图控制混杂因素的分析。这就要求作者从两个可替代的校正效应估计值间选择一个用于研究报告。校正效应估计值的Meta分析可通过逆方差加权平均进行，例如在Revman中使用的Generic inverse-variance结果类型（见第9章9.4.3）。原则上，随机试验Meta分析中使用的任何效应测量指标也可用于非随机研究的Meta分析中（见第9章9.2），尽管比值比（OR值）通常是病例对照研究的二分类结局的唯一效应测量指标，但是还可用于Logistic回归校正混杂因素。

存在一种危险是一项方法学质量低下（如常规方式收集数据）的大规模非随机研究可能主导其它多个较小偏倚风险（可能使用个性化的数据收集）非随机研究的结果。作者需要谨记：与小规模非随机试验相比，大规模非随机试验效应估计值的可信区间不太可能显示研究效应真实的不确定性（见13.5.1.2），尽管没有办法估计或校正这个问题。

13.6.2 支持系统评价作者的指南和可用资源

13.6.2.1 控制混杂

非随机试验中分布不平衡的预后因素必须纳入统计分析。有几种方法可以用来控制混杂。匹配，即在分组时，使各干预组的重要预后因素相似，用于研究设计阶段就减少混杂。分层和回归模型是控制混杂的统计学方法，得出校正了组间不平衡的预后因素后的干预措施的效应估计值。有些统计分析把倾向性评分法作为两阶段分析的一部分。首

先根据对象的特点采用Logistic回归模型估计一个研究对象接受实验性干预措施的可能（倾向评分），这一病例组合的综合测量结果随后用来匹配，分层或回归模型分析。

匹配

选择具有相同程度预后因素的病人能够使组间可比性好。因此，匹配可以看做是校正混杂的一种方法。匹配可以在个体水平（即对于一个干预的病人选择一个或多个具有相似特征的对照）也可以在层面水平（即，对照组在一个层面（如60岁或更大年龄）与干预组选择数目大致相同的病人）。为了获得干预措施效应估计合适的可信区间，只要采取了直接匹配，研究应用成对的数据的进行统计分析。匹配单个测量指标，如倾向性评分，比匹配一组特定特征的个体更易实现。

分层

分层是根据相关预后因素的不同种类（或定量的分类）将研究对象分为多个亚组，如将年龄分成十岁一组，或体重分成四组。估计每层的干预效果并计算层间合并的效果估计。这个过程可以解释为在个体水平上进行Meta分析。对于二分类变量结果，采用OR值（odds ratio）、RR值（risk ratio）和风险差异（risk difference）作为干预效果的测量指标，常用Mantel-Haenszel方法来评估干预的总体效果。另外，倾向性评分可以作为分层变量。

建模在建模方法中，干预和预后因素的信息都被纳入到回归方程中。回归模型的优点包括两种可能，即合并没有类别的定量因素以及构建按等级测量的混杂因素的趋势模型。对于二分类结局变量Logistic回归模型总是用来估计校正的干预效果。因此，OR值用来测量干预效果。回归模型也可以用于RR值和ARR值效应测量指标，但是这些模型在实践中很少使用。一种线性回归模型通常的用于连续性变量（可能在一个或多个变量变换后），比例风险回归模型（COX模型）通常用在时间事件数据。回归模型也可仅用倾向性评分或与参与者的其他特征一起作为解释变量。

系统评价作者应该知道在任何非随机试验中，即使试验组和对照组基线看似可比，由于其他混杂因素可能存在，效应尺度的估计仍然存在偏倚风险。这是因为所有控制混杂的方法都是不完善的，如以下的原因：

- 未知的混杂因素，因此不可测量，不能控制
- 混杂因素测量不精确，如用评估并发症基于简单的等级划分（Concato 1992）不能表现出混杂因素错误分类会出现什么不同。
- 在配对分析中，对匹配程度，受试者间可匹配的混杂因素数量的实际限制。

- 在分层分析或分析处理中，正如层的宽度（十岁）所示，测量混杂因素的方法不够精确；当混杂因素被分类并离散地引入模型，这种限制也存在于回归模型。
- 由于对混杂因素和结果之间联系的相关知识不了解，可能将混杂因素纳入回归模型。

没确切的方法判断残余混杂因素可能的大小。混杂偏倚的方向不能预测，并且不同研究间可能不同。

13.6.2.2 合并多个研究

可以预期，估计的不同研究设计类型干预措施的效果会受到不同来源偏倚不同程度的影响（见13.5）。应该看到不同研究设计的研究结果的根本性不同，导致异质性增加。因此，我们建议不同研究设计类型的非随机试验（或有不同研究设计特点）或随机试验和非随机试验不要合并进行Meta分析。

由于需要尽可能的控制混杂因素，估计的干预效果及其标准误（或可信区间）是在非随机试验Meta分析中合并的关键信息。（除非在研究设计阶段研究组进行匹配，否则只是分子和分母，均数和标准误不能在干预组和对照组控制混杂因素。）因此，基于效应估计值和标准误的Meta分析方法，尤其是通用逆方差方法，将适用于非随机试验（见第9章，9.4.3）

如果一个原始非随机试验报告了特定结果的校正估计，那么提取校正效果估计值及其标准误用于Meta分析就很容易。然而，很多非随机试验同时报告了校正的和没有校正的效应估计值，并且一些非随机试验通过对不同协变量组合的分析，报告了多个校正的估计值。系统评价撰写者应该同时记录校正的和没有校正的效应估计值，但是很难在可供选择的校正估计值间做出选择。尚无指出选择哪种校正的估计会更好的一般建议。可能的选择原则是：

- 使用模型校正的最大协变量
- 使用作者主要校正的模型的估计
- 使用包括研究开始时作者就考虑到的混杂因素和重要因素最多的模型产生的估计
- 敏感性分析可以通过分别合并每个纳入研究中的阳性和阴性结果来完成。

当用OR或HR表示时，对于解释校正和没有校正的效应量在统计学上有点不同。没有校正的效应估计是人群的平均效果，如果估计没有偏倚存在，那么这个估计效果就是对具有中等混合预后特征的人群干预产生的效果。当针对预后特点进行校正，估计的效

果就是有条件的估计效果，并且这种干预效果只能在具有特定的校正的协变量组合的组间观察得到。数学研究显示条件性的估计值通常比人群平均估计值更大。这种现象可能不会在系统评价中观察到，因为估计的研究间存在异质性。

13.6.2.3 异质性分析

探索研究间异质性的可能来源是任何Cochrane系统评价的一部分内容，在第9章（9.6节）有详细讨论。非随机试验比随机试验能够预期到更大的异质性，这是因为存在其他的方法学多样性和偏倚。显示研究间结果差异最简单的方法是制作森林图（见第11章，11.3.2）

即使在系统评价中研究之间由于异质性太大不能合并分析，通过Meta回归分析确定重要的异质性的决定因素也可能是有价值的。这种分析方法可能有助于发现与干预措施效果系统性相关的方法学特征，并有助于找出最可能得出干预措施效果的真实估计值的研究亚组。

13.6.2.4 哪些情况不适合合并

进行Meta分析前，系统评价撰写者必须反复确认原始研究是否‘足够相似’，判断是否可以合并（见第9章，9.1）。在Revman中，使用‘通用逆方差’方法表示结果，在森林图中显示出每个研究的估计值及标准误。Meta分析可以取消或在森林图中只用于亚组分析。鉴于纳入的研究的效应估计值可采用一致的测量方法，我们推荐系统评价撰写者用森林图展示每个具有相似研究设计特点的非随机试验的研究结果，作为标准特征。如果没有一致的测量方法，那么就要用另外的表格系统性的显示研究结果。

如果纳入研究的同质性不足以进行Meta分析（被认为是纳入非随机研究的系统评价的准则），NRSMG推荐用森林图展示纳入研究结果，但是不进行合并估计。在森林图中，研究可能依据研究设计特征分类（或分别用森林图显示），或一些其它能反映偏倚敏感性的特征（如Newcastle-Ottawa量表的‘星’数（Wells 2008））。甚至，当已经知道不能计算合并的效应估计值时，进行异质性诊断和调查（如异质性检验，I^2统计量和Meta回归分析）也是值得的。

但是，描述性的合成还是存在问题的，因为陈述或描述没有经过选择或与别人比较而强调的结果是很困难的。理论上，作者应该在系统评价计划书中指出他们计划如何应用描述性的合成来报告原始研究结果。

13.6.3 总结

- 非随机试验系统评价中研究间异质性比随机试验系统评价中研究间异质性要大。因此，当决定是否定量合并多个结果时（即Meta分析），作者应该仔细考虑纳入研究间异质性可能的大小。我们希望对非随机试验效应估计值的合并是特例，而不是常规。
- 非随机试验效果估计值不能与随机试验或研究设计不同的非随机试验效果估计值进行合并。
- 应该用森林图来总结纳入研究的结果。
- 不管是否合并不同研究结果的效应估计值，都要进行异质性的判断和调查分析。

13.7 解释与讨论

13.7.1 解释纳入非随机研究的有效性的Cochrane系统评价结果时所面临的挑战

系统评价作者面临着巨大的挑战，他们需要另人信服地说明非随机试验系统评价的结果能够对干预措施的可能效果给出明确的答案。（Deek 2003）。在许多情况下，非随机试验系统评价有可能得出计算“平均”效果没有用处，以及来自非随机试验的证据不足以证明有效或有害（kwan 2004）及应该进行随机试验的结论。（Taggart 2001）

非随机试验系统评价研究过程中充满挑战：决定纳入哪种类型的研究，检索文献，评估研究潜在的偏倚及决定是否合并结果。系统评价撰写者需满足读者对已充分解决这些挑战的系统评价的需求，或者应该讨论如何解决及为什么没有解决这些挑战。在这部分，这些挑战参照本章提出相应问题的不同部分。系统评价讨论部分应该指出哪些挑战已被解决。

13.7.1.1 是否所有重要和相关的研究都全已纳入？

即使选择合格研究设计的方法是合理的，它可能很难证明所有相关研究已被全部找出，这是因为低质量标引、研究者使用研究设计的分类方法不一样。如果全面检索策略仅关注健康状况和感兴趣的干预就有可能导致包括很少合格研究的大量引文出现；但是

限制性的检索策略又不可避免的遗漏合格的研究。实际上，现有资源可能使作者无法处理全面检索到的结果，尤其是作者通常要阅读全文而非摘要来确定合格研究。目前还不清楚使用一个或多或少的全面搜索策略的影响。

13.7.1.2 纳入研究的偏倚风险有没有被全面评估？

解释非随机试验系统评价的结果必须考虑偏倚大小和方向。影响随机试验的偏倚也会影响非随机试验，但是通常来说对非随机试验影响更大。在非随机研究中研究对象减少很严重（且很少报告），极少根据计划书进行干预和结果评估，对结果评估很少实施盲法。通常这些非随机试验的限制被看做是进行非随机试验的一部分，这些因素对偏倚风险的影响也没有进行适当的考虑。如一些证据使用者可能认为观察长期结果的非随机试验可能被认为比短期结果的随机试验质量要好，只是简单的根据它们相关性的判断而没有评估偏倚风险（见13.2.1.4）。

评估非随机试验中混杂因素的程度尤其存在问题。系统评价撰写者不仅需要有充足的评估方法，也要收集和报告研究者考虑到的混杂因素及用来控制混杂因素方法的详细信息。纳入的原始研究可能没有报道这些信息，妨碍了系统评价作者观察合格研究方法的不同及异质性的其他来源，在撰写计划书时考虑到这些是非常重要的。

作者必须记住以下关于混杂因素的要点：

- 由混杂导致的偏倚的方向的不确定。
- 不同研究控制混杂因素的方法不同；
- 任何特定研究中残余混杂的大小是不知道的，并且在研究间有可能变化。
- 残余混杂（和其他偏倚）意味着可信区间低估了效果估计值范围真实的不确定性
- 识别校正的或没有校正的可能混杂因素是非常重要的。

上述的挑战对所有非随机研究系统评价均有影响。然而，这些挑战在一些卫生领域可能不那么极端（如在长期观察性研究或副反应研究中混杂不是主要问题，或一些公共卫生初级预防干预措施研究中）。存在偏倚的一个表现就是研究的间异质性。尽管在研究对象、干预措施及结果评估中存在不同能引起异质性，但是非随机试验系统评价中必须认真考虑偏倚可能是异质性产生的原因。然而，没有异质性不表示没有偏倚，因为可能一种偏倚在所有研究中都存在。

能够预测偏倚的大小和方向吗？这是一个正在研究的课题，该课题试图收集能够决

定偏倚大小和方向的因素（如研究设计和干预类型）的相关经验证据。预测偏倚可能的大小和方向的能力将极大地提高对非随机试验系统评价提供的证据的使用程度。目前，有一些证据表明，至少在某些特定情况下，偏倚的方向是可以预测的。（Henry 2001）。

13.7.2 评估纳入非随机试验的系统评价的证据强度

“暴露”非随机试验对特定健康问题的证据有助于充分讨论它的意义和重要性及能用来解释它的确定性。严格上说，需要讨论观察到的结果可能产生错误的机会的大小。正式的证据分级总是将非随机试验放到较低位置，但是都高于临床意见。（Eccles 1996，National Health and Medical Research Council 1999，Oxford Centre for Evidence-based Medicine 2001）这些都强调在非随机试验中要注意偏倚，以及把因果关系归于观察效应的难度。由非随机试验系统评价提供的证据强度有可能依赖于解决13.7.1中列出的挑战的程度。解决这些挑战的能力将随着医疗环境和结果改变。在某些环境中可能没有混杂存在。如婴儿接种疫苗时因为不知道预后信息，限制了可能的混杂（Jefferson 2005）。

不管讨论得出需要随机试验还是依据非随机研究的证据足够做出明智的决策，均依赖于使用有潜在偏倚的研究设计带来的成本的不确定性和所研究效应的总效应值大小。这个值的大小可能取决于医疗环境的广泛性。纳入系统评价内部的评估值是不可能的，（系统评价评估出来的值）它作为证据只能作为后续出版的更广泛讨论的一部分。

例如，由非随机试验产生的关于罕见严重副反应的证据是否足够决定这个干预措施不能应用？证据是不确定的（由于缺乏随机试验），但是已知的值显示有可能潜在的严重危害，或许足以决定撤销这种干预措施。（但值得注意的是，有关撤销干预措施的决定可能取决于是否同样的益处可以从其它没有这种风险的干预上获得；如果不能，这种干预仍可使用但是要全面的披露潜在危害）。有益处的证据不是基于随机试验，因此这种益处是不确定的，有害性非随机试验系统评价的值可能更大。

相反的，非随机试验系统评价中，一种新的效果不明显的干预措施的证据，不足以让决策者在面对这种不确定的证据和提供干预所带来的巨大成本的情况下，做出建议广泛使用的决定。在这种情况下，决策者有可能认为：如果随机试验可行并且此时的投资在将来可能有回报，那么应该进行随机试验研究。

在Cochrane系统评价的结果总结表中，推荐使用GRADE用于评估一系列证据质量的方案，这部分内容已在12章介绍了（见12.2）。分为四个质量水平：高、中、低、非常低。

一组研究大致分类为，随机试验是高级别证据，可能由于研究的局限性（偏倚风险）、异质性、间接证据、精确度不够或发表偏倚而降级。观察性研究是低级别证据，可能由于效应强度很大、可能的混杂因素未考虑、效剂量-效应关系而升级。系统评价撰写者需要决定来自非随机试验的证据是否应该升级或可能降级（如半随机试验）。

13.7.3 对潜在系统评价作者的指导

实施非随机试验系统评价比随机试验系统评价更加困难。很可能在系统评价的每个阶段都要做出复杂的决策，需要流行病学家或方法学专家的建议。。所以潜在的系统评价撰写者应该尝试与流行病学家或方法学家合作，暂不考虑系统评价的目的是否是观察危害还是收益，短期还是长期的结果，常见还是罕见事件。

卫生保健专业人员热衷于在那些没有或很少有随机试验的领域进行非随机试验系统评价研究提高他们专业领域的证据基础（大部分Cochrane系统评价的动机）。方法学家也希望有更多的非随机试验系统评价发表，以发现在本章节中强调的方法学上不确定的领域。然而，卫生保健专家也应该意识到：a做非随机试验系统评价需要的资源有可能比随机试验系统评价需要的资源更多；b结论可能更弱，有可能对研究问题的贡献相对较少。因此，系统评价撰写者以及CRG的编辑需要在初期决定投入的资源是否值得，研究的重点问题是否有可能被证明。

整合所需的卫生保健专业人员和方法学家的团队可能会使得估计干预措施长期罕见的不良结局的非随机试验系统评价相对容易，如药物的副作用。然而，这些系统评价可能要求引入其他方面的专家作为合作者，如相应的药学专家。在很多健康条件下，迫切需要提供除了传统有效性的随机试验系统评价外，还需要不良反应的系统评价，可能这些系统评价通常将需要纳入非随机试验。

13.8 本章信息

作者：Cochrane非随机研究方法学工作组的代表Barnaby C Reeves, Jonathan J Deeks, Julian PT Higgins和 George A Wells

本章引用方式：Reeves BC, Deeks JJ, Higgins JPT, Wells GA. Chapter 13: Including non-randomized studies. In: Higgins JPT, Green S (editors), Cochrane Handbook for Systematic Reviews of

Interventions. Version 5.0.2 [updated September 2009]. The Cochrane Collaboration, 2009. Available from www.cochrane-handbook.org.

致谢：感谢Ole Olsen, Peter Gøtzsche, Angela Harden, Mustafa Soomro, Guido Schwarzer and Bev Shea 提供不同章节的草稿，同样感谢Laurent Audigé, Duncan Saunders, Alex Sutton, Helen Thomas and Gro Jamtved对终稿的审校。

Box 13.8.a Cochrane非随机研究方法学工作组

Cochrane协作网非随机研究方法学工作组（NRSMG）建议成立指导小组，针对Cochrane系统评价中纳入非随机的关于医疗保健干预措施研究的标准制定政策和指导原则,该小组成员可以是愿意为工作组做出贡献的任何人。小组工作的重点主要针对方法学，而不是特定医疗保健干预措施。

NRSMG成员工作内容包括：

- 制定相关指南以确定Cochrane系统评价何时纳入非随机研究数据。
- 进行非随机研究的方法学研究，包括检索方法、质量评价、Mete分析、缺陷与误用。
- 比较系统评价中同时使用随机和非随机研究产生的各种可能的偏倚，鉴别随机和非随机研究导致相似结论或有些结论相互矛盾产生的可能条件。
- 收集医疗保健问题的案例：（a）同时纳入随机试验和非随机试验进行研究，（b）目前还没有（或很长时间内没有）的随机试验研究。
- 在每年的Cochrane年会提供专题培训。

13.9 参考文献

Audige 2004

Audige L, Bhandari M, Griffin D, Middleton P, Reeves BC. Systematic reviews of nonrandomized clinical studies in the orthopaedic literature. Clinical Orthopaedics and Related Research 2004: 249-257.

Chan 2004

Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004; 291: 2457-2465.

Concato 1992

Concato J, Horwitz RI, Feinstein AR, Elmore JG, Schiff SF. Problems of comorbidity in mortality after prostatectomy. JAMA 1992; 267: 1077-1082.

Deeks 2003

Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG. Evaluating non-randomised intervention studies. Health Technology Assessment 2003; 7: 27.

Doll 1993

Doll R. Doing more good than harm: The evaluation of health care interventions: Summation of the conference. Annals of the New York Academy of Sciences 1993; 703: 310-313.

Downs 1998

Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. Journal of Epidemiology and Community Health 1998; 52: 377-384.

Eccles 1996

Eccles M, Clapp Z, Grimshaw J, Adams PC, Higgins B, Purves I, Russel I. North of England evidence based guidelines development project: methods of guideline development. BMJ 1996; 312: 760-762.

Fraser 2006

Fraser C, Murray A, Burr J. Identifying observational studies of surgical interventions in MEDLINE and EMBASE. BMC Medical Research Methodology 2006; 6: 41.

Furlan 2006

Furlan AD, Irvin E, Bombardier C. Limited search strategies were effective in finding relevant nonrandomized studies. Journal of Clinical Epidemiology 2006; 59: 1303-1311.

Glasziou 2007

Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ 2007; 334: 349-351.

Golder 2006a

Golder S, Loke Y, McIntosh HM. Room for improvement? A survey of the methods used in systematic reviews of adverse effects. BMC Medical Research Methodology 2006; 6: 3.

Golder 2006b

Golder S, McIntosh HM, Duffy S, Glanville J, Centre for Reviews and Dissemination and UK Cochrane Centre Search Filters Design Group. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. Health Information and Libraries Journal 2006; 23: 3-12.

Golder 2006c

Golder S, McIntosh HM, Loke Y. Identifying systematic reviews of the adverse effects of health care interventions. BMC Medical Research Methodology 2006; 6: 22.

Greenland 2004

Greenland S. Interval estimation by simulation as an alternative to and extension of confidence intervals. International Journal of Epidemiology 2004; 33: 1389-1397.

Grobbee 1997

Grobbee DE, Hoes AW. Confounding and indication for treatment in evaluation of drug treatment for hypertension. BMJ 1997; 315: 1151-1154.

Henry 2001

Henry D, Moxey A, O'Connell D. Agreement between randomized and non-randomized studies: the effects of bias and confounding. 9th Cochrane Colloquium, Lyon (France), 2001.

Higgins 2003

Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in Meta-analyses. BMJ 2003; 327: 557-560.

Jefferson 2005

Jefferson T, Smith S, Demicheli V, Harnden A, Rivetti A, Di Pietrantonj C. Assessment of the efficacy and effectiveness of influenza vaccines in healthy children: systematic review. The Lancet 2005; 365: 773-780.

King 2005

King M, Nazareth I, Lampe F, Bower P, Chandler M, Morou M, Sibbald B, Lai R. Impact of participant and physician intervention preferences on randomized trials: a systematic review. JAMA 2005; 293: 1089-1099.

Kwan 2004

Kwan J, Sandercock P. In-hospital care pathways for stroke. Cochrane Database of Systematic Reviews 2004, Issue 2. Art No: CD002924.

MacLehose 2000

MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AM. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. Health Technology Assessment 2000; 4: 1-154.

Moher 2001

Moher D, Schulz KF, Altman DG. The CONSORT Statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. The Lancet 2001; 357: 1191-1194. (Available from www.consort-statement.org).

National Health and Medical Research Council 1999

National Health and Medical Research Council. A guide to the development, implementation and evaluation of clinical practice guidelines [Endorsed 16 November 1998]. Canberra (Australia): Commonwealth of Australia, 1999.

Oxford Centre for Evidence-based Medicine 2001

Oxford Centre for Evidence-based Medicine. Levels of Evidence [May 2001]. Available from: http://www.cebm.net/index.aspx?o=1047 (accessed 1 January 2008).

Peto 1995

Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. Journal of Clinical Epidemiology 1995; 48: 23-40.

Petticrew 2001

Petticrew M. Systematic reviews from astronomy to zoology: myths and misconceptions. BMJ 2001; 322: 98-101.

Pocock 2004

Pocock SJ, Collier TJ, Dandreo KJ, de Stavola BL, Goldman MB, Kalish LA, Kasten LE, McCormack VA. Issues in the reporting of epidemiological studies: a survey of recent practice. BMJ 2004; 329: 883.

Reeves 2004

Reeves BC, Gaus W. Guidelines for reporting non-randomised studies. Forschende Komplementärmedizin und klassische Naturheilkunde 2004; 11 Suppl 1: 46-52.

Reeves 2006

Reeves BC. Parachute approach to evidence based medicine: as obvious as ABC. BMJ 2006; 333: 807-808.

Reeves 2007

Reeves BC, Langham J, Lindsay KW, Molyneux AJ, Browne JP, Copley L, Shaw D, Gholkar A, Kirkpatrick PJ. Findings of the International Subarachnoid Aneurysm Trial and the National Study of Subarachnoid Haemorrhage in context. British Journal of Neurosurgery 2007; 21: 318-23.

Rothman 1986

Rothman KJ. Modern Epidemiology. Boston (MA): Little, Brown & Company, 1986.

Shadish 2002

Shadish WR, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston (MA): Houghton Mifflin, 2002.

Siegfried 2003

Siegfried N, Muller M, Volmink J, Deeks J, Egger M, Low N, Weiss H, Walker S, Williamson P. Male circumcision for prevention of heterosexual acquisition of HIV in men. Cochrane Database of Systematic Reviews 2003, Issue 3. Art No: CD003362.

Taggart 2001

Taggart DP, D'Amico R, Altman DG. Effect of arterial revascularisation on survival: a systematic review of studies comparing bilateral and single internal mammary arteries. The Lancet 2001; 358: 870-875.

Vandenbroucke 2007

Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. PLoS Medicine 2007; 4: e297.

von Elm 2007

von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. PLoS Medicine 2007; 4: e296.

Wells 2008

Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in Meta-analyses. Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm (accessed 1 January 2008).

Wieland 2005

Wieland S, Dickersin K. Selective exposure reporting and Medline indexing limited the search sensitivity for observational studies of the adverse effects of oral contraceptives. Journal of Clinical Epidemiology 2005; 58: 560-567.

（成岚、崔晓华、肖晓娟、王小琴译，陈耀龙、田金徽、岑啸、秦天强初审）

第十四章　不良反应

作者：代表 Cochrane 不良反应方法学组的 Yoon K Loke, Deirdre Price 和 Andrew Herxheimer

本文选自工作手册 5.0.2 版本。有关如何引用它的指南，见 13.8 节。这些材料还刊登于 Higgins JPT 和 Green S 编辑的《关于干预措施的 Cochrane 系统评价手册》（书号 978-0470057964）。该手册由 John Wiley＆Sons 出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 为了实现研究的全面客观，所有系统评价作者应该考虑到干预措施可能存在不良反应
- 当证据显示潜在的不良反应严重影响临床治疗或临床决策时，一份详细的相关不良反应分析尤为重要
- 干预措施可能存在多种不良反应，系统评价作者可能需要重点详细分析几种重要不良反应，以及总结性概括其他潜在的更为宽泛的不良反应。
- 处理不良反应数据不如处理研究中的主要结局数据严谨，因此，需要仔细审查不良

反应监测强度及其报告透明度

- 不良反应的数据通常是零散的，但不良反应相关信息的缺乏并不意味干预措施的安全性好。

14.1 引言

14.1.1 研究不良反应的必要性

每一种卫生干预措施都伴随着或大或小不良反应的风险。Cochrane 系统评价仅考虑干预措施有利结果，而忽视评价其不良反应，这样的系统评价易夸大干预措施的疗效。因此，所有系统评价应该包括干预措施不良反应方面的研究，将这种偏倚最小化。

本章主要讨论 Cochrane 系统评价中与不良反应相关的问题，因不良反应的研究方法不同于一般干预结果的研究方法，故不良反应的研究方法将是本章重点讨论内容。原则上，不良反应大多是通过随机试验评估。实际上，许多不良反应，或因太罕见，或因需长期观察方能发现，或在试验计划阶段尚未知，故在随机试验期间未能观察到。Cochrane 系统评价可能用一种或多种策略来研究不良反应的问题，在某种程度上，这些策略与用于评价研究预期（有利）结果或非预期（有利或不利）结果的那些方法存在差异。本章将着重讨论非预期的不良事件（Miettinen 1983）。在 14.2 部分将讨论用于系统评价中的各种策略。

14.1.2 概念和术语

许多术语用于描述卫生干预措施相关的损害。系统评价的作者可能易混淆，尤其当已经发表的文章常随意的交替使用这些术语时。一些常见的术语包括“不良事件”（adverse event）（使用药物或其它干预措施期间及使用药物或干预措施后发生的不利结局，但不一定是由该干预措施引起的），“不良反应”(adverse effect)（干预措施和不良事件间的因果关系存在一定合理的可能性），“药物不良反应”（adverse drug reaction）（某种特定药物的不良反应），“副作用”（side effect）（在治疗剂量时发生的，治疗作用以外的，任何非预期的，有利或有害的效果）和“并发症”（complications）（术后或实施其他侵入性干预措施后发生的不良事件或效果）。

14.1.3 何时需考虑干预措施的不良反应

在系统评价中是否花费资源去纳入干预措施的不良结果时，应该考虑该干预措施本身的重要性。如果已明确该干预措施无效，或效果非常有限且并未广泛推广使用，则可能不值得花费资源去详细地评价其不良反应。另一方面，如果干预措施潜在的伤害可能是影响临床医生，患者或卫生决策者决策的关键信息，则须详细评价分析该干预措施的不良反应。

表 14.1.a 关于不良反应分析在治疗决策中重要作用的例证

表 14.1.a 详细审查干预措施不良反应的内容及举例

当利弊间差距较小时	
治疗带来的益处是很小或不确定的，但发生不良反应的可能性很大	• 在健康人群中用阿司匹林预防心血管事件，但增加了出血的风险 • 抗生素治疗儿童急性中耳炎，但存在出现皮疹和腹泻的风险 • 紧急直流电复律用于心血管稳定的病人新发房颤，但存在发生电复律中风的风险
治疗能带来潜在的高益处，但存在重要的安全问题	• 阿司匹林治疗有胃肠出血病史的中风患者 • 颈动脉内膜切除术用于治疗伴有缺血性心脏病的老年中风患者
治疗措施存在潜在的长期益处或对群体健康有益，但对个体不存在即刻直接的益处	• 提高疫苗摄入以提高人群免疫力，减低人们对早期严重神经不良反应的恐惧情绪
许多有效治疗措施在安全性方面存在差异	
治疗措施干预疗效相当，但安全性存在差异	• 患有癫痫的育龄妇女应用抗癫痫药 • 一种新的胰岛素注射装置比现有装置引起的疼痛轻微
利弊平衡存在明显差异。如，最有效的干预措施可能存在严重的不良反应，而疗效相对较差的干预措施可能更安全	• 在侵蚀性风湿性关节炎治疗中的改变病情抗风湿药，如，使用羟氯奎（相对安全）或甲氨喋呤（可能更有效，但安全性更差） • 对转移性乳腺癌进行多种药物联合化疗或单一药物序贯化疗
不良反应使患者决定终止进行有效治疗措施	
治疗措施具有很大益处，但不良反应严重影响患者依从性，需要更多证据以帮助进一步决策	• 治疗措施有效但不良反应使患者无法继续治疗。是否降低干预强度（如减少剂量或缩短疗程）有助于避免不良反应或是否存在某种治疗方法以阻止不良反应发生（如质子泵抵制剂对阿司匹林引起的消化性溃疡的预防作用），诸如上述方面的证据有必要进一步研究

14.2 系统评价中不良反应的研究范围

14.2.1 用同种方法研究有利结果和不良反应

在本节，14.2.2 及 14.2.3 节，我们将描述系统评价中可能用于研究不良反应的三种大体方法。方法一，用同种方法评价预期效果（益处）和非预期结果（损害），也就是说，采用统一的合格纳入标准（在研究类型，研究对象和干预措施方面）。

这种方法意指检索中可能使用单一检索策略。关键问题是看系统评价作者如何处理可能出现的以下三种数据集：

(a) 研究结果包括有利结果和不良反应两方面数据

(b) 研究结果仅包括有利结果数据

(c) 研究结果仅包括不良反应数据

上述(a)类研究有重要的优点，其数据来源于同一研究群体及环境，利弊可以直接比较，并且，利弊的证据分析来源于相似设计和质量的研究，但是，因其纳入的研究多为研究期限相对较短的研究，关于不良反应的数据可能非常有限，尤其可能局限于某些短期损害作用。

相比单独上述(a)类研究而言，联合上述三种研究类型评价干预措施利弊将会增加可得的信息量。比如，上述(a)和(b)类研究可用于评价有利效果，而(a)和(c)可用于评价不良反应。但是，针对不良反应的研究与针对有利结果的研究存在差异，系统评价作者应注意很难进行利弊的直接比较。

14.2.2 用不同方法研究有利结果和不良反应

第二种方法是和研究有利结果的研究相比，针对非预期结果（不良反应）的研究采用不同于针对预期结果（有利结果）研究的纳入标准。

评价不同的研究结果可能需要用不同的研究类型（Glasziou 2004）。通过大部分实验性研究（比如随机试验）无法解决罕见，长期或既往未被发现的不良反应的评价问题，此时尤其需要使用不同的纳入标准以解决该类问题（见 14.4 节）。此种方法允许对不良反应进行更严格的评价，但是需要花费更多时间和资源，同时意味着有利结果和不良反应常不能直接比较。随机试验优点是通过随机方法分配干预措施，而非随机研究则通过其他不同机制分配干预措施，作者在进行系统评价时应对此进行详细检查。

14.2.3 针对不良反应的独立系统评价

第三种方法，仅对不良反应进行单独系统评价。这可能适用于一种干预措施应用于多种疾病或临床症状而其不良反应可能在不同人群及使用环境是相似的情况。例如，阿司匹林广泛用于多种疾病患者，如卒中患者，周围血管病患者，冠状动脉疾病患者。阿司匹林针对前述不同病症的主要效果需要不同的系统评价分别分析讨论，但是其不良反应（如，脑出血或肠道出血）在不同的疾病群体都是非常相似的，可将上述相同问题集中在一个系统评价中进行研究。事实上，除非试验是基于一个组合人群，这样一个问题很难通过任何其他方式来解决。

类似地，针对某些特定亚人群（如，儿童）的干预措施，其不良反应数据可能非常有限。即便试验是针对不同的疾病状况，但对该特定群体（如，5-羟色胺再摄取抵制剂对儿童的不良反应）分析所有可得的数据仍是有价值的。

单独研究不良反应的系统评价作者必须提供充分的与干预措施相关的研究其有利结果的系统评价交叉访问信息（最好通过电子链接）。如干预措施效果的系统评价更新，确认存在新的安全问题，则关于该干预措施不良反应的系统评价也应该尽快进行相应更新。

14.3 选择纳入哪些不良反应

14.3.1 狭小关注与广泛关注

在系统评价中选择需纳入的不利结果是困难的。与干预措施相关的某些特定不良反应在进行系统评价前可能已经获知，但仍可能存在其他尚未发现的不良反应。预先可能无法确定与系统评价最相关的不良反应。根据所研究的问题，结合治疗或预防的背景，以下一般策略可能有参考意义。

狭小关注详细描述一种或两种已知的，或几种患者和医务人员特别关注的严重不良反应

优点：最容易的方法，尤其是数据收集。能重点研究几种重要的不良反应，得出对治疗决策有重大影响的有意义的结论(McIntosh 2004)。

缺点：范围可能太窄。该方法仅适用于预先已知的不良事件.

广泛关注

研究预先已知或未知的多种不良反应

优点：覆盖范围更广，可评价我们可能未知的新不良反应

缺点：工作量大，尤其在数据收集阶段困难重重。一些研究人员所发现的广泛的，非特指的评价是非常消耗资源但极少能获得有用的信息 (McIntosh 2004)。这些研究人员同时指出，过去未被认识的不良反应最好的发现方式可能是通过监测，而不是通过系统评价。

为了以更有组织的方式研究不良反应，系统评价作者可以选择将研究的范围缩小至以下一些领域：

- 前 5 到 10 种最常见不良反应
- 医生或患者认为的所有严重不良反应
- 根据类别，例如：
  - 通过实验室结果进行诊断的病症（如低钾血症）；
  - 患者报告的症状（如，疼痛）

14.3.2 退出或脱落作为不良反应的结局测量指标

退出或脱离常用于试验报告的结局测量指标。系统评价作者应审慎解释诸如安全性和耐受性这些替代性指标的数据，因为可能存在以下潜在偏倚：

- 中止实验的原因非常复杂，可能由于轻微的但却令人烦恼的副作用、毒性、缺乏疗效、非医学原因，或综合原因（Ioannidis 2004）。
- 在试验条件下，患者和研究者要保持低数量的退出或脱落的压力，，可能导致研究结果不能反映不良事件在研究人群中的真实状况。
- 在试验中未进行盲法时，更易发生研究对象退出的情况，这将导致干预措施在退出患者身上的效果被高估。例如，安慰剂对照组患者的症状不太可能导致治疗中断。相反，在积极治疗组主诉不良反应症状的患者可能更容易退出试验。

14.4 研究类型

大部分 Cochrane 系统评价关注随机试验，随机试验对疗效提供最可靠的估计。但是，在临床试验中很少能观察到罕见或长期的不良事件。一个全面彻底的调查研究可能需要纳入队列研究，病例-对照研究甚至是病案报告或病例系列。在 14.2.2 节和 14.2.3 节所

概述的方法可供选用，以致不同的研究设计类型都被纳入以研究不良反应。关于在 Cochrane 系统评价中纳入非随机试验研究（包括病例对照和队列研究）的更详细论述见 13 章（13.2 节）。关于纳入病例报道的一些问题见 14.6.3 节。

14.5 不良反应的检索方法

14.5.1 药物不良反应的信息源

除了第 6 章描述的证据来源外，为了尽可能全面地搜索药物不良反应的数据，系统评价作者可能需要要考虑核查以下数据资源：

- 药物不良反应的标准参考书，如，梅氏药物副作用（Meyler’s Side Effects of Drugs），药物副作用年鉴（SEDA），欧洲药典（马丁代尔）：完整药物参考，Davies 药物不良反应教科书及他们所概括的论文。
- 监管部门基于产品生产厂家所提交信息（这些信息可能是未公开出版或在别处无法获取）发布的安全警报。安全公告范例可通过以下渠道获取：
  - 英国：当前药物警戒问题（www.mhra.gov.uk）；
  - 澳大利亚：澳大利亚药物不良反应公告（www.tga.gov.au/adr/aadrb.htm）；
  - 欧洲药品评估机构的欧洲公共评价报告（www.emea.eu）；
  - 美国：食品药品管理局 FDA 药物监控（www.fda.gov/medwatch）。
- 专业的药物信息数据库：如全文数据库（药物新闻和爱荷华州药物信息查询台（IDIS），书目数据库（如德文特药物档案，毒理学数据库，药物学数据库）和摘要数据库（如 Drugdex，XPhram）但是，系统评价作者不得不考虑这些专业数据库的订购成本，尤其是这些数据资源在系统评价中的实用性或额外收益尚未得到正式评价

系统评价作者也能向世界卫生组织（WHO）乌普萨拉监测中心（UMC; www.who-umc.org）申请检索（通常需要付费）他们的自发（spontaneous）报告数据库（Vigibase）；这是一个关于褪黑素系统评价的例子（Herxheimer 2002）。但是，在 UMC 关于某种特定药物最常见不良反应的排序结果不同于来源于双盲随机试验 meta-分析的排序结果（Loke 2004）：在 UMC 关于胺碘酮的不良反应资料显示最常见的是甲状腺问题，其次为皮肤反应，而 Meta-分析表明心脏问题最常见，其次为甲状腺问题。

原始监测数据（以自发病例报告的形式）也可通过加拿大，美国，英国和荷兰监管部门的网络免费获得。但是，数据发布格式存在较大差异，并且解释和分析这些数据需要专业的技能（见 14.6.3 章节）

14.5.2 不良反应检索策略

确定不良反应研究的最佳检索策略有待建立（Golder 2006）。主要有两种检索方法：主题词检索和自由词检索。两种检索都有局限性。因此，可进行二者组合检索使其敏感性最大化（即，使相关研究漏检的可能性降至最小）。检索策略的最终确立可能需要将检索过程反复数次。例如，在电子检索时，可能需要将先前确定相关的研究中所描述和标引的主题词、副主题词和自由词进行重复组合检索。作者可能需要综合考虑检索的全面性（敏感度）和精确度，以决定使用哪种检索词的组合。使用主题词和自由词时应注意一些问题。

14.5.2.1 用主题词检索电子数据库中的不良反应

如 MEDLINE 的医学主题词表（MeSH）和 EMBASE 的 EMTREE 等主题词（也称受控词汇或分类词）在电子数据库里是用于描述研究特征。MEDLINE 和 EMBASE 用于不良反应的主题词较少，在 MEDLINE 里包括药物毒性（DRUG TOXICITY）和药物不良反应系统（ADVERSE DRUG REACTION SYSTEMS），在 EMBASE 里包括药物毒性（DRUG TOXICITY）和药物不良反应（ADVERSE DRUG REACTION）。然而，检索不良反应最有用的方式是运用副主题词检索（Golder 2006）。副主题词描述主题词某一特定的方面，比如，药物的“副反应”，或手术“并发症”，或他们能用于检索任意主题词的某一方面（漂浮副主题词检索）。重要数据库 MEDLINE 和 EMBASE 表示不良反应信息的副主题词不同，例如：

阿司匹林/不良反应(MEDLINE)

乙酰水杨酸/药物不良反应(EMBASE)

在上述例子中，阿司匹林（Aspirin）是 MEDLINE 里的医学主题词，不良反应（adverse effects）是副主题词；乙酰水杨酸（Acetylsalicylic-acid）是 EMBASE 主题词表系统的主题词，药物不良反应（adverse-drug-reaction）是副主题词。

在数据库中，研究可能：(1)用干预措施的名称和不良反应的副主题词一起作标引，

如阿司匹林/不良反应，或乳房切除术/并发症；或(2)不良事件本身与干预措施一起作标引，如阿司匹林和胃肠出血，或淋巴水肿和手术； 或(3)某些情况，一篇文章可能仅用不良事件作索引，如出血/化学诱导

因此，单独用索引词或副主题词不能检索出所有不良反应的数据，但是，主题词和副主题词的组合检索却能检索出标引者认为重要的主要不良反应信息(Derry 2001)。

能与特定干预措施或多种干预措施组合检索的副主题词（漂浮），并且在 MEDLINE 中证明有用的是：

/不良反应（注意，如果这个副主题词扩展检索，检索结果将会包括副主题词中毒和毒性在内的结果）

/中毒

/毒性

/禁忌症

能与不良结局或所有的结果一起检索的副主题词（漂浮），并在 MEDLINE 里被证明有用的是：

/化学诱导

/并发症

在 EMBASE 数据库中，副主题词与干预措施一起组合检索也可能是一种有效的检索：

/药物不良反应

/药物毒性

在 EMBASE 数据库中，副主题词与不良结局组合检索也可能是一种有效检索策略

/并发症

/副反应

14.5.2.2 用自由词检索电子数据库中的不良反应

自由词（也称文本词）是作者在发表的杂志文章的标题和摘要中使用的词；这些词在数据库的标题和摘要字段中能被检索到。以下两个重要问题使自由词检索严重受限：

1. 作者用于描述不良反应的词太过宽泛，无论是一般意义的描述（毒性，副反应，不良反应）还是具体某种不良反应（如，昏睡，疲倦，不适都是同义）的描述都是如此。

2. 自由词检索使得在标题和摘要中未提及的不良反应检索不到，即便整篇报道都在描述该不良反应，只要标题和摘要中未提及该不良反应，该研究便可能被漏检（Derry 2001）。

一个高敏感的自由词检索应该整合其潜在的各个同义词及其他们不同的拼写状态（如，名词的单复数形式）进行组合检索，干预措施所有相关的自由词都应组合检索，如：

（阿司匹林或乙酰水杨酸）和（副作用或不良反应或出血或失血）

14.6 不良反应偏倚风险评估

14.6.1 临床试验

尽管在第 8 章有临床试验偏倚风险评估的一般内容，但是作者还须考虑其他可能影响不良反应信息的特定因素。特别关注的方面包括监测和发现不良反应的方法，利益冲突（Jüni 2004），报告偏倚（Chan 2004）及盲法（Schulz 2002）。

在安慰剂对照盲法分配方案充分隐藏的随机对照试验中可能已经对干预措施主要结果进行评价。然而，不良反应数据的收集可能是回顾性的。如，在研究后期通过发放调查问卷给曾在试验组接受积极治疗的研究对象进行信息收集。虽然主要结果的偏倚风险可能较低，但所监测干预措施的有害效应的偏倚风险可能不在同一水平。在 RevMan 软件中推荐的偏倚风险评价工具对以下方面也予以考虑：盲法，结局缺失数据，或系统评价作者定义的结局类别

我们知道，用于监测发现不良反应的方法对不良反应的发生频率有重要影响：在进行研究时，仔细寻找将会比那些在研究时未仔细寻找的研究报告的不良反应发生数更多。比如，一群高血压患者，在自发报告的基础上监测不积极主动的话，其不良反应报告率为 16%，而用特制的问卷积极监测，则其不良反应发现率为 62%（Olsen 1999）。不同的不良反应监测方法将会产生不同的结果，因此，这样的研究很难进行比较，做 Meta-分析也没任何意义（Edwards 1999）。监测的持续时间和监测频率也应该记录。

短期随访或低监测频率的研究对不良反应监测的结果也可能不可靠；但信息的缺乏并不表示该干预措施是安全的。相反，经过对已知不良反应严格随访和积极监测的研究能产生表明干预措施有较少不良反应的真实证据。

最后，干预措施的应用年限及其使用进展可能与所监测到的不良反应的类型及数量有关。如致癌作用显然是个长期效应，但也有些干预措施，如手术，随时间推移变化几乎不大。

评价不良反应证据质量可能有用的问题，有：

关于实施：

- 对不良反应进行定义了吗？
- 报告了不良反应的监测方法吗？前瞻性或常规的监测方法；自发报告；患者检查清单，问卷或日记；患者的系统调查表？

关于报告

- 不良反应分析中排除了某些患者吗？
- 报告中提供了干预组的数值资料吗？
- 调查者报告了干预组的哪一类不良反应？

14.6.2 病例对照和队列研究

干预措施的有效性需要随机试验来证实，而干预措施不良反应常常在非随机研究中能被有效发现（Miettinen 1983）。Vandenbroucke 指出，观察性研究最有望提供关于医学干预措施不良反应无偏倚的观察性研究证据（Vandenbroucke 2004）。这个观点经比较不良反应的随机试验和观察性研究的结果得以证实。结果表明，源于观察性研究的不良反应风险估计往往较低（Papanikolaou 2006）。一些提及观察性研究得出不良反应风险更高的研究，能更好地反应患者真实状况（Vandenbroucke 2006）。像其他的研究，病例对照和队列研究潜在的更易受偏倚的影响，这些数据的局限性应该予以更严格地讨论。这类研究偏倚风险评估的进一步讨论见 13 章（13.5 章）。Jick 起草了一个最有可能发现到不良反应的研究类型分类，研究的类型也需证实（Jick 1977）。

14.6.3 病案报告

不良反应的病案报告在各种文献中广泛存在，并由监管部门核对整理。这些病案报告的评价存在特定方法学问题。对这些数据感兴趣的系统评价作者需要考虑以下问题。

这些报告有良好的预测价值吗？

轶事报告经后续调查可能被证明是个假警报，而非真正由干预措施引起的不良反应。

虽然有研究表明自 1963 年以来四分之三的该类资料都是正确的（Venning 1982），最近一个包括 63 个可疑不良反应的系统调查表明大部分不良反应（63 例中有 52 例，占 82.5%）未经详细评价（Loke 2006）。仅有三个报道，对照试验支持药物与不良事件间的假定联系，而两个对照试验无法证实这种联系。然而，产品说明书或药品的专题论文可能已经修订，并列出这些不良事件。因此，很难确定该病例报道是一个真正警报还是个假警报。即便如此，病案报道仍然是首先发现新不良反应一个基础性工作（Stricker 2004）。无论是过去还是现在，从市场撤下某药物主要是基于病例报道和病例系列（Venning 1983，Arnaiz 2001）。由于显著的效果，因此从市场撤下药物并不需要正式的对照组（Glasziou 2007）。

因果关系的决定

通常难以确定不良事件是否由特定干预措施引起（尤其当患者采取多种治疗措施的时候）。系统评价作者必须判定干预措施产生该作用的可能性大小，或者在治疗期间发生该不良事件仅仅是偶合事件。然而，两个独立的系统评价作者对同一个病案报告可能不能达成一致判断。几个研究评价了系统评价作者对不良反应报告的反应。其中一个研究，用因果关系标准评价可疑的不良反应，两个研究者意见一致率仅 35%（Lanctot 1995）。另外一个研究，三个临床药理学家，评价了 500 份可疑药物不良反应报告，在确定导致不良反应的元凶时，对其中 36%的报告中未能达成一致意见（Koch-Weser 1977）。

干预措施与不良反应间是否存在合理的生物学机制？

如果不良事件能用易理解的生物学机制进行解释，则该不良反应更具合理性。如，胺碘酮具有碘样化学结构，有助于解释该药在甲状腺机能方面的常见不良反应。

报告提供的信息是否足够详实以便进一步评价该证据？

一个基于 1520 个发表的可疑的不良反应报告的研究发现：这些报告提供的信息存在明显的差异（Kelly 2003）。关于患者特征的详细信息的报告，仅 3 个变量的报告超过了 90%，而其它 12 例变量不到 25%。在评价该可疑药物时，Kelly 发现超过 90%仅报告一个药物变量（比如药物剂量，疗程，或使用频率或确切的配方）的信息；其他 6 个变量报告在 14%至 74%之间。对系统评价作者而言，信息报告的明显差异意味着很难对这些进行详细具体评价。

使用报告中的数据是否存在任何潜在的问题，它可能高估其总体益处吗？

在纳入全部研究避免发表偏倚与不可靠信息可能导致假警报间存在一个平衡。麻腮风（MMR）疫苗接种项目曾因一篇发表在一个声誉较好杂志中的轶事报告而被迫中断，

同时由于疫苗接种的减少，人们由于麻疹暴发而受到健康危害（Asaria 2006）。不良反应额外信息的纳入（可能存在不可靠性）能产生有害的效果。系统评价的作者需要仔细考虑这些负面影响和传播这类信息的法律分歧。

14.7 本章信息

作者：Yoon K Loke，Deirdre Price 和 Andrew Herxheimer，代表 Cochrane 不良反应方法学组

本章引用格式如下：Loke YK, Price D, Herxheimer A. Chapter 14: Adverse effects. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org .

致谢：以下同仁（按字母顺序列出）为 Cochrane 不良反应方法学组提供了专业意见，并帮助发布该指南：Jeff Aronson, Anne-Marie Bagnall, Andrea Clarke, Sheena Derry, Anne Eisinga, Su Golder, Tom Jefferson, Harriet MacLehose, Heather McIntosh and Nerys Woolacott.

框 14.7.a Cochrane 不良反应方法组

不良反应方法学组（）为确定和系统评价不良反应提供方法学指导。AEMG 的起源追溯到约十年前一个关于系统评价干预措施不良反应的非正式会议。作为非随机试验研究方法学组的小分支，AEMG 形成于 2001 年 1 月。于 2007 年 6 月正式注册成立。

AEMG 的基本原则是每一种卫生干预措施存在一定的有害风险。为实现充分完全的知情决策，治疗决策需要有系统的利弊分析。那些主要关注治疗益处的系统评价往往缺乏有害效果的信息，对于需要全面信息进行综合决策的人往往造成困难。AEMG 致力于解决这种信息不对称状况，致力于与系统评价小组及方法学组合作，以提高不良反应分析的方法学和质量。AEMG 将乐于研究方法学上存在不确定性，需要进一步研究的领域，希望发展和传播已经发现的能填补这些空白的知识。

网址：aemg.cochrane.org

14.8 参考文献

Arnaiz 2001

Arnaiz JA, Carne X, Riba N, Codina C, Ribas J, Trilla A. The use of evidence in pharmacovigilance. Case reports as the reference source for drug withdrawals. European Journal of Clinical Pharmacology 2001; 57: 89-91.

Asaria 2006

Asaria P, MacMahon E. Measles in the United Kingdom: can we eradicate it by 2010? BMJ 2006; 333: 890-895.

Chan 2004

Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004; 291: 2457-2465.

Derry 2001

Derry S, Kong LY, Aronson JK. Incomplete evidence: the inadequacy of databases in tracing published adverse drug reactions in clinical trials. BMC Medical Research Methodology 2001; 1: 7.

Edwards 1999

Edwards JE, McQuay HJ, Moore RA, Collins SL. Reporting of adverse effects in clinical trials should be improved: lessons from acute postoperative pain. Journal of Pain and Symptom Management 1999; 18: 427-437.

Glasziou 2007

Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ 2007; 334: 349-351.

Glasziou 2004

Glasziou P, Vandenbroucke JP, Chalmers I. Assessing the quality of research. BMJ 2004; 328: 39-41.

Golder 2006

Golder S, McIntosh HM, Duffy S, Glanville J, Centre for Reviews and Dissemination and UK Cochrane Centre Search Filters Design Group. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. Health Information and Libraries Journal 2006; 23: 3-12.

Herxheimer 2002

Herxheimer A, Petrie KJ. Melatonin for the prevention and treatment of jet lag. Cochrane Database of Systematic Reviews 2002, Issue 2. Art No: CD001520.

Ioannidis 2004

Ioannidis JPA, Evans SJ, Gøtzsche PC, O'Neill RT, Altman DG, Schulz K, Moher D. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Annals of Internal Medicine 2004; 141: 781-788.

Jick 1977

Jick H. The discovery of drug-induced illness. New England Journal of Medicine 1977; 296: 481-485.

Jüni 2004

Jüni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative Meta-analysis. The Lancet 2004; 364: 2021-2029.

Kelly 2003

Kelly WN. The quality of published adverse drug event reports. Annals of Pharmacotherapy 2003; 37: 1774-1778.

Koch-Weser 1977

Koch-Weser J, Sellers EM, Zacest R. The ambiguity of adverse drug reactions. European Journal of Clinical Pharmacology 1977; 11: 75-78.

Lanctot 1995

Lanctot KL, Naranjo CA. Comparison of the Bayesian approach and a simple algorithm for assessment of adverse drug events. Clinical Pharmacology and Therapeutics 1995; 58: 692-698.

Loke 2004

Loke YK, Derry S, Aronson JK. A comparison of three different sources of data in assessing the frequencies of adverse reactions to amiodarone. British Journal of Clinical Pharmacology 2004; 57: 616-621.

Loke 2006

Loke YK, Price D, Derry S, Aronson JK. Case reports of suspected adverse drug reactions--systematic literature survey of follow-up. BMJ 2006; 332: 335-339.

McIntosh 2004

McIntosh HM, Woolacott NF, Bagnall AM. Assessing harmful effects in systematic reviews. BMC Medical Research Methodology 2004; 4: 19.

Miettinen 1983

Miettinen OS. The need for randomization in the study of intended effects. Statistics in Medicine 1983; 2: 267-271.

Olsen 1999

Olsen H, Klemetsrud T, Stokke HP, Tretli S, Westheim A. Adverse drug reactions in current antihypertensive therapy: a general practice survey of 2586 patients in Norway. Blood Pressure 1999; 8: 94-101.

Papanikolaou 2006

Papanikolaou PN, Christidi GD, Ioannidis JP. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. Canadian Medical Association Journal 2006; 174: 635-641.

Schulz 2002

Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. The Lancet 2002; 359: 696-700.

Stricker 2004

Stricker BH, Psaty BM. Detection, verification, and quantification of adverse drug reactions. BMJ 2004; 329: 44-47.

Vandenbroucke 2004

Vandenbroucke JP. When are observational studies as credible as randomised trials? The Lancet 2004; 363: 1728-1731.

Vandenbroucke 2006

Vandenbroucke JP. What is the best evidence for determining harms of medical treatment? Canadian Medical Association Journal 2006; 174: 645-646.

Venning 1982

Venning GR. Validity of anecdotal reports of suspected adverse drug reactions: the problem of false alarms. British Medical Journal (Clinical Research Edition) 1982; 284: 249-252.

Venning 1983

Venning GR. Identification of adverse reactions to new drugs. II (continued): How were 18 important adverse reactions discovered and with what delays? British Medical Journal (Clinical Research Edition) 1983; 286: 365-368.

