

Meta 分析实践

基于 BCG 疫苗试验的方法演练与 R 语言实现

作者：晨瀚宇（Chanw）

时间：May 22, 2026

版本：1.0

同一份数据、同一个临床问题，十种 meta 分析方法的演练与比较。

© 2026 晨瀚宇. 保留所有权利。

版权声明

书名：Meta 分析实践——基于 BCG 疫苗试验的方法演练与 R 语言实现

作者：晨瀚宇（小红书：Chanw）

版本：v1.0（2026 年 5 月）

© 2026 晨瀚宇. 保留所有权利。

本书内容（包括但不限于文字、代码、图表、排版设计）均为作者原创。

禁止事项：

- 未经作者书面授权，禁止以任何形式转载、复制、传播本书的全部或部分内容
- 禁止将本书内容用于商业用途或二次售卖
- 禁止去除、修改本书中的版权标识和作者署名

允许事项：

- 个人学习和研究用途的合理引用（需注明出处）
- 在社交媒体上分享本书的购买链接

如发现侵权行为，请联系作者。

购买渠道：小红书搜索「Chanw」

前言

这本书的写作思路是：选定一份具体的数据集，把所有相关 meta 分析方法都在它身上跑一遍，让读者亲手看到每一种方法在同一个问题上能给出什么答案、各自的假设和局限在哪里。

本书用的数据是 metafor 包内置的 *dat.bcg*，13 项 1948 至 1980 年间在世界各地完成的 BCG 卡介苗 RCT，回答的核心问题是把多项独立研究合并起来做证据综合时，方法的选择会如何影响最终结论。这份数据是 meta-analysis 教学的事实标准样本，临床故事性强，纬度跨度从赤道到北纬 55°，异质性显著，适合把每种方法的表现差异都展示出来。

读者画像是在读研究生与青年研究者，能用 R，看过几篇 meta 分析的论文，但没有亲手跑过完整流程的那一批。本书不打算把 Borenstein 的两卷本和 Cochrane Handbook 重复一遍，更不打算把所有方法都讲到学术教科书的精度。这本书的野心比较有限：让读者跑通一项 BCG 试验级别的 meta 分析、看懂常见输出、知道每个数字背后的假设是什么、知道踩坑的常见地方在哪儿。读完之后愿意去翻 Cochrane Handbook 第 10 章作为参考、愿意去看 Schwarzer 的 R 教程把更复杂的情形深入处理，这本书的任务就完成了。

晨瀚宇
2026 年 5 月

目录

前言 **i**

第 1 章 问题与数据：BCG 卡介苗 13 项随机对照试验 **1**
1.1 BCG 卡介苗的全球效果争论 . . . 1
1.2 一项 BCG 试验的数据结构：2x2 表 . . . 1
1.3 风险比 RR 的定义与计算 . . . 2
1.4 13 项研究的答案一字排开 . . . 3
1.5 朴素合并：把 13 张 2x2 表直接相加 . . . 4
1.6 全书章节与方法路线 . . . 4

第 2 章 PRISMA 与系统综述流程 **6**
2.1 PRISMA 2020 与四阶段流程 . . . 6
2.2 PICO 框架与检索式构造 . . . 8
2.3 筛选流程与 Cohen's κ . . . 9
2.4 数据提取与 Cochrane RoB 2 . . . 10

第 3 章 效应量及其方差的计算 **12**
3.1 对数风险比 log RR . . . 12
3.2 log RR 的方差 . . . 12
3.3 比值比 OR 与对数比值比 . . . 13
3.4 风险差 RD . . . 14
3.5 escalc 函数：一站式计算 . . . 15

第 4 章 固定效应模型与逆方差加权 **16**
4.1 为什么不能按样本量加权 . . . 16
4.2 固定效应模型的假设 . . . 16
4.3 逆方差加权法 . . . 17
4.3.1 每项研究的权重分布 . . . 17
4.4 Mantel-Haenszel 加权 . . . 18
4.5 Forest plot . . . 19
4.6 累积对比表（首次出现） . . . 20

第 5 章 随机效应模型与 τ^2 估计 **21**
5.1 从固定效应到随机效应 . . . 21
5.1.1 随机效应模型的权重 . . . 21
5.2 DerSimonian-Laird 估计量 . . . 22
5.3 REML 估计 . . . 23
5.4 Hartung-Knapp 调整 . . . 23
5.5 预测区间 PI . . . 24
5.6 Forest plot 带预测区间 . . . 24

第 6 章 异质性诊断：Q、I^2、τ^2 与预测区间 **26**
6.1 异质性的两个来源 . . . 26

6.2 Cochran's Q 检验 . . . 26
6.3 I^2 与 H^2 . . . 27
6.4 τ^2：异质性的绝对量 . . . 27
6.5 异质性的视觉诊断 . . . 28
6.6 累积对比表与异质性指标完整报告 . . . 29

第 7 章 亚组分析与 meta 回归 **31**
7.1 亚组分析与 meta 回归的关系 . . . 31
7.2 亚组分析：按分配方式 alloc . . . 31
7.2.1 亚组之间的差异检验 . . . 32
7.3 Meta 回归：纬度 ablat . . . 32
7.3.1 Bubble plot：协变量与效应的散点 . . . 33
7.4 多变量 meta 回归 . . . 34
7.5 条件 RR 报告 . . . 34

第 8 章 发表偏倚的诊断与校正 **36**
8.1 发表偏倚的产生机制 . . . 36
8.2 漏斗图 . . . 36
8.2.1 Contour-enhanced 漏斗图 . . . 37
8.3 Egger 回归检验 . . . 38
8.4 Begg 秩相关检验 . . . 39
8.5 Trim-and-fill 校正 . . . 39
8.6 发表偏倚不显著但仍要警惕的场景 . . . 40

第 9 章 敏感性分析与影响诊断 **41**
9.1 Leave-one-out 分析 . . . 41
9.2 累积 meta 分析 . . . 42
9.3 影响诊断 . . . 43

第 10 章 方法对比与 GRADE 证据等级 **46**
10.1 全书方法终极对比 . . . 46
10.1.1 结论的一致与分歧 . . . 46
10.1.2 该报告哪一个？ . . . 47
10.2 GRADE 证据等级框架 . . . 47
10.2.1 五个降级因素在 dat.bcg 上的评估 . . . 47
10.2.2 综合评级 . . . 48
10.3 时代局限与证据更新 . . . 48
10.4 终结对比表 . . . 49

Bibliography **50**

第 1 章 问题与数据：BCG 卡介苗 13 项随机对照试验

内容提要

- ❑ 看清 BCG 在不同地区给出矛盾答案的真实数字
- ❑ 学会读 2x2 表：每一格代表的是哪一群人
- ❑ 掌握风险比 RR 的算法与含义
- ❑ 理解为什么把 13 张 2x2 表直接相加不可行

1.1 BCG 卡介苗的全球效果争论

1948 年，全世界每年还有几百万人死于结核病。一种叫做卡介苗的活减毒疫苗已经在欧洲使用了二十多年，英文称 BCG，全称 Bacillus Calmette-Guérin。

BCG 的工作原理可以这样理解。免疫系统平时不认识结核菌长什么样，等到结核菌真的进入身体，免疫系统才开始慌乱反应。BCG 是把活的、被减毒过的结核菌一次性给免疫系统看一眼，让它”认识”这种细菌的家族特征。等真正的结核菌进来时，免疫系统能一眼认出来，立刻处理掉。

这套机制的合理性看起来无可辩驳，但 BCG 在全球的实际效果争论极大。法国的研究说效果惊人，美国的研究说几乎没用，印度南部一项最大规模的随机对照试验干脆得出”完全无效”的结论。同一种疫苗，不同的研究地点，结论却天差地别。

为了让”全球结论矛盾”不只是一个抽象说法，先看三项最具代表性的试验真实给出的数字：

- Aronson 1948 在美国土著儿童 262 人里发现 BCG 把结核发病风险**降低了 60%**。
- Hart & Sutherland 1977 在英国 26,465 人里发现 BCG 把发病风险**降低了 76%**。
- TPT Madras 1980 在印度南部 176,782 人里发现 BCG **几乎完全无效**。

三项都是大型随机对照试验，执行质量都不差，但答案差出了天与地。如果一位医生只读到 Aronson 这篇论文，会得到”BCG 是个非常有效的疫苗”的结论；只读到 TPT Madras 那篇，会得到”BCG 不值得推广”的结论。临床决策不能这样靠运气挑论文。要回答”BCG 在全球到底有用吗”，必须把所有相关研究放在一起整理、合并。这件事就是 **meta 分析**（中文常译为”元分析”或”汇总分析”）。

到了 1990 年代初，Colditz 和同事把 13 项 1948–1980 年间在世界各地完成的 BCG 随机对照试验整理成一份汇总数据集，发表在 JAMA 上 [1]。这份数据后来被收进 Viechtbauer 编写的 metafor 包，称为 *dat.bcg*，成为 meta 分析教学的事实标准样本，相当于机器学习领域的 iris、回归分析教学里的 mtcars。**本书全书都用这一份数据**。每一章用一种 meta 分析方法处理这同一份数据，到第 10 章把所有结论汇总比较。

1.2 一项 BCG 试验的数据结构：2x2 表

每项 BCG 试验做的事其实都一样：把一群健康人随机分成两组，一组打 BCG（叫**处理组**），一组不打或打安慰剂（叫**对照组**）。隔几年看哪组发病的人更多。

每项试验完整的核心结果都可以装进一张 2 行 2 列的小表格，行业通称 **2x2 表**。以 Aronson 1948 这项美国土著儿童研究为例，全员 262 人，跟踪一年多结束之后的结果整理成表 1.1。

表 1.1: Aronson 1948 BCG 试验的 2x2 表

	后来发病了	后来没发病	这一组合计
打了 BCG（处理组）	$a = 4$ 人	$b = 119$ 人	$a + b = 123$ 人
没打 BCG（对照组）	$c = 11$ 人	$d = 128$ 人	$c + d = 139$ 人

读这张表的方法是一格一类人。左上角红色深的格子 $a = 4$，意思是这项试验里 **打了 BCG 又发病**的孩子有 4 个。右上角蓝色浅的格子 $b = 119$，意思是 **打了 BCG 没发病**的孩子有 119 个。左下角红色浅的格子 $c = 11$，意思是 **没打 BCG 又发病**的孩子有 11 个。右下角蓝色深的格子 $d = 128$，意思是 **没打 BCG 没发病**的孩子有 128 个。最右边那一列是每组的总人数：处理组 $a + b = 123$，对照组 $c + d = 139$，加起来正好是这项试验的全部 262 人。

为什么要画这张表？因为本书后面所有 meta 分析方法，本质上都从这四个数字开始。**13 项研究就是 13 张这样的 2x2 表**。metafor 包内置的 *dat.bcg* 里四列数字 *tpos*、*tneg*、*cpos*、*cneg* 依次对应表 1.1 里的 a、b、c、d。本书后续章节里出现的字母 $a/b/c/d$ 都回到这张图。

停下来想一想。如果你能告诉别人某项试验的 a、b、c、d 四个数字各是多少，你就完整复述了这项试验的核心结果。这就是 meta 分析整理研究数据的最小单位。

1.3 风险比 RR 的定义与计算

知道了 2x2 表的四个数字，怎么算”BCG 到底有没有用”？

最直觉的做法是看两组的**发病率**（也叫” 事件率”），就是每组里发病的人占该组总人数的比例。Aronson 1948 里：

$$\text{打了 BCG 的人发病率} = \frac{a}{a+b} = \frac{4}{123} = 3.25\%,$$

$$\text{没打 BCG 的人发病率} = \frac{c}{c+d} = \frac{11}{139} = 7.91\%.$$

打了 BCG 的发病率明显比没打的低。低多少？把打了的发病率除以没打的发病率：$3.25\% \div 7.91\% = 0.41$。

这个 0.41 就叫**风险比**（risk ratio，简称 **RR**）。它的临床含义是：**打了 BCG 的人发病风险只有没打 BCG 的人的 0.41 倍，相对降低了 59%**。

用一个完全不需要医学背景的例子理解 RR：100 个吃辣的人里 5 个胃疼，100 个不吃辣的人里 1 个胃疼。吃辣的人胃疼” 风险” 是不吃辣的人的 $5 \div 1 = 5$ 倍。RR 就是这种” 风险倍数”。

定义 1.1 (风险比 RR)

对一项试验的 2x2 表（如表 1.1），风险比定义为

$$\mathrm{RR} = \frac{a/(a+b)}{c/(c+d)}.$$

其中 a 是处理组里发生事件的人数，$a+b$ 是处理组总人数，c 是对照组里发生事件的人数，$c+d$ 是对照组总人数。分子 $a/(a+b)$ 是处理组的发病率，分母 $c/(c+d)$ 是对照组的发病率。二者的比值即风险比。♣

RR 的解读规则只有三条：等于 1 表示处理对结局没有影响；小于 1 表示处理有保护作用，越接近 0 保护越强；大于 1 表示处理反而增加风险。Aronson 1948 算出 RR = 0.41，落在保护区，BCG 在这项试验里把发病风险压低到原来的 41%。

1.4 13 项研究的答案一字排开

按上面的方法给每项研究算一个 RR。13 项的结果按试验地点的纬度从低到高排，列在表 1.2。

表 1.2: 13 项 BCG 试验的真实结果与 RR（按纬度从低到高排序）

编号	作者	年份	纬度	BCG 组发病/总数	对照组发病/总数	RR
5	Frimodt-Moller	1973	13°	33 / 5,069	47 / 5,808	0.80
8	TPT Madras	1980	13°	505 / 88,391	499 / 88,391	1.01
11	Comstock et al	1974	18°	186 / 50,634	141 / 27,338	0.71
7	Vandiviere et al	1973	19°	8 / 2,545	10 / 629	0.20
9	Coetzee & Berjak	1968	27°	29 / 7,499	45 / 7,277	0.63
12	Comstock & Webster	1969	33°	5 / 2,498	3 / 2,341	1.56
13	Comstock et al	1976	33°	27 / 16,913	29 / 17,854	0.98
3	Rosenthal et al	1960	42°	3 / 231	11 / 220	0.26
10	Rosenthal et al	1961	42°	17 / 1,716	65 / 1,665	0.25
1	Aronson	1948	44°	4 / 123	11 / 139	0.41
6	Stein & Aronson	1953	44°	180 / 1,541	372 / 1,451	0.46
4	Hart & Sutherland	1977	52°	62 / 13,598	248 / 12,867	0.24
2	Ferguson & Simes	1949	55°	6 / 306	29 / 303	0.20

13 项的 RR 散布从最低 0.20（强保护，发病风险压到原来的 20%）到最高 1.56（反而增加风险）跨度极大。这就是 §1.1 提到的”全球结论矛盾”在数字上的真实表现。

把同一份数据画成图 1.1，每一行是一项研究，红点是 BCG 组发病率，蓝点是对照组发病率，中间用淡灰线连起来代表两组的差距。按纬度从低到高排。

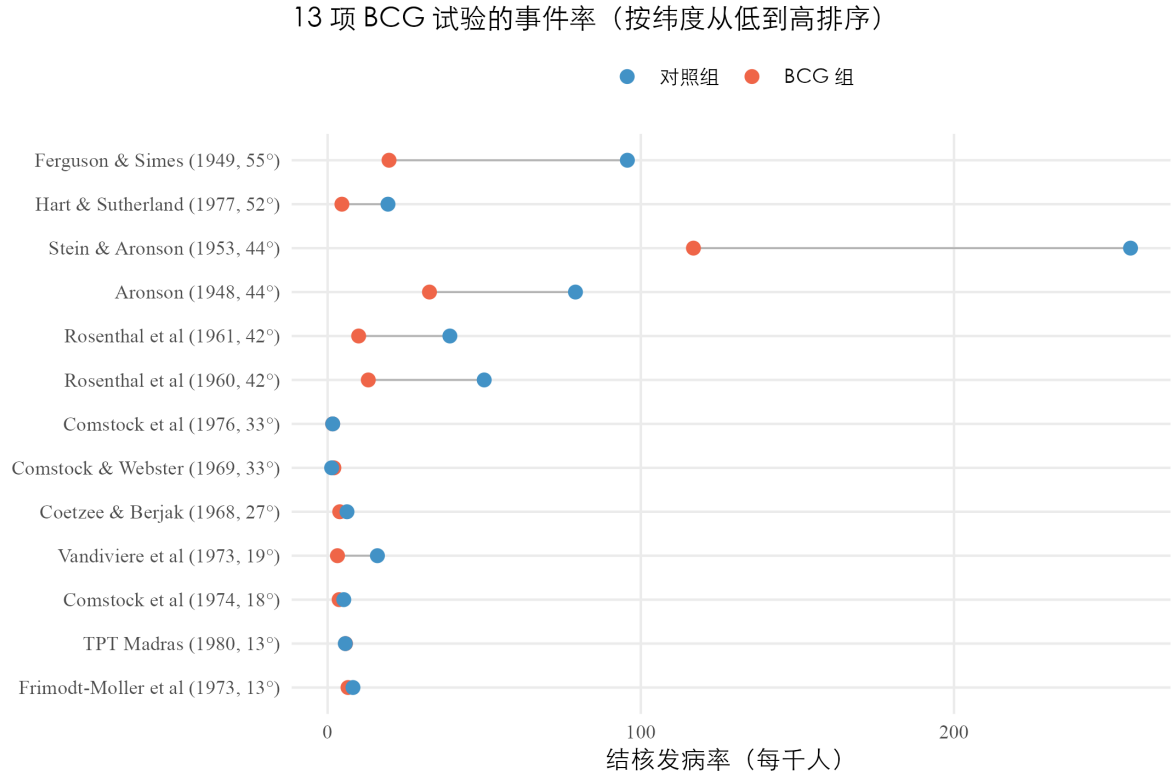


图 1.1: 13 项 BCG 试验在两组上的结核发病率（每千人）。在低纬度试验里，红蓝两点几乎贴在一起；纬度越高，差距越明显。

扫一眼这张表与图，会注意到一个隐藏的规律：表 1.2 按纬度排序后，上半部分（低纬度试验）RR 接近 1 或大于 1，下半部分（高纬度试验）RR 普遍在 0.20–0.50 之间。BCG 在赤道附近几乎没用，在高纬度地区效果显著。这个规律不是巧合，第 7 章会用纬度作为协变量解释 75% 的研究间差异。这里先记下这个伏笔。

1.5 朴素合并：把 13 张 2x2 表直接相加

13 项研究的 RR 跨度太大，没法用任何一项单独下结论。最直觉的" 合并" 方法是：**把 13 张 2x2 表的对应格子加起来**，得到一张涵盖全部 357,347 名受试者的" 超级 2x2 表"，再算一个总的 RR。结果列在表 1.3。

表 1.3: 朴素合并：13 张 2x2 表对应格子相加得到的" 超级 2x2 表"

	发病了	没发病	合计
打了 BCG（13 项汇总）	1,065	189,999	191,064
没打 BCG（13 项汇总）	1,510	164,773	166,283

按 §1.3 的 RR 定义代入这张超级表：

$$\mathrm{RR}_{\text{朴素}} = \frac{1{,}065/191{,}064}{1{,}510/166{,}283} = \frac{0.557\%}{0.908\%} = 0.61.$$

看起来 BCG 平均把全球结核风险降低了 39%，结论相当乐观。这种" 把所有 2x2 表直接相加" 的合并方法叫**朴素合并**（naïve pooling）。

但 0.61 这个数字其实靠不住，问题出在样本量分布上。13 项研究最大的 TPT Madras 一项就有 176,782 人，最小的 Aronson 才 262 人，差出 670 倍。把所有 2x2 表加起来时，TPT Madras 一项就贡献了将近一半的样本。最终的合并 RR 几乎被这一项研究单独决定。

回顾表 1.2：TPT Madras 给出的 RR 是 1.01（几乎无效），但其他 12 项里有 8 项给出了强保护信号（RR < 0.5）。朴素合并把 TPT Madras 一项的" 无效信号" 放到最大权重，结论被它拉向接近无效的方向。其他 12 项研究的强保护信号被淹没了。

停下来想一想。如果一项研究因为样本量极大就单独决定全球答案，那么做 meta 分析就失去了意义——直接报告 TPT Madras 一项就行了，何必整理另外 12 项。要做合理的合并，必须想办法让每项研究" **按精度发言**"，不让一项大研究把其他 12 项的信号全部淹没。这是第 4 章固定效应模型的任务。

1.6 全书章节与方法路线

到这里第 1 章把数据基本面摆清楚了：13 项试验装成 13 张 2x2 表，每张表算一个 RR，13 项 RR 跨度从 0.20 到 1.56，朴素合并被一项大研究主导得到 0.61 但靠不住。后续章节按一条主线把这件事一步步解决。

第 2 章先暂停统计本身，讲 13 项 RCT 是怎么从全球文献库里筛选出来的。这个流程叫 PRISMA，是任何系统综述的开场白。第 3 章回到数字，把 RR 之外的几种常用效应量（OR、风险差、Hedges' g）也讲清楚。第 4 章解决朴素合并的问题，引入 **固定效应模型**，按方差倒数加权让每项研究" 按精度发言"。但 BCG 数据上会出现一个新问题：四项研究就占了 86% 的权重，结论仍然被它们主导。第 5 章承认 13 项研究本来就不一样，引入 **随机效应模型**。第 6 章把" 研究本来就不一样" 的程度量化成 Q、I^2、τ^2 三个数字，叫**异质性诊断**。第 7 章发现纬度能解释 75% 的研究间差异，这是 BCG 数据最重要的方法学发现。第 8 章质疑：13 项研究就是全部吗，还是有阴性结果根本没发表（**发表偏倚**）。第 9 章去掉一项研究后结论是否还成立（**敏感性分析**）。第 10 章把所有方法的结论汇总，给一份临床证据等级（**GRADE**）。

