试验序贯分析（TSA）

工作者手册

User manual for

Trial Sequential Analysis (TSA)

原著：**Kristian Thorlund, Janus Engstrøm, Jørn Wetterslev, Jesper Brok, Georgina Imberger, and Christian Gluud**

译者：刘建平、李迅、刘兆兰、张颖、李文元、任君、杨国彦、王禹毅、王聪聪、柴倩云、陈妮妮、张抗、王丽琼、梁宁、苏春香、周芬、张云皎、熊文婧、孙瑾

免责申明

本软件自发行不再更改,不附有任何明确或暗示的，包括但不仅限于软件适用于销售、特殊目的和不侵犯任何专利与法权的担保。任何情况下软件作者或版权所有者不对任何索赔、损害或其他责任负责，无论协议行为、侵权行为以及由连接、使用或其他与本软件有关的处理方式引起的、产生的或在这些过程中发生的各种行为。

在任何情况和法律理论下，无论在侵权行为中、协议内还是其他情况中，哥本哈根临床试验中心（Copenhagen Trial Unit）都不对您或他人的利益、信誉损失及任何间接的、特殊的、偶然的或必然的损害以及任何人的重大疏忽造成的损害负责，包括但不仅限于信誉损失、工作中断、电脑系统失效或硬件故障及其他任何损失造成的损害。

本手册与试验序贯分析（The Trial Sequential Analysis software，TSA）软件测试版配套。哥本哈根临床试验中心（Copenhagen Trial Unit）已经对 TSA 软件进行了详细的测试，但仍然可能有错误发生。使用者反馈是纠错过程和实现其他改进的一个重要部分，因此我们鼓励您告诉我们您使用这款软件的体验，您可以通过发送反馈到 tsa@ctu.dk 来实现与我们的沟通。

团队成员的分工和贡献

丹麦哥本哈根的哥本哈根临床试验中心（Copenhagen Trial Unit）开发了 TSA 软件。开发团队成员包括 Kristian Thorlund (KT)，Janus Engstrøm (JE)，Jørn Wetterslev (JW)，JesperBrok (JB)，Georgina Imberger (GI)和 Christian Gluud (CG)。每个团队成员的分工与贡献列述如下：

项目经理：KT
主要软件应用程序开发者：JE
合作软件应用程序开发者：KT、JW、JB、CG
统计程序员：KT
内部测试人员：JB、GI、JW、KT、CG
手册作者：KT（主要）、GI、JW、JB、JE、CG
工程监理：JW 和 CG

哥本哈根 DK-2100
国王医院 3344 部
电话：+45 3545 7171 传真：+45 3545 7101
电子邮箱：tsa@ctu.dk

翻译、审校：
北京中医药大学循证医学中心：
刘建平、李迅、刘兆兰、张颖、李文元、任君、杨国彦、王禹毅、王聪聪、柴倩云、陈妮妮、张抗、王丽琼、梁宁、苏春香、周芬、张云皎、熊文婧、孙瑾

前言

本手册对哥本哈根临床试验中心（Copenhagen Trial Unit）研发的试验序贯分析（Trial Sequential Analysis, TSA）软件使用提供了理论与实际操作的指导。第一章介绍了试验序贯分析的相关概念与理论，第二章提供了实践运用中几种具体方法的技术概述，第三到五章则介绍了软件的安装、使用和实际应用。

TSA软件可从哥本哈根临床试验中心网站www.ctu.dk/tsa下载。您可以免费使用该软件进行累积性meta分析，然而在您的分析报告或发表物中，请对软件以及我们对方法进行介绍的相关文章进行恰当的援引。

关于 TSA 软件的使用，如果您需要任何帮助，请通过 email tsa@ctu.dk 与我们联系。

目录

免责申明 1
团队成员的分工和贡献 2
前言 3
1. 试验序贯分析的概念和原理 6
1.1 Meta-分析中的随机误差 6
1.2 定义证据强度—信息量 7
1.3 达到信息量之前的统计显著性检验 8
1.4 达到信息量之前的无效性检验 8
1.5 概要 10
2 TSA 的方法学 11
2.1 综合多个临床试验结果的方法 11
2.1.1 二分类变量和连续性变量的效应值测量 11
2.1.2 固定效应模型和随机效应模型设置 12
2.1.3 随机效应模型 meta-分析方法 13
2.1.4 零事件试验的处理方法 14
2.2 累积 Meta 分析中校正的显著性检验和无效性检验 15
2.2.1 确定性 Meta 分析结论所需的信息量 16
2.2.2 累积检验统计量（Z 曲线） 20
2.2.3 Meta 分析中显著性检验存在的问题 21
2.2.4 α-消耗函数和试验序贯监测界值 22
2.2.5 试验序贯分析中校正的可信区间 26
2.2.6 迭代对数法则 28
2.2.7 β 消耗函数与无效界值 29
3 TSA 软件的安装和启动 31
3.1 安装条件 31
3.2 安装 32
3.3 启动 TSA 程序 32
3.3.1 TSA 不能正常启动的原因 32
3.4 启动 Revman 5.0 (RM5)转换器 33
4 如何使用 TSA 33
4.1 启动 33
4.1.1 创建一个新的 Meta 分析 33
4.1.2 保存和打开 TSA 文件 35
4.1.3 从 RevMan5 文件中导入 Meta 分析数据 35
4.2 添加、编辑和删除试验 38
4.2.1 添加试验 38
4.2.2 编辑和删除试验 39
4.3 自定义 Meta-分析设置 40
4.3.1 选择关联测量 40
4.3.2 选择统计模型 40
4.3.3 选择零事件数据处理方法 41
4.3.4 选择可信区间的类型 41

4.4 校正后的显著性检验应用（应用 TSA）.....42
4.4.1 添加显著性检验.....42
4.4.2 编辑和删除显著性检验.....48
4.4.3 添加和加载显著性检验模板.....48
4.4.4 显著性检验的运算.....49
4.5 TSA 图示选项.....50
4.6 探索试验间的差异性.....53
5 TSA 应用实例.....53
5.1 数据集.....54
5.2 避免假阳性.....54
5.3 阳性结果的确证.....55
5.3.1 “答案是”的确证.....55
5.3.2 避免过早的高估.....57
5.4 无效性检验.....59
5.5 估算一个新的临床试验的样本量.....60
5.6 其它已发表的试验序贯分析的应用实例.....62
6 附录.....64
6.1 二分类变量和连续性变量 Meta 分析中的效应量估计.....64
6.2 随机效应的处理.....65
6.2.1 Biggerstaff-Tweedie 法的公式.....65
6.3 试验序贯分析.....65
6.3.1 多重显著性检验夸大 I 型错误.....65
6.3.2 未应用 TSA 软件的其它备选方法.....66
7 缩略语和统计符号列表.....67
7.1 常用缩略语.....68
7.2 统计符号.....68
7.2.1 小写字母符号.....68
7.2.2 大写字母符号.....69
7.2.3 希腊字母符号.....71
参考文献.....73

1. 试验序贯分析的概念和原理

1.1 Meta-分析中的随机误差

Meta-分析的某些“阳性”结果可能源于机遇（随机误差）的作用，而非干预措施潜在的“真实”效应 [1-10]。同理，某些中性或“阴性”（“非阳性”）Meta-分析结果也可能由于欠缺统计学把握度和精确度，呈现“机遇产生的结果” [9-13]。这两种类型的误差即常见的假阳性结果（或 I 类错误）和假阴性结果（或 II 类错误）。Meta-分析的结果往往通过一些统计检验（检验统计量），结合 P 值或相应的可信区间，判断为“阳性”或“阴性”。

当某个 Meta-分析中的试验数目较小且患者样本量较小时，随机误差就可能导致出现错误的结果 [1;2;4-6;9;11;12;14;15]。相反地，如果患者数目很大，那么检验统计量和干预措施效应估计值通常会趋向于“真实值”[1;2;4-6;9;11;12;14;15]。图 1（A）和 1（B）为检验统计量这种趋势的一个实例。在这两种情况下，在某个较早的阶段推断统计量具有统计显著性是错误的，但是最终均趋向于“真实的”统计显著性一侧。

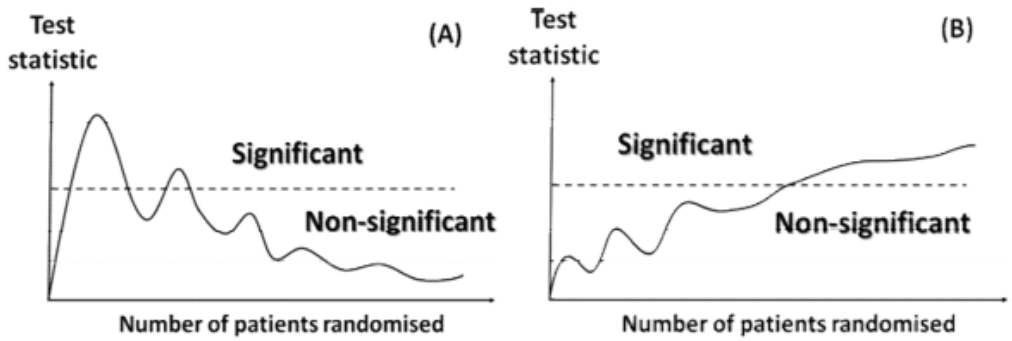


图 1 实例：在两项随机临床试验 A 和 B 中，随着纳入试验进行随机分配的患者数目（Number of patients randomised）变化，结局测量后（例如：死亡），检验统计量（Test statistic）趋势变化。（Significant：显著的，Non-significant：非显著的）

只有当随机误差或不精确性“足够大”而使统计推断出现错误时，在统计学检验（和干预措施效应值估计）的过程中随机误差和不精确性才会带来问题。如图 2(A)所示，在 X_1 和 X_3 时的统计学检验，将导致错误推断结果为具有统计显著性（如假阳性结果），但在 X_2 和 X_4 检验时则不会出现这种情况。因此，只有在 X_1 和 X_3 检验时，随机误差才“足够大”，以致出现错误的具有统计显著性的结果。如图 2(B)所示，在 X_1 和 X_2 时的统计学检验，将导致错误推断所观察的干预措施效应值不具有统计显著性（即假阴性结果），但在 X_3 和 X_4 检验时则不会出现这种情况。因此，只有在 X_1 和 X_2 检验时，这种不精确性才足以导致出现错误的不具有统计显著性的结果。

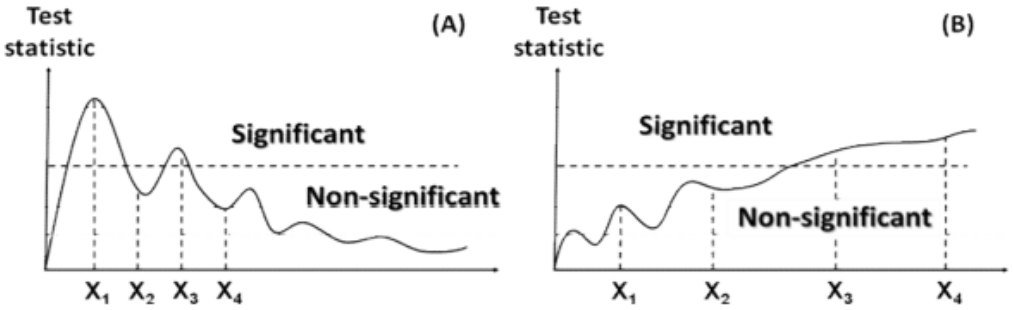


图2 实例：两项随机临床试验A和B中，假阳性和假阴性统计检验结果随时间发生变化。（Test statistic：检验统计量，Significant：显著的，Non-significant：非显著的）

数据累积性分析过程中采用的统计检验越多，得到假阳性或假阴性结果的几率就越高。这种现象通常称为*‘重复统计学意义检验导致的多重性’*[10;16-18]。

在Meta-分析中，将得出假阳性或假阴性结论的风险降到最低是非常重要的[3]。Meta-分析中，干预措施的合并效应值通常采用P值进行检验。Meta-分析人员必须决定P值的阈值，一个足够小的在此阈值之下的P值，才能表明“阳性”的结论。低于该阈值，该结果被认为具有统计显著性。在特定的时候，任何阈值都将决定着观察到假阳性结果（I类错误）或假阴性结果（II类错误）的风险。例如，如果图2中统计显著性的阈值（水平虚线）上移，那么观察到假阳性结果（图2(A)）的几率会降低，而观察到假“阴性”结果（图2(B)）的几率则会增加。当传统的显著性检验在较“早”阶段和/或多次进行时，这些累积的最大化的风险将使得结果失真（如图2所示）[16-18]。因此，任何统计显著性推断都必须结合证据强度进行推断。证据强度应当根据纳入研究中应统计的患者数目、观察到的事件数目和多重性的影响进行评估[1;2;4;6;10;19-21]。

1.2 定义证据强度—信息量

随机对照试验的Meta-分析增加了干预措施效应估计值的把握度和精确度[13]。如果系统综述和Meta-分析纳入了所有现有的临床试验，那么这些系统综述和Meta-分析可被视为现有的最佳证据[13]。然而，“现有的最佳证据”并不等同于“充分的证据”或“有力的证据”[1;2;4;6;11;12]。