（郭琴译，秦天强、岑啸初审）

第十五章　整合经济学证据

作者：代表 Campbell 和 Cochrane 经济学方法学组的 Ian Shemilt, Miranda Mugford, Sarah Byford, Michael Drummond, Eric Eisenstein, Martin Knapp, Jacqueline Mallender, David McDaid, Luke Vale 和 Damian Walker

本文选自工作手册 5.0.1版本。有关如何引用它的指南，见15.10节。这些材料还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容摘要

- 经济学是研究如何优化配置有限的资源使其产出有益社会的学科，因此它与任何卫生保健决策相关。
- 最佳决策也要求有关效果的最佳证据。
- 本章描述 Cochrane 系统评价中引入经济学观点和证据的方法，重点在严格评价卫生经济学研究。

- 在 Cochrane 评价中引入经济学观点和证据可加强其在卫生保健决策和新经济学分析中的价值和适用性。

15.1 经济学证据在Cochrane评价中的作用和联系

15.1.1 引言

Cochrane评价从多个有关卫生保健干预措施效果和其它方面的研究中收集、筛选、评价和合并可靠的数据。Cochrane评价可提供关于干预措施效果的强有力证据，生产选择性偏倚更少和更具统计效能的信息，因此Cochrane评价比单个研究的证据更容易说服决策者。

然而，面对稀缺资源，决策者不仅要经常考虑干预措施是否有效，还要考虑该干预措施能否更高效地利用资源。Cochrane评价的主题涵盖了各种问题，这些问题的解决对改善有限资源下个体和公众的健康和福利起着重要作用。Cochrane评价若包含了干预措施的经济方面内容，则能加强其作为卫生保健决策基础要素之一的价值和适用性（Lavis 2005）。

若不考虑卫生保健成本和任何健康获益的价值，推动有效的保健可导致低效使用卫生保健的公共和私人资金，这可能间接伤害了个体和公众（Williams 1987）。对这一问题的争论持续了多年。Archie Cochrane对系统评价发展影响深远（当然还包括Cochrane Collaboration），他支持决策要根据干预措施的经济学证据及效果证据。Cochrane最著名作品是根据Rock Carling系列讲座整理的书——《效果和效力》（Cochrane 1972）。框15.1.a中的两段引言摘自该书，阐述了经济学证据在卫生保健决策中的重要性。

框15.1.a Archie Cochrane关于卫生经济学的论述（Cochrane 1972）

> “资金和设备几乎都是按照高级顾问们的观点来分配的，但越来越多的是，申请额外设备必须以‘铁证’的详细论据为基础，即病人的角度和成本产生的预期获益。几乎没有人反对这样做。”（p.82）
>
> “若我们想从国家拨给NHS的经费中获得最优结果，那么我们最后必须采取以某一具体行动给人群带来的效益和成本的形式来陈述结果，并且如果钱投入越多，则获益更高。”

15.1.2 经济学和经济学评价

经济学是研究如何优化配置有限资源以得到对社会有益的产出的学科（Samuelson 2005）。资源包括人们的时间和技能、设备、房屋、能源及履行和维持一定活动（如，转诊病人的治疗及后遗症和并发症的后期处理）所需的其他投入。这里把卫生经济学研究定义为完整经济学评价研究、部分经济学评价研究和单纯效果研究（包含更少描述、测量或评估干预措施相关的资源利用的信息）。

完整经济学评价是比较分析不同备选措施的成本（资源利用）和产出（结果、效果）（Drummond 2005）。这有别于只关注成本和资源利用的经济学描述或部分经济学评价。完整经济学评价不是单一的研究方法，而是构建具体决策问题的框架。这意味着恰当类型的完整经济学评价，即数据收集分析方法，主要由决策问题或经济学问题决定，从决策者的问题和角度出发（见15.2.1）。完整经济学评价的目的在于描述、测量和评估所有相关备选措施价值（如干预措施X与对照措施Y比较）及他们的投入和产出。成本-效益分析（cost-benefit analysis，CBA）属于这类范畴。一些方法并未对所有结果赋值，但仍被视为完整经济学评价，包括成本-效果分析（cost-effectiveness analysis，CEA）和成本-效用分析（cost-utility analysis，CUA）。所有类型的完整经济学评价都使用边际分析方法。换句话说，他们目的都是测量增量资源利用、成本和/或成本-效果。框15.1.b将简短描述CEA、CUA和CBA（见Drummond的第二章（Drummond 2005））。

其它类型的卫生保健资源利用研究并未明确比较不同备选干预措施间的成本（资源利用）和产出（效果）两方面。这样的研究不是完整经济学评价，而是部分经济学评价。部分经济学评价可为理解不同干预措施的经济学方面提供有用证据。被称为部分经济学评价的卫生经济学研究包括成本分析、成本-描述研究和成本-结果描述。除完整和部分经济学评价外，随机试验和其他类型的单纯效果研究也可能包含更少描述、测量或评价干预措施相关资源利用的信息。同时这类信息并非总能构成完整或部分经济学评价方法，它仍然可能对理解干预措施的经济学的内容提供有用证据。

经济学评价既可利用，也可用于干预措施效果的系统评价。首先，系统评价可以包含经济学内容，即严格评价了发表和未发表的卫生经济学研究（见15.1.3）。其次，随着与单纯效果研究（如随机试验）同时进行（或者整合进效果研究）的完整和部分经济学评价数量增加（Maynard 2000，Neumann 2005），基于系统评价产生的效果证据开展的完整经济学评价也在增加。事实上，上述各类完整经济学评价（CEAs，CUAs，CBAs）

都能与干预效果的系统评价同时进行或者整合进去，包括采用决策分析法对干预措施成本和效果的证据进行建模或合并（Briggs 2006）。此时的经济学评价可被看作是建立在系统评价之上的更深层次的证据合成。

无论Cochrane系统评价和其他系统评价是否包含了干预措施经济方面更宽的内容，都可为随后或者同时进行的完整经济学评价模型分析提供有用的数据来源。尤其是一篇做得很好的关于聚集了效应量、不良反应和并发症的Meta分析，因其采用了随机试验的系统评价方法，可作为偏倚最小的数据来源为经济学模型提供效应量和不良反应参数（Cooper 2005）。这需系统检索恰当的数据源以便补充构建成本-效果方程或经济学模型的其他重要参数值范围（Weinstein 2003，Philips 2004，Cooper 2005）。

框15.1.b　完整经济学评价的类型

所有类型的完整经济学评价都是比较一个或多个备选措施（如干预措施X与Y的比较）的成本（资源利用）和产出（结果、效果）。它们均以相同方式对资源利用赋值（即以单位成本乘以资源利用量）。不同之处主要在于对效果的说明和赋值上。这些差异反映了不同决策问题（或经济学问题）的目的和观点不同。

成本-效果分析（cost-effectiveness analysis，CEA）：干预措施（及其对照措施）的效果用同一单位的结局指标测量（如死亡率、心肌梗死、肺功能、体重、出血量、二次感染和修复手术）。备选措施间用‘单位效果成本’进行比较。

成本-效用分析（cost-utility analysis，CUA）：当备选措施的效果在生存时间和生存质量上不同（或不同的效果指标），这些效果可用效用表示。效用的测量指标包括生存时间和主观健康感受。最熟悉的测量指标为质量调整生命年（quality-adjusted life year），或QALY.。备选措施间用获得每个单位效用的成本（如一个QALY的成本）进行比较。

成本-效益分析（cost-benefit analysis，CBA）：资源投入和备选措施效果均用货币单位表示，以利于在卫生保健系统内直接比较或者跨项目比较，或与卫生保健系统外的项目比较（如卫生保健干预措施VS 司法系统干预措施）。

15.1.3　Cochrane 评价中涵盖的经济学问题

本章总体目的在于描述Cochrane系统评价和其他系统评价的作者如何制作干预措施除效果的最佳证据外的经济学方面的最佳证据。

当前对于Cochrane评价中涵盖的经济学问题并没有正式的要求。因此Cochrane评价的作者可考虑将该指导原则作为考虑纳入经济学问题的系列备选方法之一。严格评价卫生经济学研究是其所概括的方法学体系的首要要素，它可作为Cochrane评价的完整构成要素。这包括对相关卫生经济学研究的数据进行收集、筛选、严格评价、总结和可能情况下合成。该指导原则的三个核心前提如下：

1. 对于Cochrane评价的国际终端用户来说，系统评价的经济学内容总体目的在于总结不同环境下干预措施的经济学方面，帮助他们理解如何在不同备选治疗或检验措施之间进行关键的经济学取舍。
2. 第二个重要目的在于为Cochrane系统评价提供一个框架，规范临床和经济学数据，以利于后续的或同步的经济学分析。
3. 即使干预措施效果的证据不明朗，经济学问题仍然与决策相关。首先，终端用户通常需要关注干预措施与对照措施相比相关资源利用和成本增量的证据，这有助于弄清未来是否要投入效果和成本-效果研究。其次，对终端用户来说，弄清现有的完整经济学评价是否是基于稳定的效果证据很重要。

Cochrane评价的作者需要从制定计划书的最早期阶段，详细考虑干预措施涵盖的经济学方面及经济学问题怎样与评价主题相关联。使用本章描述的方法也要求至少经过卫生经济学方法的一些培训。因此，一旦决定纳入经济学问题，最好尽早咨询有系统评价方法学经验的卫生经济学家。

一些Cochrane评价小组（Cochrane Review Groups，CRGs）已经拥可以找到一个或多个经验丰富的卫生经济学家，他们定期参加系统评价中经济学问题方面的工作。Campbell和Cochrane经济学方法学组（Cochrane Economics Methods Group，CCEMG）将尝试着帮助Cochrane评价作者联系愿意参加，或提供建议或同行评议支持的卫生经济学家（见框15.10.a）。

15.2 计划Cochrane评价中经济学构成要素

15.2.1 构建经济学问题

决定在Cochrane评价纳入涵盖经济学方面的干预措施后，研究的第一阶段是构建一个或多个问题，或目的，即该系统评价的经济学内容拟解决的问题。每一个经济学问题

或目的都将决定采用何种方法进行后续阶段的卫生经济学研究的严格评价。

构建经济学问题要求详细考虑经济学问题对所评价特定主题的作用和联系。为了帮助作者和编辑明确这一概念，我们准备以下初步问题作为开始。

- 干预措施试图解决的疾病或问题对社会(如卫生系统、卫生或社会服务提供者、个体、家庭、雇主）的经济负担是什么？
- 与对照措施比较，应投入什么类型的增量资源可以实施和维持该干预措施？（如员工、设备、药物、住院病人护理）？
- 与对照措施比较，实施该干预措施投入到增量资源产出是什么？或该干预措施对后续资源利用会产生哪些影响(如并发症、再次手术、门诊病人复诊、误工)？
- 干预措施与对照措施比较，产生的与资源利用变化相关的增量成本是什么？（如直接和间接医疗成本、病人自费费用、雇佣护理人员费用）
- 与对照措施比较，与该干预措施引起的收益或损害效应增量相关的经济价值是什么？（如意愿支付或效用的测量）
- 采纳或拒绝某一干预措施时，需要考虑如何取舍成本（资源利用）和获益或不良反应？

除了上述问题，同样需要考虑下面的关键问题：

- 重要性：当干预措施和对照措施比较，与之相关的不同增量资源利用或增量成本条目的重要性顺序是什么？换句话说，要在不同备选措施中选择时，哪些条目的资源利用（资源投入和资源产出）和成本可能最重要？
- 时间跨度：可能累计的重要成本（资源利用）和效果（结果）的时间跨度是什么？Cochrane评价通过确定中间指标和终点指标为目标结局指标，明确地建立了一个效果的时间跨度。当考虑所有相关成本（资源利用）和效果时，也要同时考虑这个相同的时间跨度是否适用。
- 分析角度：当干预措施和对照措施比较时，相关增量成本的可能承担者是谁，增量获益的可能接受者是谁（如患者、患者家庭、医疗服务提供者或第三方支付、卫生保健系统、社会）？某些成本（资源利用）可能与某个分析角度相关，但与其他分析角度可能无关。例如：非正式护理的成本可能与患者或社会角度有关，而选择更窄的分析角度，如卫生保健系统角度，则可能排除这类成本。更复杂的情况是不同的分析角度间一些资源利用或者成本分类可能有交叠。考

虑到Cochrane评价终端用户的范围时，务实的方法是考虑所有分析角度，然后不仅报告资源利用和成本的测量，同时报告成本的承担者或资源利用的引发者。

临床事件路径可提供更加有用的工具帮助明确经济学问题对所评价特定主题的作用和联系。临床事件途径用系统、清晰的方法描述了不同医疗和社会保健的过程和结果。该方法从介绍干预措施开始，通过病人管理的后续变化，直到最终结果，描述了相关的具有独特资源意义或结果价值的主要事件途径。（见Donaldson第二章（Donaldson 2002））。表15.2.a举例说明临床事件“中风”的临床事件途径。制定临床事件途径时，要着重考虑重要性、时间跨度和分析角度这些关键问题。

一旦仔细考虑了经济问题的作用和相关性后，就能构建起一个或多个经济问题或目的。系统评价作者应避免询问这类经济学问题“干预措施X(与Y或Z比较)的成本-效果是什么？”，因为卫生经济学研究的严格评价对这类用于不同临床情况的问题不可能提供可信的答案。经济学问题或目的应在系统评价计划书的目的部分与其他研究问题和目的一起清晰地陈述。

考虑经济学问题的作用和相关性也能用于评述干预措施的经济学方面内容，并在系统评价的背景部分提出。

无论作者是否打算将卫生经济学研究的严格评价整合进系统评价，都可以被纳入‘经济学评述’。这有助于系统评价的终端用户通过考虑干预措施潜在的经济后果，以确定在经济学背景下进行研究的干预措施。‘经济学评述’也许强调了疾病的经济负担或实施干预措施的医疗条件、实施和维持干预措施所需的资源类型（资源投入），以及干预措施对后续资源利用的潜在影响（资源产出）和成本-效果问题。评论应有相应的参考文献、严格评述和相关文献支持。框15.2.a从当前Cochrane系统评价背景中摘录了这类评论的一些例子。

图15.2.a　临床事件路径

事件路径	举例
临床事件	中风
↓	↓
临床事件处理＋后续临床事件	急性期治疗和康复＋后遗症和并发症的治疗
↓	↓
用于处理事件的资源以及事件的结果	住院时间长短、康复治疗强度、后遗症和并发症的处理（如，出血的二级预防）及疾病每一阶段相关的健康结果
↓	↓
资源利用的成本和结果的效用	采用医疗（或其他）费用和价格对资源利用赋值，及结果赋值，如采用质量调整生命年（QALYs）或意愿支付（WTP）

框15.2.a　背景中强调干预措施经济方面内容的评述

“大便失禁…在医学、社会和经济上都是一个很大的问题…美国每年花费4亿多美元用于尿失禁和大便失禁相关产品…1991年英国整个医疗机构和长期护理机构用于失禁的垫子、用具和其他处方药的直接成本约6800万英镑…随着全球老年人口的增长，其医疗保健服务和家庭护理将同样面临着不断增长的挑战。”（Brown 2007）

“若一个新的且相对昂贵的治疗【拉莫三嗪】能被常规应用，必须清楚的知道其与标准抗癫痫药（AED）如卡马西平的比较疗效。英国西北部的癫痫服务调查突出显示了潜在的成本意义，该调查提到近40%的药物成本（占癫痫治疗所有直接成本构成因素中最大比例）用于购买新药拉莫三嗪和氨己烯酸，虽然仅有7%的病人服用这类药物。”（Gamble 2006）

“晚期结肠直肠癌的姑息化疗成本不仅包括化疗成本，还包括化疗相关并发症处理的成本。若姑息化疗能够改善症状和生活质量，那么这将减少病人对其他症状/支持疗法的依赖性和需要，以抵消姑息治疗的成本。另一方面，若化疗相关的毒性高，且降低生活质量，则姑息治疗的成本将比单独的支持治疗更高。(Best 2000)

15.2.2　纳入资源利用、成本和成本-效果指标作为结局指标

形成经济学问题的过程也有助于在系统评价中确定把资源利用、成本或成本效果（或全部）指标作为目标结局指标纳入。这些结果应该包括系统评价中“纳入研究标准”部分的“结局指标类型”中。尽可能地将资源利用和成本分解为具体事项或具体类别（如，住院天数、手术时间、门诊人数、二级预防六个月随访的出血量、误工天数、直接医疗资源利用、直接医疗成本、间接资源利用或间接医疗成本、患者自付费用），避免使用

一般的描述性术语作为测量结果（如，‘成本’、‘资源利用’、‘卫生经济学’）。成本效果指标可作为目标结果纳入系统评价中，这包括增量成本-效果比（ICERs），每获得1个QALY的增量成本和成本-效益比（见15.1.2部分）。

15.2.3 卫生经济学研究的具体类型和系统评价中经济学内容的范围

卫生经济学研究的严格系统评价首先应该考虑纳入哪些具体类型的研究（见15.1.2部分）。这由已经构建的经济学问题或目的及作为目标结局指标的资源利用、成本和成本-效果指标决定。

纳入哪些类型的研究还应咨询卫生经济学专家，因为并不需要在不同形式的经济学问题、‘经济’结果指标和不同类型的卫生经济学研究之间绘制分析路径。如，一项成本-效果分析纳入了所有中间阶段分析结果和最终结果，它可能提取到与资源利用、成本和成本-效果指标相关的数据；然而若仅仅报告了最终结果，它仅可能提取到与成本-效果指标相关结果数据。

系统评价中纳入的卫生经济学研究类型应该在“系统评价纳入标准”部分的“研究类型”中陈述。下面陈述了所有类型经济学研究的特征：

研究类型

应考虑在卫生经济学研究的严格评价中纳入以下研究类型：

［干预措施 VS 对照措施］的完整经济学评价研究（即，成本-效果分析、成本-效用分析、成本-效益分析）；［干预措施 VS 对照措施］的部分经济学评估研究（即，成本分析、成本-描述研究、成本-结果描述）；报告信息更有限的随机试验，如缺乏与［干预措施 VS 对照措施］相关的资源利用或成本估计值。

当拟做卫生经济学研究的严格评价时，关键的方法学决策是制定系统评价中卫生经济研究的范围。这至少包括三个方案：

1. 仅考虑与效果研究同时进行的相关卫生经济学研究，其效果研究符合效果系统评价的纳入标准；
2. 考虑与效果研究同时进行的和采用效果研究数据的相关卫生经济学研究，其效果研究符合效果系统评价的纳入标准；
3. 考虑所有相关的卫生经济学研究，无论是否与效果研究同时进行或是否基于效果研究的数据，只要这些效果研究符合效果系统评价的纳入标准。

第一个方案一般仅考虑纳入以高质量随机试验同时进行的卫生经济学研究。第二个方案还纳入了以高质量随机试验数据为基础的Meta分析的经济模型研究。这类经济学评价的系统评价一个较好的例子是Campbell和他的同事所做的关于腹主动脉瘤筛查的系统评价（Campbell 2007）。第三个方案更广泛，纳入了所有相关的卫生经济学研究，包括如基于观察性研究或大型管理数据库的经济学分析，或回归为基础的成本和资源利用分析。

纳入这么多不同类型的卫生经济学研究对严格系统评价结果的影响如何目前还知之甚少。然而，考虑经济学研究“范围”对结果至少有潜在影响还说的过去，因为不同方案也许涉及不同类型的研究（见15.5.2部分）。另外，如果系统评价纳入以单个研究（如，随机试验）为基础的经济学评价和以模型为基础的经济学评价，最佳选择是分别考虑不同类型的研究，以保持研究之间的可比性。

在实践中，目前大部分打算纳入卫生经济学研究证据的Cochrane系统评价都限制为与效果研究同时进行的经济学研究，且效果研究符合系统评价效果部分的纳入标准（即，第一选择），但并未清晰描述出来（Shemilt 2007）。因为关于经济学研究范围的决策潜在地排除了一些没有经过严格方法学质量评价的卫生经济学研究，所以这个决策的结果应该在“系统评价的纳入排除标准”中“研究类型”的部分陈述，同时还有拟纳入的经济学研究类型的细节，如通过补充“本系统评价将只考虑纳入与系统评价纳入的效果研究同时进行的卫生经济学研究”作为以上的解释说明。

15.3 查找研究

15.3.1 电子检索过滤器的使用

相关卫生经济学研究的检索方法取决于考虑纳入的研究类型及对其严格评价的范围（见15.2.2和15.1.2）。然而，所有研究检索策略第一阶段的目的均相同：初筛检索到效果研究，以及确定其中能够潜在纳入包含相关卫生经济学研究的Cochrane系统评价的研究。

使用用于获取卫生经济学研究的检索策略可以滤过从电子文献数据库检索到的效果研究的电子记录。这在摘要和全文初筛之前进行，目的在于通过限制拟评估记录的数量来帮助找到经济学研究。电子检索过滤器对于电子文献数据库检索后文献量很大的系

统评价最有用（即，文献量相对较少，可能就不必使用电子过滤器，但仍需要应用明确的标准）。

评价和传播中心（The Centre for Review and Dissemination，CRD）已经制定出了一系列电子检索策略以获取经济学研究，包括英国国家卫生服务部（National Health Service, NHS）经济学评价数据库（NHS Economic Evaluation Database，NHS EED）。MEDLINE（Ovid CD-ROM），CINAHL（Ovid CD-ROM), EMBASE（Ovid online）和PsychINFO（Ovid online）的版本已发表在NHS EED手册（Craig 2007）中以及www.york.ac.uk/inst/crd/nfaq2.htm. 这些检索策略都可使用“AND”运算符添加到具体的系统评价用于检索相应数据库，过滤包括“economics”条目的检索结果记录。

这些NHS EED检索策略是非常宽泛的，可检索到经济学的方法研究和经济学研究的系统评价，还有所有类型的卫生经济学研究（见15.1.2）。对于更多的具体检索，建议采用缩窄检索策略和MeSH词的范围。这些检索策略也可在信息检索专家的指导下调整后检索其他电子文献数据库。

调整检索策略时需考虑不同数据库之间卫生经济学研究的不同索引或分类方法。目前可获得一个实用的电子文献数据库注释列表，它涵盖了卫生经济学文献和相关灰色文献的网站（Napper 2005）。

在使用检索卫生经济学研究的电子检索过滤器时，一个重要的程序是Cochrane系统评价也经常使用其他的检索过滤器来检索其他特定设计的研究，如随机试验。这些“研究设计检索过滤器”也可通过“AND”运算符来组合到具体系统评价的检索策略中。因此，若严格评价的范围没有限制为仅纳入与效果研究同时进行的卫生经济学研究（如，也包括模型为基础的经济学评价：见15.2.3），则‘经济学检索过滤器’应使用‘OR’运算符来组合其他任何‘研究设计过滤器’，以确保可以检索到所有类型的卫生经济学研究。此外，若严格评价的范围限制为仅纳入与效果研究同时进行的卫生经济学研究，且效果研究为该系统评价纳入的效果部分，则不必使用‘经济学检索过滤器’，因为大部分经济学研究都可以通过使用‘研究设计检索过滤器’检索到（虽然这种情况检索结果可能仍然遗漏了一些相关的经济学研究，如以随机试验为基础的经济学研究，但这些研究经常与效果研究独立发表或者在其后发表）。

15.3.2 专题数据库的使用

NHS EED是作为Cochrane Library（www.thecochranelibrary.com）的一部分来发行的。因此，无论用户何时检索Cochrane Library，NHS EED的记录都像Cochrane系统评价一样突出显示出来。NHS EED也可免费在线检索，从CRD网站登陆（见www.york.ac.uk/inst/crd/crddatabases.htm）。Cochrane图书馆中NHS EED版本每个季度定期更新，CRD网站版本每个月更新。

推荐所有Cochrane系统评价，尤其是那些包含了卫生经济学研究严格评价的系统评价都检索NHS EED，并使用其检索结果。NHS EED包含所有语言出版的卫生保健领域内完整经济学评价的结构式摘要，和部分经济学评价、方法学研究和经济研究综述的书目记录。NHS EED结构式摘要格式包括同行独立评价的卫生经济学家撰写的严格评论，并且以概要形式提供了方法学、结果及其它数据的详细信息，这对卫生经济学研究系统评价的严格评价和数据收集均有用（见15.5.2和15.4.2）。

某些时候在已发表的Cochrane系统评价中纳入相关完整经济学评价研究的NHS EED摘要作为补充是有用的，如，Rodgers等和Fayter等发表的（Rodgers 2006, Fayter 2007）（见15.6.2）。Cochrane系统评价在检索时若未在NHS EED找到完整经济学评价的结构式摘要，那么系统评价作者可以通报Campbell和Cochrane经济方法学组（Cochrane Economics Methods Group）（15.10.a框）使得NHS EED研究者意识到需要制作相关结构式摘要。

除了‘经济学检索过滤器’和其他的‘研究设计过滤器’，可通过调整具体系统评价的检索策略来检索NHS EED和其他专题数据库的卫生经济学文献（见下）。检索Cochrane图书馆时，也默认检索NHS EED（除非采用高级检索时排除了该数据库）。如何检索CRD网站中的NHS EED版本，可在www.crd.york.ac.uk/crdweb/html/help.htm的CRD帮助界面找到相关信息。

将英国NHS EED数据库的原则扩展到其他欧洲国家的愿望已促进建立了欧洲卫生经济评价数据库网络（European Network of Health Economic Evaluation Databases，EURONHEED），该数据也是在线免费的（见http://infodoc.inserm.fr/euronheed/）。NHS EED只提供了EURONHEED完整摘要的链接（截止到2000），因此虽然检索NHS EED将检索到两个数据库的所有完整经济学评价摘要，但它无法检索到只在EURONHEED中有的部分经济学评价、方法学研究或经济学系统评价的书目记录。

NHS EED项目组在一篇文章中详细描述了Cochrane系统评价如何在NHS EED、EURONHEED和其他专题数据库（包括CEA 注册库、卫生经济评价数据库HEED和Econlit）中检索卫生经济学文献（Aguiar-lbanez 2005）。CRD也在www.york.ac.uk/inst/crd/econ4.htm网站发行了在线注释列表，包含了这些数据库的详细信息和每个数据库的网站链接，作为‘卫生经济学信息资源’（www.york.ac.uk/inst/crd/econ.htm）页面下的一部分。这个注释列表也包括了卫生经济文献常用数据库的详细信息（见15.3.1）。

若卫生经济学研究严格系统评价的范围限制在与系统评价效果研究同时进行的卫生经济学研究上（见15.2.3），则补充检索NHS EED和其他专业数据库，目的在于核查这些数据库是否包括了与纳入的效果研究同时进行的完整经济学评价的所有结构式摘要。然而，若卫生经济学研究的严格评价范围更广泛（见15.2.3），则需要进一步查找更多符合该系统评价纳入标准的经济学研究。

15.4 筛选研究和收集数据

15.4.1 评价与研究主题的相关性

一旦初步纳入的卫生经济学研究全文（和完整经济学评价的结构式摘要，可用的情况下），则下一步就是评价这些研究与特定系统评价主题的相关性，这是评价偏倚风险的预备阶段。决定是否纳入或排除相关卫生经济学研究，应看它们是否满足该系统评价方案中的所确定的目标研究人群、干预措施、对照措施和结局的标准。应在‘排除研究的特征’表格中报告排除卫生经济学研究的原因。

15.4.2 数据收集

Cochrane系统评价中经济学内容的准确数据收集需根据每个系统评价具体化，这取决于特定的经济问题或目的和目标结局指标，如增量资源利用、成本或成本效果。一般来说。需要收集两种类型的数据：纳入的卫生经济学研究特征和结果的详细信息。从发表的报告中提取数据可能受限于卫生经济学研究的报告质量（若信息缺失，可进一步联系作者获取额外的详细信息）。

收集每个经济学研究特征的有用数据可能包括：研究年代；干预措施和对照措施的详细信息；研究设计和资源利用的来源、单位成本和效果数据（见15.1.2和15.2.3）；决策权限、地理位置和组织机构情况；分析角度和成本和效果的时间范围（见15.2.1）。

对于结果，应该分别提取与干预措施和对照措施相关资源利用的具体条目及其单位成本的估计值，如果还报告了资源利用成本信息，也应提取（即资源利用量X单位成本）。每个资源利用的类型和数量也应该根据自然单位提取（如，住院天数，手术时间，六个月随访的门诊病人数，工作天数）。收集价格年和货币类型也很重要，用以估计成本和增量成本。只要可能，也应该收集个体患者水平上增量资源利用和成本的测量（即，每个患者的资源利用，每个病人的成本）。若报道了增量资源利用、成本和成本效果，也应该收集点估计值和不确定测量（如，标准误或可信区间）。另外，收集所有敏感性分析数据和任何不同假设对结果大小和方向影响的信息也有用。

CRD报告6（Craig 2007）包括一个NHS EED完整经济学评价的结构式摘要模板（见15.3.2）和指导数据收集和严格评价的注释。这些材料可为Cochrane系统评价经济学内容设计数据收集表提供模板。

若一个完整经济学评价已有相应的NHS EED结构式摘要，研究者也许可不用在研究中进一步收集相关数据。同时，只要是Cochrane系统评价需要对没有完整NHS EED摘要的经济学研究进行严格评价和数据收集，鼓励Cochrane系统评价作者注册NHS EED，生产摘要，以便避免重复。请联系CCEMG获得更多信息或向NHS EED要求注册生产结构式摘要（见15.3.2）。

15.5 偏倚风险的处理

15.5.1 按研究设计进行研究分类

在评价偏倚风险之前需将纳入的卫生经济学研究按研究设计分类。卫生经济学研究方法学质量的严格评价随研究设计有轻微变化。

分类包括下面两个阶段：

1. 卫生经济学研究设计的分类。
2. 生产效果数据的研究设计的分类，如果可行，卫生经济学研究基于该效果数据。

每一个卫生经济学研究可以按完整经济学评价、部分经济学评价或报告资源利用或与干预措施相关成本的更有限信息的效果研究（如，随机试验）进行分类（阶段1）（见15.1.2）。生产效果数据的研究设计分类（阶段2），卫生经济学研究基于该效果数据，仅适用于阶段1中分类为完整经济学评价或成本-结果描述性研究的卫生经济研究。生产效果数据的研究也许是单个研究设计（如，随机试验，非随机试验，观察性研究）或几个研究的整合（如，随机试验的Meta分析）（见15.1.2）。

进行卫生经济学研究分类时，咨询卫生经济专家可能有用。这是因为报告的是使用一种研究设计（如，成本-效益分析）的卫生经济学研究，更仔细观察会也许发现使用的是另外一种设计（如，成本-效果分析）。这意味着系统评价时，经济学研究分类需要具体分析（Zarnke 1997）。

根据卫生经济学研究的严格评价范围和拟考虑纳入的研究类型（见15.2.3），也许在这一阶段会基于研究设计分类而排除某些卫生经济学研究。这一阶段排除的原因同样应该在‘排除研究特征’表中报告。

15.5.2 方法学质量的严格评价

下一步是对剩余的卫生经济学研究进行方法学质量的严格评价，以便解决偏倚风险。已有文章很好地描述了卫生经济学分析实施和报告质量的差异（Neumann 2005）。严格评价卫生经济学研究的核心目的在于评价这些研究是否以透明的方式来描述方法、假设、模型和可能的偏倚以及是否有证据充分支持，是否任何一个认真的读者都容易获得该证据的强度（Rennie 2000）。

可采用已制定出的方法学质量评价量表来严格评价卫生经济学研究。只要使用了改量表/清单对Cochrane系统评价中卫生经济学研究进行了严格评价，就应在Cochrane系统评价中的‘数据收集和分析’部分描述用于严格评价卫生经济学研究量表的详细书目信息。无论使用哪种量表，在发表的系统评价中采用额外的表格来总结纳入卫生经济学研究的完整质量评价量表结果是有用的。

完整经济学评价的可靠性在一定程度上可通过使用可信的效果数据来部分预测，所以与单个效果研究（如，随机试验）同时进行的完整经济学研究的部分严格评价涉及到可能在效果研究中会出现的所有潜在偏倚来源（见第8章）。对于这种类型的完整经济学评价研究，严格评价由下面两个部分组成。

1. 若完整经济学评价研究基于单个效果研究进行，则评价完整经济学评价研究偏倚风险，可通过广泛认可的效果研究评价量表进行。
2. 评价完整经济学评价研究的方法学质量，可通过广泛认可的针对与单个研究设计同时进行的经济学评价研究的量表进行。

已制定了大量的量表来指导卫生经济学研究的严格评价。尽管尚无正式被验证的量表，但有两个量表比大多数量表接受了更多的审查：

- 英国医学杂志（BMJ）用于指导作者投稿和同行评议专家的量表（Drummond 1996）；
- 针对经济学评价方法学质量评价的CHEC清单（Evers 2005）。

图15.5.a和15.5.b再次列出了这些量表。Cochrane 系统评价推荐使用‘Drummond量表’和‘Evers量表’来严格评价与单个效果研究同时进行的完整经济学评价的方法学质量，并使用量表适合条目的组合来严格评价部分经济学评价（见15.1.2）。

若卫生经济学研究严格评价的范围包含相关经济学模型研究（见15.2.3），则需要不同的量表来评价这些研究的方法学质量，因为‘Drummond量表’和‘Evers量表’与之相关但不足够。推荐使用‘Phillips量表’来严格评价经济学模型研究的方法学质量（Philips 2004）。可采用已经发表的数据来源的分级作为这一量表的补充，该分级系统被认为是获得经济学模型中相关参数的最佳数据来源（Cooper 2005）。

可通过相应的NHS EED结构式摘要来严格评价所有类型的完整经济学评价的方法学质量，若可得话，来补充量表（见15.3.2）。因为NHS EED结构式摘要纳入的质量评价与以上推荐量表所反映的质量维度相同。

尚无经过广泛验证的最低方法学质量标准用于筛选纳入系统评价中的经济学研究。因此决定纳入或排除研究需要做出全面判断，包括研究的方法学质量、与该经济学问题的相关性、干预措施、研究人群和研究结果（见15.4.1）。应在‘数据收集和分析’部分陈述与卫生经济学研究的方法学质量合格的标准。

需要注意的是，迄今，只有相对很少的经验性研究调查纳入满足部分而非全部方法学标准的经济学研究对卫生经济学研究严格评价结果的影响。然而，与选择与效果研究质量和设计及与卫生经济学研究设计相关的纳入标准一样（见15.2.3），资源利用、成本和/或成本-效果指标使用不同的数据来源对结果至少有潜在影响也是合理的（见15.2.3）。

图15.5.a　Drummond 量表（Drummond 1996）

条目	是	否	不清楚	不恰当
研究设计				
1. 是否陈述了研究问题	□	□	□	
2. 是否陈述了研究问题的经济学重要性	□	□	□	
3. 是否清晰陈述了分析的角度及其合理性	□	□	□	
4. 是否陈述了选择替代项目或对照干预措施的合理性	□	□	□	
5. 是否清晰陈述了替代的对照措施	□	□	□	
6. 是否陈述了所使用的经济学评价形式	□	□	□	
7. 是否选择了正确的经济学评价形式以解决相关问题	□	□	□	
数据收集				
8. 是否陈述了所评估效果的来源	□	□	□	
9. 是否给出了效果研究设计和结果的详细信息（若基于单个研究）	□	□	□	
10. 是否给出了合成效果估计值的的方法或 meta 分析的详细信息（若基于多个效果研究的证据合成）	□	□	□	
11. 是否清晰陈述了经济学研究的主要结局指标	□	□	□	
12. 是否陈述了效益赋值的方法	□	□	□	
13. 是否给出了评估对象的详细信息	□	□	□	
14. 是否分别报道了生产力变化（若纳入）	□	□	□	
15. 是否讨论了生产力变化与研究问题的相关性	□	□	□	
16. 是否分别报告了单位成本及资源利用数量	□	□	□	
17. 是否描述了单位成本和数量的评估方法	□	□	□	
18. 是否记录了当前货币和价格数据	□	□	□	
19. 是否给出了因通货膨胀所致货币价格调整或货币换算的详细信息	□	□	□	
20. 是否给出了任何模型的详细信息	□	□	□	
21.模型选择及其关键参数是否恰当	□	□	□	
结果分析和解释				
22. 是否陈述成本和获益的时间界限	□	□	□	
23. 是否陈述了折现率	□	□	□	
24. 折现率的选择是否恰当	□	□	□	
25. 若成本和效益没有折现，是否给出了解释	□	□	□	
26. 是否给出随机数据的统计学检验和可信区间的详细信息	□	□	□	
27. 是否给出敏感性分析方法	□	□	□	
28. 敏感性分析变量的选择是否恰当	□	□	□	
29. 变量变化范围是否恰当	□	□	□	

30. 是否比较了相关替代措施	□	□	□
31. 是否报道增量分析	□	□	□
32. 主要结局指标是以总数的形式还是非总数的形式呈现	□	□	□
33. 是否回答了研究问题	□	□	□
34. 是否报道了数据得出的结论	□	□	□
35. 是否与结论信息同时提供了恰当的防止错误理解的说明	□	□	□

图15.5.b　Evers量表（Evers 2005）

条目	是	否
1. 是否清晰描述了研究人群？	□	□
2. 是否清晰描述了替代措施？	□	□
3. 是否以可回答的形式准确定义了研究问题？	□	□
4. 经济学研究设计对于研究目的是否恰当？	□	□
5. 选择的时间界限对纳入相关成本和结果是否恰当？	□	□
6. 实际选择的角度是否恰当？	□	□
7. 是否确定了每个替代措施的所有重要相关成本？	□	□
8. 是否所有成本都以恰当的物理单位进行了测量？	□	□
9. 成本的赋值是否恰当？	□	□
10. 是否确定了每个替代措施的所有重要相关结局？	□	□
11. 是否恰当测量了所有结局指标？	□	□
12. 结局指标的评估是否恰当？	□	□
13. 是否采用了替代措施的成本和结局的增量分析?	□	□
14. 所有未来成本和结局的折现是否恰当?	□	□
15. 是否敏感性分析包含了所有重要的、不确定的变量？	□	□
16. 是否报道了数据得出的结论？	□	□
17. 是否讨论了研究结果在其他背景和人群/客户团体中的可推广性？	□	□
18. 文章是否声明研究者和资助者没有潜在的利益冲突？	□	□
19. 是否恰当讨论了伦理和分配问题？	□	□

15.6 结果分析和描述

指导Cochrane系统评价中经济学内容分析方法的重点在纳入卫生经济学研究的特征和结果的表格上。还可通过对纳入研究的严格评价及其主要结果讨论对描述性总结来补充。另外，某些情况下，也许要考虑资源利用或成本数据的Meta分析或制定经济学模型。以下章节将详细阐述这些方法。需要更进一步的方法学研究来评价分析卫生经济学研究及其结果描述的更多方法（见15.9）。

15.6.1 以表格形式描述结果

系统评价中‘纳入研究特征表’描述纳入的卫生经济学研究特征的详细信息，如研究年代；干预措施和对照措施的详细信息；研究设计；数据来源；研究权限和背景；分析角度和时间界限（见15.4.2）。作者也可考虑了额用另外的表格来总结评价纳入卫生经济学研究方法学质量的量表结果（见15.5.2）。

可使用‘纳入研究特征’表、其他表格或两者都可来总结纳入的卫生经济学研究的结果。任一种情况下，只要可能，应报告目标干预措施和每个对照措施比较，每项资源利用或者成本和增量成本和/或成本-效果指标的点估计值及其不确定性。陈述与成本和/或增量成本估计值相关的当年货币类型和价格年也很重要。

可能将成本估计值换算为通用货币和价格年，以利于不同研究估计值的比较。以购买力平价（Purchasing Power Parities，PPPs）为基础的国际汇率可用于将成本估计值换算为目标货币，用国民生产总值（gross domestic product，GDP）平减指数（或GDP物价平减指数）将成本估计值转换为固定价格年。包含PPP换算率和GDP平减指数的数据集可从国际货币基金组织（International Monetary Fund）的世界经济展望数据库（World Economic Outlook Database）获得（半年更新一次：登陆www.imf.org/external/data.htm）。成本估计值换算为通用货币和价格年应咨询经验丰富的卫生经济学家后进行。CCEMG将在适当的时候将发布解决这一主题的方法学指导。

15.6.2 结果的描述性总结

Cochrane系统评价也许包含了纳入的经济学研究主要的特征和结果的描述性总结，包括增量资源利用、成本和成本-效果，并用表格形式补充和提供评论。可与效果研究

结果的描述性总结一起放到结果部分（见11章，11.7）。

描述性总结的核心目的在于为终端用户弄清来源于多个研究间的成本和资源利用估计值的同质性程度。可通过描述评估方法及比较组间、纳入的研究之间和内部的资源利用和成本模式的差异来解释研究间结果的不一致性。正如本章前面讨论的，经济学评价研究以不同方式、为不同目的构建（见15.1.2）。这也许是导致研究之间方法和结果异质性的因素之一。经济学研究间方法和结果的异质性及对统计学上异质性可能来源的关注，有助于以简明扼要的方式总结国际经济学文献，而这种简要的方式可能对系统评价的终端用户有用（Gibody 1999）。避免将这部分当作关于成本-效果推荐的分析形式很重要（见15.8）。

纳入的卫生经济学研究的描述性总结的其他重要特征包括：

- 按研究设计报告系统评价纳入卫生经济学研究的总数量；
- 列出纳入研究想要解决的经济学问题；
- 报告纳入研究的设计类型；
- 报告纳入研究的分析角度；
- 报告纳入研究的时间跨度；
- 纳入研究报告的增量资源利用、成本和/或成本-效果指标的讨论；
- 报告纳入研究的增量资源利用、成本和/或成本-效果的不确定性指标
- 报告纳入研究成本估计值的货币类型和价格年；
- 如果可能，将每个研究报告的成本估计值折算为通用货币和价格年；
- 强调纳入研究间和敏感性分析中的敏感性分析和结果一致性的关键特征；
- 讨论纳入研究的总体方法学质量和局限性；
- 讨论纳入研究对其他地区和背景的相关性和适用性；和
- 讨论纳入卫生经济学研究的效果数据的质量，经济学研究中使用的结果和Cochrane系统评价中效果估计值之间的关系。

进一步的选择是提供完整的NHS EED的链接或完整经济学评价研究的其它结构式摘要。如果可用的话，NHS EED结构式摘要包括完整卫生经济评价的特征和结果的信息（见15.3.2）。一些系统评价在附件部分包含了纳入完整经济学评价的NHS EED摘要，及摘要在系统评价主要内容部分的描述性概括（Rodgers 2006, Fayter 2007）。

15.6.3 资源利用和成本数据的Meta分析

目前尚无达成一致的方法，如使用Meta-分析或其他定量合成方法，合并从多个经济学评价提取的成本-效果估计值（如增量成本-效果、成本-效用或成本-效益比）。然而，原则上来说，对于干预措施和对照措施，如果从两个或多个纳入的研究中获得的资源利用和成本指标的估计值采用相同的度量标准，则可使用Meta-分析进行合并。实践中，在Cochrane系统评价中考虑是否进行资源利用或成本数据的Meta-分析要十分谨慎。在决定是否采用Meta分析合并前，应特别注意不同研究间的度量标准是否一致。

资源利用和成本对同一国家内和不同国家间当地背景的特征变化很敏感，如当地价格或服务机构和服务提供方面（Drummond 2001，Sculpher 2004）。这限制了成本、资源利用以及相应成本-效果的估计值在不同背景下的外推性和可转化性。经济学评价中资源利用和成本与具体目标人群和地区密切相关，因此这也是作为当地特定背景下用于资源分配决策的最佳可得数据资源的根本原因（Cooper 2005）。跨地域和政治界限的资源利用或成本的Meta分析是否能产生有意义的结果，以及Meta-分析的结果如何解释，对Cochrane系统评价终端用户来说这些结果有什么额外价值？对这些问题都有争议。（Hutubessy等和Kumaranayake和Walker的文章中进一步讨论了卫生经济学评价的适用性和可转化性问题（Kumaranayake 2002，Hutubessy 2003）。）

另一方面，特定背景下的资源利用或成本的估计值是否可以外推或转化到不同背景下是个经验问题。当有证据证明不同研究间资源利用或成本的变化很小，这种情况下可以正当地合并。另外，清晰的描述成本的分布也很重要。许多已经完成的Cochrane系统评价包含了资源利用数据的Meta-分析。少量Cochrane系统评价包含了成本数据的meta-分析，虽然并非总是同时提供了这些数据产生方法的严格评价。

若Cochrane系统评价进行了资源利用或成本数据的Meta-分析，应提供产生这些估计值的相应卫生经济学研究的方法学的彻底的严格评价来支撑（见15.5.2，和15.6.2），同时采用统计学方法来调查和整合研究间异质性（如，I^2，chi-squared;随机效应模型：见9章，9.5）。来自多个研究的成本估计值应在合并数据前调整为通用货币和价格年（见15.6.1）。作者应参考第9章介绍的关于Meta-分析统计程序的进一步指导。

若进行资源利用或成本数据的Meta分析，在结果部分应包含一个描述性总结来评论结果的方向和大小及其精度。

同样，若一个系统评价纳入了两个或更多的卫生经济学研究，但（在Meta-分析中）

并没有合并这些研究的资源利用和/或成本数据，需在方法学部分陈述（见框15.6.a这类陈述举例）。

框15.6.a　对资源利用或成本数据不做meta-分析的陈述

> “［资源利用和成本结局指标］因认为不同试验间资源利用和成本数据不可比，故不对这类结局指标做合并分析…因为不同公共卫生系统的差异，这些结果与研究实施的具体国家相关。详细的报告显示在不同国家，不同条目之间的成本分摊明显不同。”（Birks 2006）。

15.6.4　建立经济学模型

Cochrane系统评价为后续或者同时进行的完整经济学评价提供所需证据的关键部分，包括使用决策分析方法合并或建模分析干预措施成本和效果的可得证据（见15.1.2和15.1.3）。与特定人群和背景下相关替代方法以及与所描述的具体分析角度（如，病人、卫生保健提供者和第三方支付者、卫生保健系统、社会）认为是相关的成本和结局指标相比，这一方法通常涉及点估计的估计和联合分布及干预措施引起的增量成本与效果（即成本-效果、成本-效用、成本-效益）的描述。

我们不在这里详细介绍经济学建模方法，因为并不推荐其作为Cochrane系统评价的常规内容。但鼓励希望追求干预措施的‘深层次’经济问题的Cochrane系统评价作者与制作经济学模型的专家合作。虽然有时可能制定一个经济学模型的通用结构，作为Cochrane系统评价中的经济学模型部分，不同背景下这一基本模型的输入和输出都是类似的，但要求用于模型的部分（甚至全部）数据都与当地具体背景相符。

尽管已经讨论过关于跨越地域和跨背景的经济学评价结果外推性和转化的问题（见15.6.3），也不能排除有时值得（尽管需要集中的时间、资源和专家人员）为Cochrane系统评价制定一个或者多个经济学模型。例如，若希望直接使用该系统评价作为未来整合有经济学评价的研究设计的一部分，则有利于推动为该Cochrane系统评价制定经济学模型。这时制定模型有助于弄清经济学评估中需要考虑哪些结构性假设和参数，及在研究中需要收集的数据。若在Cochrane系统评价使用这类方法，就需确定每个经济学模型实例的目的在于提供了在具体的地域和给定的时间点，干预措施与对照措施比较的成本-效果的解释性评估。

鼓励经济学模型制定者使用Cochrane系统评价包含的证据，以促进制定经济学模型。使用本章介绍的方法将经济学证据引入Cochrane系统评价中，一定程度上有利于提高在随后或同时进行的完整经济学评价中使用Cochrane系统评价的相关性和适用性。

15.7 解决报告偏倚

来自商业和其它类型的压力也许会影响研究的资金和关注卫生干预措施经济价值的研究结果的报告，这已广泛达成共识（Drummond 1992）。尽管如此，与效果研究相比，直到最近才有相对较少的研究关注卫生经济学评价研究中的发表问题和相关偏倚。然而，一些最近的研究开始使用系统评价和研究合成的方法来处理这一问题。

Bell和他的同事做了一个关于已发表的卫生保健领域中成本效果研究的系统评价，发现由公司资助的研究相对于无公司资助的研究，更倾向于报告在建议的成本-效果可接受性阈值以下或者附近的增量成本效果比（Bell 2006）。Miners和他的同事做了一个系统评价，比较相关卫生保健技术公司和签约的大学评估工作组分别提交给英国国家卫生与临床优质研究所（the National Institute of Health and Clinical Excellence，NICE）的关于卫生保健技术的成本效果证据（Miners 2005）。这项研究发现针对同一技术的评估，由厂家提供的增量成本效果比的估计值的平均值明显低于由大学工作组所提供的。Friedberg和他的同事发现已发表的由医药公司资助的肿瘤新药的经济学分析得出不利定量结论的可能性是非营利基金资助研究的八分之一，且得出有利定性结论的可能可能性是1.4倍（Friedberg 1999）。对这一问题，其他系统评价也普遍得出了类似结论（Freemantle 1997，Azimi 1998，Lexchin 2003）。这些方法学系统评价研究中讨论的常见主题是作者怀疑对研究结果的观察模式倾向于导致报告或发表偏倚。普遍的假设是，如果经济学分析结果暗示也许该干预措施在经济上没有吸引力，则赞助商、作者或杂志编辑自觉或不自觉地不予发表。

然而，以上所有的方法学系统评价受到研究设计限制（作者通常承认和讨论局限性）。调查是否存在报告和发表偏倚的理想的和最稳定的研究设计，涉及研究内发表结果与未发表结果的直接比较，或发表的研究及结果和未发表的研究结果的直接比较（Song 2000）。要实现这样系统的、广泛的比较难度很高，因为找到所有相关的、未发表的经济学分析的固有困难。由于缺乏这样的数据，所以不能排除对结果的观察模式其它的替

代解释（如，能反映增量成本-效果比真实分布的结果）。

系统评价中解决发表偏倚的方法可以应用于卫生经济学研究的系统评价中，且有相同的解释说明，这已在第10章介绍。已经提出了在经济学评价研究中帮助解决发表偏倚方法的建议，如那些在Cochrane系统评价中也许会遇到的问题：

1. 鼓励更透明、更一致性的方法来开展和报告经济学分析，通过发布指导这类研究严格评价的良好实践指南和量表——特别是基于综述的研究和模型研究；
2. 增强对杂志投稿中研究赞助商和作者之间潜在利益冲突的审查；
3. 增强经济学评价中所有基础数据的可及性以提高方法学的透明度。

15.8 结果的解释

卫生经济学研究系统评价的结果解释取决于具体的经济学问题及与卫生保健服务决策相关的背景。Cochrane系统评价中——针对国际用户——明显存在大量潜在的经济学问题和不同决策机构需要考虑的背景因素需要考虑。放在全球背景下，对多个经济学评价研究的严格系统评价结果进行单纯的解释以便于得出采纳或反对某项卫生保健治疗措施或诊断性试验的结论，并不可行。虽然在这种情况下Cochrane系统评价也不可能提供任何政策评估的中心要素，但它仍然有利于完善经济学讨论并在国际背景下进行讨论（Gilbody 1999）。

一个系统评价中很少或者没有与该主题相关的高质量经济学评价研究，对卫生经济学研究的严格评价可高度提示缺少经济学证据，需要未来的研究解决。需要进一步开展经济学评价研究应该在“对未来研究的启示”和“作者结论”中陈述。框15.8.a列出两个这类例子。因为完整经济学评价也取决于干预措施效果可靠数据的可得性，因此也要考虑缺乏严格的效果研究将显著影响完整经济学评价的可行性和可得性。再者，Cochrane系统评价和其他评价无法克服这个局限性，都需在结论部分提醒注意。

框15.8.a 结论部分强调需要更多经济研究

“在大部分时间中，［纳入研究中］并未计算干预措施的成本。这极其重要。未来研究中应该计算节约的成本，并且与提供干预措施的潜在成本相权衡……是否能提供该项具有成本效果的服务是当前卫生保健环境中的关键问题。因此需要关于成本和药剂师干预措施效果指标的研究。”（Beney 2000）。

15.9 结论

本章概述了将经济学证据引入Cochrane系统评价的方法学框架。同时不可能实践，也不推荐生产关于“干预措施X是否具有成本效果”的论述，它仅可帮助决策者理解他们需要解决的资源分配问题的结构，需要考虑主要的参数，不同资源利用背景、成本和成本-效果间的差异，及这些差异潜在的原因（Drummond 2002）。引入经济学证据也可加强Cochrane系统评价对随后（或平行）进行的完整经济学评价提供数据的有用性和适用性。可以预计的是，这个指南将在更广泛的读者批评建议下继续完善和更新，该方法也会随其在Cochrane系统评价中应用经验的基础上和进一步方法学研究中发展。

制定这个指南的过程也有助于为进一步明晰未来研究的重要优先次序，以制定和测试如何识别、评价、分析和报告干预措施经济方面证据的替代方法。关键优先次序包括：进一步完善资产负债表法来总结系统评价中经济学成分的结果；评估经济学系统评价中应用不同方法学质量标准或不同经济学评价研究纳入标准对结果的影响；及应用个体数据调查和处理不同资源利用背景、成本和效用（及其他健康状态偏好）间异质性的方法。CCEMG网站的‘研究’页面列出了这些和其他方法研究的优先次序。（见框15.10.a）

15.10 本章信息

作者：Ian Shemilt, Miranda Mugford, Sarah Byford, Michael Drummond, Eric Eisenstein, Martin Knapp, Jacqueline Mallender, David McDaid, Luke Vale, Damian Walker on behalf of the Campbell and Cochrane Economics Methods Group.

本章引用格式：Shemilt I, Mugford M, Byford S, Drummond M, Eisenstein E, Knapp M, Mallender J, McDaid D, Vale L, Walker D. Chapter 15: Incorporating economics evidence. In: Higgins JPT, Green S(editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.

致谢：Dawn Craig, Julian Higgins, Kevin Marsh和John Nixon对草稿的意见。

框15.10.a　Campbell和Cochrane 经济学方法学组

Campbell和Cochrane经济学方法学组（CCEMG）于1998年正式注册为Cochrane协作网 方法学组，2004年联合注册为Campbell协作网方法学组。在可得资源下该方法学组的核心目的包括：

- 促进和支持在系统评价中考虑经济学问题；
- 为系统评价相关用户开发恰当的、无偏倚的、客观的Cochrane系统评价的经济学方法，和
- 为系统评价作者和编辑联系可以帮助其评价的经济学家，或提供专家建议和同行评议。

许多Cochrane系统评价已经包含了干预措施的经济学方面。然而，本章是本手册第一次纳入了在Cochrane系统评价中应用经济学方法的详细指南。未来本章将在正在开展的方法学研究项目和Cochrane系统评价引入经济学证据的更多经验基础上更新。

E-mail: research@c-ceng.org

Web site: www.c-cemg.org

15.11　参考文献

Aguiar-Ibanez 2005

Aguiar-Ibanez R, Nixon J, Glanville J, Craig D, Rice S, Christie J, Drummond MF. Economic evaluation databases as an aid to healthcare decision-makers and researchers. Expert Review of Pharmacoeconomics and Outcomes Research 2005; 5: 721-722.

Azimi 1998

Azimi NA, Welch HG. The effectiveness of cost-effectiveness analysis in containing costs. Journal of General Internal Medicine 1998; 13: 664-669.

Bell 2006

Bell CM, Urbach DR, Ray JG, Bayoumi A, Rosen AB, Greenberg D, Neumann PJ. Bias in published cost effectiveness studies: systematic review. BMJ 2006; 332: 699-703.

Beney 2000

Beney J, Bero LA, Bond C. Expanding the roles of outpatient pharmacists: effects on health services utilisation, costs, and patient outcomes. Cochrane Database of Systematic Reviews 2000, Issue 3. Art No: CD000336.