整本书是**一份数据 + 一条主线**：每章解决主线上的一个具体问题，章末把那一步的结论汇入对比表。读完全书，你应当能清楚地说出 BCG 在全球的整体效应是多少、不同地区有多大差异、这份证据值多少。

第 2 章 PRISMA 与系统综述流程

内容提要

- 看清第 1 章那 13 项 RCT 是按什么标准从全球文献库里筛出来的
- 学会 PICO 框架把研究问题写清楚
- 学会双人独立筛选与 Cohen's κ 一致性度量
- 理解 Cochrane RoB 2 五个偏倚域，给每项研究的执行质量打分

第 1 章把 13 项 BCG 试验的数字一字排开，给了我们答案矛盾的真实图景，但跳过了一个关键的前提问题：这 13 项试验是怎么从全球文献库里挑出来的？为什么是 13 项，不是 30 项或 5 项？换一个研究团队复现同样的检索流程，会得到一样的研究列表吗？

这些问题不回答，再精巧的统计合并都建在沙地上。本章讲的不是统计方法，而是 meta 分析”动手算之前”必须走完的检索、筛选、提取、报告流程，行业称为 PRISMA 方案。

1990 年代之前的所谓”综述”，多数是 narrative review。作者根据自己的阅读经验挑选若干”代表性”研究讲故事，挑选标准不公开、过程不可复现、不同作者得出相反结论是常态。1993 年成立的 Cochrane 协作网开始系统性地推动综述流程的标准化，1999 年 QUOROM 声明问世，2009 年扩展为 PRISMA 声明，2020 年 Page 等人发布 PRISMA 2020 更新版 [3]。今天任何想发表 meta 分析的论文，几乎都被要求附上一张 PRISMA 流程图。

2.1 PRISMA 2020 与四阶段流程

PRISMA 的全称是 Preferred Reporting Items for Systematic Reviews and Meta-Analyses，中文常译为系统综述与 meta 分析报告规范。它不是一种统计方法，而是一份报告清单与一张流程图，规定了一份合规系统综述需要交代清楚哪些信息。PRISMA 2020 把整个研究过程划分成四个阶段，每个阶段都有具体的数字记录。

定义 2.1 (PRISMA 2020 四阶段流程)

识别阶段记录从所有数据源检索到的候选记录数量。筛选阶段按标题、摘要排除明显不相关的记录。合格性评估阶段对剩余记录进行全文阅读，按预设的纳入与排除标准做最终判断。纳入阶段是最终进入定性合成与定量合成（meta 分析）的研究集合。每个阶段的数字与排除原因都记录在 PRISMA 流程图里。 ♣

把这套流程画成图就是经典的 PRISMA 流程图，形状像一个倒置的漏斗：上方是大量候选记录，下方是少数最终纳入的研究。本书数据集 *dat.bcg* 是 Colditz 1994（JAMA 271(9):698–702）所选定的 13 项 BCG vs no-BCG 随机对照试验。Colditz 1994 检索了 MEDLINE 1966–1992 数据库并通过文献追溯补充，最终纳入 14 项试验，其中 13 项具备完整 2x2 计数数据进入 metafor 包的 *dat.bcg*。Colditz 1994 发表于 PRISMA 标准建立之前，原文未按 PRISMA 报告每个阶段的具体记录数。

R 包 PRISMA2020 提供了 PRISMA 2020 规范的流程图绘制工具，常用于研究者完成自己的系统综述后生成最终流程图。下面这段代码读入官方 CSV 模板、把 *dat.bcg* 已知的真实事实填入（最终纳入 13 项研究），中间阶段保留为模板占位符，生成图 2.1。读者在做自己的系统综述时，把每个阶段的真实数字依次填入即可。

```
library(PRISMA2020)

# 读入 PRISMA2020 官方 CSV 模板
csv_path  <- system.file("extdata", "PRISMA.csv", package = "PRISMA2020")
prisma_df <- read.csv(csv_path)

fill_n <- function(df, key, val) {
```

```
  df$n[!is.na(df$data) & df$data == key] <- val; df
}

# 中间阶段保留模板占位符，提示读者按自己的检索过程填入
prisma_df <- fill_n(prisma_df, "database_results",          "n = ?")
prisma_df <- fill_n(prisma_df, "duplicates",                "n = ?")
prisma_df <- fill_n(prisma_df, "records_screened",          "n = ?")
prisma_df <- fill_n(prisma_df, "records_excluded",          "n = ?")
prisma_df <- fill_n(prisma_df, "dbr_sought_reports",        "n = ?")
prisma_df <- fill_n(prisma_df, "dbr_assessed",              "n = ?")

# dat.bcg 真实事实：Colditz 1994 最终纳入 13 项 RCT
prisma_df <- fill_n(prisma_df, "new_studies",   "13")
prisma_df <- fill_n(prisma_df, "total_studies", "13")

prisma_obj <- PRISMA_data(prisma_df)
plot <- PRISMA_flowdiagram(prisma_obj,
                           previous = FALSE, other = FALSE,
                           detail_databases = TRUE, fontsize = 10)
PRISMA_save(plot, filename = "figure/chap02_prisma_flow.pdf",
            filetype = "PDF", overwrite = TRUE)
```

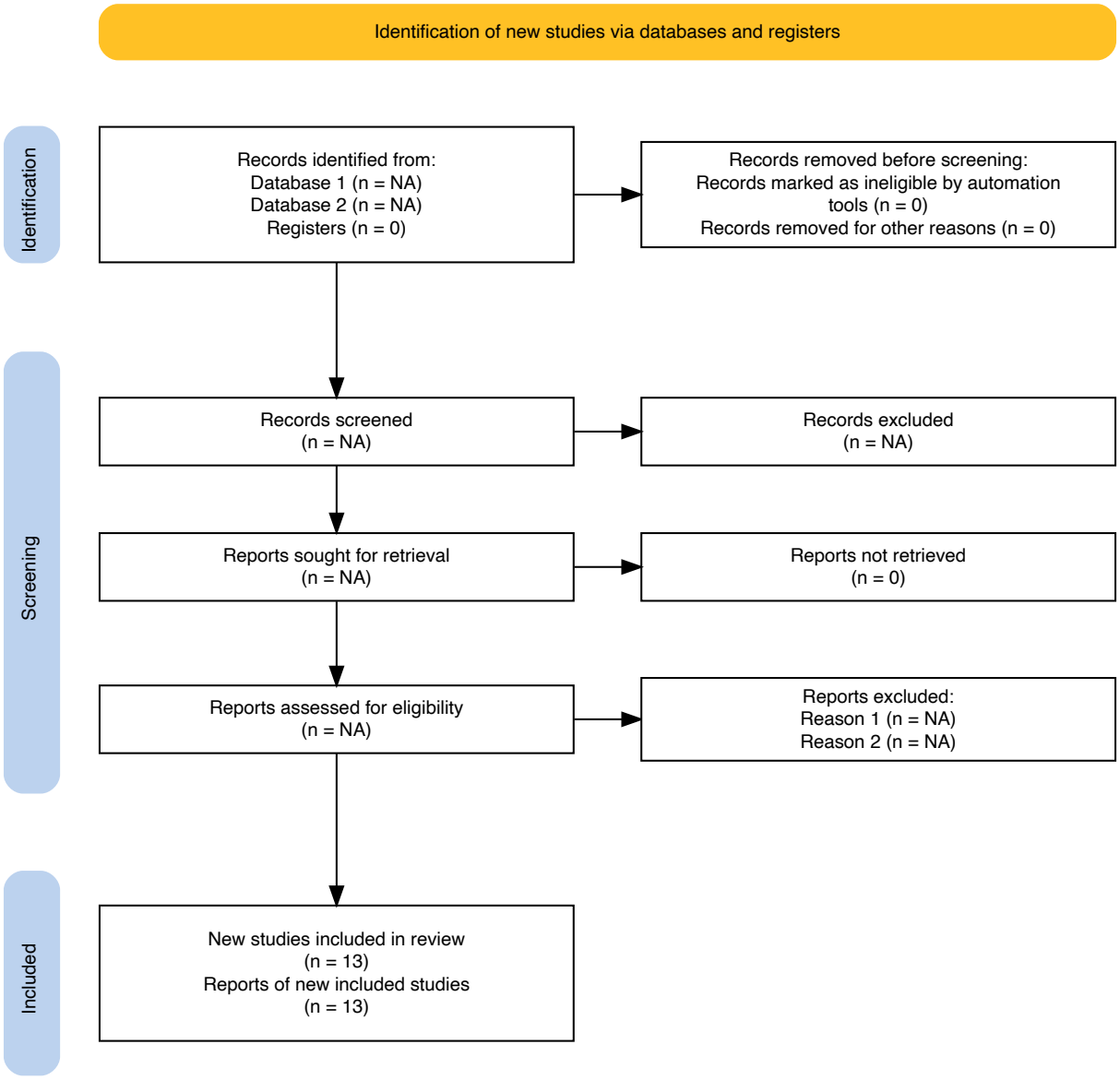


图 2.1: PRISMA 2020 流程图模板。本书 *dat.bcg* 的 13 项最终纳入研究对应底部 `n = 13` 框；中间阶段 `n = ?` 为模板占位符，读者在做自己的系统综述时填入对应数字。本图由 R 包 PRISMA2020 直接生成，符合 Page et al. 2021 [3] 的报告规范。

PRISMA 流程图的价值有两层。第一层是**可复现性**：换一个团队按同样的检索式和筛选标准重新走一遍流程，应该能得到接近的最终研究列表。如果两次结果差异巨大，说明流程的某个环节没记录清楚。第二层是**透明审稿**：审稿人和读者扫一眼流程图就能定位到可疑环节。比如最后纳入只有 3 项研究但识别阶段就有 5000 条记录，意味着排除率超过 99%，纳入标准可能过严，或者检索关键词不准确。

2.2 PICO 框架与检索式构造

要从茫茫文献库里准确找到与你的研究问题相关的论文，得先把研究问题本身写清楚。”BCG 有效吗” 这种宽泛问题没法直接翻译成检索式。临床流行病学界常用的结构化模板叫 PICO 框架，把研究问题拆成四个维度。

定义 2.2 (PICO 框架)

P = Population，目标人群；**I** = Intervention，要评估的干预；**C** = Comparator，对照条件；**O** = Outcome，关心的结局。任何 meta 分析的研究问题都要先用 PICO 写清楚，再翻译成检索式与纳入排除标准。 ♣

PICO 的好处可以用点餐场景来打比方。你跟服务员说"我想吃点东西"，服务员没法办；说"中餐、辣口、人均 50 块、午餐时段"，服务员才知道往哪儿推。PICO 就是把研究问题具体化到能直接送进数据库检索的程度。BCG 这本书的研究问题用 PICO 写出来如下：

表 2.1: BCG meta 分析的 PICO 填充

维度	具体内容
P：目标人群	健康人群，无活动性结核病史，包括婴儿、儿童、青少年与成人
I：干预	接受 BCG 卡介苗接种
C：对照	不接种或接种安慰剂
O：结局	任意形式的结核病发病，包括肺结核与肺外结核

PICO 填好之后，下一步是把每个维度翻译成数据库的检索关键词。一个好的检索式必须覆盖同义词与不同写法，否则会漏掉相关研究。BCG 这个干预在文献里既写作"BCG"，也写作"Bacillus Calmette-Guérin"、"卡介苗"、"卡介苗接种"。结核病既有"tuberculosis"，也有"TB"、"pulmonary tuberculosis"、"Mycobacterium tuberculosis"。把这些同义词用 OR 连起来，再用 AND 把不同维度连起来，得到一个 PubMed 检索式雏形：

```
((BCG OR "Bacillus Calmette-Guerin"[MeSH])
AND (tuberculosis OR "Mycobacterium tuberculosis"[MeSH] OR TB)
AND (random* OR "controlled clinical trial"[Publication Type]))
```

里面的 `[MeSH]` 标记表示 PubMed 的 Medical Subject Headings 主题词检索，比关键词检索更精确，因为 PubMed 已经把同义词归到同一个 MeSH 词下。`random*` 里的星号是通配符，能同时匹配 randomized、randomised、randomization 等多种变形。

回到 BCG 的具体实践。如果只用 PubMed 一个数据库，检索覆盖会偏向英语文献，漏掉非英语研究和未发表的灰色文献。Cochrane Handbook 推荐至少检索三个数据库：PubMed、Embase、Cochrane CENTRAL。重要研究还应做**文献追溯**：把已纳入研究的参考文献列表过一遍，把引用过这些研究的后续论文也找出来。这些步骤决定了图 2.1 顶部"识别阶段 n =?"框需要填入的数字。

定理 2.1 (雷区)

检索式过窄是 meta 分析最容易踩的坑之一。常见症状是：用了精确的方法学限定（比如只搜 RCT），又限制了语言（比如只英语），又只用一个数据库，最终识别阶段就只有几十条记录。这种情况下纳入研究极有可能是有偏的子集，发表偏倚被放大。诊断方法有两条：一是看识别阶段记录数与最终纳入数的比例，比例小于 10:1 通常意味着检索过窄；二是把已知的若干"种子文献"放进检索结果里看是否被检出，没检出就说明关键词缺失。稳健做法是先用宽口径检索保证覆盖，再用筛选阶段做精度控制。 ♡

2.3 筛选流程与 Cohen's κ

识别阶段拿到的几百到几千条候选记录，多数是不相关的。筛选阶段的任务是把不相关的排除掉。直接做法是一个研究者一条条读过去打勾，但单人筛选有两个问题：判断标准会随读者疲劳漂移，主观偏好会悄悄进入决策。Cochrane 推荐**双人独立筛选**：两个研究者背对背各自打勾，对比两份结果，分歧时第三人裁定或共同讨论。

两个人独立打勾之后，第一个要回答的问题是：他们的判断到底有多一致？最朴素的指标是**观察一致率**，记作 P_o，等于两人判断完全相同的记录数除以总记录数。但 P_o 有一个明显的缺陷：即使两人完全瞎打勾，纯靠运气也会有相当比例的一致。

举个具体例子。100 篇候选论文里，两位审阅员 A 和 B 独立判断是否纳入。表 2.2 把他们的结果交叉列出。

表 2.2: 两位审阅员对 100 篇候选论文的纳入判断交叉表

	B 纳入	B 排除	合计
A 纳入	25	5	30
A 排除	3	67	70
合计	28	72	100

观察一致率 $P_o = (25+67)/100 = 0.92$，看起来非常好。但纯靠运气也能拿到不少一致。如果 A 平均纳入率是 30%，B 平均纳入率是 28%，假设两人独立瞎打勾（互不相干），”两人都纳入”的随机概率是 $0.30 \times 0.28 = 0.084$，”两人都排除”的随机概率是 $0.70 \times 0.72 = 0.504$，**偶然期望一致率** $P_e = 0.084 + 0.504 = 0.588$。也就是说，即使两人完全瞎打，单纯由于”纳入率本身偏低”，也会有大约 59% 的记录恰好被两人一致判断。0.92 的观察一致率扣掉这 0.59 的偶然部分，剩下的才是真一致。

定义 2.3 (Cohen’s κ)

Cohen’s κ 是把偶然一致率扣除之后的一致性度量：

$$\kappa = \frac{P_o - P_e}{1 - P_e}.$$

其中 P_o 是两位审阅员判断完全相同的记录数占总记录的比例（观察一致率），P_e 是按两位审阅员各自纳入率独立计算的偶然期望一致率。$\kappa = 1$ 表示完全一致，$\kappa = 0$ 表示一致水平等同于纯随机，$\kappa < 0$ 表示比随机还糟。

把表 2.2 的数据代进去：$\kappa = (0.92 - 0.588)/(1 - 0.588) = 0.332/0.412 = 0.806$。0.806 在 Landis & Koch 1977 的常用解读里属于”几乎完美一致”。

κ 与 P_o 在不同边际分布下会拉开。把同一个 $P_o = 0.92$ 配上不同的纳入率：两人都打 70%（$a = 66, b = 4, c = 4, d = 26$），$P_e = 0.70^2 + 0.30^2 = 0.58$，$\kappa = (0.92 - 0.58)/(1 - 0.58) = 0.81$，与 0.806 接近。换一种极端：两人都很挑剔，纳入率都只有 5%（$a = 1, b = 4, c = 4, d = 91$），$P_e = 0.05^2 + 0.95^2 = 0.905$，$\kappa = (0.92 - 0.905)/(1 - 0.905) = 0.158$，跌到”差”一档。同样的 $P_o = 0.92$，因为绝大多数文献本来两人都不会纳入，”瞎打”也能拿到 0.905 的偶然一致，扣除偶然部分后 κ 几乎为零。这就是 Feinstein & Cicchetti 1990 命名的”kappa 悖论”：高观察一致不必然意味着高真实一致。

Landis-Koch 的常用区间标准（简化版）：$\kappa < 0.40$ 一致性差，0.40–0.60 中等，0.60–0.80 良好，> 0.80 几乎完美（原文进一步把 < 0.40 细分为 slight 与 fair 两档，本书简化为一档）。系统综述的筛选阶段一般要求 $\kappa > 0.60$，否则说明纳入排除标准本身太模糊，需要重新讨论。$\kappa > 0.80$ 是大多数 Cochrane 综述报告的目标值。

2.4 数据提取与 Cochrane RoB 2

筛选完成、最终纳入的研究列表确定下来之后，下一步是从每篇研究里提取需要的数字。对一份 meta 分析来说，至少要提取每项研究的样本量、处理组与对照组的事件数或均值方差、随机化方式、随访时间、关键基线协变量。提取这一步同样要求双人独立操作，理由与筛选阶段相同。

数据提取之外，每项研究还要做**风险偏倚评估**。同样是 RCT，执行质量可以差很远：随机化序列是用电脑生成还是用奇偶日期决定？盲法是双盲、单盲还是开放？失访比例多少？预设的主要结局有没有变？Cochrane 在

2019 年发布了第二版风险偏倚工具，英文称 Risk of Bias 2，简称 RoB 2，把风险偏倚拆成五个域来分别评估。

定义 2.4 (Cochrane RoB 2 五个偏倚域)

RoB 2 把试验的偏倚风险拆成五个独立域：随机化过程偏倚、对预期干预的偏离偏倚、缺失结局数据偏倚、结局测量偏倚、选择性结果报告偏倚。每个域评级为低风险（low）、有些担忧（some concerns）、高风险（high），最后综合成总体偏倚风险 [4]。

♣

每个域查的是什么。随机化过程域查的是序列生成与分配隐藏：电脑生成的伪随机序列加密码信封是低风险，按入院日期奇偶分组就是高风险，因为研究者能预测下一个进来的病人去哪一组。对预期干预的偏离域查的是参与者与研究人员是否盲法，盲法不全的研究里研究者可能为治疗组病人额外多做几次检查，这种" 小心呵护" 会污染结果。缺失结局数据域查的是失访比例与失访原因，失访超过 20% 且原因与处理相关时结果可信度就要降级。结局测量域查的是测量人是否盲法、测量方法是否一致。选择性报告域查的是研究是否预先注册了主要结局，发表的论文是否报告了所有预先承诺的结局。

回到 BCG 13 项 RCT。从第 1 章表 1.2 可以看到，13 项研究中 8 项（包含 `random` 与少量 `alternate` 实际为 random 的）属于标准随机化路径，2 项是 `alternate`（按入院顺序交替分配），3 项是 `systematic`（按某种系统规则分配，如出生日期奇偶）。后两类按今天的标准都属于" 准随机"，分配方式可被预测，按 RoB 2 是有些担忧或高风险。这意味着这 13 项研究的偏倚风险并不均匀，第 5 章随机效应模型与第 9 章敏感性分析都会回到这个事实。

定理 2.2 (雷区)

RoB 2 评估最常见的误用是把它当成填表打勾的形式主义，每个域随手填个"low" 或"some concerns" 了事。这种做法让审稿人看到 RoB 表打的全是绿勾，但实际偏倚风险并未受控。正确的做法是每个域写一两句" 为什么是这个评级" 的依据：随机化过程域引用论文里写的具体方法描述，缺失数据域报告失访比例的具体数字，结局测量域说明测量者是否盲法。RoB 2 表的每一个评级都应该是有依据的判断，不是直觉打分。

♡

第 2 章到这里收束。整理 13 项研究的检索、筛选、提取、偏倚评估流程本身不产生新的合并估计，对比表暂不更新。从第 4 章固定效应模型开始，每章会在对比表里加一行。下一章先回到统计本身，把第 1 章只讲过 RR 之外的几种常用效应量（OR、RD、log RR 与方差）讲清楚。

第 3 章 效应量及其方差的计算

内容提要

- 学会把 RR 取对数变成 log RR，看清”对称合并”的需要
- 掌握 log RR 的方差公式与”事件数决定方差”的事实
- 理解比值比 OR 与风险差 RD 各自适合什么场景
- 学会用 `escalc()` 一站式算出每项研究的效应量与方差

第 1 章给每项研究算了一个 RR，把 13 项的结果摆在桌面上。第 2 章讲了 13 项试验是怎么被筛进 meta 分析的。从本章起进入”动手合并”前的最后准备：每项研究不仅要有一个效应量数字，还要带上一个不确定性数字（方差）。第 4 章固定效应模型按方差倒数加权时，需要的就是这两件东西。