如果单个随机对照试验使用了二分类变量，则需要估计所需的事件发生数量和患者人数，以得出可靠的统计推断。也就是说，需要进行样本量估算以确保“足够的”事件发生数量和纳入的患者人数[22]。进行Meta分析时也需要设定类似的“目标值”[1;2;6;23]。该目标值被称为Meta-分析所需信息量（information size, IS）或最佳信息量[1;2;4;6;11;12;14;15;19;23-25]。图3显示了A和B两种典型的Meta-分析情况，当达到所需信息量后统计量趋于稳定。

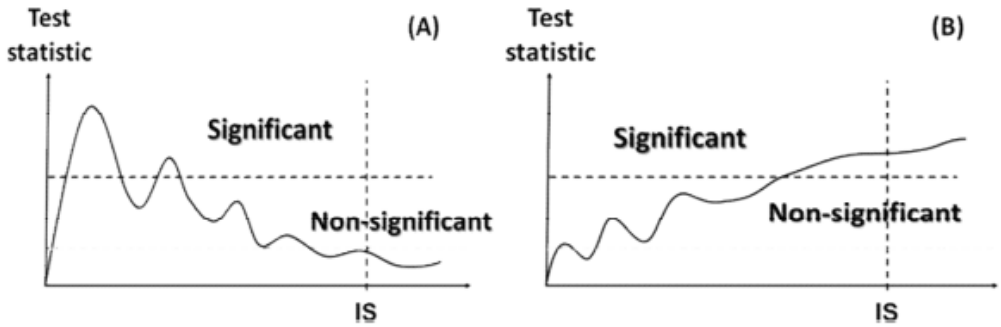


图3 实例：在A和B两项累积性Meta-分析中，所需信息量大小如何确保得出可靠的显著性检验结果。（Test statistic：检验统计量，Significant：显著的，Non-significant：非显著的，IS：所需信息量或最佳信息量）

单个临床试验的样本量估算通常基于期望的对照组事件发生率，期望的试验组干预措施相对危险度降低，以及期望的I类错误和II类错误的最大风险[26]。在Meta-分析中，可能存在由于试验人群、干预措施和方法不同而产生的异质性。因此，Meta-分析样本量估算需要进行校正——即提高，以确保异质性产生的方差在允许范围内[4;6;11;12;23]。这种校正类似于多中心临床试验中对各中心之间的差异进行校正[4;6;23]。

传统的Meta-分析方法，例如RevMan5.1.软件[27]，不考虑现有证据的数量[13]。反而理所当然地认为干预措施效应的统计显著性是可靠的，而不考虑应计入的事件发生数目和患者人数。相反地，干预措施效应未达到统计显著性则通常被视为不可靠，而进一步推测“需要更多的证据”[28]。

基于有限的事件发生数目和患者人数得出干预措施效应和P值的实验证据通常是不可靠的[1;2;4-6;9;11;12;29]。可能约有25%纳入小样本事件发生数目和患者人数的传统Meta-分析错误地将干预措施估计效应值视为具有统计学显著性[4;5]。有证据显示，在早期的阳性Meta-分析研究中，观察到的大量干预措施效应值的合并结果，在累积更多的证据后易于改变[4;5;9]。

1.3 达到信息量之前的统计显著性检验

Meta-分析的目的是尽早且尽可能可靠地鉴别出某干预措施的利弊[4;11-13;20]。因此，Meta-分析在出现新的临床试验后通常会进行更新。例如，Cochrane系统综述作者必须至少每两年对他们的系统综述进行更新[13]。Meta-分析更新时，结果也会反复接受显著性检验。在随机对照试验中，众所周知，对累积性数据反复进行显著性检验会扩大I类错误的总体风险[30]。模拟试验显示如果在Meta-分析中反复进行显著性检验且P值小于0.05即认为证据具有“统计显著性”，那么I类错误的实际风险在10%到30%之间[7;8;10;31]。如果应用某干预措施作为治疗手段的医疗决策是基于这样的结果而得出，将意味着10项这样的治疗决策中有1至3项是不合理的。

为解决这个问题，可以通过校正临界值的方式，即设定认为结果具有统计显著性和不具有统计显著性的临界值[1;2;4;6;11;12;14;15;24;25]。或者，可以根据证据强度和显著性检验进行的数量（“重对数律”）惩罚检验统计量[7;8]。TSA软件为上述两种方式提供了实施方法，每种方法均基于先进的概率论定理而建立。第一种方法采用随机对照试验中反复显著性检验的方法学（即统计监控界限值）[4;6;11;12]。第二种方法惩罚——即根据Meta-分析中现有信息量的强度和显著性检验进行的数量而降低检验统计量[7;8]。

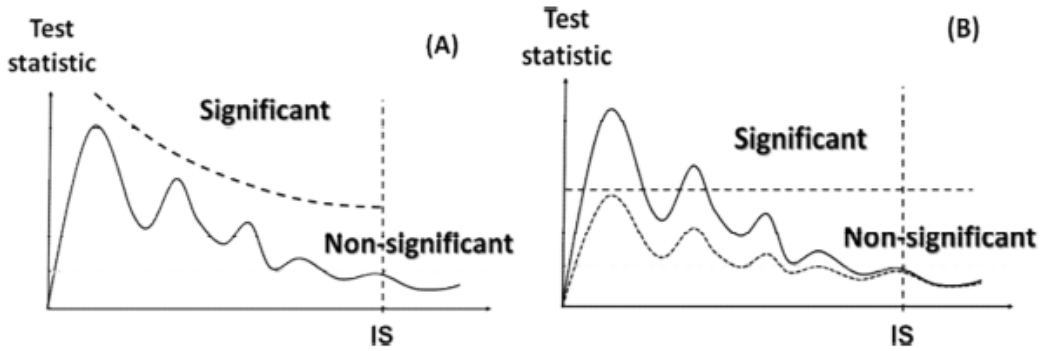
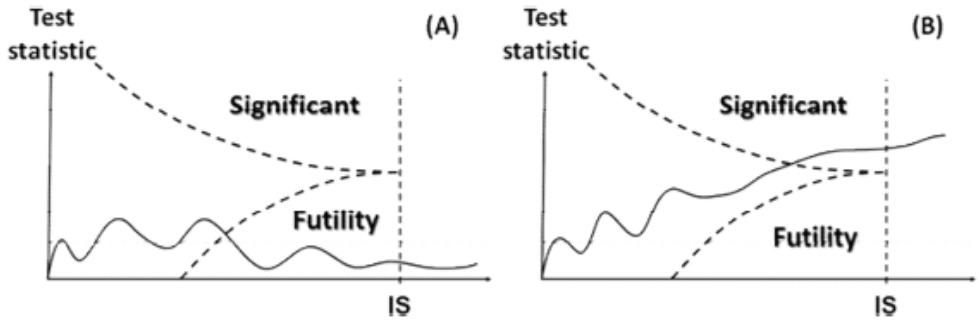


图4 实例：两项累积性Meta-分析A和B中，显著性临界值校正（约定的监控界限值）（A）以及校正检验统计量（约定的）（B），以避免出现假阳性统计检验结果。（Test statistic：检验统计量，Significant：显著的，Non-significant：非显著的，IS：所需信息量或最佳信息量）

图4（A）显示了Meta-分析中通过监控界限值校正统计显著性的临界值以避免出现假阳性结果的实例。图4（B）显示了Meta-分析中通过适当地校正检验统计量以避免出现假阳性结果的实例。

1.4 达到信息量之前的无效性检验


Test
statistic
(A)
Significant
Futility
IS
Test
statistic
(B)
Significant
Futility
IS

当干预不可能取得预期效应或者干预的效应量比临床上对患者有用的最小效应量更小时，可以用试验序贯分析软件来评估。Meta分析常用来指导未来的研究。在着手开始一个未来的临床研究前，研究者需要了解目前知识的准确总结。如果一个meta分析发现某干预措施无（重要）效应，研究者需要辨别该发现是因为把握度不够还是因为干预确实可能无效应。除非达到一个合适的信息量，通常我们认为“无效”是因为把握度不够。在某些情况下，我们可以提前得出治疗效应不可能达到预期那么大效应的结论，因此阻止研究者花费大量的资源在不需要的研究上。当然，预期效应可以从新考虑，未来的研究也可以探讨更小的效应值。

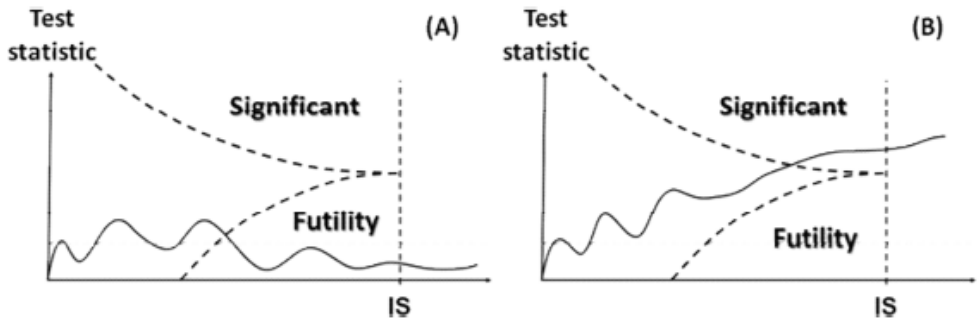


图5 试验干预不优于对照干预（已经开展了许多试验）（A）和试验干预统计学显著优于对照干预（已经开展了许多试验）（B）无效性界限的实例。（Test statistic：检验统计量，Significant：显著的，Futility：无效的）

试验序贯分析提供一种能尽早得出无效应结论的技术。最早被随机临床试验用来中期分析的“无效性界限”，现在已经被构想和用于为“无效”提供界值[30]。

如果试验干预确实优于对照干预，我们期望检验统计量围绕一个向上倾斜的直线波动，最后获得统计学意义（当 Meta 分析有足够的把握度）。当一个真正有效干预措施的 meta 分析纳入患者和事件都很小，获得有统计学意义结果的可能性会因为把握度不够而很低。然而，当积累更多的证据时，获得阴性结果的可能性就会降低。无效性界限是一系列界值的集合，反映了由于可能证据的强度而得到阴性结果风险的不确定性（例如患者数量的积累）。在界值之上，检验统计量可能因为把握度不够而没有统计学意义，但仍有机会在 meta 分析超过信息量前发现有统计学意义的效应。在界值之下，检验统计量非常小以至于几乎不可能发现有统计学意义的效应。在后一种情形下，对患者的随机化是无用的；干预措施不具有假定的效应。

图 5（A）举例说明了试验干预不优于对照干预。在超过需要信息量前，检验统计量穿过无效线界限（向上倾斜的凹形曲线）。图 5（B）举例说明了试验干预统计学显著优于对照干预。在该例中，检验统计量一直在无效性曲线之上（因为有潜在的效应），最后获得统计学意义。

1.5 概要

试验序贯分析是一种结合多种技术的方法。它把需要的证据量化，并提供需要信息量的具体数值。统计学显著性门槛根据证据的量化强度和多样性来校正[4;6;11;12]。无效性界值也通过一个相似的统计框架来建立。

总之，试验序贯分析能提供一个信息量，即干预效应的统计显著性界值以及无效性界值。用试验序贯分析方法得出的结论比传统meta分析的结论更可靠。经验证据显示信息量的考量和校正显著性界值能减少因为meta分析不精确和重复显著性检验造成的早期假阳性结果[4;6;11;12]。

另外一种方法是，根据证据的强度和显著性检验的样本量来校正检验统计量

[7;8]。模拟研究显示校正检验统计量能有效控制 meta 分析中的一类错误 [7;8]。

本手册接下来的内容包括哥本哈根试验中心试验序贯分析软件的理论与实践指南。第 2 章是试验序贯分析软件应用的方法学技术（中等水平）的介绍。3-5 章是描述怎样安装、使用和运用试验序贯分析软件。

2 TSA 的方法学

试验序贯分析结合了传统的 meta 分析方法、meta 分析样本量考量（例如需要信息量）和随机临床试验累计数据重复显著性检验的方法 [1;2;4;6;11;12]。在第 2 章，首先描述了多个试验数据合并的 meta 分析方法。2.1 描述了连续数据和分类数据的效应指标，meta 统计分析模型（固定效应模型和随机效应模型的一些变量），处理零事件数据的方法。2.2 描述了当随机误差的风险增加（证据质量低和重复显著性检验）时校正显著性水平的方法。这一章我们试图为读者提供关于第一章解决的问题一个中等水平的概念性理解，没有深入描述方法学的高级部分。

2.1 综合多个临床试验结果的方法

2.1.1 二分类变量和连续性变量的效应值测量

试验序贯分析项目利于二分类数据和连续数据的 meta 分析。二分类数据是指数据属于两个类别之一（例如死亡或生存）。连续数据是能用数值测量的数据（例如血压或生存质量分数）。每一类的数据，都有多种指标可用于干预效应比较 [13]。