Best 2000

Best L, Simmonds P, Baughan C, Buchanan R, Davis C, Fentiman I, George S, Gosney M, Northover J, Williams C, Colorectal Meta-analysis Collaboration. Palliative chemotherapy for advanced or metastatic colorectal cancer. Cochrane Database of Systematic Reviews 2000, Issue 2. Art No: CD001545.

Birks 2006

Birks J, Harvey RJ. Donepezil for dementia due to Alzheimer's disease. Cochrane Database of Systematic Reviews 2006, Issue 1. Art No: CD001190.

Briggs 2006

Briggs A, Sculpher M, Claxton K. Decision Modelling for Health Economic Evaluation. Oxford (UK): Oxford University Press, 2006.

Brown 2007

Brown SR, Nelson RL. Surgery for faecal incontinence in adults. Cochrane Database of Systematic Reviews 2007, Issue 2. Art No: CD001757.

Campbell 2007

Campbell H, Briggs A, Buxton M, Kim L, Thompson S. The credibility of health economic models for health policy decision-making: the case of population screening for abdominal aortic aneurysm. Journal of Health Services Research and Policy 2007; 12: 11-17.

Cochrane 1972

Cochrane AL. Effectiveness and Efficiency: Random Reflections on Health Services. London (UK): Nuffield Provincial Hospitals Trust, 1972.

Cooper 2005

Cooper N, Coyle D, Abrams K, Mugford M, Sutton A. Use of evidence in decision models: an appraisal of health technology assessments in the UK since 1997. Journal of Health Services Research and Policy 2005; 10: 245-250.

Craig 2007

Craig D, Rice S. CRD Report 6: NHS Economic Evaluation Database Handbook (3rd edition). York (UK): Centre for Reviews and Dissemination, University of York, 2007.

Donaldson 2002

Donaldson C, Mugford M, Vale L. From effectiveness to efficiency: an introduction to evidence-based health economics. In: Donaldson C, Mugford M, Vale L (editors). Evidence-based Health Economics: From Effectiveness to Efficiency in Systematic Reviews. London (UK): BMJ Books, 2002.

Drummond 1992

Drummond MF. Economic evaluation of pharmaceuticals: science or marketing? Pharmacoeconomics 1992; 1: 8-13.

Drummond 1996

Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. BMJ 1996; 313: 275-283.

Drummond 2001

Drummond M, Pang F. Transferability of economic evaluation results. In: Drummond M, McGuire A (editors). Economic Evaluation in Health Care: Merging Theory with Practice. New York (NY): Oxford University Press, 2001.

Drummond 2002

Drummond M. Evidence-based medicine meets economic evaluation – an agenda for research. In: Donaldson C, Mugford M, Vale L (editors). Evidence-based Health Economics: From Effectiveness to Efficiency in Systematic Reviews. London (UK): BMJ Books, 2002.

Drummond 2005

Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes (3rd edition). Oxford (UK): Oxford University Press, 2005.

Evers 2005

Evers S, Goossens M, de Vet H, van Tulder M, Ament A. Criteria list for assessment of methodological quality of economic evaluations: Consensus on Health Economic Criteria. International Journal of Technology Assessment in Health Care 2005; 21: 240-245.

Fayter 2007

Fayter D, Nixon J, Hartley S, Rithalia A, Butler G, Rudolf M, Glasziou P, Bland M, Stirk L, Westwood M. A systematic review of the routine monitoring of growth in children of primary school age to identify growth-related conditions. Health Technology Assessment 2007; 11: 22.

Freemantle 1997

Freemantle N, Mason J. Publication bias in clinical trials and economic analyses. Pharmacoeconomics 1997; 12: 10-16.

Friedberg 1999

Friedberg M, Saffran B, Stinson TJ, Nelson W, Bennett CL. Evaluation of conflict of interest in economic analyses of new drugs used in oncology. JAMA 1999; 282: 1453-1457.

Gamble 2006

Gamble CL, Williamson PR, Marson AG. Lamotrigine versus carbamazepine monotherapy for epilepsy. Cochrane Database of Systematic Reviews 2006, Issue 1. Art No: CD001031.

Gilbody 1999

Gilbody SM, Petticrew M. Rational decision-making in mental health: the role of systematic reviews. Journal of Mental Health Policy and Economics 1999; 2: 99-106.

Hutubessy 2003

Hutubessy R, Chisholm D, Edejer TT. Generalized cost-effectiveness analysis for national-level priority-setting in the health sector. Cost Effectiveness and Resource Allocation 2003; 1: 8.

Kumaranayake 2002

Kumaranayake L, Walker D. Cost-effectiveness analysis and priority setting: Global approach without local meaning? In: Lee K, Buse K, Fustukian S (editors). Health Policy in a Globalising World. Cambridge (UK): Cambridge University Press, 2002.

Lavis 2005

Lavis J, Davies H, Oxman A, Denis JL, Golden-Biddle K, Ferlie E. Towards systematic reviews that inform health care management and policy-making. Journal of Health Services Research and Policy 2005; 10 Suppl 1: 35-48.

Lexchin 2003

Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. BMJ 2003; 326: 1167-1170.

Maynard 2000

Maynard A, Kanavos P. Health economics: an evolving paradigm. Health Economics 2000; 9: 183-190.

Miners 2005

Miners AH, Garau M, Fidan D, Fischer AJ. Comparing estimates of cost effectiveness submitted to the National Institute for Clinical Excellence (NICE) by different organisations: retrospective study. BMJ 2005; 330: 65.

Napper 2005

Napper M, Varney J. Etext on Health Technology Assessment (HTA) Information Resources. Chapter 11: Health Economics Information. Available from: http://www.nlm.nih.gov/archive//2060905/nichsr/ehta/ chapter11.html (accessed 1 January 2008).

Neumann 2005

Neumann PJ, Greenberg D, Olchanski NV, Stone PW, Rosen AB. Growth and quality of the cost-utility literature, 1976-2001. Value in Health 2005; 8: 3-9.

Philips 2004

Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, Woolacoot N, Glanville J. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technology Assessment 2004; 8: 36.

Rennie 2000

Rennie D, Luft HS. Pharmacoeconomic analyses: making them transparent, making them credible. JAMA 2000; 283: 2158-2160.

Rodgers 2006

Rodgers M, Nixon J, Hempel S, Aho T, Kelly J, Neal D, Duffy S, Ritchie G, Kleijnen J, Westwood M. Diagnostic tests and algorithms used in the investigation of haematuria: systematic reviews and economic evaluation. Health Technology Assessment 2006; 10: 18.

Samuelson 2005

Samuelson PA, Nordhaus WD. Economics. London (UK): McGraw-Hill, 2005.

Sculpher 2004

Sculpher MJ, Pang FS, Manca A, Drummond MF, Golder S, Urdahl H, Davies LM, Eastwood A. Generalisability in economic evaluation studies in healthcare: a review and case studies. Health Technology Assessment 2004; 8: 49.

Shemilt 2007

Shemilt I, Mugford M, Byford S, Drummond M, Eisenstein E, Knapp M, Mallender J, McDaid D, Vale L, Walker D. Where does economics fit in? A review of economics in Cochrane Reviews. 15th Cochrane Colloquium, Sau Paulo (Brazil), 2007.

Song 2000

Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. Health Technology Assessment 2000; 4: 10.

Weinstein 2003

Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, Luce BR. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices - Modeling studies. Value in Health 2003; 6: 9-17.

Williams 1987

Williams A. Health economics: The cheerful face of the dismal science? In: Williams A (editors). Health and Economics. London (UK): Macmillan, 1987.

Zarnke 1997

Zarnke KB, Levine MA, O'Brien BJ. Cost-benefit analyses in the health-care literature: don't judge a study by its label. Journal of Clinical Epidemiology 1997; 50: 813-822.

（袁强译，王莉、秦天强、岑啸初审）

第十六章　统计学中的特殊问题

编辑：代表 Cochrane 统计学方法学组的 Julian PT Higgins, Jonathan J Deeks 和 Douglas G Altman。

本文选自工作手册5.0.1版本。有关如何引用它的指南，见16.10节。这些材料还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容摘要

- 因数据缺失限制了某项研究被纳入 Meta 分析（并且试图与作者联系也未能获取相应数据）时，任何数据填补方法都应加以描述，同时应用敏感性分析评估其影响。
- 对于非标准设计研究，如整群随机对照试验和交叉试验应使用与设计相符的分析方法。即使研究者不能解释结果数据之间的相关性，系统评价作者同样可以采用近似的方法加以分析。
- 若 Meta 分析中的某项研究包含了两个以上干预组，推荐的方法是先将相关组合并

成一组后，再按照典型的一个试验组与一个对照组进行比较分析。

- 尽管干预措施间的间接比较有可能造成一些误导，但目前已有一些不破坏原有随机化的方法，包括由此演化而来的多臂试验 Meta 分析。
- 为减少多元统计分析可能造成的错误结论，系统评价作者应在计划书中事先规定好将要进行的统计分析，将重复比较的次数控制在最小；同时在解释有统计学意义的结果时，应说明一共进行了多少次假设检验。
- 使用贝叶斯方法和分层（或多水平）模型可以进行较为复杂的 Meta 分析，并在技术与结果解释上比 RevMan 中的标准方法更具优势。
- 零阳性事件研究无法直接计算 RR 或 OR。对于罕发事件研究，Peto 法由于偏倚低、检验效能高，成为 Meta 分析方法的首选。

16.1 缺失数据

16.1.1 缺失数据的类型

在系统评价和Meta分析中缺失数据的表现形式多种多样（见表16.1.a）。例如：在系统评价中可能是缺失了一个完整研究，或者一个研究中可能缺失了一个结局指标，或一个结局指标的汇总数据缺失，或者汇总数据中缺失了某个体病人数据。本章仅对数据缺失的可能来源进行讨论，更详细的讨论详见本手册相关章节。

一个系统评价中可能会漏掉某个研究，原因很多，如从未公开发表、或发表在冷门杂志、或很少被引用、或数据库标引引不当等。因此，作为系统评价作者应时刻提醒自己发生漏检的可能性。一个研究文献被漏检，有很大的可能性是由于其研究结果‘令人不感兴趣’或‘不受欢迎’（即出现了发表偏倚）。对这个问题的深入讨论详见第10章，有关详尽检索方法的描述见第6章。

有些研究中可能没有报告系统评价者所感兴趣的结局指标。例如，生存质量或严重的副作用等。究竟是未测量还是测量了但未报告，通常很难判断。此外，这些结局指标未报告，有可能与选择性报告结果有关（选择性报告偏倚见第8章，第8.13节）。同样地，结局指标的汇总数据也可能会缺失，这些汇总数据常可以直接用于Meta分析。如，最常见的就是连续性变量资料的标准差缺失。特别是在干预前后差别分析比较时，前后变化差值的标准差常常缺失，我们在第16章1.3节中专门介绍了缺失标准差的估算方法。可能

出现缺失的汇总数据还有：样本含量（特别是各干预组的样本量）、阳性事件数、标准误、计算率时所需的随访时间、以及时间事件结果的详细信息等。对于一些特殊设计的研究（如整群随机对照试验和交叉试验），若统计方法使用不恰当，也可能得到缺失的汇总结果，有时需要使用相关系数或标准差进行校正分析，相关内容和方法可参考本章第3节整群随机对照试验以及第4节的交叉试验。总的来说，大多数方法学家认为在系统评价中不应将汇总数据（如无可用数据）缺失作为研究的排除标准，相反，这类研究更适合纳入到系统评价中，并探讨其缺失对Meta分析结果的潜在影响。

在有些纳入的研究中（当然不是全部纳入研究），可能会出现个别个体的数据缺失。须知，只要不是对随机对照试验中所有随机分组研究对象的分析，就不能算作意向性分析。即使原始研究者并未采用，有时也可能进行意向性分析，相关方法见本章第2节。

数据缺失也能影响亚组分析。如果计划使用亚组分析或Meta回归（见第9章第6节），则需要能与其它研究区别的研究水平的、详细特征信息。若原始研究中未报告这些信息，则需要向其作者索要。

表 16.1.a　Meta分析中的数据缺失类型

缺失数据类型	**缺失数据的可能原因**
项研究缺失	发表偏倚; 检索不充分.
结局指标缺失	结局未测量; 选择性报告偏倚.
汇总数据缺失	选择性报告偏倚 报告不完全
单个病例数据缺失	缺少意向性分析 失访 选择性报告偏倚.
研究水平的特征信息缺失（用于亚组分析或Meta回归）	特征未测量 报告不全

16.1.2 数据缺失的一般处理原则

用于处理数据缺失的统计方法很多，这里我们只简要地为Cochrane系统评价作者介绍一些基本概念和给出一些一般的建议。考虑数据为什么会缺失？这一点很重要。统计学家常用“随机缺失”和“非随机缺失”表示不同的缺失类型。

如果丢失与数据本身无关，则称为“随机缺失”。例如，如果在邮寄过程中丢失一些生存质量调查问卷，这种丢失与调查对象的生存质量并无关联。在某些情况下，统计学家又将这种缺失进一步细分为“随机缺失”和“完全随机缺失”，但在系统评价中这种划分并无多大的实际价值。尽管缺失会造成样本量减少，但由于基于可用数据分析仍属于无偏估计，随机性的数据缺失对结果的影响有限。

如果丢失与数据本身有关，就属于“非随机缺失”。例如，在一项研究抑郁的试验中，抑郁复发患者非常有可能放弃随访，从而造成结局数据缺失。这类数据缺失对结果的影响“不可忽视”，否则会造成基于可用数据典型的有偏估计。发表性偏倚、选择性报告偏倚以及原始研究中的个体病人退出和脱落等，根据这个定义都属于非随机性缺失。

数据缺失的主要处理方法如下：

1 仅分析现有数据（即忽略缺失数据）

2 用一些可以当作观测值的数据进行替代（如用最近一次的随访观测值、数据缺失个体均按最坏结局估计、使用均数替代、或用回归分析的预测值估计）

3 用不确定数据替代（如多重估计、或结合标准误进行简单校正估计，方法同上）

4 基于现有数据结构与缺失数据的关系，使用能处理缺失数据的统计模型分析。

第1种方法可能适用于随机性数据缺失的处理。第2-4种方法主要用于处理非随机性数据缺失。第2种方法在很多情况下实用并在系统评价中常用，但需注意替代值往往是不确定的，并且其可信区间过窄。第3和第4种方法较为复杂，需要有经验丰富的统计人员参与。

针对Cochrane系统评价中的缺失数据有如下四个处理的一般建议：

- 只要可能，就应联系原始研究者、获取相关缺失数据
- 对缺失数据的任何处理方法均应做出明确的假设：如，假设数据是“随机丢失”的，或者缺失数据均假设按照最坏结局处理。
- 实施敏感性分析用来评估在所假设的条件下缺失数据替换对结果的影响以及结果的稳定性。（见第9章，9.7）

- 在系统评价的讨论部分应说明缺失数据对结果的潜在影响。

16.1.3 标准差缺失

16.1.3.1 填补标准差

标准差缺失在计量资料Meta分析中常见，一种解决方法是需要进行估计。在估计标准差之前，作者应仔细寻找用于估算标准差所需要的基础数据（如：可信区间、标准误、t值、P值、F值等），见第7章（第7.3节）相关内容。

最简单的估算方法就是直接从一个或多个类似原始研究中借用标准差。Furukawa等发现不同借用途径的结果是很接近的，如从同一Meta分析的其它研究中借用，与从其它Meta分析的原始研究中借用（Furukawa 2006）。若有多个标准差可供选择，是使用它们的平均值、最大值、相对较大值或是其它，需要系统评价者自己决定。对于均数差（MD）的Meta分析，若选择较大标准差，可能会降低该研究的权重并得到精度较差的可信区间。对于标准化均数差（SMD）的Meta分析，若选择的标准差过大，则会使结果更偏向于无效。当然，若同时有多个候选标准差，可以使用较为复杂的方法估计。例如，鉴于对数均数与对数标准差间存在强的线性关系，Marinho等曾建立以log（均数）为x，以log（SD）为y的直线回归方程，用以估计标准差（Marinho 2003）。

由于上述所有估算方法均涉及了对未知统计量的假设，除非是不得已而为之，否则最好避免使用。若Meta分析中大多数研究的标准差缺失，再估计这些值已无必要。相反，若只有少数研究中的个别数据缺失，就可以进行估算，并与其它有完整数据的研究一起进行合并分析。同时统计量假设带来的变化对结果的影响可用敏感性分析进行评估。

16.1.3.2 前后变化差值的标准差估算

与基线相比，可以计算前后变化差值，但其标准差往往被忽略，通常只能得到以下信息：

	基线（治疗前）	**治疗后**	**前后变化值**
试验组 （样本量）	均数±标准差	均数±标准差	均数差值
对照组 （样本量）	均数±标准差	均数±标准差	均数差值

注意：各组均数差值一般由各组治疗前后测量值直接相减得到（即使在原始研究中没有报告，也可以通过手算获取）。然而根据上表信息无法计算出前后变化差值的标准差，以至于不能判断前后变化差值的变异大小。但如果原始研究中报告了其他一些信息，将有助于计算前后变化差值的标准差。如果给出了差值组间比较的统计分析结果(可信区间，t值、P值或F值等)，可以用第7章描述的方法来计算差值的标准差，详见（第7.3节）。

当可用于计算变化值的标准差的信息缺乏时，则需要估计标准差。如在同一个系统评价中其它原始研究有前后变化差值的标准差，可以合理地用来替代缺失的标准差。但其恰当性取决于以下条件：相同测量尺度、相同程度的测量误差、相同的时间段（基线和结果测量期间）。

采用以下方法也可估算变化差值的标准差（Follmann 1992，Abrams 2005）。方法之一就是借助相关系数来估算。相关系数可以用来描述受试者基线测量值与结局测量值间的相似程度的，但在临床试验中较少使用。这里我们分两步来估计前后变化差值的标准差。（1）先利用一个报告充分的原始研究计算出相关系数。(2)再利用相关系数估算出另一个报告不全原始研究的变化差值的标准差。注意这里使用的相关系数既可通过(1)中的方法获取，也可通过其他方法(如理论推导法)得到。但使用这种方法应慎重，原因在于我们不能确定所估计的相关系数是合适的(例如，基线值与终点值间的相关性可能会随着观察时间间隔的延长而降低，同时该相关系数也可能与结局指标性质及受试者特征有关)。另一种简单方法就是直接使用干预后测量值进行比较分析，这是因为在随机对照试验中组间的基线是均衡可比的，使用干预后测量值比较与干预前后差值比较，理论上两者具有相同的分析价值。

(1) 利用报告完整的研究计算相关系数

假设某研究提供了基线测量值、干预后测量值、干预前后变化差值的均数及标准差，例如：

	基线测量值	干预后测量值	前后变化差值
试验组 (样本量 129)	15.2± 6.4	16.2±7.1	1.0±4.5
对照组 (样本量 135)	15.7±7.0	17.2±6.9	1.5±4.2

根据上表中最后一列的数据即可对终值相对于基线的变化值进行分析。但是，我们可以利用此研究中的其它数据计算两个相关系数，即每个干预组的。我们使用下面的表示法：

	基线测量值	干预后测量值	前后变化差值
试验组 (样本量 NE)	$M_{E,baseline} \pm SD_{E,baseline}$	$M_{E,final} \pm SD_{E,final}$	$M_{E,change} \pm SD_{E,change}$
对照组 (样本量 NC)	$M_{C,baseline} \pm SD_{C,baseline}$	$M_{C,final} \pm SD_{C,final}$	$M_{C,change} \pm SD_{C,change}$

试验组的相关系数CorrE计算公式如下：

$$Corr_E = \frac{SD^2_{E\ baseline} + SD^2_{E\ final} - SD^2_{E\ change}}{2 \times SD_{E\ baseline} \times SD_{E\ final}}$$

对照组的相关系数CorrC的计算同上。

在本例中计算如下：

$$Corr_E = \frac{6.4^2 + 7.1^2 - 4.5^2}{2 \times 6.4 \times 7.1} = 0.78$$

$$Corr_C = \frac{7.0^2 + 6.9^2 - 4.2^2}{2 \times 7.0 \times 6.9} = 0.82$$

若基线值或干预后测量值的标准差有一方不能获得,可用另一方替代(假定干预措施不影响结果变量的变异性)。相关系数的值在–1 和1之间。如果相关系数小于0.5，则使用前后变化差值分析的价值不大，相反，直接使用干预后测量值分析将更精确。假设各组的相关系数近似，可以通过一个简单的平均，获得该研究所有个体类似的基线和结果测量的相关系数（例如，0.78和0.82的平均是0.80）。但如果出现各组的相关系数差异明显，一者可能与用于有效估计的样本量过小有关且干预措施影响测量指标的变异大小，二可能是，干预效应的大小与基线水平的高低有关，最好放弃估算。同时在估算之前，建议先尝试性地计算出Meta分析中多数研究的相关系数，看看是否一致。若不一致，那

么估算就权当做了一次试探性分析。

(2) 用相关系数估算前后变化值的标准差

现在假设一个研究，没有干预前后变化值的标准差，但是，当知道基线和结果的标准差时，我们可以通过所估计的相关系数对缺失的标准差进行估计。利用下列公式估算前后变化差值的标准差。公式中的Corr值为相关系数，既可由同一Meta分析中的其它研究（上述（1）方法）得到，也可以从其它地方估算或者根据理论推断进行假定。但无法是哪种途径，均应进行敏感性分析，用以评价不同Corr值对估计结果稳定性的影响。

试验组干预前后差值的标准差估计公式：

$$SD_{E\ change} = \sqrt{SD^2_{E\ baseline} + SD^2_{E\ final} - (2 \times Corr \times SD_{E\ baseline} \times SD_{E\ final})}$$

对照组的计算同上。同样，若基线值或干预后测量值的标准差有一方不能获得,可用另一方替代（如果能够假定干预措施不影响结果变量的变异性是合理的）。

以下面数据为例：

	基线测量值	干预后测量值	前后变化差值
实验组 （样本量 35）	12.4±4.2	15.2±3.8	均数=2.8
对照组 （样本量 38）	10.7±4.0	13.8±4.4	均数=3.1

如果相关系数为0.8，则对照组变化差值的标准差为：

$$SD_{E\ change} = \sqrt{4.0^2 + 4.4^2 - (2 \times 0.80 \times 4.0 \times 4.4)} = 2.68$$

16.2 意向性分析相关问题

16.2.1 引言

在随机对照试验中常会有一些受试者因为各种原因被排除，如失访而导致无法观测结局指标，或者因与研究方案背离（如误用其他干预措施或未接受治疗、依从性差、不符合纳入标准等）。另外，若某些特定结局指标变量的测量需要依赖其他变量，这样可能会导致全部受试者的某些结局指标无法测量（详见本章第2.4节）。在分析时，如果将部分受试者排除在外，那么对干预措施效果的估计就可能产生偏倚，具体描述详见第8章（第12节）。意向性分析则是指不论在试验后续情况如何，当初所有参与随机分组的受试者均统统纳入分析（Newell 1992，Lewis 1993）。意向性分析之所以被广泛接受，关键在于其本身是无偏倚的，同时又能解决非常贴近临床实际情况的问题。

以下为意向性分析的处理原则，详见第8章第12节。

1. 无论受试者实际接受了何种干预措施，均按照当初随机分组统计；
2. 测量所有受试者的结局指标；
3. 将所有参与随机分配的受试者纳入分析。

这些原则是否应同时使用，目前还没有定论（Hollis 1999）。但第一条原则已被广泛接受，第二条原则通常认为不可行，而第三条因涉及对结局未知的受试者缺失数据的估计，仍存有争议（详见本章第1.2节）。

针对数据完整的受试者分析常被称为“可用病例分析”。有些临床试验只报告了那些完成试验受试者的分析结果以及那些依从（或部分接受）干预方案受试者的分析结果，一些作者误认为这类分析就是意向性分析，但实际上这只是“符合方案分析”。此外，还有一些研究者只对那些实际接受干预的受试者进行分析而忽略其随机分配情况（即“完成治疗分析”）。因此，作者在临床试验报告中所描述的ITT分析，通常需要根据其提供的详细资料进行核实判断。

许多人（不是所有人）认为“可用病例分析”和ITT分析并不适用于不良反应分析，理由就是有的受试者实际上并未接受某种治疗，却将发生的不良反应归因于该种治疗显然不合理。同时由于ITT分析可能使结果趋向为无差别，并不适合等效性检验或非劣效检验。

大多数情况下，作者至少应该从文献中提取出可用于“可用病例分析”的数据。如

有可能，尽量重新纳入可避免被排除的受试者数据。在个别情况下，有可能将文中和表格提供的信息以及从作者处获取的被试验报告排除但有随访数据的受试者信息结合起来，从而实现真正意义的ITT分析。若有这样可能，不需要对研究结果进行填补，应尽量采用此种方法进行真正的ITT分析。

否则，只能通过填补数据的方法来进行意向性分析，这就涉及对没有记录结局的受试者结局做出假定，但多数情况下，这样的估算分析与“可用病例分析”相比，只有在精确性中存在明显的毫无根据的扩大的情况下才会有不同。对缺失数据过多的试验结果进行评价终究要涉及主观判断、以权衡利弊，详见第8章（第12节）。须知，统计分析并不能完全弥补数据本身的缺陷（Unnebrink 2001）。因为没有假设能充分反映真实的情况，所以应使用多种方法进行敏感性分析、以全面评估假设可能造成的影响（详见第9章，第9.7节）。

在接下来的两节中，我们将介绍一些针对二分类变量或连续性变量缺失数据的处理方法。尽管可以估计缺失数据，但在目前最明智的选择还是只对数据完整者进行分析，并评估这些缺失数据对偏倚风险的潜在影响(见第8章，第12节)。通常在研究方案及系统评价的方法学部分应详细报告缺失数据的填补方法及其基于的假设。

如果单个受试者数据可用，则可以考虑进行周密地敏感性分析。系统评价者在这个阶段可以参阅处理临床试验中缺失数据的大量文献（Little 2004）。若有可能，一旦获取个体受试者的数据，就应把那些在研究报告中被排除分析的受试者重新纳入（Stewart 1995）。若在分析中对这部分重新纳入对象的详细信息报告不全，应向临床试验者索取必要信息资料。

16.2.2 二分类数据的意向性处理

通常应在纳入研究的偏倚风险表中报告无结局观察指标的受试者比例；注意该比例可能会因结局指标和随机组别的不同而有所变化。但在数据分析中如何处理这些受试者还没有公认的好办法。下面介绍的两种基本方法，并且合理地选择应该是使用这两种方法作为主要分析和敏感性分析（见下文和第九章第7节）的基础。

- 可用（完整数据）病例分析：只纳入有结局观察指标的受试者，将研究中有特定结局的数据的总人数作为分母；同时将各纳入研究中缺失数据的变异程度作为异质性的潜在来源。

- 估算缺失数据、实施ITT分析：无论原始研究者如何分析处理，将所有参加随机分组的受试者统统纳入分析。这涉及对失访病人结局的估计，在一些用于估计二分类结局数据的方法中，最常用的方法就是假设所有失访者均出现结局事件或假设失访者均未出现结局事件。另外也可利用对照组事件发生率或者不同组中所有完成试验者的结局事件发生率进行估算（后者可得到相同的干预效应量估计值，但导致效应估计值精度无根据的扩大）。如何估算缺失数据必需借助临床经验加以权衡。将估算了缺失二分类数据的原始研究纳入RevMan分析时，相应的权重可能比实际高，建议向统计专家咨询后、再决定更为合适的权重值（Higgins 2008）。但是，没有任何假设能够反映真实情况，除非在一些特定情况下，如戒烟试验中估计‘失败’，因此，在一般情况下，缺失数据估算方法应慎用。

缺失数据对结果的潜在影响应在系统评价的结果解释中加以讨论。这些影响常与数据缺失程度、事件发生频率以及合并效应量大小等有关。Gamble和Hollis认为对二分类结局指标的敏感性分析应根据“最好”结局和“最坏”结局两种情况分别进行分析（Gamble 2005）。所谓“最好”结局是指试验组中的失访对象均未发生不良结局事件，而对照组中的失访者均发生了不良结局事件；“最坏”结局的分析，则刚好相反。若“最好”结局分析结果与“最坏”结局结果相比，差别过大，那么该原始研究的权重应该降低，但要注意该种做法可能有些极端。

除此以外，另有一种看似更合理的敏感性分析就是尝试确定失访对象的事件发生率。例如，假设在一个已采用可用病例分析的结果中，某项研究的试验组事件发生率为20%而对照组是15%。假设失访对象也遵从相似的比例，那么就有三种合适的敏感性分析来进行比较，即假设两组失访对象的不良事件发生率均为15%，或假设试验组和对照组失访者事件发生率分别为15%和10%，或两组依次为20%和10%。另一种假设，就是在主要分析中所有失访者均发生了不良结局事件，那么就有两种敏感性分析来进行比较，即，两组失访对象的事件发生率均为10%，或干预组为10%，对照组为5%。可以考虑使用图形法展示不同假设下的敏感性分析结果（Hollis 2002）。

Higgins等从失访或缺失原因入手，提出一种替代方法，认为数据缺失的受试者中事件发生的风险与观察结局指标本身的风险有关（Higgins 2008）。另外，White 等建议使用贝叶斯方法，该法自动地降低那些有较多缺失数据研究的权重（White 2008a，White 2008b）。

16.2.3 连续性数据的意向性处理

在完全的ITT分析中，应纳入所有参与随机分组的研究对象，包括未接受干预方案者和失访者。同时要得到所有参与随机受试者结局变量的均数和标准差。与二分类数据处理原则一样，在纳入研究的偏倚风险表中报告受试者的退出率。同样，下面介绍两种方法，可供敏感性分析时参考（见第9章，第9.7节）

- 可用（完整数据）病例分析：只纳入结果已知的病例数据。缺失数据对结果的潜在影响应在系统评价的结果解释中讨论。这些影响同样与数据缺失程度，合并效应量估计值大小以及结局指标的变异程度等有关。同时将各纳入研究中缺失数据的变异程度作为异质性的潜在来源。
- 估算缺失数据、实施ITT分析：无论原始研究者如何分析处理，将所有参加随机分组的受试者统统纳入分析。需要估计失访病人的结局指标。然而专门介绍Meta分析中连续性数据缺失估计方法的方法学文献目前还不多见。在有些情况下可能会利用标准（尽管通常是有问题的）估计方法，如根据失访者最后一次观察的结果或用其与基线相比的变化值进行替代，但这种方法通常需要病人的个体初始观察数据。通过对缺失数据的估计，变相地扩大了可用样本量，可能会导致结果精度被人为夸大，因此，不推荐使用这种方法。

对连续性数据进行敏感性分析的一个简单方法就是假设缺失数据均数与分析所得均数之间存在一个固定的差值。例如，在对可用病例数据进行分析后，可假设试验组缺失数据比实际观测数据平均多2个观察单位，而对照组的缺失数据则比实际观测数据平均少2个单位，然后再次进行比较，分析结果的敏感性。另外，White等建议使用贝叶斯方法，对于有较多缺失数据研究，该法将自动降低其权重值（White 2007）。

16.2.4 适用于部分受试者的结局观察指标

有些试验的结局观察指标仅适用于部分受试者。例如，在研究不孕不育的临床试验中，通常会给出治疗后临床妊娠流产的比例。根据这个结局指标的定义，那些未达到一个临界状态（临床妊娠）的受试者将被排除，这将破坏试验自身的随机性。一般地，最好能够重新调整结局观察指标，使得能够纳入所有随机分组的受试者进行分析，如在本例中，可将结局变量改为受试妇女是否‘成功妊娠’（即开始妊娠并满24周或足月）。另一个例子是观察受试者的中期或长期预后（如出现慢性肺部疾病）时，若部分受试者出现

死亡将会低估该预后指标，简便的解决的方法就是将这两种结局合二为一，如改为“出现死亡或慢性肺部疾病”。

在观测连续性数据时也会遇到另外一些棘手问题，如随访结束时只能得到存活受试者的结局观察数据（如观察卒中后病人的活动能力或生存质量）。可审慎使用两种不成熟的方法：（a）将死亡受试者活动能力估计为0（注意这有可能没有反映出受试者死亡时的真实身体状态，同时也会导致数据严重失衡；（b）只对存活受试者进行分析（但必须注明：用于幸存者的非随机对照试验结果）。结果解释时应重点考虑干预组间的死亡比例的任何差异。

16.3 整群随机对照试验

16.3.1 引言

在整群随机对照试验中，受试者以小组而非个体为单位随机分组。整群随机对照试验也被称为“组群随机对照试验”。这里的分配单位不再是个体，而是相似个体组成的“群”或“小组”。例如，“组”可以是学校、村镇、诊所或家庭。使用这种设计主要基于以下几个考虑：一是评价一项干预措施的组群效应（如疫苗的群体免疫），二是避免相同环境下的试验个体间、干预措施相互沾染（如评估饮食干预效应的试验中，可按家庭随机分配而非个体），三是整群随机设计使用简单方便。

整群设计研究中同一个群体内受试者的反应往往比较接近，就是说这些数据不再被认为是相互独立的。在许多这样的研究中，依然以个体受试者为分析单元进行分析，就会犯“分析单元错误”（Whiting-O'Keefe1984），因为此时分析单元与分配单元不同。如果在整群随机试验中，忽略群体特征而以个体作为分析单元，其结果会造成P值人为的偏小，容易得到干预措施有效的假阳性结果。同样在Meta分析中，整群特征被忽略，该研究的可信区间会过于狭窄，并被赋予不恰当的、较大的权重。这种情况也会出现在以人体局部为研究单元（如以眼睛或牙齿为研究单位），或者对一个受试者进行重复测量的试验之中，如果按研究单元（如牙齿或每次观测）进行分析，不考虑数据在受试者内部的集群特征，将会发生分析单元错误。

有关整群随机对照试验的详细介绍可参阅Murray 1995和Donner 2000等参考文献。在Meta分析中整合整群随机试验的详细讨论（Donner 2002）以及相应的技术问题的处

理方法（Donner 2001）。White和Thomas讨论了整群随机对照试验中标准化均数差进行分析的相关问题（White 2005）。

16.3.2 整群随机对照试验的偏倚风险评估

在整群随机对照试验中，比较独特的偏倚包括：(i)招募偏倚；(ii)基线不平衡；(iii)群组缺失；(iv)分析不当；(v)可比性差。

(i) 招募偏倚：当群组已实施随机分组后，再招募个体受试者到各个群组，其中可能出现招募偏倚。一个群组一旦被获知分配到“干预”组或“对照”组，会直接影响后续的受试个体招募。Farrin等发现在一项康复治疗腰背痛的整群随机试验中，大量病情轻微者，被招募到“积极治疗”组（Farrin 2005）。Puffer等也回顾分析了36个整群随机对照试验，发现其中14个（39%）可能存在招募偏倚（Puffer 2003）。

(ii) 整群随机对照试验中是一次性的、将所有群组进行随机分配，因此分配隐藏不再是一个问题。但是，因为只有少数的群体做到了随机分配，则试验组与对照组之间在群组水平及个体水平上都有可能出现基线不平衡的情况。这种基线不均衡可以通过在群体水平进行分层或配对来加以控制，尽管这还不是真正意义上的偏倚。如果报告了群体间的基线可比性或在统计分析中对基线特征进行了校正等，将有助于减少基线失衡带来的影响。

(iii) 在整群随机对照试验中有时会遇到某个群组的数据缺失，将不得不在分析中删掉此类数据。这和以个体为单位的随机对照试验中缺失结局数据一样，可能会导致偏倚。此外，在整群随机对照试验中，即使某个群体内单个个体缺失结局数据，也可能诱发偏倚风险。

(iv) 许多整群随机对照试验因未考虑资料的群组特征，所使用的统计分析方法并不恰当。例如，Eldridge等对初级保健领域的152个整群随机对照试验进行了回顾分析，结果发现有41%的研究在分析中未考虑群体（Eldridge 2004），这会产生“分析单位错误”问题，造成结果过于精确（即干预效应估计值的标准误偏小）以及过小的P值出现。当然效应量的估计仍为无偏，但若不加以纠正，则在Meta分析中会被赋予较大的权重，不允许用于对整群结果进行校正的类似方法，可参阅第16.3.6节，其中一些工作可以由系统评价作者自己来实施。

(v) 若在一个Meta分析中同时纳入整群和个体随机对照试验，或者所纳入的整群随

机对照试验中，群组类型并不统一，则需要考虑干预效应的估计值有可能不一致。例如，在研究传染病疫苗的试验中，一个社区所有人都接种疫苗将被认为比只有一半人接种的社区效果更好。Hahn等人完成的一个对髋关节保护器的Cochrane系统的评价（Hahn 2005），也说明了同样的问题。个体随机试验结果可能显示没有任何明确的收益，但整群随机对照试验却却显示出得到干预有效的强阳性结论，这可归结为整群随机对照试验中可能存在“羊群效应”（这往往发生在疗养院，接受干预的依从性会因相互间的影响而得以强化）。一般来说，这种“沾染”会导致效应的低估。因此，若一项可能存在沾染的非整群随机试验结果显示干预有效，那么基本可以断定干预效应是存在的，但效应大小很可能被低估。沾染和“羊群效应”在不同类型的群体中可能有不同表现。

16.3.3 整群随机对照试验的分析方法

在整群随机对照试验中若要避免“分析单位错误”，最直接的方法就是以群组为分析单位，对每一群组采用总和测量。群组数作为样本量，具体分析过程如同个体随机对照试验分析（只不过群体变成了个体）。但这种根据群体数量和群组大小来分析，会损失大量信息，导致检验效能大幅降低。

另外，现在已经有一些统计方法，可以在个体水平上进行分析并解释数据中的群体效应。对整群随机对照试验提取的理想信息，就是利用与群组设计相匹配的统计分析方法对所要求效应测量指标的直接估计值（如比值比及其可信区间）。这些方法比较复杂，可能基于“多水平模型”、“方差成分分析”或采用“广义估计方程”（GEE）等其它技术。需要注意的是方法的选择应该恰当合适。整群随机对照试验中采用正确的分析方法得出的效应估计值及其标准误，可使用RevMan中的倒方差法（GIV）进行Meta分析。

16.3.4 整群随机试验Meta分析的近似方法：有效样本含量

遗憾的是，过去许多整群随机试验的分析方法选用不当，往往忽略群体特征，直接采用以个体为分析单位的统计方法。若属于这种情况，如果可提取以下信息则可以进行校正分析：

- 随机分配到每个干预组的群组数；或每个群组平均的大小；
- 忽略整群设计的所有个体的结局数据（如发生结局事件的个体数量或比例，或均数和标准差）；

- 群内（或组内）相关系数（ICC）的估计值。

ICC是群体内和群体间相对变异的估计值（Donner 1980）。它描述了同一个群组内个体的‘相似程度’。然而，该数据在发表文献中很难直接获取，常见的方法是从类似研究中估计得到，下列文献中提供了ICCs估计的具体范例（Ukoumunne 1999，Campbell 2000，Health Services Research Unit 2004）。与其他类型的相关系数相比，ICC的值看起来可能比较小：通常小于0.05。然而，即使是很小的ICC值也可能对可信区间的宽度产生较大影响（继而影响Meta分析中的权重），特别是在群组较大时，影响尤为明显。一些经验研究表明，较大群组的ICC往往较小（Ukoumunne 1999）。

校正方法就是使用‘有效样本含量’（Rao 1992）进行估计。整群随机试验中单个干预组的有效样本含量就是其原来的样本量除以‘设计效应校正因子’。设计效应校正因子为：1+（M-1）ICC，

M是群组大小的平均值，ICC为群内相关系数。通常在干预组间假定一个共同的设计效应调整因子。对于二分类变量，各干预组的受试者数量、阳性事件数，均除以设计效应调整因子，结果四舍五入取整后，录入RevMan中，但要注意，对于小样本试验可能不适用。对于连续性变量数据，均数和标准差保持不变，只需对样本量加以校正。

16.3.5 整合整群随机试验的实例

假如在一个整群随机试验中，随机抽取10个班的295名学生进入干预组，随机抽取另11个班的330名学生进入对照组。学生中成功的数量（忽略整群抽样）分别为：干预组：63/295；对照组：84/330

设想从一个可靠的外部资源获得群内相关系数为0.02，则该研究的群体平均大小为(295+330)/(10+11) = 29.8。设计效应调整因子为1 + (M – 1) ICC = 1 + (29.8 – 1)×0.02 = 1.576。则干预组的有效样本含量为295 / 1.576 = 187.2，对照组为330 / 1.576 = 209.4。对于各组的阳性事件发生数同样进行调整：干预组：40.0/187.2；对照组：53.3 /209.4。将上述校正后的数据依次输入RevMan，如二分类结果或连续性结果，例子中的试验数据可按以下输入：干预组：40/187；对照组：53/209。

16.3.6 整群随机试验Meta分析的校正分析：标准误调整法

16.3.4中的方法一个明显的缺点就是校正结果必须四舍五入取整。另一种较为灵活

的方法是调整增大标准误，即将原来效应估计值的标准误（忽略整群抽样的估计结果）乘以设计效应调整因子的平方根。标准误可由可信区间计算（参见第7章，7.7.7部分）。二分类或连续性结局变量的标准分析方法，可用于计算这些用于RevMan的可信区间，使用标准误调整法进行Meta分析可通过RevMan软件中的倒方差法实现。例如：干预组：63/295；对照组：84/330，其OR = 0.795 (95% CI 0.5481.154)。使用第7章（7.7.7.3部分）中的方法，可计算比值比的对数（lnOR）= –0.23，标准误为0.19。使用16.3.5中的设计效应调整因子（1.576）进行标准误校正：$0.19\times\sqrt{1.576}=0.24$。将比值比对数（–0.23）及其校正标准误（0.24）输入RevMan，使用普通倒方差结果加以分析。

16.3.7 整合整群随机试验需注意的问题

原则上，同一Meta分析中，可以同时纳入整群随机试验和个体随机试验。但应注意不同试验类型的效应估计可能存在重要差异。实施整群随机试验一般都有很好的理由，这些理由同样需要加以审查。例如在治疗传染病的试验中，干预社区内所有个体可能比只干预随机抽取的部分个体更有效，这是因为前者降低了再感染的可能性。

在系统评价中，作者需要找出所有的整群随机试验并详细说明怎样处理相应的数据。还应该进行敏感性分析以验证结果的稳定性，尤其是外部来源的ICCs（见第9章9.7）。在此过程中，建议向统计人员寻求帮助。

16.3.8 个体随机试验中的整群抽样问题

个体随机试验也可能涉及整群抽样问题。同一类型的医护人员（例如：内科医生、外科医生、护士或治疗师等）各自负责一群受试者地干预，也会发生整群抽样的问题，如Lee和Thompson等曾讨论并提出了与整群随机试验类似问题（Lee 2005a）。

16.4 交叉试验

16.4.1 引言

平行分组试验是将每个受试者分配给一种干预措施、与一种或多种替代干预措施比较。与此相反，交叉试验是将每个受试者分配到干预措施的系列中。一个简单的随机交

叉设计就是‘AB/BA’设计，在试验中受试者刚开始先接受干预A或干预B，然后分别‘交叉’到干预B或干预A。在每个阶段相当于一个平行试验，相关内容参见16.4.5。为保持一致，同样使用E和C、而不是A和B代表不同干预措施。

与平行试验相比，交叉设计有很多可能的优势，包括：（i）每个受试者都是他或她自己的自身对照，消除了受试个体间的变异；（ii）检验效能不变时，样本量只有平行试验的一半；（iii）每个受试者接受每种干预措施，有利于筛选出最佳干预措施。Senn等对交叉试验进行了系统论述（Senn 2002）。Elbourne等则介绍了交叉试验的相关Meta分析方法（Elbourne 2002），Lathyris等则给出了在系统评价中包含交叉试验的一些经验证据（Lathyris 2007）。

16.4.2 交叉试验的适用性评价

交叉试验适合于治疗措施稳定、慢性疾病的短期效果评估，如应用于缓解哮喘和癫痫的干预措施的研究。但若干预措施的效果持久、导致试验周期较长或疾病发展迅速时，交叉试验就不再适用。考虑交叉试验时应权衡利弊。交叉试验的主要问题是不能处理携带效应（干预效应随时间的变化），即干预措施的效果持久，在一时期内给予的干预措施，其产生的效应要持续到后一阶段，影响后续干预措施的效果评估。交叉试验的不同阶段间应有一个洗脱期，以减少了前一种干预对后一种干预效果的影响。另外，如果主要结局是不可逆的（如死亡率或生育研究中的怀孕率），那么交叉设计也不适用。交叉试验的另一个问题就是存在退出风险，因为与平行分组试验相比，交叉试验周期长，而处理交叉试验中缺失值的方法目前很有限。有关交叉试验风险偏倚的评故将在16.4.3讨论。

Meta分析中是否纳入交叉试验，作者应首先考虑交叉试验是否适合于所研究问题的条件和干预措施。例如，尽管在老年痴呆症领域开展交叉试验较多，但由于该病属于不可逆的退行性改变，Qizilbash等认为开展交叉试验并不恰当，因此，在Meta分析时仅纳入了第一阶段的数据（Qizilbash 1998）。其次，需要考虑的是存在严重携带效应的可能性，这主要是通过主观判断，因目前能够处理携带效应的统计方法还不成熟。此外，干预措施本身以及洗脱期的长短也是要考虑的因素。

尽管系统评价中排除交叉试验的唯一理由是研究疾病不适合交叉设计，但在很多情况下是很难或无法从交叉试验中提取合适的数据。在16.4.5部分，专门就在系统评价中纳入交叉试验给出了一些注意事项并提出了相应的建议。现首先讨论如何将第8章‘偏

倚风险’工具推广应用于交叉试验的风险评估。

16.4.3 交叉试验的偏倚风险评估

交叉试验中的主要偏倚风险包括：（i）交叉设计是否合适;（ii）是否存在携带效应;（iii）是否数据不完整、仅有第一阶段的数据；（v）分析不正确；（iv）交叉试验结果与平行试验结果相比，缺乏可比性。

（i）交叉设计的适用条件是疾病（如哮喘）相对稳定且不需长期随访。因此第一个要考虑的问题就是交叉设计是否适合研究条件。

（ii）尤其要考虑的是携带效应是否存在，即效应从一个阶段持续到下一个阶段的可能性。若存在携带效应，则干预组间效应的差异大小与接受干预措施的顺序有关，因此，效应估计将由此受到影响（一般会低估效应，向无效偏倚，容易出现假阴性结果）。

因此，若存在携带效应，交叉设计将不再适用。然而，在试验完成之前，是否存在携带效应往往是未知的。系统评价者应在试验报告中寻求有关携带效应的评估信息，但实际情况不容乐观。在一个未公开发表的系统评价（Mills 2005）中，从2000个研究中纳入了116个公开发表的交叉试验，结果仅有30%试验讨论了携带效应，但只有12％的研究报告了分析结果。

（iii）假如存在携带效应，常规的处理方法是仅将第一阶段的数据纳入Meta分析。虽然交叉试验的第一阶段实际是一个平行分组试验，但只使用第一阶段的数据，可能会产生偏倚，特别是在携带效应被证实存在的情况下；同时这样的处理也破坏了交叉设计本身，受试个体间变异将不受控制。这种“两阶段分析”结果的可信度虽严重下降（Freeman 1989），但仍在普遍使用。

交叉试验若只能得到第一阶段数据，应充分考虑其发生偏倚的潜在风险，尤其是当研究者明确说明使用了“两阶段分析”策略。

（iv）交叉试验的统计分析应体现自身配对的优势，使用一些类似配对分析的方法（Elbourne 2002）。大多数情况下，尽管研究者采用了配对分析方法，但结果未报告或报告不充分，系统评价者无法提取配对数据。虽然可以获取一些非配对数据，但一般与所估计的效应或统计学显著性无关。因此，尽管这算不上偏倚的来源，但在Meta分析中通常会导致其权重比其应有的权重低的多。在上述的系统评价中（Mills 2005），116个交叉试验中只有38%的研究进行了配对分析。

（v）若携带效应不存在，交叉试验应像平行试验一样得出相同的治疗效应。尽管有一项研究显示交叉试验和平行试验的治疗效应存在差异（Khan 1996），但他们研究的是不孕症的干预措施，而不孕症其实是不适合做交叉试验的，数据经再次分析后，所得结果并不支持原有结论（te Velde 1998）。

交叉试验中需要考虑偏倚风险的其它问题还有：

- 接受第一阶段治疗后受试者可能会退出试验，不接受第二阶段治疗。在分析时一般将其剔除。
- 试验的两个阶段可能存在系统性差异，若阶段效应不是太严重，可以认为它对两种干预措施的影响是同等的，但它同时也提示前后两个阶段的研究条件不稳定。
- 干预措施数量或研究阶段数可能不清楚。如Lee等纳入了64个交叉试验，结果有12个具体的设计信息无法核实（Lee 2005b）。
- 交叉试验中干预措施的顺序不随机。偶尔会遇到一些研究的所有受试者接受干预措施的顺序完全一致，这种实验无法进行干预措施间的有效比较，这是因为除了干预效应外，还可能叠加结局指标的时间趋势变化。
- 退出的报告率低，特别是那些完成了一个治疗阶段的受试者。Lee 的系统评价，所纳入的64个研究中只有9个说明了受试者的退出数量（Lee 2005b）。

评价交叉试验的偏倚风险时可尝试回答下列问题：

- 使用交叉设计是否合适？
- 接受干预措施的顺序是否随机？
- 是否可认为该研究无携带效应?
- 无偏倚数据是否可及？

16.4.4 交叉试验的分析方法

若携带效应或时间效应都不存在，那么对两阶段两种干预措施的交叉试验的连续性数据可采用配对t检验分析。以受试个体为单位，计算差值，即用每个受试者的‘试验干预（E）测量值’减去‘对照干预（C）测量值’。这些测量差值的均值和标准差是估计效应量和进行统计检验的基础。在RevMan中可利用普通方差倒置法进行Meta分析。

只要能获取下列任一方式中的相关数据便可进行配对分析：

- 通过查阅文献或联系试验实施者处获取每个受试者数据；
- 受试者试验干预（E）和对照干预（C）测量差值的均数和标准差（或标准误）；
- 差值均数和下面其中之一：(i)配对t检验的t统计量；(ii) 配对t检验的P值；(iii) 差值可信区间；
- 可提取个体病人配对测量值的描述试验干预（E）和对照干预（C）测量值的图表。

详细信息请参考 Elbourne等相关文献（Elbourne 2002）。

若可及结果中包含每个受试者接受干预措施的顺序信息，则可直接进行分析来校正时间效应（如第3章Senn所述，Senn 2002）。

16.4.5 将交叉试验纳入Meta分析的方法

不幸的是，交叉试验的报告方式各种各样，并且Meta分析所需的配对数据却少有报告。一般只能得到试验组和对照组各自测量值的均数和标准差（标准误）。在Meta分析中，纳入交叉试验结果的一个简单方法就是把该试验看成E和C比较的平行试验，然后分析各阶段E和C的测量结果。这种方法会产生分析单位偏倚（见第9章，9.3部分），应避免使用，除非可证明该方法的估计结果与配对分析结果（如16.4.4部分所述）十分接近。因为忽略配对数据的特征，按照完全随机的数据进行处理可能导致可信区间变得更宽、相应权重将会很小、重要的临床异质性也可能被掩盖等。尽管如此，这种方法虽不合理，但估计比较保守，研究结果是被低估而不是高估。但也有人坚持认为可以这种方式纳入交叉试验，缘于这种分析单位错误没有其他类型的分析单位错误严重。

纳入交叉试验进行Meta分析的第二种处理方法就是只纳入第一阶段的数据。如果认为交叉试验中携带效应比较严重或存在其他原因认为交叉设计不适用，那么只纳入第一阶段的数据进行分析是可行的。但如果是研究者发现携带效应有统计学意义后，而放弃其他阶段数据，这时所纳入的第一阶段数据可能是有偏数据子集，代表性差。

纳入未恰当报告的交叉试验进行Meta分析的第三种处理方法就是通过填补缺失标准差进行近似配对分析。这种方法将在16.4.6部分详述。

交叉试验中二分类变量结果处理方法较为复杂，需要有统计学家帮助（Elbourne 2002）。

16.4.6 Meta分析中交叉试验的近似分析

表16.4.a展示了一些可从交叉试验报告中获得的数据，表中的符号也会在本部分的下面内容中使用。利用这些数据可对交叉试验进行直接近似估计，以获得Meta分析所需的均数差或标准化均数差。系统评价者在Meta分析中应考虑缺失数据填补法是否优于交叉试验完全排除法。这种取舍主要依赖于填补数据的可信度以及不同填补数据下Meta分析结果的稳定性。

表16.4.a 从交叉试验报告中可能获得的一些数据

相关数据	重要统计量	相关的常用统计量
干预E	N, M_E, SD_E	M_E的标准误
干预C	N, M_C, SD_C	M_C的标准误
E 和 C的差值	N, MD, SD_{diff}	MD的标准误
		MD的可信区间
		配对*t*统计量
		配对*t*检验的P值

16.4.6.1 均数差

与平行分组试验分析一样，一般可直接得到配对均数差的点估计值，即差值均数等同于均数之差：$MD = M_E - M_C$。其标准误为：$SE(MD) = \frac{SD_{diff}}{\sqrt{N}}$。

N是试验中受试对象的数量，SDdiff 是E和C的受试对象测量值差值的标准差。如16.4.4部分所示，标准误也可直接从MD的可信区间中得到，或从配对t统计量、配对t检验的P值中获取。然后将MD和SE（MD）的数值输入RevMan中采用普通方差倒置法进行分析。

当无法直接得到标准误且没有报告差值的标准差时，一个简单的方法就是对标准差进行填补，这与16.1.3部分处理其它缺失标准差方法无异。Meta分析中也可借用其他研究的差值的标准差，只要研究性质类似、测量尺度相同等即可。同其他填补缺失值的方法一样，也应进行敏感性分析以评价填补数据对Meta分析结果的影响（见16.1和第9章9.7）。

若从任何研究中都无法得到均数差值标准差的信息，可通过假定一个特殊的相关系数来估算标准差。该相关系数描述单个受试者干预E和C测量值之间的相似性，相关系数变化范围在–1和1之间。在交叉试验中，该相关系数的期望值在0到1之间，这是因为一个受试者在E中测量值高于平均结果，那么与其在C中测量值也很可能高于平均结果。若相关系数为0或负数，那么与平行分组试验相比，交叉设计已无任何统计学优势。

鉴于交叉试验结果的报告形式大多类似平行分组试验，分别报告各干预组的标准差（SDE和SDC，见表16.4.a）。那么差值的标准差可利用各干预组的标准差和相关系数（Corr）通过下列公式进行估计：

$SD_{diff}=\sqrt{SD_E{}^2+SD_C{}^2-(2\times Corr\times SD_E\times SD_C)}$ 。

6.4.6.2 标准化均数差

交叉试验结果的最合理报告形式是采用标准化均数差（SMD），相当于均数差除以测量标准差（注意不是差值的标准差）。SMD可由合并干预组标准差来计算，计算如下：

$SMD=\dfrac{MD}{SD_{pooled}}$，而

$SD_{pooled}=\sqrt{\dfrac{SD_E{}^2+SD_C{}^2}{2}}$ 。

SMD标准误的计算也需要相关系数：$SE(SMD)=\sqrt{\dfrac{1}{N}+\dfrac{SMD^2}{2N}}\times\sqrt{2(1-Corr)}$ 。

另外，使用估算的相关系数，也可将MD转化为SMD：$SMD=\dfrac{MD}{SE(MD)\times\sqrt{\dfrac{N}{2(1-Corr)}}}$

这种情况下，估算的相关系数不仅对标准误有影响，同时可影响SMD自身效应强度的估计（如16.4.6.1部分MD的分析）。

16.4.6.3 相关系数估计

相关系数（Corr）值可由Meta分析中的其它研究估算得到（如下），也可用Meta分析以外的资源估算，或者基于合理假设得来。所有这些，均应进行敏感性分析，尝试不同的Corr值，判断分析使用不同Corr结果的稳定性。若Meta分析中所纳入的其它研究报告了表16.4.a中的三个标准差，就可据此估计相关系数。计算时假定第一阶段和第二