本章把”动手合并”前需要的术语补齐：log RR 与方差（第 4 章会直接用），OR 与 RD 作为 RR 的两个亲戚（不同研究有时报告不同效应量）。

3.1 对数风险比 log RR

第 1 章的 RR 用起来很顺，但有一个数学上的别扭。”风险降一半”对应 RR $= 0.5$，”风险翻倍”对应 RR $= 2$。直觉上这两个效应”大小相等、方向相反”，可在原始数轴上 0.5 离 1 有 0.5 的距离，2 离 1 有 1 的距离，远远不对称。如果你把若干研究的 RR 排在数轴上，”保护”和”有害”两边的视觉重量是不一样的。

解决办法是给 RR 取自然对数。原数轴上 0.5 和 2 不对称，取对数后变成 $\ln(0.5) = -0.69$ 和 $\ln(2) = +0.69$，正负完全对称。原数轴上 0.25 和 4 不对称，取对数后变成 -1.39 和 $+1.39$，也对称起来。”保护一半”和”风险翻倍”在数轴上就是镜像。

对数尺度还有第二个好处。统计学里有一条经验：一个量是若干小因素相乘得到的（如风险比是事件率的比值），它在原始尺度上分布偏斜，取对数之后近似对称、近似正态。后续合并、计算置信区间都依赖正态近似，必须在对数尺度上做才合法。

定义 3.1 (对数风险比 log RR)

对风险比取自然对数：

$$\log \mathrm{RR} = \ln(\mathrm{RR}).$$

其中 RR 是处理组发病率与对照组发病率之比（见第 1 章定义）。$\log \mathrm{RR} = 0$ 对应 $\mathrm{RR} = 1$（无效），< 0 对应保护作用，> 0 对应风险增加。Meta 分析在 log RR 这个尺度上做合并，最后再把结果取指数变回 RR 报告。

♣

回到 Aronson 1948。第 1 章算出 $\mathrm{RR} = 0.41$，取对数得到 $\log \mathrm{RR} = \ln(0.41) = -0.89$。第 4 章要合并 13 项 log RR 而不是 13 项 RR。

3.2 log RR 的方差

每项研究给出的 log RR 都是一个估计，估计就有不确定性。这种不确定性怎么衡量？

用一个完全不需要医学背景的例子建立直觉：抛一枚均匀硬币，正面概率是 0.5。抛 10 次，你看到 6 次正面（比例 0.6）不算意外；甚至看到 4 次或 7 次都很常见。再抛 1000 次，正面比例几乎一定在 0.48 到 0.52 之间，你看到 0.6 就会怀疑硬币不均匀。同样是估计”正面概率 $= 0.5$”这个事实，10 次的估计”晃”得很厉害，1000 次

的估计稳得多。

方差是这种"晃动幅度"的数学度量。方差越小说明估计越稳，方差越大说明估计越晃。对一个估计的 log RR 来说，统计推导给出的方差近似公式是 2x2 表四个格子的部分倒数和。

定义 3.2 (log RR 的方差近似)

对每项研究的 log RR 估计 $\hat{y}_i$，方差近似为

$$v_i^{\mathrm{RR}} = \frac{1}{a_i} - \frac{1}{a_i + b_i} + \frac{1}{c_i} - \frac{1}{c_i + d_i}.$$

其中 a_i, b_i, c_i, d_i 是第 i 项研究 2x2 表的四个格子（见第 1 章表 1.1）。事件数 a_i 或 c_i 接近零时方差会被放大，这是稀疏事件下 RR meta 分析的主要限制。

把 Aronson 1948 代入：$a = 4, b = 119, c = 11, d = 128$，

$$v = \frac{1}{4} - \frac{1}{123} + \frac{1}{11} - \frac{1}{139} = 0.250 - 0.008 + 0.091 - 0.007 = 0.326.$$

四个分项保留三位小数累加得 0.326，与 metafor 用 `escalc()` 跑出的高精度结果 0.3256 一致。

停下来想一想。决定方差大小的不是总样本量，是事件数。一项研究招了 10 万人但 BCG 组只发生 1 例结核、对照组也只 2 例，方差仍然很大，因为四个分项里有两个分母极小。这是临床流行病学常踩的坑：动辄几十万人的大样本研究在罕见结局上仍可能统计精度很差。

3.3 比值比 OR 与对数比值比

二分类结局除了 RR，还有一种常用效应量叫**比值比**（odds ratio，简称 OR）。理解 OR 之前先理解**比值**（odds）本身。

日常语言里"概率"和"比值"经常混用，但它们是两个不同的量。100 个人去做体检，5 个人查出胃部不适。"概率"是 $5/100 = 5\%$——查出不适的人占总人数的比例。"比值"是 $5/95$——查出不适的人对没查出的人。比值没有上限，因为分母可以任意小。事件率低时，5/95 与 5/100 非常接近；事件率高时，比如 50 个查出不适，比值 $50/50 = 1$，概率 50%，二者拉开差距。

OR 就是两组比值的比。继续吃辣的例子：100 个吃辣的人 5 个胃疼（吃辣组的 odds = 5/95），100 个不吃辣的人 1 个胃疼（不吃辣组的 odds = 1/99）。$\mathrm{OR} = (5/95)/(1/99) \approx 5.21$。回想 RR：$\mathrm{RR} = (5/100)/(1/100) = 5$。OR 比 RR 略大一点，原因是事件率不为零，吃辣组的分母从 100 缩到 95（少了 5），不吃辣组的分母从 100 缩到 99（只少 1），分母收缩不对称把 OR 推得比 RR 更远离 1。

定义 3.3 (比值比 OR 与对数比值比)

对一项试验的 2x2 表，OR 是两组比值的比：

$$\mathrm{OR} = \frac{a/b}{c/d} = \frac{a \cdot d}{b \cdot c}.$$

其中 a, b, c, d 是 2x2 表的四个格子（同表 1.1）。对数比值比 $\log \mathrm{OR} = \ln(\mathrm{OR})$ 在大样本下近似正态，方差近似为

$$v_i^{\mathrm{OR}} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.$$

注意 OR 的方差是四个格子倒数的完全和，与 log RR 方差的"部分和"形式略有不同。

OR 与 RR 在罕见事件下接近。事件率低时 $p/(1-p) \approx p$，OR 与 RR 数值接近。BCG 数据里大多数研究的结核发病率都在百分之几以下，所以 OR 与 RR 应该非常接近。在 dat.bcg 上跑一遍验证：

```
library(metafor)
d <- read.csv("data/bcg.csv")

```

```r
# 一次性算 RR 和 OR
es_rr <- escalc(measure = "RR", ai = tpos, bi = tneg,
                ci = cpos, di = cneg, data = d)
es_or <- escalc(measure = "OR", ai = tpos, bi = tneg,
                ci = cpos, di = cneg, data = d)

# 取指数变回原尺度，对比 RR 和 OR
data.frame(
  trial    = d$trial,
  author   = d$author,
  RR       = round(exp(es_rr$yi), 3),
  OR       = round(exp(es_or$yi), 3),
  diff_pct = round((exp(es_or$yi) - exp(es_rr$yi)) /
                   exp(es_rr$yi) * 100, 1)
)
```

OR 与 RR 的偏差

13 项研究里多数 OR 与 RR 的偏差在 5% 以内。偏差最大的两项是 Stein & Aronson 1953（OR $=0.384$，RR $=0.456$，OR 比 RR 小约 16%）和 Ferguson & Simes 1949（OR $=0.189$，RR $=0.205$，OR 比 RR 小约 8%）。这两项研究事件率较高：Stein & Aronson 1953 对照组结核发病率约 25.6%（372/1451），处理组也有 11.7%（180/1541），事件率离"罕见"已经很远，OR 比 RR 偏离零的距离更明显。其余 11 项研究事件率都在 1%–6%，OR 与 RR 几乎重合。

定理 3.1 (雷区)

把 OR 直接当 RR 解读是临床论文最常见的偏差来源之一。读者看到 OR $=0.5$ 常理解成"风险减半"，但 OR $=0.5$ 只在事件率很低时才约等于 RR $=0.5$；事件率接近 50% 时 OR 比 RR 偏离零更远，OR $=0.5$ 实际上对应的 RR 可能是 0.7 或更大。识别这个雷区有两条诊断方法：看研究的事件率，事件率超过 10% 就要警惕；看作者用了 RR 还是 OR，混用术语的论文要逐句核对。

3.4 风险差 RD

RR 与 OR 都是**相对**效应量，回答"风险是几倍"的问题。临床决策很多时候需要的是**绝对**效应量，回答"打疫苗能少多少人发病"的问题。这就是 **风险差**（risk difference，简称 RD）。

继续吃辣的例子：100 个吃辣的人 5 个胃疼，100 个不吃辣的人 1 个胃疼。RD $=5/100-1/100=4$ 个百分点。意思是吃辣比不吃辣绝对增加了 4% 的胃疼风险。如果一个医院每天 1000 个人吃辣，每天就多出 40 个胃疼患者。RD 的好处是它能直接翻译成"打多少人能预防一例发病"，这个倒数叫做 **NNT**（number needed to treat），临床决策最关心的指标。

定义 3.4 (风险差 RD)

RD 是两组事件率的差：

$$\mathrm{RD}=\frac{a}{a+b}-\frac{c}{c+d}=p_t-p_c.$$

其中 $p_t=a/(a+b)$ 是处理组发病率，$p_c=c/(c+d)$ 是对照组发病率。方差近似为

$$v_i^{\mathrm{RD}}=\frac{p_t(1-p_t)}{n_t}+\frac{p_c(1-p_c)}{n_c},$$

其中 $n_t = a + b$ 是处理组总人数，$n_c = c + d$ 是对照组总人数。RD 的取值范围是 $[-1, 1]$，单位是事件率本身的单位（比例或百分点）。

RD 与 RR / OR 有一个根本不同：RD 不需要取对数。RR 与 OR 是比值，原尺度上保护和有害不对称，必须取对数才能让正负效应对称合并。RD 是差值，原尺度上 -0.05 与 $+0.05$ 已经天然对称，直接合并即可。

把 Aronson 1948 代入 RD 公式：$p_t = 4/123 = 0.0325$，$p_c = 11/139 = 0.0791$，RD $= 0.0325 - 0.0791 = -0.0466$。意思是接种 BCG 比不接种绝对降低了约 4.66 个百分点的发病风险。

回到 BCG 数据。13 项研究的 RD 跨度从 -0.140（Stein & Aronson 1953）到 $+0.0007$（Comstock & Webster 1969）。Stein & Aronson 的 RD 含义是：每 100 人接种 BCG，比不接种少 14 人发生结核——这是绝对意义上巨大的预防效应。Hart & Sutherland 1977 的 RD 是 -0.0147（每 100 人少 1.47 人发生结核），RR $= 0.24$ 看起来效果很大，但绝对差小得多，因为这项研究对照组基线发病率本来就只有 1.93%。这说明同一项研究 RR 看起来效果很大、RD 看起来效果很小，是**基线风险**低的直接结果。

定理 3.2 (雷区)

RD 在跨研究合并上有一个被低估的麻烦。不同研究的基线发病率差距巨大时，RD 在合并后的解读会很尴尬：合并 RD 等于"加权平均的绝对效应"，但这个平均对应不到任何具体的临床场景。BCG 数据里基线发病率从 0.49‰ 到 25.6% 跨度极大，合并 RD 给出的数字是几个百分比的混合，难以直接用于临床决策。Cochrane Handbook 推荐：跨研究异质性大时优先用 RR 或 OR 做主分析，RD 仅作为辅助报告，并按基线风险分层呈现。

3.5 escalc 函数：一站式计算

前面演示了 `escalc()` 计算 RR 和 OR。它支持的 measure 选项覆盖了 meta 分析常用的所有效应量。

表 3.1: `escalc()` 常用 measure 参数与对应效应量

measure	效应量	适用场景
`"RR"`	log 风险比	二分类结局，RCT 默认选择
`"OR"`	log 比值比	二分类结局，病例对照或 logistic 回归输出
`"RD"`	风险差	二分类结局，临床绝对效应
`"PETO"`	Peto OR	罕见事件（事件率通常 $< 5\%$ 且组间事件数接近时）
`"SMD"`	Hedges' g	连续型结局，跨量表合并
`"MD"`	均值差	连续型结局，量表一致
`"COR"`	Pearson r	相关系数 meta 分析
`"ZCOR"`	Fisher z	相关系数（对数尺度合并）

`escalc()` 返回一个数据框，包含 `yi`（效应量估计）和 `vi`（方差估计）两列。这两列是后续所有 meta 分析函数（`rma()`、`rma.mv()`）的标准输入。本书 BCG 数据是二分类结局，所有合并都用 `measure = "RR"`。

第 3 章到这里收束。每项研究的 $\hat{y}_i$ 与 v_i 是合并的输入。第 4 章用这两个数字按"方差越小、发言权越大"的逆方差加权，把 13 项研究合并成一个全局估计 RR $= 0.65$。但 13 项研究的方差极不平衡，最大的 TPT Madras 一项就占 41% 权重，前 4 项合计占 86%，结论仍然被少数大研究主导，异质性指标 $I^2 = 92\%$ 也直接证伪 FE 模型的核心假设。这是第 4 章的主要发现，也是引出第 5 章随机效应模型的起点。

第 4 章 固定效应模型与逆方差加权

内容提要

- 理解固定效应模型的核心假设：所有研究估同一个真实效应
- 看清逆方差加权的设计动机：让”按精度发言”取代”按样本量发言”
- 学会用 `rma()` 与 `rma.mh()` 拟合两种固定效应估计
- 在 dat.bcg 上读懂 forest plot 与异质性 $I^2 = 92\%$ 的红灯信号

第 1 章朴素合并把 13 张 2x2 表叠加得到 RR $= 0.61$，但被一项 17.6 万人的 TPT Madras 单独主导，结论被它拉向接近无效的方向。本章引入 meta 分析的第一种正式合并方法——**固定效应模型**（fixed-effects 或 equal-effects），让每项研究”按精度发言”而不是”按样本量发言”。

4.1 为什么不能按样本量加权

回顾朴素合并失败的原因：把 2x2 表对应格子相加，相当于把每项研究按”样本量”加权。TPT Madras 一项 17.6 万人，权重接近一半，直接把合并结果决定了。

最自然的修补思路是：**别按样本量加权，按精度加权**。精度高的研究权重大，精度低的研究权重小。第 3 章已经讲过，每项研究的 log RR 估计有方差 v_i，方差越小代表估计越精确。所以”按精度发言”在数学上就是**权重** $= 1/v_i$，叫做逆方差权重。

用一个完全不需要医学背景的例子建立直觉。班里讨论”今天食堂的菜咸不咸”，吃了三大碗的同学（精度高、方差小）说话有分量，只闻了一下饭味的同学（精度低、方差大）声音被淡化。逆方差权重做的就是这件事：让样本量大、事件多的研究主导合并结果，但对真正”精度差”的研究压低声音，不像朴素合并那样让单一巨研究决定一切。

对照一下 BCG 数据。TPT Madras 的方差 $v_8 = 0.0040$，权重 $1/0.0040 = 250$；Aronson 1948 的方差 $v_1 = 0.3256$，权重 $1/0.3256 = 3.07$。两项研究权重比 81:1。Hart & Sutherland 1977 的 $v_4 = 0.0200$，权重 50。三项的相对发言权 250 : 50 : 3.07，比”按总样本量发言”176 : 26.5 : 0.262 的 670:1 比例温和不少。

4.2 固定效应模型的假设

逆方差加权法的合理性建立在一条统计假设上：13 项 BCG 试验估的是**同一个真实效应**，研究之间的差异完全来自抽样波动。

定义 4.1 (固定效应假设)

设 k 项研究各自给出对真实效应 θ 的估计 $\hat{y}_i$，方差 v_i。固定效应模型假设

$$\hat{y}_i = \theta + \varepsilon_i, \qquad \varepsilon_i \sim N(0, v_i).$$

其中 $\hat{y}_i$ 是第 i 项研究的对数风险比估计，θ 是全部研究共享的真实效应（所有 θ_i 都等于这个 θ），ε_i 是抽样误差，服从均值零、方差 v_i 的正态分布。v_i 是第 i 项研究的方差（第 3 章给出过算法）。

这条假设在临床流行病学里很少真正成立。Hart & Sutherland 1977 的英国受试者和 TPT Madras 1980 的印度南部受试者，年龄构成、营养状况、当地结核流行株系、环境暴露都可能不同，BCG 在两群人身上的”真实效应”很难被论证为完全相同。作为统计基线，FE 模型在它的假设下有解析最优解（最小方差线性无偏估计），是理解后续随机效应模型的起点。

4.3 逆方差加权法

定义 4.2 (逆方差权重与合并估计)

在 FE 假设下，加权平均

$$\hat{\theta} = \frac{\sum_i w_i \hat{y}_i}{\sum_i w_i}$$

对真实效应 θ 是无偏估计（任意正权重 w_i 都能使均值无偏）。要在无偏估计中达到方差最小，权重必须取

$$w_i = \frac{1}{v_i}.$$

其中 $\hat{y}_i$ 是第 i 项研究的 log RR 估计，v_i 是其方差，w_i 是该研究在合并里的权重。合并估计的方差为 $\mathrm{Var}(\hat{\theta}) = 1/\sum_i w_i$，对应 95% 置信区间为 $\hat{\theta} \pm 1.96\sqrt{1/\sum_i w_i}$。

♣

为什么是逆方差而不是别的权重形式？方差小的研究噪音少，把它的估计”压低声音”等于浪费信息；方差大的研究噪音多，把它的估计”放大声音”等于让噪音主导。两项研究 $v = 0.04$ 与 $v = 0.16$，逆方差权重 $1/v$ 给出 25 与 6.25，方差差 4 倍权重就差 4 倍。这种”按精度的反比”分配恰好让总方差最小。

下面用 metafor 的 `rma()` 函数在 BCG 数据上拟合固定效应模型。`rma()` 是 metafor 的核心函数，参数 `method = "EE"` 表示 equal-effects（与传统的 fixed-effects 同义）。

```
library(metafor)
d  <- read.csv("data/bcg.csv")
es <- escalc(measure = "RR",
             ai = tpos, bi = tneg, ci = cpos, di = cneg,
             data = d)

fe_iv <- rma(yi, vi, data = es, method = "EE")
print(fe_iv)
```

FE 模型输出

IV 加权固定效应模型给出合并 $\widehat{\log \mathrm{RR}} = -0.4303$，标准误 0.0405，95% CI $[-0.5097, -0.3509]$。换算到 RR 尺度，合并 RR $= 0.65$，95% CI $[0.60, 0.70]$。从这个数字看，BCG 整体把结核风险降低了约 35%。

异质性检验给出 $Q = 152.23$，自由度 $k - 1 = 12$，$p < 0.0001$，$I^2 = 92.13\%$。I^2 把研究间观察变异的超过九成归因于真实的效应异质性，只有不到一成可以用抽样波动解释。这个结果与 FE 模型的核心假设直接冲突：13 项研究几乎可以肯定不是在估同一个真实效应。I^2 与 τ^2 的正式定义在第 6 章展开。这里先记住一个红灯信号：I^2 超过 50%，FE 模型给出的 CI 几乎一定过窄，结论应当保留质疑。

4.3.1 每项研究的权重分布

逆方差权重让方差小的研究主导结果。表 4.1 列出 dat.bcg 13 项研究在 IV 模型下的权重，按权重从大到小排序。

表 4.1: 13 项研究在固定效应（IV）模型下的权重

编号	作者	年份	方差 v_i	权重 (%)	累计 (%)
8	TPT Madras	1980	0.0040	41.40	41.40
6	Stein & Aronson	1953	0.0069	23.75	65.15
11	Comstock et al	1974	0.0124	13.21	78.36
4	Hart & Sutherland	1977	0.0200	8.20	86.56
5	Frimodt-Moller et al	1973	0.0512	3.20	89.76
9	Coetzee & Berjak	1968	0.0564	2.91	92.67
13	Comstock et al	1976	0.0714	2.30	94.97
10	Rosenthal et al	1961	0.0730	2.25	97.22
2	Ferguson & Simes	1949	0.1946	0.84	98.06
7	Vandiviere et al	1973	0.2230	0.74	98.80
1	Aronson	1948	0.3256	0.50	99.30
3	Rosenthal et al	1960	0.4154	0.39	99.69
12	Comstock & Webster	1969	0.5325	0.31	100.00

停下来想一想。前 4 项研究合计权重 86.56%，剩余 9 项研究合计权重 13.44%。FE 合并结果几乎是这四项研究的加权平均。回想第 1 章每项研究的 log RR：TPT Madras 是 0.012（几乎无效），Stein & Aronson 是 -0.786，Comstock 1974 是 -0.339，Hart & Sutherland 是 -1.442。最大权重的 TPT Madras 把合并拉回到接近零，最强保护的 Hart & Sutherland 因为权重小（8.2%）影响有限。这就是为什么 FE 合并 RR $= 0.65$，看起来比" 按试验平均" 温和很多。

朴素合并把 TPT Madras 的" 无效信号" 放到接近 50% 的样本量份额；FE-IV 把它压到 41% 的精度份额。情形改善但仍不理想——实质问题在 $I^2 = 92\%$ 表明的" 研究本来就不一样"，第 5 章随机效应模型才能正面解决。

4.4 Mantel-Haenszel 加权

逆方差加权在事件数充足时表现良好，但当某些研究的事件数很少（比如对照组只发生 3 例事件），$1/v_i$ 会因为方差爆炸而权重过小，甚至 v_i 因为零格而无法计算。Mantel & Haenszel 1959 提出了一种不同思路的加权方案，直接基于 2x2 表的格子计数构造合并估计，避开了 $1/v_i$ 的不稳定性。