分类数据效应指标

假设有 k 个独立的试验采用分类数据结局指标来比较两种干预措施（干预措施 A vs 干预措施 B）。这类试验一般会报告两组发生事件（例如死亡）的数量 e_A and e_B，和总的样本量 n_A and n_B,。对于分类数据，两组的干预效应可用 RD，RR，OR 来表示 [13]，下面的公式是指标的计算方法。

$$\mathrm{RD}=\frac{e_A}{n_A}-\frac{e_B}{n_B}$$

$$\mathrm{RR}=\frac{(e_A/n_A)}{(e_B/n_B)}$$

$$\mathrm{OR}=\frac{e_A/(n_B-e_B)}{e_B/(n_A-e_A)}$$

相对危险度和比值比常用对数的形式来表示，因为采用对数转化后数据能更好的用于统计分析（例如数据分布均匀和近似正态）[13]。也能通过对数形式的数据来获得标准差、方差和干预效应比的权重。附录 6.1 提供了 RD，log(RR)，和 log(OR)的标准误的计算公式。

当两组的事件发生率很低的时候（罕见事件数据，rare-event data），可采用 Peto’s 比值比来代替传统的比值比 [13]，计算公式如下：

$$\mathrm{ORPeto}=\exp(((e_A-E(e_A))/v))$$

公式中的$E(e_A)$是干预组A发生事件的期望值，v是e_A的方差（超几何分布）。附录6.1提供了$E(e_A)$和v的计算公式。

假设有k个独立的试验采用连续数据结局指标来比较两种干预措施（干预措施A vs 干预措施B）。这些试验通常报告两组指标的均数（例如，生活质量得分均数）m_A 和 m_B,均数标准差为sd_A and sd_B，两组的样本量为n_A和 n_B。当所有试

验的均数是采用相同标准测量，效果差异可以用均差（MD）来计算，MD=m_A - m_B. 均差的标准误采用下面的公式计算：

$$\mathrm{SE（MD）}=\sqrt{\frac{sd_A^2}{n_A}-\frac{sd_B^2}{n_B}}$$

当均数不是用相同标准测量时，可采用相同的标准使其标准化，这样多个试验的数据就能合并了[11]。常用方法是用每个试验的均数除以估计标准偏差，因此提供一个标准差形式的效应估计，平均差除以标准偏差得到标准化平均差（SMD）[13]。

试验序贯分析对于标准均差（SMD）的meta分析并没太大帮助。标准均差meta分析的校正显著性检验需要信息量计算，而信息量计算是建立在以标准差形式报告期望均差的基础上。临床医生一般不这样报告效应值，因此容易产生不现实的信息量需求。

2.1.2 固定效应模型和随机效应模型设置

假设有k个独立的试验。用Y_i表示第i个试验的观察效应值。对于分类数据meta分析，Y_i或是估计风险差、对数相对风险、对数比值比，或是对数peto比值比。对于连续数据meta分析，Y_i是第i个试验的估计均差。用μ_i表示第i个试验的真正效应，用μ表示整个meta分析人群的真正潜在干预效应。用σ_i^2表示第i个试验观察干预效应的方差（样本误差）。

对于固定效应模型，纳入试验的特征（包括患者纳入排除标准，干预的方式，试验设计，方法学质量，随访时间等）假定是相似的[13]。用数学公式表示为$\mu_1=\mu_2=\ldots\mu_k=\mu$。多个独立试验的观测干预效应假定满足分布$Y_i\sim N(\mu,\sigma_i^2)$。固定效应模型中每个试验的权重$w_i$根据每个试验的变异来确定，因此，固定效应模型的权重计算为$w_i=\sigma_i^{-2}$。合并干预效应$\hat{\mu}$是根据每个试验观测干预效应的加权平均计算的。

$$\hat{\mu}=\frac{\sum w_i Y_i}{\sum w_i}$$

方差为

$$\mathrm{Var}(\hat{\mu})=\frac{1}{\sum w_i}$$

对于随机效应模型，试验间的干预效应假定是不同的，但是用μ表示潜在真正的效应。用τ^2表示试验间的变异，随机效应模型定义为

$$Y_i=\mu_i+\varepsilon_i,\ \varepsilon_i\sim N(0,\sigma_i^2)$$
$$\mu_i=\mu+E_i,\ E_i\sim N(0,\tau^2)$$

ε_i表示第i个试验的残差，E_i表示“真实的”总体效应和“真实的”潜在效应的差异。在上面的等式中，Y_i假定满足分布$Y_i\sim N(\mu,\sigma_i^2+\tau^2)$。同样的，试验权重根据变异确定，所以随机效应模型的权重为$w_i^*=(\sigma_i^2+\tau^2)^{-1}$。干预效应meta分析的$\hat{\mu}$是通过每个试验的观测效应加权平均获得的。

$$\hat{\mu}=\frac{\sum w_i^* Y_i}{\sum w_i^*}$$

方差表示为

$$\mathrm{Var}(\hat{\mu})=\frac{1}{\sum w_i^*}$$

统计学显著性检验采用的是Wald检验统计量，计算方法为meta分析的干预效应（对数相对危险度或比值比）除以标准误。

$$Z = \frac{\hat{\mu}}{\sqrt{\mathrm{Var}(\hat{\mu})}}$$

该检验量归类为Z检验量或Z值。假定条件为两个具有相似Z值的干预近似服从标准正态分布（均数为0，标准差为1）。这个假定也被归类为空假设或是H_0。相应的双侧p值可根据下面的公式计算：

$$P = 2 \cdot \left(1 - \Phi(|Z|)\right)$$

其中，|Z|表示Z值的绝对值，Φ表示累计标准正态分布[13]。P值表示获得Z值或者更大是因为机遇的几率。P值越小说明观察到两组差异是由于机遇引起的几率就越少，P值越大说明观察到两组差异是由于潜在“真实”治疗。

2.1.3 随机效应模型meta-分析方法

如上所述，随机效应模型会包括一个描述试验之间异质性的统计量[13]。常用的方法是采用一些试验间方差估计量对试验间的方差，τ^2进行估算[13]。

Meta-分析使用最常见的试验间方差估算量（也是Cochrane 协作网*Review Manager*软件采用的唯一方式）是由DerSimonian和Laird (DL)提出的估计量[13;27;32]。DL估计量计算公式如下：

$$\tau_{DL}^2 = \max\left(0, (Q - k + 1)/\left(S_1 - (S_2/S_1)\right)\right)$$

Q是由$Q = \sum W_i(Y_i - \hat{\mu})^2$计算出的Cochrane同质性检验统计量，$S_r = \sum w_i^r$, r = 1,2, k即为纳入meta-分析中试验的数量[13;32]。

由于DL估计量往往低估试验间方差[33-40]，因此在TSA软件中我们采用另外两种随机效应模型——Sidik和Jonkman (SJ) 以及Biggerstaff 和Tweedie(BT) [33;34;41]。

SJ随机效应模型试验间方差采用简单（非迭代）估计量，基于观察到的干预措施效应估计量Y_i的总方差的重复参数[33;34]。计算公式如下:

$$\tau_{SJ}^2 = \sum v_i(Y_i - \mu_0)^2/(k-1)$$

此处$v_i = r_i + 1$，$r_i = \sigma_i^2/\tau_0^2$且$\tau_0^2$是试验间方差的初步估算，可以表示为：

$$\tau_0^2 = \sum (Y_i - \mu_{uw})^2/k$$

μ_{uw}是观察到的试验效应估计值的加权平均数，μ_0是采用τ_0^2作为试验间方差估计值的加权随机效应估计值。模拟试验显示，与DL估计量相比，SJ估计量产生低估试验间方差偏差估计的概率较小[34;37]。也就是说，SJ方法低估试验间异质性的可能性更小，特别是在meta-分析的数据存在中度或较大异质性的情况时。基于SJ估计量得出的可信区间范围接近于期望水平（例如：95%可信区间将包含所有meta-分析大约95%的真实值）[34;37]。而经常报告的基于DL估计值的可信区间范围则往往低于期望水平[33;35-38]。例如，一些模拟试验显示基于DL估计值的95%可信区间所涵盖的实际范围为80%-92%[34;37]。这些可信区间的范围等同于假阳性率为8%到20%，明显大于通常可接受的5%。

因为大多数meta-分析仅纳入有限数量的临床试验，所以估计试验间方差往往存在随机误差[41]。因此需要综合考虑估计随机效应模型试验间方差的不确定性。Biggerstaff和Tweedie (BT)提出了一种达到这种综合的方法[41]。他们为DL估计量的τ^2推断出了一种近似的概率分布f_{DL}。将试验权重定义为$w_i(t) = (\sigma_i^2 + t)^{-1}$，$t$是一个变量，可以假定所有可能的$\tau^2$值, 他们利用$f_{DL}$以及考虑到估计$\tau^2$的不确定性而求出的试验权重。这样通常会形成一种加权方式，即相对于DL方法，样本量较大的临床试验将被赋予更大的权重，而样本量较小的临床试验将被赋予更小的权重。Biggerstaff和Tweedie也提出了一个meta-分析中干预措施效应方差的校正公

式，因此可以促进可信区间的校正（见附录，6.2.1节）。

哪种方式是最佳的随机效应方法？

与DL方法相比，SJ和BT方法各具优劣，还需要视情况而定。SJ估计量可能在meta-分析出现中度异质性时高估试验间方差，导致人为地形成较宽的可信区间[34;37]。有研究证明当meta-分析中的临床试验为小样本、无偏倚的情况时，BT方法所形成的可信区间范围与DL方法所形成的可信区间范围类似[35]。然而，当纳入研究的样本量不同，并且小样本的临床试验存在偏倚时，BT方法可以恰如其分地在考虑异质性的前提下赋予样本量较大的临床试验更大的权重。这一点非常重要。因为DL随机效应模型广受批评的一点就是当meta-分析存在异质性时，小样本的临床试验通常被不自然地赋予较大的权重。一种通常采用的差强人意的解决办法是使用固定效应模型代替。这样做可能降低由于不恰当的加权方式所致的估计值合并的偏倚，但由于未考虑异质性，依然形成不自然地较窄的可信区间。BT方法则可以降低由于不恰当的随机效应模型加权而产生的偏倚，同时考虑异质性。

选择随机效应模型应当包含采用不同方法进行比较的敏感性分析。如果DL、SJ和BT方法均得出类似的统计推断（例如，点估计和可信区间），那么可以认为使用DL方法是合理的，且有把握认为试验间方差的估计是可靠的。

如果三种方法中有两种（或所有）方法形成的结果不同，那么需要采用这两种（或所有）方法进行meta-分析，并且根据每种方法的内在特点来考虑这些结果。例如，如果采用DL和SJ方法形成的结果不同，那么应当考虑以下两种可能的解释：1) 该meta-分析出现了中度或较大的异质性，因此DL估计量低估了试验间方差而形成不自然的较窄的可信区间；2) 该meta-分析中异质性较小，因此SJ 估计量高估了试验间方差而呈现出不自然的较宽的可信区间。在这种情况下，应当分别采用这两种方法进行meta-分析，且考虑这两种情况提示的结果是“真实的”。

2.1.4 零事件试验的处理方法

在结局为二分类变量的试验中，我们关注的结局发生率有可能很低。例如由激素替代疗法引起的心脏病的发生率[42]。有时一组数据中也可能存在零事件结局的记录。在这种情况下，比率效应的测量指标（RR值和OR值）对干预效应的估计没有意义[42]。对此问题的一类解决方法就是给所有干预组和对照组中发生事件数和未发生事件数都加上一个（或几个）常量[42]。这种方法被称为连续校正[42]。很多连续性校正的方法在meta-分析文献中已有提及。

常数连续校正

常数连续校正法在meta-分析文献中是最常用的一种简便方法[42]。这种方法是给每个干预组和对照组的事件发生数和未发生数增加一个连续修正的因素（一个常量）。

表1 零事件试验举例

组别（Group）	发生事件数（Events）	未发生事件数（No）	总事件数（Events）
干预（Intervention）	0	20	20
对照（Control）	5	20	25

零事件试验举例见表1。如果我们使用一个校正因素0.5来进行常量连续校正，则干预组的发生事件数为0+0.5=0.5，干预组未发生事件数为20+0.5=20.5，对照组的发生事件数为5+0.5=5.5，对照组中未发生事件数为20+0.5=20.5。因为患者总数是发生事件数加上未发生事件数，所以患者总数（在使用常数0.5连续校正之后）在干预组中是20.5+0.5=21，在对照组中则是20.5+5.5=26。

如果我们使用的校正因素是0.1，则在干预组中，发生事件数和患者总数（连续校正之后）分别是0.1和20.2，在对照组中则分别是5.1和25.2。

对干预组连续性校正的倒数

另一个连续校正的方法是在事件和非事件数上添加一个常数，这个常数是对干预组患者的总数的倒数[42]。这种类型的连续校正也被称为“治疗组”连续校正[42]。在表1的例子中，干预组的校正因素是1/25=0.04，对照组的校正因素是1/20=0.05。用这种连续性校正方法得出，在干预组中患者数是20.04，事件数是0.04。对照组中，患者数是25.05，事件数是5.05。