阶段干预E测量的均数、标准差相同（干预C也是一样）。

$$Corr = \frac{SD_E{}^2 + SD_C{}^2 - SD_{diff}{}^2}{2 \times SD_E \times SD_C}$$ 。

填补之前，建议尽可能使用多个研究数据来估算相关系数并进行比较。若这些相关系数差别较大，实施敏感性分析就显得尤为重要。

16.4.6.4 实例

假设一个交叉试验报告了如下数据：

干预E （样本含量10）	M_E = 7.0， SD_E = 2.38
干预C （样本含量10 ）	M_C = 6.5， SD_C = 2.21

计算均数差以及估算差值的SD（SD_{diff}）

均数差MD = 7.0 – 6.5 = 0.5。假设从其它试验中得到差值的标准差通常是2。进而可估计MD的标准误：

$$SE(MD) = \frac{SD_{diff}}{\sqrt{N}} = \frac{2}{\sqrt{10}} = 0.632$$ 。

可将0.5和0.632作为均数差及其标准误依次输入RevMan，使用普通方差倒置法加以分析。

计算均数差以及估算相关系数（Corr）

均数差的估计值仍为0.5，假设已估算相关系数为0.68，可估算差值的标准差为：

$$SD_{diff} = \sqrt{SD_E{}^2 + SD_C{}^2 - (2 \times Corr \times SD_E \times SD_C)}$$

$$= \sqrt{238^2 + 2.21^2 - (2 \times 0.68 \times 2.38 \times 2.21)} = 1.8426$$

那么，MD的标准误为： $SE(MD) = \frac{SD_{diff}}{\sqrt{N}} = \frac{1.8426}{\sqrt{10}} = 0.583$

可将0.5和0.583作为均数差及其标准误依次输入RevMan，使用普通方差倒置法分析。注意相关系数0.68应进行敏感性分析。

计算标准化均数差以及估算相关系数（Corr）

标准化均数差可直接从数据中计算：

$$SMD=\frac{MD}{SD_{pooled}}=\frac{MD}{\sqrt{\frac{SD_E{}^2+SD_C{}^2}{2}}}=\frac{0.5}{\sqrt{\frac{2.38^2+2.21^2}{2}}}=0.218$$。

那么，SMD的标准误为：

$$SE(SMD)=\sqrt{\frac{1}{N}+\frac{SMD^2}{2N}}\times\sqrt{2(1-Corr)}=\sqrt{\frac{1}{10}+\frac{0.218^2}{20}}\times\sqrt{2(1-0.68)}=0.256$$。

可将0.218和0.256作为标准化均数差及其标准误依次输入RevMan，使用普通方差倒置法分析。

同样，也可利用MD及其标准误计算SMD：

$$SMD=\frac{MD}{SE(MD)\times\sqrt{\frac{N}{2(1-Corr)}}}=\frac{0.5}{0.583\times\sqrt{\frac{10}{2(1-0.68)}}}=0.217$$

由于两种计算标准化合并标准差方法的公式稍有不同，标准化均数差也有轻微差别。

16.4.7 纳入交叉试验应注意的问题

原则上，在同一Meta分析中交叉试验可与平行分组试验合并。但应考虑到不同试验类型的其它重要特征有可能存在重要差异。例如，交叉试验的干预周期可能比较短或纳入的受试者病情较轻。建议单独对平行试验和交叉试验进行Meta分析，而无论两者是否能够合并在一起。系统评价者应明确陈述如何处理交叉试验中的数据，同时应进行敏感性分析以观察其结论的稳定性，在相关系数从外部资源获得的情况下，敏感性分析尤为重要（见第9章，9.7部分）。相关统计处理最好寻求统计学家的帮助。

16.5 多个干预组的研究

16.5.1 引言

在随机临床试验中，受试者随机进入多个处理组还是比较常见的。例如2000年12月发表的一篇关于随机试验的系统评价发现：四分之一的随机试验有两个以上干预组（Chan 2005）。例如，一个试验可能有两个或多个试验干预组和一个共同对照组，或有

两个对照组，如安慰剂组和标准治疗组。这类研究统称为“多个干预组研究”或“多臂试验”。析因试验设计就是一个典型例子，它将两个或多种干预按照组合设计成四组或更多组别进行同期比较（见16.5.6部分）。

尽管在一个系统评价中可对多个干预进行比较（并因此进行多个Meta分析），几乎所有的Meta分析都是进行的两两比较。当遇到多个干预组的研究时，如何处理应考虑以下三个独立的问题：

1. 确定哪些干预组与系统评价相关。
2. 确定哪些干预组与某一Meta分析相关。
3. 若多个干预组均有关，确定以何种形式将该研究纳入Meta分析。

16.5.2 确定哪些干预组与系统评价相关

对于某一多臂试验，与系统评价相关的干预组都是那些可进行两两比较的干预组，若对这些干预组单独分析，也应符合系统评价的纳入标准。如在一个研究戒烟效果的系统评价中，目的是比较尼古丁替代疗法和安慰剂的戒烟效果，可能查找到‘尼古丁口香糖VS行为疗法VS安慰剂口香糖’比较的三臂研究。这三个干预组组合的两两比较中，只有一组（‘尼古丁口香糖VS安慰剂口香糖’）与系统评价的目的有关，行为疗法组就与系统评价无关。但是，假如该研究是比较‘尼古丁口香糖+行为疗法VS行为疗法+安慰剂口香糖VS单用安慰剂口香糖’，那么前两种干预的比较就与系统评价相关，而与安慰剂口香糖组无关。

举一个有多个对照组的例子，在一个涉及‘针灸VS无针灸’比较的系统评价中，可能查找到‘针灸VS假针灸VS无干预’比较的研究。一方面，系统评价者会问是否应纳入‘针灸VS假针灸’比较结果，另一方面，‘针灸VS无干预’的比较结果是否也应纳入？若这两种情况都纳入，那该研究的所有三个干预组都与系统评价有关。

为避免读者对各纳入研究的特征和性质产生混淆，一般建议在纳入研究特征表中，‘干预组’单元格或‘备注’单元格应给出多臂试验中所有的干预组。对那些与系统评价相关的干预组以及用于分析的干预组应详加描述。

当判断特定的Meta分析中应纳入哪些干预组时，需要考虑同样的相关性问题。鉴于一个Meta分析只能处理一种两两比较，系统评价者应考虑某研究中各种可能的两两比较干预是否都应纳入该Meta分析。为了区别系统评价水平的决策与Meta分析水平的决策，仍以‘尼古丁替代疗法VS安慰剂VS其他比较’的系统评价为例，‘尼古丁口香

糖VS行为疗法VS安慰剂口香糖’研究中的所有干预组都可能与系统评价相关。但多个干预组的同时存在可能不会影响Meta分析，这是因为‘尼古丁口香糖VS安慰剂口香糖’和‘尼古丁口香糖VS行为疗法’很可能用于不同的Meta分析之中。相反地，倘若一个Meta分析中同时纳入‘针灸VS假针灸’的研究和‘针灸VS无干预’的研究，那么该‘针灸VS假针灸VS无干预’研究中的所有干预组均可进入同一个Meta分析。对于后者的处理方法将在16.5.4部分描述。

16.5.3 评价多个干预组研究的偏倚风险

选用何种统计分析方法应在设计阶段加以确定，若在看到数据之后再临时选定统计分析方法，就有可能导致偏倚风险。例如，同一干预措施的不同剂量组，研究者看到包括P值的数据分析结果后，决定将不同剂量组合并成一组加以分析；或者在不同干预组两两比较时，可能会有选择地只报告有统计学意义的结果。这些做法均存在偏倚风险。

Juszczak 等对60个多种干预的随机试验进行系统综述，发现超过三分之一的研究设置了至少四个干预组（Juszczak 2003）。但只有64%的研究报告了针对所有结局的比较的结果，表明存在选择性报告结果的可能。另有20%的研究在分析时合并了干预组。尽管如此，倘若研究中报告了每个干预组的汇总数据，就可以实施Meta分析了，至于如何合并干预组就变得无关紧要；系统评价者不需要像试验研究者那样分析数据。

评价多臂试验的偏倚风险可以采用如下推荐的问题：

- 当受试者是随机分组时，是否相应报告了各组的数据？
- •对于某些结局指标，干预组间比较的结果，是否不存在选择性报告的可能性?

若第一个问题回答为‘是’，那么第二个问题就不重要了（因为答案很可能也为‘是’）。

16.5.4 如何从一个研究中纳入多个干预组

在Meta分析中处理纳入多臂试验的方法可能很多，但当研究具有一个或多个共同的干预组时，应避免将多臂试验简单地分拆为多个双臂试验纳入Meta分析，这样可导致重复计数、人为扩大样本量，并由由于多重比较所得干预效应估计值间未知的相关性产生分析单位错误（见第9章，9.3部分）。注意区分两种情况，一种是多臂试验中无

共同干预组，任意两两比较间是相互独立的，不存在分析单位偏倚；另一个情况是多臂试验中具有共同的干预组，因此有共同的受试者两两比较间不独立，就存在一定程度的相关性，就可能出现分析单位偏倚。例如，假设将一个研究随机受试者分为四组：‘尼古丁口香糖’、‘安慰剂口香糖’、‘尼古丁贴片’、‘安慰剂贴片’。若Meta分析的问题是一个较为宽泛的问题——尼古丁替代疗法是否有效，那么该Meta分析可纳入‘尼古丁口香糖VS安慰剂口香糖’比较，也可纳入‘尼古丁贴片VS安慰剂贴片’比较。就像来自不同的研究一样，只要是相互独立的比较，将其纳入Meta分析一般是可行的，但需要注意可能会涉及采用随机效应分析（见16.5.5部分）。

当研究具有一个或多个共同的干预组时，为避免分析单位误差，可选择下列方法：

- 分组合并、将多臂试验转换成双臂试验（推荐使用）。
- 排除其他干预组、仅选择两个干预组比较。
- 把共同组拆分成两个或多个小样本组，纳入两个或多个（相互独立）比较Meta分析比较。
- 纳入两个或多个相关比较并对相关性做出解释。
- 进行多臂试验的Meta分析（见16.6部分）。

在多数情况下，推荐的方法就是将所有相关的试验干预组合并为一个组，将所有相关的对照干预组合并为一个对照组。举一个例子，假设一个关于‘针灸VS无针灸’的Meta分析将会考虑既纳入‘针灸VS假针灸’研究，也纳入‘针灸VS无干预’的研究，那么比较‘针灸VS假针灸VS无干预’的研究，将‘假针灸’与‘无干预’组合并后可纳入Meta分析。这种合并后的对照组就可按照常规方法与‘针灸’组比较。对于二分类变量结果，可将组间的样本含量和事件发生人数分别合并。对于连续性变量结果，可使用第7章（7.7.3.8部分）描述的方法合并均数和标准差。

另一替代方法就是只摘选其中两个干预组进行比较（如上例中只选择‘假针灸’或‘无干预’作为对照组），但这样做会损失信息同时有选择相关结果之嫌，因此一般不作推荐。

第三种处理方法是将共同干预组均匀分割成多个小样本组后，然后将所有两两比较组合纳入Meta分析。例如，一个三臂试验比较了121例接受针灸的患者、124例接受假针灸的患者和117例无针灸的患者，那么先将共同组121例分成两个小组，分别包括61例患者和60例患者，然后组合成两对比较纳入Meta分析，即61例‘针灸’VS 124例‘假针灸’比较，以及60例‘针灸’VS 117例‘无针灸’比较。对于二分类变量结果，

事件发生数和总数都可以如此分割。对于连续性变量结果，只将受试者总数进行分割，原来的均数和标准差保持不变。但注意，这种方法只能部分消除、不能完全克服分析单位错误（这是因为分割产生的比较仍然存在关联），所以一般也不推荐这种方法。尽管如此，这种方法的潜在优势就是可以用来近似估计多种干预间的异质性（如在本例中使用假针灸和无干预作为对照组的差异）。

除此之外，还有两个最后的选择，但都需要统计学的支持，一个是在分析中要考虑同一研究中相关的比较结果之间的相关性，另一个是进行多臂试验Meta分析。前者需要在考虑比较间相关性的基础上，计算相关的两两比较结果的平均值（或加权平均值）以及方差（和权重）。与上述合并方法相比，该法通常也可获得相似结果。多臂试验Meta分析详见16.6部分。

16.5.5 多臂试验中的异质性考虑

处理研究间异质性有两种方法，一是使用随机效应Meta分析，二是通过亚组分析或Meta回归分析（第9章，9.6部分）。当这些分析涉及多臂试验时，也会产生其它一些问题。首先，若使用16.5.4部分推荐的方法合并干预组，那么与干预有关的一些异质性来源将无法进行分析。如在Meta分析之前已将‘假针灸’和‘无干预’合并为一组，将不能再按照‘假针灸’或‘无干预’组单独进行亚组分析。分析和探讨异质性的最简单方法就是在研究中建立多对两两比较组（如‘针灸VS假针灸’和‘针灸VS无干预’），尽管如此，若这些比较中包含共同干预组（这里是针灸），这些比较间的独立性将不再满足，即使将共同干预组分割成多个组，仍会产生一定程度的分析单位错误，不过，将共同干预组分割仍不失为一种研究异质性的可行方法。

若将同一研究中的多对比较纳入随机效应Meta分析时，又会产生另一个疑问。随机效应Meta分析允许研究间变异的存在，它是假定Meta分析中各研究效应量可以不同、但服从某一特定分布（一般为正态分布）。但是，如果两个或多个效应量估计值来自同一个多臂试验，那就可以认为研究内部和研究之间的不同比较具有相同的变异，而随机效应Meta分析时却将它们当作不同的独立研究来处理，不论比较间是独立的还是相关的（见16.5.4部分）。克服这一缺点方法之一就是对同一研究内的不同比较使用固定效应Meta分析，对不同研究间的比较使用随机效应Meta分析。建议请统计学家帮忙，实施相应地分析。事实上，采用不同模型的分析结果往往相差不大。

16.5.6 析因试验

在析因试验中，可同时针对2种（或多种）干预措施进行研究。因此，如受试者可随机接受阿司匹林或安慰剂，也可随机接受行为干预或标准护理。多数析因试验都像这样具有两个‘因子’，每个因子包含2个水平，这就是经典的2×2析因试验。偶尔也会遇到3×2试验或同时研究3种、4种或多种干预的析因设计。在系统评价中往往利用有限信息，所有的比较中常只有一个与系统评价有关。下面将重点讨论2×2析因试验，但其原则同样适用于更复杂的研究设计。

多数析因试验的目的就是达到做一个试验获得两个试验的效果。常假定不同干预措施的效果是相互独立的，也就是不存在交互作用（协同作用）。偶尔也会专门开展观察两种治疗是否存在交互作用的临床试验，但初衷往往是探索试验中两种积极治疗各自的效果及其相结合的效果，不设安慰剂组。这种试验设计严格来讲不属于析因试验。

2×2析因设计可用2×2表来表示，行表示一个比较（如阿司匹林VS安慰剂），列表示另一比较（如行为干预VS标准护理）：

		B 随机	
		行为干预（B）	**标准护理（非 B）**
A 随机	**阿司匹林(A)**	A 和 B	A，非 B
	安慰剂(非 A)	B，非 A	非 A，非 B

2×2析因设计可看作是研究不同干预措施的两个平行分组试验，相关结果可同时报告，既可以观察到阿司匹林VS安慰剂的结果（包括所有受试者不管他们是否接受行为干预或标准护理），又可观察到行为干预的结果。这些结果往往以2×2表的周边合计展示（行、列的合计数），同时还可进一步探讨疗法间的交互作用（即是否接受B或‘非B’时A的效果）。对此，表内四个格子内的结果应同时关注（McAlister 2003）。需要注意的是，如果将2×2析因设计试验分拆成两个平行试验，有可能发表两个独立的试验报告，倘若发表在不同杂志上，将无法看到全部结果。

McAlister等综述了44个已发表的析因试验报告（McAlister 2003），发现只有34%报告了全部四个格子的结果。在同一系统评价中59%的试验报告了交互作用的检验结果。再次分析后发现，2/44（6%）试验的交互作用的检验结果为P<0.05，可以归结为机遇所致（McAlister 2003）。因此，尽管担心无法识别的交互作用，研究者大多将析因设计限制应用在那些治疗间并无潜在的实质性交互作用的情况。遗憾的是，很多系统评价

者在Meta分析中并未利用考虑很少存在交互作用的实际情况，只纳入了一半的可及数据（如只纳入了不接受B中有关A VS 非A的比较数据，排除了接受B中A VS 非A的比较结果）。

评价析因试验的偏倚风险时推荐使用如下问题：

- 不同干预效果间是否不存在重要交互作用？

16.6 间接比较和多臂试验Meta分析

16.6.1 引言

可替代性干预措施间的相互比较，可能是Cochrane干预性系统评价的关注重点或是其次要研究目的，同时也是对Cochrane系统评价进行再评价的主要标志，再评价是指对同一领域、多个Cochrane干预性系统评价，尤其是同样条件下不同干预措施的系统评价再次进行综合分析（参见22章）。多个替代干预措施的直接比较可以采用随机对照试验加以实现，但此类研究并不多见。间接比较，顾名思义，就是将两个无直接比较结果的竞争性的干预措施放在一起分析，见16.6.2。多臂试验Meta分析（MTM）可以看作是一种广义的间接比较方法，它既允许把直接和间接比较合并在一起分析，又可同时分析多种干预措施相比较的效果，见16.6.3。

16.6.2 间接比较

当缺乏直接比较的随机对照试验时，干预措施间可采用间接比较。例如，假设一些试验比较了‘营养师和医生’给出的饮食干预效果，而另外一些试验则比较了‘营养师与护士’，但没有试验对‘医生和护士’的饮食干预效果进行直接比较。事实上，通过‘营养师VS医生’试验以及‘营养师VS护士’试验结果间的间接比较，也可实现医生和护士饮食干预效果相互比较的目的。

注意千万不要将试验中相关组别拿来直接比较，这样做会破坏试验原有的随机性，产生类似两个独立队列比较的选择性偏倚（通常是极端的）。如在上述实例中，将‘营养师VS护士’试验中的接受护士建议病人，直接与‘营养师VS医生’试验中的接受医生建议病人相互比较，这样的错误应尽量避免。

间接比较有多种方法，但需仔细考虑各方法的基本假设。其中较为简单的一种方

法就是进行亚组分析，假设试验条件、类型、对象基本相同，按不同的比较定义为不同的亚组。对于典型的两个亚组（三种干预、两对比较;）的分析，可使用Bucher介绍的简便方法估计亚组间的差异并确定其统计学意义（Bucher 1997）。如上述实例，两个试验分别当作两个亚组处理，一个亚组是‘营养师VS医生’，另一个亚组是‘营养师VS护士’。两个亚组合并效应的差值即为所需比较‘医生VS护士’的效应量估计，该分析可在RevMan中使用亚组间差异检验来实现（见第9章9.6.3.1部分）。间接比较结果是否真实可靠，取决于不同亚组来源试验的相似程度，其实就是那些可能影响结局观察的相关因素的相似程度。有关间接比较的更多内容，可参考Song和Glenny发表的相关文献（Song 2003，Glenny 2005）。

间接比较本身不是随机化比较，不能按照随机来解释。类似观察性研究，间接比较本质上属于不同试验的结果观察，同样会受到观察性研究有关偏倚的影响，例如，可能存在混杂因素，具体处理，可参见第9章9.6.6部分。在系统评价过程中，若同时存在直接比较和间接比较，且直接比较试验不存在设计缺陷，建议直接比较和间接比较分开进行，最终结论的形成还是要基于直接比较的结果。

16.6.3 多臂试验Meta分析

在一个Meta分析中可以同时分析比较三种或多种干预措施，这通常是指‘多臂试验Meta分析’（‘MTM’），‘网络Meta分析’或‘综合干预比较’（‘MTC’）Meta分析。多臂试验Meta分析可用于多个干预组研究的分析处理，并可用于不同干预比较研究的合并分析。Caldwell等对此提供了一个简易读本（Caldwell 2005），更详细的内容介绍可参考Salanti等发表文献（Salanti 2008）。注意多臂试验Meta分析均保留每个干预的特征，允许多种干预间相互比较。这与16.5部分讨论的处理单个研究中多组比较的合并方法不同，该法通过合并方式，将多组比较最终转换成两组比较，原来的干预分组不再保留。

多臂试验Meta分析，最简单的应用就是16.6.2部分描述的间接比较。三种干预（如营养师的建议、医生的建议和护士的建议）中的任意两种均可与第三方进行间接比较。例如，医生和护士干预效果相比，可通过‘营养师VS医生’试验和‘营养师VS护士’试验的间接比较加以实现。这种分析还可以多种方式进一步扩展延伸，用于解决其他类似问题。例如，倘若有‘医生VS护士’直接比较的临床试验，那么可以实现直接效

应和间接效应的合并。倘若超过三种干预，将会同步进行多个直接比较和间接比较。

若纳入研究均只比较了两种干预，也可通过亚组分析来实现多臂试验Meta分析，亚组分析差别检验详见第9章（9.6.3.1部分）。尽管如此，最好使用随机效应模型以允许各亚组间存在异质性，此时也可使用Meta回归，参见第9章（9.6.4部分）。倘若研究中干预组不止两个，标准的亚组分析和Meta回归将不再适用，最好选择多元Meta分析方法来进行数据合成，或者使用WinBUGS中的贝叶斯框架（见16.8.1部分）。贝叶斯框架的独特优势在于使用概率的而不是粗略的排序方法、处理所有干预措施。

多臂试验Meta分析特别适合于系统评价的再评价（第22章）。需要注意的是，多臂试验Meta分析的假设比较苛刻：除比较的干预措施不同外，要求纳入的临床试验在其他所有方面应尽量相似。另外，间接比较不是随机比较，可能存在观察性研究中的常见偏倚，如混杂偏倚等（见第9章9.6.6部分）。若系统评价中可同时进行直接比较和间接比较，要注意多臂试验Meta分析只能作为补充分析使用，不能完全替代直接比较。当然，实施多臂试验Meta分析不仅需要统计专家的专业知识，还需要他们的经验。

16.7 多重比较及机遇的作用

16.7.1 引言

系统评价过程中常会涉及多重比较的问题。例如，同时有多个结局观察指标，同一结局指标重复测量多次，进行多个亚组分析，或多个干预措施相互比较等等，这些均需要进行多重比较。然而，比较的次数越多，越容易出现假阳性错误（由机遇导致的‘统计学显著性’）。按传统的检验水准0.05，即使比较的干预组间实际上没有差别，也有1/20机会出现有统计学意义的结果。随着假设检验数量的增加，犯假阳性错误率会大幅提高，即使实际上没有效应差别，倘若进行14个假设检验，就会有超过50%（51.2%）机会检验出至少一个有差别。同时，多重独立性假设检验可能出现假阳性的机会更多一些。例如，一个临床试验有多个结局指标，由于测试的对象相同，各指标间存在一定程度的关联，独立性不满足。由于不同亚组的对象不同，不同亚组间的多重分析，就属于多重独立性假设检验，产生的假阳性问题要比多指标多重比较更为严重。

临床试验、流行病学、公共卫生研究（Bauer 1991，Ottenbacher 1998）及系统评价（Bender 2008）中都会涉及多重比较显著性检验的问题。关于多重比较问题已有大量

统计文献可供参考，许多统计方法可以用来校正各种情况下的多重检验结果（Bender 2001，Cook 2005，Dmitrienko2006）。然而，在何时考虑多重性问题、以及采用何种方法校正，尚未形成共识。例如，当多个假设检验不独立时，使用独立性多重检验的校正法，将会导致P值过大。对于证实性临床试验，由于假设事先已确定，使用多个假设检验时、为避免假阳性错误，必须对检验结果进行校正（Koch1996），相应要求已被写入有关统计指南（CPMP Working Party on Efficacy of Medicinal Products 1995）。而探索性研究，由于没有预先设定重要的假设，不需要也无法对其进行多重检验校正（Bender 2001），因此，探索性研究中有统计学意义的结果，不管是否进行了多重检验校正，均应理解为‘假设产生’。

16.7.2 系统评价中的多重比较

系统评价中一般不涉及多重检验校正问题，通常也不作推荐使用。然而，与其它研究类型一样，系统评价中也经常遇到多重比较问题。系统评价作者应记住，Cochrane系统评价目的在于估计干预措施的效应，重点不是对其进行假设检验，当然，多重比较问题肯定会影响到区间估计和假设检验（Chen 2005，Bender 2008）。

系统评价过程中也会碰到与多重比较有关的其它问题。例如，一项研究结果的报告中，通常不会报告其做了多少个检验或分析，一些研究很有可能有选择性地将那些有统计学意义或者所关注的结果拿来发表，至于那些‘无趣的’结果则被省略，这种选择性报告，会得出一些误导的结果和虚假结论，这方面内容详见第8章（8.13部分）。

理想情况，是在设计阶段就应有合理的统计假设检验分析计划（包括任何对多重检验的校正）。但这对于系统评价来说难度很大，这是因为在开始阶段，并不知道从纳入研究中可得到哪些结果和效应指标，使得在系统评价中预先计划多重检验方法变得不可行。况且目前只有一些针对单个研究的多重比较方法能够用于Meta分析，供系统评价使用的多重比较方法尚需进一步研究（Bender 2008）。

总之，对于系统评价中的多重假设检验和多重区间估计问题，目前还没有简单的或完全令人满意的解决方法。但是，以下基本建议可供参考，更多详细建议可参见有关文献（Bender 2008）。

- 在系统评价计划书中说明主要分析结果指标及其方法有哪些（越少越好）。应事先按主要和次要对结局指标进行分类，在‘结果汇总表’中应事先设定主要结

果指标。若主要假设已知，并能够实施多重检验校正，这样的校正将会大大增强所得结论的可信度。

- 尽管Cochrane系统评价要求纳入尽可能多对使用者有用的结局观察指标，但进行多重分析，将很难形成最终的结论。下结论时应牢记，即使组间实际上无差别，机遇也会造成大约1/20假设检验有统计学意义（检验水准为0.05）。
- 不要根据有统计学意义的P值来选择重要结果（如在摘要中）。
- 若多次测量结局观察指标，应全面展示所有时间点的效应估计，或者选择一个较为合理的时间点（尽管要获得所有试验的合适数据本身就是一个问题）加以报告，应避免对不同时间点上的效应进行多重检验。
- 将亚组分析的数量降低到最小，并谨慎地解释结果。
- 对任何无事先假设的结果的解释，均应慎重。即使有‘统计学意义’时，这种发现也只用于产生假设，而不是证实假设。

16.8 Meta分析中的贝叶斯和分层方法

16.8.1 贝叶斯方法

贝叶斯统计是有别于“经典统计”一类分析方法，同样可以进行假设检验和置信区间估计。与仅基于样本统计量及其抽样分布的“经典统计”方法不同，它是在“先验”的基础上，利用样本信息对“先验”更新后，得到统计推断结果。贝叶斯分析中，最初的不确定性使用“先验分布”来表达，当前数据和关于数据如何生成的假设可用似然估计来总结，两者结合后得到后验分布，后验分部可用非常像经典的参数估计和置信区间的多个点估计和可信区间进行概括。贝叶斯分析不能在RevMan中进行，需要专用分析软件：WinBUGS软件（Smith 1995，Lunn 2000），该软件目前可免费下载。

在Meta分析中，使用先验分布描述所分析对的特定效应量的不确定性，如OR或MD。一般为临床试验前、对效应大小的主观判断，也可来自Meta分析纳入研究之外的证据资源，如来自非随机研究的信息。先验分布的宽度反映了效应量的不确定程度。当先验信息很少或无时，可采用‘无信息’先验分布，在可及范围内的所有效应值都等可能。似然估计包括对纳入研究的数据总结（如随机试验的2×2表）以及Meta分析模型的假定（如固定效应或随机效应模型）。

贝叶斯统计中，对使用先验分布还存在争议。尽管它可用效应的主观判断来表示先验分布，但将客观试验数据和主观判断结合在一起，看起来很奇怪。因此，Meta分析中的常见做法就是直接使用“无信息”先验分布来反映事先未知情况，对于主要比较尤其适用。当然，先验分布也可用Meta分析中其它的一些具体数值来描述，如随机效应模型中纳入研究结果的变异程度。尤其是Meta分析中若纳入的研究数量比较少时，引入对其它这些参数的主观判断或外部证据可能很有有用，注意进行敏感性分析，可以分析假设对结果的影响。

贝叶斯分析和经典Meta分析的一个区别是对区间可信度的解释不同：OR的95%置信区间就是我们相信OR值有95%的可能性包含在该区间，这也是许多研究者对经典可信区间的实际解释，但严格意义上讲，经典的95%可信区间指的是假如做无数次的区间估计，区间范围包括真值的频率是95%。贝叶斯分析还可以计算OR在特定范围内的概率，这对于“经典统计”是无法实现的。如系统评价者可确定OR小于1的可能性（这可能表示试验干预的有益效应），或者OR不大于0.8的概率（0.8可能表示一个重要的临床效应），应该说明的是，这些概率与先验分布的选择是相对应的。不同Meta分析者因使用不同的先验分布分析相同的数据，可能获得不同的结果。

与许多经典的Meta分析方法相比，贝叶斯方法有一些特殊优势。例如，贝叶斯方法可用于：

- 直接合并外部证据，例如关于干预措施的效应或研究间变异的可能程度；
- 可利用效用的概念来综合分析各种临床结局，将Meta分析扩展到决策分析；
- 允许Meta分析中各研究间方差的估计存在一定的变异（见第9章9.5.4部分）；
- 可以进行利弊综合分析（潜在风险和治疗获益）（见第9章9.6.7部分）；
- 鉴于WinBUGS软件比较灵活，可以实施较为复杂的分析（如多臂试验Meta分析）；以及
- 定量分析观察主观先验因样本数据而改变的程度（Higgins 2002）。
- 那些在Meta分析中准备采用贝叶斯方法的系统评价者，最好能向统计专家寻求帮助，以避免方法误用。已有一些有价值的文献值得一读（Sutton 2000，Sutton 2001，Spiegelhalter 2004）。

16.8.2 分层模型

一些复杂技术可为Meta分析提供一种新的统计分析模型，称为分层模型或多水平模型（Thompson 2001）。这是因为Meta分析中的数据通常来自两个层次：处在较高层次的研究，研究中处于较低层次的受试者。有时也可能涉及其它水平，如多中心临床试验中的中心层面或整群随机试验中的群层面。无论是基于汇总结果的Meta分析（如输入对数OR及其方差），还是基于病人个体病人数据（IPD）的Meta分析，分层模型都是适用的（Turner 2000）。使用随机效应模型来描述研究效应估计间难以解释的变异时，分层模结构模型尤为适合（见第9章9.5.4）。

分层模型而不是较为简单的Meta分析方法，在很多情况下都很有用。如可用于：

- 允许研究内治疗效应方差估计存在一定程度的变异；
- 允许研究间方差估计存在一定程度的变异，T2（见第9章9.5.4部分）；
- 对二分类结局变量可实现精确地模拟（而不仅是汇总统计量）；
- 研究潜在风险和治疗效益之间的关系（见第9章9.6.7部分）；
- 可同时处理研究水平特征变量（见第9章9.6.4部分）和个体水平的变量（见第18章）。

当病人个体数据（IPD）结果变量和协变量都具备时，尤其适合于分层模型分析（Higgins 2001）。然而使用此方法时仍要小心，不要将研究内关系和研究间关系相互混淆。

使用分层模型分析也需要一些复杂软件，可采用经典的统计分析软件（如SAS的mixed过程或MlwiN）或贝叶斯方法专用软件（如WinBUGS）。目前Meta分析中的许多方法学研究的分层模型分析，使用的是贝叶斯方法。

16.9 罕见事件（包括0频数）

16.9.1 罕见事件的Meta分析

对于罕见结果，Meta分析可能是获得卫生保健干预效果可靠证据的唯一方法。一般情况下，单个研究发现罕见事件结果差异的能力不足，但多个研究的Meta分析也许有足够检验效能去观察干预措施对罕见事件的发生率是否有影响。然而，Meta分析的许多方法都是基于大样本近似法，当事件较罕见时就不适用了。那么选择Meta分析方法时，作者应慎重。

事件发生是否‘罕见’，采用的判定标准可能不止一个。如可以将1/1000的风险定义为罕见事件，以相同的方法也可将1/100的风险算作罕见事件，但是，即便风险达到1/10，可能仍会受到在此讨论的问题的影响。典型的是，Meta分析中经常会遇到大部分研究的某个组或多个组事件发生为0的情况。

16.9.2 格子计数为零的研究

在单个研究中，当一个或两个组观察到无事件发生时就会出现计算问题。倒方差Meta分析方法（倒方差固定效应和D-L随机效应方法）计算每个研究的干预效应估计及其标准误。对于一个或两个组无事件发生的研究，这些计算通常涉及到除以0计数，这将导致计算错误。多数Meta分析软件（包括RevMan）可自动检查有问题的0计数，当出现此问题时会对研究结果表格中的所有计数为0的格子添加一个固定值（通常是0.5）。只有所有纳入研究中的相同格子均为0，M-H方法才需要校正0格子，因而需要校正的情况更少。然而，在许多软件中用于M-H法的校正方法与倒方差法一样。与差值法相比，OR和RR法常需0格子校正，Peto OR法除外，该法只在所有研究的所有组出现0事件这种极端情况下才会涉及校正计算问题。

使用固定校正值可以达到避免计算错误的目的，但同时也会造成结果偏向于无差别以及高估研究测量方差（由此造成其在Meta分析中的权重被低估）。当研究组的大小不等时（这种情况通常非随机研究比随机研究更为常见），将会在治疗效应估计中产生方向性偏倚。Sweeting等提出了校正值不固定的一种新方法，按照对照组样本大小倒数的一定比例进行校正，当研究组的大小不平衡时，该法优于固定的0.5校正（Sweeting 2004）。

16.9.3 无事件发生的研究

如果某个研究中两个组的事件发生数均为0时，在以OR和RR为效应量的Meta分析中，常规做法就是将其从Meta分析中排除。因为这类研究提供信息量有限，特别是无法提供有关效治疗应量的大小及或方向信息。同时，如果试验干预组和对照组的事件发生数都非常少，将无法判断哪组有较高的风险或两组的风险是否属于相同或不同的数量级（当风险极低时，相应地比值要么非常大、要么非常小）。能否可以认为样本较大组的风险可能要低一些呢（可信区间的上限同样会减低），这种看法欠妥，样本大小是由研究者决定的，同时它也与事件发生率无关。

率差（RD）法表面上看起来优于OR法，当任何一组均无事件发生时，RD可以计算（率差为0），因此，可以纳入Meta分析。Bradburn等进行模拟研究发现当事件发生数较少时，RD法估计的可信区间太宽，同时检验效能也低，得出RD法并不适合罕见事件的Meta分析（Bradburn 2007）。因为正确识别(或试图反驳)严重不良事件的能力是药物开发中的一个关键问题，因此对药物治疗安全性研究结果的正确处理显得尤为重要。许多随机试验报告中很可能会漏掉那些“无事件发生”指标及其结果，由此排除在Meta分析之外。当对研究结果进行Meta分析时，如果纳入研究未报告不良事件，一种可能是确实未发生这种事件，另一种也可能是未将这种事件作为终点测量指标，但无论哪种可能，RD法Meta分析结果将会受影响；而OR和RR法Meta分析，由于不会纳入那些无事件发生的研究，其结果不会受影响。

16.9.4 无事件发生研究的可信区间

当无观察事件发生时，可把可信区间上限作为事件风险替代值，特别是在估计严重不良事件风险时，该方法非常有用。其中“3替代”法最为常用，若一个组中无观察事件发生，那么该组事件发生数可信区间的上限就定为3，若该组样本大小为N，则风险估计为3/N（Hanley 1983）。虽然没有此种方法用于系统评价的提议或评估，但在一系列研究（包括随机试验、非随机对照试验或病例系列研究等）中如果任何干预组均未观察到严重不良反应事件发生时，使用此法估计较为合理，可将N作定义为接受干预措施的患者总例数。然而对于两组事件发生率相互比较时，不能提供任何信息，该法不适用。

数值3法与泊松分布单侧95%可信区间上限一致（或等同于双侧90%可信区间）。对于风险估计，可以采用更常见的单侧97.5%可信区间（等同于双侧95%可信区间）上限，此时选取的替代值不再是3而变成了3.7（Newcombe 2000）。另一种推荐使用的替代值为4，即‘4替代’法，风险估计的上限为：4/(N+4)。实施Cochrane系统评价时，这些方法均可推荐使用。例如，若10例观察对象无事件发生，事件发生数的可信区间上限为3.7，风险就是3.7除以10（即0.37）；若100例观察对象无事件发生，事件发生数的上限仍是3.7，风险就是3.7除以100（即0.037）。

16.9.5 罕见事件Meta分析方法的有效性

模拟研究表明许多Meta分析方法分析罕见事件时，可能给出错误的结论。这是因为大多数Meta分析方法是基于渐进统计理论的，无法处理罕发事件。强行使用这些方法，会出现一系列问题，如结果为有偏估计、可信区间过宽或检验效能太低不能发现实际存在的差别而出现假阴性结果。

下面以OR法Meta分析为例，介绍如何选择有效的统计方法。方法的选择取决于对照组的基础风险、治疗潜在效应量的大小、治疗组和对照组受试者数量是否平衡等。虽然没有研究来直接估计RR大小，但RR意义与相应的OR值非常接近，特别是当事件罕发时，OR和RR都可解释为发生概率之比。

Bradburn等发现当事件罕发时，很多最常用的一些Meta分析方法都存在偏倚（Bradburn 2007）。特别是方差倒置、OR随机效应估计法（D-L）、RD法、使用0.5校正的M-H方法，偏倚最大。正如上面所述，当事件风险较低时，RD法Meta分析估计的可信区间范围较为保守，同时检验效能较低。

当事件发生率低于1%、同时试验组和对照组样本大小比较均衡、效应量不是特别大时，Peto-OR法是偏倚最小、检验效能最高的方法，它能提供可靠的区间范围。这一发现在三种不同情况下的Meta分析中得到验证，同时Sweeting等的研究也证实了上述结论（Sweeting 2004）。

但应注意Peto-OR法只是对OR值的一种近似估计，当知到近似值很小，效应量非常大时（如RR=0.2），治疗效应会被低估。但当事件发生风险为1/1000时，Peto法仍为所有Meta分析方法中最佳的选择，误差不会超过对照组风险的6%。

在其它情况下（即事件风险高于1%，效应明显、事件风险在1%左右，许多组间不均衡的Meta分析等），较为合理的方法有未进行零格子校正的M-H-OR法、Logistic回归和确切计算法，注意这些方法目前尚不能在RevMan中实现。

罕见事件的Meta分析应避免使用方差倒置法（包括D-L随机效应法），该法基于大样本方差估计，不适合于罕见事件的处理，直接用研究的方差来估计其对Meta分析的贡献。由于D-L法是Meta分析软件中所能提供的、唯一随机效应分析法，我们建议在治疗效应研究时，可以将异质性纳入分析，但处理稀疏数据、进行效应估计时，建议重点先放在效应估计上，其次才考虑异质性处理问题。

16.10 本章信息

编辑：代表Cochrane统计学方法学组的Julian PT Higgins, Jonathan J Deeks和Douglas G Altman。

本章引用格式：Higgins JPT, Deeks JJ, Altman DG (editors). Chapter 16: Special topics in statistics. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.

作者：Doug Altman, Deborah Ashby, Ralf Bender, Catey Bunce, Marion Campbell, Mike Clarke, Jon Deeks, Simon Gates, Julian Higgins, Nathan Pace and Simon Thompson.

致谢：我们尤其要感谢Joseph Beyene, Peter Gøtzsche, Steff Lewis, Georgia Salanti, Stephen Senn和Ian White对本章初稿的有益意见。有关Cochrane统计方法学组的详细信息见第9章（方框9.8.a）。

16.11 参考文献

Abrams 2005

Abrams KR, Gillies CL, Lambert PC. Meta-analysis of heterogeneously reported trials assessing change from baseline. Statistics in Medicine 2005; 24: 3823-3844.

Bauer 1991

Bauer P. Multiple testing in clinical trials. Statistics in Medicine 1991; 10: 871-889.

Bender 2001

Bender R, Lange S. Adjusting for multiple testing - when and how? Journal of Clinical Epidemiology 2001; 54: 343-349.

Bender 2008

Bender R, Bunce C, Clarke M, Gates S, Lange S, Pace NL, Thorlund K. Dealing with multiplicity in systematic reviews. Journal of Clinical Epidemiology 2008; 54: 343-349.

Bradburn 2007

Bradburn MJ, Deeks JJ, Berlin JA, Russell LA. Much ado about nothing: a comparison of the performance of Meta-analytical methods with rare events. Statistics in Medicine 2007; 26: 53-77.

Bucher 1997

Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in Meta- analysis of randomized controlled trials. Journal of Clinical Epidemiology 1997; 50: 683-691.

Caldwell 2005

Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ 2005; 331: 897-900.

Campbell 2000

Campbell M, Grimshaw J, Steen N. Sample size calculations for cluster randomised trials. Changing Professional Practice in Europe Group (EU BIOMED II Concerted Action). Journal of Health Services Research and Policy 2000; 5: 12-16.

Chan 2005

Chan AW, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. The Lancet 2005; 365: 1159-1162.

Chen 2005

Chen T, Hoppe FM. Simultaneous confidence intervals. In: Armitage P, Colton T (editors). Encyclopedia of Biostatistics (2nd edition). Chichester (UK): John Wiley & Sons, 2005.

Cook 2005

Cook RJ, Dunnett CW. Multiple comparisons. In: Armitage P, Colton T (editors). Encyclopedia of Biostatistics (2nd edition). Chichester (UK): John Wiley & Sons, 2005.

CPMP Working Party on Efficacy of Medicinal Products 1995

CPMP Working Party on Efficacy of Medicinal Products. Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products. Statistics in Medicine 1995; 14: 1659-1682.

Dmitrienko 2006

Dmitrienko A, Hsu JC. Multiple testing in clinical trials. In: Kotz S, Balakrishnan N, Read CB, Vidakovic B (editors). Encyclopedia of Statistical Sciences (2nd edition). Hoboken (NJ): John Wiley & Sons, 2006.

Donner 1980

Donner A, Koval JJ. The estimation of intraclass correlation in the analysis of family data. Biometrics 1980; 36: 19-25.

Donner 2000

Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. London (UK): Arnold, 2000.

Donner 2001

Donner A, Piaggio G, Villar J. Statistical methods for the Meta-analysis of cluster randomized trials. Statistical Methods in Medical Research 2001; 10: 325-338.

Donner 2002

Donner A, Klar N. Issues in the Meta-analysis of cluster randomized trials. Statistics in Medicine 2002; 21: 2971-2980.

Elbourne 2002

Elbourne DR, Altman DG, Higgins JPT, Curtin F, Worthington HV, Vaillancourt JM. Meta-analyses involving cross-over trials: methodological issues. International Journal of Epidemiology 2002; 31: 140-149.

Eldridge 2004

Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. Clinical Trials 2004; 1:80-90.

Farrin 2005

Farrin A, Russell I, Torgerson D, Underwood M, UK BEAM Trial Team. Differential recruitment in a cluster randomized trial in primary care: the experience of the UK back pain, exercise, active management and manipulation (UK BEAM) feasibility study. Clinical Trials 2005; 2: 119-124.

Follmann 1992

Follmann D, Elliott P, Suh I, Cutler J. Variance imputation for overviews of clinical trials with continuous response. Journal of Clinical Epidemiology 1992; 45: 769-773.

Freeman 1989

Freeman PR. The performance of the two-stage analysis of two-treatment, two-period cross-over trials. Statistics in Medicine 1989; 8: 1421-1432.

Furukawa 2006

Furukawa TA, Barbui C, Cipriani A, Brambilla P, Watanabe N. Imputing missing standard deviations in Meta-analyses can provide accurate results. Journal of Clinical Epidemiology 2006; 59: 7-10.

Gamble 2005

Gamble C, Hollis S. Uncertainty method improved on best-worst case analysis in a binary Meta-analysis. Journal of Clinical Epidemiology 2005; 58: 579-588.

Glenny 2005

Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, Bradburn M, Eastwood AJ. Indirect comparisons of competing interventions. Health Technology Assessment 2005; 9: 26.

Hahn 2005

Hahn S, Puffer S, Torgerson DJ, Watson J. Methodological bias in cluster randomised trials. BMC Medical Research Methodology 2005; 5: 10.

Hanley 1983

Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA 1983; 249: 1743-1745.

Health Services Research Unit 2004

Health Services Research Unit. Database of ICCs: Spreadsheet (Empirical estimates of ICCs from changing professional practice studies) [page last modified 11 Aug 2004]. Available from: http://www.abdn.ac.uk/hsru/epp/cluster.shtml://www.abdn.ac.uk/hsru/epp/cluster.shtml (accessed 1 January 2008).

Higgins 2001

Higgins JPT, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. Statistics in Medicine 2001; 20: 2219-2241.

Higgins 2002

Higgins JPT, Spiegelhalter DJ. Being sceptical about Meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. International Journal of Epidemiology 2002; 31: 96-104.

Higgins 2008

Higgins JPT, White IR, Wood AM. Imputation methods for missing outcome data in Meta-analysis of clinical trials. Clinical Trials 2008; 5: 225-239.

Hollis 1999

Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 1999; 319: 670-674.

Hollis 2002

Hollis S. A graphical sensitivity analysis for clinical trials with non-ignorable missing binary outcome. Statistics in Medicine 2002; 21: 3823-3834.

Juszczak 2003

Juszczak E, Altman D, Chan AW. A review of the methodology and reporting of multi-arm, parallel group, randomised clinical trials (RCTs). 3rd Joint Meeting of the International Society for Clinical Biostatistics and Society for Clinical Trials, London (UK), 2003.

Khan 1996

Khan KS, Daya S, Collins JA, Walter SD. Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. Fertility and Sterility 1996; 65: 939-945.

Koch 1996

Koch GG, Gansky SA. Statistical considerations for multiplicity in confirmatory protocols. Drug Information Journal 1996; 30: 523-534.

Lathyris 2007

Lathyris DN, Trikalinos TA, Ioannidis JP. Evidence from crossover trials: empirical evaluation and comparison against parallel arm trials. International Journal of Epidemiology 2007; 36: 422-430.

Lee 2005a

Lee LJ, Thompson SG. Clustering by health professional in individually randomised trials. BMJ 2005; 330: 142-144.

Lee 2005b

Lee SHH. Use of the two-stage procedure for analysis of cross-over trials in four aspects of medical statistics (PhD thesis). University of London, 2005.

Lewis 1993

Lewis JA, Machin D. Intention to treat--who should use ITT? British Journal of Cancer 1993; 68: 647-650.

Little 2004

Little RJA, Rubin DB. Statistical Analysis with Missing Data (2nd edition). Hoboken (NJ): John Wiley & Sons, 2004.

Lunn 2000

Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 2000; 10: 325-337.

Marinho 2003

Marinho VCC, Higgins JPT, Logan S, Sheiham A. Fluoride toothpaste for preventing dental caries in children and adolescents. Cochrane Database of Systematic Reviews 2003, Issue 1. Art No: CD002278.

McAlister 2003

McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. JAMA 2003; 289: 2545-2553.

Mills 2005

Mills EJ, Chan AW, Guyatt GH, Altman DG. Design, analysis, and presentation of cross-over trials. 5th Peer Review Congress, Chicago (IL), 2005.

Murray 1995

Murray DM, Short B. Intraclass correlation among measures related to alcohol-use by young-adults - estimates, correlates and applications in intervention studies. Journal of Studies on Alcohol 1995; 56: 681-694.

Newcombe 2000

Newcombe RN, Altman DG. Proportions and their differences. In: Altman DG, Machin D, Bryant TN, Gardner MJ (editors). Statistics with Confidence (2nd edition). London (UK): BMJ Books, 2000.

Newell 1992

Newell DJ. Intention-to-treat analysis: implications for quantitative and qualitative research. International Journal of Epidemiology 1992; 21: 837-841.

Ottenbacher 1998

Ottenbacher KJ. Quantitative evaluation of multiplicity in epidemiology and public health research. American Journal of Epidemiology 1998; 147: 615-619.

Puffer 2003

Puffer S, Torgerson D, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. BMJ 2003; 327: 785-789.

Qizilbash 1998

Qizilbash N, Whitehead A, Higgins J, Wilcock G, Schneider L, Farlow M. Cholinesterase inhibition for Alzheimer disease: a Meta-analysis of the tacrine trials. JAMA 1998; 280: 1777-1782.

Rao 1992

Rao JNK, Scott AJ. A simple method for the analysis of clustered binary data. Biometrics 1992; 48: 577-585.

Salanti 2008

Salanti G, Higgins J, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. Statistical Methods in Medical Research 2008; 17: 279-301.

Senn 2002

Senn S. Cross-over Trials in Clinical Research (2nd edition). Chichester (UK): John Wiley & Sons, 2002.

Smith 1995

Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects Meta-analysis: A comparative study. Statistics in Medicine 1995; 14: 2685-2699.

Song 2003

Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published Meta-analyses. BMJ 2003; 325: 472-475.

Spiegelhalter 2004

Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester (UK): John Wiley & Sons, 2004.

Stewart 1995

Stewart LA, Clarke MJ. Practical methodology of Meta-analyses (overviews) using updated individual patient data. Statistics in Medicine 1995; 14: 2057-2079.

Sutton 2000

Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for Meta-analysis in Medical Research. Chichester (UK): John Wiley & Sons, 2000.

Sutton 2001

Sutton AJ, Abrams KR. Bayesian methods in Meta-analysis and evidence synthesis. Statistical Methods in Medical Research 2001; 10: 277-303.

Sweeting 2004

Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in Meta-analysis of sparse data. Statistics in Medicine 2004; 23: 1351-1375.

te Velde 1998

te Velde ER, Cohlen BJ, Looman CW, Habbema JD. Crossover designs versus parallel studies in infertility research. Fertility and Sterility 1998; 69: 357-358.

Thompson 2001

Thompson SG, Turner RM, Warn DE. Multilevel models for Meta-analysis, and their application to absolute risk differences. Statistical Methods in Medical Research 2001; 10: 375-392.

Turner 2000

Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for Meta-analysis of clinical trials with binary outcomes. Statistics in Medicine 2000; 19: 3417-3432.

Ukoumunne 1999

Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. Health Technology Assessment 1999; 3: 5.

Unnebrink 2001

Unnebrink K, Windeler J. Intention-to-treat: methods for dealing with missing values in clinical trials of progressively deteriorating diseases. Statistics in Medicine 2001; 20: 3931-3946.

White 2005

White IR, Thomas J. Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to Meta-analysis. Clinical Trials 2005; 2: 141-151.

White 2007

White IR, Carpenter J, Evans S, Schroter S. Eliciting and using expert opinions about dropout bias in randomized controlled trials. Clinical Trials 2007; 4: 125-139.

White 2008a

White IR, Higgins JPT, Wood A. Allowing for uncertainty due to missing data in Meta-analysis. Part 1: Two-stage methods. Statistics in Medicine 2008; 27: 711-727.

White 2008b

White IR, Welton N, Wood A, Ades AE, Higgins JPT. Allowing for uncertainty due to missing data in Meta-analysis. Part 2: Hierarchical models. Statistics in Medicine 2008; 27: 728-745.

Whiting-O'Keefe 1984

Whiting-O'Keefe QE, Henke C, Simborg DW. Choosing the correct unit of analysis in medical care experiments. Medical Care 1984; 22: 1101-1114.