定义 4.3 (Mantel-Haenszel 合并 RR)

MH 合并 RR 不写成 $\sum w_i \hat{y}_i / \sum w_i$ 的标准形式，而是直接在 2x2 计数上构造比值：

$$\widehat{\mathrm{RR}}_{\mathrm{MH}} = \frac{\sum_i a_i (c_i + d_i)/n_i}{\sum_i c_i (a_i + b_i)/n_i}.$$

其中 a_i, b_i, c_i, d_i 是第 i 项研究 2x2 表的四个格子，$n_i = a_i + b_i + c_i + d_i$ 是该项研究的总样本量。分子可以读成" 在 RR $= 1$ 的零假设下处理组应该看到的事件数" 按各研究累加，分母是对照组应该看到的事件数累加；二者的比值度量了实际观测相对于零假设的偏离。 ♣

MH 估计量在稀疏事件下偏倚比 IV 小，是 Cochrane Collaboration 推荐的二分类 meta 分析默认方法。在 BCG 数据上跑 MH 加权：

```
fe_mh <- rma.mh(measure = "RR",
                ai = tpos, bi = tneg, ci = cpos, di = cneg,
                data = d)
print(fe_mh)
```

IV 与 MH 的对比

MH 合并给出 $\widehat{\log \mathrm{RR}}_{\mathrm{MH}} = -0.4537$，$\mathrm{RR}_{\mathrm{MH}} = 0.635$，95% CI $[0.588,\ 0.686]$。与 IV 模型的 $\mathrm{RR}_{\mathrm{IV}} = 0.65$ 相比，MH 估计稍稍偏低（保护效应略大），CI 几乎重合。两种方法在 BCG 数据上结论高度一致，原因是 BCG 的事件数普遍充足（最少的 Comstock & Webster 1969 也有 8 例事件），没有触发 IV 加权的稳定性问题。MH 估计的 $I^2 \approx 92\%$，与 IV 几乎一致。

定理 4.1 (雷区)

固定效应模型最常见的误用是在异质性显著时仍坚持用 **FE** 报告主结果。IV 加权对方差最小的研究给出超大权重，FE 假设又强行假设所有研究估同一个数。当 $I^2 > 50\%$、Q 检验 $p < 0.05$ 时，研究之间存在真实效应异质性，FE 合并的 CI 会显著偏窄，结论看起来比真实证据更确定。诊断方法只有一条：报 FE 之前先看 I^2 与 Q，$I^2 > 50\%$ 或 Q 检验 $p < 0.05$ 就改报随机效应。在 dat.bcg 上 $I^2 = 92\%$，FE 模型只能作为对比基线，主结果应当用第 5 章的随机效应模型给出。 ♡

4.5 Forest plot

Forest plot 是 meta 分析最重要的可视化工具，把每项研究的效应估计与合并结果画在同一张图上。读这张图能直接看到三件事：每项研究的效应方向与精度（横线长短）、最大权重的研究是哪几项（方块大小）、研究间散布是否远超单条横线（异质性的视觉证据）。

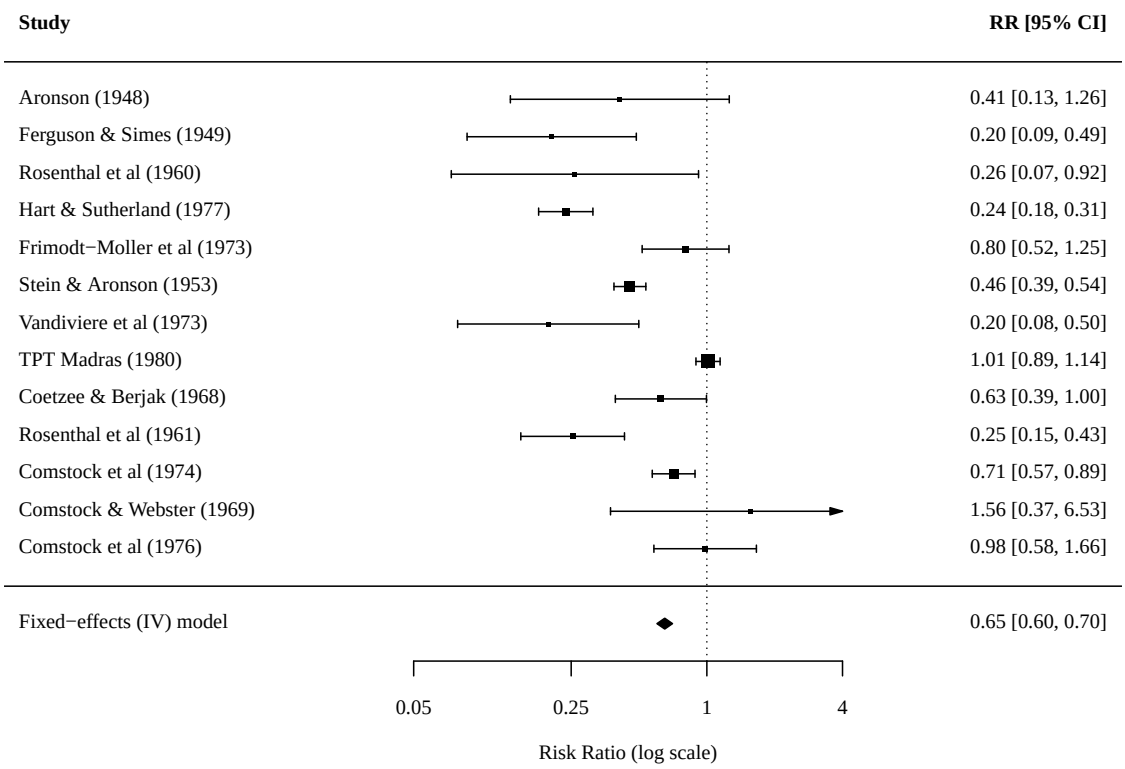


图 4.1: 13 项 BCG 试验固定效应（IV）模型 forest plot。每项研究方块大小与权重成正比，横线为 95% CI；底部菱形为合并 RR 估计 0.65（95% CI 0.60–0.70）。最大方块的 TPT Madras 1980 落在零线右侧（RR ≈ 1），最强保护的 Hart & Sutherland 1977 方块小、影响有限。13 项研究的横线散布范围远超过任何单条横线的宽度，是异质性显著的视觉证据。

4.6 累积对比表（首次出现）

从本章起每章在对比表里加一行，到第 10 章合并为终极版。表的设计是把每种方法的合并估计放在同一份模板里，便于横向比较。

表 4.2: 方法对比表（截至第 4 章）

方法	合并 logRR	合并 RR (95% CI)	核心假设	局限
朴素合并	-0.488	0.61 (—)	13 个 2x2 表可直接相加	被大样本研究主导，洗掉异质性
FE-IV	-0.430	0.65 (0.60, 0.70)	13 项研究估同一个真实效应	$I^2 = 92\%$，假设明显违反
FE-MH	-0.454	0.64 (0.59, 0.69)	同 FE-IV，对稀疏事件更稳健	异质性显著时 CI 仍偏窄

两种 FE 估计相互验证，但 $I^2 = 92\%$ 表明 FE 假设在这份数据上不成立。第 5 章随机效应模型会松开” 同一个真实效应” 这条假设，给出一个更合理的合并估计。

第 5 章 随机效应模型与 τ^2 估计

内容提要

- 理解从 FE 到 RE 的核心松绑：每项研究的真实效应不再相同
- 掌握 DerSimonian-Laird 与 REML 两种 τ^2 估计方法
- 学会 Hartung-Knapp 调整在小研究数下的纠偏作用
- 学会读预测区间（PI），把它与置信区间区分开

第 4 章给出 FE-IV 模型的合并 RR $= 0.65$，但 $I^2 = 92\%$ 表明 13 项研究并不在估同一个真实效应。本章松开” 同一个真实效应” 这条假设，把研究间的真实差异用一个新的方差参数 τ^2 估计出来。BCG 数据是异质性显著的典型案例，随机效应模型给出的合并结果与 FE 差异巨大，是 meta 分析教学的标准范例。

5.1 从固定效应到随机效应

固定效应模型假设 $\hat{y}_i = \theta + \varepsilon_i$，所有研究共享同一个真实效应 θ。随机效应模型把这个假设松开：每项研究自己有一个真实效应 θ_i，13 个 θ_i 不再要求相等，而是从一个总体分布里抽出来的。

定义 5.1 (随机效应模型)

$$\hat{y}_i = \theta_i + \varepsilon_i, \quad \theta_i \sim N(\mu, \tau^2), \quad \varepsilon_i \sim N(0, v_i).$$

其中 $\hat{y}_i$ 是第 i 项研究的 log RR 估计，θ_i 是第 i 项研究的真实效应（每项研究自己一个），μ 是平均真实效应（meta 分析合并的目标）。ε_i 是抽样误差，方差 v_i；θ_i 自身的分布方差是 τ^2，叫研究间方差。$\tau^2 = 0$ 时模型退化为固定效应。

♣

把 τ^2 翻译成生活语言。假设 13 项 BCG 试验各自的真实效应分别是 $\theta_1, \theta_2, \ldots, \theta_{13}$。固定效应模型说这 13 个数都相等（共有一个 θ）；随机效应模型说这 13 个数不相等，但围绕一个总体均值 μ 分布，分布的宽度就是 τ。BCG 数据里，伦敦 Hart & Sutherland 1977 的真实效应可能是 log RR $= -1.4$，赤道附近 TPT Madras 1980 的真实效应可能是 0，二者本来就不一样。τ^2 把这种本来就有的差异量化出来，让合并 CI 老老实实变宽。

5.1.1 随机效应模型的权重

随机效应模型的权重也按方差倒数计算，但方差包含两部分：研究内的抽样方差 v_i 和研究间的真实方差 τ^2。

定义 5.2 (随机效应权重)

RE 模型下每项研究的权重为

$$w_i^{\mathrm{RE}} = \frac{1}{v_i + \tau^2}.$$

其中 v_i 是第 i 项研究自身的抽样方差（同 FE），τ^2 是研究间真实效应的方差。合并估计 $\hat{\mu} = \sum_i w_i^{\mathrm{RE}} \hat{y}_i / \sum_i w_i^{\mathrm{RE}}$，方差 $\mathrm{Var}(\hat{\mu}) = 1/\sum_i w_i^{\mathrm{RE}}$。

♣

把 τ^2 加进权重的分母会带来两个直接后果。其中一个：所有研究的权重都比 FE 模型小，因为分母变大了，导致合并 CI 比 FE 宽。另一个：原本权重悬殊的研究权重被” 摊平”。FE 模型下 TPT Madras 一项就占 41%，到 RE 模型下这个比例会显著缩水，因为 τ^2 把它的相对优势削弱了。

5.2 DerSimonian-Laird 估计量

τ^2 不是直接观测的，必须从数据里估计。DerSimonian & Laird 1986 提出了最早也是最常用的 τ^2 矩估计量。

定义 5.3 (DerSimonian-Laird τ^2 估计量)

$$\hat{\tau}^2_{\mathrm{DL}} = \max\left\{0,\ \frac{Q-(k-1)}{\sum_i w_i - \sum_i w_i^2 / \sum_i w_i}\right\}.$$

其中 Q 是异质性 Q 统计量（第 4 章给出）、k 是研究数、$w_i = 1/v_i$ 是 FE 模型下的逆方差权重。分子 $Q-(k-1)$ 是观察 Q 超出”在 FE 假设下应有期望 $k-1$”的部分，分母把这个超额量还原到 log RR 平方尺度。如果 $Q \le k-1$（异质性弱于纯抽样波动），τ^2 截断为零。

在 BCG 数据上手算一遍。先把第 4 章给出的 13 项研究 IV 权重 $w_i = 1/v_i$ 全部累加：$\sum w_i = 609.70$。再算 $\sum w_i^2 = 94{,}820.58$，所以 $\sum w_i^2 / \sum w_i = 94{,}820.58/609.70 = 155.52$。DL 公式分母 $\sum w_i - \sum w_i^2/\sum w_i = 609.70 - 155.52 = 454.18$。Q 已经在第 4 章算出 $Q = 152.23$，自由度 $k-1 = 12$，分子 $Q - 12 = 140.23$。代入：

$$\hat{\tau}^2_{\mathrm{DL}} = \frac{140.23}{454.18} = 0.3088.$$

分子 140.23 是观察到的研究间散布超出抽样波动的部分，分母 454.18 是把这个超额量还原到 log RR 平方尺度的换算因子。

```r
library(metafor)
d  <- read.csv("data/bcg.csv")
es <- escalc(measure = "RR", ai = tpos, bi = tneg,
             ci = cpos, di = cneg, data = d)

# 手算 DL τ² 的中间量
w       <- 1 / es$vi
sum_w   <- sum(w)               # 609.70
sum_w2  <- sum(w^2)             # 94820.58
denom   <- sum_w - sum_w2 / sum_w   # 454.18
Q       <- rma(yi, vi, data = es, method = "EE")$QE   # 152.23
tau2_DL <- (Q - 12) / denom
cat("tau^2 (DL, 手算) =", round(tau2_DL, 4), "\n")

# DerSimonian-Laird τ² 估计（与手算应当一致）
re_dl <- rma(yi, vi, data = es, method = "DL")
print(re_dl)
```

DL 估计结果

RE-DL 给出 $\hat{\tau}^2_{\mathrm{DL}} = 0.3088$，$\hat{\tau} = 0.556$。合并 $\widehat{\log \mathrm{RR}} = -0.7141$，标准误 0.1787，95% CI $[-1.0644,\ -0.3638]$，对应 RR $= 0.49$，95% CI $[0.35,\ 0.70]$。

与 FE-IV 的 RR $= 0.65$ 相比，RE-DL 的合并 RR 偏小（保护效应更强），CI 宽度从 FE 的 $[0.60,\ 0.70]$ 变成 $[0.35,\ 0.70]$。这种 CI 大幅变宽是合理的：随机效应承认研究之间存在真实差异，CI 必须把这种差异纳入不确定性。

DL 是矩估计——它要求 Q 的期望值在 FE 模型下等于 $k-1$，超出 $k-1$ 的部分就归因于研究间方差。这个推导不需要假设 θ_i 服从正态，是一种”分布无关”的估计。代价是 DL 在小研究数（$k < 20$）和高异质性下**偏低**：它会低估 τ^2，进而高估权重精度，让 CI 偏窄。

5.3 REML 估计

REML（Restricted Maximum Likelihood，限制性最大似然）是基于正态假设的最大似然估计，理论上效率更高，在小研究数下偏倚比 DL 小。Cochrane Handbook 与 metafor 包都把 REML 作为默认 τ^2 估计量。

定义 5.4 (REML τ^2 估计)

在假设 $\theta_i \sim N(\mu, \tau^2)$ 与 $\hat{y}_i \mid \theta_i \sim N(\theta_i, v_i)$ 下，REML 通过对"抹去 μ 后的边际似然"求极值得到 τ^2 估计。形式上没有解析解，由迭代算法求解，metafor 默认使用 Fisher scoring 算法。

♣

REML 的优势在小样本下尤为明显。DL 的矩估计在 k 小的时候 Q 本身的方差就不可忽略，DL 估计波动很大。REML 用整个似然函数的形状信息，对 τ^2 的估计更稳。代价是 REML 假设 θ_i 服从正态，DL 不假设。实务上正态假设的违反在 k 较大时影响有限。

```
re_reml <- rma(yi, vi, data = es, method = "REML")
print(re_reml)
```

REML 估计结果

REML 给出 $\hat{\tau}^2_{\text{REML}} = 0.3132$，与 DL 的 0.3088 几乎一致。合并 $\widehat{\log \text{RR}} = -0.7145$，95% CI $[-1.0669,\ -0.3622]$，$\text{RR} = 0.489$，95% CI $[0.34,\ 0.70]$。BCG 数据上 DL 与 REML 结果几乎重合，说明 $k = 13$ 已经足够让两种估计趋同。在更小的 k（如 $k = 5$）下，二者会拉开。

5.4 Hartung-Knapp 调整

DL 与 REML 都用大样本正态近似算合并 CI——把 $\hat{\mu}$ 的方差当作正态分布的方差，乘以 1.96 算 95% CI 半宽。这个近似在 k 大时合理，k 少时偏窄。Hartung & Knapp 2001 提出用 t 分布替代正态、自由度 $k-1$，让 CI 更稳健。

定义 5.5 (Hartung-Knapp 调整)

HK 调整修改两个东西。其一，把合并估计的标准误从 $\sqrt{1/\sum_i w_i^{\text{RE}}}$ 替换为

$$\text{SE}_{\text{HK}} = \sqrt{\frac{\sum_i w_i^{\text{RE}} (\hat{y}_i - \hat{\mu})^2}{(k-1)\sum_i w_i^{\text{RE}}}},$$

其中 w_i^{RE} 是 RE 权重，$\hat{\mu}$ 是 RE 合并均值。其二，置信区间用 t_{k-1} 分布的临界值替代 1.96：

$$\text{CI}_{95\%} = \hat{\mu} \pm t_{k-1,0.975} \cdot \text{SE}_{\text{HK}}.$$

metafor 通过参数 `test = "knha"` 启用 HK 调整。

♣

HK 为什么更稳？两件事都是把 CI 拉宽。第一件事是用每项研究偏离合并均值的实际散布替换理论方差，让 CI 反映"实际看到的散布"而不是"理论应该有的散布"。第二件事是把 $z_{0.975} = 1.96$ 换成 $t_{k-1,0.975}$。当 $k = 13$，$t_{12,0.975} = 2.179$，比 1.96 大约 11%；$k = 5$ 时 $t_{4,0.975} = 2.776$，比 1.96 大约 42%。研究数越少，HK 调整把 CI 拉得越宽，避免小样本下的过度自信。

```
re_hk <- rma(yi, vi, data = es, method = "REML", test = "knha")
print(re_hk)
```

REML + HK 调整结果

HK 调整后合并 $\widehat{\log \mathrm{RR}} = -0.7145$（点估计与 REML 完全相同），标准误 0.1808（略高于 REML 的 0.1798），95% CI $[-1.1084,\ -0.3206]$，$\mathrm{RR} = 0.489$，95% CI $[0.33,\ 0.73]$。p 值用 t 分布算，$p = 0.0019$（仍显著）。Cochrane Handbook 自 2019 年起推荐随机效应模型默认使用 REML + HK 组合，尤其在 $k < 20$ 时。本书后续的随机效应主结果均采用这组配置。

5.5 预测区间 PI

随机效应模型有一个比 CI 更直接对应临床决策的指标——**预测区间**（prediction interval，简称 PI）。CI 描述合并均值的不确定性，PI 描述未来一项新研究的真实效应可能在哪里。在异质性显著时 PI 比 CI 重要得多。

定义 5.6 (预测区间)

$$\mathrm{PI}_{95\%} = \hat{\mu} \pm t_{k-2,0.975} \cdot \sqrt{\hat{\tau}^2 + \mathrm{SE}(\hat{\mu})^2}.$$

其中 $\hat{\mu}$ 是 RE 合并均值，$\hat{\tau}^2$ 是研究间方差估计，$\mathrm{SE}(\hat{\mu})$ 是合并均值的标准误。PI 比 CI 宽，因为根号里把 $\hat{\tau}^2$（未来研究自身的真实异质性）和 $\mathrm{SE}(\hat{\mu})^2$（合并估计的不确定性）两部分加在一起。自由度 $k-2$ 而非 **HK CI** 的 $k-1$：PI 公式根号里多估了一个 τ^2 参数，按 IntHout-Higgins-Tovey 2016 的推导多扣一个自由度。metafor 的 `predict()` 函数对带 `test = "knha"` 的对象会自动用对应自由度计算 CI 与 PI。 ♣

PI 与 CI 的差别在临床上至关重要。CI 告诉你的是” 全球 BCG 的平均效应”；PI 告诉你的是” 在某一个具体地区开展 BCG 项目，预期效应的范围”。如果 PI 跨过零线，意味着未来一项研究完全可能给出无效甚至反向的结果，临床推广就要谨慎。

```
pi_hk <- predict(re_hk)
cat("合并 RR =", round(exp(pi_hk$pred), 3),
    "  95% CI =", round(exp(pi_hk$ci.lb), 3), "to",
                  round(exp(pi_hk$ci.ub), 3), "\n")
cat("95% PI  = [", round(exp(pi_hk$pi.lb), 3), ",",
                   round(exp(pi_hk$pi.ub), 3), "]\n")
```

BCG 数据的预测区间

REML + HK 给出合并 $\mathrm{RR} = 0.489$，95% CI $[0.33,\ 0.73]$，95% PI $[0.136,\ 1.763]$。CI 不跨过 1，” 全球平均 BCG 有保护效应” 这个结论在 5% 错误率下成立。但 PI 跨过 1，意味着未来在某个新地区开展的 BCG 试验可能给出从 $\mathrm{RR} = 0.14$（强保护，把风险压到原来的 14%）到 $\mathrm{RR} = 1.76$（反向，比对照组风险高 76%）的任何结果。这就是 BCG 数据真正的临床信息：BCG 平均有效，但在不同地区效果可能截然不同。第 7 章的纬度 meta 回归会回答这种” 具体地区效应” 如何用协变量解释。

5.6 Forest plot 带预测区间

随机效应 forest plot 通常在合并菱形下方加一根额外的横线代表 PI，方便临床读者直接读出” 未来研究的可能效应范围”。metafor 通过参数 `addpred = TRUE` 自动添加这一行。