经验性连续性校正

常数连续校正法和“治疗组”连续校正法都将干预效果估计偏向于“零效应”（比如，危险差为0和率测量为1）[42]。另一种连续校正法是经验性连续校正，这种方法的干预效果估计更偏向于Meta分析的效果[42]。例如，设$\hat{\theta}$为Meta分析的优势比，并不包括零事件，设*R*为所需要连续校正试验的随机化比例。干预组中的连续校正因子是CF_I，对照组中的连续校正因子是CF_C，可以近似于用下面的公式：

$$\mathrm{CF_r} = \frac{\mathrm{R}}{\mathrm{R}+\hat{\theta}} \cdot \mathrm{C}$$

$$\mathrm{CF_C} = \frac{\hat{\theta}}{\mathrm{R}+\hat{\theta}} \cdot \mathrm{C}$$

由公式可以看出，这两种连续校正的校正因子相加之和是一个常数C[42]。

2.2 累积 Meta 分析中校正的显著性检验和无效性检验

积累的Meta分析中校正的显著性检验有两个目标：第一，它必须测量和计算现有的证据强度；第二，当对累积数据进行重复显著性检验时，必须控制统计误差（1类错误和2类错误）的风险。

量化证据强度就必须定义“目标位置”[1;2;4;6;11;12;23]。这个试验序贯分析（TSA）软件，可以将现有的证据强度进行测量，并通过计算所需要的信息量进行考虑。这个信息量类似于再一个临床试验中所需的样本量[1;2;4;6;11;12;23]。

控制1类错误的风险的方法，是改变我们测量统计学意义的方式。如果一个Meta分析进行显著性检验之前，它的样本量已经超出了所需的信息量大小，那么这个为统计显著性所设的阈值会被校正，去考虑随机误差的升高的风险[1;2;4;6;11;12;23]。另外，检验的统计量本身可能会和证据强度一起改变。TSA软件就是用这两种方法来控制1类错误。

Meta分析的样本量超过所需信息量时，控制2类错误风险的方法就是设置阈值（规则），使得干预组相比于对照组被视为没有优势（和/或劣势）。

这种校正显著性阈值（即控制1类错误）的方法是由 Armitage和Pocock引进的，这些方法被称作“分组序贯分析”[18;43;44]。在Armitage和Pocock的分组序贯分析中，必须要在数据中了解患者在一个期中分析中随机的大致数量[30]。在随机对照试验中，期中分析中的累积数据都是提前计划好的，因此，在每一个interim look之间，组的大小是确定的[30]。在Meta分析中，当需要更新数据时，一个interim look就会出来从新的临床试验中添加数据。不过，Meta分析很少能够定期更新，更新的速度和添加的患者数量都不可预测。因此，Armitage和Pocock的方法并不适用于Meta分析。

Lan and DeMets 扩展了Armitage 和Pocock的理论，允许期中分析灵活多

变，且没有计划性。Lan 和 DeMets用这个方法重复单个随机试验的显著性检验[16;17;30]。期中分析的时间的灵活性使得这个方法适用于Meta分析。因此，这个方法也被TSA软件引用，包括促进在Meta分析中定义“有统计学意义”的敏感界值的监测边界的构建。

同样地，构建无效边界，使得Meta分析当中合理阈值的“无用”更加容易[30]。2.2.1节至2.2.5节描述了这个方法的方法学基础和理论解释。

有一种用于控制2类错误的方法，是对Lan-DeMets方法的扩展，它可以进行非优效和非劣效性检验。即不同于构造来校正阈值的统计显著性，该方法用构造校正阈值进行非优势和非劣效性（即无差异性）检验。总之，校正后的非优势和非劣势边界填补无效边界或内部楔形边界。在2.2.7节，我们对这个方法的基本方法和理论解释进行了描述。

根据先前所述，另一种改变阈值的方法是惩罚统计量本身。这种方法用于惩罚应用的统计检验是一个相对较新的方法，建立在先进概率论的定理上，尤其是用在被称为“重对数定律”的定理上[7;8]。2.2.2节至2.2.6节描述了这个方法的基本方法和理论解释。

2.2.1 确定性 Meta 分析结论所需的信息量

为确凿可靠的Meta分析确定所需信息量的大小（例如，所需患者的数量）是构建校正阈值用TSA的“统计学意义”的先决条件[1;2;4;6;11;12]。阈值的水平则必须按照证据的强度来构造[1;2;4;6;11;12]。该统计方法相关的TSA是基于这样一种假设，即数据会累积直到超过所需的信息量的大小[30]。关于这个假设的进一步说明，请参阅在对此问题较早的方法论的论文[16;17;30;43;44]。

传统的信息量考量

一般认为，一个确凿可靠的Meta分析所需的样本量至少要和一个可探测的、大型的、真实合理的、有把握度的干预性研究所需的样本量一样大[1;2;4;6;11;12]。在这个条件下，一个Meta分析所需的最小信息量（患者数量）可以用下面这个人们所熟知的公式计算得出：

$$\mathrm{IS}_{\mathrm{Patients}} = 2 \cdot \left(\mathrm{Z}_{1-\alpha/2} + \mathrm{Z}_{1-\beta}\right)^2 \cdot 2 \cdot \sigma^2/\delta^2 \qquad (1)$$

在这个公式中，α为获得假阳性结果所需的最大风险（I类错误），β是得到假阴性结果（II类错误）的所需的最大风险，其中$\mathrm{Z}_{1-\alpha/2}$ 和$\mathrm{Z}_{1-\beta}$为$(1-\alpha/2)$和$(1-\beta)$的标准正态分布分位数[1;2;4;6;11;12]。用$\alpha/2$而不是用均值说明这个信息量的大小是被假设成一个双侧检验。对二分类变量来说，$\delta = \mathrm{P_C} - \mathrm{P_E}$ 表示对现实的或者最低限度的重要干预作用的预先估计（$\mathrm{P_C}$和$\mathrm{P_E}$分别表示对照组和干预组的结果的比例）。若$\mathrm{P}^* = (\mathrm{P_C} + \mathrm{P_E})/2$（即干预组和对照组信息量相同）时，$\sigma^2 = \mathrm{P}^*(1 - \mathrm{P}^*)$表示相关联的方差。对连续性变量，δ则表示两组之间方式的差异的预先估计，而σ^2表示相关联的方差。

累积患者数的替代

二分类变量的Meta分析中，Meta分析的信息量和精确度显著的依赖于事件发生或结局发生的数量。因此，在计算Meta分析的信息量时，选择事件数要比患者样本量进行估算更适合。如果所有的试验里，被随机分配接受两种干预措施的患者数量都是相等的，那么所需事件数可以通过下面的式子进行计算：

$$IS_{Events} = P_C * IS/2 + P_E * IS/2$$

其中IS_{Events}是Meta分析所需的事件数，P_C 和P_E和上段的意义是一样的。

这些统计信息（Fischer信息）是将包含在数据集（给出一些假设统计模型）中的信息进行了统计测量[45]。在比较两种干预措施的标准的Meta分析中，统计信息是简单的合并方差的倒数。Meta分析中的统计信息是一个理论上优势测量方法，因为它将三个因素合并在一次测量中：患者人数，事件数和试验次数[46]。这种测量途径为Meta分析提供了一个简单的方法进行信息量考量。Meta分析数据类似于单个试验的累积数据，所需的数据信息可以由下式给出：

$$IS_{Statistical} = \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2/\delta^2$$

在这个式子$IS_{Statistical}$实际获得的统计信息的 Meta 分析，α是 I 类错误所需的最大风险，$Z_{1-\alpha/2}$是标准正态$(1-\alpha/2)$百分位数，β是 II 类错误所需的最大风险，$Z_{1-\beta}$标准正态$(1-\beta)$百分位数，δ是一些预先指定的（最低限度相关）的干预效果[30;45]。

异质性校正因子

Meta分析中涉及的临床试验所包含的研究对象通常覆盖宽泛的人群，并且使用不同的干预方案，不同的研究设计以及具有不同层次的方法学质量（例：风险偏倚或者“系统误差”）[13;47]。考虑这些原因，很容易想到相对于一个独立试验来说，meta分析会产生额外的变异度。这些额外的变异被称为异质性（或者试验间的变异）[13; 47]。由于增加变异会减弱结果的精确度，所以对信息量的考量必须包含meta分析中包括异质性在内的所有变异来源[4;6;11;12]。考虑信息量时有一种合并异质性的方法是在meta分析中通过异质性校正因子将信息量之间相乘[6; 23]。近期，一种相似的异质性校正因子已经被推荐用于独立临床试验中的样本量估算[48]。

异质性校正因子是我们为了meta分析模型而制定的概念化的潜在假设。在固定效应模型中，假设所有的试验都可以被看作可重复的相同试验（在这里指设计和实施过程）。因此，固定效应的meta分析中所需求信息量可以像独立临床试验样本量计算方式那样算出来。在随机效应模型中，我们假设纳入的试验来源于合理的试验分布（在这里指设计和实施过程）。很显然，随机效应模型的变化总是大于固定效应模型。那么，异质性校正因子必须解释由于meta分析导致的从固定效应假设到随机效应假设过程所增加的变异。一个精确的校正可以使得异质性校正因子等于随机效应模型meta分析的总变异与固定效应模型meta分析的总变异之间的比值[6; 23]。所以，异质性调节因子总是等于或者大于1。通过等式（1），让IS_{Fixed}代表固定效应meta分析中所需求的信息大小，V_R代表随机效应模型meta分析中的总变异，V_F代表固定效应模型meta分析的总变异，那么异质性校正信息大小可以通过以下公式得出：

$$IS_{Random} = \frac{V_R}{V_F} IS_{Fixed}$$

假定期望的干预效应在固定(δ_F)和随机(δ_R)效应模型是近似相等的（也就是假设$\delta_F = \delta_R$），从数学上可以看出meta分析中所有的试验的个案都被赋予相同的权重，一致性校正因子（AF）为以下形式：

$$AF = \frac{V_R}{V_F} = \frac{1}{1 - I^2}$$

这里，I^2是meta分析中常见的测量异质性的不一致因素[47]。

还应注意一个要点，无论如何试验的权重都不相等，使用I^2会导致对校正因子的过低估计，从而低估了所需求的信息量[23]。这种情况下，我们可以定义一个离散量(D^2)满足等式成立。

$$AF = \frac{V_R}{V_F} = \frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{k} w_i^*} = \frac{1}{1 - D^2}$$

在这里，w_i表示试验在固定效应模型中的权重，w_i^*表示试验在随机效应模型中的的权重。使用D^2列出的方程，我们得出：

$$D^2 = \frac{V_R - V_F}{V_R} = 1 - \frac{V_F}{V_R} = 1 - \frac{\sum_{i=1}^{k} w_i^*}{\sum_{i=1}^{k} w_i} = \frac{\sum_{i=1}^{k} (w_i^{-1} + \tau^2)^{-1}}{\sum_{i=1}^{k} w_i}$$

此处τ^2代表试验间差异。D^2是差异测量的一个有利性能，它可以将以上推导概括的用于任何给出的meta分析模型。因此，如果我们想使用带有总变异V_R的随机效应模型去meta分析一些试验，那么差异的测量和共同校正因子可以简单地如下表示：

$$D^2 = \frac{V_R - V_F}{V_R} \text{ 且 } AF = \frac{V_R}{V_F}$$

变异性的估计，尤其是试验间的变异，可能会受随机误差和偏倚的左右[41;47; 49; 50]。为此，有些情况下，根据现有数据选择使用D^2或I^2可能不太适合。在只包含有限试验数量（例如小于10个试验）的meta分析中，异质性和试验间变异的估计可能就像小样本临床试验（例如小于100个受试者）干预效应估计一样不可靠。当一个meta分析受到延迟偏倚（例如，大多数阳性结果的试验已经被发表）影响时，会导致典型地低估试验间变异。这种低估是由于早期纳入研究可能已经产生相似（阳性的）的干预效应估计[50]。此后的meta分析（分析的更新）可能包含了很多中立的或阴性结果，在这种情况下，异质性估计将会增大。

对于一个预期小样本（纳入临床研究数量少）的meta分析，我们建议做一个关于异质性预期程度的先验估计。如果我们让H代表一个D^2的概念估计，我们可以使用下面的公式进行先验计算。

$$AF = \frac{1}{1 - H}$$

举例说明，我们预估一个meta分析将出现轻度异质性——基于我们知道的临床问题，试验间的观察差异，当今与未来的预期差异，以及回顾性审视——可以选择定义H为25%，则AF约为1.33。如果预期一个中等程度的异质性，可以选择定义H为50%，则AF为2.00左右。如果有很强的异质性，H可设为75%，则AF可估计为4.00。

当meta分析只纳入很少的试验而异质性程度很难估计时，我们推荐TSA使用者对变量进行敏感性分析。例如，可以为已给出的meta分析设想一个现实的或者可接受的最小和最大的异质程度。举个例子，可以推断最小的较为合理的统计学异质性程度为20%，也可以觉得是否统计学异质性超过了60%，那么亚组的效应测量要比整体综合干预效果估计要好，这种情况就不要做整体的meta分析了。在这个例子中，对于主要信息量计算，我们可以使用二者的均值，（60%+20%）/2=40%，但是要承认所需信息量可能要像基于60%那样大的异质性校正，也可能像20%那样低的异质性校正。再举一个例子，你可以通过添加“想象”未来试