（李玲、何佳译，康德英、秦天强、岑啸初审）

第十七章　病人报告的结局

作者：代表病人报告结局方法学组的 Donald L Patrick, Gordon H Guyatt 和 Catherine Acquadro Cochrane。

本文选自工作手册 5.0.1版本。有关如何引用它的指南，见17.9节。这些材料还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 病人报告的结局是指直接来自病人报告的关于病人健康状态的主观感受或功能情况及治疗情况，而不包括医务人员或其他人的解释。
- 病人报告的结局包括症状、体征、功能状态、认知或其他方面，如便利性（Convenience）和耐受性。
- 病人报告的结局量表所包含的反映相关概念的条目来源于目标人群；患者的参与是确保问卷内容效度至关重要的一部分。

- 为作者提供的术语词汇表见病人报告的结局方法学组网站（www.cochrane-pro-mg.org）。
- 病人报告的结局不仅在很多客观疾病结局指标不可用时很重要，而且在病人表述某种健康状态和治疗措施对于他们最重要的方面时也很重要。
- 病人报告的结局可能是连续资料或分类资料。有技术可用于合并这两种指标。
- 系统评价作者可能需要阅读与病人报告结局相关的背景资料，确保他们懂得纳入试验的选择，尤其是试验的有效性及检测变化的能力。
- 本章节提供了作者在将病人报告的临床结局纳入其系统评价之前应考虑的相关问题，以及“结果总结”的表格清单。
- 如果在系统评价评价方案中病人报告的临床结局被选为重要的结局指标，而完成的系统评价中又没有记录其结局指标，则应该强调这是当前关于治疗有效性研究中的缺陷。

17.1 什么是病人报告的临床结局？

病人报告的临床结局（Patient reported outcomes, PROs）是指直接来自病人报告的任何关于病人健康状态的主观感受或功能和治疗情况，不包括临床医生或其他人对病人身体反应的解释。病人报告的临床结局（PROs）所包括的任何治疗或结局评估都是通过访谈、自填问卷、日志或其他数据收集工具如便携设备和网络表单（2006年美国食品药品监督管理局）获得。来自看护人、卫生保健专业人员或父母和监护人（在一些情况下很有必要，如晚期癌症和认知功能障碍）的替代报告不能认为是病人报告的临床结局，而应该是一类单独的结局资料。

病人报告的临床结局是获取从病人角度的治疗效果；除了生存、疾病和生理学指标外直接评估治疗效果；而且往往是对病人来说很重要的结局指标。来自病人的报告可能包括日志中记录的体征和症状、主观感受的评估（最常被归类为症状）、行为能力的报告（最常被归类为功能状态）、一般认知或幸福感，其他报告包括对治疗的满意度、总体或健康相关生活质量和对治疗的依从性。报告也可能包括不良反应或副作用（见第14章）。

病人报告的临床结局有时用作临床试验的主要结局指标，尤其是对病人健康状况的直接受益没有可用测量指标时。通常PROs是主要结果指标的补充，如生存、疾病指标、

临床评分和生理或实验室指标。图17.1a展示了对病人来说很重要的所有结局的分类。

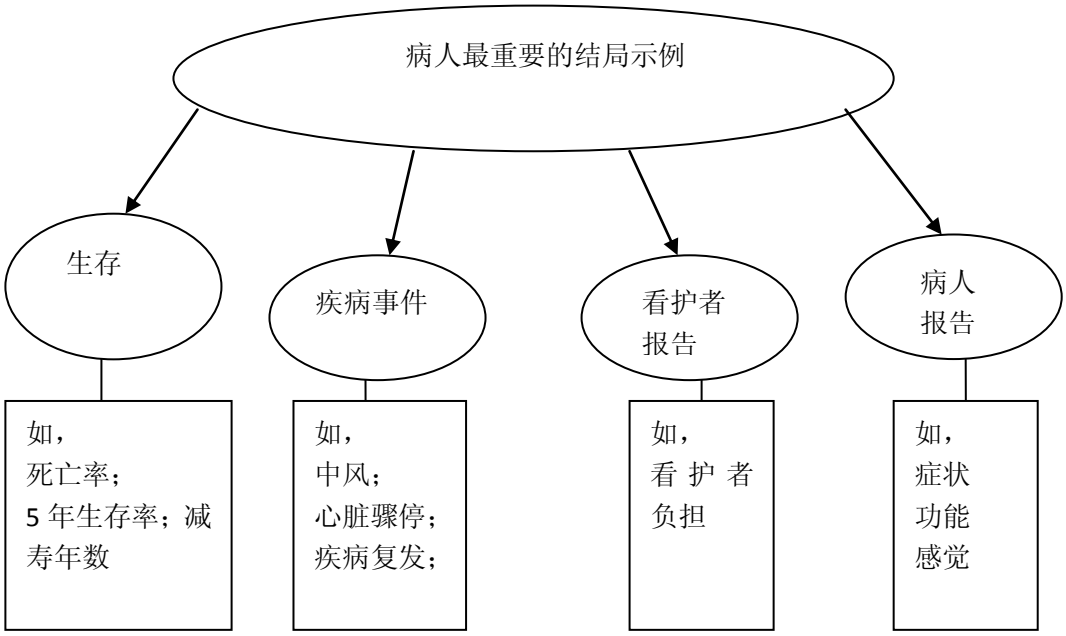


图17.1.a　病人最重要的临床试验结局分类说明

可能利用某种特定疾病、特定条件或一般情况下的测量指标（或工具）收集病人报告的临床结局。特定疾病的测量指标描述了某种特定疾病状况、条件或诊断分组的严重程度、症状或功能受限（如，关节炎或糖尿病）。特定条件的测量指标描述了病人关于某种特定条件或问题（如，后背疼痛）、特定干预或治疗措施（如，膝关节置换术或冠状动脉旁路移植术）的症状或体验。一般测量指标是专为任何疾病组或样本人群设计的。

病人报告的临床结局的术语词汇表见Cochrane病人报告的临床结局方法学组网站（见17.9a框图）。

17.2　病人报告的临床结局和Cochrane系统评价

系统评价作者将根据评价目的和范围选择纳入病人报告的临床结局（PROs）。当病人重要的外部观察结局难以获得或很少时，病人报告的临床结局就最重要。很多情况下，包括疼痛、功能障碍、性功能障碍和失眠，都没有令人满意的生物学测量指标可供选择。对于只有病人自己知道的情况就要求PROs作为主要的结局指标，如疼痛的强度和情绪。

当观察结局指标可以获得时，PROs同样也很重要，因为它直接反映了病人认为重要的问题。

系统评价早期过程很重要的部分就是定义及列出全部与研究问题相关的病人重要结局（Guyatt 2004）（见第5章，5.4.1节）。这一步与PROs的测量方法密切相关。许多原始研究都未测量对病人来说非常重要的自觉健康和生命质量。如果这样，与干预措施对疾病指标的影响相比，如发病率或死亡率，干预措施对病人报告的临床结局的影响的证据质量可能就差得多。在极端情况下，“结果概要”表中有一栏是空白的，例如，健康相关生存质量（Health reported quality of life，HRQOL）这一栏就是空白的因为没有研究直接涉及这个问题。需要事先仔细考虑病人所有很重要的结局，并将其作为一个空白行纳入在“结果概要”表中，这样将重点突出在合格的随机试验及其他研究中缺失的结局测量指标。

很重要的是系统评价作者要理解所纳入的研究中的PROs的本质，并向读者传达这一信息。临床试验中，研究人员使用多种工具获得病人报告的临床结局，并且制定、验证和分析PROs数据方法多种多样。

17.3 作为病人报告临床结局的健康状况和生命质量

健康状况和生命质量是PROs很重要的一类结局指标。已发表的论文常常很随便且相互替代使用“生命质量（Quality of life，QOL）”、“健康状况”、“功能状态”、“健康相关生命质量（HRQOL）”和“幸福感”（well-being），尽管它们都有明确的定义（见表17.3a）。

不同类型的工具可用于评估健康状况和生命质量（见表17.3b）。这些可能产生一个总分或指标数（如，表示干预措施对躯体或情绪功能的影响）、一个指标（也是一个总分，但是综合死亡和全面健康进行加权）、一个概况（维度或领域的个体得分）或者一系列测试（多重结果评估不同的概念）：见17.3b。

可以采用一般或特殊的工具或两者的组合测量HRQOL。如果研究人员对具体疾病和比较治疗措施在不同疾病和健康状况下对HRQOL影响之外的问题感兴趣，就可能选择一般HRQOL测量工具，该工具涵盖了HRQOL的所有相关领域（包括，活动能力、自理、和躯体、情绪及社会功能），并且是为管理有任何潜在健康问题（或者没有健康问题）的人所设计的。这些工具有时也叫做健康档案（health profiles）；最常用的健康档案

是医学结果研究（Tarlov 1989，Ware 1995）中所使用工具的简化形式。此外（或另外）随机试验和其它研究可能依靠针对特定功能（如，睡眠或性功能）、某个问题（如，疼痛）或某个疾病（如，心脏衰竭、哮喘或肠易激综合征）的工具。

PROs问卷具体概念和条目的得出应该来源于对病人、家庭成员、临床专家和相关文献的定性研究。定性方法的应用指南见第20章。病人的参与PROs问卷的制定必不可少，以确保内容效度。纳入的研究中使用和测量的概念只能通过检查声称是测量生命质量或健康相关生命质量工具中实际条目的内容和问题来决定。‘概念’就是被测量的“事物”。概念可能涉及到个别条目或相同概念条目的条目集，常常称作维度。例如，一个测量疼痛的条目，一种只有病人自己知道的感觉，就是一个症状而且这个被测量的症状概念就叫疼痛。一个评估爬楼梯困难程度的条目就是关于躯体功能的概念，并且可能就被归类为爬楼梯或作为躯体功能的一部分。研究者间概念界定的差异很大，而且没有一致认可的概念分类。然而，每一个条目、子维度、维度或总分表示了一个或多个概念，作者能从内容上识别它们，如，用于标记一个条目、维度或总分的语言。

系统评价作者可从原PRO研究作者所用的某个特定工具纳入条目的属性或来源中获得重要的见解。然而系统评价作者常常想只通过阅读已发表的临床试验结果来试着得到对概念或构思的准确认识。为了获得充分理解，他们至少得写一篇描述原始研究中纳入的PRO工具制定及前期使用情况的简要文章。

例如，一篇耳鸣的认知行为治疗（CBT）Cochrane系统评价作者纳入生命质量作为结局指标（Martinez-Devesa 2007）。对于生命质量评价，有四个试验使用耳鸣障碍问卷，一个使用耳鸣问卷，一个使用耳鸣反应问卷评估生命质量。系统评价中引用了原始来源。在MEDLINE中针对所有三个工具的心理测量属性的文章的引文均可查到，并且使用Google搜索引擎也能轻松找到。。这些文章包含对所测量条目和概念的信息，系统评价作者可以进行工具间内容的比较。

在理解工具在测量什么时，应考虑的另一个问题就是PRO工具怎样赋予权重的。很多特定工具都是赋予每个条目均等的权重而得到一个总分。主要针对经济分析设计的实用工具十分强调条目的加权，并试图把HRQOL在死亡和全面健康之前作为一个连续锚定。对我们之前段落罗列的问题感兴趣的读者，可以看到很经典且很有用的总结（Guyatt 1993）。

表17.3a　所列关于生命质量词汇的定义

词汇	定义
功能状态	个体能有效履行或有能力履行那些职责、任务或有价值的活动（如，上班、进行体育运动或维修房屋）。
健康相关生命质量（HRQOL）	个人健康状况。HRQOL 通常指生活中受精神或身体的健康状态主导的或显著影响的方面
生命质量（QOL）	对我们生活所有方面的评估，如，包括我们在哪里生活、怎样生活及生活得怎样。它围绕着这些生活因素，如家庭环境、经济情况、住房和工作满意度。（另见健康相关生命质量）。
幸福	主观的身体和情绪状态；个体感觉如何；一种区别于涉及行为和活动功能的精神状态

表17.3b　根据Patrick和Erickson改编而来的健康状态和生命质量测量方法的分类（Patrick 1993）

测量方法	优势	劣势
得分类型		
单个指标数	全球评估； 对人群有用	可能很难解释
单个指标	表示净影响； 对成本效益有用	有时不可能分开各个维度对总体得分的贡献
相关分数的概况	单一工具； 各维度对总分可能的贡献	长度可能是一个问题； 可能没有总分
独立评分测试	范围广泛的相关结局的可能	不能将不同的结果与常用的测量量表联系起来； 可能需要校正多重比较； 可能需要确定主要结果
人群范围和概念		
一般的：应用于不同的疾病，环境，人群和概念	广泛适用； 概念范围的概述； 检测到未预期结果的可能	可能不能反应变化； 可能没有关注病人利益； 长度可能是一个问题； 效果可能难以解释

特殊的：适用于个体，疾病，条件，人群或概念/维度	调查对象更易于接受； 可能更能反应变化	无法比较不同的环境或人群； 无法检测到未预期的结果
权重体系		
实用的：侧重病人、提供者或社区	区间尺度； 包含病人或消费者的观点	可能很难获得权重； 可能与更容易得到的均等权重相同的结果
均等的权重：各条目权重相同或根据频率/答复加权	更熟悉的技术； 似乎更容易使用	可能受患病率的影响； 不能进行权衡

17.4 测量病人报告的临床结局中的问题

17.4.1 工具的有效性

有效性是指该工具是否在测量计划测量的事物。PROs维度测量的经验证据要求强有力的正确性推论。为了提供这样的证据，调查人员借鉴了由多年致力于确定评估智力和态度的问卷是否能测量目的事物的心理学家制定的验证策略。

验证策略包括：

内容相关：工具的条目和维度恰当的证据、且与其预期的测量概念、人群和用途全面相关；

结构相关：条目、维度和概念之间的关系符合先前的逻辑关系假设的证据，这种关系应与病人和病人团体的特点或其他测量指标共存；

标准相关（针对某个用作诊断工具的PRO工具）：在何种程度上PRO工具的得分与标准测量相关。

确定有效性涉及检查应存在于评估指标间的逻辑关系。例如，一般来说日常生活中，跑步运动能力低下的病人比起该方面运动能力强的人更易发生气短，并且我们希望看到一个测量情绪功能的新问卷和现有的情绪功能问卷之间存在实质性相关性。

当我们对评估随着时间发生的改变感兴趣时，我们要评估评分变化的相关性得分。例如，一般而言，跑步运动能力下降的病人应在呼吸困难方面呈现增长态势，反之运动能力提高的则呈现减弱态势。同样，一个新的情绪功能测量指标应当在现有情绪功能测量指标改善的病人中体现出有所改善。该过程的技术术语就是测试工具的结构效度。

系统评价作者应该寻找和评估运用于其纳入的研究的PROs的有效性的证据。不幸的是，运用PROs的随机试验及其他研究的报告很少审查其所使用工具有效性的证据，但是系统评价作者能够从那些问卷已经预先经过验证的声明中（引文的支持）获得一些保证。

关于效度的最后一个问题是，如果测量工具运用于不同的人群或文化和语言不同的环境下，而不是问卷制作的环境（典型就是，使用了一个英语问卷的非英语版本）。理论上讲，应该有在纳入随机试验人群中有效性的证据。理想的情况是，PRO测量工具用于每一个研究都应该重新验证不管什么样的数据对验证有用，例如，测量的其它终点指标。作者评估证据有效性时应引起注意，当试验中要评估的人群与用于验证研究的人群不同时。

17.4.2 一个工具测量变化的能力

当我们运用工具评估治疗效果时，它们必须要能够测量出组间的差异，如果差异确实存在。随机应确保试验组和对照干预组的受试者在试验开始时处于相同的状态，不管是PRO计划测量的概念或结构。PROs必须能够检测对于病人重要的方面和区别试验过程中哪些参加人员保持不变、改善或恶化。这有时被称为应答性或对变化的敏感性。

在实验干预措施提高病人主观感受，而工具不能发现这种提高的研究中，测量变化能力差的工具会导致假阴性结果。这个问题在涵盖HRQOL所有相关领域的，但是对每个领域的涵盖都是很表浅的一般问卷中可能特别突出。试验组和对照干预组间PROs显示无差异的研究中，工具缺乏应答性是一个可能的原因。

17.5 定位并选择有病人报告的临床结局的研究

PROs与其他结局的检索方法相同（见第6章）。通常要检查研究者检索到的所有研究以找出那些包括PROs的研究。有时一个单独的、额外的PRO检索方法可用于补充标准策略。例如，如果哮喘领域一个纳入随机试验和其它研究的系统评价尚未找到运用PROs的研究，可进行一个单独的检索，包含用于哮喘的PROs的特定检索词，如“哮喘特异的生命质量”。然而，这依赖于所检索的数据库中存在关于PROs的电子记录。

PROs的索引词在主要的书目数据库中是不同的。系统评价作者不能依赖单一的索引词或副主题词检索包括PROs的研究。多个检索词通常很必要。例如，Maciejewski等 在一个关于评估减肥干预措施对健康相关生命质量影响的随机试验的系统评价中运用以下MEDLINE索引词（Maciejewski 2005）：“条件价值评估”；“健康状态”；“健康相关生命质量”；“心理方面”；“社会心理”；“生命质量”；“自我效能”；“SF-36”；“效用”；“幸福感”；“支付意愿”。自由词检索也应该包括尽可能多的相关同义词。检索需要把索引词和自由词组合起来且有可能需要多次反复。

系统评价作者可能发现在纳入对PRO方法和结果进行评价的系统评价中，设计和运用一个单独的数据收集表格很有用。这种格式的例子可以在我们的网站上找到：www.cochrane-pro-mg.org/documents.html.系统评价作者应该注意从工具中收集数据的其它方法：尤其是，对于连续性变量和二分类结局变量，是否能以有利于数据分析的形式收集数据。

17.6 评估和描述病人报告的临床结局

表17.6a介绍了针对PROs的精选问题，系统评价作者在将PROs融入系统评价中时应予以考虑。根据这个一览表，作者可考虑在“纳入研究的特征”表中详细地描述PROs或作为一个附表。

基于Patrick 和Erickson的第7章，医学文献的一个用户指南、疾病预防控制中心关于社区预防服务评估的指南和及医疗结局量表使用的标准（Patrick 1993，Guyatt 1997，Zaza 2000，Lohr 2002）。

表17.6a　临床试验中描述和评估PROs一览表

1. 什么是 PROs 测量？ a) 研究测量中 PROs 的概念是什么？ b) 作者提供的概念和结构选择的基本原理（如果有的话）是什么？ c) c. 涉及到结局指标选择的患者是否都测量了 PROs？
2. 遗漏 a) 从病人、临床医生、重要人物、支付者或其他管理人员和决策者的角度来看，研究中是否有生命质量（如，综合评价、生活满意度）或健康（如，症状、功能、认知）的重要相关方面被遗漏？
3. 如果随机试验和其他研究测量了 PROs，工具的测量方法是什么？ a) 调查人员采用了产生单一指标或指数、一个概况的工具还是一系列的工具？ b) 如果调查人员测量 PROs，他们运用的是一般工具还是特定工具，或者两者兼而有之？ c) 谁最后完成了工具？
4. 测量工具是以设想的方式运作吗—有效性/效度？ a) 所运用的工具之前已经被验证过吗（提供参考）？之前被验证应用于该人群的证据被介绍过吗？ b) 该研究中工具被重新验证了吗？
5. 测量工具是以应该工作的方式运作吗—测量变化的能力？ a) 即使这些变化很细微，PROs 也能够检测病人状态的变化吗？
6. 你能够使效应的强度（如果有的话）让读者易于理解吗？（你必须做到!） a) 你能够提供达到功能阈值或状况改善病人的差异估计值和相关的需要治疗人数（NNT）吗？

17.7　病人报告的临床结局不同测量指标间的可比性

调查人员可能会选择不同的工具评估PROs，因为他们对某种PRO使用不同的定义或者运用不同的工具评估同一PRO。例如，某个调查人员可能会选择使用一个通用的工具评估功能状态或使用一个不同疾病的特定工具评估功能状态。结局的定义可能相同或不同。系统评价作者必须决定如何对研究间的PROs分类及何时合并结果。这些决定将基于PRO的特点，并需要在系统评价中提取并报告。

在许多场合，运用PROs的研究会确定基线和随访测量，而且关注的结局在干预组和对照组间从基线到随访情况的变化将有所不同。理想的状况是，为了合并两种概念相关的PROs，要有单个病人数据在两种测量方法中的变化具有很强的纵向相关性的证据和对工具相似应答的证据。进一步的支持证据可能来自治疗组和对照组间差异或前后测量间差异的相关性。如果不能找到这些数据，则可以依靠某个时间点单个病人间的横断面相关性。

例如，用于测量慢性阻塞性疾病病人健康相关生命质量的两个主要工具是慢性呼吸问卷（CRQ）和St. George's呼吸问卷（SGRQ）。这两种问卷在个体研究中的相关性在横断面（某一时间点的相关性）比较和纵向（变化的相关性）比较的变化范围均是0.3-0.6（Rutten-van Mölken 1999，Singh 2001，Schünemann 2003，Schünemann 2005）。

在随后的调查中，调查人员对包括23个病人组的15个研究中CRQ和SGRQ问卷的平均变化的相关性进行了评估，并得出相关系数为0.88（Puhan 2006）。尽管有极强的相关性，证明了CRQ比SGRQ应答更好：CRQ的标准化应答平均值（标准化应答平均值的中位数为0.51，IQR为0.19-0.98）显著高于（P＜0.001）SGRQ（标准化应答平均值的中位数为0.26，IQR为-0.03-0.40）。也就是说，当两种工具都用于相同的研究，CRQ系统地得出较大的治疗效果。因此，合并运用这两种工具的试验的结果可能导致低估运用SGRO的研究的治疗效果。

常常不幸的是之前段落提到的详细数据不可用。调查人员必须依靠不同工具测量相同潜在结构的程度的直觉作出决定。例如，社会心理干预措施治疗经前综合征的一篇Meta分析的作者面临大量的结局测量指标，9个纳入的研究中的25个PROs。他们处理这个问题通过两名调查人员独立评估每一个工具（包括全部维度）并把它们分成6个独立的概念范畴；通过讨论解决分歧以达成共识。对包含不止一个研究（2个至6个研究）的每个类别进行合并分析。

对运用不同测量尺度的研究的Meta-分析通常会使用标准化均差（SMDs；见第9章，9第.2.3节）。然而当关注的是比较干预组和对照组相对于基线的变化时，SMDs是很有问题的，因为变化的标准差不能反应病人间的变异（它们也依赖于最终测量值和基线间的相关关系；见第9章，第9.4.5.2节）。

类似的原则也适用于系统评价作者关注以二分类形式存在的可用数据的研究，或者系统评价作者容易从中提取二分类结局数据的研究。例如，研究黄酮类物质对痔疮症状影响的调查人员发现纳入的随机试验没有坚持用类似的症状测量指标；但是，全部14个

试验除一个以外都记录了完全没有症状、症状有所改善、仍然有症状或症状加重的病人的比例（Alonso-Coello 2006）。初始分析的时候，调查人员把病人完全没有症状、有些症状或有所改善的临床结局同等看待，并且基于对治疗效应的强度和方向的预先的期望合并每个结局指标。

这留下一个问题，即怎样处理那些报告病人体验“有所改善”的研究。调查人员进行了比较二分方法的分析，即将“有所改善”作为阳性结果和阴性结果（类似于没有改善）。二分法的结果往往非常有用，尤其是很容易对临床医生和病人作出结果解释。虚构但仍严格的二分方法会得出综合统计量为临床实践提供有用的指导。

测量某种具体PRO多种工具的应用和用多种方法分析的试验可能导致选择性地报告最令人关注的结果并把严重的偏倚带入系统评价中。集中于PROs研究的系统评价作者应警惕这一问题。当只有少数纳入的研究报告了某种特定结果，特别是当它是一个显著的结果时，做为认真负责的调查人员应对它进行测量，同时作者应注意到报告偏倚的可能性（见第10章）。

17.8 结果解释

17.8.1 关注单个病人报告的临床结局的研究总结

当Meta-分析纳入的研究只是报告单个PRO时，作为一个连续变量，合并的结果将会产生均数差。该均数差存在的问题就是临床医生可能对其解释存在困难。例如，如果告诉运用慢性呼吸问卷的一些列随机试验中康复和标准护理的均数差为1.0（95% CI 0.6 – 1.5），很多读者会不清楚这是否表示不重要、细微但很重要、中等程度还是很大的效应。

系统评价作者通过报告可能的结果范围和研究中治疗组和对照组结果均数的范围有助于结果解释。最有用的是（如果是有的话）是一个最小差异的估计值，并是病人有可能认为是重要的（最低重要差异，MID）。有许多方法能得出MID的估计值，包括运用变化的总体评定等级（Guyatt 2002）。理想情况下，系统评价作者会在摘要里介绍MID的估计值。例如，调查人员评估呼吸康复治疗对慢性肺部疾病患者报告的健康相关生命质量的影响，摘要中，“HRQL两个重要特征，呼吸困难和控制，与MCID0.5相比，总体效果大于MCID，分别为：1.0（95% CI 0.6-1.5）和0.8（0.5-1.2）”（Lacasse 1996）。

尽管这非常有用，但它可能诱导临床医生作出不恰当的推论。如果MID是0.5，处理间均数差为0.4，临床医生可能推断没有人从这种干预措施中受益。如果均数差是0.6，他们可能得出的结论是所有人均受益。这两种推论可能都是错误的。首先，他们忽略了点估计值的不确定性（可信区间）。更重要的是，忽略了个体对治疗反应的变异性（标准差）。

可能对于研究者来说，给出‘响应者’的定义有助于解释结果（见第12章，第12.6.1节）。知道单个病人作为治疗应答的定义非常有用。这个应答者的定义是基于由经验证据支持事先规定的的标准，并支持应答者定义作为受益的衡量。定义应答者的方法包括：（1）相对于基线在一个或多个尺度上事先确定的变化；（2）一定大小或更大的得分变化（如，在8分的范围内改变2分）；（3）相对于基线的百分比变化。

17.8.2 运用不止一个病人报告的临床结局进行研究总结

第17.8.1节的论述指出当合并PROs时均数差不再是一个可能测量效果的指标，因此我们用SMD替代（见第9章，第9.2.3节）。不幸的是，当必须依靠SMD生成汇总数据时没有一种完全令人满意方法可提供对PRO的效应强度。可提供给读者标准经验法则解释效应值（如，0.2表示很小的效应，05表示中等程度的效应，08表示很大的效应（Cohen 1988））或一些变异（<0.41 =小，0.40至0.70 = 中等，>0.70 =大）。另一种可能更差的方法就是，在大多数情况下，使用大约0.5的标准化均数差作为有显著差异的最小值（Norman 2003）。

报告和解释PROs和其他临床结局的及做出推论和结论一般方法在第12章中有论述（见12.6节）。

17.8.3 当研究并没有涉及病人报告的临床结局

许多原始研究无法测量对病人来说很重要的自觉健康状况及生命质量的各个方面。这种情况下，干预措施对PROs影响的证据远弱于对疾病指标发病率或患病率影响的证据。极端情况下，尚无研究涉及PROs。优先考虑对病人重要的全部临床结局测量指标，将突出在纳入的随机试验和其他研究中所缺失的结局测量指标。这个遗漏应该在系统评价作者的结论中作为对将来研究的启示而重点突出。

17.9 本章信息

作者：Donald L Patrick, Gordon H Guyatt and Catherine Acquadro，代表Cochrane 病人报告的临床结局方法组

本章引用格式：Patrick D, Guyatt GH, Acquadro C. 第17章：病人报告的临床结局。In: Higgins JPT, Green S（编辑）。Cochrane 干预系统评价手册。版本5.0.1 [2008年9月更新]。Cochrane 协作网，2008年。Available from www.cochrane-handbook.org.

致谢：Jason Busse, Peter Fayers, Toshi Furukawa, Madeleine King 和Milo Puhan对草案提出了建议。

框17.9a Cochrane病人报告的临床结局方法组

病人报告的临床结局方法组（PRO MG）的主要目的是建议 Cochrane 作者何时及如何把健康状况和生命质量数据并入系统评价中。一些 Cochrane 评价组已经在系统评价中合并 PRO 数据时遇到了一些困难。这些困难包括合并和解释数据及评估 PRO 尺度的有效性。

PRO MG旨在：

完善PRO研究的文献检索方法；

开发系统评价HRQL研究的方法；

晚上PRO研究的Meta-分析方法（与统计方法组合作）；

完善PRO测量工具运用于经济学评价的方法，与Campbell-Cochrane方法组合作；

软件开发。

该小组根据要求向 Cochrane 协作指导组提供建议，召开关于健康和病人报告的临床结局问题和方法的专题会，回应协作网的需要以及为该手册准备建议。该小组成员将参加 Cochrane 系统评价的准备并会通过书面材料和培训研习会给作者一些建议。该小组成员还会帮助已决定纳入 PRO 结局的系统评价作者制定方案计划书和审查。

网站：www.cochrane-pro-mg.org/

17.10 参考文献

Alonso-Coello 2006

Alonso-Coello P, Zhou Q, Martinez-Zapata MJ, Mills E, Heels-Ansdell D, Johanson JF, Guyatt G. Meta-analysis of flavonoids for the treatment of haemorrhoids. British Journal of Surgery 2006; 93: 909-920.

Cohen 1988

Cohen J. Statistical Power Analysis in the Behavioral Sciences (2nd edition). Hillsdale (NJ): Lawrence Erlbaum Associates, Inc., 1988.

Guyatt 1993

Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Annals of Internal Medicine 1993; 118: 622-629.

Guyatt 1997

Guyatt GH, Naylor CD, Juniper E, Heyland DK, Jaeschke R, Cook DJ. Users' guides to the medical literature. XII. How to use articles about health-related quality of life. Evidence-Based Medicine Working Group. JAMA 1997; 277: 1232-1237.

Guyatt 2002

Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. Mayo Clinic Proceedings 2002; 77: 371-383.

Guyatt 2004

Guyatt G, Montori V, Devereaux PJ, Schünemann H, Bhandari M. Patients at the center: in our practice, and in our use of language. ACP Journal Club 2004; 140: A11-A12.

Lacasse 1996

Lacasse Y, Wong E, Guyatt GH, King D, Cook DJ, Goldstein RS. Meta-analysis of respiratory rehabilitation in chronic obstructive pulmonary disease. The Lancet 1996; 348: 1115-1119.

Lohr 2002

Lohr K. Assessing health status and quality-of-life instruments: attributes and review criteria. Quality of Life Research 2002; 11: 193-205.

Maciejewski 2005

Maciejewski ML, Patrick DL, Williamson DF. A structured review of randomized controlled trials of weight loss showed little improvement in health-related quality of life. Journal of Clinical Epidemiology 2005; 58: 568-578.

Martinez-Devesa 2007

Martinez-Devesa P, Waddell A, Perera R, Theodoulou M. Cognitive behavioural therapy for tinnitus. Cochrane Database of Systematic Reviews 2007, Issue 1. Art No: CD005233.

Norman 2003

Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Medical Care 2003; 41: 582-592.

Patrick 1993

Patrick DL, Erickson P. Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation. New York (NY): Oxford University Press, 1993.

Puhan 2006

Puhan M, Soesilo I, Guyatt GH, Schünemann HJ. Combining scores from different patient reported outcome measures in Meta-analyses: when is it justified? Health and Quality of Life Outcomes 2006; 4: 94.

Rutten-van Mölken 1999

Rutten-van Mölken M, Roos B, Van Noord JA. An empirical comparison of the St George's Respiratory Questionnaire (SGRQ) and the Chronic Respiratory Disease Questionnaire (CRQ) in a clinical trial setting. Thorax 1999; 54: 995-1003.

Schünemann 2003

Schünemann HJ, Griffith L, Jaeschke R, Goldstein R, Stubbing D, Guyatt GH. Evaluation of the minimal important difference for the feeling thermometer and the St. George's Respiratory Questionnaire in patients with chronic airflow obstruction. Journal of Clinical Epidemiology 2003; 56: 1170-1176.

Schünemann 2005

Schünemann HJ, Goldstein R, Mador MJ, McKim D, Stahl E, Puhan MA, Griffith LE, Grant B, Austin P, Collins R, Guyatt GH. A randomised trial to evaluate the self-administered standardised chronic respiratory questionnaire. European Respiratory Journal 2005; 25: 31-40.

Singh 2001

Singh SJ, Sodergren SC, Hyland ME, Williams J, Morgan MD. A comparison of three disease-specific and two generic health-status measures to evaluate the outcome of pulmonary rehabilitation in COPD. Respiratory Medicine 2001; 95: 71-77.

Tarlov 1989

Tarlov AR, Ware JE, Jr., Greenfield S, Nelson EC, Perrin E, Zubkoff M. The Medical Outcomes Study. An application of methods for monitoring the results of medical care. JAMA 1989; 262: 925-930.

US Food and Drug Administration 2006

US Food and Drug Administration. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims [February 2006]. Available from: http://www.fda.gov/cber/gdlns/prolbl.htm (accessed 1 January 2008).

Ware 1995

Ware JE, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. Medical Care 1995; 33: AS264-AS279.

Zaza 2000

Zaza S, Wright-De Agüero LK, Briss PA, Truman BI, Hopkins DP, Hennessy MH, Sosin DM, Anderson L, Carande-Kulis VG, Teutsch SM, Pappaioanou M, Task Force on Community Preventive Services. Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. American Journal of Preventive Medicine 2000; 18 (Suppl 1): 44-74.

（高霑译，秦天强、岑啸初审）

第十八章　个体病人数据的系统评价

作者：代表 Cochrane 协作网个体病人资料 meta 分析方法学组的 Lesley A Stewart, Jayne F Tierney 和 Mike Clarke。

本文选自工作手册 5.0.1版本。有关如何引用它的指南，见18.6节。这些材料还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 个体病人数据（Individual patient data, IPD）Meta 分析是从每一临床试验负责人处直接收集每一个体病人的原始数据。
- 通过“原始”资料获取能够进行数据核对，深入探讨，并以统一方法重新分析数据。
- 在已发表信息不能进行高质量系统评价或要求的特定分析类型不能使用汇总数据的情况下，IPD Meta 分析显示其独特优势。
- 大部分 IPD Meta 分析由协作组实施并发表，包括一个项目组或秘书处，及提供资

料的研究者与顾问组。

- IPD 系统评价通常比基于已发表的汇总数据的传统系统评价更耗时耗力。
- 收集 IPD 的收益在某些情况下也许微不足道，在另一些情况却可能至关重要。

18.1 引言

18.1.1 什么是IPD系统评价?

个体病人数据（IPD）Meta分析是系统评价的一种特殊类型，是直接从纳入研究的原始研究者处收集每一个研究对象的原始数据，而非从已发表的研究结果中提取数据。这些资料可重新集中分析，在适当条件下可进行Meta分析。Cochrane系统评价可由个体病人数据评价组承担，但IPD Meta分析通常需要专职人员，需要大量的时间去开展。IPD系统评价需要特殊的方法，比基于已发表或汇总数据的传统系统评价需要更多的时间和成本。但是，IPD系统评价在数据质量和可进行的分析类型方面有独特优势（Stewart 1995，Stewart 2002）。因此，IPD系统评价被视为系统评价的“金标准”。本章目的在于提供IPD系统评价方法的概述，帮助作者确定在他们的系统评价中收集IPD是否有用且可行。本章不提供详细方法学指导，若首次制作可联系IPD Meta分析方法学组，工作组人员可提供咨询和指导。

18.1.2 何时需制作IPD系统评价？

若根据已发表信息无法进行高质量评价，或要求的特定分析类型不能使用标准方法的情况下，可考虑IPD系统评价。当原始资料已损坏、丢失或研究者不配合时，不适合IPD分析。另外，如果所有数据都以适当格式发表且容易获得，不需要进行IPD系统评价。IPD分析有利条件的详细信息见框18.1.a.

框18.1.a　IPD分析可能的有利条件

IPD meta 分析在以下情况可能有用：

- 未发表或灰色文献的研究
- 研究报告质量较差（如资料介绍不恰当，有选择性或不明确等）
- 发表的分析中已删去大量个体病人资料
- 除已报道的结果外希望获得长期结果数据（如死亡率或儿童发育结果数据）
- 不同研究结果测量方法定义不同
- 需用时间事件分析法
- 需用多元分析或其他复杂分析法
- 重点探讨干预措施与病人特征之间的关系

18.1.3　IPD系统评价方法有何不同？

IPD Meta分析所采用的一般方法与其他系统评价方法相同，但是在数据收集、核对与分析阶段有较大区别。与Cochrane系统评价相同，IPD分析需要撰写计划书，说明其评价目的、想要解决的问题，确定纳入排除标准，收集个体病人数据的理由及方法以及确定数据分析方法。尽管原始研究者参与到项目中，可能更易于找到他们所做或所知的研究，但是不论是否能收集到个体病人数据，用于合格研究筛选的方法应均相同。项目内容应以结构式报告进行准备和发表。IPD系统评价可能还包括结果展示和与合作研究者的讨论会议。

18.1.4　如何组织一项IPD系统评价？

IPD系统评价通常以项目合作形式开展，即包括提供原始研究资料的研究者与项目管理者所构成的一个积极的协作组。项目由较小的地方项目组或秘书组管理，其重大战略性决策由更大的顾问小组提供咨询。结果以协作组名义发表。秘书处同时负责协作者间讨论会议的组织，召集所有成员来讨论初步结果。

18.1.5　那些卫生保健领域使用过IPD分析方法？

自从20世纪80年代末期，随着心血管疾病和癌症研究领域方法学的稳步发展，开始使用IPD Meta分析。目前癌症领域有超过50篇IPD Meta分析，包括大范围实体瘤位置和血液系统恶性肿瘤的筛查与治疗（Clarke 1998）。IPD Meta分析也用于其它领域系统评

价（Simmonds 2005），包括HIV感染，痴呆，癫痫，抑郁症，疟疾，疝和哮喘等。Cochrane IPD Meta分析方法学组网站有正在进行和已完成的IPD系统评价数据库，可查找详细信息。（见框18.6.a）

18.1.6 制作IPD系统评价的首要步骤

制作一个IPD 系统评价前，应仔细考虑项目成功所需的技术和资金支持，接受相关培训与收集意见，最好事先联系Cochrane IPD Meta分析方法学组(框18.6.a)。

18.2 IPD Meta分析的协作性质

18.2.1 协作组

大部分IPD Meta分析由协作组完成并发表。协作组由以下部分构成：项目管理组或秘书组，顾问小组成员（若有），提供数据的研究者。

18.2.2 协商合作

建立合作关系需要一定的时间和精力。联系原始研究负责人比较困难，可能有些研究者不愿加入项目中来。第一步通常是信函联系，邀请合作人员，说明项目情况，说明需要的受试者信息及该Meta分析如何开展和发表。信函通常由项目小组代表顾问组发出。这一阶段需提供计划书以说明更多信息，但第一次通讯联系并不收集数据。另外可能有必要建立与负责研究数据管理和提供数据查询的数据中心或研究机构的单独联系。鼓励原始研究者参与到IPD系统评价中来，重点是尽量提供支持，建立必要的联系，保证所有合作者都参与并了解进展。定期发送内部电子邮件是保证合作组信息更新，相互联系，特别是长期项目开展的有效途径。

18.2.3 保密

原始研究者通常与IPD Meta分析小组应签署保密协议，为数据的使用与储存提供保障。协议内容大概为数据进行秘密保存，仅项目组成员有权访问，不能在其他地方复制与传播，具体细节可有所不同。越来越多的国家资料保护法律要求所提供的数据不能明

确受试者身份，一个较好的办法是用某些数据进行个体资料的去识别，如用研究特定代码代替姓名。另外，通过电子邮件发送的信息应尽量加密。

18.3 数据处理

18.3.1 确定需要收集的资料

计划书中应明确计划分析的病人特征和结局指标。但在开始收集资料前，最好应向原始研究者确认实际可用的资料。确定需要收集的变量时，应慎重考虑需要分析的数据和分析方法，尽量避免收集不必要信息和遗漏必要信息。调查人员存在可能遇到提供不用于分析和报道的数据的烦恼或质疑是可以理解的。

多数情况下，是能够收集个体研究所定义的结局与病人特征变量，但同时应考虑是否存在进一步分析需收集的数据。如果研究中对结局指标有不同定义，想要以统一方式对所有研究中的每一例病人重新定义，则可能需要额外的变量。比如重新定义子痫需要收集收缩压、舒张压和蛋白尿相关数据。

18.3.2 数据格式

原始研究者答应合作后，下一步即是告知其需要提供的数据及参考的格式。项目组应做好接受数据提供者方便的任何格式的准备，无论是电子版，还是纸质版，必要时重新编码。20世纪80年代最初的IPD meta分析多为纸质版数据，现在很多信息都是通过email或光盘提供，调查人员通常按照指定格式转换或编码数据。

18.3.3 变量的重新编码与定义

来自不同研究以个体水平收集的数据经不同阶段、等级、序列或其他评分系统转换后可进行汇总，否则由于数据收集工具的差异，数据无法合并。因此，收集恰当的，经统一编码或转化后反应相同定义的数据非常重要。例如，如果感兴趣的结局是先兆子痫，除收集血压和蛋白尿的数据外，还要考虑根据计划书定义是否观察到先兆子痫。

18.3.4 核对数据

核对数据的目的在于提高数据的准确性，确认试验使用正确的随机方法，适当情况下，尽量保证数据是最新的。进行准确的数据核对取决于卫生领域的需解决的问题及数据性质，主要有以下四方面：

18.3.4.1 缺失或重复数据核对

收到数据后，应立即检查数据是否能被中央分析系统读取和加载。例如，如果数据以电子邮件附件形式发送，应检查文件能否打开，信息是否正确。在此阶段，确定已经收集来自所有合适（通常是随机的）的病人的数据是很有用的，并检查提供的数量是否与其他信息或任何出版物一致，比如病人记录序列或研究的ID号码是否有明显遗漏或重复。

18.3.4.2 数据合理性核对

合理性核查包括变量范围检查，要求原始研究者确认任何离群值或异常值：如确认异常年老或年轻患者，或异常高或低的胆固醇水平。资料还应与任何相关的已发表的研究核对，如确认基线特征分布，与结果一致的研究人数（后续纳入或额外随访可能改变后续发表的结果）。

18.3.4.3 随机方法的核对

通常核实随机方法是否被恰当地使用非常有助于判断。如果随机日期可得，可以通过时间-受试者累积入组数量示意图判断：恰当的随机应保证各干预组入组人数一致，各组入组的曲线应该频繁交叉。也可通过观察每周内每天各组入组人数的随机分布进行判断。这里，假设合理数量的受试者被随机分配并且在正常的临床工作时间进行随机分组的试验只有很少的受试者是在此之外纳入的，那么在任一工作日可看到同样数量的受试者被分配到每个干预组。核实干预组间和重要的亚组中重要基线特征是否均衡可比同样有助于判断，但应注意，具有统计学差异的不均衡可能是由机遇所致。

18.3.4.4 最新数据的核实

针对事件长时间观察得出的结局，如癌症生存率，很重要的是核实随访在个干预组保持一致并且所得的是最新数据。基于未经历观察事件病人（截尾值看做事件发生）所

做的‘反’K-M曲线可用于检查各组间随访的均衡性。

每个研究的所有核查结果应有总体情况概览，包括所提供数据的质量及任何潜在问题，任何担心可以婉转地向研究负责人说明。简单的错误或误解通常可通过讨论解决，重要问题不能解决的情况比较少见。

数据转换或修改前应备份存档，数据核对的整个过程中，对数据所做的任何修改或转换都应妥善记录。

18.4 数据分析

18.4.1 数据分析优势

通过收集原始研究数据可对数据进行统一核查、充分挖掘和再分析。因此，IPD数据分析可不依靠已发表报告中的解释信息和分析方法，不受制于以表格格式提供的汇总数据，或不需要对用不同方法计算出来的统计量进行合并。同时避免了原始分析的问题，如可根据意向性治疗原则分析，即使原始研究无此分析。

18.4.2 一般方法

目前大多数IPD meta分析采用二阶段分析方法。第一阶段，每个研究按照计划书中设定的方法分析。第二阶段，每个研究的分析结果或概括性的统计量按传统系统评价方法进行合并，得出一个总的效应估计值（Simmonds 2005）。二分类数据（Turner 2000）、连续性数据（Higgins 2001）、等级资料（Whitehead 2001）和时间事件数据（Tudor Smith 2005b）采用更加复杂的多水平模型分析方法，但目前应用并不常见。若研究无异质性，对时间事件数据的二阶段分层对数秩法最好避免以免夸大干预的效果（Tudor Smith 2005a）。

18.4.3 时间-事件分析

收集随机分组和结局发生间的时间的IPD可以进行时间事件分析，如包括恢复时间，癫痫缓解时间，怀孕时间，死亡时间等。IPD meta分析在癌症领域如此重要的主要原因之一是对生存的时间事件分析对治疗评价至关重要。除了治愈之外，大多数干预都可能

延长生存时间，因此，不仅要衡量是否发生死亡，同时应衡量死亡发生的时间。为了进行这类分析需确认每一个体的“无结局”时间，通常收集随机分组时间，结局状态（即无论是否观察到特定事件）和结局最后评估日期。有时需收集随机分组日期与结局最近评估时间之间的间隔时间。采用时间事件分析是为了计算Meta分析中每个试验的风险比（hazard ratios）。

18.4.4 长期随访结果分析的更新

针对结局为持续发生的事件如生存数据，IPD Meta分析可提供研究随访期之外的干预措施效果，同时还可提供相关结局的更新数据，如研究发表后的死亡率。

18.4.5 亚组分析

搜集IPD是分析探讨某项干预措施是否在不同类型的受试者中具有同样效果的最有效方式，例如女性的治疗效果是否优于男性。基于发表的汇总数据的传统分析，通常难以提取足够合适的数据进行亚组分析，特别是难以一次用多种因素描述个体。IPD允许对个体进行直接分类，可由单一或多种因素定义亚组分析（研究分层）。收集IPD还可进行如多水平模型等更复杂的分析，探讨干预效果和病人特征间的关系。

18.4.6 其他分析

使用IPD可对无论有无干预的病人本身特点进行深入研究。例如，大量资料集可用于建立预后指标，即可以根据病人特点预测治疗结果（International Germ Cell Cancer Collaborative Group 1997）。

18.4.7 软件支持

IPD不能直接用RevMan分析。如果进行二阶段分析，IPD数据需先进行RevMan软件以外的分析，得出每个研究可能进入RevMan的概括性统计量二分类和连续性数据可以常规方式录入。时间事件分析结果，事件发生数的观察值减去预期值的结果和方差数字可使用“O – E and Variance”选项。另外“generic inverse-variance”选项可以用于效果估计，如风险比，率比或调整估计

尽管许多标准统计软件包可进行必要的单个研究的IPD分析，每次分析一个研究的一个结果较费时费力，并且目前没有能够在IPD Meta分析中进行直接分析及数据整合和作图的商业软件。非商业软件包“SCHARP”可分析单个研究，可计算二分类、连续性和时间事件IPD的汇总结果、结果输出表格和森林图，并对于非营利性组织免费。该软件包基于SAS由英国医学研究委员会临床试验Meta分析组开发，可通过与IPD Meta分析方法学组联系作者获得（见Box 18.6.a）。

18.5 局限性与注意事项

18.5.1 IPD系统评价不能解决的问题

虽然IPD方法有助于避免研究中分析报告相关的问题，但通常它不能避免与研究设计或实施相关的偏倚。如有此类问题（同时也可能出现在研究的出版物和基于此的任何系统评价中），该研究应从Meta分析中排除。

18.5.2 不能获得的研究

收集IPD时，往往能够纳入传统系统评价不会纳入的研究，因为这些研究要么未发表要么没有报告允许其纳入的足够信息，因此有助于避免多种发表偏倚（Stewart 2002）。但是，必须确保通过限制对提供IPD的研究的分析，因选择性使用研究数据不会带来偏倚。

IPD方法成功应用的重点在于所有（或接近所有）的研究数据均可用。如果无法获得相关研究结果，如原始研究单位倾向于提供阳性结果数据，那么忽视这些不可用的研究会导致IPD系统评价出现偏倚。如果大部分数据已获得， 可能90％或更多进行了随机分配的个体，我们就会比确信所得结果。然而，如果信息较少，则需谨慎下结论。合并任何不可用研究（自出版物提取或以表格的形式获得）的敏感性分析，并与主要IPD结果比较，有助于数据的解释。在不能获得所有研究IPD的IPD系统评价报告中应说明无法获得个体病人数据的原因以及由此导致的偏倚的可能性。

与其它的Cochrane系统评价一样，IPD Meta分析应清楚描述未纳入的研究及相关原因。如果仅有少数研究能提供IPD用于分析，则该方法的价值值得商榷。在癌症研究方面有较好经验，大多数病例坚持治疗使符合纳入标准的研究的很高比例的数据能够获得，因此，在项目早期判断原始研究这提供IPD的意愿和能力非常重要。

18.5.3 何时制作IPD系统评价

制作任何系统评价一开始就需要考虑好合适的方法以及数据类型，应特别注意引起偏倚的各种因素。收集个体病人资料的优势在某些情况下可能微不足道但是在其他方面却至关重要。

18.6 本章信息

作者：代表Cochrane个体病人资料Meta分析方法学组的Lesley A Stewart, Jayne F Tierney and Mike Clarke

本章引用格式：Stewart LA, Tierney JF, Clarke M. Chapter 18: Reviews of individual patient data. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.

致谢：感谢Paula Williamson对文章起草的帮助。

框18.6.a Cochrane个体病人资料Meta分析方法学组介绍

Cochrane 个体病人资料 Meta 分析方法学组（IPD MA MG）包括参与 IPD 系统评价或对此感兴趣的人员及有关方法学研究人员。该协作组的目标是为制作 Cochrane 系统评价 IPD meta 分析提供指导。

IPD MA MG 成员的工作包括：

- 承担 IPD meta 分析
- 承担实证研究，如 IPD meta 分析与其他形式的系统评价比较的相对好处；利用收集的信息探讨随机试验和系统评价设计方法，分析和报告偏倚和异质性来源
- 帮助 Cochrane 系统评价作者确定是否适合做 IPD 系统评价，若可以，提供相关建议
- 开展 Cochrane 专题讨论培训并提供培训材料
- 维护 IPD 系统评价注册，方法学研究项目和 Meta 分析数据库

网址：www.ctu.mrc.ac.uk/cochrane/ipdmg

18.7 参考文献

Clarke 1998

Clarke M, Stewart L, Pignon JP, Bijnens L. Individual patient data Meta-analysis in cancer. British Journal of Cancer 1998; 77: 2036-2044.

Higgins 2001

Higgins JPT, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. Statistics in Medicine 2001; 20: 2219-2241.

International Germ Cell Cancer Collaborative Group 1997

International Germ Cell Cancer Collaborative Group. International Germ Cell Consensus Classification: a prognostic factor-based staging system for metastatic germ cell cancers. Journal of Clinical Oncology 1997; 15: 594-603.

Simmonds 2005

Simmonds MC, Higgins JPT, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clinical Trials 2005; 2: 209-217.

Stewart 1995

Stewart LA, Clarke MJ. Practical methodology of Meta-analyses (overviews) using updated individual patient data. Statistics in Medicine 1995; 14: 2057-2079.

Stewart 2002

Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. Evaluation and the Health Professions 2002; 25: 76-97.

Tudor Smith 2005a

Tudor Smith C, Williamson PR. Meta-analysis of individual patient data with time to event outcomes. International Conference of the Royal Statistical Society, Cardiff (UK), 2005.

Tudor Smith 2005b

Tudor Smith C, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data Meta-analysis of time to event outcomes. Statistics in Medicine 2005; 24: 1307-1319.

Turner 2000

Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for Meta-analysis of clinical trials with binary outcomes. Statistics in Medicine 2000; 19: 3417-3432.

Whitehead 2001

Whitehead A, Omar RZ, Higgins JPT, Savaluny E, Turner RM, Thompson SG. Meta-analysis of ordinal outcomes using individual patient data. Statistics in Medicine 2001; 20: 2243-2260.