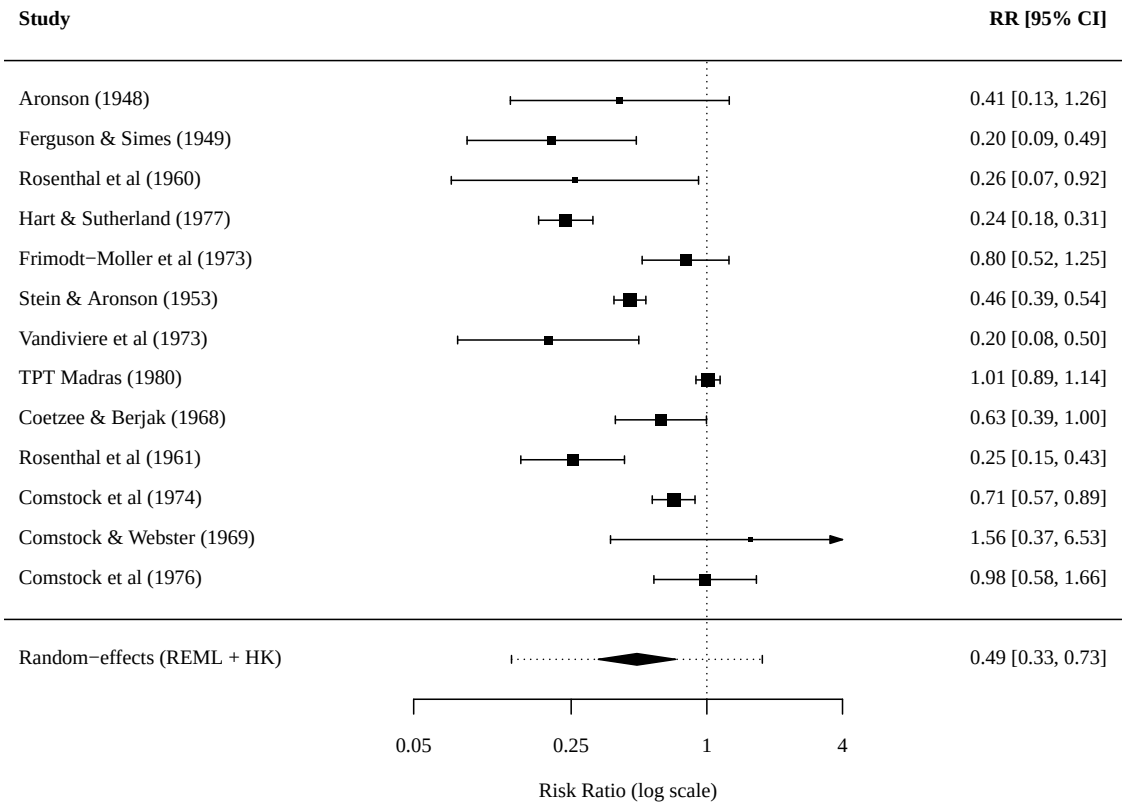


图 5.1: 13 项 BCG 试验随机效应（REML + HK）模型 forest plot。底部菱形为合并 RR = 0.49（95% CI 0.33–0.73），其下方水平横线为 95% 预测区间 [0.136, 1.76]，跨过 1，提示未来研究的不确定性远大于合并 CI 暗示的程度。

定理 5.1 (雷区)

随机效应模型有两个常被混淆的雷区。第一是把 **PI** 当 **CI** 解读。论文里读到"95% interval = [0.13, 1.76]"，不知道是 CI 还是 PI 就草率引用，会得到完全相反的结论。CI 描述合并均值的不确定性，PI 描述未来研究的不确定性。报告时必须明确标注。第二是在异质性极高时报告 **RE** 合并均值。当 $I^2 > 85\%$，不同研究的真实效应很可能本质上不同，合并的 μ 可能对应不到任何实际场景。BCG 数据 $I^2 = 92\%$，单一合并 RR = 0.49 只是统计意义上的全球平均，不能直接解读为"在我所在的地区 BCG 把风险降低 51%"。正确做法是配合第 7 章的亚组与 meta 回归，解释异质性来源。 ♡

第 5 章到这里收束。三种 RE 估计（DL、REML、REML+HK）给出几乎相同的点估计 RR = 0.49，CI 宽度从 DL 的 $[0.35, 0.70]$ 一路扩到 REML+HK 的 $[0.33, 0.73]$。比 FE-IV 的 $[0.60, 0.70]$ 显著更宽是合理的：RE 模型承认研究本来就不一样，CI 必须把这种差异纳入不确定性。

但合并 RR = 0.49 与 PI $[0.14, 1.76]$ 跨度极大的反差说明，**异质性本身需要被正面量化**。第 6 章就用 Q、I^2、τ^2 这一组指标系统刻画异质性，PI 不再重复定义，按本章 Definition 5.6 直接引用。第 6 章末把所有方法（FE 与 RE）的合并结果汇入一张更新的对比表。

第 6 章 异质性诊断：Q、I^2、τ^2 与预测区间

内容提要

- 理解异质性的两个来源：抽样波动与真实效应差异
- 掌握 Cochran's Q 检验的功效局限
- 学会区分 I^2、H^2、τ^2 三个相关但不同的指标
- 学会用 Galbraith 与 Baujat 图做异质性的视觉诊断

第 5 章已经把 τ^2 与预测区间用起来了，本章把" 异质性" 这一组诊断指标系统讲清楚。第 5 章告诉我们 $\tau^2 = 0.31$ 是研究间方差的绝对量，本章要回答：这个 τ^2 在异质性强弱坐标里处于什么位置？$I^2 = 92\%$ 这个百分比怎么算出来？为什么报告 meta 分析时必须配齐 Q、I²、τ² 三个数字？

Meta 分析里" 异质性" 这个词被用来指代两类不同的现象：研究在临床或方法学上的差异（叫**临床异质性**和**方法学异质性**），与统计上效应估计之间的散布（叫**统计异质性**）。本章只讲统计异质性，因为它有现成的量化指标。临床和方法学异质性的判断要靠领域知识，本书不展开。

6.1 异质性的两个来源

13 项 BCG 研究的 log RR 散布在 -1.62 到 $+0.45$ 之间。这种散布的来源有两个：每项研究因为样本有限会有抽样波动（同样的真实效应，重复抽样会得到不同点估计）；研究之间的真实效应可能本来就不一样（伦敦人群与马德拉斯人群的真实 BCG 效应可能根本不是一个数）。统计异质性诊断的核心任务是：把观测到的总散布拆成" 抽样波动" 与" 真实差异" 两部分，看真实差异占多大比例。

理解了这个拆解，后续 Q、I^2、H^2、τ^2 之间的关系就一目了然——它们都是这个拆解的不同表达方式。

6.2 Cochran's Q 检验

Cochran 1954 提出的 Q 统计量是最早的异质性检验。它把每项研究的 log RR 偏离 FE 合并估计的距离按权重加权平方求和。

定义 6.1 (Cochran's Q)

$$Q = \sum_{i=1}^{k} w_i (\hat{y}_i - \hat{\mu})^2.$$

其中 $w_i = 1/v_i$ 是 FE 模型的逆方差权重，$\hat{y}_i$ 是第 i 项研究的 log RR 估计，$\hat{\mu}$ 是 FE 合并估计，k 是研究数。在原假设" 所有研究估同一个真实效应"（$\tau^2 = 0$）下，$Q \sim \chi^2_{k-1}$，期望值等于自由度 $k-1$。Q 显著超过 $k-1$ 提示存在真实效应异质性。 ♣

为什么 Q 在原假设下期望等于自由度？这是 χ^2 分布的性质：χ^2_{k-1} 的均值就是 $k-1$。直觉上，每一项 $w_i(\hat{y}_i - \hat{\mu})^2$ 的期望大约是 1（标准化平方），k 项加起来减去用于估计 $\hat{\mu}$ 损失的 1 个自由度，剩下 $k-1$。

把 Q 换算成 p 值是用 χ^2_{k-1} 分布查表。BCG 数据 $Q = 152.23$，$df = 12$，$\chi^2_{12,0.95} = 21.03$，远小于 152.23，$p < 10^{-16}$。异质性极度显著。

定理 6.1 (雷区)

Q 检验在小研究数下功效低。$k = 5$、$\tau = 0.3$ 时 Q 检验的功效常常不到 50%，意味着即使有真实异质性也可能漏报。$p > 0.05$ 不能解读为”无异质性”，只能解读为”Q 检验没拒绝原假设”。判断异质性应当看 I^2 和 τ^2 的点估计，而不是只盯着 Q 的 p 值。Cochrane Handbook 从 2008 年起就明确反对”$p > 0.05$ 就用 FE 模型”的实务做法。

♡

6.3 I^2 与 H^2

Higgins & Thompson 2002 提出 I^2，把 Q 换算成”研究间异质性占总变异的比例”，更易解读。

定义 6.2 (I^2 与 H^2)

Higgins & Thompson 2002 的原始公式直接由 Q 推得：

$$H^2 = \frac{Q}{k-1}, \qquad I^2 = \frac{Q-(k-1)}{Q} = 1 - \frac{1}{H^2}.$$

其中 Q 是 Cochran's Q、k 是研究数。H^2 是观测总变异与抽样变异的比，I^2 是研究间真实差异占总变异的比例。$I^2 = 0$ 表示研究间无真实差异；$I^2 = 1$ 表示几乎所有差距都来自真实异质性。

♣

两种 I^2 公式的小差异。把 BCG 的 $Q = 152.23$、$k - 1 = 12$ 代入原始公式得 $I^2 = 140.23/152.23 = 92.12\%$，与 metafor 报的 92.22% 差了 0.10 个百分点。差异不是 bug，是 metafor 在 RE 模型上用了一个等价但更直接的版本：$I^2 = \tau^2/(\tau^2 + \tilde{s}^2)$，其中 $\tilde{s}^2$ 是 13 项研究抽样方差按权重综合后的”典型抽样方差”。这个 generalized 形式直接用 τ^2 与抽样方差的比例度量异质性，与原始公式在异质性显著时数值几乎一致，差异常常在小数点后第二位。两个形式都对，本书所有 I^2 数字以 metafor 输出为准。

把 I^2 翻译成生活语言。13 项 BCG 研究的 log RR 散布有 92% 来自真实差异，只有 8% 来自抽样波动。这就像 13 个不同地区报告同一种疫苗的效果，地区之间的差异比每个地区内部的随机波动大十倍以上，意味着”全球平均效应”是个统计意义上的合并，对应不到任何具体地区。

I^2 的常用解读区间是 Higgins 2003 提出的：$< 25\%$ 低、25%–50% 中等、50%–75% 较高、$> 75\%$ 高。这是经验区间，不是绝对标准。临床决策应当结合 τ^2 与预测区间一起判断。

定理 6.2 (雷区)

I^2 有两个常见误用。其中一个是把 I^2 当成异质性的绝对量。I^2 是相对比例，受抽样方差大小影响：研究内方差越小，I^2 越接近 1；研究内方差越大，即使真实异质性相同，I^2 也会偏小。BCG 数据 13 项研究方差跨度从 0.004 到 0.53，$I^2 = 92\%$ 这个数字依赖于方差分布的具体形态。另一个是在小研究数下相信 I^2 的点估计。$k = 5$ 时 I^2 的 95% CI 常常宽到 $[0\%, 95\%]$，单看点估计没有信息量。metafor 的 `confint()` 函数会同时报 I^2 与它的 CI，必须一起读。

♡

6.4 τ^2：异质性的绝对量

τ^2 是研究间真实效应方差的绝对量，单位与 log RR 平方一致。它是 RE 模型的核心参数，第 5 章已经介绍过两种估计方法（DL 和 REML）。

定义 6.3 (τ 在原尺度上的解读)

τ^2 的开方 τ 在 log RR 尺度上是研究间真实效应的标准差。把 τ 取指数 $\exp(\tau)$ 给出 RR 尺度上的”真实效应散布因子”。例如 $\tau = 0.56$ 时 $\exp(\tau) = 1.75$，意思是若两个研究真实 log RR 相差 1 个标准差 τ，二者的 RR 比为 $\exp(\tau) \approx 1.75$。

♣

回到 BCG 数据。下面这段代码用 REML + HK 配置算 τ^2 与它的 95% CI（基于 Q-profile 方法）。

```
re <- rma(yi, vi, data = es, method = "REML", test = "knha")
ci_tau <- confint(re)
print(ci_tau)
```

BCG 数据的 τ^2 估计与 CI

$\hat{\tau}^2 = 0.313$，95% CI $[0.120, 1.112]$。CI 不包含零，说明 τ^2 显著大于零，存在真实异质性。$\hat{\tau} = 0.56$，对应原 RR 尺度上 $\exp(0.56) \approx 1.75$：13 个真实效应的 RR 差一个标准差就差出 1.75 倍。

I^2 的 95% CI 是 [81.9%, 97.7%]，下界 82% 已经超过 Higgins 的" 高异质性" 阈值 75%。无论从点估计还是区间估计看，BCG 数据的异质性都毫无疑问。

6.5 异质性的视觉诊断

数字之外，两张图能直观展示异质性。第一张是 Galbraith plot（也叫 radial plot），横轴是研究的精度 $1/\mathrm{SE}_i$，纵轴是标准化效应 $\hat{y}_i/\mathrm{SE}_i$。同质时所有点应该集中在一根过原点的直线两侧 ± 2 范围内；超出 ± 2 的点是异质性的视觉证据。

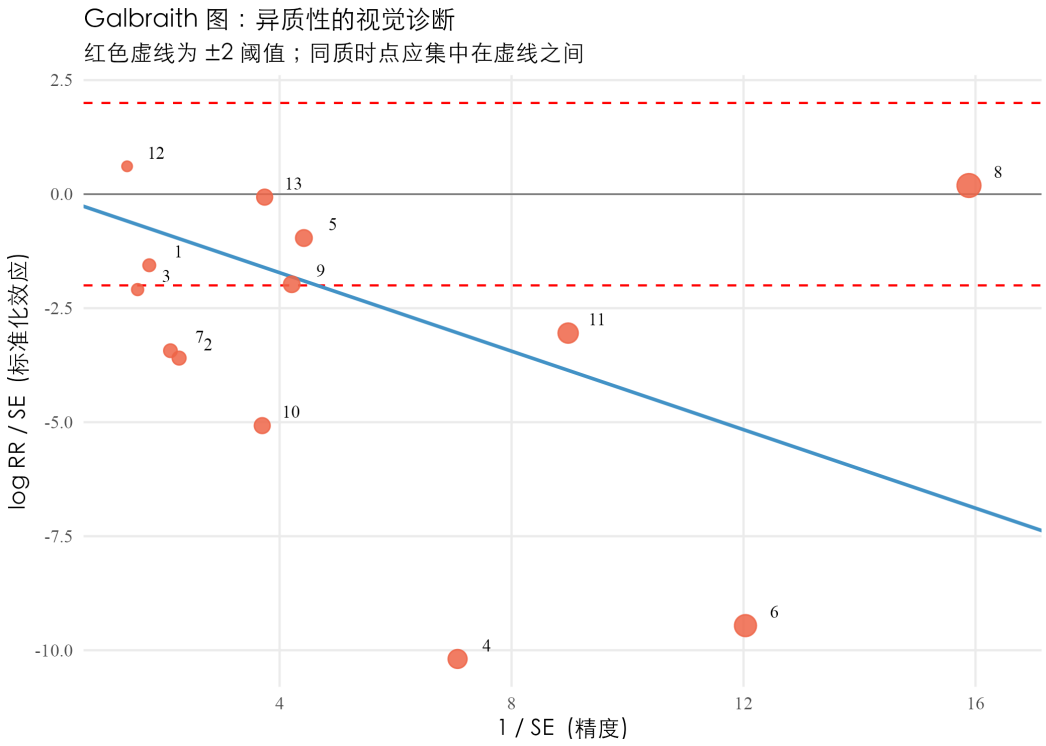


图 6.1: Galbraith 图。蓝色直线是 FE 合并估计对应的斜率（截距过原点）。红色虚线为标准化效应的 ± 2 阈值。多项研究偏离虚线之外，是异质性显著的视觉证据。

第二张是 Baujat plot，横轴是每项研究对总 Q 的贡献（异质性贡献），纵轴是每项研究对合并估计的影响力（去掉它之后合并估计变化多少）。右上角的研究既贡献异质性又主导合并，是 meta 分析里" 最有问题" 的研究。

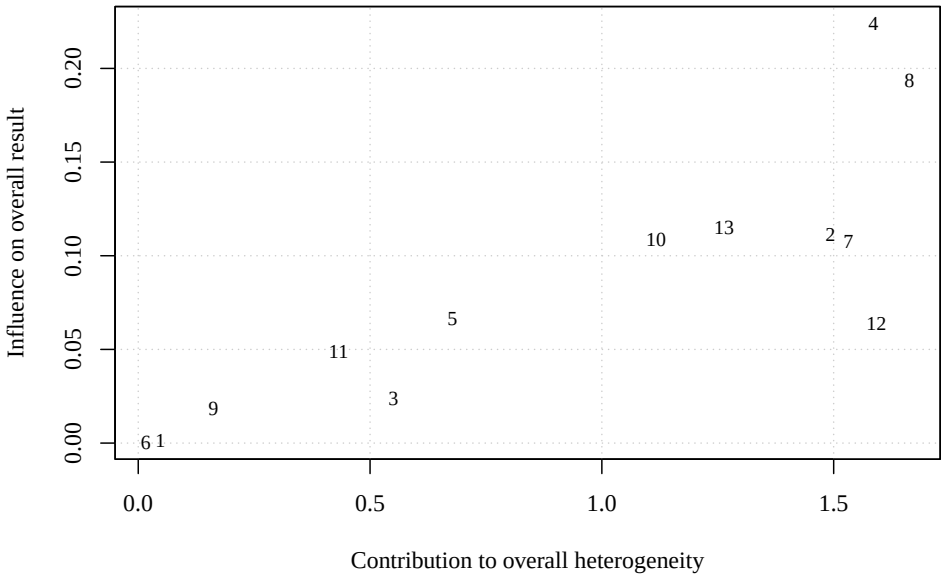


图 6.2: Baujat 图。横轴是研究对 Q 的贡献，纵轴是去掉该研究后合并估计的变化量。右上角的研究同时贡献异质性与影响合并，是敏感性分析的优先关注对象。

Baujat 图的右上角主要是 4 号 Hart & Sutherland 1977（高纬度强保护）与 8 号 TPT Madras 1980（赤道附近无效）。这两项分别是异质性的” 两极”，正是第 7 章 meta 回归要解释的核心。

6.6 累积对比表与异质性指标完整报告

把 BCG 数据上的全部异质性指标整理成一份完整报告供后续章节引用：

表 6.1: dat.bcg 异质性指标完整报告

指标	估计	解读
Q	152.23	自由度 12，$p < 10^{-16}$，异质性极度显著
H^2	12.86	总变异是抽样变异的 12.86 倍
I^2	92.22%	总变异中超过 92% 来自真实异质性
I^2 95% CI	[81.9%, 97.7%]	下界仍在 Higgins” 高异质性” 阈值之上
τ^2	0.313	log RR 尺度上的研究间方差
τ^2 95% CI	[0.120, 1.112]	不含零，τ^2 显著大于零
τ	0.56	log RR 尺度的研究间标准差
$\exp(\tau)$	1.75	RR 尺度上一倍标准差对应 1.75 倍效应差
PI (RR)	[0.136, 1.763]	未来研究的 RR 可能跨度极大（见第 5 章）

第 6 章是合并方法引入的中段节点。截至本章，全部六种合并估计（朴素合并 + 两种 FE + 三种 RE）的累积对比表如下。

表 6.2: 方法对比表（截至第 6 章）

方法	logRR	RR (95% CI)	τ^2	核心假设	局限
朴素合并	-0.488	0.61	—	2x2 表可直接相加	被大样本主导
FE-IV	-0.430	0.65 (0.60, 0.70)	0	共有真实效应	$I^2 = 92\%$，假设违反
FE-MH	-0.454	0.64 (0.59, 0.69)	0	同 FE-IV，稀疏稳健	异质性 CI 偏窄
RE-DL	-0.714	0.49 (0.35, 0.70)	0.309	真实效应正态分布	τ^2 小样本偏低
RE-REML	-0.715	0.49 (0.34, 0.70)	0.313	同 RE-DL	依赖正态假设
RE-REML+HK	-0.715	0.49 (0.33, 0.73)	0.313	同 RE-REML	k 小时 t 近似仍可能偏窄

绿色高亮的 RE-REML+HK 是本书后续章节的主分析方法。但 $\mathrm{RR} = 0.49$ 与 PI $[0.14, 1.76]$ 跨度极大的反差，意味着合并均值本身需要进一步分解：13 项研究为什么差异这么大？第 7 章用纬度作为协变量做 meta 回归，看能否把 $\tau^2 = 0.31$ 解释掉相当一部分。

第 7 章 亚组分析与 meta 回归

内容提要

- 理解亚组分析与 meta 回归的关系：前者是后者的离散化特例
- 学会用 `mods` 参数把研究层级协变量喂给 `rma()`
- 用纬度 *ablat* 解释 dat.bcg 的研究间方差
- 用分配方式 *alloc* 检验亚组差异
- 警惕协变量过拟合：13 项研究里加几个协变量已是上限

第 6 章给出 $I^2 = 92\%$、$\tau^2 = 0.31$，意味着 13 项 BCG 试验之间的真实效应差异极大。本章的任务是把这种差异**用协变量解释**。如果我们能找到一个研究层级的变量，使得在该变量条件下不同研究的真实效应趋于一致，那么 τ^2 就会显著下降，meta 分析的合并均值也能换成"在协变量取某个值时的条件均值"，比单一全球平均更有临床意义。