验，设想或构建一些“最好”和“最坏”的情况（无论这些可能是什么）给现在的meta分析。对于D^2估计值，以及构建的用于敏感性分析的真实或可接受的异质性程度，这种方法可以评估它们的稳定性和可靠性。

估计对照组事件比例和预期干预效应

做TSA时，对照组事件比例和预期干预效应的估计是计算所需信息量的重要决定因素。所做的一切都是尽可能的使这个估计值精确、现实。

对于二分类变量数据，对照组事件比例可以通过临床经验和相关领域的证据来估计。真实干预效应的先验估计通常以相对危险度降低（RRR）来表达。当只有像调查研究这样有限的证据时，可以通过临床经验和相关领域的证据估计临床相关干预效应。Pogue和Yusuf的文章是个例子[1;2]，对照组事件比例P_C和先验的RRR就是基于相关心脏病领域的经验。Pogue和Yusuf给两个著名的心脏病学meta分析应用了信息量考量：“静脉溶栓酶在急性心肌梗塞的应用”（IntravenousStreptokinase in Acute Myocardial Infarction），“静脉镁在急性心肌梗塞的应用”（Intravenous Magnesium inAcute Myocardial Infarction）。假设大多主要血管结局，例如死亡，可以真实的估计对照组有10%死亡率。Pogue和Yusuf进一步考虑了理论干预对心肌梗塞后死亡率的预防。他们注意到为减少主要心血管事件的风险，如死亡，真实有效的治疗有先前产生的10%，15%，或者最好的是20%RRR。

对于给出的任何临床问题都需要考虑什么值适合P_C和RRR。在（实验）试验组期望的事件比例P_E，可以通过公式$P_E = P_C (1- RRR)$获得。之后，假设的P_E和P_C可以代入公式计算需求信息量。

关于从一个干预领域到另一个的期望真实干预效应的推理可能会有问题，因为先验估计对“实质”表达的近似值欠佳。临床试验文章计算的样本量都是基于过于乐观的估计干预效应。没有原因表明为什么这个与meta分析的信息量计算有任何差别。

如果随机对照试验已经调查了干预效应，那么这种估计的集合可能会用于更好的量化期望干预效应。然而，不是所有试验都提供了正确的估计，在估计一些期望干预效应值时，要注意确保干预效应估计的正确性。

很多试验过高估计调查干预效应归咎于选择性报告结局和风险偏倚（即，由于随机序列生成，随机隐藏，盲法，失访，或其他机理的不足导致的系统误差）[13;51-58]。这样的试验被归为高风险偏倚[13]。相反的，采用正确的干预效应估计的试验被归为低风险偏倚[13]。如果证据是来源于很多低偏倚风险的文章得出的调查干预效应值，它就可以在meta分析时适合使用先前的期望干预效应[6;11;12]。然而，当证据比较少时，meta分析是经常需要信息量计算的。即使有大量低偏倚风险的文章能够得出接近预期真实的干预效应，而这些文章的汇总估计仍然会受限于需考虑的随机误差，延迟偏倚和发表偏倚。因此，我们说低偏倚风险临床研究合并的先验的预期干预效应估计是可靠的，也仅在这个meta分析不存在大量随机误差的前提下才成立。此外，满足这个规律还需要在不存在延迟偏倚和发表偏倚的条件。

推荐一种方法来确定干预效应meta分析的信息量并不太现实。当然，信息量考量应该基于合理的对照组事件比例的范围，干预效应以及合适的I类和II类错误。独立临床试验的样本量估计不仅是算一个数。合理的样本大小不仅取决于一系列合理的治疗效果,对照组事件率,和I型和II型错误，而同时要提供一个合理的估计区间数量的病人以产生确证性的结果。从产生的样本量范围来说，可以选择一个

主要终点，其余的作为样本量的敏感度（把握度）计算。我们推荐meta分析信息量考量遵循同一个规则。低偏倚风险P_C和RRR估计，可以轻易将先验“真实的”最好和最差的干预效果合并，如此，提供一个合理的区间使得meta分析信息需求列于最终的meta分析推论里。

局限性

meta分析的需求信息量（无论是决定患者需求量，事件需求量还是统计信息需求量）计算会产生很多局限性。在随机临床试验中，可以合理假设与目标人群相似的随机患者的预后因素分布。在系统综述meta分析中，草案阶段就已决定试验是基于很少的纳入标准地纳入进来。由于临床试验的纳入标准（和排除标准）几乎都不一样，也因为样本量不同，所以meta分析和系统综述的作者不太可能去控制预后因素的分布。甚至，当一些系统综述纳入标准被更新了，作者也无法精确的预测新发表的试验的预后因素分布。基线预后因素在对照组会对发生率有很重要的影响。这种情况适合做一个先验分析，试图将meta分析中人群的基线发病率与目标人群的基线发病率量化，如果有必要还可以做事后敏感性分析。

最小的重要比较干预效果（又称为最小重要差异）可能不会总是与纳入的研究相类似。例如，如果调查患者人群在试验中经历不同不良事件的风险，那么最小重要差异在试验中可能也会有差异。这种变异是临床意图的结果。任何医疗干预，获益的机会都需要超过任何伤害的风险。一组有很大伤害风险的人群需要有跟多获益的机会才有治疗的意义。当最小重要差异在试验中变化时，信息量考量可能仍然是合理的。然而，需要记住的是推理描绘的meta分析结局只能推广于应用了先验最小重要差异的患者人群。

当通过患者或事件数定义所需样本量时，不可预知的异质性问题可能通过预期一些最大化异质程度来处理，并且相应的校正所需样本量[4]。这种方法有个明显的局限是当仅有很少的临床研究可用时，期望的异质性程度是很难猜测和估计的。尽管我们推荐采用敏感性分析对异质性进行校正，但随着meta分析纳入的试验增多，实际发生的临床异质性随之增加，则仅用预估的异质性也不能够反映的实际情况。

当meta分析的信息量是通过所需统计信息定义时，所需信息量公式就不再需要异质性估计值了。此时，meta分析中实际的信息（或统计信息估计值）通过试验间估计变异值体现在了异质性中。

然而这种情况存在一种局限性，即只有当试验间估计变异可靠时，累积统计信息的估计才可靠。可能的解决办法是使用更复杂的方法学去校正与试验间变异估计相关的不确定因素。方法之一是使用随机效应办法，当使用传统的DerSimonian-Laird 估算法(见第2.1.3节)时[41]，通过Biggerstaff-Tweedie方法合并试验间变异估计的不确定性因素。另一种方法是当试验间变异参数能够通过先验分布推导出时，可采用贝叶斯meta分析（Bayesianmeta-analysis）。

2.2.2 累积检验统计量（Z 曲线）

正如2.2.2节提到的，Meta分析的“统计学意义”用的是Wald-type检验统计量。该统计信息是通过汇集干预作用除以标准误来计算[13]，通常被称为Z-statistic或 Z-value。如果两个调查的干预没有不同（即零假设），则Z值将近似地服从标准正态分布（正态分布，均值为0，标准偏差为1）。可观察到的Z值的绝对值越大，两个调查的干预不同的证据强度越强。如果观察的Z值的绝对值基本上大于0，通常可以表明两种干预措施的效果之间是存在差异的，不能简单的用偶然出现来解释。在这种情况下，两种干预措施之间的差异被称为“统计学显著”。在

定义当中，或者在某些极端情况，如零假设为真的情况下，P值表示可观察到的差异的概率。在实际当中，我们通常用P值来评估统计显著性。用Z值求出的P值（细节详见 2.1.2节）；这两个测量值表现的是同一个的信息的不同方法，是可以互换的。例如，双侧P值小于5％和Z值的绝对值大于1.96是同一个意思，反过来也是一样。

每当更新了一个Meta分析，其中的Z值也会被重新计算。因此，随着Meta分析的更新，同时会产生一系列Z值。观察显着性检验的演变，一系列Z值可根据累计信息（累积的患者，事件或统计信息）绘制，从而产生一个曲线，该曲线通常被称为Z-曲线[1;2;4;6;11;12]。

2.2.3 Meta 分析中显著性检验存在的问题

第一章我们提到，meta分析中传统的显著性检验无法体现观测统计值、P值和证据强度，以及显著性检验重复次数之间的关系[1-4;6;11;12]。这会增加meta分析得出假阳性结果的风险。本节内容将对初级到中级meta分析中显著性检验统计学及概念上的意义进行描述，同时介绍如何解决无法合并证据强度以及显著性检验重复次数的问题。

显著性检验的一般标准

传统的显著性检验会得到I类错误关于界值α的最大风险,当P值作为统计显著值时，α也是P值的临界值。在统计学显著性评估中，P值和Z值是可互换的。如上所述，对每个P值的临界值α，都存在一个相应的Z值临界值Z_α。例如，如果我们允许一个双侧检验I类错误的最大风险为5%，则意味着Z绝对值大于1.96作为统计显著值。但如果我们预设一个双侧检验I类错误最大风险为1%，即Z绝对值大于2.58作为统计显著性值。

$\Pr(X|Y)$ 表示事件Y为真（或已发生）的情况下事件X发生的概率，用|Z|表示Z的绝对值，则我们需要解决的问题是选取一个合适的临界值c，使得以下等式成立：

$$\Pr(|Z| \geq c|H_0 \text{ is true}) \leq \alpha \qquad (2)$$

$$\Pr(|Z| = c|H_0 \text{ is true}) = \alpha \qquad (3)$$

我们假设其余理论部分关于重复显著性检验（第2.2.2节—第2.2.5节）的显著性检验都是双侧的，所有检验Z值都是绝对值。我们假设后者是因为这样做其涉及的代数更简单。例如，在对一个非特定检验统计量的双侧临界值作界定时，人们可能需要考虑$\Pr(Z \leq -c \text{ or } Z \geq c \mid \ldots)$ 而不是$\Pr(|Z| \geq c \mid \ldots)$

重复性显著性检验中出现的问题

传统的单一显著性检验如果累积了“足够的”数据则可认为是可靠的。在荟萃分析中，一个单一的显著性检验一旦超过了所需信息量大小则被认为是可靠的[1-4;6;11;12;20;59]。如果我们对荟萃分析超过所需信息量大小时或之后进行single test显著性检验，统计学显著性检验就仅仅需要确定一个合适的临界值c，这样可使(2) 和(3)等式成立。例如，如果之前没有对荟萃分析的数据进行显著性检验，我们则认为c=1.96是适合的。

当一个累积meta分析经过一次或多次显著性检验（在超出其信息量大小之前）时，这种情况变得更加复杂。考虑示例中荟萃分析校正一次和用I 类错误最大为5%。这种情况下，第一个荟萃分析产生一个Z值，Z_1，校正后又产生另一个

Z_2。如果荟萃分析产生Z值大于1.96，那么表明研究的两个干预措施有显著性差异。而如果第一个荟萃分析的结果并不显著(i.e., Z1<1.96)，那么在meta分析的校正中产生的Z值大于1.96时两个干预措施依然可以认为是有显著性差异的。根据概率论的规律，两个干预措施在零假设下被表明有显著性差异:

$$\begin{aligned}\Pr(H_0 \text{ rejected}) &= \Pr(|Z_1| \geq 1.96 \text{ or } |Z_2| \geq 1.96) \\ &= \Pr(|Z_1| \geq 1.96) \cdot \Pr(|Z_2| \geq 1.96 | |Z_1| < 1.96)\end{aligned}$$

上述公式可以表明此等式始终大于5%(见附录A.3.1)。一般，使用单一检验界值进行重复显著性检验常会扩大I类错误，（重复的）显著性检验使用的累计数据越多，I类错误就越大[30]。对荟萃分析中的数据，模拟研究已经证实，当用传统*α*=5%的临界值来检验每次校正后的显著性差异，则重复显著性检验产生I类错误的可能性为10%到30%[7;8;10;31]。

2.2.4 α-消耗函数和试验序贯监测界值

解决2.2.3部分所述问题的方法是校正Z值的临界值，允许I类错误风险恢复到所期望的最大风险值[1;2;6;17]。在两个检验的例子中，在无效假设条件下我们需要两个临界值，c_1和c_2。

$$\Pr(|Z_1| \geq c_1 \text{ or } |Z_2| \geq c_2) \leq \alpha$$

这就等于去发现两个最大I类错误风险值α_1和α_2，在无效假设条件下合计为α。

$$\Pr(|Z_1| \geq c_1) \leq \alpha_1$$

$$\Pr(|Z_2| \geq c_2 | |Z_1| < c_1) \leq \alpha_2$$

通常情况下，运行k次重复显著性检验 (i.e.,进行初始meta分析和修正k-1)时，我们需要无效假设条件下确保为每个k检验找到临界值$c_1, \ldots, c_k$。

$$\Pr(|Z_1| \geq c_1 \text{ or } |Z_2| \geq c_2 \text{ or} \ldots \text{or } |Z_k| \geq c_k) \leq \alpha$$