（成岚、崔晓华译，李江、岑啸、秦天强初审）

第十九章　前瞻性 Meta 分析

作者：Cochrane 前瞻性 Meta 分析方法学组的 Davina Ghersi, Jesse Berlin 和 Lisa Askie

本文选自工作手册 5.0.1版本。有关如何引用它的指南，见19.6节。这些材料还刊登于Higgins JPT和Green S编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 前瞻性 Meta 分析是对结果已知的符合 Meta 分析要求的研究（通常是随机试验）进行的 Meta 分析。
- 前瞻性 Meta 分析能够假设个体研究的结果、前瞻性应用选择标准、预设分析报告。作为 Meta 分析而非多中心试验，它允许计划书中纳入的研究的改变，尽管按原计划的 Meta 分析能获得最大的效能。
- 前瞻性 Meta 分析常由一个协作组开展，且常收集和分析单个病例数据。

- 计划书对于前瞻性 Meta 分析非常重要，并可能作为 Cochrane 系统评价的计划书被发表。Cochrane 前瞻性 Meta 分析方法学组负责前瞻性 Meta 分析的注册，并有能力对前瞻性 Meta 分析的实施提供建议。

19.1 引言

19.1.1 什么是前瞻性Meta分析？

一个合理实施的系统评价在确定潜在的符合标准的研究之前应该确定将要研究的问题。然而，系统评价从本质上来讲是回顾性的，因为试验是在已经完成和报道了之后才被纳入的（Pogue 1998，Zanchetti 1998）。如果系统评价研究问题的主要内容的选择是基于一个或者多个阳性试验的结果报道，那么应该认识到单个随机对照试验的结果可能会对回顾性系统评价的结果带来偏倚。这可能会影响：

- 研究的选择标准 （即认为合格的试验类型）
- 目标人群的选择
- 干预的性质
- 对照的选择
- 将进行评估的结局及其测量

例如：一个系统评价中一个研究的结果与其它研究的结果相反，作者对这种明显异质性的可能解释进行了讨论，认为临床上可以解释。基于此，作者决定排除该研究。这也许是一个合理的决定，但这是在所有研究的结果已知的情况下做出的，因此在本质上是有问题的。

如第10章（10.2节）所述，对一个试验结果的认识可能会影响相关结果的发表。即使已发表的试验，其结果也有可能存在选择性报告的情况，从而在系统评价中引入更隐蔽的发表偏倚。（Chan 2004）

前瞻性Meta分析（PMA）是结果还未知的符合meta分析要求的研究（通常是随机试验）进行的meta分析，与累积Meta分析、单个病人数据meta分析具有共同点（Egger 1997）。前瞻性Meta分析可以克服回顾性Meta分析中一些公认的问题（see also Chapter 18, Section 18.5），通过：

- 可在单个研究结果未知情况下做出特定假设

- 可前瞻性应用研究的选择标准
- 可在单个试验得到结果之前给出采用预计分析方法（包括亚组分析）的分析结果报告，这就避免了解释依赖于数据的亚组的分析结果的潜在困难。

系统评价还取决于作者获得所有随机分组患者相关结局数据的能力。如果发表的论文对这些数据没有报道，这可能难以获取相关数据。由于大部分前瞻性Meta分析会收集和分析单个病人数据，这就能够克服这个问题，同时如果数据允许，可以进行时间事件分析是它的一个优点。基于病人因素的亚组分析在只依赖于汇总数据的情况下可能会得到误导性的结果，这也突出了单个病人数据的又一个优点。前瞻性Meta分析也为试验的设计、数据收集和其它临床试验过程的标准化提供了独特的机会。例如，研究者可能会同意在每一个试验中使用相同工具来测量一个特定的结果并且在同一时间点测量。在一个预防儿童肥胖的Cochrane干预性系统评价中，某些结局指标测量的异质性和不可靠性使试验间的这些数据难以合并（Summerbell 2005）。前瞻性Meta分析对于这个问题提出了公认的一整套标准，以解决缺乏标准化的一些问题（Steinbeck 2006）。

19.1.2 前瞻性Meta分析和大型多中心试验的区别是什么？

对于认识到单个足够大的试验的好处但又没有能力开展临床试验的人来说，前瞻性Meta分析是一个具有吸引力的选择（Simes 1987，Probstfield 1998）。它是一种有用的方法，如需大样本量来保证足够的检验效能，但单中心、大规模试验却不可行时。这可能是因为当研究信息被认为将会“流失海外”时，出于局部利益的考虑而阻止研究者参与。这同样导致了进行罕见病研究时，想要得到大量的试验研究对象是非常困难的。

因此，对研究者而言，另一种选择就是在当地实施自己的研究，同时与开展类似研究的研究者合作，在每个试验结束后将结果合并起来。这使单个研究者保持一定自主权的同时也可以开展合适的Meta分析。在另一种情况下，特别是在缺乏强制性的随机试验注册时，它可能是有用的，即对同样的问题存在两个或多个的试验时，研究者却不知道有其它的相似试验。一旦发现有类似的试验，研究者可以合作（如有必要可对数据搜集进行调整）并计划在Meta分析中前瞻性地合并他们的研究结果。

前瞻性Meta分析和多中心试验的另一区别是前瞻性Meta分析没有规定计划书在不同研究间保持一致。研究设计的多样性可能被一些人视为前瞻性Meta分析的应有特征，因此在研究人群或干预措施方面的预期程度的变异是可以接受的。体弱和损伤干预技术

的合作研究（FICSIT，Frailty and Injuries：Cooperative Studies of Intervention Techniques）就是一个前瞻性meta分析的例子，包括8项在体弱老年人群中进行以运动为基础的干预的研究（Schechtman 2001）。FICSIT的8个研究单位，采用各自特定的指标来确定干预措施和评估标准以及不同的纳入标准（除了所有参与者是老年人之外）。这种在研究设计中引入系统变异性方式，也被称为“Meta试验设计”，也是前瞻性Meta分析的一种可行方法（Cholesterol Treatment Trialists'（CTT）Collaborators 2005）。

19.1.3 哪些卫生保健领域使用过前瞻性Meta分析方法？

近年来，前瞻性Meta分析在心血管疾病（Simes 1995，世界卫生组织 – ISI降血压治疗试验协作组1998），儿童白血病（Shuster 1996，Valsecchi 1996）和儿童和青少年肥胖（Steinbeck 2006）中被使用过。另外，在一些已知领域，例如传染病，使用前瞻性Meta分析的机会被大量错过（Ioannidis 1999）。Cochrane前瞻性Meta分析方法学组网页上有正在进行和已完成的前瞻性Meta分析列表，也可以找到其它更多的信息（Ghersi 2005）。

19.1.4 我们需要什么资源？

前瞻性Meta分析是有意义的工作，不应该草草展开。它可能需要多年的时间来完成，也需要一个固定的、持续的、适当人员配置和足够资金支持的秘书处。一旦前瞻性Meta分析协作组成立（见19.2节），就需要诸多资源以确保该小组多年运行，这通常比回顾性的单个病人数据的系统评价（见18章）需要更长的时间。秘书处需要定期的举行电话会议、面对面会议（至少每年一次）、信件通讯、更新联络信息以及运用其它机制来维持协作组，这类似于多中心随机试验的协调中心所从事的活动。建立秘书处的一个好处是有助于促进遵守前瞻性Meta分析计划书，并鼓励完成单个研究的随访。

19.2 前瞻性Meta分析的协作本质

19.2.1 协作小组

正如单个病人数据的Meta分析（见第18章，第18.2.1节），大部分前瞻性Meta分析是由协作小组来完成和发表的。协作小组应包括参与每个参与的试验的代表，它通常有一

个指导小组或秘书处进行日常项目管理。协作小组可以选择建立一个解决具体问题，并给指导小组或秘书处提供关于临床的、技术的或其它问题的建议的特别小组。

19.2.2 协商合作

和单个病人数据的Meta分析（见第18章，第18.2.2节）一样，建立和协调参与试验的研究者之间的协商合作关系对前瞻性Meta分析的成功是至关重要的。然而，前瞻性Meta分析首先关注的并不是从单个的试验中查找和获得数据。因为协作组是在得出结果数据前建立的，所以协作工作重点，至少在最初阶段是对每个参与研究的研究人群、设计和数据收集方法达成一致。当前瞻性Meta分析协作小组成员同意参与项目时，他们必须同意共同的核心计划书和收集共同的核心数据资料。单个试验可以修订本地计划书或增加其他数据项目，但是他们必须保证这些不会影响核心的共同计划书的要素。

在一个前瞻性的Meta分析中，要努力查找所有的在研试验，即最大限度地保证精确性又能避免可能由于排除已知结果的研究而产生的偏倚。为了保证单个研究适合纳入前瞻性meta分析，这需要证据来支持，试验结果在试验同意进入前瞻性Meta分析后不被试验数据监控委员会以外的知道。这应该是理想证据形式，即试验被预先注册（Laine 2007）。协作组从每个协作的试验小组获得一个明确的（并签署的）协议同样是可取的。这样做是为了鼓励单个研究者们提供实质性的贡献、并认同前瞻性Meta分析的概念和计划书的细节。

19.2.3 保密

关于数据匿名性和安全性的保密问题类似第18章单个病人数据的Meta分析中所描述的问题（第18.2.3节）。前瞻性Meta分析的具体问题包括：做出详细计划关于如何处理前瞻性Meta分析中已经完成并即将发表结果的试验以及与数据及其安全性监控有关的问题（包括单个试验中期分析或可能的前瞻性Meta分析合并的中期分析的影响）。（见19.5.2节）

19.3 前瞻性Meta分析的计划书

19.3.1 计划书应包括的内容？

所有的前瞻性Meta分析都应该有一个公开的计划书。前瞻性Meta分析的计划书在内容上和一个单个试验的类似。它的基本要素如下并总结于框19.3.a。

目标、纳入标准和结局指标

所有计划书的第一个重要步骤均是确定研究假设，然后建立纳入标准。例如，尽管可能纳入其它的研究设计类型，前瞻性Meta分析可能要求纳入的研究对研究对象和干预措施进行随机分配。如果是随机的，则单个试验可能需要选择一个共同的随机方法，或者至少使用相同的分层因素。研究人群的特征需要明确，这是对试验组和对照组的最低要求。计划书还应该具体说明需要测量哪些结局指标，哪些是首要哪些是次要的，何时和如何进行测量，以及其它必需的研究设计特征。如果前瞻性Meta分析是新建的，则其中的每个试验可能使用完全相同的试验方案。

检索方法

计划书应该详细描述如何检出在研试验，包括潜在的合作者是如何（将）被找到并加入的。

试验细节

检出的纳入试验的详情应在计划书中列明。此列表可能包括每个试验中的预期参与者数量和进度表。计划书应该在提交注册时应包括一份对结果已知（对于试验数据监测委员会之外的任何人）的试验的声明。

在确定纳入到前瞻性Meta分析时，只有其试验结果未知的研究才能被纳入。如果合适的试验已被确定，但却因为它们的研究结果已知而没有被纳入到前瞻性Meta分析中，则计划书中应该说明如何处理这些数据。例如，可以采用这些研究中合并或单个病人数据来进行次敏感性分析（secondary sensitivity analyses）。计划书中还应该描述如何处理在前瞻性Meta分析的进行中后续发现的试验。

分析计划

计划书应像单个病人数据Meta分析（见18章）一样概述数据收集和分析的计划。这包括样本大小和效能计算（对前瞻性Meta分）、任何将要进行的中期分析以及计划的亚组分析的详细信息。在前瞻性Meta分析中，对于解决主要假设以外的其他问题的策略也

可应该阐述。只要将要纳入分析的研究的结果未知，这些问题也可以被添加，即不是“基于数据”的研究问题。值得注意的是，这里可能有只对前瞻性Meta分析适用而不能对单个试验使用的分析方法，如亚组分析。

一般要求纳入前瞻性Meta分析的试验研究者们同意提供单个病人的数据。计划书中应该描述如果其中某些研究者也许因为考虑保密性或知情同意而不能（或不愿意）提供个体病人数据，将出现何种情况。例如，前瞻性Meta分析秘书处会接受合适的汇总数据（可进行两阶段的分析：使用个体病人数据计算出每个研究的效应估计值，然后使用标准化Meta分析方法合并。）吗？计划书应说明是否会进行周期性（例如每5年）的数据收集来定期更新前瞻性Meta分析，以及以后试验者需要在何时提供最新的、长期的结果数据。

管理与协调

前瞻性Meta分析计划书应概述项目管理结构（包括所有委员会，见19.2.1节）、数据管理的程序（如何收集数据、要求的格式、何时提交以及质量保证程序等，见第18章，第18.3节）和由谁将负责统计分析。

发表政策

前瞻性Meta分析的一个关键内容是发表政策。必须有一项关于作者（如确定以团体名义发表，同时列出单个作者）。关于原稿撰写的政策也是很重要的。例如，需要说明稿件在出版前应该发给所有作者并征求意见。像那些合作研究中经常成立的一样，这可能也有一个编委会。

前瞻性Meta分析中一个特殊的问题是是否应该发表单个研究以及发表的时间（这一般不会出现在多中心研究和单个病人数据的Meta分析中）。大多数研究者希望在对前瞻性Meta分析做出贡献的同时也单独发表自己的研究，并且很可能是在前瞻性Meta分析发表之前，以避免相同数据的重复发表。相类似的，任何前瞻性Meta分析发表的文章都应该明确指出纳入数据的来源以及已经发表的相同数据。前瞻性Meta分析的计划书也需要说明如果参与试验的结果不能在指定的时间内发表会有什么结果。这可能由于资金不足提前终止或仅仅是在预定指定的时间内仍未发表。计划书还应该说明如何处理那些违背前瞻性Meta分析参与协议的试验。

框19.3.a　前瞻性Meta分析计划书的要素

目的：

明确假设或目标

方法：纳入标准：

- 试验设计的要求（例如需要随机、最少随访时间）
- 患者人群的要求
- 干预和对照措施的要求
- 结局指标：明确的主次要结局指标、定义、测量方法和时间
- 亚组详情

方法：检索方法：

- 详细描述如何检出在研试验

方法：数据收集和分析:

- 试验详情：
  - 纳入试验的详情一览表
  - 一份对结果已知（对于试验数据监测委员会之外的任何人）的试验的声明。只有试验结果未知的试验才能被纳入到前瞻性Meta分析。
  - 是否同各个试验的代表（如：赞助者或主要研究者）签署了协作协议。
- 分析计划：
  - 详述样本大小和效能计算（针对前瞻性Meta分析）、中期分析、亚组分析等
- 管理与协调：
  - 管理结构和委员会的详细信息
  - 数据管理（数据收集、数据格式、何时提交以及质量保证程序等）
  - 统计分析的负责.
- 发表政策：
  - 作者政策（例如：以‘组’的名字发表）
  - 编委会（成员和职责）
  - 定稿政策（例如：征求所有试验者意见）

19.3.2　计划书的发表

如果准备做一个Cochrane系统评价，则前瞻性Meta分析的计划书需要提交给恰当的Cochrane系统评价小组，以便能在Cochrane系统评价数据库中找到。否者，计划书需要发表在其它地方（如：CTT/PPP Protocol（胆固醇治疗试验协作组2005））。前瞻性Meta

分析计划最好在Cochrane前瞻性Meta分析方法学组注册（见框19.6.a）并且必须至少每年更新。按照国际规定，前瞻性Meta分析中的每个试验在纳入第一个受试者前都必须先在一个公开访问的、世界卫生组织承认的一级注册机构注册（www.who.int/ictrp/network/list_registers）。

19.4 前瞻性Meta分析的数据收集

通常前瞻性Meta分析中的单个试验一旦完成和发表就可以提供单个病人数据。前瞻性Meta分析的优势在于试验者可以预先决定需要收集那些数据及其数据形式，使得对数据的重新定义和编码与回顾性的单个病人数据相比，较少出现问题。前瞻性Meta分析应该结合现有数据交换标准（data interchange standards）制定一个数据传输协议，如临床数据交换标准协会（Clinical Data Interchange Standard Consortium （CDISC；www.cdisc.org））所制定的标准。

前瞻性Meta分析秘书处一旦接收了数据，将使用同单个病人数据Meta分析的一样程序来严格审查数据，包括缺失和重复数据、数据合理性的核对、随机化类型的评估和确保所提供的信息是最新的（见第18章，第18.4.4节）。数据查询将在数据被纳入到最后的分析数据集之前同单个试验者直接协商解决。

19.5 前瞻性Meta分析的问题

19.5.1 一般方法

大多数前瞻性Meta分析使用类似于回顾性单个病人数据Meta分析的一般分析技术。这些技术在第18章（第18.4节）有详细描述，包括了一般分析方法和时间事件分析方法（如果合适）。病人水平数据允许使用统计效能更高的的亚组分析和多水平模型来探讨干预措施和病人特征之间的关系，及在某些情况下的预后模型。第18章（第18.4.7节）描述了一些可以用来处理这类数据的软件包。

19.5.2 中期分析和数据监测

在单个临床试验中使用中期数据分析和安全性监测已经越来越普遍。前瞻性Meta

分析提供了使用所有试验中的中期数据的一个独特机会。这些数据可能被合并分析或是单独使用，然后在参与试验的数据监测委员之间共享结果。

使用中期分析也引起了一些伦理上的问题。例如，在干预措施的总效益（如主要结局指标）被证实之后还适合再继续做随机吗？当关注的临床亚组结果未知或对于不常见的结局指标的，例如，且结果显示一个结局有益或有害，但另一结局不明确时，研究者应该为获得进一步的净临床效益而继续试验吗？

如Hillman and Louis建议，如果每个试验都有自己的数据监测委员会，那么他们之间交流对解决这个问题是有用的(Hillman 2003)。各委员会要了解前瞻性Meta分析中纳入的其它试验及其结果，因为这些外部因素可能会影响某一监测委员会的决定。例如：当证据显示干预措施有效时是否应该提前结束试验。相反地，可能会认为由于考虑中期安全性结果，来自所有参与试验的新的安全性数据可能会减少单个试验错误地提前中止的几率。 因此，它将有助于各试验的数据安全监测委员会达成共识，即在亚组、少见结局指标（或是净临床效益）和前瞻性Meta分析的目的达到之前单个试验不能终止。

另一种可能的选择是如果现有亚组已能得出在临床和统计上有意义的结果，应考虑限制病人进入试验。在任何情况下，当单个试验的结果可用时，都适合使用Whitehead描述的序贯临床试验方法（Whitehead 1997）去得到前瞻性Meta分析严格的终止规则。

19.6 本章信息

作者：Cochrane前瞻性Meta分析方法学组Davina Ghersi, Jesse Berlin 和 Lisa Askie

本章引用格式：Ghersi D, Berlin J, Askie L. Chapter 19: Prospective Meta-analysis. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook. org.

框19.6.a　Cochrane前瞻性Meta分析方法学组

Cochrane 前瞻性 Meta 分析方法学组（PMA MG）的作用是：

- 为前瞻性 Meta 分析提供一个注册平台
  - Cochrane（通过 Cochrane 系统评价组）
  - 非 Cochrane（通过 Cochrane 前瞻性 Meta 分析方法学组）
- 为提交注册的计划书提供一个评估平台，以确保它们确实是前瞻性 Meta 分析。这可能又以下方式实现：
  - 为 Cochrane 系统评价组成员提供培训（如：编辑和同行评审）
  - 前瞻性 Meta 分析方法学组同行评审计划书
  - 前瞻性 Meta 分析的研究者和同行评审一览表
- 为前瞻性 Meta 分析制定合适的方法学标准
- 为正在（打算）从事前瞻性 Meta 分析的人们提供建议和支持

该组是对正在进行、已经完成或是有兴趣进行前瞻性 Meta 分析的卫生保健研究的所有领域的人敞开的。若要加入，则需要在前瞻性 Meta 分析方法学组问卷（可到 PMA 网站下载，见下）上做出详细承诺。成员被要求每年更新这些信息。

网址: http://www.cochrane.org/docs/pma.htm.cochrane.org/docs/pma.htm

19.7　参考文献

Chan 2004

Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004; 291: 2457-2465.

Cholesterol Treatment Trialists' (CTT) Collaborators 2005

Cholesterol Treatment Trialists' (CTT) Collaborators. Efficacy and safety of cholesterol-lowering treatment: prospective Meta-analysis of data from 90 056 participants in 14 randomised trials of statins. The Lancet 2005; 366: 1267-1278.

Egger 1997

Egger M, Davey Smith G. Meta-analysis: potentials and promise. BMJ 1997; 315: 1371-1374.

Ghersi 2005

Ghersi D. Cochrane Prospective Meta-analysis Methods Group. About the Cochrane Collaboration (Methods Groups) 2005, Issue 2. Art No: CE000132.

Hillman 2003

Hillman DW, Louis TA. DSMB case study: decision making when a similar clinical trial is stopped early. Controlled Clinical Trials 2003; 24: 85-91.

Ioannidis 1999

Ioannidis JPA, Lau J. State of the evidence: current status and prospects of Meta-analysis in infectious diseases. Clinical Infectious Diseases 1999; 29: 1178-1185.

Laine 2007

Laine C, Horton R, DeAngelis CD, Drazen JM, Frizelle FA, Godlee F, Haug C, Hebert PC, Kotzin S, Marusic A, Sahni P, Schroeder TV, Sox HC, Van der Weyden MB, Verheugt FW. Clinical trial registration: looking back and moving ahead. Canadian Medical Association Journal 2007; 177: 57-58.

Pogue 1998

Pogue J, Yusuf S. Overcoming the limitations of current Meta-analysis of randomissed controlled trials. The Lancet 1998; 351: 47-52.

Probstfield 1998

Probstfield J, Applegate WB. Prospective Meta-analysis: Ahoy! A clinical trial? Journal of the American Geriatrics Society 1988; 43: 452-453.

Schechtman 2001

Schechtman K, Ory M. The effects of exercise on the quality of life of frail older adults: a preplanned Meta-analysis of the FICSIT trials. Annals of Behavioural Medicine 2001; 23: 186-197.

Shuster 1996

Shuster JJ, Gieser PW. Meta-analysis and prospective Meta-analysis in childhood leukemia clinical research. Annals of Oncology 1996; 7: 1009-1014.

Sim 2006

Sim I, Chan AW, Gulmezoglu M, Evans T, Pang T. Clinical trial registration: transparency is the watchword. The Lancet 2006; 367: 1631-1633.

Simes 1987

Simes RJ. Confronting publication bias: a cohort design for Meta-analysis. Statistics in Medicine 1987; 6: 11-29.

Simes 1995

Simes RJ. Prospective Meta-analysis of cholesterol-lowering studies: the Prospective Pravastatin Pooling (PPP) Project and the Cholesterol Treatment Trialists' (CTT) Collaboration. American Journal of Cardiology 1995; 76: 122c-126c.

Steinbeck 2006

Steinbeck KS, Baur LA, Morris AM, Ghersi D. A proposed protocol for the development of a register of trials of weight management of childhood overweight and obesity. International Journal of Obesity 2006; 30: 2-5.

Summerbell 2005

Summerbell CD, Waters E, Edmunds LD, Kelly S, Brown T, Campbell KJ. Interventions for preventing obesity in children. Cochrane Database of Systematic Reviews 2005, Issue 3. Art No: CD001871.

Valsecchi 1996

Valsecchi MG, Masera G. A new challenge in clinical research in childhood ALL: the prospective Meta-analysis strategy for intergroup collaboration. Annals of Oncology 1996; 7: 1005-1008.

Whitehead 1997

Whitehead A. A prospectively planned cumulative Meta-analysis applied to a series of concurrent clinical trials. Statistics in Medicine 1997; 16: 2901-2913.

WHO - ISI Blood Pressure Lowering Treatment Trialists' Collaboration 1998

WHO - ISI Blood Pressure Lowering Treatment Trialists' Collaboration. Protocol for prospective collaborative overviews of major randomised trials of blood-pressure-lowering treatments. Journal of Hypertension 1998; 16: 127-137.

Zanchetti 1998

Zanchetti A, Mancia G. Searching for information from unreported trials - amnesty for the past and prospective Meta-analysis for the future. Journal of Hypertension 1998; 16: 125.

（何佳、李玲译，李伦、秦天强、岑啸初审）

第二十章　定性研究与 Cochrane 系统评价

作者：代表 Cochrane 定性研究方法学组的 Jane Noyes, Jennie Popay, Alan Pearson, Karin Hannes 和 Andrew Booth。

本文选自工作手册 5.0.1版本。有关如何引用它的指南，见20.4节。这些材料还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 定性研究的证据在提高那些针对方针政策、循证实践和用户决策而制定的系统评价的价值方面有重要作用。
- Cochrane 系统评价所纳入的结局研究大多数都包含了定性研究或与自身相关的定性研究。
- 定性研究通过以下 4 种方式提高 Cochrane 干预性系统评价的质量：
  - 指导系统评价：定性研究的证据有助于定义并精炼系统评价的问题，从而确保

系统评价纳入恰当的研究并关注重要的结局指标。

- 充实系统评价：在搜索有效性证据的同时对定性研究的证据进行整合。
- 拓展系统评价：通过对特定定性研究证据的检索来解决那些和有效性系统评价直接相关的问题。
- 补充系统评价：通过对其它独立但具有补充性质的定性系统评价的定性证据的综合，从而解决效力/有效性之外的其他方面的问题。

- 目前已有大量适用于 Cochrane 干预性系统评价目的和范围的定性证据综合方法。
- 定性研究的综合是一个充满争议和亟待改进的领域。Cochrane 定性研究方法学组专门创建了一个用于讨论和改进该领域方法学的论坛。

20.1 引言

本章的主旨是概述可能用于指导、充实、拓展和补充Cochrane系统评价的定性研究的方法。定性研究并不是用来测量干预的效果，而帮助对Cochrane系统评价结果的进行解释、说明和应用。因此，源于定性研究的证据可以完善定量研究的系统评价。

本章的目的是使作者能：

1、明确系统评价的类型及系统评价中能够用定性证据充实和拓展的问题；

2、在决定用定性研究证据合并来补充Cochrane系统评价时，明确证据来源和方法学上的问题；

3、了解一些合并定性研究证据的途径和方法；

4、如有需要，能获得更多的信息、建议和资源。

本章分为两个部分：第一部分（20.2节）提供了一些关于Cochrane系统评价中纳入定性研究证据的思考和指导，包括资源的影响。第二部分（20.3节）则着重针对方法学问题，阅读的重点和Cochrane定性研究方法学组扮演的角色及其详细情况展开了广泛的讨论。我们也提供了一个相应的范例，用以说明综合的定性研究证据如何补充已有的关于效果的Cochrane系统评价。

20.2 Cochrane系统评价中整合定性研究证据：概念和问题

20.2.1 定性研究的定义

定性研究者研究自然环境中的事物，从而试图理解或解释这些事物的相关现象人们所赋予的意义（Denzin 1994）。定性研究旨在发掘所研究的事物对于研究课题更深层的意义。定性研究运用说明和自然的方法来进行研究，同时优先考虑定性研究数据对重要研究问题或现有信息做出的贡献。

在卫生保健中，理解定性研究的证据对于系统评价的价值时，我们必须考虑到证据的多样性和弥散性（Popay 1998b，Pearson 2005）。定性研究包含以下诸多方面：一系列原理，研究设计及涉及深入定性访谈的特定技术；参与性与非参与性的观察研究；核心团队；文献分析及许多其它的数据收集方法（Pope 2006）。针对不同类型的数据，我们可采用诸如现象学、人种学、扎根理论、行为研究、案例研究和许多其它等多样化的理论和方法来进行相关研究的设计和数据分析。同时，理论依据和研究者的观点对定性研究的资料分析影响重大，并且基于此对其它研究进行概括。

在经验科学中，一个理论或假说的立足点完全取决于对其有利的证据的数量和特征。支持性证据的相对权重使我们能在相互竞争的理论中做出选择。在自然科学中，认识的产生涉及用假设所得出的结果对假设或一组假设进行验证并通过实验和观察来验证这些结果的真实性。

医务人员需要寻找证据来证明各种活动和干预措施的价值，因此，所需证据类型必须基于该行为的性质和目的。对许多研究问题，如父母亲的观念和儿童疫苗的接种（Mills 2005a，Mills 2005b），定性研究是一种适当和可取的方法。

20.2.2 Cochrane系统评价中定性研究证据的使用

Cochrane干预性系统评价的主要目的是研究干预措施较对照措施的效果如何，若干预有效，则评估其效果的大小。而高质量的随机试验是Cochrane协作网在这个方向上努力的核心所在。但是在所有Cochrane系统评价中都纳入定性研究的证据却是不恰当也不可能的。

但是，人们逐渐认识到在确证系统评价是否充分体现了其对于决策、实践以及用户决策的最大价值的过程中，来自以下两种研究的证据至关重要：针对那些干预的提供者

和接受者的体验的定性研究；评估干预实施过程中诸多影响因素的研究（Mays 2005，Arai 2005，Popay 2005）。

不久前，卫生领域才承认定性研究证据对于干预评价的现实意义，但目前在评价健康干预措施的研究中加入定性成分较为常见（Pope 2006）而评价复杂的干预措施（如不同的医疗保健服务模型）时则会选择“混合方式”的途径。因此， Cochrane系统评价所纳入的结局研究越来越倾向于嵌入定性研究或与之相结合。而Cochrane系统评价的作者也越来越希望了解如何利用定性研究的证据来提高自己的系统评价对潜在使用者的实用性和利用度。

定性研究的综合证据通常探讨如下的问题：人们怎样经历疾病过程？为什么干预措施有效（或无效）？对那些人有效？在什么条件下有效？在一些系统评价中（特别是针对医护服务问题的系统评价）亟需通过利用定性研究证据解决以下问题：在获取医疗服务过程中存在哪些阻碍因素和促进因素，或是这些因素对人群及其经历和行为会产生怎样的影响？这些证据可以通过诸如人种学和求助行为的访谈研究等方法产生。定性研究证据可以通过增加对以下方面的了解来帮助解释系统评价结果：了解哪些参与了完善、实施或接受干预措施的人群的体验；了解他们重视或不重视干预的哪些方面及原因。这些类型的定性研究证据能帮助我们深入了解包括如其它政策的发展的影响，促进或阻碍项目、服务或治疗顺利进行的因素，这些影响干预措施的外部因素以及如何对需要大规模推行的干预措施进行调整（Roen 2006）。

我们总结了定性研究影响针对卫生政策和实践的Cochrane干预性系统评价的4种方式：

1、指导系统评价（informing）：定性研究的证据有助于定义并精炼系统评价的问题。这确保系统评价纳入了恰当的研究并针对重要的结局指标，从而使系统评价对潜在使用者的利用价值得到最大化。

2、充实系统评价（enhancing）：在搜索有效性证据的同时对定性研究的证据进行整合。与试验相关的定性研究证据可被用于探讨干预措施实施的问题。在20.2.3部分详细探讨了将定性研究与随机试验一起实施的问题。

3、拓展系统评价（extending）：通过对特定地检索和整合定性研究证据来解决和有效性系统评价直接相关的问题。

4、补充系统评价（supplementing）：通过整合定性研究证据来解决效力/有效性之外的其它方面的问题。

用于拓展和补充系统评价的定性研究的综合均采用了多水平或平行的综合方式（详见20.3.2.5）。目前尚无仅针对定性研究证据的单一Cochrane系统评价模板。

Cochrane公共卫生和健康促进组已经专门制定了一份指南来指出定性研究对哪些类型的系统评价和问题有价值（见21章）。上述系统评价主要针对以下一些问题：1）干预措施是否有效（效力）；2）为什么干预有效或无效——包括研究其如何生效（可行性，适当性和意义）；3）参与者对干预过程的体验如何？

当使用定性研究来充实和拓展Cochrane干预性系统评价时，我们应在系统评价的方法部分的‘数据收集和分析’下设置独立的标题来描述规范、筛选、评价和综合定性研究所使用的方法。

20.2.3 考虑那些随机对照试验内部或与之并列的定性研究

随着“混合方法”演进至用于评估复杂干预（如卫生服务提供策略）的效果时，Cochrane干预性系统评价所纳入的研究也越来越多地嵌入或结合定性研究，即便是相应定性研究的证据不会和试验的证据本身的结果发表在同一处。例如，我们在Box 20.3a中总结的系统评价范例里，虽然并不是所有的定性研究资料都得到了分析或发表，但该Cochrane干预性系统评价所纳入的6项试验中有5项包含了定性成分或存在与之相关的定性研究。值得注意的是，定性的成分在试验研究报告中总是不会被提及。事实上，一些研究只有通过联系主要的调查员后才被了解到的。

在研究在随机对照试验内部或与之并列的定性研究时，需要考虑以下一些问题：

1. 检索定性研究证据：通过主题检索限定相关试验而获得定性研究证据的检索策略既不具有全面性，也无代表性。上述检索策略的目的不是为了检索定性研究无误，而实际上是一种特异性地有目的地排除其它多种定性研究类型的措施。
2. 用于探讨患病体验的定性研究证据的综合：如果疾病体验是关注的焦点，那么通过相关试验的检索策略得到的定性研究资源就并不非得提供一个整体全面的观点。在这些情况下，我们应该考虑并提倡进行多水平或平行的综合（见20.3.2.5）。理想情况下，作者应与定性研究者和信息专家共同制定定性研究的检索策略，确保检索出其它相关的研究。
3. 借助定性研究的综合探讨干预措施实施过程中的问题：如果有关实施的问题是关注的焦点，那么试验所嵌入或并列的定性研究的证据与之相关程度最高。这

些实施证据最有可能是由混合方法研究产生的，同时包含定性和定量研究的证据。需要采取必要的步骤，如进行另外的针对性的检索和联系试验的研究者，来检索与试验相关的所有定性研究资源。

4. 被Cochrane干预性系统评价排除的研究中的定性证据：有时候，一项试验虽然没有达到Cochrane干预性系统评价的纳入标准（例如由于不能接受的偏倚的风险），但其中嵌入的或结合高质量的定性研究。根据指导原则，如果定性研究证据质量可靠，那么就可以将定性证据纳入系统评价。

20.2.4 关于资源

今后在Cochrane系统评价中应用定性研究证据时将不可避免地对系统评价作者和协作网工作组（CRGs）造成很多影响。资源的限制可能左右补充性定性证据的综合与系统评价结合的程度。将定性研究结果纳入到Cochrane系统评价时作者应考虑以下一些情况：

- 研究团队是否配备有具备定性研究专业知识的专家或能获得来自资深定性综合研究者的建议？
- 是否需要额外的培训？
- 预算是否够应对对时间和资源额外要求？
- 研究团队是否能获得适当的数据库和期刊？
- 研究团队是否能和熟悉定性研究检索的信息学专家取得合作？
- 负责该系统评价的协作网系统评价小组能否支持定性研究证据的整合以及是否有足够的资源支持系统评价编辑的整个过程？

20.3 定性研究证据的综合

20.3.1 综合定性研究证据来补充Cochrane干预性系统评价的实例：直接督导疗法与结核病（TB）

在探讨定性研究证据综合的方法学之前，我们先提供一个范例（Box 20.3.a）。该系统评价的全文刊登在Journal of Advanced Nursing杂志上（Noyes 2007）。其平行的定性研

究证据综合用于拓展和补充一篇关于以直接观察疗法（即在监督下服药）作为提高病人对结核治疗方案依从性的干预措施（Volmink 2007）的Cochrane系统评价，该系统评价纳入了6个随机对照试验，但发现直接观察疗法（DOT）较家庭治疗没有显著的统计学差异。而与之伴随的定性研究证据的综合关注的是对于结核病治疗的经历和认知，探讨来源于这些研究的证据是否有助于解释随机对照试验的结果，从而促进结核治疗政策的制定。在上述过程中定性研究证据的综合所解决了Cochrane系统评价所不能解决的问题诸如直接督导疗法的适宜性及其促进干预实施的方式等一类问题。

Box 20.3.a　直接督导疗法与肺结核：定性研究证据综合的总结

背景：直接观察疗法是世界卫生组织诸多为了加强肺结核患者管理和依从性的干预措施中的一种（Maher 1999）。直接观察疗法要求肺结核患者定期在卫生工作者或其他指定人员监督下服药。一项针对直接观察疗法的 Cochrane 干预性系统评价发现了该疗法和病人家庭自主治疗的效果比较时证据相互冲突。为了完善这篇系统评价，我们综合了那些涉及患有或可能患有肺结核的人群、医疗服务提供者及政策制定者的定性研究的证据，来探讨他们对肺结核及其治疗的体验和观点。研究结果用来深入解释和探讨 Cochrane 干预性系统评价本身及其对研究、政策和实践的影响。

问题回顾：需要解决的两大研究问题：

1、在获取和开展肺结核治疗的过程中，促进和阻碍因素有哪些？

2、能否通过探索定性研究和（或）干预性系统评价中定性研究的部分来解释结果的异质性？

方法：

检索方法：对英文文献进行广泛而系统的检索，使用了以下检索词：DOT; DOTS; Directly observed therapy; Directly observed treatment; supervised swallowing; self-supervis*；与 TB；tuberculosis 并列检索。我们尝试用“qualitative”进行方法学的筛选，但发现这种方法是无用的，因为“Qualitative Research”（“定性研究”）2003 年之后才编入 Medline MeSH 主题词库，但即便是 2003 年之后的很多文献也没有以“定性研究”划分出来。我们检索了 MEDLINE, CINAHL, HMIC, Embase, British Nursing Index, International Bibliography of the Social Sciences, Sociological Abstracts, SIGLE, ASSIA, Psych Info, Econ lit, Ovid, Pubmed, the London School of Hygiene and Tropical Medicine database of TB studies（特别感谢 Dr Simon Lewin），和 Google Scholar。同时查看了相关论文的参考文献。我们还利用了一个私人网络来筛选文献。我们与系统评价中纳入的 6 个随机对照研究的主要研究人员取得联系并获取相关的定性研究。

文献筛选和评估：按照以下定义筛选文献：“主要的研究重点是肺结核患者或潜在患者及卫生服务提供者对肺结核的体验和/或看法的文献。”不论是独立的研究还是一个更大规模并混合了不同方法的研究的一部分，在数据收集和分析的过程中必须使用定性研究方法。在评价研究的方法学和理论方面的质量时， JN 和 JP 会用到两种不同的对比框架（Popay 1998a, Critical Appraisal Skills Programme 2006）。研究质量高低并不作为排除标准，但低质量的研究要被重新审核，从而判断其是否改变了综合的结果——结果是没有改变。

分析：使用主题分析法对 1990-2002 年和更新至 2005 年 12 月的数据进行综合。主题的确立通过收集资料中所有的观点，体验，见解——构建主题是为了对参与人群的体验拥有全面直观的认识。叙述性总结的方法被用来解释实验结果。

结果：纳入了来自 53 个研究的 58 个报告。从 1990-2002 年的数据综合中产生了 5 个主题，包括：社会经济环境、物质资源和主观能动性；与肺结核及其治疗有关的的解释模型和知识系统；关于肺结核的屈辱经历和公众的态度；对社会团体及社会关爱的惩处，激励和支持。2005 年更新文献中增加了两个主题：项目实施过程中的障碍，质疑了那些造成治疗失败的主要原因中起决定性作用的因素。

结论：Cochrane 系统评价并没有发现直接观察疗法和病人家庭自主治疗的效果之间有显著统计学差异，因而可以认为并不是直接观察疗法本身改善了治疗结局。上述 6 个随机对照试验比较了直接观察疗法与自主疗法之间的 8 个变量，发现这些变量随着结核病人的需要程度不同而差异显著。直接观察疗法的变量根据监督的人员，监督地点和监督频率的不同而变化显著。定性研究的综合指出，在决定一种特定类型的直接观察疗法效力强弱时，上述因素至关重要。研究还强调了社会经济因素和药物的副作用对病人求医习惯养成和治疗依从性的影响。具体来说，设置监督员的方法并没有增加结核病人对医疗卫生服务和药物的依从性。在治疗中有时还会需要监督员，但是，当主要的研究重点是人员监督而不是实际支持时，观察起到的效果将会最小。设立监督员直接进行观察将会对那些害怕泄露隐私的人群产生最消极的影响，例如那些曾遭受过性别歧视的妇女。相反，强调以病人为中心的支持治疗方案将更有可能增加病人对医疗的寻求和依从性。定性研究证据还会帮助我们洞悉那些肺结核患者认为最有帮助的支持疗法的类型。一般情况下，设置观察者的价值大小主要取决于观察者和服务能否适用于被观察者多种多样的个人情况（年龄，性别，工作单位，地理位置，收入等）。鉴于这些不同，研究结果指出，应根据当地情况以病人为中心开展治疗而不是单一地全球性干预。

20.3.2 方法学问题

定性研究证据综合方法学上主要的挑战在于检索策略的设计和执行，研究质量的评价和适当的综合方法。

20.3.2.1 检索策略

定性研究的数据库索引系统的分析已经有了重大的进步。McMaster大学的Hedges Project已经将其测试过的方法学筛选程序的范围扩展到能够纳入来自以下数据库的定性研究：MEDLINE（Wong 2004）CINAHL（Wilczynski 2007）， PsycINFO（McKibbon 2006）和EMBASE（Walters 2006）。但是随机对照研究中收集和报告的或作为相关研究中某部分的定性研究证据依然很难检索到（Evans 2002）。MEDLINE在2003年才将“qualitative research”纳入其MeSH主题词表。CINAHL 1988引入“qualitative research”，反映护理领域研究人员定性研究特别关注，相应地关注‘生命质量’相关问题（详见17

章，17.3）。另外，因为对‘qualitative’不同的使用，所以对定性研究的定位依然存在很多问题（Grant 2004）。

另一方面，现有的针对与定性研究的相关索引词的检索策略和计划书驱动的检索策略价值有限（Evans 2002，Barroso 2003，Greenhalgh 2005）。系统评价作者必须认识到，在各大数据库里限定检索条件会导致很多有用信息的丢失。一项对综合干预性（包含定性研究证据）系统评价证据资源的调查发现，只有30%的证据能通过数据库或手工检索找到。将近半数的研究是通过“滚雪球”法找到的，另外的24%则是通过个人的知识和关系网获得（Greenhalgh 2005）。不同方法检索定性研究的检索策略还有待进一步发展。

尽管对于系统和明确地检索定性研究的检索策略的需要已达成共识，但近来也出现了关于定性研究证据的综合是否需要被全面详尽的检索的争论。有人认为，为了能够全面地解释特定现象，一种目的性更强的抽样方法可能更加适用，检索的程度是受达到理论饱和度（即理论上的相应定性研究数量）和发现‘反驳案例’的需要决定的（Dixon-Woods 2006）。而这就迫切需要提高检索方法的报告标准的质量（Booth 2006）。

20.3.2.2 严格评价

研究质量的评价（严格评价）对定性研究证据综合而言是一个相当有争议的问题。目前，人们对常规质量评价的价值产生了分歧，而且尚无足够证据来断定各种方法的严密性或附加价值。

这是一个不断发展的领域，Cochrane定性研究方法学组成员也在积极参与贡献该方面的知识和经验。但我们认为对那些反对和支持对定性研究证据的综合进行严格评价的观点进行思考和讨论很重要。

有超过100种的工具和构架可用于定性研究的评价，它们主要反映在对随机对照试验和其他形式定量研究的方法学质量的评价上（Vermeire 2002，Cote 2005）。但是，重点是要认识到关于“质量”的问题和定性研究的内容是不同的。以形成‘纳入排除’决定固定清单所呈现的正式的证据评价程序和标准可能对定性研究是不适用的（Popay 1998a，Barbour 2001，Spencer 2003）。相反，这样的工具作为探索和解释定性研究一部分可能是最合适的。基于特定数据，使用严格公式化的方法且方法学质量较低的研究可能会得出新的见解；而方法学合理的研究也可能会由于阐述不清，导致对研究现象的了解不深入。Dixon-Woods等人比较了3种结构化评价方法发现：结构化评价方法无法对一篇系统评价是否需要纳入定性研究做出很一致的判断。（Dixon-Woods 2007）。

另一个问题关注的是质量评价的时机，以及在整个过程中哪一阶段的结果具有参考意义——严格评价是该被视为建立质量门槛的障碍还是权衡纳入的研究的结果信息不同强度的过滤器？

如果作者决定将质量评价作为系统评价的一个步骤，那么他们需要使用那些具体方法（如EPPI（Evidence for Policy and Practice Information）或JBI（Joanna Briggs Institute）的方法）的框架，或者选用已经出版的定性评价的工具，框架或清单。Spencer等人写过一篇关于现有评价框架和清单的综述，可能有助于系统评价作者们选择适合的方法（Spencer 2003）。专家意见也是评价研究质量时的一个重要因素。

有关这一部分的讨论的关键参考文献见20.6.6：延伸阅读。

20.3.2.3 定性研究证据的综合

定性研究证据综合是一个通过分析和比较关注同一主题的不同证据来源对应的定性研究的概念和结果，合并单个的定性研究证据从而给出新的理解的过程。因此，定性研究证据综合本身可被理解为一个完整的研究，可以和任何干预措施或诊断试验效果的系统评价中的Meta-分析相比。它可以是一个综合或解释的过程，但是要求过程的透明化并且作者应对系统评价纳入的研究的证据进行筛选和提取；将证据分类，并合并这些类别最终得出综合的结果。但是，使用这个种方法是，重要的是要认识到定性研究证据综合的真正价值并不是描述人们对某个问题或某种治疗措施的体验，而是理解“为什么”他们有这样的感受和行为方式（Popay 2005）。

举例来说，针对慢性病人体验的原始定性研究反映了他们对自身患病原因的解释。但是此类工作也超出了单纯去解释这些原因的社会目的——它还说明了人们通过这些叙述如何在所有疾病都有到的色彩的社会背景下“重建”价值观（Williams 1984）。同样的，最近的一篇关于用药的定性研究系统评价（Campbell 2003，Pound 2005），使用了人种学的Meta分析作为对综合的拓展，通过总结那些研究间循环往复的“研究主题”，解释了为什么人们以特定方式（不用）用药。

20.3.2.4 选择适当的方法

定性研究综合中，定性研究证据纳入标准的选取需要参考许多因素，包括：

- 系统评价及其研究问题的类型和范围；
- 合并可用证据；

- 专业的团队；
- 可用的资源。

有许多不断发展的用于综合定性研究和混合方法证据的方法。与其他对此感兴趣的个人和系统评价组织一样，Cochrane定性研究方法学组成员积极参与制定并在近期着手开始评价多种现有的方法。成员已经制定出两个用于综合定性和定量卫生证据的核心文本，为方法学和流程提供更多更细致的信息和指导（Petticrew 2006，Pope 2007）。

我们认为，任何方法学上高质量的定性研究证据综合都可以用于其最适合的那一类干预性系统评价。

详细描述一系列定性研究和混合方法学证据的综合方法超出了本章的讨论范围。而且多种方法已经在许多已发表的系统评价中被使用。例如：Bayesian Meta-分析（Bayesian Meta-analysis），说明性研究综合（critical interpretive synthesis），EPPI（Evidence for Policy and Practice Information Coordinating（EPPI）Centre approach），JBI方法（Joanna Briggs Institute（JBI）approach），人种学meta分析（Meta-ethnography），meta综合（Meta-synthesis），Meta-研究（Meta-study）（后设研究），meta总结（Meta-summary），叙述综合（narrative synthesis），基于基础理论的定性证据合并（qualitative evidence synthesis drawing on grounded theory），实用性综合（realist synthesis），二次主题分析（secondary thematic analysis）。

大多数方法都有附有详细的指导（见Noblit and Hare 的人种学meta分析和Papay等的叙述总结（Noblit 1988, Popay 2006b））。Dixon-Woods 等人详细介绍了几种方法的潜力和相关问题（Dixon-Woods 2005, Dixon-Woods 2006）。但迄今为止，几乎没有针对不同方法可靠性的评估。更多内容参见20.6。

20.3.2.5 定性和定量证据综合的方法

用于合并定性和定量研究结果的大体方法有两种：

1、多水平合并（Multilevel syntheses）：定性证据（synthesis 1）和定量证据（synthesis 2）可作为独立分支或分开进行，但存在联系，最后合并其每个综合成分（1和2）的综述和结果（synthesis 3）（参见Thomas et al.（Thomas 2004））。

2、平行合并（Parallel syntheses）：定性证据（synthesis 1）和定量证据（synthesis 2）作为独立分支或独立但有联系的系统评价。之后定性研究综合（1）以平行和并列关系来帮助解释综合试验（synthesis 2）（参见Noyes and Popay （Noyes 2007））。

多水平和平行的证据综合都需要独立进行证据的系统评价，随后将其与综合试验进行合并或是两者并存。叙述综合（narrative synthesis）指南（Popay 2006b）提供了一套用于整合不同研究设计的研究结果的工具。接下来的方法学工作要求在使用不同的定性研究方法和产出一系列证据的过程中，不用考虑偏倚最小化就能将其与定量研究进行综合。（Lucas 2007）。

20.3.2.6 结论

由于人们逐渐意识到定性研究有助于提高系统评价的现实意义和实用性，所以对使用系统评价研究更多形式的证据（尤其是定性研究证据）给予了越来越多的关注。但是，在用于临床之前，无论设计还是严格生成的证据都应该考虑其质量。对于Cochrane干预性系统评价，定性研究证据也必须满足该系统评价严格的方法学要求。评价和分析定性研究证据的方法已经出现并将随着时间不断完善。系统评价过程中质量评价的各个途径的严谨性和价值还有待于通过更多的证据来确立。

20.4 本章信息

作者：代表Cochrane定性研究方法学组的Jane Noyes, Jennie Popay, Alan Pearson, Karin Hannes和Andrew Booth。

本章引用格式：Noyes J, Popay J, Pearson A, Hannes K, Booth A. Chapter 20: Qualitative research and Cochrane reviews. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.

框20.4.a　Cochrane定性研究方法学组

Cochrane定性研究方法学组（QRMG）致力于完善和支持以下的方法学工作：在系统评价中纳入定性研究证据并将这些工作成果在协作网系统评价小组内外传播。

QRMG将发挥如下作用：

- 确定定性研究证据在Cochrane系统评价中应扮演的角色。
- 整理，开发和宣传适当的方法学标准，有助于：
  - 检索Cochrane系统评价中的定性研究证据；
  - 严格评价定性研究；
  - 在系统评价中将定性研究证据与其他数据结合；
  - 通过在手册中指导作者等多种方法宣传这些方法学标准。
- 举办论坛来讨论和分析系统评价中定性研究证据的作用，并且完善严格和系统的方法加强这一作用：
  - 提高方法发展的透明度，鼓励学习了解；
  - 鼓励和促进与其他方法学组的联系和资源共享。
- 为Cochrane系统评价小组成员提供与定性研究专家的联系方式：
  - 为需要将定性研究与系统评价相结合的人员提供建议和帮助；
  - 为系统评价计划书的评估和发展提供方法。
- 培训Cochrane和Campbell系统评价小组成员。
- 维护相关的方法学数据库和存储器。
- 维护纳入定性研究证据综合的系统评价计划书或单纯关注定性研究证据的系统评价的数据库和存储器。
- 维护纳入定性研究证据综合的系统评价和单纯定性研究证据系统评价的数据库和存储器。
- 每年调查成员以确定关注的热点和进行中的课题。

小组成员参与了由York大学宣传中心（the Centre for Reviews and Dissemination at the University of York）完成的系统评价讨论和指导指南的撰写，并且一直在参与由Cochrane卫生促进和公共健康部门所制定的指南的完善工作。

网址: www.joannabriggs.edu.au/cqrmg

20.5 参考文献

Arai 2005

Arai L, Roen K, Roberts H, Popay J. It might work in Oklahoma but will it work in Oakhampton? Context and implementation in the effectiveness literature on domestic smoke detectors. Injury Prevention 2005; 11: 148-151.

Barbour 2001

Barbour RS. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? BMJ 2001; 322: 1115-1117.

Barroso 2003

Barroso J, Gollop CJ, Sandelowski M, Meynell J, Pearce PF, Collins LJ. The challenges of searching for and retrieving qualitative studies. Western Journal of Nursing Research 2003; 25: 153-178.

Booth 2006

Booth A. "Brimful of STARLITE": toward standards for reporting literature searches. Journal of the Medical Library Association 2006; 94: 421-429.

Campbell 2003

Campbell R, Pound P, Pope C, Britten N, Pill R, Morgan M, Donovan J. Evaluating Meta-ethnography: a synthesis of qualitative research on lay experiences of diabetes and diabetes care. Social Science and Medicine 2003; 56: 671-684.

Cote 2005

Cote L, Turgeon J. Appraising qualitative research articles in medicine and medical education. Medical Teacher 2005; 27: 71-75.

Critical Appraisal Skills Programme 2006

Critical Appraisal Skills Programme. 10 questions to help you make sense of qualitative research [2006]. Available from: http://www.phru.nhs.uk/Pages/PHD/resources.htm (accessed 1 January 2008).

Denzin 1994

Denzin NK, Lincoln YS. Introduction. Entering the field of qualitative research. In: Denzin NK, Lincoln YS (editors). Handbook of Qualitative Research. Thousand Oaks (CA): Sage Publications, 1994.

Dixon-Woods 2005

Dixon-Woods M, Agarwal S, Jones D, Young B, Sutton A. Synthesising qualitative and quantitative evidence: a review of possible methods. Journal of Health Services Research and Policy 2005; 10: 45-53.

Dixon-Woods 2006

Dixon-Woods M, Bonas S, Booth A, Jones DR, Miller T, Sutton AJ, Shaw RL, Smith JA, Young B. How can systematic reviews incorporate qualitative research? A critical perspective. Qualitative Research 2006; 6: 27-44.

Dixon-Woods 2007

Dixon-Woods M, Sutton A, Shaw R, Miller T, Smith J, Young B, Bonas S, Booth A, Jones D. Appraising qualitative research for inclusion in systematic reviews: a quantitative and qualitative comparison of three methods. Journal of Health Services Research and Policy 2007; 12: 42-47.

Evans 2002

Evans D. Database searches for qualitative research. Journal of the Medical Library Association 2002; 90: 290-293.

Grant 2004

Grant MJ. How does your searching grow? A survey of search preferences and the use of optimal search strategies in the identification of qualitative research. Health Information and Libraries Journal 2004; 21: 21-32.

Greenhalgh 2005

Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. BMJ 2005; 331: 1064-1065.

Lucas 2007

Lucas PJ, Baird J, Arai L, Law C, Roberts HM. Worked examples of alternative methods for the synthesis of qualitative and quantitative research in systematic reviews. BMC Medical Research Methodology 2007; 7: 4.

Maher 1999

Maher D, Mikulencak M. What is DOTS? A Guide to Understanding the WHO-recommended TB Control Strategy Known as DOTS. Geneva (Switzerland): World Health Organization, 1999.

Mays 2005

Mays N, Pope C, Popay J. Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. Journal of Health Services Research and Policy 2005; 10 (Suppl 1): 6-20.

McKibbon 2006

McKibbon KA, Wilczynski NL, Haynes RB. Developing optimal search strategies for retrieving qualitative studies in PsycINFO. Evaluation and the Health Professions 2006; 29: 440-454.

Mills 2005a

Mills E, Jadad AR, Ross C, Wilson K. Systematic review of qualitative studies exploring parental beliefs and attitudes toward childhood vaccination identifies common barriers to vaccination. Journal of Clinical Epidemiology 2005; 58: 1081-1088.

Mills 2005b

Mills EJ, Montori VM, Ross CP, Shea B, Wilson K, Guyatt GH. Systematically reviewing qualitative studies complements survey design: an exploratory study of barriers to paediatric immunisations. Journal of Clinical Epidemiology 2005; 58: 1101-1108.

Noblit 1988

Noblit GW, Hare RD. Meta-ethnography: Synthesising Qualitative Studies (Qualitative Research Methods). London: Sage Publications, 1988.

Noyes 2007

Noyes J, Popay J. Directly observed therapy and tuberculosis: how can a systematic review of qualitative research contribute to improving services? A qualitative Meta-synthesis. Journal of Advanced Nursing 2007; 57: 227-243.

Pearson 2005

Pearson A, Wiechula R, Court A, Lockwood C. The JBI model of evidence-based healthcare. JBI Reports 2005; 3: 207-216.

Petticrew 2006

Petticrew M, Roberts H. Systematic Reviews in the Social Sciences: A Practical Guide. Oxford (UK): Blackwell, 2006.

Popay 1998a

Popay J, Rogers A, Williams G. Rationale and standards for the systematic review of qualitative literature in health services research. Qualitative Health Research 1009; 8: 341-351.

Popay 1998b

Popay J, Williams G. Qualitative research and evidence-based healthcare. Journal of the Royal Society of Medicine 1998; 91 (Suppl 35): 32-37.

Popay 2005

Popay J. Moving beyond floccinaucinihilipilification: enhancing the utility of systematic reviews. Journal of Clinical Epidemiology 2005; 58: 1079-1080.

Popay 2006a

Popay J. Incorporating qualitative information in systematic reviews. 14th Cochrane Colloquium, Dublin (Ireland), 2006.

Popay 2006b

Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, Britten N, Roen K, Duffy S. Guidance on the conduct of narrative synthesis in systematic reviews. Results of an ESRC funded research project. (Unpublished report, 2006, University of Lancaster, UK).

Pope 2006

Pope C, Mays N. Qualiative methods in health research. In: Pope C, Mays N (editors). Qualitative Research in Health Care (3rd edition). Malden (MA): Blackwell Publications/BMJ Books, 2006.

Pope 2007

Pope C, Mays N, Popay J. Synthesising Qualitative and Quantitative Health Research: A Guide to Methods. Maidenhead (UK): Open University Press., 2007.

Pound 2005

Pound P, Britten N, Morgan M, Yardley L, Pope C, Daker-White G, Campbell R. Resisting medicines: a synthesis of qualitative studies of medicine taking. Social Science and Medicine 2005; 61: 133-155.

Roen 2006

Roen K, Arai L, Roberts H, Popay J. Extending systematic reviews to include evidence on implementation: methodological work on a review of community-based initiatives to prevent injuries. Social Science and Medicine 2006; 63: 1060-1071.

Spencer 2003

Spencer L. Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence. London (UK): Government Chief Social Researcher's Office, Cabinet Office, 2003. Available from www.gsr.gov.uk/downloads/evaluating_policy/a_quality_framework.pdf.

Thomas 2004

Thomas J, Harden A, Oakley A, Oliver S, Sutcliffe K, Rees R, Brunton G, Kavanagh J. Integrating qualitative research with trials in systematic reviews. BMJ 2004; 328: 1010-1012.

Vermeire 2002

Vermeire E, Van Royen P, Griffiths F, Coenen S, Peremans L, Hendrickx K. The critical appraisal of focus group research articles. European Journal of General Practice 2002; 8: 104-108.

Volmink 2007

Volmink J, Garner P. Directly observed therapy for treating tuberculosis. Cochrane Database of Systematic Reviews 2006, Issue 4. Art No: CD003343.

Walters 2006

Walters LA, Wilczynski NL, Haynes RB. Developing optimal search strategies for retrieving clinically relevant qualitative studies in EMBASE. Qualitative Health Research 2006; 16: 162-168.

Wilczynski 2007

Wilczynski NL, Marks S, Haynes RB. Search strategies for identifying qualitative studies in CINAHL. Qualitative Health Research 2007; 17: 705-710.

Williams 1984

Williams G. The genesis of chronic illness: narrative re-construction. Sociology of Health and Illness 1984; 6: 175-200.

Wong 2004

Wong SS, Wilczynski NL, Haynes RB, Hedges Team. Developing optimal search strategies for detecting clinically relevant qualitative studies in MEDLINE. Medinfo 2004; 11: 311-316.

20.6 定性研究方法

20.6.1 一般的定性研究

Boulton M, Fitzpatrick R. Qualitative methods for assessing health care. Quality in Health Care 1994; 3: 107-113.

Britten N, Jones R, Murphy E, Stacey R. Qualitative research methods in general practice and primary care. Family Practice 1995; 12:104-114

Esterberg KG. Qualitative Methods in Social Research. Boston (US): McGraw-Hill, 2002.

Giacomini MK. The rocky road: qualitative research as evidence. Evidence-Based Medicine 2001; 6: 4-5

Grbich C. Qualitative Research in Health: An Introduction. London (UK): Sage Publications, 1999.

Green J, Britten N. Qualitative research and evidence-based medicine. BMJ 1998; 316:1230-2. Guba RG, Lincoln YS. Competing paradigms in qualitative research. In: Denzin NK, Lincoln YS (Eds) Handbook of Qualitative Research. Thousand Oaks (CA): Sage Publications, 1994.

Miller S, Fredericks M. The nature of "evidence" in qualitative research methods. International Journal of Qualitative Methods 2003; 2: Article 4. Retrieved 1 January 2008 from http://www.ualberta.ca/~ijqm.

Murphy E, Dingwall R, Greatbach D, Parker S, Watson P. Qualitative research methods in health technology assessment: a review of the literature. Health Technology Assessment 1998; 2: 1–274.

Popay J, Williams G. Qualitative research and evidence based healthcare. Journal of the Royal Society of Medicine 1998; 91(Suppl 35):32–37.

Pope C, Mays N. Qualitative research: reaching the parts other methods cannot reach: an introduction to qualitative methods in health and health service research. BMJ 1995; 311: 42-45.

Pope C, Van Royen P, Baker R. Qualitative methods in research on healthcare quality. Quality and Safety in Health Care 2002; 11:148-152.

20.6.2 定性研究方法

Fetterman DM. Ethnography. Step by step. Newbury Park (CA): Sage Publications, 1989.

Glaser BG, Strauss AL. The Discovery of Grounded Theory: Strategies for Qualitative Research. Chicago (IL): Aldine, 1967.

Hammersley M. Reading Ethnographic Research. New York (NY): Langman, 1990.

Hammersley M, Atkinson P. Ethnography: Principles in Practice. London (UK): Routledge, 1995 .

Lambert H, McKevitt C. Anthropology in health research: from qualitative methods to multidisciplinarity. BMJ 2002; 325: 210-213.

Maggs-Rapport F. Combining methodological approaches in research: ethnography and interpretive phenomenology. Journal of Advanced Nursing 2000; 31: 219-225.

Meyer J. Using qualitative methods in health related action research. In: Pope C, Mays N (Eds). Qualitative Research in Health Care. London (UK): BMJ Books, 1999.

Savage J. Ethnography and health care. BMJ 2000; 321:1400-1402.

Strauss A, Corbin J. Grounded Theory in Practice. Thousand Oaks (CA): Sage Publications, 1997.

Strauss A, Corbin J. Basics of Qualitative Research Techniques and Procedures for Developing Grounded Theory. Thousand Oaks (CA): Sage Publications, 1998.

Taylor SJ, Bogdan R. Introduction to Qualitative Research Methods: A Guidebook and Resource. New York (NY), John Wiley & Sons, 1998.

Yin RK. Case Study Research: Designs and Methods. Newbury Park (CA): Sage Publications, 1989.

20.6.3 定性研究文献检索

Flemming K, Briggs M. Electronic searching to locate qualitative research: evaluation of three strategies. Journal of Advanced Nursing 2007; 57: 95-100.

Shaw RL, Booth A, Sutton AJ, Miller T, Smith JA, Young B, Jones DR, Dixon-Woods M. Finding qualitative research: an evaluation of search strategies. BMC Medical Research Methodology 2004; 4: 5

InterTASC Information Subgroup, University of York web site: http://www.york.ac.uk/inst/crd/intertasc/

20.6.4 定性研究证据综合

Jensen LA, Allen MN. Meta-synthesis of qualitative findings. Qualitative Health Research 1996; 6: 553-560.