7.1 亚组分析与 meta 回归的关系

亚组分析是按某个离散协变量（如分配方式 *alloc*）把研究分组，每组单独跑随机效应模型。Meta 回归是把协变量（离散或连续）作为自变量进入回归方程。两者的本质相同：都是用研究层级协变量解释 $\hat{y}_i$。亚组分析是 meta 回归在分类协变量上的特例。

定义 7.1 (Meta 回归)

在随机效应模型基础上加入研究层级协变量 x_i：

$$\hat{y}_i = \beta_0 + \beta_1 x_i + u_i + \varepsilon_i, \quad u_i \sim N(0, \tau^2_{\text{res}}), \quad \varepsilon_i \sim N(0, v_i).$$

β_1 是协变量的回归系数，τ^2_{res} 是残差异质性，即协变量未能解释的研究间方差。模型用加权回归（权重 $1/(v_i + \tau^2_{\text{res}})$）迭代估计，可加入多个协变量。 ♣

β_1 的解读类似线性回归：x 每增加一个单位，效应量（log RR）平均变化 β_1。τ^2_{res} 与基线模型的 τ^2_{base} 之差就是协变量**解释掉的异质性**，常用比例 $R^2 = 1 - \tau^2_{\text{res}}/\tau^2_{\text{base}}$ 报告（meta 分析里的 R^2 与传统线性回归 R^2 含义类似但定义略不同）。

7.2 亚组分析：按分配方式 alloc

dat.bcg 的 *alloc* 列记录了每项研究的随机化方式：`random`（标准随机化）、`alternate`（按入院顺序交替分配，方法学等级偏低）、`systematic`（按某种系统规则分配，比如按出生日期）。13 项研究里 7 项是 `random`，2 项 `alternate`，4 项 `systematic`。理论上 `random` 的偏倚风险最低，结果应当最可信。

```
re_random     <- rma(yi, vi, subset = (alloc == "random"),
                     data = es, method = "REML", test = "knha")
re_alternate  <- rma(yi, vi, subset = (alloc == "alternate"),
                     data = es, method = "REML", test = "knha")
re_systematic <- rma(yi, vi, subset = (alloc == "systematic"),
                     data = es, method = "REML", test = "knha")
```

按 alloc 的亚组合并

三个亚组的合并结果如下：

alloc	k	合并 logRR	合并 RR (95% CI)	解读
random	7	−0.971	0.379 (0.20, 0.70)	标准随机化下 BCG 把风险压低 62%
alternate	2	−0.541	0.582 (0.02, 20.85)	仅 2 项研究，CI 失去意义
systematic	4	−0.424	0.654 (0.21, 2.09)	4 项研究，CI 跨过 1

从点估计看，`random` 亚组保护效应最强（RR = 0.38），`systematic` 亚组弱很多（RR = 0.65）。但 `alternate` 组仅 2 项研究，τ^2 估计极不稳定，CI 宽到 [0.02, 20.85]，几乎不携带信息。Cochrane Handbook 推荐亚组分析每组至少 3 项研究，`alternate` 组实际上不该单独报。

7.2.1 亚组之间的差异检验

亚组结果摆出来后，下一个问题是：三个亚组的合并均值差异是不是统计上显著？metafor 通过 `mods = ~ alloc` 把 *alloc* 当作分类 moderator 喂进模型，得到组间差异检验。

```
re_alloc <- rma(yi, vi, mods = ~ alloc, data = es,
                method = "REML", test = "knha")
```

alloc 的 moderator 检验

组间差异检验给出 $Q_M = 0.97$，$df = 2$，$p = 0.413$。在 $\alpha = 0.05$ 水平下，`alloc` 不能解释 13 项研究的异质性。点估计上 `random` 确实保护效应更强，但样本量不足以让差异显著。

7.3 Meta 回归：纬度 ablat

alloc 没解释掉异质性，让我们试连续变量 *ablat*。第 1 章的散点图已经预告：纬度越高，BCG 保护效应越强。本节把这个观察用 meta 回归正式化。

```
re_ablat <- rma(yi, vi, mods = ~ ablat, data = es,
                method = "REML", test = "knha")
print(re_ablat)
```

ablat 单变量 meta 回归

截距 $\beta_0 = 0.252$（赤道处的预期 log RR，对应 RR = 1.29），斜率 $\beta_1 = -0.0291$，$t_{11} = -3.55$，$p = 0.0046$。95% CI $[-0.0472, -0.0111]$。意思是纬度每增加 1 度，log RR 下降 0.029，对应 RR 乘以 $\exp(-0.029) \approx 0.97$。从赤道到北纬 50 度，RR 累计变化 $\exp(-0.029 \times 50) \approx 0.23$ 倍，是个巨大的效应修饰。

最关键的数字是 τ^2 的变化。基线 $\tau^2_{\text{base}} = 0.313$，加入 *ablat* 后残差 $\tau^2_{\text{res}} = 0.076$。$R^2 = 1 - 0.076/0.313 = 75.6\%$。**单单一个纬度变量，就解释了超过四分之三的研究间异质性**。这是临床流行病学里 meta 回归的经典展示，由 Colditz et al. 1994 首次报告 [1]。

7.3.1 Bubble plot：协变量与效应的散点

Meta 回归的标准可视化是 bubble plot，横轴是协变量、纵轴是效应估计、点的大小代表权重。metafor 没有内置 `bubble.plot()` 函数，需要手动用 ggplot2 画。下面这段代码画 *ablat* vs $\hat{y}_i$ 的 bubble plot，加上回归拟合线与 95% 置信带。

```
pred_x <- seq(min(es$ablat), max(es$ablat), length.out = 100)
pred_y <- predict(re_ablat, newmods = pred_x)

ggplot(es, aes(x = ablat, y = yi)) +
  geom_ribbon(data = data.frame(x = pred_x,
                                lo = pred_y$ci.lb,
                                hi = pred_y$ci.ub),
              aes(x = x, ymin = lo, ymax = hi),
              fill = "#EF6548", alpha = 0.15,
              inherit.aes = FALSE) +
  geom_line(data = data.frame(x = pred_x, y = pred_y$pred),
            aes(x = x, y = y), color = "#EF6548",
            inherit.aes = FALSE) +
  geom_point(aes(size = 1 / (vi + 0.076)),
             color = "#4292C6", alpha = 0.85) +
  labs(x = "绝对纬度（度）", y = "log RR")
```

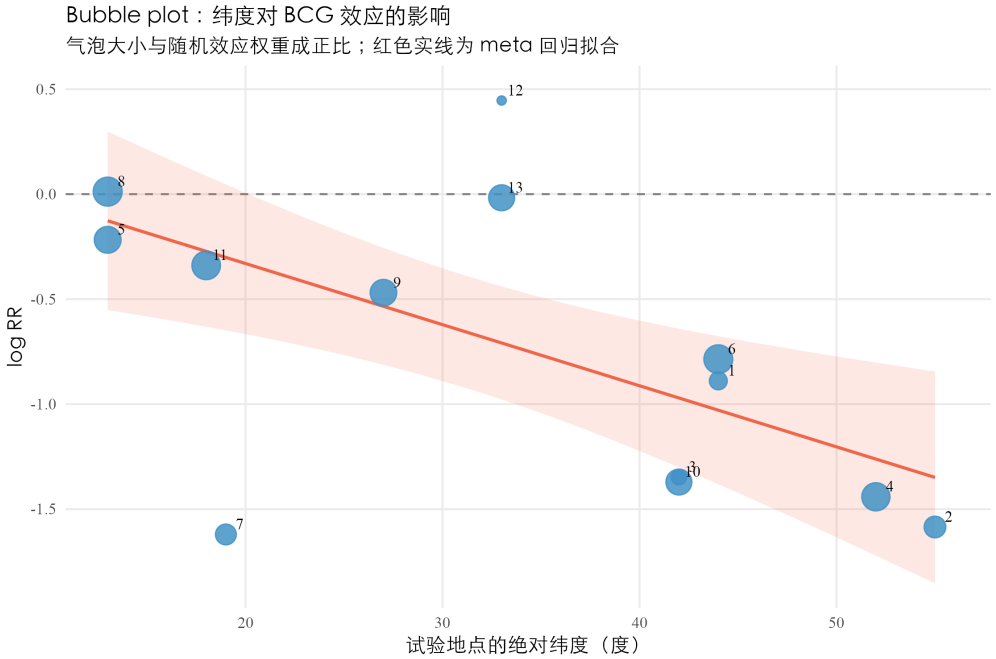


图 7.1: Bubble plot 展示纬度对 log RR 的影响。气泡大小与随机效应权重成正比；红色实线为 meta 回归拟合，浅红带为 95% CI。13 项研究沿一条向下的斜率分布，$R^2 = 75.6\%$。

为什么纬度能解释这么多？Colditz 1994 给出的生物学解释是：BCG 是把活的减毒分枝杆菌打进人体诱导免疫记忆，但低纬度热带地区的环境里有大量与结核分枝杆菌长得很像的非结核分枝杆菌，当地居民从小通过呼吸、皮肤接触等渠道暴露在这些”近亲”细菌里，免疫系统已经”自然认识”了结核杆菌的家族特征。BCG 多打一针的边际收益就很有限。高纬度地区缺乏这种环境暴露，BCG 才能从零开始建立免疫。这是一个生物学假设，meta 回归只是给这个假设提供数字支持。

定理 7.1 (雷区)

Meta 回归用的是研究层级的协变量与研究层级的效应，结论严格来说也只在研究层级成立。把”纬度每增加 1 度 log RR 下降 0.029”误解成”个体层纬度每增加 1 度个体的 log RR 下降 0.029”，是著名的生态谬误（ecological fallacy）。BCG 数据里”纬度”的真实角色是研究开展地点的总体特征（环境暴露、人群构成、疫苗株系）的一个代理，与个体的纬度位置无关。把研究层级关联误读为个体层级关联会得到错误的临床推广。诊断方法只有一条：报 meta 回归结论时必须明确写”研究层级关联”，不能省略这个限定词。 ♡

7.4 多变量 meta 回归

把 *ablat* 和 *alloc* 放进同一个回归看。

```
re_joint <- rma(yi, vi, mods = ~ ablat + alloc, data = es,
                method = "REML", test = "knha")
print(re_joint)
```

ablat + alloc 联合回归

联合模型给出 *ablat* 系数 -0.0273（$p = 0.0186$），*alloc* 各类型系数都不显著。$R^2 = 53.8\%$，反而比仅 *ablat* 模型的 75.6% 低。原因是联合模型多了 2 个 alloc 类别参数，REML 在小样本（$k = 13$）下对噪声敏感，加入不重要的协变量会让 τ^2_{res} 估计反弹。

这是 meta 回归的实务规则：**协变量个数应远小于研究数**。Borenstein 与 Higgins 2013 推荐每 10 项研究最多加 1 个协变量。13 项研究的上限就是 1 个连续协变量加 1 个二分类协变量。BCG 数据上单变量 ablat 模型已经解释了 75.6%，足够强；继续加变量会过拟合。

定理 7.2 (雷区)

Meta 回归过拟合是研究数 k 小时的常见错误。每加一个协变量，模型多估一个系数，τ^2_{res} 的稳定性下降。$k = 13$ 时加 3 个协变量已经接近极限，加 5 个就开始拟合噪声。识别过拟合有两条诊断方法：R^2 反而下降（如本章 *ablat* + *alloc* 模型）；个别协变量的 p 值剧烈波动（加入或剔除其他协变量会让 p 值从 0.01 跳到 0.5）。稳健做法是预先指定 1–2 个最关键协变量，不做”加几个看看”的探索式回归。 ♡

7.5 条件 RR 报告

Meta 回归的合并 RR 不是单一的”全球平均”，而是某个 *ablat* 取值下的条件均值。报告时通常给出协变量在几个有意义的取值（如样本两个端点与中位数）下的条件 RR。

```
# 在 ablat = 13°（赤道）, 33°（中位数）, 55°（高纬度）上的条件均值
predict(re_ablat, newmods = c(13, 33, 55), transf = exp)
```

输出（含 95% CI）：

表 7.1: Meta 回归在三个纬度上的条件 RR

纬度	条件 RR	95% CI	临床解读
13°（赤道附近）	0.88	(0.58, 1.35)	CI 跨过 1，效果不显著
33°（中位数）	0.49	(0.38, 0.64)	风险减半
55°（高纬度）	0.26	(0.16, 0.43)	风险压到原来的四分之一

把” 全球平均 RR $= 0.49$” 换成” 赤道附近 RR ≈ 0.88、中纬度 RR ≈ 0.49、高纬度 RR ≈ 0.26”，临床决策上更可操作。第 6 章累积对比表里 RE-REML+HK 的全球 RR $= 0.49$ 是中纬度场景的近似，赤道地区的实际效果可能完全不同。第 10 章会把这件事写进 GRADE 评级。

第 7 章到这里收束。$\tau^2 = 0.31$ 被纬度解释了 75.6%，研究间差异不再是黑箱。下一章检查 meta 分析方法学的另一个隐患：13 项研究是不是就是全部？还是有一些阴性结果根本没发表？这是**发表偏倚**问题。

第 8 章 发表偏倚的诊断与校正

内容提要

- 理解发表偏倚的产生机制与对 meta 分析的影响
- 学会用漏斗图直观判断小样本研究是否对称缺失
- 掌握 Egger 回归检验与 Begg 秩相关检验的差异
- 用 trim-and-fill 估计缺失研究并校正合并估计
- 警惕"漏斗图不对称等于发表偏倚"的常见误读

第 7 章用纬度解释了 75.6% 的研究间异质性，但 meta 分析的可信度还有一道关，纳入分析的 13 项研究本身可能就是有偏的子集。如果阴性结果或小样本无显著结果的研究系统性地不被发表，meta 分析的"原料"就被预先过滤过，合并估计会偏向显著方向。这种偏倚叫**发表偏倚**，英文称 publication bias，是 meta 分析最难完全控制的偏倚类型。

8.1 发表偏倚的产生机制

发表偏倚的根源是学术发表的筛选过程。期刊编辑、审稿人、作者本人都倾向于发表"有发现"的研究：阴性结果的研究投稿成功率明显低于阳性结果，作者本身也更可能把阴性结果归档而不投稿（这种现象叫 file drawer problem）。结果是 meta 分析能检索到的研究池里，阴性结果系统性地少于真实存在的数量。

定义 8.1 (发表偏倚)

当研究是否发表与研究结果（效应方向、统计显著性）相关时，文献库里的研究集合对真实研究总体不再是随机样本。系统综述检索到的子集就成为一个有偏样本，meta 分析的合并估计因此向"显著"或"阳性"方向偏移。识别与校正发表偏倚的核心工具是漏斗图及其相关检验。

8.2 漏斗图

漏斗图把每项研究画成一个点，横轴是效应估计，纵轴是研究的精度（通常用标准误 SE_i 或 $1/\mathrm{SE}_i$ 表示，纵轴反向使大研究在上、小研究在下）。在没有发表偏倚的理想情况下，所有研究的点应当围绕真实效应对称分布，形状像一个倒置的漏斗：精度高的大研究散布窄，集中在漏斗顶端；精度低的小研究散布宽，铺开在漏斗底部。

定义 8.2 (漏斗图的几何直觉)

设真实效应为 θ。每项研究的估计 $\hat{y}_i$ 在 θ 周围按 $N(\theta, v_i)$ 分布。精度 $1/\sqrt{v_i}$ 越高，分布越窄。在 θ 两侧 $\pm 1.96\sqrt{v_i}$ 的边界就是 95% 散布带，画在精度 vs 效应坐标系上形成两条向漏斗顶端汇聚的斜线。所有研究应落在这两条斜线包围的漏斗内。

发表偏倚的视觉特征是漏斗一侧（通常是阴性结果一侧）小研究系统性缺失。BCG 数据的漏斗图给出第一道诊断。

```
re <- rma(yi, vi, data = es, method = "REML", test = "knha")
funnel(re,
       xlab    = "log Risk Ratio",
       atransf = exp,
       at      = log(c(0.05, 0.25, 1, 4)),
```

```
       refline = re$beta,
       label   = "out")
```

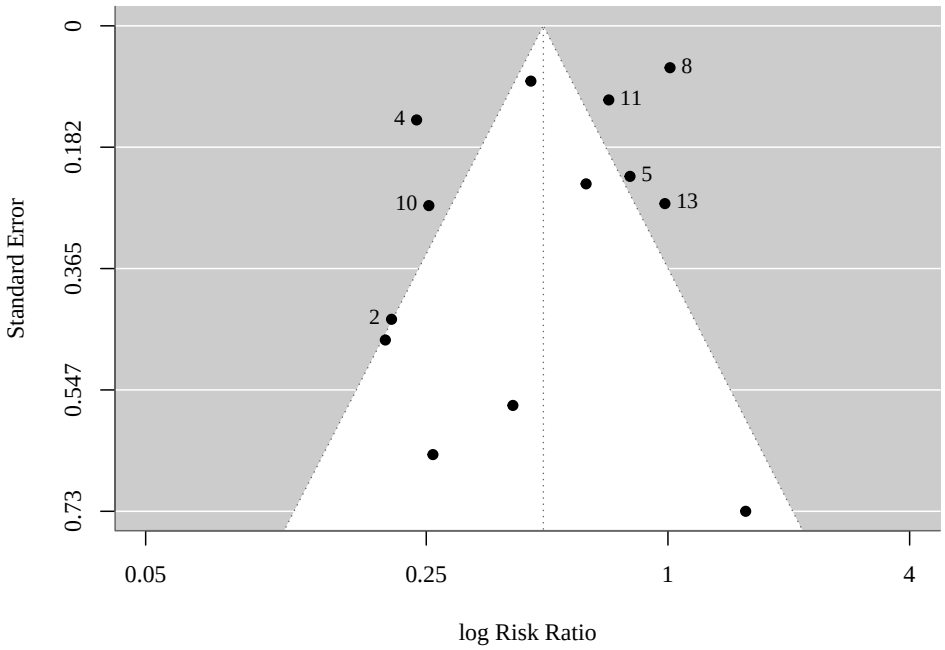


图 8.1: dat.bcg 的漏斗图。横轴为 log RR，纵轴为标准误（小研究在下）。虚线为合并估计 $\widehat{\log \mathrm{RR}}$，斜线为伪 95% CI 带。13 项研究落在漏斗内，左右两侧大致对称，没有一侧系统性缺失。

8.2.1 Contour-enhanced 漏斗图

标准漏斗图只看分布形状，无法区分” 漏斗一侧缺失是因为发表偏倚还是因为该侧本就没有研究”。Peters 2008 提出 contour-enhanced 漏斗图：在图上用阴影标注不同显著性区域（$p > 0.10$、$0.05 < p < 0.10$、$p < 0.05$），便于判断缺失的研究是否落在” 显著区” 还是” 非显著区”。

```
funnel(re,
       level    = c(90, 95, 99),
       shade    = c("white", "gray85", "gray70"),
       refline  = 0,
       atransf  = exp,
       legend   = TRUE)
```

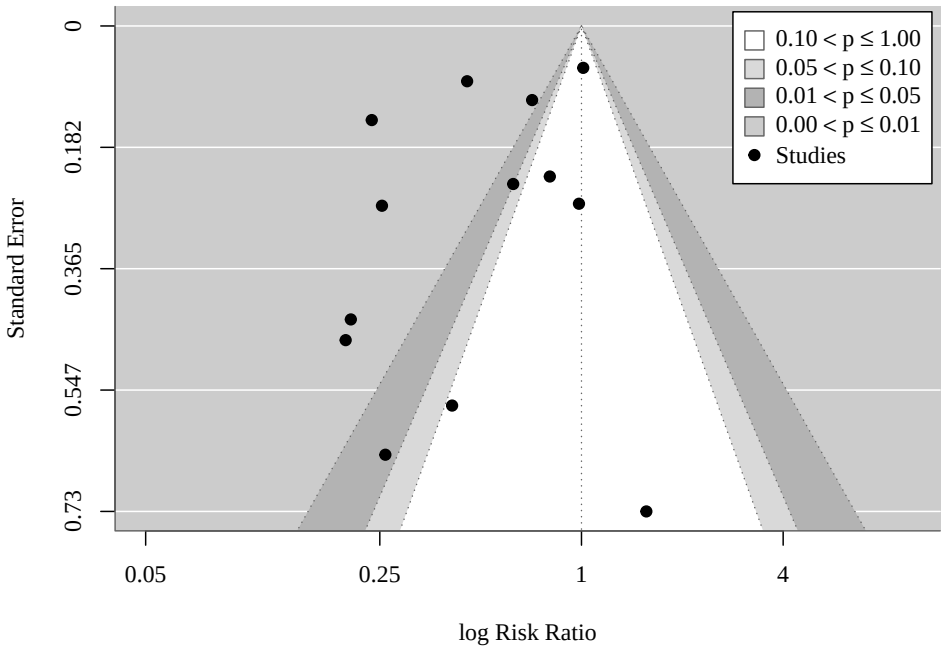


图 8.2: Contour-enhanced 漏斗图。阴影区代表不同显著性水平（白色 $p > 0.10$，浅灰 $0.05 < p < 0.10$，深灰 $p < 0.05$）。如果缺失研究都集中在白色非显著区，提示发表偏倚；如果分布与显著性区域无关，则不对称可能源于异质性。

8.3 Egger 回归检验

视觉判断有主观性，需要客观的统计检验。Egger 1997 把”漏斗图对称性”翻译成一个回归检验：把每项研究的 standardized effect（$\hat{y}_i/\mathrm{SE}_i$）对精度（$1/\mathrm{SE}_i$）做线性回归，截距应当为零（如果分布对称）。截距显著偏离零就提示不对称。