这就等于找到k的最大I类错误风险$\alpha_1, \ldots, \alpha_k$，在无效假设条件下合计为α。

$$\Pr(|Z_1| \geq c_1) \leq \alpha_1$$

$$\Pr(|Z_2| \geq c_2 | |Z_1| < c_1) \leq \alpha_2$$

$$\Pr(|Z_3| \geq c_3 | |Z_1| < c_1 \text{ and } |Z_2| < c_2) \leq \alpha_3$$

$$\ldots$$

$$\Pr(|Z_k| \geq c_k \text{ or } |Z_1| < c_1 \text{ and} \ldots \text{and } |Z_{k-1}| < c_{k-1}) \leq \alpha_k$$

对Z曲线的校对（排序）被称为*监测界值，*或*group 序贯监测界值*（用一系列临界值在对临床试验中随机的累积的群体病人检验序列中）[17;30;44]。在meta分析中，这样的临界值用在一系列的试验中，因此我们把它们称作*试验序贯监测界值（sequential monitoring boundaries）*[6]。meta分析和试验序贯监测界值结合被称作试验序贯分析（*trial sequential analysis*）[6]。

试验序贯监测界值的运用需要对k的最大I类错误风险值和intensive数值积分作出预先说明[60]。设置I类错误风险值$\alpha_1, \ldots, \alpha_k$的一个简单的方法就是α-消耗法（或α-消耗函数）[1;2;17;30]。这一方法在TSA程序中执行。α-消耗函数是有关时间的单调递增函数，可以适当根据累计信息量设定每个显著性检验的最大I类错误风险$\alpha_1, \ldots, \alpha_k$[16;17]。自变量由信息部分决定；根据需要的信息大小区分累计信息（例如，根据需要的病人数量来区分累积病人数）[6;15;17]。因变量（函数）是累

积的I类错误；这就使得当明确界定信息边界的显著性时，误差值应为最大。随着信息部分增加-例如，随着累积信息的增加-可接受的I类错误值也会增加。α-消耗函数提供了允许在任何IF下量化随机误差风险的方法，以确保总随机误差风险在达到信息量大小是可以小于5%。单调递增函数与单调递减函数Z检验下的统计显著性临界值相当。

α-消耗函数为0到1之间（0是0名患者随机分配的点，1为累积信息量等于所需信息量时的点）[16;17]。0的α-消耗函数的值总是等于0：$\alpha(0) = 0$此时没有累积的信息量。1的α-消耗函数总是等于α: $\alpha(1) = \alpha$，这时所有所需信息都已累积，总α误差就是可接受的I类错误总和（总是5%）。在0和1之间任意一点（i显著性检验时的信息部分(IF_i)）α-消耗函数的值与最大I类错误风险值的总和相等，总的最大I类错误风险由第i次显著性检验的临界值得出。换言之，α-消耗函数与I类错误消耗多少相等。即：$\alpha(IF_i) = \alpha_1 + \alpha_2 + \cdots + \alpha_i$。因此：

$$\Pr(|Z_1| \geq c_1) \leq \alpha_1 = \alpha(IF_1)$$
$$\Pr(|Z_2| \geq c_2 | |Z_1| < c_1) \leq \alpha_2 = \alpha(IF_2) - \alpha(IF_1)$$
$$\Pr(|Z_3| \geq c_3 | |Z_1| < c_1 \text{ and } |Z_2| < c_2) \leq \alpha_3 = \alpha(IF_3) - \alpha(IF_2)$$
$$\ldots$$
$$\Pr(|Z_k| \geq c_k \text{ or } |Z_1| < c_1 \text{ and} \ldots \text{and } |Z_{k-1}| < c_{k-1}) \leq \alpha_k = \alpha(IF_k) - \alpha(IF_{k-1})$$

实际α-消耗函数可是是任何的单调递增函数[16;17]。一个众所周知的例子是$\alpha(t) = t \cdot \alpha$[16;17;30]。当所有显著性检验在等距（关于信息部分量表）运行时，此α-消耗函数会为Z值产生相等的界值(如：$c_1 = c_2 = \ldots = c_k$)。这个校正首先由Pocock提出。一个更加常用的α-消耗方法就是power family -α-消耗函数定义为$\alpha(t) = t^{\rho} \cdot \alpha$[16;17;30]。基于把握度的α-消耗函数，r>1和所有显著性检验在等距运行时，前期显著性检验要比后期产生的界值更为保守。总之，Z-曲线的界值（绝对值）在α-消耗函数为凸函数且所有显著性检验在等距下运行时是单调递减的[16;17;30]。单调递减函数界值（由单调递增函数得出）是最令人满意的，因为随机误差的影响通常与累积信息量成反比。尽管递减界值有无数种组合存在，可能有些界值组合是可取的。

根据高等概率论，α-消耗函数产生的理论最优界值由以下表达式得出：

$$\alpha(IF) = 2 - 2\Phi\left(Z_{\alpha/2}/\sqrt{IF}\right)$$

这里，Φ是个标准正态累积分布函数[16;17;30]。由此α-消耗函数得出的边界类型最早由O'Brien和Fleming提出来用于等量信息部分[61]。Lan and DeMets后来提出α-消耗函数信息部分可以灵活增量[16;17;30]。因此，上述α-消耗函数通常被称为Lan-DeMets式O'Brien-Flemingα-消耗函数。由α-消耗函数得出的监测界值通常简称为Lan-DeMets 监测界值或O'Brien-Fleming 监测界值。在手册的剩余部分，我们将称之为O'Brien-Fleming 监测界值。当前，O'Brien-Flemingα-消耗函数是唯一一个能在TSA软件里运行的α-消耗函数。

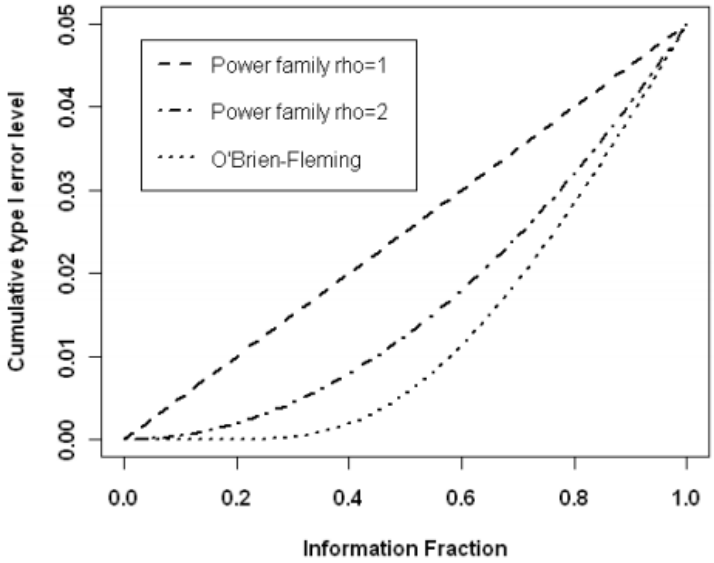


图6 当ρ=1或ρ=2时，基于把握度的α-消耗函数形态，和O'Brien-Fleming α-消耗函数形态。（Information Fraction：信息量分数，Cumulative type I error level：累积I类错误水平，Power family rho=1：ρ=1时基于把握度的α-消耗函数，Power family rho=2：ρ=2时基于把握度的α-消耗函数，O'Brien-Fleming：O'Brien-Fleming α-消耗函数）

图6所示，O'Brien-Fleming α-消耗函数是一个指数递增函数。它在早期阶段边界比较保守，此边界下只累积了有限的数据，但随后边界会更大，累积的数据也会越多。

方法学专家建议将O'Brien-Fleming边界作为多数随机对照试验中对累积数据进行重复显著性检验的首选[30;62]。在meta分析中，早期阶段尤其关注随机误差风险（和时间趋势偏倚）(如，在meta分析中包括一个小样本的患者和事件)，O'Brien-Fleming边界也作为首选[1;2;4;6;11;12]。

原因有两个。第一，对所需信息量的异质性校正是基于预期异质性程度的合理的先验估计，O'Brien-Fleming边界可以自然解释由于随机误差和异质性而引起meta分析的推论发生波动的程度。第二，只要后来的显著性检验在信息轴的合理间距间执行（如，距所需信息量至少1%），O'Brien-Fleming边界就相对不会受到之前显著性检验数量的影响。第二个特性适合用在设计meta分析中，因为在结果更新时，meta分析受到显著性检验影响的频次并不清楚。例如，有些meta分析可能包括虽不同但高度重叠的数据，因为纳入标准根据系统综述的更新而改动。其它监测界值，如一组基于把握度族的α-消耗函数监测界值rho=2可能产生关于统计学显著性方面的矛盾推论，如果，例如，为2个预先的更新估计的监测界值不是4.

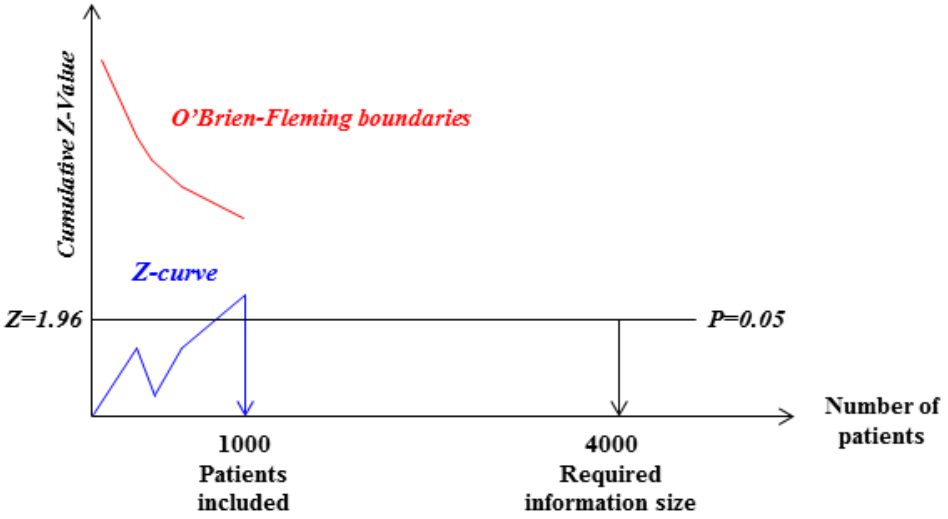


图7 示例：进行四次累积meta分析后仍无结论。（Cumulative Z-Value：累计Z值，Number of patients：患者数，1000 Patients included：纳入1000例患者，4000 Required information size：需要4000例的信息量，O'Brien-Fleming boundaries：O'Brien-Fleming边界，Z-curve：Z-曲线）

图7为O'Brien-Fleming 边界的使用示例。在此meta分析中，所需信息量为4000名患者，但可获得的信息为1000名。Z的终值大于1.96。使用传统单一检验界值，此Z值显示统计学有显著性。使用O'Brien-Fleming 边界，则此样本量下需要更大的Z值，以推断出统计学显著性。因其未与边界相交，因此meta分析尚无法得到肯定结论。

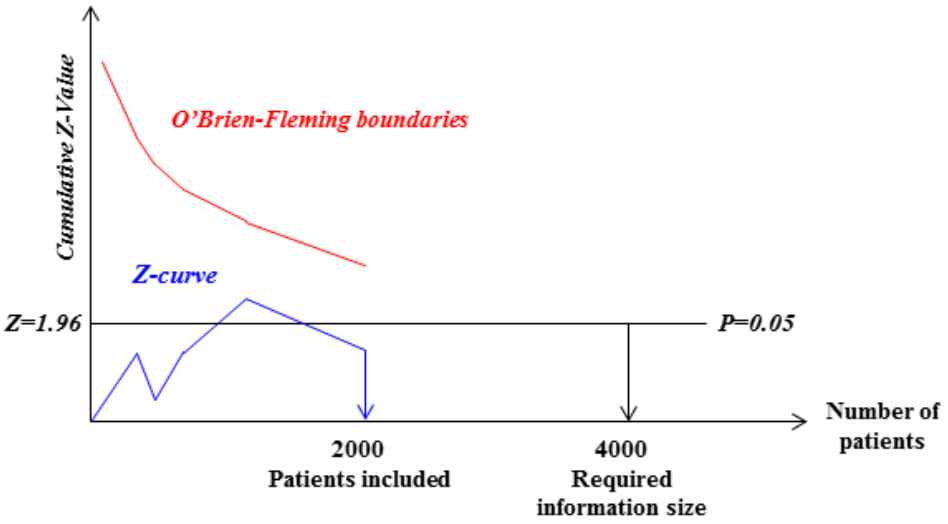


图8 示例：该meta分析在第五次累积显著性检验中有假阳性Z值（Cumulative Z-Value：累计Z值，Number of patients：患者数，2000 Patients included：纳入2000例患者，4000 Required information size：需要4000例的信息量，O'Brien-Fleming boundaries：O'Brien-Fleming边界，Z-curve：Z-曲线）

图8所示例子中，所需信息量还是4000名患者，现在可获得样本量有2000名。终点Z值小于1.96；使用传统的或边界技术时，结果都是不确定的。然而前面Z值已经在累积过程中计算出来了，包括一个比1.96大的值。这个例子说明一个累积Z-曲线如何在早期meta分析中与显著性传统界值相交，那就是只需说明