Noblit GW, Hare RD. Meta-Ethnography: Synthesising Qualitative Studies. Newbury Park (CA): Sage Publications, 1988.

Paterson BL, Thorne SE, Canam C, Jillings C. Meta-Study of Qualitative Health Research. A Practical Guide to Meta-Analysis and Meta-Synthesis. Thousand Oaks (CA): Sage Publications, 2001.

Pearson A. Balancing the evidence: incorporating the synthesis of qualitative data into systematic reviews. JBI Reports 2004; 2 :45-64.

Sandelowski M, Barroso. Creating metasummaries of qualitative findings. Nursing Research 2003; 52: 226-33.

Sandelowski M, Barroso J. Handbook for Synthesising Qualitative Research. New York (NY): Springer, 2007.

Sandelowski M, Docherty S, Emden C. Focus on qualitative methods. Qualitative Meta-synthesis: issues and techniques. Research in Nursing and Health 1997; 20: 365-371.

Thorne S, Jensen L, Kearney MH, Noblit G, Sandelowski M. Qualitative metasynthesis: reflections on methodological orientation and ideological agenda. Qualitative Health Research 2004; 14: 1342-1365.

Zhao S. Metatheory, metamethod, qualitative Meta-analysis: what, why and how? Sociological Perspectives 1991; 34: 377-390.

20.6.5 定性定量证据合并

Dixon-Woods M, Cavers D, Agarwal S, Annandale E, Arthur A, Harvey J, Hsu R, Katbamna S, Olsen R, Smith L, Riley R, Sutton AJ. Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. BMC Medical Research Methodology 2006; 6: 35.

Dixon-Woods M, Fitzpatrick R, Roberts K. Including qualitative research in systematic reviews; opportunities and problems. Journal of Evaluation in Clinical Practice 2001; 7: 125-133.

Dixon-Woods M, Fitzpatrick R. Qualitative research in systematic reviews. BMJ 2001; 323: 765-766

Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O, Peacock R. Storylines of research in diffusion of innovation: a Meta-narrative approach to systematic review. Social Science and Medicine 2005; 61: 417-430.

Harden A, Garcia J, Oliver S, Rees R, Shepherd J, Brunton G, Oakley A. Applying systematic review methods to studies of people's views: an example from public health research. Journal of Epidemiology and Community Health 2004; 58: 794-800.

Pawson, R. Evidence-based policy: the promise of 'realist synthesis'. Evaluation 2002; 8: 340-358.

Pawson R. Evidence Based Policy: A Realist Perspective. London (UK): Sage Publications, 2006.

Pearson, A, Field, J, Jordan, Z. Evidence-based Clinical Practice in Nursing and Healthcare: Assimilating Research, Experience and Expertise. Oxford (UK): Blackwell, 2007.

Petticrew M, Roberts H. Systematic Reviews in the Social Sciences: A Practical Guide. Oxford (UK): Blackwell, 2006.

Pope C, Mays N, Popay J. Synthesising Qualitative and Quantitative Health Research: A Guide to Methods. Maidenhead (UK): Open University Press, 2007.

Popay J (Ed). Moving beyond Effectiveness in Evidence Synthesis: Methodological Issues in the Synthesis of Diverse Sources of Evidence. London (UK): NICE, 2006.

Roberts K, Dixon-Woods M, Fitzpatrick R, Abrams K, Jones D. Factors affecting uptake of childhood immunisation: a Bayesian synthesis of qualitative and quantitative evidence. The Lancet 2002; 360: 1596-1599.

Webb C, Roe B (Eds). Reviewing Research Evidence for Nursing Practice. Oxford (UK): Blackwell, 2007.

20.6.6 定性研究严格评价

Blaxter M. Criteria for evaluation of qualitative research. Medical Sociology News 1996; 22: 68-71.

CASP (Critical Appraisal Skills Programme). 10 Questions to make sense of qualitative research [2006]. Available from: http://www.phru.nhs.uk/pages/phd/resources.htm (accessed 1 January 2008).

Dixon-Woods M, Shaw RL, Agarwal S, Smith JA. The problem of appraising qualitative research. Quality and Safety in Healthcare 2004; 13: 223-225.

Elder NC, Miller WL. Reading and evaluation qualitative research studies. Journal of Family Practice 1995; 41: 279-285

Forchuk C, Roberts J. How to critique qualitative research articles. Canadian Journal of Nursing Research 1993; 25: 47-55.

Horsburgh D. Evaluation of qualitative research. Journal of Clinical Nursing 2003; 12: 307-312.

Malterud K Qualitative research: standards, challenges, and guidelines. The Lancet 2001; 358: 483-488.

Popay J, Rogers A, Williams G. Rationale and standards for the systematic review of qualitative literature in health service research. Qualitative Health Research 1998; 8: 341-351.

Secker J, Wimbush E, Watson J, Milburn K. Qualitative methods in health promotion research: some criteria for quality. Health Education Journal 1995; 54: 74-87.

Spencer L, Ritchie J, Lewis J, Dillon L. Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence. London (UK): Government Chief Social Researcher's Office, 2003.

Vermeire E, Van Royen P, Griffiths F, Coenen S, Peremans L, Hendrickx K. The critical appraisal of focus group research articles. European Journal of General Practice 2002; 8: 104-108.

20.6.7 相关网址（Accessed 1 January 2008）

Campbell Collaboration

A Campbell Review can include evidence from studies of the implementation of an intervention.

www.campbellcollaboration.org

Centre for Reviews and Dissemination (CRD), University of York, UK

In addition to a handbook, CRD has an online resource centre.

www.york.ac.uk/inst/crd

Evidence for Policy and Practice Information and Coordinating (EPPI) Centre

The EPPI Centre provides links to methods, tools and databases.

eppi.ioe.ac.uk/cms

Joanna Briggs Institute (JBI)

JBI offers a variety of evidence-based healthcare resources concerning the synthesis of evidence.

www.joannabriggs.edu.au

National Institute for Health and Clinical Excellence (NICE)

NICE has produced guidance on methods for development of NICE public health guidance which incorporate diverse study designs.

www.nice.org.uk

Social Care Institute for Excellence (SCIE)

SCIE has produced guidance on the conduct of knowledge reviews which incorporate diverse study designs. www.scie.org.uk

（陈燕玲、汪泽皓、肖晓娟译，陈耀龙、申希平、岑啸、秦天强初审）

第二十一章　公共卫生与健康促进系统评价

本文选自工作手册 5.0.1 版本。有关如何引用它的指南，见 21.9 节。这些材料还刊登于 Higgins JPT 和 Green S 编辑的《关于干预措施的 Cochrane 系统评价手册》（书号 978-0470057964）。该手册由 John Wiley＆Sons 出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 公共卫生和健康促进干预是一种定义广泛的行为。这种行为能用多种方法和多种研究类型进行评价，其中包括整群随机试验。但是针对某些问题，可得的最佳证据可能源于非随机试验。
- 检索公共卫生与健康促进文献可能是个非常复杂的工作，要求作者使用除了数据库检索以外的方法进行检索。

- 公共卫生和健康促进干预的系统评价有能力调查具有不同不利因素层次群体间结

果差异。然而，由于收集不同组间信息的局限性及研究中弱势组研究对象的有限性，决定了解决公共卫生与健康促进中不平等问题的复杂性。

- 评价公共卫生和健康促进干预更深层次的问题是如何从研究背景的影响中剥离出干预的实际效果
- 进一步寻求环境因素和干预措施特点的信息可能有助于解释干预措施或干预结局的持续情况

21.1 引言

公共卫生和健康促进干预的系统评价的指南由 Cochrane 卫生促进和公共卫生组（HPPH）（现在转到公共卫生系统评价组）制作并在 2005 年发布，于 2007 年更新。本章仅提供未在手册其他处讨论的关于公共卫生和健康促进与干预问题的概要。在 Cochrane 公共卫生系统评价组网页（www.ph.cochrane.org.）上可获取公共卫生和健康促进干预系统评价指南的完整版本

21.2 纳入的研究类型

公共卫生和健康促进干预是一种定义广泛的行为。这种行为能用多种方法和多种研究设计类型进行评价。尚无哪种单一的方法能回答所有公共卫生和健康促进与干预措施的所有相关问题。如果已经清晰地界定了系统评价研究的问题，其后需要回答的问题便是关于研究设计类型的问题（Petticrew 2003）。一个初步范围的搜索有助于发现可能已经用于研究该干预措施的研究类型。选择研究的标准应该首先能反应特定研究问题或系统评价待回答的问题，而不是事先决定的研究的等级（Glasziou 2004）。决定纳入的研究类型将会影响制作系统评价的后续阶段，尤其是检索、偏倚风险评估和分析（尤其是 Meta-分析）

虽然随机试验结果的推广性存在一定程度的局限性，但它能为效果评价提供有用的证据资源（Black 1996）。由于可行性及伦理问题，公共卫生和健康促进干预的随机试验可能不易获得。在公共卫生领域，整群随机试验逐渐被采纳；一些干预措施要求在整群水平得以应用（Donner 2004）。如果研究单位足够且能够随机分组，确保潜在混杂因素在

组间分布均衡，这类试验就能够提供有价值的证据（见第 16 章 16.3 节）。对于某些问题，非随机试验能代表当前可得最佳研究证据（有效性方面）。非随机试验的系统评价能估计干预措施是否有效及效果的大小。论证从不同研究设计类型所得证据可能导致检验干预措施效果的后续研究设计（包括随机试验）的产生从不同的研究中抽提出的证据模式可能引致后续研究类型（包括随机试验）的形成以检验干预措施。得出定性数据的研究也可能与其他研究效果外的问题相关。比如说，数据的收集可能集中在可能接受特定干预的参与者和限制或促进特定干预的产生预期结果的因素上。有研究者不断对公共卫生和健康促进干预试验的随机与非随机试验间差异进行研究（如英国方法学项目）。第 13 章讨论了在 Cochrane 系统评价纳入非随机试验的一般问题。第 20 章阐述了定性研究问题。

21.3 检索

由于文献分布广泛而分散，检索关于公共卫生和健康促进干预的研究比检索一般医学文献要复杂得多（Peersman 2001）。公共卫生和健康促进的多学科特点意味着能在许多不同领域，通过大量广泛的电子数据库查找到相关研究（Beahler 2000，Grayson 2003）。由于术语不精确及不断变化也会产生许多检索方面的问题（Grayson 2003）。因此，检索关于公共卫生和健康促进干预的研究是个非常复杂的工作，要求作者使用除了数据库检索以外的方法进行检索。为了克服检出定性研究的困难，当前最好的方法要求研究者执行广泛的文献检索（如，针对多种资源的高敏感度检索）。但是，这种试图最大化检索相关条目的方法导致检索出大量的无关记录（Shaw 2004）。由于在书目数据库中定性研究的索引词不充分，我们不推荐应用研究类型过滤器。我们意识到在执行高敏感度检索策略时，常需要平衡所花时间和所需资源与所得研究相关与不相关比率来进行实际的决策。研究者可能决定他们需要应用研究类型进行文献过滤，如果这样，在描述检索策略时他们需要如实报告以便表明研究潜在的局限性。表 21.3 列出了一些公共卫生和健康促进问题相关的电子数据库。

表 21.3.a　公共卫生和健康促进相关电子数据库（所列网站通过网络可免费进入）

领域	资源
心理学	PsycINFO/PscyLIT
生物医学	CINAHL, LILACS (Latin American Caribbean Health Sciences Literature, www.bireme.br/bvs/I/ibd.htm), Web of Science, Medline, EMBASE, CENTRAL, SCOPUS
社会学	Sociofile, Sociological Abstracts, Social Science Citation Index, Social Policy and Practice
教育	ERIC (Educational Resources Information Center), C2-SPECTR (Campbell Collaboration Social, Psychological, Educational and Criminological Trials Register, www.campbellcollaboration.org), REEL (Research Evidence in Education Library, EPPI-Centre, eppi.ioe.ac.uk)
运输	NTIS (National Technical Information Service), TRIS (Transport Research Information Service, ntl.bts.gov/tris), IRRD (International Road Research Documentation), TRANSDOC (from ECMT (European Conference of Ministers of Transport)).
体育活动	SportsDiscus
健康促进	BiblioMap, TRoPHI (Trials Register of Promoting Health Interventions) and DoPHER (Database of Promoting Health Effectiveness Reviews) (EPPI-Centre, eppi.ioe.ac.uk), Public Health Electronic Library (National Institute for Health and Clinical Excellence, www.nice.org.uk/guidance) Database of abstracts of reviews of effectiveness (DARE)
其他	Popline (population health, family planning) db.jhuccp.org/popinform/basic.html, Enviroline (environmental health) – available on Dialog, Toxfile (toxicology) – available on Dialog, Econlit (economics), NGC (National Guideline Clearinghouse, www.guideline.gov)
定性研究	ESRC Qualitative Data Archival Resource Centre (QUALIDATA, www.qualidata.essex.ac.uk), Database of Interviews on Patient Experience (DIPEX, www.dipex.org)

21.4 研究质量和偏倚风险评估

由于相关研究的研究类型广泛，评价公共卫生和健康促进研究的质量及其偏倚风险较为困难。在系统评价的计划阶段，作者需要考虑质量评价的标准。评价标准取决于系统评价纳入研究的类型。作者应该在 Cochrane 系统评价小组（CRG）的指导下制作系统评价和选用评价工具。以下描述的工具可能有助于评价公共卫生和健康促进干预研究

- 随机试验的偏倚风险应该用第 8 章描述的协作网“偏倚风险”工具来评价见第 8 章（8.5 节）。
- 关于整群随机试验的问题在第 16 章讨论（见 16.3.2 章）。
- 关于非随机试验研究的偏倚风险，作者可以参考第 13 章（见 13.5 章）。
- 作者可选用‘定量研究的质量评价工具’（有效的公共卫生实践项目 2007）。由加拿大有效公共卫生实践项目组发布的这个工具可以应用于评价任何定量研究设计。这个工具约花费 10-15 分钟完成。在他们的网页上有评估工具的详细说明（ http://www.myhamilton.ca/myhamilton/CityandGovernment/HealthandSocialServices/Research/EPHPP/）。这个工具包括完整性的干预因素，Deeks 等的系统评价表明它适用于干预性研究效果的系统评价(Deeks 2003）
- 关于断点时间序列和前后对照研究指南在 Cochrane 有效实践和护理机构组（the Cochrane Effective Practice and Organisation of Care Group）可获得（Cochrane EPOC 组 2008）。
- 应该谨慎对待非对照研究（也称为没有对照的前后研究）的结果。对照组的缺乏使我们无从知道没有干预时的状况。解释来自非对照研究数据，所面临的一些特殊问题包括混杂因素（包括季节性）的易感性和向均数回归。

21.5 伦理和不平等问题

公共卫生和健康促进干预有助于促进人群健康。系统评价能确定这些干预措施在达成理想结果方面的效力。在评价公共卫生和健康促进干预效果时有几条特定伦理原则需要考虑。效力是根据从干预措施中受益的人群来测量。这种结果主义方法未考虑受益的分布情况（Hawe 1995)，因此，未涉及健康公平性的问题。健康相关行为或结局的全面

提高可能确实掩盖了组间健康相关结局的差异（Macintyre 2003）。那些在中上社会经济阶层中有效的干预措施可能在弱势人群中并非也有效。即便出于善意的干预可能实际上增加不公平性。组间健康差异可能源于许多和弱势相关的因素间复杂的相互关系（Jackson 2003）。

公共卫生和健康促进干预的系统评价有潜力研究不同弱势水平组间的不同结果。为了有效减少健康的不公平性及不平等性，确认干预措施在弱势群体中的干预效果就变得极为重要。健康不平等性是指“个体或群体获得健康的差异、变化和差距”（Kawachi 2002）。健康公平性是一个伦理概念，指特定健康不平等性的公平性或不公平性问题。国际健康公平协会定义健康公平为：不同社会、经济、人口统计、地域界定的人群或亚人群在健康状况的一方面或多方面不存在潜在的、可补救的系统差异（Macinko 2002）。换言之，健康不公平性是指不公平或不公正的那些健康不平等性，或源于某些不公正（Kawachi 2002）。公共卫生和健康促进干预措施有效性的系统评价能提供干预措施对于健康不平等性的影响的信息。这些信息能用于处理这些卫生不公平现象。

弱势可从居住地（place of residence），种族（race），职业（occupation），性别（gender），宗教（religion），教育（education），社会经济地位（socio-economic position）和社会资本（social capital）方面去考虑，即 PROGRESS（Evans 2003）。作者应该认真考虑这些与所研究人群相关的因素。可从这些因素方面去提取数据。Cochrane 健康公平领域和 Campbell 公平方法学组正在研究与 Cochrane 系统评价相关的公平的定义：www.equity.cochrane.irg.au/en/index.html .

系统评价基于充分详细的研究数据，允许找出相关亚组进行与健康不平等性有关的分析。这要求不仅注意利益和损害的水平，也要注意利益和损害的分布；谁受益，谁受害，谁被排除？

健康不平等性相关的干预措施有效性的系统评价要求三个计算要素：

- 健康状况的有效测量指标（或卫生状况变化）
- 社会经济地位（或弱势）测量指标；
- 用于概括不同组间人群健康状况差异强度的统计学方法

系统评价作者应该决定与系统评价主题相关的不利状况或状态的指标。有许多因素与不利状况有关（即 PROGRESS，residence, race or ethnicity, occupation, gender, religion, education, socio-economic position (SES) and social capital），作者需要收集所研究人群中可能与 PROGESS 因素相关的数据。

进行关于卫生不平等性的系统评价是复杂的。这种复杂性不仅因为收集两组人群差异信息的局限性，而且在研究中有限的弱势群体的参与也是导致该工作复杂的原因。尽管存在这些困难，系统评价在唤起人们对于健康不平等性意识方面仍起着重要作用。Cochrane 健康公平领域和 Campbell 公平方法学组已经确认出许多与卫生公平相关的系统评价，将为作者提供一些额外的指导。

为了查找健康不公平性的研究，系统评价作者需要进行全面的网络检索，对一些社会经济学数据需要联系原始研究作者以获得详细信息。因为原始研究常常未呈现研究对象的社会经济学信息，因此联系作者的工作尤其必要（Oakley 1998, Jackson 2003，Ogilvie 2004）。一旦研究评价和数据提取完成，需要根据是否有效地减少健康不平等性对研究进行分类。一个减少不公平性的有效干预措施通常对弱势群体或个体更有效。一个潜在可有效减少不公平性的干预措施对不同的社会经济群体（由于弱势群体的卫生问题更严重，可能减少卫生不平等性）都同样有效。当干预的目标人群只针对弱势群体或个体时，则判断会变得更困难。一篇有关学校饮食问题的 Cochrane 系统评价，仅针对弱势儿童的有效干预在减少由社会经济状况所致的健康不平等性方面，被归类为‘潜在有效’（Kristjansson 2007）。如果研究包括优势和弱势群体的混合水平，但未包括依据社会经济分组（或类似的分组）的分解的结果，则不可能评判不同的有效性。

21.6 背景

所采用干预措施的类型及其干预结果成功与否主要依赖于社会、经济和政治背景（见图 21.6.a 的例证）。系统评价公共卫生和卫生促进干预的一个问题是如何从称之为“根据环境交互作用的编程”中剥离出“干预”效果(Hawe 2004)。传统上认为，干预措施导致结局的产生。但实际上研究中的结果可能是由于采取干预措施的环境中事先便存在的因素所导致的。因此，在研究中应该把背景作为效应修饰因子进行考虑和测量（Eccles 2003，Hawe 2004）。这些背景因素可能与项目‘主办者’有关。更广泛的背景可能包括主办者能够操作的所有系统。一些调查者认为背景因素也适用于目标调查人群的特点。多年来，上述问题已达到共识（但未明确界定）：决策者认为来源于其他国家的研究结果并不适用于自己国家。

自 Israel 及其同事所做的系统评价（Israel 1995）发表后，在健康促进领域，“背景

评价”这一术语的使用越来越广泛。然而对作为社区或组织水平干预的随机试验设计的一部分的背景交互的系统调查是未知的（Eccles 2003，Hawe 2004）。相反，已存在背景方面的探索研究，即作为持续性研究或项目制度化研究开展更好的领域的一部分，见 21.7 章。一个相关的和不断增长的多学科研究领域是实现和集成科学，这种科学正引导研究者更多地研究干预措施的变化过程的复杂性（Ottoson 1987，Bauman 1991，Scheirer 1994）。目前，定量研究落后于定性的背景分析。

想要从干预措施的效果中分离出背景的作用是极其困难的，除外针对这一目的而设计的研究。有时一些项目从一个背景转到另一个背景，有些能够观察到益处（Resnicow 1993），有些则没有（Lumley 2004）。理论上，在样本含量足够的情况下，整群随机设计有望平衡重要的背景因素。然而，现在很少有调查者测量或报告任何对我们评估而言可能是很重要的的背景因素。我们也注意到最近呼吁更多关注外部真实性（Glasgow 2006，Green 2006）。编辑，研究者共同努力鼓励报告更多和检查更多的关于干预背景方面的信息（Armstrong 2008）。这点应该在未来的 Cochrane 系统评价中有所反映。

图 21.6.a　干预的成功与否取决于背景相关因素的例证（Frommer 2003）

通过媒体干预促进水果和蔬菜的消费	
↓	取决于以下的背景因素：
水果蔬菜的可获得性和相关价格	
↓	取决于以下背景因素：
地理因素，食物的配送系统和零售价	

21.7　可持续性

可持续性指干预或干预效果持续的一般现象（Shediac-Rizkallah 1998，Swerissen 2004）。干预的可持续性在系统评价中应该是个重要的考虑因素。由于决策者、实施者和资助者日益关注对稀缺资源有效且高效地分配，可能使得对健康干预的长期可行性关注增加（Shediac-Rizkallah 1998）。系统评价用户有兴趣了解健康收益，如减少某种特定的疾病，促进健康，是否在干预措施结束后依然持续存在。

不幸的是，常常无法收集干预措施及干预效果持续程度的数据、限制了对干预措施长期影响程度的评估。在 Cochrane 系统评价中，认真考虑以前的研究是如何关注（未关注）可持续性的问题，将增加我们对该领域的理解，并有望促进未来的研究在可持续性评估方面的设计进一步提高。

一个长期的项目并不一定要产生持久的结果，为了产生有用或有效的结果并不是所有的干预措施需要持续(Shediac-Rizkallah 1998)。同样，系统评价作者也应该考虑结果的持续性是否与干预的目的相关。如果相关，作者应该考虑已经测量了（或本应该测量）什么结果，持续多长时期，结果随时间变化模式如何。

应该寻求背景因素的信息和可能有助于解释干预措施或结果持续到什么程度的干预措施特点。凡是未测量的结果的可持续性，作者应该探寻干预结果潜在的持续性。以下四个框架可能有助于帮助判定干预效果的持续性：

1. Bossert 列出下列五个影响持续性的因素（Bossert 1990）：
    - 执行和评价干预措施时经济和政治因素
    - 机构执行干预措施的力度
    - 将行动完全整合至现有的项目、服务或课程中
    - 项目是否包括一个良好的训练（能力构建）内容；
    - 社区参与项目。
2. Swerissen 和 Crisp（Swerissen 2004）制定的框架对确定在不同社会组织水平时干预和效果可能的可持续性进行了指导。该框架概述了干预水平，干预策略与干预及其效果可能具有稳定性的关系。
3. Shediac-Rizkallah 和 Bone 对概念化可持续性提出一个有用的框架（Shediac-Rizkallah 1998）。在该框架中，项目持续性的关键方面定义为以下几方面：1）从项目中获得健康收益；2）在某个组织内项目的制度化；3）在接受的社区进行能力构建。影响持续性的关键因素定义如下：1）广义的环境因素；2）组织环境内的因素；3）项目设计和实施因素
4. 多伦多大学健康促进中心曾在一份文件中概括了关于干预效果持续性的四个整合因素（Health Communication Unit 2001）。

21.8 适用性和可转移性

当决定将一个特定研究或系统评价的结果运用到某一特定人群、干预措施或背景时，应该考虑其适用性（见 12 章，12.3 部分）。可转移性或转移的可能性是相似术语。适用性与完整性、背景及本章前部分讨论到的可持续性密切相关。

公共卫生和健康促进干预系统评价包括几个问题，这几个问题使得适用性的确定远比临床试验文献更复杂。首先，许多公共卫生干预措施不涉及随机化。虽然没有非随机设计的内在特点，这些研究可能有不明确的合格标准、环境和干预措施，使得适用性的确定更加困难。由于干预的代表性或研究参与者无法典型地代表目标人群，因此，源于随机试验的结果可能推广性相对要差些（Black 1996）。其次，公共卫生和健康促进干预往往有多个干预元素。因而使以下诸方面变得困难：1）判定哪个特定的干预元素有显著的效果；2）评估两种元素间的协同作用。再次，在社区干预中，干预实施和依从性可能更难实现和测量。这也使得解释和应用这些结果更困难。最后，在公共卫生和健康促进干预的研究中，社区潜在的社会文化特点是复杂的且难以测量。因此，难以决定干预措施可以在何种程度应用于何种人，难于判断适用性。另一方面，这种异质性可能增加适用性，因为原始研究中的人群，研究背景和干预措施可能十分多样，增加了干预措施广泛应用的可能性。

系统评价的作者被委以总结与所有潜在用户相关的不同方面证据的责任。这样用户便可将系统评价中的环境与自身的具体状况和环境进行比较，以发现相同与差异。这样，用户能够明确证据本身与他们具体情况的差别。

以下几个问题可能有助于作者考虑公共卫生与健康促进相关的适用性及可转移性的问题（Wang 2006）。

适用性

- 当地社会的政治环境允许执行该干预措施吗？
- 存在执行这项干预措施的政治障碍吗？
- 一般的公众和目标人群（亚群）会接受这项干预措施吗？干预措施的每一方面都未违背当地的社会规范吗？伦理上可接受吗？
- 干预措施能调整至适应当地文化吗？
- 在当地环境下存在执行该项干预措施的必要资源吗？（列出必要资源的清单可能有助于回答该问题）

- 当地的目标人群有足够的教育水平以理解干预措施的内容吗？
- 在当地哪个组织将负责提供该干预措施？
- 执行该干预措施存在由于组织结构因素的任何困难吗？
- 当地的干预措施提供者有给予该干预措施的技能吗？如果没有，可得到相应的培训吗？

可转移性

- 在当地感兴趣的健康问题的基线患病率如何？研究背景与当地的患病率差异是什么？
- 研究中的背景与当地背景相比，目标人群的特点是否可比？干预措施将涉及的特定方面，目标人群的特点，如种族，社会经济地位，教育水平等，将会对干预效果产生影响吗？
- 由于政治环境，社会接受度，资源，组织结构及提供者的技能的问题，干预的执行能力可比吗？

21.9 本章信息

编辑：Rebecca Armstrong, Elizabeth Waters 和 Jodie Doyle

本章引用格式：Armstrong R, Waters E, Doyle J (editors). Chapter 21: Reviews in health promotion and public health. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org .

编著者：Rebecca Armstrong, Elizabeth Waters, Nicki Jackson, Sandy Oliver, Jennie Popay, Jonathan Shepherd, Mark Petticrew, Laurie Anderson, Ross Bailie, Ginny Brunton, Penny Hawe, Elizabeth Kristjansson, Lucio Naccarella, Susan Norris, Elizabeth Pienaar, Helen Roberts, Wendy Rogers, Amanda Sowden 和 Helen Thomas.

21.10 参考文献

Armstrong 2008

Armstrong R, Waters E, Moore L, Riggs E, Cuervo LG, Lumbiganon P, Hawe P. Improving the reporting of public health intervention research: advancing TREND and CONSORT. Journal of Public Health (Oxford) (in press, 2008).

Bauman 1991

Bauman LJ, Stein RE, Ireys HT. Reinventing fidelity: the transfer of social technology among settings. American Journal of Community Psychology 1991; 19: 619-639.

Beahler 2000

Beahler CC, Sundheim JJ, Trapp NI. Information retrieval in systematic reviews: challenges in the public health arena. American Journal of Preventive Medicine 2000; 18: 6-10.

Black 1996

Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996; 312: 1215-1218.

Bossert 1990

Bossert TJ. Can they get along without us? Sustainability of donor-supported health projects in Central America and Africa. Social Science and Medicine 1990; 30: 1015-1023.

Cochrane EPOC Group 2008

Cochrane EPOC Group. Cochrane Effective Practice and Organisation of Care Group. Available from: http://www.epoc.cochrane.org (accessed 1 January 2008).

Deeks 2003

Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG. Evaluating non-randomised intervention studies. Health Technology Assessment 2003; 7: 27.

Donner 2004

Donner A, Klar N. Pitfalls of and controversies in cluster randomization trials. American Journal of Public Health 2004; 94: 416-422.

Eccles 2003

Eccles M, Grimshaw J, Campbell M, Ramsay C. Research designs for studies evaluating the effectiveness of change and improvement strategies. Quality and Safety in Health Care 2003; 12: 47-52.

Effective Public Health Practice Project 2007

Effective Public Health Practice Project. Effective Public Health Practice Project [Updated 25 October 2007]. Available from: http://www.city.hamilton.on.ca/PHCS/EPHPP (accessed 1 January 2008).

Evans 2003

Evans T, Brown H. Road traffic crashes: operationalizing equity in the context of health sector reform. Injury Control and Safety Promotion 2003; 10: 11-12.

Frommer 2003

Frommer M, Rychetnik L. From evidence-based medicine to evidence-based public health. In: Lin V, Gibson B (editors). Evidence-based Health Policy: Problems and Possibilities. Melbourne (Australia): Oxford University Press, 2003.

Glasgow 2006

Glasgow RE, Green LW, Klesges LM, Abrams DB, Fisher EB, Goldstein MG, Hayman LL, Ockene JK, Orleans CT. External validity: we need to do more. Annals of Behavioral Medicine 2006; 31: 105-108.

Glasziou 2004

Glasziou P, Vandenbroucke JP, Chalmers I. Assessing the quality of research. BMJ 2004; 328: 39-41.

Grayson 2003

Grayson L, Gomersall A. A Difficult Business: Finding the Evidence for Social Science Reviews. London (UK): ESRC UK Centre for Evidence Based Policy and Practice, 2003.

Green 2006

Green LW, Glasgow RE. Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. Evaluation and the Health Professions 2006; 29: 126-153.

Hawe 1995

Hawe P, Shiell A. Preserving innovation under increasing accountability pressures: the health promotion investment portfolio approach. Health Promotion Journal of Australia 1995; 5: 4-9.

Hawe 2004

Hawe P, Shiell A, Riley T, Gold L. Methods for exploring implementation variation and local context within a cluster randomised community intervention trial. Journal of Epidemiology and Community Health 2004; 58: 788-793.

Health Communication Unit 2001

Health Communication Unit. Overview of Sustainability [Version 8.2, 30 April 2001]. Available from: http://www.thcu.ca/infoandresources/sustainability.htm (accessed 1 January 2008).

Israel 1995

Israel BA, Cummings KM, Dignan MB, Heaney CA, Perales DP, Simons-Morton BG, Zimmerman MA. Evaluation of health education programs: current assessment and future directions. Health Education Quarterly 1995; 22: 364-389.

Jackson 2003

Jackson T, Aldrich R, Dixon J, Furler J, Turrell G, Wilson A, Duell N, Robertson L, Leonard J. Using Socioeconomic Evidence in Clinical Practice Guidelines. Canberra (Australia): National Health and Medical Research Council, 2003.

Kawachi 2002

Kawachi I, Subramanian SV, Almeida-Filho N. A glossary for health inequalities. Journal of Epidemiology and Community Health 2002; 56: 647-652.

Kristjansson 2007

Kristjansson EA, Robinson V, Petticrew M, MacDonald B, Krasevec J, Janzen L, Greenhalgh T, Wells G, MacGowan J, Farmer A, Shea BJ, Mayhew A, Tugwell P. School feeding for improving the physical and psychosocial health of disadvantaged elementary school children. Cochrane Database of Systematic Reviews 2007, Issue 1. Art No: CD004676.

Lumley 2004

Lumley J, Oliver SS, Chamberlain C, Oakley L. Interventions for promoting smoking cessation during pregnancy. Cochrane Database of Systematic Reviews 2004, Issue 4. Art No: CD001055.

Macinko 2002

Macinko JA, Starfield B. Annotated Bibliography on Equity in Health, 1980-2001. International Journal for Equity in Health 2002; 1: 1.

Macintyre 2003

Macintyre S. Evaluating the evidence on measures to reduce inequalities in health. In: Oliver A, Exworthy M (editors). Health Inequalities: Evidence, Policy and Implementation. Proceedings from a meeting of the Health Equity Network. London (UK): The Nuffield Trust, 2003.

Oakley 1998

Oakley A, Peersman G, Oliver S. Social characteristics of participants in health promotion effectiveness research; trial and error? Education for Health 1998; 11: 305-317.

Ogilvie 2004

Ogilvie D, Petticrew M. Reducing social inequalities in smoking: can evidence inform policy? A pilot study. Tobacco Control 2004; 13: 129-131.

Ottoson 1987

Ottoson JM, Green LW. Reconciling concept and context: theory of implementation. In: Ward WB (editors). Advances in Health Education and Promotion Volume 2. Greenwich (CT): JAI Press, 1987.

Peersman 2001

Peersman G, Oakley A. Learning from research. In: Oliver S, Peersman G (editors). Using Research for Effective Health Promotion. Buckingham (UK): Open University Press, 2001.

Petticrew 2003

Petticrew M, Roberts H. Evidence, hierarchies, and typologies: horses for courses. Journal of Epidemiology and Community Health 2003; 57: 527-529.

Resnicow 1993

Resnicow K, Cross D, Wynder E. The Know Your Body program: a review of evaluation studies. Bulletin of the New York Academy of Medicine 1993; 70: 188-207.

Scheirer 1994

Scheirer MA. Designing and using process evaluations. In: Wholey JS, Hatry HP, Newcomer KE (editors). Handbook of Practical Program Evaluation. San Francisco: Jossey Bass, 1994.

Shaw 2004

Shaw RL, Booth A, Sutton AJ, Miller T, Smith JA, Young B, Jones DR, xon-Woods M. Finding qualitative research: an evaluation of search strategies. BMC Medical Research Methodology 2004; 4: 5.

Shediac-Rizkallah 1998

Shediac-Rizkallah MC, Bone LR. Planning for the sustainability of community-based health programs: conceptual frameworks and future directions for research, practice and policy. Health Education Research 1998; 13: 87-108.

Swerissen 2004

Swerissen H, Crisp BR. The sustainability of health promotion interventions for different levels of social organization. Health Promotion International 2004; 19: 123-130.

Wang 2006

Wang S, Moss JR, Hiller JE. Applicability and transferability of interventions in evidence-based public health. Health Promotion International 2006; 21: 76-83.

（郭琴译，文进、岑啸、秦天强初审）

第二十二章　系统评价再评价

作者：Lorne A Becker 和 Andrew D Oxman

本文选自工作手册 5.0.1版本。有关如何引用它的指南，见22.4节。这些材料还刊登于Higgins JPT和Green S 编辑的《关于干预措施的Cochrane系统评价手册》（书号978-0470057964）。该手册由John Wiley＆Sons出版有限公司发行。公司地址：The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England。公司电话：（+44）1243 779777。订购及客户服务查询电子邮件地址：cs-books@wiley.co.uk。公司主页：www.wiley.com。

内容提要

- 系统评价再评价（即再评价）旨在针对某一健康问题的两个或多个潜在干预措施效果的多个 Cochrane 系统评价进行总结。
- 缺乏相关的干预性 Cochrane 系统评价时，Cochrane 再评价可纳入其他已发表的系统评价。
- 再评价应首选已有大量 Cochrane 系统评价存在的研究领域。
- 再评价结构上与系统评价相似，但纳入的是系统评价而非原始研究。
- 再评价包括的“纳入系统评价一览表”相当于 Cochrane 系统评价的“结果总结表”

- 再评价应随纳入系统评价更新而更新。

22.1 引言

22.1.1 系统评价再评价的定义

Cochrane 系统评价再评价（Cochrane 再评价）是基于多个干预性系统评价整合证据，使其更易获取和利用。本章概括了Cochrane再评价的原理以及协作网系统评价小组（CRGs）、评价员应遵循的具体方法。

22.1.2 Cochrane再评价的原理

首先，Cochrane再评价旨在对同一健康问题的两个或多个潜在干预措施的相关系统评价进行再次评价。Cochrane再评价着重关注某些潜在的干预措施的Cochrane系统评价并对其重要测量指标结果进行汇总。

需重点指出的是，有多种原因需开展系统评价再评价，Cochrane 再评价仅涵盖部分而非全部目的。表22.1.a概括了进行再评价的不同原因，并指出了哪些情况适合发表Cochrane 再评价。在注册和发表Cochrane再评价之前，CRGs会确认拟进行的再评价是否适合发表。

正如表 22.1 a所总结，Cochrane再评价核心目标是作为Cochrane 图书馆的“用户友好接口”（user-friendly front end），使读者可快速浏览有关特定决策的现有干预措施的Cochrane系统评价（及全面的清单）。首要的预期读者是决策者（如，临床医生、政策制定者、知证用户），他们可针对某一特殊问题进入Cochrane图书馆查询证据。再评价一旦完成，将发表于Cochrane 系统评价数据库，读者可依据其格式快速地将其与Cochrane干预性、诊断性或方法学系统评价区分开来。

22.2 制作Cochrane 再评价

22.2.1 组织事宜

进行Cochrane再评价应优先考虑已有诸多Cochrane干预性系统评价存在的重要领域。

是否需要再评价取决于评价员、CRG成员的兴趣所在，Cochrane中心和相关领域也会设定Cochrane再评价的优先领域，并试图寻求评价员开展再评价。Cochrane系统评价评价员如有意愿，也可担当再评价作者，但并非顺理成章。再评价作者应熟悉Cochrane系统评价的方法学，最好曾经参与过Cochrane系统评价的制作。

每个CRG将对每一个再评价进行编辑控制；与干预性系统评价流程一样需提交题目和计划书。大多数情况下，再评价纳入的Cochrane系统评价应来源于同一个CRG组，并由该组负责编辑。如果再评价纳入的系统评价涉及多个CRG组，如，再评价涉及的干预措施应用于多种情况，其编辑程序将由多个相关CRG组协商解决，并确定哪个CRG组担任责任编辑，当前类似问题就如此解决。

再评价作者所关注的研究尚缺乏Cochrane系统评价时，则可考虑联系相关CRG组，在更广的范围内开展新的系统评价、更新已存在系统评价、或针对Cochrane系统评价尚未纳入的干预措施展开新的系统评价。

表22.1.a　进行再评价的理由和Cochrane 发表的条件

目的	纳入标准	文章示例	是否适合纳入 Cochrane 再评价	点评
对针对同一条件下或同一健康问题的不同干预措施的多个系统评价进行再评价，以总结证据	Cochrane 系统评价	A Cochrane Overview of interventions for nocturnal enuresis (Russell 2006)	是	这是 Cochrane 再评价的首要目的（在摘要和正文的目的部分应写明 Cochrane 系统评价再评价）
	Cochrane 和非 Cochrane 系统评价	某些 BMJ 临床证据章节和卫生技术评估（HTA）报告.	也许是	#有时同时纳入 Cochrane 和非 Cochrane 系统评价是合理的。如，某重要干预措施存在高质量的系统评价，但无 Cochrane 系统评价。但鼓励 CRGs 首先关注针对 Cochrane 系统评价的再评价： • 检索和纳入非 Cochrane 系统评价需承担额外的工作量和挑战 • Cochrane 图书馆的用户可能无法获得非 Cochrane 系统评价 • Cochrane 再评价的首要目的是总结 Cochrane 系统评价并提供用户友好接口。
针对同一健康问题的同一干预措施，而不同系统评价关注了不同结局的多个系统评价进行再评价	Cochrane 干预性系统评价	更年期激素替代疗法（HRT）的 Cochrane 系统评价再评价结局可能包括骨密度、绝经期症状、心血管风险/事件、认知功能等。	偶尔是	理论上，一篇系统评价应包括一项干预措施能对决策产生影响的所有结局。但有时，对于 HRT，不同系统评价对研究结局的考虑程度有所不同。

	Cochrane 和非 Cochrane 干预性系统评价	某些 BMJ 临床证据章节和卫生技术评估（HTA）报告.	很少是	理由同上
对某一干预措施应用于不同疾病或健康问题，不同人群多个系统评价进行再评价，以总结证据	Cochrane 系统评价	不同人群及背景的维生素A的Cochrane系统评价再评价	偶尔是	相同或相似的干预措施有时针对于不同的健康问题，或关注于不同人群的不同研究和系统评价。在这种情况下，系统评价再评价不太可能被关注具体临床问题的医生和患者所感兴趣，而更多被政策决策者或关注跨越不同的系统评价问题的人员所关注。
	Cochrane 和非 Cochrane 系统评价		很少是	理由同上
从针对多个健康问题的某一干预措施的多个系统评价中总结不良反应相关证据	仅纳入 Cochrane 系统评价，或同时纳入 Cochrane 和非 Cochrane 系统评价	NSAIDs 用于骨关节炎或风湿性关节炎的不良反应的再评价	很少是	许多 Cochrane 系统评价报告不良反应，但少有系统评价专门针对不良反应而作。许多重要的不良反应发生很罕见，所以对照试验无法精确地估计其真正发生率。基于这些原因，基于 Cochrane 或其他系统评价的再评价也很难准确概括某干预措施的不良反应情况，除非纳入的系统评价专门讨论不良反应的发生率（详见第十四章）
针对某领域所有问题提供一个全面的再评价，同时也包括未纳入系统评价的研究。	系统评价和一些没有纳入系统评价的研究	某些 BMJ 临床证据章节、卫生技术评估（HTA）报告或某杂志的大综述（a synoptic review article for a journal）	不是	包括未纳入系统评价的研究可能适用于许多情况，如，实施 HTA 报告、制定临床实践指南、BMJ 临床证据资源。但都超越了 Cochrane 再评价的范围。Cochrane 再评价的作者应在纳入的系统评价已经过时，尤其是发表了新的相关研究，及有相关干预措施的系统评价尚未发表时高度注意。然而，在再评价中不必进行新的系统评价或更新已有系统评价。

22.2.2 方法学

Cochrane再评价不同于系统评价的方法学，它是基于系统评价而非原始研究进行归纳和总结。系统评价和再评价方法学的主要不同点的总结见表22.2.a。

再评价目的并非基于纳入的系统评价重新检索文献、制定纳入标准、评价文献质量、进行Meta分析。此外，再评价无需系统查找其他研究或从纳入文献中额外提取结局指标。再评价更强调的是评价纳入系统评价的局限性，针对某结局进行Meta分析以提供不同干预措施效果的间接比较。这并不表示再评价实行更详细地分析，如严格的评价、新的检索和分析是不恰当的，这些并非Cochrane再评价的初衷。

表22.2a 干预措施的Cochrane系统评价和再评价的方法学比较

	Cochrane 系统评价	Cochrane 再评价	关于 Cochrane 再评价的备注
目的	基于干预措施效果研究中总结证据	基于干预措施疗效的系统评价中总结证据	适用于同一疾病或健康问题的多种干预措施效果研究分散在不同的Cochrane系统评价
纳入标准	描述原始研究的纳入和排除标准	描述系统评价的纳入和排除标准	• 主要纳入 Cochrane 系统评价。 • 有时选择纳入 Cochrane 系统评价和在 Cochrane 图书馆（疗效评价摘要数据库和卫生技术评估(HTA)数据库）检索到的其他系统评价。 • 偶尔纳入其他出版来源的系统评价。
检索	全面检索相关原始研究	重点检索相关的 Cochrane 系统评价	偶尔可检索非 Cochrane 系统评价
数据收集	从纳入原始研究中收集	从纳入系统评价中收集	如有必要，再评价的作者可从已纳入的系统评价作者方面获取其他信息，或者偶尔也可从纳入系统评价纳入的原始研究中自行提取相关数据。
评价局限性	针对纳入原始研究，即，偏倚风险	针对纳入的系统评价	Cochrane 再评价的作者应应用明确标准严格评价系统评价。一般局限性（如，该系统评价是否更新）和特殊局限性（即，相对于再评价的具体目标，纳入系统评价是否存在局限性）都应考虑。

证据质量评价	不同研究间的每个重要结局指标	目前质量评价仍依赖于纳入系统评价的报告	• 推荐每篇再评价应包括每个重要结局指标的证据质量评价。 • 如纳入系统评价未对其纳入原始研究进行质量评价，再评价作者应补充实行。 • 如果纳入系统评价已有质量评价，再评价作者应严格评价纳入系统评价的判断，并确保这些判断在纳入的系统评价中保持一致。
分析	对纳入研究中每个重要结局的结果予以合并	总结纳入系统评价的结果。当对照分散在不同系统评价中，尤其是多种干预措施的间接比较时，需进行额外分析。	目前再评价作者尽可能应依据纳入系统评价报告的分析。数据偶尔需要重新分析，如不同系统评价分析了不同的人群或亚组。如有必要应进行各系统评价间的比较分析。

22.2.3 更新Cochrane 再评价

定期更新再评价是至关重要的，且应遵循Cochrane系统评价更新的常规流程（详见第三章）。只要纳入的系统评价有更新，再评价就应随之更新。通常，Cochrane再评价只需进行微小的变动。如，如果Cochrane系统评价未检索到新的原始研究，再评价只需将此信息更新至最近的日期。若纳入系统评价更新的结果和结论有所变化时，再评价需随之大范围修改。

22.3 Cochrane再评价的格式

22.3.1 标题和再评价的内容（或计划书内容）

再评价的题目形式应为：[干预措施或对照] 对 [某卫生问题] 在或针对 [某类人群、疾病或问题、特定场所]。

“干预措施或对照”有多种表达格式，取决于系统评价的范围。如果系统评价考虑到了所有潜在的干预措施，那么此部分应为“对于…的干预措施”。如果再评价仅局限于部分潜在的干预措施，那么题目就应提及“部分”之意，如，“对于…的手术措施”。

如两种干预措施比较，题目应提及“比较”之意，如，“对于…的手术与药物比较”

再评价其他内容与系统评价相同。（详见第四章4.2）

22.3.2 摘要

摘要所包含以下每一条目内容应如下所述：

背景：应简明扼要的说明研究内容或详述再评价的原理和目的。

目的：最好用一句话精确的阐明再评价的主要目的。可能的形式应为“总结Cochrane系统评价以评价［干预措施或对照］对于［某卫生问题］对于或在［某类人群、疾病和卫生问题、特定场所］的效果”

方法：此部分应简明提及用于识别符合在评价纳入标准的检索策略以及数据收集和分析的方法。后者应该限制说明提取数据、评价数据质量和真实性的指导原则，而无需罗列提取数据的细节。应阐明指导原则应用的方法（如，由多位评价员独立提取数据）。

主要结果：此部分应首先报告纳入系统评价的数量，而后简明扼要地陈述结果解释的相关细节（如，纳入系统评价的质量，可比性说明，如果恰当的话）。结果应以着重描述主要定性和定量结果（通常主要结果不超过七个）为主要目的。结局指标的选择应基于辅助决策者判断是否采用某种特殊干预措施的期望值。如果相关，应提及针对每一结局指标，纳入的研究数量、受试者总数及相应证据的质量。结果应以叙述的方式表述，如计量结果不清楚或不直观时也可采用定量方式表述（如，标准化均值差分析的结果）。摘要中提及的统计方法应与正文强调的一致，均应以标准方法表述，如“RR= 2.31（95%CI［1.13，3.45］）”。若可能，应同时报告结果的绝对值和相对值。当对照组某些结局的风险随不同的原始研究或系统评价而不同时，作者应谨慎报告结果的绝对效应量（详见第11章，11.5.5）。如果纳入的系统评价没有计算合并结果，再评价应给出定量评价或描述结果的类型和范围。但阳性和阴性结果的研究或系统评价的数量应避免报告。

作者的结论：再评价的首要目的是陈述信息并非提供建议。作者应简明、直接地从再评价的结果中得出结论，使其能直接反应主要结果。作者应注意不要混淆“缺乏证据”和“缺乏疗效”。作者不应就实践的应用环境、价值观、喜好、权衡条件作出假设；避免给出建议或推荐意见。应指出数据和分析的重要局限性。如相关，应包括关于研究具体意义的重要结论（包括系统评价）。作者不应作出诸如“有待更多的研究”的俗套声明。

22.3.3 通俗语言摘要（plain language summary）

原称“概要”，旨在以简洁的形式总结再评价，以便于卫生保健用户能够理解。（见第4章，4.4部分）。

22.3.4 Cochrane再评价正文

Cochrane再评价的目标读者是卫生保健决策者（如，临床医生、知证者、政策决策者），他们已经对潜在的疾病和卫生问题有基本的理解，并在某种程度上期望发现在Cochrane图书馆上所界定的关于某卫生问题的潜在干预措施重要信息。再评价应提供相关问题的Cochrane系统评价结果的概述，指引读者从具体单个系统评价中获得更多的细节信息。

Cochrane再评价的正文包括很多固定的标题。作者根据需要也可自行添加子标题。某些特定的标题被设计为“推荐”。推荐部分的内容应被包含在再评价文中，但子标题并非必须使用，如子标题下内容过于简短时可不使用。应列出与特殊系统评价有关或无关的附加子标题。此剩余部分将分别描述标题的相关种类（固定、推荐、自选）。

背景 ［固定，一级标题］

此部分应陈述再评价中的Cochrane系统评价关于背景的内容。背景有助于确定再评价的原理，并应明确说明再评价关注的研究问题，包括清楚地描述所关注的疾病状况、干预措施、对照措施和结局指标。还应阐明所关注问题的重要性。背景应简明扼要（长度通常为一页）并以实施卫生保健调查研究人员能理解的方式呈现。背景部分推荐包含以下内容，但并非强制。

状况描述 ［推荐，2级标题］

背景开头应简单描述关注的疾病，并说明其重要性。可包含生物学、诊断、预后、公共卫生指标（包括患病率或者发病率）等方面的重要信息。

干预措施描述 ［推荐，2级标题］

此部分应提及当前应用于此疾病的所有干预措施，不管其是否已在Cochrane系统评价中进行过讨论。合理的干预措施分组将简化正文部分（如，列出非甾体抗炎药而非提供此类药物名称的详尽清单）。如可能，应对应用当前不同干预措施（如放疗加化疗）的可能性加以讨论。应提及当前临床实践中各种潜在的干预措施相对地位（如果可行的话）。

干预措施的作用原理 ［推荐，2级标题］

系统评价收集证据以评价干预措施是否会真的产生预期效果。此部分应阐明所评价的干预措施为何会对卫生保健的潜在使用者产生影响的理论性推理。如，将药物干预与生物学状况联系起来。作者可提及一组经验证据，如相似或相同的干预措施对其他人群有影响。作者也可提及证实干预措施可能有效的文献。对现有文献的参考不应包括再评价纳入的系统评价任何对于结果的讨论，也不应包括系统评价纳入的原始研究对结果的讨论；这些均属于结果部分的内容。

再评价的意义 ［自选，2级标题］

背景有助于确立再评价进行的原因，并可解释所提研究问题的重要性。此部分应阐明为何要进行此篇再评价，目标读者是谁，支持何种决策。

目的 ［固定，1级标题］

开门见山地明确阐明再评价的主要目的，包括关注的干预措施和目标问题。也可进一步阐明针对于不同受试者、不同干预对照、不同结局测量的研究目的。

方法 ［固定，1级标题］

计划书中方法部分应用将来时态书写。方法学部分应描述获得当前再评价的结果和结论所采用的方法，而不是去讨论所要总结分析的系统评价用过的方法。关于纳的入系统评价采用的方法应放在“纳入系统评价描述”部分。方法学部分有许多的子部分。

纳入标准 ［固定，2级标题］

应依据再评价所研究的问题纳入系统评价，包括受试者（身体条件或健康问题），干预措施、对照组和结局指标的详细描述。再评价一般应包括与其研究目的相关的、针对某疾病或卫生问题的、涉及多种干预措施的所有的Cochrane系统评价。但某些情况下，再评价作者希望以某种方式限制研究范围，如，关注某几种特定的干预措施（如，所有的药物治疗措施，排除非药物治疗措施）。适当的限制尤其适合于当现有的Cochrane系统评价讨论了不同的受试者（如，不同年龄、种族、性别、疾病分期、并存症的群体）的情况。在考虑干预措施是否汇总分析或分开分析时，把阅读该再评价的决策者的观点和独立进行决策所需的信息加以考虑将会很有帮助。如，由于预防和治疗决策针对的受试者不同，因此再评价针对同一疾病的干预性系统评价和预防性系统评价时应区别对待，并在相应的纳入标准部分详细阐明。

如果纳入了非Cochrane系统评价，此部分应先界定判断非Cochrane系统评价就是系统评价的标准，尤其是当针对某一研究问题同时存在两篇以上的系统评价时，哪些系统评价将被纳入的标准。

检索策略 ［固定，2级标题］

此部分应确定检索Cochrane系统评价或其他系统评价的方法。检索策略应比Cochrane系统评价中的策略简单，因为对相关文章的基本检索策略在系统评价时早已实行。如果仅纳入了Cochrane系统评价，那么只需检索Cochrane系统评价数据库（Cochrane Database of Systematic Reviews），无需再检索其他数据库。如果纳入了其他系统评价，此部分应清楚列出检索的数据库资源（如，疗效评价摘要数据 （Database of Abstracts of Reviews of Effects）（Petticrew 1999））、检索策略和检索方法。

数据收集和分析 ［固定，2级标题］

此部分应简单的描述再评价所用到的方法，应提及以下问题：

筛选系统评价 ［推荐，3级标题］

此部分应说明筛选系统评价的方法、是否由一名以上的评价员独立实施筛选、遇到分歧时如何解决。

数据提取与管理 ［推荐，3级标题］

此部分应描述从纳入的系统评价提取和获得数据的方法（如，利用数据提取表）、是否由多名评价员独立提取数据、遇到分歧如何解决。准备分析时如何处理数据的相关方法也应明确描述。对于纳入评价的缺失数据的处理方法也应加以描述。

纳入系统评价的方法学质量评价 ［推荐，3级标题］

再评价作者必须同时进行两种不同的质量评价：方法学质量评价和证据质量评价，具体如下：

此部分应描述两种不同类型评价的方法。每种评价均建议由一名以上的评价员应用评价标准、独立进行评价，并声明遇到分歧意见如何处理。应描述或引用所采用的评估工具（如，GRADE），并指明评估结果如何在再评价结果中进行解释。

纳入系统评价的质量 ［推荐，4级标题］

需描述纳入系统评价采用的方法学质量的评价方法。因对系统评价的质量评价或偏倚风险的相关研究有限，我们尚不能推荐具体的评价工具。但可参考一些可用的问卷和清单(Oxman 1994, Shea 2006)。

纳入系统评价的证据质量 ［推荐，4级标题］

即便采用了所谓的最好的方法总结证据， Cochrane干预性系统评价依然可能存在很大的局限性，因为纳入研究内部和纳入研究之间存在的潜在偏倚，单个研究结果之间存在冲突，所关注的问题缺乏证据支持或不是直接证据（见Chapter 12，Section 12.2）。

此部分应总结再评价为支持结论而确定的证据质量评价方法。理想情况下，基于此评价的信息应列在“纳入研究特征一览表”、“偏倚风险”、“研究总结”中。现在推荐偏倚风险的评价报告应按照Cochrane系统评价标准方法进行（详见第8章），在Cochrane干预性系统评价和再评价中针对每一重要结局的证据质量评价应使用GRADE评价体系。（详见第11章，11.5和第12章，12.2）