定义 8.3 (Egger 检验)

对每项研究计算 $z_i = \hat{y}_i/\mathrm{SE}_i$ 与 $p_i = 1/\mathrm{SE}_i$，拟合 $z_i = \alpha + \beta p_i + \epsilon_i$。原假设 $\alpha = 0$（漏斗对称）；α 显著不为零提示发表偏倚。检验统计量为 α 的 t 统计量，自由度 $k - 2$。

```
egger <- regtest(re, model = "lm")
print(egger)
```

BCG 数据的 Egger 检验

BCG 数据上 Egger 检验给出 $t = -1.40$，$df = 11$，$p = 0.189$。在 $\alpha = 0.05$ 水平下不能拒绝漏斗对称的原假设。”limit estimate as SE $\to$ 0” 给出极小研究的预期效应估计 $b = -0.19$（RR ≈ 0.83）。如果存在严重发表偏倚，这个 limit estimate 会显著小于合并估计；这里 limit $= -0.19$ 比合并 -0.71 小但 CI 跨过零，证据不足以指控发表偏倚。

Egger 检验在小研究数下功效低。$k = 13$ 时 Egger 的功效大约只有 30%–40%，意味着即使存在中度发表偏倚也可能漏报。Cochrane Handbook 推荐 $k \geq 10$ 才做 Egger 检验，本数据刚好在阈值线上，结论”无显著证据”应解读为”证据不足”，不能解读为”无偏倚”。

8.4 Begg 秩相关检验

Begg & Mazumdar 1994 提出另一种思路：用秩相关代替线性回归。对每项研究的标准化效应与方差排序，计算 Kendall's τ。τ 显著不为零提示效应大小与精度相关，间接提示不对称。

定义 8.4 (Begg 秩相关检验)

对每项研究的标准化效应 $\hat{y}_i/\sqrt{v_i^*}$（其中 v_i^* 是 Begg 修正方差）与方差 v_i 计算 Kendall's τ。原假设 $\tau = 0$；τ 显著不为零提示发表偏倚。

♣

Begg 检验比 Egger 检验对极端值更稳健，但功效更低。Cochrane Handbook 推荐两个检验都报，相互验证。

```
begg <- ranktest(re)
print(begg)
```

BCG 数据的 Begg 检验

BCG 数据上 Begg 检验给出 Kendall's $\tau = 0.026$，$p = 0.952$。这个 p 值远高于显著性阈值，与 Egger 检验的结论方向一致：没有发表偏倚的统计证据。Begg 与 Egger 双重检验都不显著，给临床读者的信号是"BCG meta 分析没有可识别的强发表偏倚"。但 $k = 13$ 这个样本规模本身就让两种检验功效不足，结论的严谨表述是" 证据不足以指控偏倚"，不是" 证明没有偏倚"。

8.5 Trim-and-fill 校正

Egger 与 Begg 是诊断检验，不直接给出" 如果存在偏倚，校正后的合并估计是多少" 的答案。Duval & Tweedie 2000 提出了 trim-and-fill 算法：假设漏斗图缺失的研究是" 真实但未发表" 的研究，按对称镜像把它们填补回去，重新合并。

定义 8.5 (Trim-and-fill 算法)

算法分两步。trim 步：从合并估计两侧识别" 明显异常" 的研究（贡献最多不对称的研究），临时移除这些研究重新计算合并估计。fill 步：以新合并估计为对称中心，把 trim 步移除的研究镜像填回到对侧。最终合并估计是包含所有原始研究与镜像填补研究的加权平均。metafor 通过 `trimfill()` 函数实现。

♣

```
tf <- trimfill(re)
print(tf)
```

BCG 数据的 trim-and-fill

Trim-and-fill 估计 BCG 数据右侧缺失 1 项研究（$k_0 = 1$）。填补一项后合并 $\widehat{\log \mathrm{RR}} = -0.657$，RR $= 0.518$，95% CI $[0.365,\ 0.736]$。与未校正的合并 RR $= 0.489$ 相比，保护效应略弱，但临床意义上几乎一致。

这个结果的解读必须谨慎。trim-and-fill 假设漏斗图所有不对称都来自发表偏倚，把异质性、方法学差异等其他不对称源都计入" 缺失研究"。BCG 数据的 $I^2 = 92\%$ 说明异质性极强，trim-and-fill 估计的"1 项缺失" 很可能是真实异质性的镜像偏差，而非真实的发表偏倚。

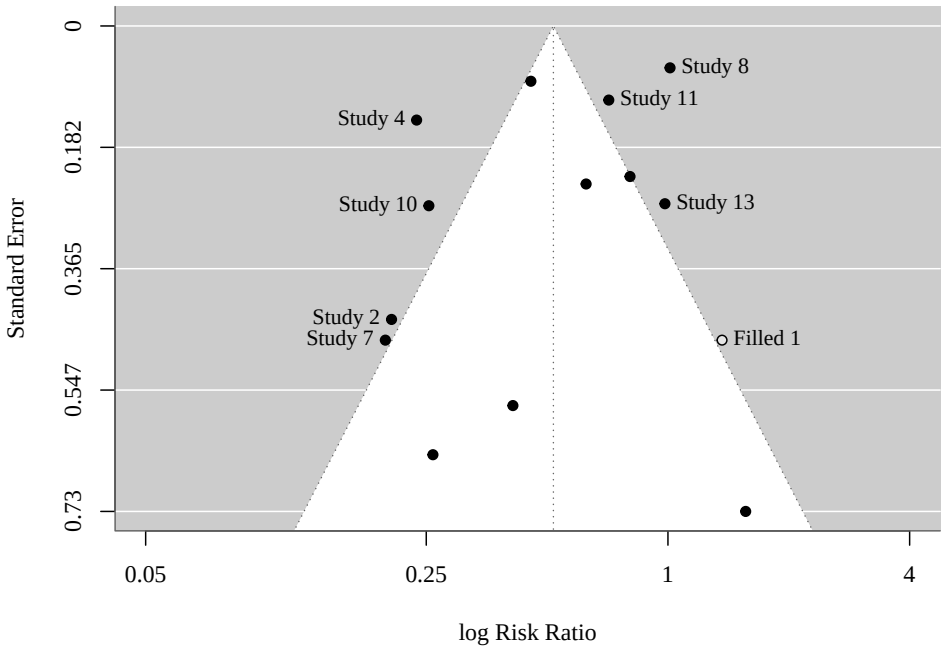


图 8.3: Trim-and-fill 后的漏斗图。空心点为算法填补的镜像研究。BCG 数据仅 1 项填补，对合并估计的影响有限。

定理 8.1 (雷区)

漏斗图不对称的解读是 meta 分析最容易过度推断的环节。不对称的可能来源至少有四类：发表偏倚（阴性结果不发表）；真实异质性（小样本研究恰好在某个特殊亚组里，与大样本研究估的不是同一效应）；方法学差异（小样本研究质量低、随机化不严，效应估计偏倚）；偶然（k 小时随机散布看起来不对称）。把所有不对称都归因于发表偏倚是 trim-and-fill 的内在假设，但临床现实里这四类来源经常并存。诊断方法是配合 contour-enhanced 漏斗图、子集分析、对小样本研究做单独的方法学评估，逐一排除其他来源后再下”发表偏倚”的结论。 ♡

8.6 发表偏倚不显著但仍要警惕的场景

BCG 数据上 Egger、Begg 都不显著，trim-and-fill 仅填补 1 项，从统计指标看”没有强发表偏倚”。但这个结论有两个限定。第一，$k = 13$ 让所有检验功效不足，”未拒绝原假设”不是”接受原假设”。第二，BCG 数据是 1948–1980 年间的早期 RCT，那个年代发表偏倚的强度可能不像现代医学文献那样有系统性研究——很多 1950 年代的 BCG 阴性结果可能根本没有进入文献库（即使作者愿意投稿，期刊体系也未建立完整）。所谓”漏斗对称”只能说”在已知文献库内对称”。

回到本数据。本书把 BCG 当作教学数据，结论”无强发表偏倚证据”足够支撑后续章节的合并分析继续进行。如果是真做政策决策的系统综述，应当配合灰色文献检索（会议摘要、技术报告、试验注册库 ClinicalTrials.gov）补全研究池，再重新评估漏斗对称性。

第 8 章到这里收束。Egger 与 Begg 检验在 BCG 数据上都不显著，trim-and-fill 仅填补 1 项研究，校正后合并 RR 从 0.49 变到 0.52，临床意义上几乎一致。但 $k = 13$ 的检验功效有限，结论的严谨表述是”证据不足以指控偏倚”，不是”证明没有偏倚”。

第 9 章接着查另一个稳健性问题：如果去掉 13 项研究里的某一项，主结论是否还成立？这是**敏感性分析**。

第 9 章 敏感性分析与影响诊断

内容提要

- 学会用 leave-one-out 检验合并结论是否被某项研究主导
- 用累积 meta 分析观察证据如何随时间累积、何时趋于稳定
- 掌握影响诊断的核心指标：Cook's distance、DFFITS、studentized residuals
- 在 dat.bcg 上识别影响最大的研究并讨论稳健性
- 区分"影响大"与"应该剔除"，避免事后剔除研究

第 4 至第 8 章给出了 BCG 数据的合并估计、异质性诊断、协变量解释、发表偏倚检验。本章是审稿人的"最后一关"：如果删掉某一项研究，主结论是否还成立？证据累积到什么时候已经足以下结论？哪项研究在统计上最有影响力？这三个问题对应的工具是 leave-one-out 分析、累积 meta 分析、影响诊断。

9.1 Leave-one-out 分析

Leave-one-out 是最直接的稳健性检验：每次去掉一项研究，用剩余 $k-1$ 项重新合并，看合并 RR 与 CI 变化多少。如果某一项研究的去除让结论方向翻转或显著性消失，就说明合并结论被这项研究主导，证据脆弱。

定义 9.1 (Leave-one-out 合并)

对每项研究 $i \in \{1,\dots,k\}$，去掉该项后用剩余 $k-1$ 项跑随机效应模型，得到 k 个合并估计 $\hat{\mu}_{(-i)}$ 与对应 95% CI。把这 k 个估计排成 forest plot，能直观看出"哪项研究的去除让合并漂移最大"。 ♣

```
loo <- leave1out(re, transf = exp)
print(loo)
```

BCG 数据的 leave-one-out 结果

完整 13 项合并 RR $= 0.489$。13 个 LOO 合并 RR 跨度从 0.452 到 0.533，所有 95% CI 都不跨过 1，结论"BCG 有保护效应"在去掉任意一项后仍然成立。

最有影响力的两项是去掉 8 号 TPT Madras 1980（RR 从 0.489 下降到 0.452，保护效应增强）与 4 号 Hart & Sutherland 1977（RR 从 0.489 上升到 0.533，保护效应削弱）。这两项分别是"反向极端"（赤道附近 RR ≈ 1）与"正向极端"（高纬度 RR ≈ 0.24），它们的去除导致合并估计向相反方向漂移。但漂移幅度都在 0.04–0.05 范围内，临床决策上相同。

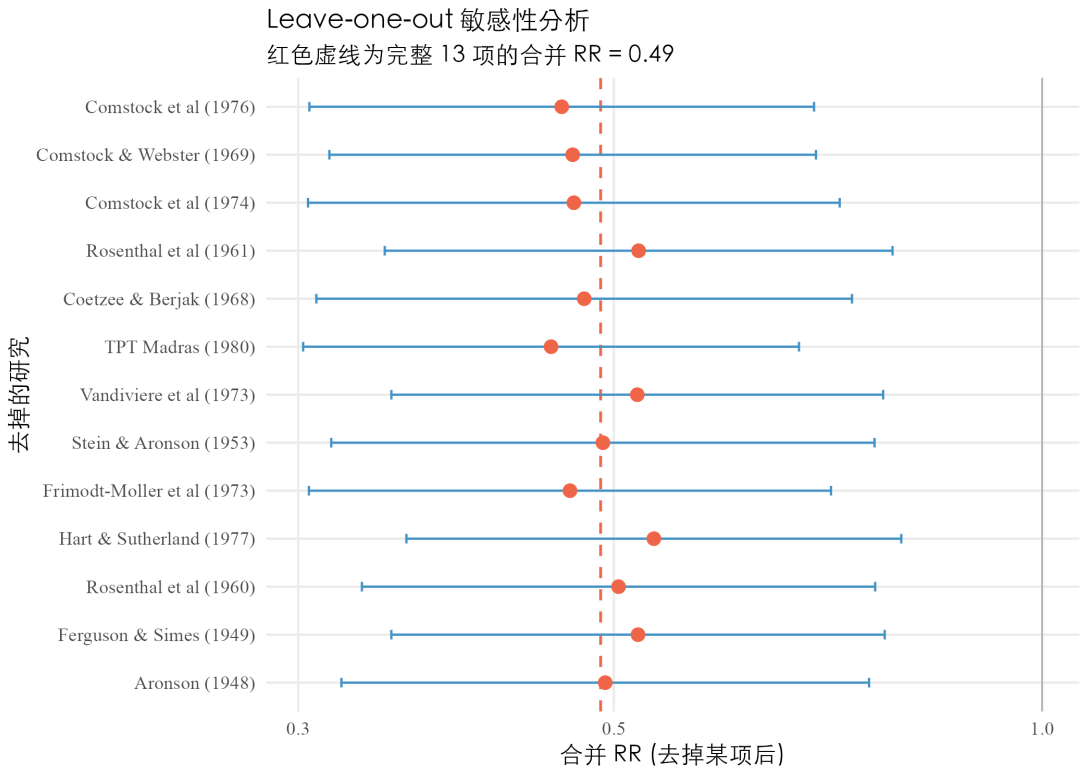


图 9.1: Leave-one-out 敏感性分析。每一行是去掉对应研究后的合并 RR 与 95% CI。红色虚线为完整 13 项合并 RR = 0.49。所有 LOO 估计都在 CI 内徘徊，结论稳健。

9.2 累积 meta 分析

累积 meta 是另一种稳健性检验：按某种顺序（通常是发表年份）逐项加入研究，每加一项重新跑合并，看证据累积过程。这种分析能回答两个问题：什么时候证据已经足以下结论？后续研究是确认了早期发现还是颠覆了早期发现？

定义 9.2 (累积 meta 分析)

把研究按时间顺序排列，从最早的 1 项开始，逐项加入新研究，每次都跑一次随机效应合并。第 i 步给出"截至时间 t_i 已有证据"的合并估计。把累积估计画成 forest plot，能直观看到证据何时收敛、何时出现转折。

♣

```
cum <- cumul(re_sorted, transf = exp)
print(cum)
```

BCG 数据的累积 meta

按发表年份累积合并：

累积 k	最新加入研究	年份	合并 RR	95% CI	τ^2	I^2
1	Aronson	1948	0.41	(0.13, 1.26)	0	0%
2	Ferguson & Simes	1949	0.27	(0.004, 19.20)	0	0%
3	Rosenthal 1960	1960	0.38	(0.14, 1.01)	0.087	41%
5	Frimodt-Moller	1973	0.33	(0.21, 0.52)	0.086	48%
8	TPT Madras	1980	0.45	(0.28, 0.74)	0.178	72%
13	Comstock 1976	1980	0.49	(0.33, 0.73)	0.313	92%

累积到 1973 年（5 项研究，全部为高/中纬度），合并 RR $= 0.33$，CI $[0.21, 0.52]$，看起来 BCG 是个相当强的疫苗。1980 年加入 TPT Madras 这项低纬度大样本无效研究后，合并 RR 立刻从 0.33 升到 0.45，I^2 从 48% 跳到 72%，异质性指标显著恶化。如果一个 1973 年的政策制定者读到当时的累积 meta，会得到一个相当乐观的结论；但 1980 年的更新会立刻打破这个乐观。这是累积 meta 最有价值的地方——它揭示了”我们以为的稳定结论何时被新证据颠覆”。

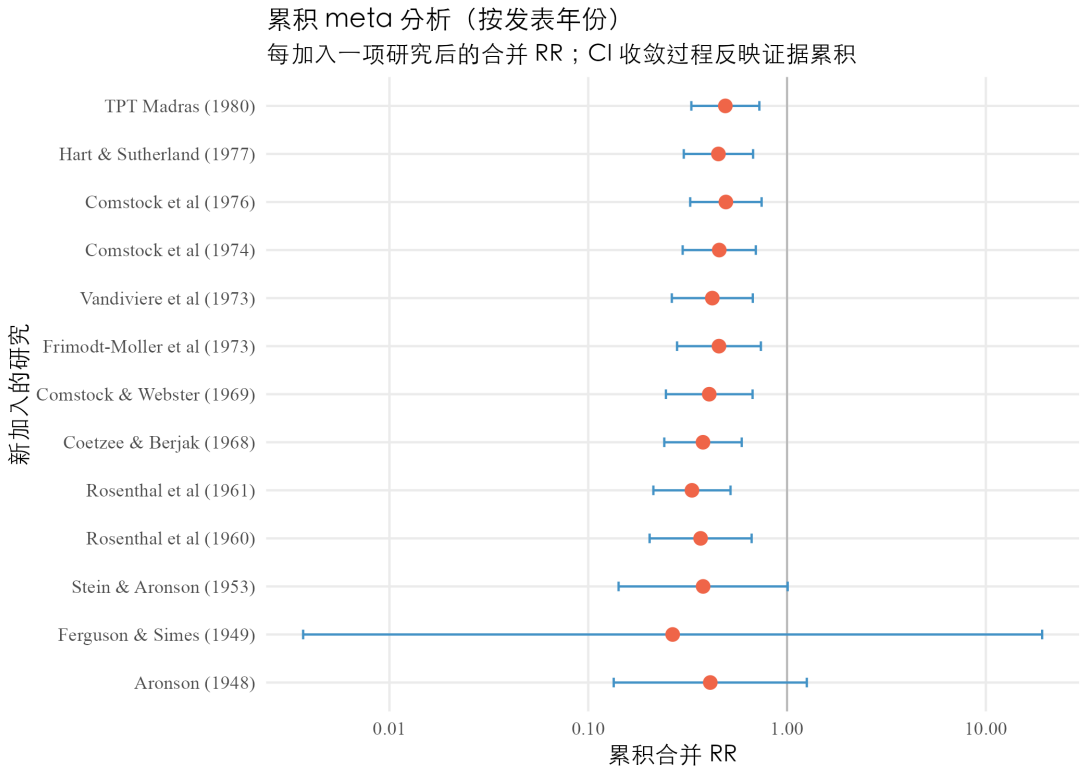


图 9.2: 累积 meta 分析。每一行是截至该研究为止的合并 RR 与 95% CI。证据累积过程显示 1973 年前点估计稳定在 RR ≈ 0.33，1980 年 TPT Madras 加入后升到 0.45，最终稳定在 0.49。

9.3 影响诊断

回归分析里有一套成熟的影响诊断指标，能识别哪些观测对模型估计有不成比例的影响。Meta 分析借用了这套工具，每项研究类比线性回归里的一个观测。把这四个指标想象成一张回归桌：13 项研究是桌面上的 13 个砝码，合并均值是桌子的重心。

定义 9.3 (Meta 分析的影响诊断指标)

Studentized residual r_i^*：第 i 项研究偏离合并均值的标准化距离。直觉上是问”这个砝码离桌子重心有多远”。$|r_i^*| > 2$ 表示该研究偏离整体趋势两个标准差以上，需要核查。

Cook's distance D_i：把第 i 项研究拿走再放回，整张桌子的重心位置变化的总量。直觉上是问”这个砝码一拿走，桌子整体晃动多大”。D_i 大说明这项研究主导了重心。常用阈值 $D_i > 4/k$。

DFFITS DFFITS_i：与 Cook's D 类似，但只看合并均值这一个数值的标准化变化量，不考虑其他估计。直觉上是问”这个砝码一拿走，桌子重心在主轴上挪动几个标准差”。常用阈值 $|\mathrm{DFFITS}_i| > 3\sqrt{p/(k-p)}$，$p$ 是参数数。

Hat value h_i：第 i 项研究的”杠杆”。直觉上是问”这个砝码本身有多重、放在多远的位置”。砝码越重（权重越大）或者放得越远（离合并均值越远），杠杆越大。注意高 hat 不一定有偏倚，只说明该研究对模型有重要贡献。常用阈值 $h_i > 3p/k$。

♣

四个指标各有侧重：rstudent 看”研究偏离主流多远”，hat 看”研究本身有多大分量”，Cook's D 与 DFFITS 看”拿掉它整张桌子怎么晃”。一个研究 hat 大但 rstudent 小（在主流方向上的大研究），通常没问题；hat 小 rstudent 大（小研究但极偏离），影响有限；hat 大 rstudent 大（大研究又偏离），就是 Cook's D 与 DFFITS 同时报警的危险组合。

metafor 的 `influence()` 函数一次性输出全部影响诊断指标，并标记被任意阈值识别的”influential” 研究。

```
inf <- influence(re)
print(inf)
```

BCG 数据的影响诊断

全部 13 项研究的关键指标如下：

编号	作者	rstudent	DFFITS	Cook's D	hat
1	Aronson	−0.21	−0.04	0.002	0.051
2	Ferguson & Simes	−1.29	−0.34	0.112	0.064
3	Rosenthal 1960	−0.74	−0.16	0.026	0.044
4	Hart & Sutherland	−1.41	−0.50	**0.227**	0.097
5	Frimodt-Moller	0.85	0.27	0.075	0.089
6	Stein & Aronson	−0.12	−0.02	0.000	0.101
7	Vandiviere	−1.30	−0.34	0.108	0.060
8	TPT Madras	1.43	0.47	**0.197**	0.102
9	Coetzee & Berjak	0.41	0.14	0.022	0.087
10	Rosenthal 1961	−1.11	−0.34	0.116	0.084
11	Comstock 1974	0.67	0.23	0.057	0.099
12	Comstock & Webster	1.31	0.26	0.063	0.038
13	Comstock 1976	1.19	0.36	0.121	0.084