它在后期的meta分析中没有显著性差异。O'Brien-Fleming边界可以防止这种较早出现的假阳性结果。

图9所示例子中，所需信息大小和可获得信息大小与图8中一样。在这里，Z值在第五个显著性检验中计算显示“非常充足”；Z-曲线与O'BrienFleming边界相交，表明对干预效果评估所需的信息量大小达到要求，则meta分析能够得肯定的结论。

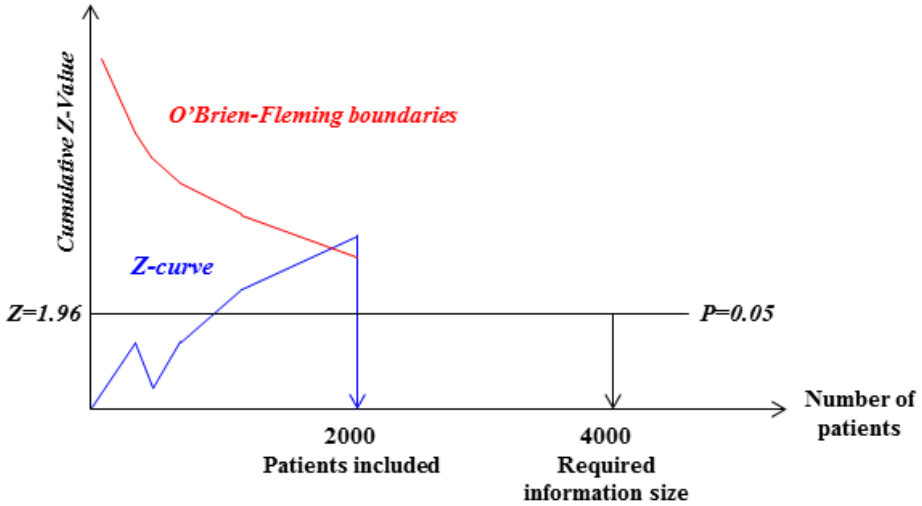


Figure 9

图9 示例：该meta分析在经第五次累积显著性检验后与O'BrienFleming边界（O'BrienFleming boundaries）相交得出确定结论。（Cumulative Z-Value：累计Z值，Number of patients：患者数，2000 Patients included：纳入2000例患者，4000 Required information size：需要4000例的信息量，O'Brien-Fleming boundaries：O'Brien-Fleming边界，Z-curve：Z-曲线）

上述例子中（图7-9），监测界值在Y轴上半轴构建。双侧对称显著性检验界值可以在y轴正负两侧构建。TSA程序可以同时进行单侧或双侧显著性检验。当二分类数据meta分析测量结果为失败（如死亡或复发），Z值在Y轴上半轴表示试验措施有益，而Z值在下半轴则表示不利。

Z曲线监测界值是一个α消耗函数；它们根据Reboussin et al的数值回归积分计算出来[60]。尽管所有的界值都是为meta分析中每个累计修正的离散点，TSA软件将这些点连接起来从而新建一个连续性界线以更好的可视化解读。

2.2.5 试验序贯分析中校正的可信区间

如同重复的显著性检验影响总的I类错误，它也影响了可信区间的构成。比如，当我们假设综合效应值是正态分布的，通常在meta分析中也是这么假设，我们得出“朴素”对称的的95%可信区间$\hat{\mu} \pm 1.96 \cdot se(\hat{\mu})$，其中，$(\hat{\mu})$表示meta分析干预效应，$se(\hat{\mu})$表示其相关的标准差。然而，如果meta分析经过重复统计评估后产生了一系列的可信区间，其包涵“真”的总效果的机率必然小于95%。也就是说，如果得到一系列理想对称$(1-\alpha)\%$的可信区间，$\hat{\mu} \pm z_{1-\alpha/2} \cdot se(\hat{\mu})$，所有可信区间包涵“真”的总效应的机率必然小于$(1-\alpha)\%$。因此，当meta分析经过重复统计评估后，理想可信区间得出虚假推论的风险将增大。当一些潜在的“真”的干预效果存在时，由可信区间得出的虚假推论将产生如图10所示的

其中一种情况。

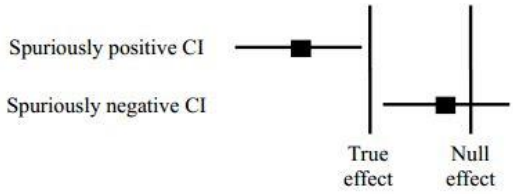


图 10 虚假阳性和阴性可信区间推论的例子。（Spuriously positive CI：假阳性可信区间，Spuriously negative CI：假阴性可信区间，True effect：真效应，Null effect：零效应）

当没有干预效果时，如果排除了零效应，可信区间将产生虚假推论。这种情况与假阳性显著性检验相当（参考2.2.4节）。

与重复显著性检验的校正类似，可信区间可根据可获得的信息强度（比如，患者数）和统计评估数量进行校正。如果我们让l和U表示一些理想的覆盖1-α的可信区间的低限和高限，我们得出

$$\Pr(l \le \mu \le u) = 1 - \alpha$$

$$Pr(l_1 \le \mu \le u_1, l_2 \le \mu \le u_2, \ldots, l_k \le \mu \le u_k) \ge 1\text{-}\alpha$$

当系统综述经过重复统计评估时，重复的理想可信区间不会产生期望范围。因此，我们需要建立一系列可达到期望范围的区间。假设一meta分析经过k次统计评估直到它超过了所需的信息量。让$l_1,l_2,\ldots,l_k$和$u_1,u_2,\ldots,u_k$分别表示k次meta分析统计评估可信区间的低限和高限。为了保证期望范围，这些限制必须满足：

$$\Pr(l_1 \le \mu \le u_1, l_2 \le \mu \le u_2, \ldots, l_k \le \mu \le u_k) \ge 1 - \alpha$$

因此，这些k个区间中第j个区间必须满足

$$\Pr(l_j \le \mu \le u_j) \ge 1 - \alpha$$

由上我们可以清楚地看到，每个重复可信区间的α水平不能超过总的最大值α。也就是说，每个重复可信区间α水平的总和方能达到总的最大值α。因此，通过控制总的α水平，我们可以控制总范围。控制总α水平的理论已经在2.2.4节中讲述过，且较容易被重复可信区间应用。理想可信区间可通过公式$\hat{\mu} \pm z_{1-\alpha/2} \cdot se(\hat{\mu})$获得，已知$z_{\alpha/2} \le \hat{\mu}/se(\hat{\mu}) \le z_{1-\alpha/2}$和在零假设下接近$(1-\alpha)\%$的机率，可得到：

$$z_{\alpha/2} \le Z \le z_{1-\alpha/2},$$

其中Z表示显著性检验的Z值。通过统计监测临界值替换$z_{\alpha/2}$和$z_{1-\alpha/2}$，c_1，$c_2, \ldots, c_k$,和孤立的$\hat{\mu}$，得出可控范围的重复可信区间的简单表达式。在k个可信区间中的第j个，该表达式为

$$\hat{\mu} \pm c_j \cdot se(\hat{\mu})$$

由$\Pr\left(\hat{\mu} - c_1 \cdot se(\hat{\mu}) \le \mu \le \hat{\mu} + c_1 \cdot se(\hat{\mu}), \ldots, \hat{\mu} - c_k \cdot se(\hat{\mu}) \le \mu \le \hat{\mu} + c_k \cdot se(\hat{\mu})\right) \ge 1 - \alpha$可推导出单侧可信区间。

TSA软件提供了计算系列统计评估中最后可信区间的选项（参考第4章）。

2.2.6 迭代对数法则

根据现有证据强度和显著性检验重复次数去校正Z值是2.2.3节中提及的重复显著性检验问题的另一个解决方法[7;8]。运用高级概率论的迭代对数法则，将标准正态分布Z值用数据中观察对象数量对数的对数来划分，分数值肯定在和-$\sqrt{2}$和$\sqrt{2}$之间。在统计检验中，这个法则可用来在meta分析重复显著性检验中控制Ⅰ类错误的扩大。通过现有信息对数的对数来划分标准正态分布检验得到的足够数据积累后可以很好地控制统计检验的“行为”。Lan等应用该理论介绍了每个显著性检验得到的Z值，校正Z值Z*，公式为：

$$Z_j^* = \frac{Z_j}{\sqrt{\lambda \ln(\ln(I_j))}}$$

Z_j是常规Z值，I_j是第j个显著性检验的累积统计信息（见2.2.1节中累积患者数的替代方案），λ是确保控制I类错误最大值的常量[8]。在连续性变量[8]及二分类变量[7]的meta分析中，Lan等和Hu等各用模拟法来估算选择适当的常量λ。在连续性变量的meta分析中，当对双侧统计检验设定I类错误最大值α为5%时（单侧α为2.5%），Lan等发现λ = 2通常能较好地控制I类错误[8]。Z*是评价统计学差异的常规标准，比如，|Z*|≥1.96表示具有双侧α=5%的统计学差异。对于二分类变量的meta分析，Hu等为不同I类错误的最大值和不同效果测量方法选择合适的λ[7]。他们类似的结果得出了表2中所示的λ推荐值。

表2 迭代对数法则计算Z值的λ推荐值

效应测量（Effect measure）	最大I类错误对应临界值（Max. type I error(cprresponding threshold)）		
	α=0.01(c=2.33)	α=0.025(c=1.96)	α=0.05(c=1.65)
危险度差（Risk difference）	λ=3	λ=1.5	λ=1.5
相对危险度（Relative risk）	λ=3.5	λ=2	λ=2
比值比（Odds ratio）	λ=3.5	λ=2	λ=2

这些λ值只适合于研究对象的数量，对照组事件比例，模拟用到的试验间变异，因此有可能不适用于所有meta分析[7;8]。例如，模拟中对照组事件比例发生的最小值是0.05，而在许多重要临床状况中则小于0.05。另外，没有任一模拟考虑了诸如时滞偏倚之类的时间趋势偏倚和发表偏倚，而在meta分析中，这些偏倚对显著性检验有相当大的影响。如同2.2.1节的局限性中所述，统计资料依赖于试验间方差的正确可靠的估算。如果试验间方差被低估（如因为时滞偏倚），惩罚的Z统计量将会人为地增大。可以合理使用表2中推荐的λ值构建一系列恰当的最小值。二分类变量meta分析中适当的λ值只包括少数的试验，患者和/或事件比Hu

等推荐的机率可能还高。

2.2.7 β 消耗函数与无效界值

当meta分析的结果显示没有统计学意义时，我们需要去评价这种无意义是因为缺少把握度还是由于干预间潜在的等效性。

等效性统计检验以及对于一干预措施的非优效性和非劣效性检验通常会用到无效检验[30]。由这个统计检验得出的临界值被称为无效边界。当Z曲线与无效边界相交时，认为两种干预措施疗效相差还没达到期望的干预效果。

超过预期信息的meta分析应该有足够的把握度去证实一种干预措施比另一种优越。在这一部分中，我们只考虑无统计学意义的meta分析未超过所需的信息量大小。我们不再把所有Z值都作为绝对值。Z值为正表明试验组优于对照组，Z值为负表明试验组劣于对照组。接下来首先讲述非优效性检验，然后是非劣效性检验和无效检验。

一个meta分析得到的Z值可能认为试验组的干预无统计学意义。只有当Z值充分低于具统计学意义的临界值（在用户手册的试验干预部分）时，才有足够把握说明试验组干预并不比对照组优越。为了理解它，我们首先要定义优越和“充分低于”。在重复统计检验的理论中，优越与所需信息量的假设有关。计算meta分析的所需信息量时，我们假设一个先验值：干预效果δ。它的效果大小代表了我们认为的两种干预间重要差别的最小值。理论上讲，如果差异小于δ值，认为是临床的、实际的、不重要的或不值得调查。当然δ值取决于研究内容。比如，如果结局指标是死亡率的话，10%的相对危险度也是非常重要的，但如果结局指标是恶心，那就不那么重要。

在定义重复统计检验中的“充分低于”时，首先考虑在何种情况下，meta分析中的信息量与所需信息量相等以及初次进行统计检验。首先，H_δ表示假设效果等同于δ，对比于无效假设，这是备选假设。假设H_δ为真，meta分析有统计学意义的机率等于选择的把握度1-β
。当信息数达到时，meta分析假阴性的机率等于β。在这种情况下，有统计显著性意义的临界值c满足：

$$\Pr(|Z| \geq c|H_0 \text{ is true}) \leq \alpha$$

因为c同时也满足$\Pr(Z < c|H_\delta \text{ is true}) \leq \beta$，也成为了非优效性检验的临界值。

在meta分析超过所需信息量前，我们需要做重复的统计检验，同时也能检验非优效性。这种检验是在备选假设下定义的临界值，不会导致II类错误的扩大。比如，如果我们做两次非优效检验，我们需有两个临界值c_1和c_2，相应地，Z值也有两个，Z_1和Z_2，

$$\Pr(Z_1 < c_1 \text{ or } Z_2 < c_2) \leq \beta$$

在这种情况下，Z_1小于c_1且Z_2小于c_2是“充分低于”具有统计学意义的临界值并得出非优效结论的条件。在进行k次非优效检验时，我们需要找到临界值$c_1, \ldots, c_k$来满足公式

$$\Pr(Z_1 < c_1 \text{ or } Z_2 < c_2 \text{ or} \ldots \text{or } Z_k < c_k) \leq \beta$$