数据合成 ［推荐，3级标题］

许多再评价仅从纳入的系统评价中提取数据然后重新整合成图表。但再评价也会基于正规（formal）统计分析进行间接比较，尤其无直接比较证据时（Glenny 2005）。实施间接比较和同时进行多种干预措施的Meta分析的统计方法常应用于再评价数据合成中（详见第16章，16.6）。间接比较的证据级别较直接（头对头）比较级别低。如果纳入的系统评价未实施直接比较，但直接比较的研究已被确认实施，再评价作者应不再进行间接比较。欲进行间接比较或者多种干预措施的meta分析的作者应寻求合适的统计和方法学的支持。

若使用了许多定性或描述性分析方法，再评价作者应指出纳入系统评价的结果标准化报告的方法，包括转化汇总统计量及对不同对照组风险进行标准化。若比较不同系统评价的绝对效应时对照组风险存在差异，作者应谨慎对待（详见第11章，11.5.5）。

结果 ［固定，1级标题］

纳入系统评价的描述 ［固定，2级标题］

应简洁描述纳入的系统评价，但应为读者充分提供关于纳入的系统评价的受试者的特征信息：用药剂量、疗程或者干预措施的其他特征。如果纳入的系统评价间存在明显差异（如，系统评价纳入或排除标准不一、对照组不同、结局指标的测量方法不同）应明确说明。此外，纳入的系统评价的研究目的和纳入标准与再评价目的间的任何不同也应说明。如，系统评价可能忽视了再评价作者出于某特殊的利益而考虑的重要结局或亚组分析。如果有些系统评价较之其他系统评价最近有所更新，也应说明。此部分许多问题可在“纳入系统评价一览表”中加以总结归纳。（详见22.3.6）

纳入系统评价的方法学质量 ［推荐，2级标题］

纳入系统评价的质量 ［推荐，3级标题］

应描述再评价纳入的系统评价的总体质量，包括系统评价间的质量差异和个别系统评价重大的质量缺陷。评价纳入系统评价的质量的标准应在“方法”部分提及（此部分不作提及）。如果评价员认为有必要具体提及如何依据每条标准对每一纳入系统评价进

行了评价，应以另外表格的形式列出而不应在正文中详述。

纳入系统评价的证据质量 ［推荐，3级标题］

应总结纳入系统评价证据的总体质量，如，对于最重要结局利用GRADE标准进行评价（见第13章，13.2）

干预措施效果 ［固定，2级标题］

应总结纳入的系统评价关于干预措施效果的主要研究结果。此部分的呈现应从临床的角度对结果进行分类而非单纯的依次罗列每个系统评价的研究结果。这些分类应包括干预措施类型（药物治疗、手术治疗、行为干预等）；疾病分期（症状前期、疾病早期、疾病晚期）；受试者特征（年龄、性别、种族）；或结局类型（存活、功能状况、不良反应）。为了便于阅读，建议使用亚标题。此处单个系统评价的研究结果及其汇总统计量都应列入总结图或总结表中。

应注意说明再评价作者认为重要但系统评价作者尚未找出证据的结局指标（因没有检索到相关研究或纳入研究未报告此项重要结果）。此外，对于不易用数值数据轻易总结的重要结果，应在此部分总结描述。

此部分作者应避免推断。在描述结果和作出结论时，应避免混淆“无证据有效”和“无效的证据”的常识性错误。当存在非确定性证据时，再评价不应武断地认为某干预措施“无效”或与对照措施的效果“一样”。这种情况下，与结果增加或减少相一致，报告具体数据及其置信区间较为合适。

讨论 ［固定，1级标题］

总结主要结果 ［推荐，2级标题］

简要总结主要研究结果、权衡重要利弊，强调重要的不确定问题。

证据的全面性和适用性 ［推荐，2级标题］

纳入的系统评价是否充分地达到了再评价的所有目标？如果没有，目前存在的差距是什么？是否研究了所有相关类型的受试者、干预措施和结局指标？应描述再评价问题的证据相关性。这是对再评价外部真实性的整体判断。尽管作者都知道现行的临床实践在不同国家和不同受试者间存在差异，但此处仍需指出再评价结果多大程度上适用于现行临床实践。

证据质量 ［推荐，2级标题］

根据纳入的系统评价是否可以就再评价的研究目的下强有力的结论？讨论应包括：纳入的系统评价是否纳入了所有相关的原始研究？是否提取了所有相关数据？使用的

方法（如，检索、纳入、数据提取和分析）是否导致偏倚？这些问题随着干预措施、结局指标、和临床亚组的不同而可能不同。如果是这样，讨论部分应清楚地阐明每一关键关注领域的证据质量情况。

再评价过程中的潜在偏倚 ［推荐，2级标题］

声明再评价关于防止偏倚的优势和局限性。这些可能在再评价作者的控制之内或之外的因素。讨论部分应包括：是否检索和纳入了所有相关的系统评价？是否获取了所有相关数据？使用的方法（如，检索、纳入、数据提取和分析）是否导致偏倚？

与其他研究或系统评价的一致性或不一致性 ［推荐，2级标题］

应评论纳入的系统评价与其它可证据内容的相符程度，应明确说明其他证据是否被系统性的进行了评价。

作者结论 ［固定，1级标题］

此部分应呈现再评价作者的结论，而非单纯地重述纳入的系统评价作者的结论。此部分的初衷是陈述信息而非提供建议。应分以下两部分进行：

临床实践意义 ［固定，2级标题］

再评价的临床实践意义的表述应基于系统评价所述的数据，尽可能地清楚化和实用化，不可任意发挥。也不要将“无证据有效”和“无效的证据”二者混淆。

研究的含义 ［推荐，2级标题］

此部分应讨论对系统评价进行再评价后仍未解决的关键临床问题。如存在针对研究问题潜在的重要的干预措施而尚未在Cochrane系统评价发布，此部分应清楚提及。此外，此部分可通过指明尚无定论的研究领域为临床决策者的未来研究提供方向和建议。

致谢 ［固定，1级标题］

此部分用于作者对没有列在作者名单中的任何组织和个人表达谢意。详见第4章,4.5

作者贡献 ［固定，1级标题］

描述所有作者的贡献 详见第4章,4.5

利益声明 ［固定，1级标题］

由于可能会导致真实的或可感知到的利益冲突，作者应该报告所涉及的利益相关的任何现在或过去的从属关系及任何组织或团体,（详见第4章,4.5）。作者如果与再评价所纳入的系统评价有关联，此处必须声明。

计划书和系统评价的区别 ［固定，1级标题］

有时必需使用最初计划书所描述的不同的方法 详见第4章，4.5

发表事项 ［固定，1级标题］

详见第4章，4.5

22.3.5 系统评价和参考文献

作者必须核查所有参考文献的准确性。

22.3.5.1 系统评价的参考文献

每个纳入的系统评价都应创建一个“参考文献ID号”，并使用于正文当中。一般是由“第一作者的姓+最近版本的发表年代”表示（如，Eforn 2006）。如果两篇或多篇系统评价的第一作者和发表年代相同，应标注字母加以区别（如，Eforn 2007a，Eforn 2007b）。参考文献部分包括以下两大固定标题：

纳入的系统评价

符合纳入标准并被纳入再评价的系统评价

排除的系统评价

不符合纳入标准而未被纳入再评价的系统评价

22.3.5.2 其他参考文献

文章引用的其他参考文献，包括背景、方法等部分引用的参考文献均应列出。

22.3.6 表格

再评价可考虑应用许多不同形式的表格，都可由RevMan软件实现。

22.3.6.1 纳入系统评价的特征一览表

每个再评价应包括一或多张“纳入系统评价特征表”（表22.3.a所示），以方便读者快速的了解纳入的系统评价的基本特征。

条目注释

系统评价（Review）

填写相应的“参考文献ID号”（见22.3.5.1）

最后更新时间（Date assessed as up to date）

此栏应列出纳入系统评价的最近更新时间（详见第3章，3.3.2）。系统评价更新发表时间与更新检索时间不应超过6个月，更新的系统评价应包括更新的检索结果。

受试者（Population）

此栏应说明Cochrane系统评价纳入的受试者特征，即对于年龄、性别、种族、疾病分期、并存症等的一些限制。

干预措施（Intervention）

列出Cochrane系统评价包含的所有干预措施，不管其是否检出并纳入了相应的原始研究。

对照措施（Comparison intervention）

此栏应列出对照措施的类型（如，安慰剂、空白治疗、选择性干预等）

报告的结局指标（Outcomes for which data were reported）

此栏包括系统评价所示数据的重要结局指标，不管再评价的数据总结是否涉及。

系统评价局限性（Review limitations）

此栏简要描述Cochrane（或其他）系统评价方法学上重要的局限性。此处不要总结系统评价纳入原始研究的质量（这应在系统评价再评价一表中提及，见22.3.6.2）

表22.3.a “纳入系统评价特征一览表”样板

系统评价 ID	最后更新时间	受试者	干预措施	对照措施	报告的结局	系统评价局限性

22.3.6.2 再评价表格（“Overview of reviews” Table）

每篇再评价应包含一张或多张如表22.3b格式的表格来总结结果。此格式设计上（尽可能）根据“结果总结”表（详见第11章，11.5）。如果再评价关注的不止一类临床对象（如，疾病的分期或严重程度、并发症、其它可能影响结局的因素不同），应以不同表格分别表述不同受试者人群的信息。具体形式可随系统评价主题不同而不同，但每张表格都应同时包括有利和不利结局，以及对照组出现此结局的频率或严重程度，干预措施相对和绝对效应估计值，偏倚风险的指标（随不同结局指标和对照而变化），以及其他备注。

“系统评价再评价”表格（Overview of reviews Table）样板

表22.3b 提供了“系统评价再评价”表格样板。设计该表的本意要与“结果总结表”

的用途尽可能相似，如果“结果总结表”的推荐格式发生变化，此表格式也应随之变化。

行标题（row headings）

行标题应由结局指标构成，首先是主要结局指标。在每一结局指标所在行中，应提供来自多对干预组和对照组的可提取数据。通常，应有一行或几行针对不良结果，即使纳入的系统评价未报告相应结果，即使纳入系统评价未报告不良结果，

条目注释

1. 结局指标（outcomes）

应列出主要的有利和不利指标（那些与受试者最相关的，再评价结果完成之前确定的，为避免作者基于结果的统计学意义而不是临床意义来挑选的结局指标）。结局指标数量不宜超过7个。重要的结局指标即使无法提取有效数据也应列入表格。

如有多组干预措施相比较，表格还是应优先基于结局指标来设计，每个结果应分行呈现每两种干预措施比较的数据。

2. 风险假定（与对照组）（Assumed risk（With comparator））

每一行应给出典型对照组的风险。以上可由纳入的Cochrane系统评价报告的对照组风险得到。如果对照组风险存在明显差异，表格中每行应提及两个或三个代表性等级—低危险、适中危险、高危险受试者。只要可能，在备注列或脚注中注明一列中给定的对照组风险可能适用的受试者类型。

3. 相应风险（干预措施）

此列应列出基于前列所示对照风险的干预措施的预期绝对风险。可根据同行的各假定风险的相对效应计算（见第11章11.5.4节）。

4. 相对效应

对于二分类结局指标，应用相对危险度（RR）和比值比（OR）。即便是不同的系统评价应用了不同合并统计方法，只要可能，纳入的系统评价的合并统计分析也应标准化。不论应用随机效应模型或固定效应模型，均应统一使用95%置信区间来表示不确定性。对于某特定结局指标而言，所有相关结果都应使用相同的效应模型。

5. 受试者人数和研究数量

多数情况下，对于某一结局指标和治疗对照组，数据可用的研究和受试者数量将少于从Cochrane系统评价所报道的数据中提取的研究和受试者的数量（因为Cochrane系统评价可能纳入没有报告特定结局指标或特定对照组对比的研究）。如果这样，此列的受试者人数和研究数量应该仅反映提供了结局指标和对照组数据的部分。

6. 质量

对表格中每行的证据质量进行备注（需注意，因为不同行内可能包含从不同系统评价或同一系统评价纳入的不同原始研究所提取的数据，不同行的证据质量可能因此不相同。）推荐使用GRADE工作组开发的GRADE标准（GRADE Working Group 2004），并且被整合到对Cochrane 系统评价作者开放的用于制作“结果总结表”的软件中。评价证据质量所用的系统和方法应在再评价方法学部分予以描述。

7. 备注（comment）

备注的目的旨在帮助解释每行中出现的信息和数据。如，这可能是关于结局指标测量或效应修饰的效度。此处应标记关于结果的重要说明。并非所有的行都需要备注，所以如果没有重要信息需要备注最好保持空白。

连续性结果测量

将有临床意义的连续性结果列入再评价表格中，明确注明易于解释的单位，如，疼痛天数或头痛频率。然而，许多量表不容易向非专业医生或患者解释清楚，如，贝克（Beck）忧郁量表评分或生活质量评分等。对于此类不易解释的连续性数据，如有可能，根据其风险变化程度（如，风险增加了50%）来报告结果将更有利于结果的解释（见第12章，12.6)。

结局指标的标注应简洁，如，“功能状态”好于“日常功能执行能力”。如果需要结局指标定义的具体细节，可加设脚注。

异质性

一般情况下异质性的分析讨论不应列入结果总结表格。但如果（1）异质性对临床和统计学结果有重大改变；（2）是重要的效果修饰因子，则必须在备注栏中予以报告。一项重要的效应修饰偶尔需要单列一行或单列表格加以描述，如，动脉内膜切除术对不同等级狭窄的不同效果。

表22.3.b “系统评价再评价表”样板

干预措施在【条件】下对于【受试者】的效果研究							
结局	干预措施和对照措施	说明比较风险（95%可信区间）		相对效应（95%可信区间）	受试者人数（研究）	证据质量（GRADE）	备注
		假定风险	相应风险				
		有对照措施	有干预措施				
结局#1							
	干预措施和对照措施#1						
	干预措施 和对照措施#2						
	等等						
结局#2							
	干预措施和对照措施#1						
	干预措施和对照措施#2						
	等等						
结局#3							
	干预措施和对照措施#1						
	干预措施和对照措施#2						
	等等						

22.3.6.3 其他表格

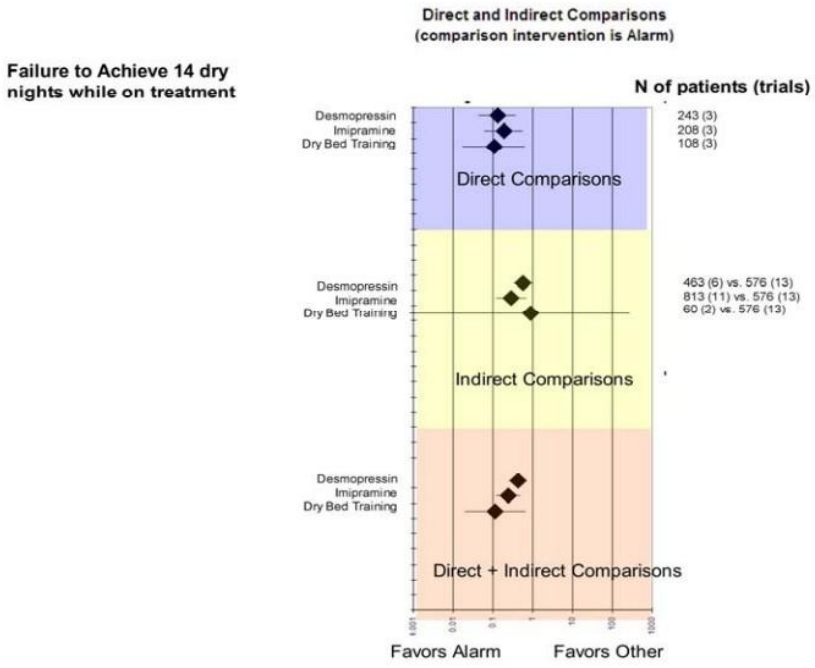
正文、“纳入系统评价特征一览表”和“系统评价再评价表”不方便放置的信息也可用其他形式表格呈现，例如：

- 支持背景的信息
- 检索策略细节
- 纳入系统评价质量评价的细节
- “结果总结表”，由再评价作者准备，针对纳入的系统评价并且其中未出现。

22.3.7 图

使用1到2（最多）幅图可帮助读者更直观的领会系统评价比较的干预措施效果的差异。再评价图首选森林图（forest top plot），图中每一行代表相比较的两项干预措施的Meta分析结果（合并效应及其95%置信区间）。每张图应呈现单一的结局，但可包括干预措施的配对比较。直接比较、间接计算比较和直接间接混合的计算比较都可一张图中展示，但必须清楚标记。正文应提供此方法的相关信息。图22.3.c列举了Russell 2006关于遗尿的再评价的图示。

图22.3.c 儿童遗尿干预措施比较的“forest top plot”图样板（由excel 制成）



22.4 本章信息

作者：Lorne A Becker and Andrew D Oxman.

本章引用格式：Becker LA, Oxman AD. Chapter 22: Overviews of reviews. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from www.Cochrane-handbook.org.

致谢：Cochrane再评价的方法由Cochrane协作网指导小组召集的工作组制定，包括Lorne Becker (召集人), Jon Deeks, Paul Glasziou, Jill Hayden, Steff Lewis, Yoon Loke, Lara Maxwell, Andy Oxman, Rebecca Ryan, Denise Thomson, Peter Tugwell和Janet Wale. 感谢他们的贡献，也感谢Lesley Gillespie, Helen Handoll和Julian Higgins对草稿的点评。

22.5 参考文献

Glenny 2005

Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, Bradburn M, Eastwood AJ. Indirect comparisons of competing interventions. Health Technology Assessment 2005; 9: 26.

GRADE Working Group 2004

GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004; 328: 1490-1494.

Oxman 1994

Oxman AD. Checklists for review articles. BMJ 1994; 309: 648-651.

Petticrew 1999

Petticrew M, Song F, Wilson P, Wright K. Quality-assessed reviews of health care interventions and the database of abstracts of reviews of effectiveness (DARE). NHS CRD Review, Dissemination, and Information Teams. International Journal of Technology Assessment in Health Care 1999; 15: 671-678.

Russell 2006

Russell K, Kiddoo D. The Cochrane Library and nocturnal enuresis; an umbrella review.Evidence-Based Child Health 2006; 1: 5-8.

Shea 2006

Shea B, Boers M, Grimshaw JM, Hamel C, Bouter LM. Does updating improve the methodological and reporting quality of systematic reviews? BMC Medical Research Methodology 2006; 6: 27.

（陈群飞译，刘雅莉、袁金秋、岑啸、秦天强初审）

附录 A　Cochrane 方法学研究方案和系统评价内容的指南

A.1　简介

Cochrane系统评价数据库中除了常见的关于医疗保健干预效果的Cochrane系统评价外，还刊载有Cochrane方法学的系统评价。他们都是由一个Cochrane系统评价小组生产制作的：位于挪威奥斯陆的Cochrane方法学系统评价小组。Cochrane方法学系统评价与干预措施系统评价的结构类似，但是在各部分标题上有一些细微的变化，以表面它涉及的是在健康与社会保健领域评估研究方法学的研究，、而不是卫生保健本身。它们以公开发表的计划书为先导，并且与卫生保健的Cochrane系统评价的制作一样地严谨和注重细节。例如，纳入或排除方法学研究的证据都是基于明确的标准。每篇系统评价都涵盖了具体并且有明确定义的方法学领域和所纳入研究的数据，这些数据可能会进行统计学的合并以增强结果的效能。在这种情况下，除了总体的均值，系统评价可包括用来描述每一个单独研究数据的图或表。

运用RevMan软件有助于根据要求的格式呈现系统评价文章。在该附录中重复了很多第4章的内容（针对干预措施的Cochrane系统评价），我们讨论整个系统评价（或计划书）的内容及涉及每一部分的大纲。手册中对其他章节的广泛引用以相应部分的指示性标注表示。

A.2　标题和系统评价信息（或方案信息）

A.2.1　标题

标题要简洁地说明所要评价的方法学，以及该方法学所针对的问题。

A.2.2　作者

所有科学论文（包括Cochrane研究方案和系统评价）的著作权包含义务、责任和信誉（Rennie 1997，Flanagin 1998，Rennie 1998）。在确定Cochrane系统评价作者署名的时

候，重要的是要分清对该篇文章做了实质性贡献（和谁应该被列出）的人和那些以其他方式给予了帮助的人，后者应该注明在感谢部分。基于“生物医学期刊投稿的统一要求”（国际医学期刊编辑委员会2006），著作权应该根据以下三个方面的实质贡献进行判断。所有作者必须签署一份“出版许可证”表格以确定这些贡献。

- 研究设计和构思，或者数据分析和解释
- 起草系统评价文章或者仔细地评阅内容
- 最后批准发表

作者可以是一个人、几个人、一个合作团队（例如，晚期膀胱癌再评价合作组）或一个或多个作者与合作团队的组合。理论上讲，作者排序应该与他们对该系统评价的相对贡献有关。贡献最多的应该被列为第一作者。

A.2.3 联系人

应提供与将要进行的系统评价的通信有关人员的联系信息。通常，这个人应：（i）负责完善和组织评价小组；（ii）与编辑沟通；（iii）确保在规定期限内准备好系统评价全文；（iv）做好标记以向编辑投稿；（v）与合作者沟通反馈；（vi）确保做好更新的准备。

联系人不一定是第一作者，联系人的选择不会影响该系统评价的引用。如果目前的联系人不再愿意对所发表的系统评价负责而且其它的评价小组成员也不愿意对此负责，那么系统评价小组协调人（RGC）的联系信息应在此列出。系统评价的联系人也不一定是作者。

A.2.4 日期

A.2.4.1 评为最新的

系统评价上一次被评为最新的日期通常是作者将系统评价提交到Cochrane系统评价数据库以供发表的日期。

另请参阅

- 描述为最新的系统评价的具体标准见第3章（第3.2节）。

A.2.4.2 检索日期

该日期用于帮助确定一篇系统评价是否已更新，并报告最近的更新日期。这将不会

公布在Cochrane系统评价数据库。

另请参阅

- 说明检索日期的具体标准见第3章（第3.3.3节）
- 检索方法的详细讨论见第6章（第6.3节）

A.2.4.3 预计完成时间

供内部使用的日期（不会公布在Cochrane系统评价数据库）仅仅是表明什么时候完成系统评价（按照研究方案），或下一次更新的日期（针对系统评价）。

另请参阅

- 系统评价的更新政策详见第3章（第3.1节）

A.2.4.4 首次发表的研究方案

Cochrane系统评价数据库中研究方案的期号是首次发表时的期号（例如，2007年第2期）。该日期在RevMan中是不可编辑的。

A.2.4.5 首次发表的系统评价

Cochrane系统评价数据库中系统评价的期号是首次发表时的期号（例如，2008年第1期）。该日期在RevMan中是不可编辑的。

A.2.4.6 最近的引用期号

Cochrane系统评价数据库中系统评价的现行引用版本的期号是首次发表时的期号（例如，2008年第2期）。该日期在RevMan中是不可编辑的。

另请参阅

- 引用版本详情见第3章（第3.2节）

A.2.5 新内容和旧内容

“新内容”部分应描述研究方案或系统评价自上一次发表于CDSR后的变化。系统评价的每次更新或修正，至少应记录一个“新内容”事件，包括事件类型、变化的日期和对所发生改变的描述。这种描述可能是，例如，一个简要的总结关于系统评价增加了

多少新信息（如，研究、受试者的数量或另外的分析）和系统评价结论、结果或方法部分所发生的任何重要变化。“新内容”表中与系统评价的现行引用版本无关的条目，应列入“旧内容”。

另请参阅

- “新内容”表格详情见第3章（第3.5节）

A.3 摘要

所有系统评价的全文必须包含一个不超过400字的摘要。摘要应当简洁但又不遗漏重要内容。Cochrane系统评价的摘要会发表在MEDLINE和科学引文索引，并在互联网上可免费获得。因此很重要的是，他们可以作为独立的文件来阅读。

另请参阅

- 摘要内容的指南详见第11章（第11.8节）

A.4 简明的言语总结

简明的言语总结（以前叫作“概要”）旨在以能被卫生保健消费者理解的简单方式总结系统评价。简明的言语总结在网络上可免费获得，因此能以独立的文件提供阅读。简明的言语总结包括两个部分：一个简明的标题（运用简明的术语重申系统评价标题）和一个不超过400字的正文总结。

另请参阅

- 简明言语摘要内容的指南详见第11章（第11.9节）。

A.5 文章正文

系统评价正文应该是简明和易读的。尽管Cochrane系统评价没有限定字数，但是系统评价作者应该把10000字作为文章字数上限，除非有需要写更长的系统评价的特殊原因。大多数系统评价文章都应该大大短于这个篇幅。根据Cochrane手册中如下的政策声明，系统评价应该能被不是该领域专家的普通人群读懂（Cochrane协作网2007）：

“Cochrane系统评价应该易于被未必是该领域专家而是对具有所评价主题的基本常识的人群阅读和理解。对一些术语和概念的解释可能是有用的，甚至是必不可少的。但是，过度的解释会损坏系统评价的可读性。简明和清晰对于可读性来说也至关重要。系统评价的可读性应该与一般医学期刊上写得很好的文章相媲美。”

一篇系统评价的正文包括了许多RevMan软件中具有的固定标题。作者可以在任何点添加小标题。某些特定的小标题可以推荐给所有作者，但这不是强制性的，如果它们使得个别部分不必要地缩短则应避免使用。与某特定的系统评价可能相关或不相关小标题的进一步讨论如下。

一些标题后面跟有固定的副标题，并因此没有紧跟其后的自由词，如：“方法”、“纳入标准”、“结果”和“作者的结论”。

背景

［固定的，1级标题］

良好构思的系统评价问题来源于已经形成的知识体系。背景部分应该阐明相关的知识，帮助制定系统评价的基本原则及解释为何提出的问题如此重要。这应该简明扼要（一般打印出来就1页左右）并让所要研究的方法的使用者易于理解。所有信息的来源都应该被引用。

就某问题或议题的描述

［推荐的，2级标题］

系统评价应该以对所要研究的方法学及其意义的简短描述作为开始。可以包括该方法在卫生保健评估中有多普遍的信息。

所调查的方法的描述

［推荐的，2级标题］

应在普遍使用的所有标准的或可供选择方法的背景下，描述所调查的方法。

这些方法怎样奏效

［推荐的，2级标题］

这部分内容可能会描述为什么在系统评价下的方法会对卫生保健的评估产生影响的理论推理。作者可能会指出一组经验证据，比如产生影响的类似方法或在其它条件下产生影响的同样方法。作者也有可能转向能证明这些方法产生可能影响的文献。

为什么这对系统评价很重要

［推荐的，2级标题］

背景应该明确说明系统评价的基本原理并解释为什么提出的问题如此重要。也可能会提到为什么进行该系统评价及怎样涉及到一个普遍问题的广泛评价。如果该系统评价是之前文章的更新版本，这样的声明是有益的，如“这是之前某年首次发表的Cochrane系统评价的更新版本，且之前是在某年更新”。这可能需补充一个对于较早版本主要结果的简要描述，以及对可能会更新系统评价的具体原因的声明。

研究目的

［固定的，1级标题］

这部分应该以对系统评价主要目的精确表述开始，理想的是用单独的一句话。它可能放在关于不同类型的卫生服务评估或其他不同方面的一系列具体目的之后。

方法

［固定的，1级标题］

研究方案中的方法部分应用将来时态。因为Cochrane 系统评价是随着证据的积累而更新的，研究方案中的方法部分应该针对好像有大量符合研究目的研究（即使在写的时候已经知道情况不是这样的）。

系统评价中的方法部分应该写成过去时态，还应描述得到现有系统评价结果及结论所做的事情。因为没有足够的证据，通常一篇系统评价是不能落实在研究方案中提出的所有方法。在这种情况下，建议在标题为“研究方案与系统评价间区别”（见下文）的部分列出没有落实的方法，这样它可以作为将来更新的系统评价的研究方案。

系统评价纳入研究标准

［固定的，2级标题］

研究类型

［固定的，3级标题］

基于研究实施或偏倚风险的任何纳入阈值，合格的研究设计及纳入标准应该列出来。例如，“不同方法的所有随机对照比较”或“前瞻性注册试验的所有临床试验队列”。

数据类型

［固定的，3级标题］

应在该部分叙述纳入系统评价中的方法学研究的原始材料，包括任何限制条件，例如，原始材料的特征（如，限制为随机临床试验）。Cochrane方法学系统评价的“数据

类型”的示例为“医疗保健试验，包括测量干预措施对一个或多个健康结局指标效果的临床干预试验和非临床干预试验”和“生物医学研究”。不应该在此描述亚组分析（见“方法”中的“亚组分析和异质性调查”）。

方法类型

［固定的，3级标题］

应在此部分定义所调查的方法，如果合适就以独立的小标题形式。应明确哪些比较是受关注的。Cochrane方法学系统评价的“方法类型”的示例为“分配隐藏充分与不充分的随机试验”。不应该在此描述亚组分析（见“方法”中的“亚组分析和异质性调查”）。

结局指标的类型

［固定的，3级标题］

请注意结局指标并不总是系统评价中纳入研究标准的一部分。如果不是，那么应当明确这个问题。应在此部分列出所关注的结局不论它们是否与纳入标准有关指标。Cochrane方法学系统评价的“结局指标的类型”的示例为“效应估计值的大小和方向（如，相对危险度降低、比值比、标准化的效应值）及预后因素的不平衡”和“摘要中结果随后的全部发表及最后发表与会议上陈述间的时间间隔”。

主要结局指标

［推荐的，4级标题］

主要结局指标应尽量少。如果已找出合格的研究，通常希望该系统评价应该能够分析这些结局指标，而且系统评价的结论将在很大程度上基于对这些结局指标的评价。

次要结局指标

［推荐的，4级标题］

应在此描述非主要结局指标。需要处理的结局指标总数的应尽可能少。

检索方法

［固定的，2级标题]

用于检索研究的方法应在此总结。以下是推荐的标题。在制定这部分之前，作者应该联系Cochrane方法学系统评价小组请求指导。

另请参阅

- 检索方法详情见第6章（第6.3节）。

电子检索

［推荐的，3级标题］

运用书目数据库检索，应说明检索的日期、时间段及其他任何限制条件，如语言。每个数据库完整的检索策略应该在系统评价后面的附录中说明。如果检索Cochrane方法学系统评价数据库（CMR），可以参考该注册库的标准描述但同时应该包括最近一次检索当前版本CMR的时间和方式，以及还应列出运用的检索词。

另请参阅

- 检索策略详情见第6章（第6.4节）。

检索其他资源

［推荐的，3级标题］

列出灰色文献来源，如内部报告和会议记录。应注意到如果杂志是专门手工检索的情况，作者完成的手工检索要帮助建立Cochrane方法学系统评价注册库（CMR），这不应列举出来因为这包含在注册库的标准化描述中。列出联系了的人（例如，研究者或主要专家）和组织。列出使用的其他资源，可能包括，例如，参考文献目录、万维网或个人收集的文章。以下可供选用的标题，可以替代“检索其他资源（这种情况下就是3级标题）”或作为小标题（4级）：

Grey literature灰色文献

Handsearching手工检索

Reference lists参考文献列表

Correspondence通信

另请参阅

- 其他检索资源详见第6章（第6.2节）。

数据收集与分析

［固定的，2级标题］

该部分应叙述数据收集与分析的方法。

研究的筛选

［推荐的，3级标题］

所使用的方法要依从筛选标准。应该说明由一位还是多位作者独立进行筛选以及怎样解决意见分歧的。

另请参阅

- 研究的筛选详见第7章（第7.2节）。

资料的提取和管理

［推荐的，3级标题］

该方法用于从已发表的研究报告或原始研究者（例如，运用资料提取/收集表）提取或获取数据。应该说明由一位还是多位作者独立完成以及怎样解决意见分歧的。如果是相关的，准备用于分析处理数据的方法就应加以说明。

另请参阅

- 数据收集详见第7章，包括收集哪些数据（第7.3节）、数据来源（第7.4节）、数据收集表（第7.5节）和从报告中提取数据（第7.6节）。

纳入研究的偏倚风险评估

［推荐的，3级标题］

该方法用于评估偏倚风险（或方法学质量）。应该说明由一位还是多位作者独立进行以及怎样解决意见分歧的。采用的工具应该加以描述或引用，并且要说明评价结果是怎样纳入结果解释的。

方法效应值的指标

［推荐的，3级标题］

应说明所选择效应指标。例如，二分类资料用OR、RR或RD；连续性资料用MD或SMD。以下可供选用的标题，可以替代“治疗效果的指标（这种情况下就是3级标题）”或作为小标题（4级）：

二分类资料

连续性资料

时间-事件资料

Unit of analysis issues分析单位问题

［推荐的，3级标题］

对于非标准设计研究中特殊问题的分析应加以说明，如交叉试验和整群随机试验。

另请参阅

- 分析单位问题详见第9章（第9.3节）。
- 交叉试验、整群随机试验和其他非标准设计的方法详见第16章。

缺失数据处理

［推荐的，3级标题］

缺失数据的处理方法应加以说明。主要包括方法学研究的信息缺失（例如，试验队列中试验信息的缺失）和统计数值缺失（如，标准差或相关系数）。

另请参阅

- 缺失数据的相关内容详见第16章（第16.1节）。

异质性评估

［推荐的，3级标题]

方法学研究间设计异质性问题的解决方法及作者怎样考虑Meta分析是否恰当均应加以说明。找出统计学异质性检验方法也应说明（例如，直观地使用卡方检验或I2）。

另请参阅

- 异质性评估详见第9章（第9.5节）。

报告偏倚评估

［推荐的，3级标题］

发表偏倚和其他报告偏倚是怎样处理的（如，漏斗图、统计学检验、估算）。作者应该明白不对称漏斗图不一定是由发表偏倚引起的（而且发表便宜也不一定会造成不对称的漏斗图）。

另请参阅

- 报告偏倚详见第10章。

数据合成

［推荐的，3级标题］

选择的Meta分析方法应该加以说明，包括固定效应或随机效应模型的选用。如果没有进行Meta分析，合并多个研究结果的系统方法应予以说明。

另请参阅

- Meta分析和数据合成详见第9章（第9.4节）。

亚组分析和异质性调查

［推荐的，3级标题］

所有预计的亚组分析都应列出来（或用于Meta回归的自变量）。任何调查效应值异质性的其他方法都应予以说明。

另请参阅

- 异质性调查详见第9章（第9.6节）。

敏感性分析

［推荐的，3级标题］

应该描述旨在确定结论与系统评价过程中所做出的决策是否稳健一致，如，从Meta分析中纳入/排除某个研究、填补缺失数据或分析方法的选择。

另请参阅

- 敏感性分析详见第9章（第9.7节）。

针对方法部分以下可供选择的标题可能有用（3级）：

未来更新系统评价的方法

另请参阅

- 更新系统评价的问题详见第3章。

结果

［固定的，1级标题］

研究描述

［固定的，2级标题］

检索结果

［推荐的，3级标题］

结果部分应以检索结果的概述开始（例如，通过电子检索到了多少参考文献）。

另请参阅

- 检索结果的呈现详见第6章（第6.6节）。

纳入的研究

［推荐的，3级标题］

对纳入研究的数目加以详细描述是必不可少的。该部分应包括对“纳入研究特征”表所含内容的简要概述，同时还应包括该表所包括的具体研究。纳入研究的关键特点应予以说明，包括方法、数据（例如，方法学研究中的临床试验类型）、对照和纳入研究的结局指标以及研究间其他任何的重要差异。作者应该注意到研究的其他特点，即是系统评价中认为很重要的要让读者明白的内容。以下可供选用的标题（4级）可能会很有帮助：

设计

样本含量

背景

方法

结局

排除的研究

［推荐的，3级标题］

这应包括“排除研究特征”表所含的内容。同时还应包括该表所包括的具体研究。从评价中排除这些研究的原因应该予以简要说明。以下供选择的标题（3级）可用于“研究的描述”部分：

进行中的研究

待归类的研究

更新检索后的新研究

纳入研究的偏倚风险

［固定的，2级标题］

该部分内容应归纳纳入研究结果的一般偏倚风险、所有研究的变异性和单个研究的重要缺陷。应在“方法”部分而不是在这个部分描述或说明用于偏倚风险评估的标准。在“偏倚风险”表中应报告怎样用每一条标准评价每一个研究，而不是以文本形式的详细描述，而该是简要的概述。

对于多数系统评价，偏倚风险评估的内容可概括为在以下标题中主要结局指标。

分配

［推荐的，3级标题］

概述所研究的方法在系统评价的研究中怎样分配。应在该部分概述可能由于这种分配所致偏倚风险的判断。

盲法

［推荐的，3级标题］

在方法学研究的分析和进行过程中对谁实施盲法或设盲要在该部分有一个简短的说明。还应在此概述可能与盲法相关的偏倚风险的判断。

随访和排除

［推荐的，3级标题］

每一个主要结局指标的数据完整性情况应在此简要描述。

选择性报告

［推荐的，3级标题］

应在此简要概述对数据的选择性使用，包括结局指标、亚组或分析的选择性报告证据。

其他潜在的偏倚来源

［推荐的，3级标题］

应在此概述任何其他潜在的偏倚来源。

方法的效应

［固定的，2级标题］

这应是对系统评价中所研究方法效应的主要结果的概述。该部分应直接针对评价的目的而不是依次把纳入研究的结果罗列出来。单个研究的结果及其统计汇总都应包含在“数据和分析”表中。通常应该按照“结局指标的类型”中的罗列顺序阐述结果。如果小标题使其更加容易理解，则鼓励使用小标题（例如，对于每个不同的数据、对照或结局指标如果系统评价关注的不止一个）。应报告进行的任何敏感性分析。作者应避免在这部分作推论。

另请参阅

- 结果描述详见第11章（第11.7节）。
- 数值结果的解释详见第12章（第12.4、12.5和12.6节）。

讨论

［固定的，1级标题］

结构化的讨论有助于对系统评价意义的考虑（Docherty 1999）。

另请参阅

- 结果解释详见第12章。

主要结果总结

［推荐的，2级标题］

总结主要结果（不是重复“方法效应”部分的内容）和明显的不确定因素，权衡重要的利弊。在“结果总结”表中明确描述。

证据的整体完整性和适用性

［推荐的，2级标题］

描述系统评价的问题与所得证据的相关性。这应引起对系统评价外部真实性的总体评估。被纳入的研究是否足以达到系统评价的所有目的？是否对所有相关的数据类型、方法和结局指标都已经进行调查了？该部分应包括对系统评价结果融入当前实际情况

程度的评论，尽管作者应考虑到当前实际情况在国与国之间可能是不同的。

证据质量

［推荐的，2级标题］

得出的一组证据本身能够做出关于系统评价目的的有力的结论么？总结已被纳入研究的证据的数量（研究的数量）、声明纳入研究重要的方法学局限性和重申其结果的一致性或不一致性。这会引起对评价内部真实性的总体评估。

系统评价过程中的潜在偏倚

［推荐的，2级标题］

声明系统评价中关于防止偏倚的优势和局限性。这些可能是系统评价作者所控制的或控制之外的因素。讨论可能包括找出所有相关研究的可能性，是否能够获得所有相关数据或是否所使用的方法会引入偏倚（例如，检索、研究筛选、数据提取、分析）。

与其他研究或系统评价的一致和分歧

［推荐的，2级标题］

该部分应包括对纳入研究怎样符合其他证据的评论，并明确声明其他证据是否经过系统地评价。

作者的结论

［固定的，1级标题］

系统评价的主要目的应该是提供信息而不是提供建议，作者的结论应分为两部分：

卫生服务评估和系统评价的意义

［固定的，2级标题］

系统评价和其他卫生服务评估的意义应尽可能实际和明确。他们不能超出被评价了的和被系统评价中的数据证明了的证据。“没有效果的证据”不应与“无效的证据”混为一谈。

方法学研究的意义

［固定的，2级标题］

当人们对今后的研究做决策的时候可能会用到Cochrane方法学系统评价的该部分内容，作者应该试着写一些对此有用的东西。正如“对于实践的意义”，其内容应基于可得的证据并避免运用没有在系统评价中讨论或纳入的信息。在本节的编写中，作者应考虑到研究的不同方面，可能利用的研究、数据、方法和结局指标的不同类型作为其框架结构。研究可能被如何实施和报告的意义应该与将来应该实施的研究区分开来。例如，

对随机对照试验而不是其他类型研究的需要，对于某些特定主题的系统评价中研究更好地描述的需要或者对特定结局指标进行常规收集的需要，应与对特殊类型的对照的比较或者在特定条件下研究的需要区分开来。

重要的是，这部分内容尽可能清楚和明确。总体陈述包括很少或没有具体信息，如“将来的研究应更好地实施”或“还需要进一步研究”对于决策者来说都没什么用，应避免这种情况。

另请参阅

- 形成结论的指导详见第12章（第12.7节）。

致谢

［固定的，1级标题］

该部分应用于感谢那些作者想要感谢的人们或组织机构，但没有在作者中列出来的人员。这可以包括以前的Cochrane系统评价作者或支持该篇系统评价的以前的资源，并且可能包括Cochrane方法学评价组的编辑小组。应当从所要感谢的人那获得许可。

作者的贡献

［固定的，1级标题］

在此应描述当前共同作者的贡献。有一名作者应当作为该篇系统评价的担保人。在系统评价投稿和发表在Cochrane系统评价数据库（CDSR）之前，所有作者应该讨论并商定各自的贡献。在系统评价更新时，这部分内容也要作必要的核实和修改以确保其是准确的和最新的。

下列采纳的可能贡献名目来自Yank等人（Yank 1999）。这是一个建议的内容清单，应描述人们做了些什么而不是找出他们的贡献属于哪一类别的内容。理论上讲，作者应该用自己的话描述其贡献：

- 构思系统评价。
- 设计系统评价。
- 协调系统评价。
- 为系统评价收集数据。
  - 制定检索策略。
  - 进行检索。
  - 筛选检索结果。
  - 整理检索文献。

    - 根据纳入标准筛选检索文献。
    - 评价文献质量。
    - 从文献中提取数据。
    - 向论文作者索取额外信息。
    - 为论文提供补充数据。
    - 获取和筛选未发表研究中的数据。
  - 系统评价的数据管理。
    - 把数据录入RevMan软件。
- 分析数据。
- 解释数据。
    - 提供方法学观点。
    - 提供临床观点。
    - 提供政策观点。
    - 提供消费者的观点。
  - 撰写系统评价。
  - 提供对系统评价的总体建议。
  - 确保系统评价的资金支持。
  - 为当前的系统评价做前期准备工作。

利益声明

［固定的，1级标题］

作者应报告任何现在的或过去的从属关系或其他任何组织机构或实体介入的利益关系从而可能引起真实存在或可能的利益冲突。可能会被其他人认为能影响系统评价作者判断的情况包括来自个人、政治、学术和其他可能的冲突以及经济利益冲突。一旦系统评价作者中涉及系统评价中纳入的研究，作者必须声明。

另请参阅

- 协作网关于利益冲突政策的总结详见第2章（第2.6节）。

经济利益冲突最受人关注，而是应该避免的，但是一旦存在，则必须报告。任何可能过度影响评价过程中判断（如研究的纳入或排除，纳入研究的真实性评估或结果的解释）的次要利益（如个人冲突）都应予以说明。

如果没有已知的利益冲突，应该明确指出，例如，注明“未知”。

研究方案与系统评价间的差异

［固定的，1级标题］

有时需要运用不同于研究方案中最初描述的方法。这可能是因为：

- 处理某特殊问题的方法没有在研究方案中详细说明；
- 研究方案中的方法不能用(例如,由于数据不充分或方法实施需要的信息缺乏)；
- 发现了一种更好的替代方法从而改变原来的方法。

从研究方案到系统评价一些方法发生改变是可以被接受的，但是必须在该部分充分说明。本节提供了系统评价方法随着时间发生一些主要改变的概述。应用于以下情况：

- 指出在最近发表的研究方案后做出的任何决定（例如，增加或改变结局指标；增加“偏倚风险”或“结果总结”表格)。
- 总结在当前系统评价中不能执行的研究方案中的方法（如没有研究进入预先定义的亚组)。
- 在解释从研究方案到系统评价方法上的任何改变时，说明是什么时候改变的及提供改变的正当理由。这些改变不应该基于所调查方法效果的结果驱使。要考虑到方法改变对系统评价结论的潜在影响及考虑进行敏感性分析来评估。

出版注释

［固定的，1级标题］

出版注释会出现在Cochrane系统评价数据库中的系统评价中。可能会包括来自Cochrane方法学系统评价小组（CMR）的编辑注释和评论，例如被编辑或评审员强调的部分是被认为在系统评价总值得发表的部分。应注明作者或这些评论的来源（例如，来自编辑或评审员)。

所有撤销的研究方案和系统评价也必须完成出版注释，并给出撤销原因。对于撤销的研究方案和系统评价，仅有基本引用信息、资助来源和出版注释被发表。

A.6 表格

A.6.1 纳入研究的特征

“纳入研究的特征”表对每个研究有5个条目的内容：方法、数据、对照、结局和备注。有多达3方面的内容用于说明不便于不包括这些类别的项目，例如，提供随访时

间的长短、资金来源或不太可能直接引起偏倚风险的研究质量指标。

表中可能会使用代码或缩略语以便清楚简洁地表述一个条目中多方面的信息内容。应用脚注解释所使用的代码或缩略语（这些将会在CDSR发表）。

A.6.2 偏倚风险

尽管被强烈推荐使用，“偏倚风险”表仍是可选择性的，而且是“纳入研究的特征”表的拓展。标准的“偏倚风险”表包括对隐藏分配的评估和作者可以进一步增加条目。对于每个条目，该表提供了研究中已报告发生的内容和关于防止偏倚的主观判断（“是”代表低偏倚风险，“否”代表高偏倚风险，否则就是“不清楚”）。

A.6.3 排除研究的特征

满足纳入标准的研究或貌似符合纳入标准的研究被排除后应罗列出来并给出排除的原因（例如，不恰当的干预对照）。这应简要说明，且通常来说一个排除的理由就足够了。

另请参阅

- 选择列出哪些研究作为排除的研究详见第7章。

A.6.4 待分级研究的特征

“待分级研究的特征”表（以前称为“待评价的研究”）与“纳入研究的特征”表是一样的结构。应用于两类研究：

因无法获得足够的信息以作出纳入或排除决定的研究。所有获取信息的合理尝试都必须在系统评价发表之前完成，但是系统评价不能为了等待获得信息而被过分延后，特别是如果纳入或排除的研究不太可能影响系统评价的结论。当表中对应条目的信息无法获得时，应填入“不详”。

已被找到但还在等待系统评价更新的研究。尤其是对系统评价结论有潜在影响的研究或受到广泛关注的研究需要在系统评价更新期间有所提及。那些表中总结的研究可能因此生成一篇修改的系统评价。应尽快完成完全纳入这些研究的所有更新。当表中对应条目的信息无法获得时，应视情况而定填入“尚未评估”或“不详”。

A.6.5 进行中研究的特征

“进行中研究的特征”表对每个研究有8个条目的内容：研究名称、方法、数据、对照、结局指标、开始日期、联系信息和备注。这些条目内容与“纳入研究的特征”表中的条目内容相当。应用脚注解释表中所使用的缩略语（这些将会在CDSR发表）。

A.6.6 结果总结表

不论证据是否可用，对于重要的结局“结果总结”表是表述结果的一种可供选择的方法。“结果总结”表在适当的条件下包括证据数量总结、不同方法特有的绝对风险、相对效应估计值（如RR或OR）、证据本身质量的描述，还有评论和注解。对证据本身质量的评估应遵循GRADE分级框架（the GRADE framework），其综合考虑了偏倚风险、直接性、异质性、精确度和发表偏倚。

另请参阅

- 关于“结果总结”表全面的说明和讨论详见第11章（第11.5节）；
- 分级制度（The GRADE system）的评价详见第12章（第12.2节）。

A.6.7 附加的表格

不便于放入文中或固定表格中的信息可使用附加的表格。如下：

- 支持背景资料的信息；
- 研究特征的总结（如对所调查方法或结局的详细说明）；
- 不适合写入“数据和分析”表中的结果，列如，偏态数据报告中位数和极差。

A.7 研究和参考文献

A.7.1 研究的参考文献

研究以下四个固定的标题进行组织：

纳入的研究

符合纳入标准并被系统评价纳入的研究。

排除的研究

不符合纳入标准被系统评价排除的研究。

待分类的研究

已被找到但直到获得更多的数据或信息才能纳入评估的研究。

进行中的研究

正在进行并且符合（或貌似满足）纳入标准的研究。

这些标题每个都可以包括多个研究（或没有研究）。每个研究都以相应的“研究身份证”（通常包括第一作者的姓和研究最初引用的年限）。每个研究涉及一年时间（通常完成的时间或最初引用的发表年限）。此外，每个研究都应该分入以下“数据来源”的类别中。

- 只是发表的数据。
- 发表和未发表的数据。
- 只是未发表的数据。
- 只是发表的数据（不使用未发表的）。

每个研究可有多个参考文献。每篇参考文献都应有识别码，如一个MEDLINE ID或一个DOI。每个研究的参考文献都应注明是“初次引用”。作者应该核实所有参考文献的准确性。

A.7.2 其他参考文献

除了以上类别的参考文献，研究的其他参考文献分为以下两类：

附加参考文献

正文中引用的其它参考文献应该在此列出，包括背景和方法部分引用的。如果某研究的报告被正文引用是因为其他一些原因而不是因为要引用该研究（例如，由于参考文献中有一些背景或方法学信息），那么它除了在相关研究中列出同时应该在此部分列出。

该系统评价的其他发表版本

期刊、书本或CDSR或其他地方发表的系统评价版本的参考文献应该在此罗列出来。

注：RevMan也包括一个“待定的分级”类别以便于准备系统评价时组织参考文献。在系统评价提交给CDSR之前，所有参考文献都应从这个类别中移出，因为如果仍保留其中是不会被发表的。

作者应该核实所有参考文献的准确性。

A.8 数据和分析

系统评价中纳入研究的结果是以等级的形式组织：研究——（可选择的）亚组——结局指标——对比组。

RevMan软件能自动生成森林图用于解释进入“数据和分析”结构中的数据、效应估计值和Meta分析结果（选用这一方法时）。作者能够控制是否进行和怎样进行Meta分析。

注：“数据和分析”应该被认为是补充信息，因为这在以某种形式发表的系统评价可能不会出现。关键的森林图（包括每个研究的数据）可能总被选入系统评价的全文作为图表（见第A.9节）。但是，发表在CDSR的Cochrane系统评价全文会包括所有“数据和分析”部分的内容作为一系列的森林图或表格。

作者应避免列出没有数据（如没有任何研究的森林图）的结局或对照组。相反，作者应该注意系统评价全文中对于对照没有数据可用。系统评价的主要结局指标应包括在一个“结果总结”表中，不论纳入研究的数据是否可用。

对照

对照应该对应于基于“目的”的假设和问题。

结局

结局的数据类型可能有5种：二分类数据、连续性数据、“O-E”和“V”统计量、一般到方差（估计值和标准误）和其他数据（只能是文本）。

亚组

亚组可能是纳入研究的分组（如及CONSORT声明发表前和发表后进行的研究）或研究结局的分组（如短期、中期、长期）。

研究数据

针对结局数据的类型每个研究的数据都必须以特定的形式被录入（如连续性数据每组的样本含量、均数和标准差）。

另请参阅

- 不同类型的数据、统计分析和Meta分析详见第9章。

A.9 图形

介绍性文字说明

一篇系统评价中可能包含有5种图形。这些图形将常常被呈现在发表的系统评价的全文中。每一个图都必须有一个标题，以提供对图的简短的说明（或解释），而且必须在系统评价的内容中提到（与之联系）。

另请参阅

- 图形的选择问题详见第11章（第11.4.2节）。

A.9.1 RevMan图形和表格

在“数据和分析结果”部分可选择森林图和漏斗图作为图表。风险偏倚判断的图形表示也能运用RevMan软件生成并纳入作为图表。

另请参阅

- 森林图详见第11章（第11.3.2节）。
- 漏斗图详见第10章（第10.4节）。
- “偏倚风险”图和“偏倚风险”总结详见第8章（第8.6节）。

A.9.2 其他图形

不能由RevMan软件生成的图形或其他图像可以被纳入作为文中的图。用RevMan的其他方式生成的图不应该用于内容，如作为森林图或其他附加表格。

作者应负责获取系统评价中图形的许可及按照指导内容确保图形适合发表。如果得到了发表受版权保护图形的许可，那么图形标题的最后必须注明：“版权所有©[年][版权持有人姓名，或其他所需的措词]：未经许可，不得转载”。

另请参阅

- 统计分析的图形应遵守Cochrane统计方法学小组的相关指导内容（见手册上附加材料的网址：www.cochrane.org/resources/handbook）。

A.10 系统评价的资助来源

作者应确认系统评价所得到的资助金和其他形式的资助，比如来自其大学或单位机构以薪水形式给予的资助。资助金来源分为“内部”（由系统评价制作机构提供）和“外部”（由其他机构或资助机构提供）。应提供每一个资助来源的原始国家和资助的项目。

A.11 反馈

系统评价中的每一条反馈信息都标有小标题和日期。概要、答复和贡献者是本部分内容的副标题。向Cochrane 方法学系统评价小组反馈编辑咨询后编写总结，必要的话，还要向提交评论的人员咨询。系统评价作者应准备一份答复。在“贡献者”一栏中应注明在回应反馈过程中作出贡献的人员姓名。

另请参阅：

- 关于反馈的进一步信息详见第3章（第3.6节）。

A.12 附件

附件提供了一个补充信息的地方，如：

- 详细的检索策略（推荐在附件中放入这些信息）；
- 非标准统计方法的详细说明；
- 数据提取表格
- 详细的结局指标（如，测量尺度）。

附件不太可能以某些正式发表的系统评价的格式出现

A.13 附录信息

编辑：Mike Clarke, Andrew D Oxman, Elizabeth Paulsen, Julian PT Higgins 和Sally Green。

该附录应被引用为：Clarke M, Oxman AD, Paulsen E, Higgins JPT, Green S（编辑）。附录A：Cochrane方法学研究方案和系统评价的内容指南。In: Higgins JPT, Green S（编辑），Cochrane干预措施系统评价手册5.0.1版本（2008年9月更新）。The Cochrane Collaboration, 2008. Available

from http://www.cochrane.org/resources/handbook.

作者：该附录以之前版本的手册为基础，Julian Higgins 和Sally Green编写了Cochraneg干预措施系统评价章节。具体贡献详见第4章（第4.3节）。

A.14 参考文献

Docherty 1999

Docherty M, Smith R. The case for structuring the discussion of scientific papers. BMJ 1999; 318: 1224-1225.

Flanagin 1998

Flanagin A, Carey LA, Fontanarosa PB, Phillips SG, Pace BP, Lundberg GD, Rennie D. Prevalence of articles with honorary authors and ghost authors in peer-reviewed medical journals. JAMA 1998; 280: 222-224.

International Committee of Medical Journal Editors 2006

International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals: Writing and editing for biomedical publication [Updated February 2006]. Available from: http://www.icmje.org (accessed 1 January 2008).

Rennie 1998

Rennie D, Yank V. If authors became contributors, everyone would gain, especially the reader. American Journal of Public Health 1998; 88: 828-830.

Rennie 1997

Rennie D, Yank V, Emanuel L. When authorship fails. A proposal to make contributors accountable. JAMA 1997; 278: 579-585.

Yank 1999

Yank V, Rennie D. Disclosure of researcher contributions: a study of original research articles in The Lancet. Annals of Internal Medicine 1999; 130: 661-670

（高霑译）