Cook's distance 最大的两项是 4 号 Hart & Sutherland 1977（$D = 0.227$）与 8 号 TPT Madras 1980（$D = 0.197$）。这与 leave-one-out 结果一致，正是这两项研究在合并均值的两侧产生最大拉拽。但所有 $|r_i^*|$ 都在 ± 2 以内，所有 D_i 都在 $4/13 = 0.31$ 以内，metafor 标记的”influential” 研究为空。BCG 数据上没有单一研究被诊断为统计意义的极端值。

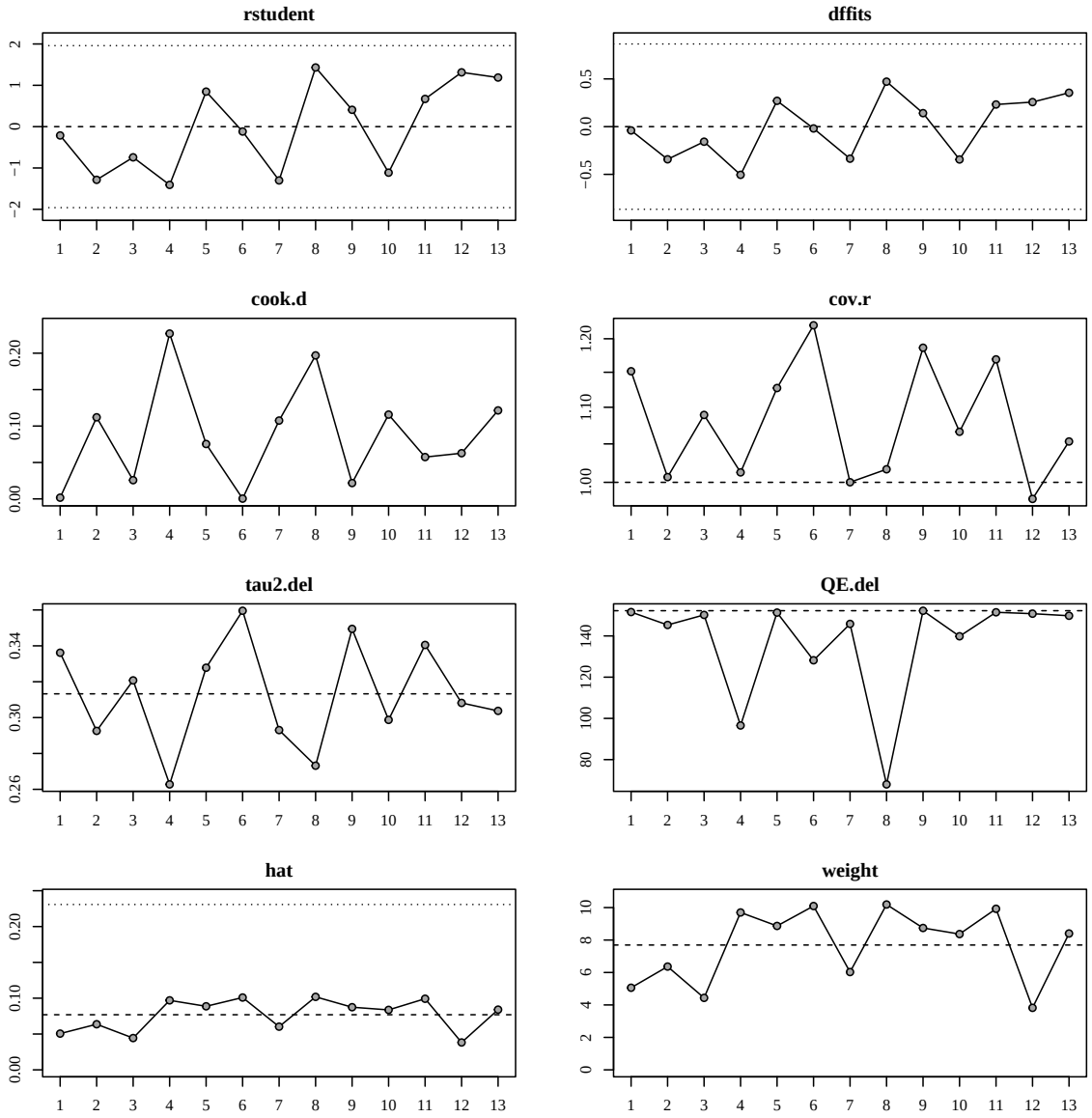


图 9.3: metafor 的 `plot.infl()` 输出。八个面板分别展示 rstudent、DFFITS、Cook's D、cov.r、tau²-deletion、QE-deletion、hat、weight。BCG 数据所有研究的指标都在常用阈值内。

定理 9.1 (雷区)

影响诊断最危险的误用是事后剔除研究换取更"好看"的结论。看到 4 号 Hart & Sutherland 1977 是 Cook's D 最大的研究，把它剔除会让合并 RR 从 0.49 升到 0.53，I^2 从 92% 降到 90%，看起来稳健性更好。这种做法在 Cochrane Handbook 里被明确禁止：影响诊断的目的是识别需要核查的研究，不是决定要剔除的研究。识别出影响大的研究后，正确流程是回到原文核查方法学：随机化是否严格、失访是否高、结局是否预先注册。如果方法学没问题，研究就保留，无论它对合并结果的影响多大。剔除研究必须基于预先指定的方法学标准，不能基于"它让我的结论更显著"。 ♡

第 9 章到这里收束。leave-one-out 表明 BCG 主结论高度稳健（13 项研究任意去掉一项后合并 RR 仍在 0.45–0.53 之间，所有 CI 都不跨过 1）；累积 meta 揭示了 1980 年 TPT Madras 加入后整体异质性显著上升的转折；影响诊断指标都在常用阈值内，没有单一研究主导结果。

第 10 章把所有方法的合并估计放进终极对比表，并用 GRADE 框架把这一份证据翻译成临床证据等级，给出"BCG 这份证据值多少"的最终答案。

第 10 章 方法对比与 GRADE 证据等级

内容提要

- 把全书九种合并方法的估计放进一张终极对比表
- 讨论方法之间结论的一致与分歧
- 学会用 GRADE 框架把统计估计翻译成临床证据等级
- 在 dat.bcg 上完整跑一遍 GRADE 五要素降级评估
- 讨论 BCG 数据的时代局限与现代证据更新方向

第 1 章到第 9 章把 13 项 BCG 试验在所有主流 meta 分析方法上各跑了一遍，本章把这些结果汇总，回答两个问题：方法之间结论的一致与分歧告诉了我们什么？统计意义上的合并估计如何翻译成临床决策的证据等级？

10.1 全书方法终极对比

把九种方法的合并 RR 与 95% CI 放进一张终极对比图。

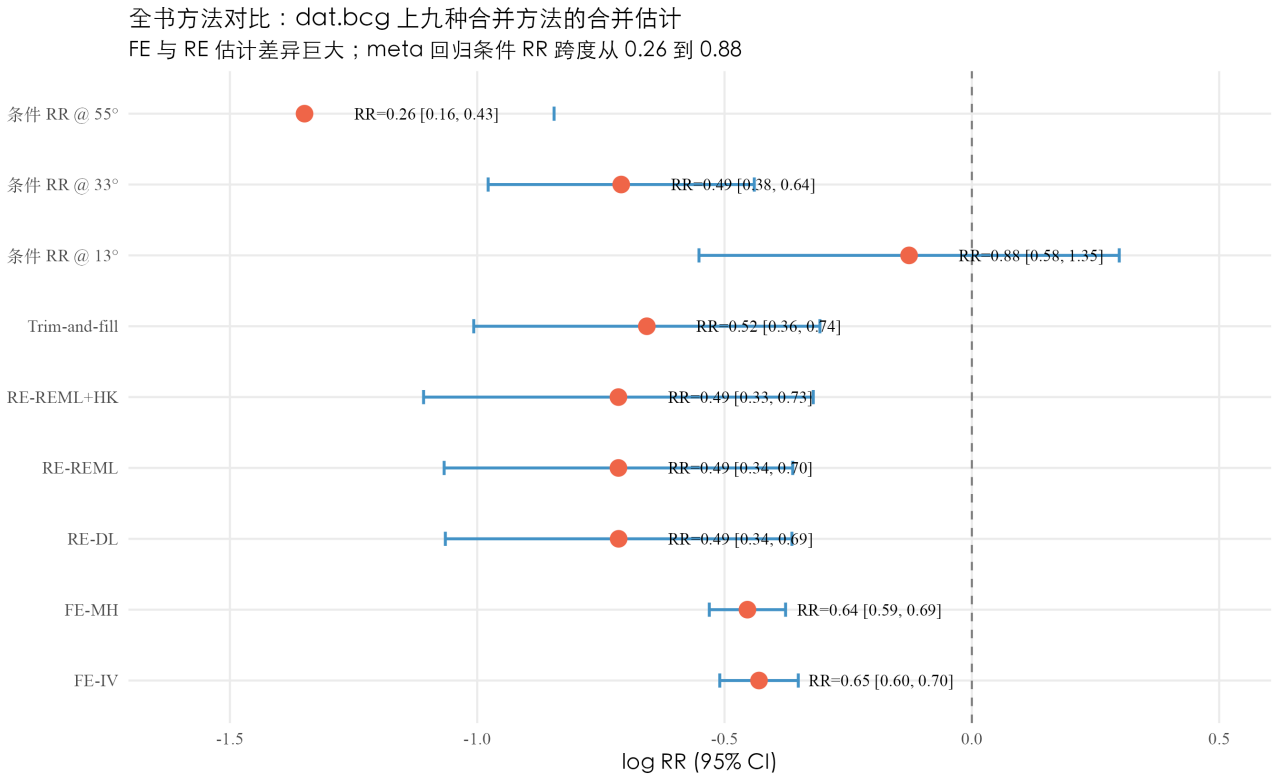


图 10.1: dat.bcg 上九种合并方法的合并估计与 95% CI 全景图。FE-IV/MH 给出 RR ≈ 0.65，三种 RE 估计给出 RR ≈ 0.49，trim-and-fill 校正后 RR $= 0.52$；meta 回归在不同纬度下的条件 RR 从赤道附近的 0.88 到北纬 55° 的 0.26。

10.1.1 结论的一致与分歧

读这张对比图能看到三层信息。第一层，所有方法都给出 BCG 有保护作用的结论：合并 RR 全部小于 1，CI 的上界（除去 13° 条件 RR 跨过 1 之外）也都小于 1。方向上 meta 分析的结论稳健。

第二层，FE 与 RE 估计在数值上相差甚远。FE-IV 给出 RR $=0.65$，看起来 BCG 把风险降低 35%；RE-REML+HK 给出 RR $=0.49$，看起来 BCG 把风险降低 51%。差异源自第 4 章和第 5 章讲过的机制：FE 模型让大样本研究主导（TPT Madras 占 41% 权重），把合并均值拉向 RR ≈ 1；RE 模型把权重摊平，让中小样本研究的强保护信号有更大发言权。$I^2 = 92\%$ 的现实下，RE 估计是更负责任的报告选择。

第三层，meta 回归的条件 RR 跨度极大：赤道附近 RR $=0.88$（CI 跨过 1），中纬度 0.49，高纬度 0.26。这意味着所谓” 全球平均 RR $=0.49$” 是个不可对应到任何具体人群的合并均值。临床决策应当基于” 在我所在地区的条件 RR”，不是” 全球平均 RR”。本章讨论 GRADE 时会回到这一点。

表 10.1: dat.bcg 全书方法终极对比

方法	logRR	RR	95% CI	适用场景
朴素合并	-0.488	0.61	—	不推荐，仅作对比基线
FE-IV	-0.430	0.65	(0.60, 0.70)	异质性低 ($I^2 < 25\%$) 时
FE-MH	-0.454	0.64	(0.59, 0.69)	同 FE-IV，稀疏事件更稳健
RE-DL	-0.714	0.49	(0.35, 0.70)	历史经典选择
RE-REML	-0.714	0.49	(0.34, 0.70)	现代默认选择
RE-REML+HK	-0.714	0.49	(0.33, 0.73)	Cochrane 推荐主分析
Trim-and-fill	-0.657	0.52	(0.37, 0.74)	发表偏倚校正
条件 RR @ 13°	-0.127	0.88	(0.58, 1.35)	赤道地区临床推广
条件 RR @ 33°	-0.709	0.49	(0.38, 0.64)	中纬度临床推广
条件 RR @ 55°	-1.349	0.26	(0.16, 0.43)	高纬度临床推广

10.1.2 该报告哪一个？

实务上 meta 分析论文必须选定一个” 主结果” 在摘要里报告，配合敏感性分析作支撑。本数据的主结果选择遵循三条标准：异质性 $I^2 = 92\%$ 显著，必须用随机效应；研究数 $k = 13$ 落在小样本区间，REML 比 DL 偏倚小且 HK 调整对小样本更稳健；异质性能被纬度解释 75.6%，meta 回归的条件 RR 比单一合并均值更有临床意义。综合下来，主结果应当是 **RE-REML+HK 给出的全球均值** RR $=0.49$ **（95% CI 0.33–0.73）**，配合 meta 回归报告 **在不同纬度下的条件 RR**。这是 Cochrane Handbook 与 PRISMA 报告规范的标准格式。

10.2 GRADE 证据等级框架

统计估计回答的是” 数据告诉了我们什么”。临床决策还要回答” 我们应该多大程度上相信这个结论”。Grading of Recommendations Assessment, Development and Evaluation 框架，简称 GRADE，是把 meta 分析的统计估计翻译成临床证据等级的事实标准 [2]。

定义 10.1 (GRADE 证据等级)

GRADE 把证据等级分为四档：高（⊕⊕⊕⊕）、中（⊕⊕⊕◯）、低（⊕⊕◯◯）、极低（⊕◯◯◯）。RCT meta 分析起点为高。从高开始，按五个降级因素（偏倚风险、不一致性、间接性、不精确性、发表偏倚）逐项评估，每命中一项可降一级。也存在三个升级因素（大效应、剂量反应、混杂会减弱观测效应），但通常只用于观察研究。

10.2.1 五个降级因素在 dat.bcg 上的评估

偏倚风险（Risk of Bias）：13 项研究中 7 项标准随机化，2 项交替分配，4 项系统分配。后两类是准随机方法，按 Cochrane RoB 2 工具的” 随机化过程” 域评估属于” 有些担忧” 或” 高风险”。整体而言 BCG 数据的偏倚风险**中等偏高**，应当降一级。

不一致性（Inconsistency）：$I^2 = 92\%$，$\tau^2 = 0.31$，预测区间 [0.14, 1.76] 跨过 1。异质性极强，应当降一级。第 7 章已经用纬度解释掉 75.6%，剩余 $\tau^2_{\text{res}} = 0.076$ 仍然不可忽略。如果按" 分层后再评级" 的做法，可以分别报告高/中/低纬度的证据等级，每一层内的异质性可能小很多。

间接性（Indirectness）：13 项研究的人群以白人为主（早期欧美研究）+ 南亚（TPT Madras）+ 北美土著（Aronson、Stein-Aronson）。BCG 株系也不完全统一（部分研究用法国 Pasteur 株，部分用丹麦 Copenhagen 株）。结局都是结核病发病，但定义略有差异（部分仅肺结核，部分含肺外结核）。整体间接性**中等**，是否降级取决于读者关心的具体临床问题。

不精确性（Imprecision）：合并 RR = 0.49 的 95% CI [0.33, 0.73]，区间排除了 RR = 1（无效），也排除了 RR > 0.8（弱保护）。结论"BCG 有显著保护作用" 在统计意义上精确。**不降级**。

发表偏倚（Publication Bias）：第 8 章给出 Egger $p = 0.19$、Begg $p = 0.95$、trim-and-fill 仅填补 1 项。但 $k = 13$ 让所有检验功效不足。保守起见可以认为发表偏倚证据**不强**，**不降级**，但报告时应当注明检验功效有限。

10.2.2 综合评级

从 RCT 的" 高" 起点，命中两条降级因素（偏倚风险一级、不一致性一级），最终 **证据等级为低（⊕⊕◯◯）**。临床决策上，BCG 对结核的保护效应在全球平均尺度上" 低等级证据支持"，但在分层后的高纬度亚组（RR = 0.26），证据等级可以提升到中等。

表 10.2: dat.bcg 的 GRADE 评级

因素	降级	依据
起点	高	13 项 RCT meta 分析
偏倚风险	−1	6 项准随机分配，按 RoB 2 偏倚风险中等偏高
不一致性	−1	$I^2 = 92\%$，PI 跨零线
间接性	0	人群与株系有差异但与目标问题相关
不精确性	0	95% CI 排除无效与弱保护
发表偏倚	0	检验不显著但 $k = 13$ 功效有限
最终等级	**低（⊕⊕◯◯）**	

定理 10.1 (雷区)

GRADE 评级最常见的误用是把统计显著当成" 高证据等级"。BCG 数据合并 RR 的 95% CI 排除了 1，统计上是" 显著有效"，但 GRADE 评级仍然只有" 低"。统计显著只对应不精确性这一项不降级，偏倚风险与不一致性两项已经把等级压到低。临床读者看到 $p < 0.001$ 就直接相信结论的做法是不分级的危险简化。GRADE 的价值在于把" 统计显著" 与" 临床可信" 分成两件事独立评估。 ♡

10.3 时代局限与证据更新

dat.bcg 的 13 项研究全部完成于 1948–1980 年。70 多年过去，结核病学、疫苗学、流行病学方法学都发生了巨大变化。本数据集作为教学样本极其经典，但作为" 今天该不该打 BCG" 的政策依据，已经过时。现代 BCG 证据来自三个新方向：

第一，对原始 RCT 的长期随访延展。Hart & Sutherland 在 BMJ 1977 之后又有多次随访更新，最长追踪到接种 60 年后。结论是 BCG 的保护效应随接种后时间逐渐减弱，但在前 10–15 年仍然显著。

第二，IPD（个体患者数据）meta 分析重新分析原始数据。Mangtani 2014（BMJ）对部分 BCG 试验做了 IPD-MA，能调整年龄、随访时间等个体协变量，得到比聚合数据 meta 分析更精细的估计。

第三，BCG 与新型结核疫苗的对比。M72/AS01E 等新疫苗在 2018–2024 年间陆续报告 II/III 期试验结果。Meta 分析的更新版本会把 BCG 与新疫苗的对比纳入证据体系，回答”BCG 是否仍是首选” 的政策问题。

这本书用 dat.bcg 作为教学数据集是为了让 meta 分析方法的演练有一份完整、干净、临床故事性强的数据。读者掌握了本书的方法之后，可以把同一套工具应用到任何当代的系统综述上。

10.4 终结对比表

表 10.3: dat.bcg 全书方法终极对比表

方法	logRR	RR (95% CI)	τ^2	说明
朴素合并	−0.488	0.61	—	不调整研究层级，反例基线
FE-IV	−0.430	0.65 (0.60, 0.70)	0	假设违反 ($I^2 = 92\%$)
FE-MH	−0.454	0.64 (0.59, 0.69)	0	同 FE-IV，稀疏稳健
RE-DL	−0.714	0.49 (0.35, 0.70)	0.309	矩估计
RE-REML	−0.714	0.49 (0.34, 0.70)	0.313	似然估计
RE-REML+HK	−0.714	0.49 (0.33, 0.73)	0.313	**主分析（推荐）**
Trim-and-fill	−0.657	0.52 (0.37, 0.74)	—	填补 1 项，结论稳健
LOO 范围	—	0.45–0.53	—	全部不跨 1，结论稳健
Meta 回归（ablat）	—	13°: 0.88; 33°: 0.49; 55°: 0.26	0.076	$R^2 = 75.6\%$，临床推广首选
GRADE 等级	低（⊕⊕◯◯）			

全书到这里收束。从第 1 章 13 项试验答案矛盾的图景，走到第 10 章 GRADE ” 低” 等级的临床证据评级，主线是清晰的：朴素合并失败 → FE 模型按精度发言但 $I^2 = 92\%$ 假设违反 → RE 模型承认研究本来就不一样 → 异质性诊断量化研究间方差 → 纬度解释了 75.6% 的差异 → 发表偏倚检验未拒绝原假设 → 敏感性分析支持稳健性 → GRADE 综合评级为低。

读者掌握了这一套工具之后，可以把它应用到任何当代的系统综述上。换一份数据、换一个临床问题，方法学骨架不变。这正是 meta 分析作为一种” 对证据的统计整理” 的核心价值。

Bibliography

[1] Graham A. Colditz et al. “Efficacy of BCG vaccine in the prevention of tuberculosis: meta-analysis of the published literature”. In: *JAMA* 271.9 (1994), pp. 698–702.

[2] Gordon H. Guyatt et al. “GRADE: an emerging consensus on rating quality of evidence and strength of recommendations”. In: *BMJ* 336.7650 (2008), pp. 924–926. DOI: 10.1136/bmj.39489.470347.AD.

[3] Matthew J. Page et al. “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews”. In: *BMJ* 372 (2021), n71. DOI: 10.1136/bmj.n71.

[4] Jonathan A. C. Sterne et al. “RoB 2: a revised tool for assessing risk of bias in randomised trials”. In: *BMJ* 366 (2019), p. l4898. DOI: 10.1136/bmj.l4898.