在备选假设H_δ下，$\beta_1, \ldots, \beta_k$,总和为β时，找到II类错误风险的最大值k是有意义的。

$$\Pr(Z_1 < c_1) \leq \beta_1$$

$$\Pr(Z_2 < c_2 | Z_1 \geq c_1) \leq \beta_2$$
$$\Pr(Z_3 < c_3 | Z_1 \geq c_1 \text{ and } Z_2 \geq c_2) \leq \beta_3$$
$$\ldots$$
$$\Pr(Z_k < c_k | Z_1 \geq c_1 \text{ and} \ldots \text{and } Z_{k-1} \geq c_{k-1}) \leq \beta_k$$

这就要求在重复性检验中控制II类错误与控制I类错误的要求是相似的。多重检验增加了实际错误数量，我们需要寻找一种方法控制错误量的增加。因为错误是由同一现象产生，II类错误可由相似的解决方案来控制。在2.2.3节，α消耗函数是建立显著性检验合理边界的方法。同样，β消耗函数可用来寻找重复非优效检验的临界值，且能确保控制II类错误。β消耗函数随着时间的推移而增长，用来合理分配随信息量积累进行非优效检验的最大II类错误β_1，…，β_k。β消耗函数是信息量的分数。所需信息量除以累积信息量的值在0到1之间，0的β消耗函数永远等于0，1的β消耗函数永远等于β。在0和1之间的任意一点，β消耗函数与总的最大II类错误风险相等，总的最大II类错误风险由第i次非优效性检验的临界值得出。换言之，β消耗函数与“消耗”的II类错误相等，即$\beta(IF_i) = \beta_1 + \beta_2 + \cdots + \beta_i$。

正如2.2.4节中描述的同样的原因，O’Brien-Fleming函数构成了β消耗函数的最佳选择。在TSA v.0.8中，惟一有效的β消耗函数是O’Brien-Fleming消耗函数。

图11展示了包括重复非优效性和显著性检验的meta分析。在这个meta分析中，所需信息量为4000名患者。当达到2000名患者时，meta分析结果是不确定的，因为它未与有统计学意义的上界或非优效的下界相交。Z值曲线的走向表明在达到3000名患者时，meta分析如何变得有确定意义。

在A例中，Z曲线与非优效边界相交，说明试验组干预并不比对照组干预优越。在B例中，Z值与O’Brien-Fleming优效性边界相交表明试验组干预比对照组干预优越。

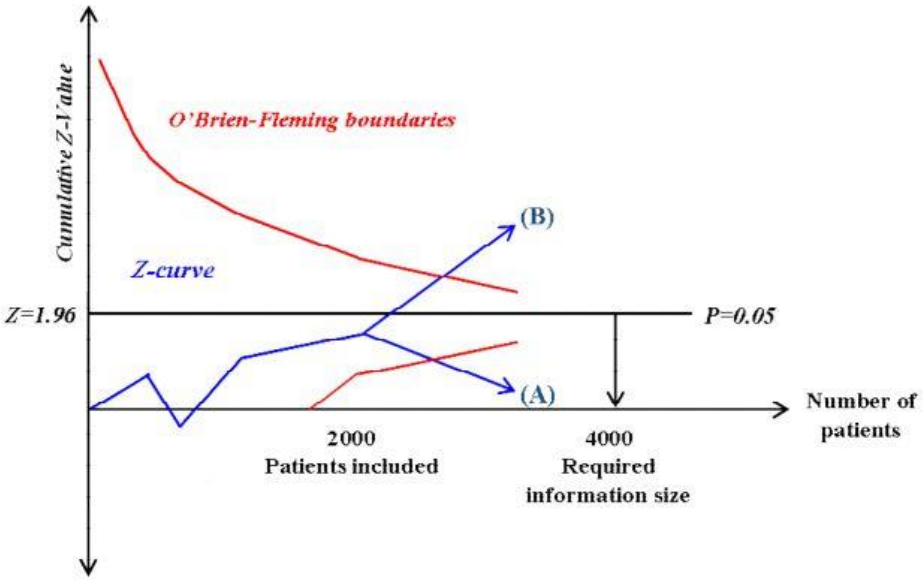


图 11 示例：包括重复非优效（红线）和显著性（棕线）检验的 meta 分析。前 2 个试验的累积 Z 曲线达到必要信息量的一半。2 个新的试验加入后，A 线表示无效（且累积 Z 分数也无意义），B 线表示干预措施显著有益（且累积 Z 分数达到显著，同时穿越一般显著性界值线和 O’Brian-Fleming 界值线）。（Cumulative

Z-Value：累计 Z 值，Number of patients：患者数，2000 Patients included：纳入 2000 例患者，4000 Required information size：需要 4000 例的信息量，O'Brien-Fleming boundaries：O'Brien-Fleming 边界，Z-curve：Z-曲线）

为了评价组间的等效性，非优效边界需与非劣效边界共同使用。假设一个两组比较的meta分析（A组和B组），如果累积的Z值低于非优效边界，A组不比B组好，但可能更差。如果累积的Z值高于非劣效边界，表示A不比B差。在这种情况下可得出A与B等效的结果。从图形上看，“等效的区域”是这两个边界相交之后的内楔的部分（见图12）。

图12是一个包含了所有TSA内容的meta分析，这些内容有：所需信息量、双侧显著性检验的边界、非优效检验及非劣效检验的边界。在该例中，所需信息量为4000。在患者数达3000左右时，Z值落在了内楔的部分，此时，可得出干预效果不比预期好的结论。

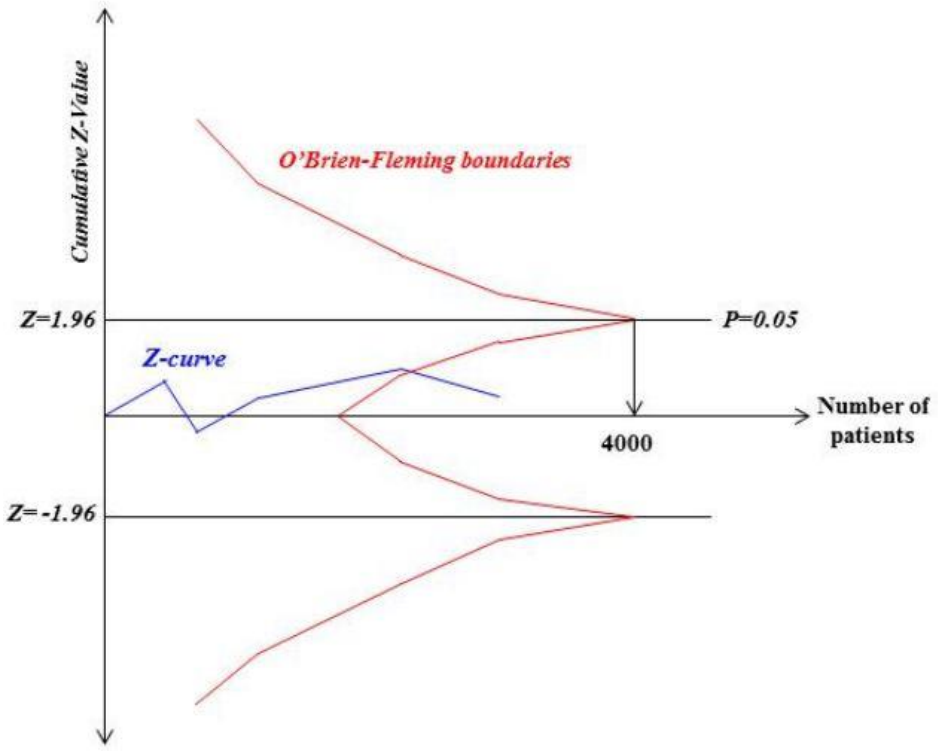


图 12 示例：meta 分析重复非优效、非劣效和显著性检验的界值。（Cumulative Z-Value：累计 Z 值，Number of patients：患者数，O'Brien-Fleming boundaries：O'Brien-Fleming 边界，Z-curve：Z-曲线）

3 TSA 软件的安装和启动

3.1 安装条件

TSA是一款基于Java™语言的程序，因此可在任何支持Java™语言的操作系统（如Microsoft Windows, Mac OS, UNIX, Linux等）上运行。运行TSA软件的电脑需要安装最新版本（或最少是较新版本）的Java（Java Runtime Environment, JRE）。Java可在这个网站：www.java.com免费下载。

截至此手册发表（2011年8月），最新Java版本为1.6。TSA软件可在JAVA1.6环境下正常运行。

3.2 安装

TSA软件安装所需的所有文件均被放在一个ZIP压缩包里提供给用户。TAS安装程序包可应用WinRAR或GZIP等软件进行解压。TAS安装程序包解压后，解压文件夹里会出现*TSA.jar*, *RM5Converter.jar*和*TEMPLATES*. TPL三个文件，同时有两个名称分别为lib和samples的文件夹。

TAS.jar压缩包中包含TSA应用程序；RM5Converter.jar压缩包中包含一个特定程序，能够实现将RevMan5中的数据（目前仅适用于二分类结局指标）转化为适用于TSA分析的数据格式。

TEMPLATES.TPL文件包含监测边界模板（monitoring boundary templates），在对不同资料分别进行序贯分析时可作为默认值使用。模板文件的内容受TSA软件控制。文件夹lib中包含TSA程序应用所需的的多种支持包。文件夹samples中则包含本手册提供的示例'.TSA'文件（详见第五章）。

将整个TSA压缩包解压至硬盘上某一个文件夹中即完成了本程序的安装。无需进一步操作。

3.3 启动 TSA 程序

要启动 TSA 软件，双击 TSA.jar 的文件即可。

另外该软件也可以通过提示符启动。首先打开命令提示符，浏览至TSA软件解压后的文件夹，键入'java –jar TSA.jar'。

如果您使用的是 Microsoft Windows 操作系统，您可以通过单击“开始”按钮（通常在屏幕左下角），然后单击“运行...”项打开 DOS 提示符。在弹出“运行”窗口时，输入'cmd“（不含引号）在文本字段中，然后单击 OK，DOS 提示符即会出现。执行 cd（更改目录）命令，浏览解压的 TSA 软件所在的文件夹。例如，如果您在 C 盘的 Program Files 文件夹内创建了一个名为 TSA 文件夹，同时解压 TSA 软件到这个文件夹，您应该首先在 DOS 提示符中将目录更改到 TSA 文件夹。这可以通过键入'cd C:\Program Files\TSA'（没有引号）来完成。在 DOS 提示符下的目录被更改后，需键入'java –jar TSA.jar'。

3.3.1 TSA 不能正常启动的原因

如果您在启动 TSA 软件时出现问题，可能有几个原因。常见的如下：

系统中是否安装了 JAVA（JRE）？

安装的 JAVA（JRE）版本是 1.6 或更高版本？

是否从 ZIP 压缩包中提取了所有文件？

是否重命名，移动或删除任何解压的文件或文件夹？

如果双击 TSA.jar 文件后时启动了不同程序（非 TSA），这说明.jar 文件扩展名没有关联到 Java（JRE）。如果发生这种情况，可以尝试通过提示符（见上文）手动启动该程序，或者尝试更改文件名关联。如果使用的是 Windows 操作系统，则可以通过如下步骤更改文件关联：

- 打开资源管理器窗口（例如，双击我的电脑），然后单击“工具”菜单
- 选择“文件夹选项...”，然后进入“文件类型”选项卡
- 在列表中找到 JAR 文件扩展名
- 点击“更改”
- 从列表中选择“Java(TM) Platform SE binary (Java（TM）平台 SE 二进制文件)”，然后单击确定（OK）
- 如果' Java（TM）平台 SE 二进制“不在列表中，单击”浏览“并在 JRE 的

bin 文件夹中找到 javaw.exe。它的默认路径是：C:\Program Files\Java\jre6\bin。

如果操作系统不是 Microsoft Windows，请参考使用者手册查询相应操作系统的处理方法。

3.4 启动 Revman 5.0 (RM5)转换器

双击 RM5 Converter.jar 文件，运行 Review Manager 5 数据转换程序（目前仅适用于二分类结局指标）。

3.4.1 RM5 转换器不能正常打开的原因

该 RM5Converter.jar 和 TSA.jar 具有相同的基本条件，所以如果您打开它有困难，请参阅第 3.3.1 节。此外，RM5 Converter.jar 要求 TSA.jar 文件能够运行。因此，TSA.jar 必须和 RM5 Converter.jar 位于同一个文件夹。

4 如何使用 TSA

4.1 启动

TSA 启动后会弹出类似于图 13 的窗口。启动窗口有一个菜单栏，包括菜单文件（menus File），设置（Settings），帮助（Help），以及 5 个灰色（不可选）的标签：Meta 分析（Meta-analysis），试验（Trials），TSA，图形（Graphs）和多样性（Diversity）。

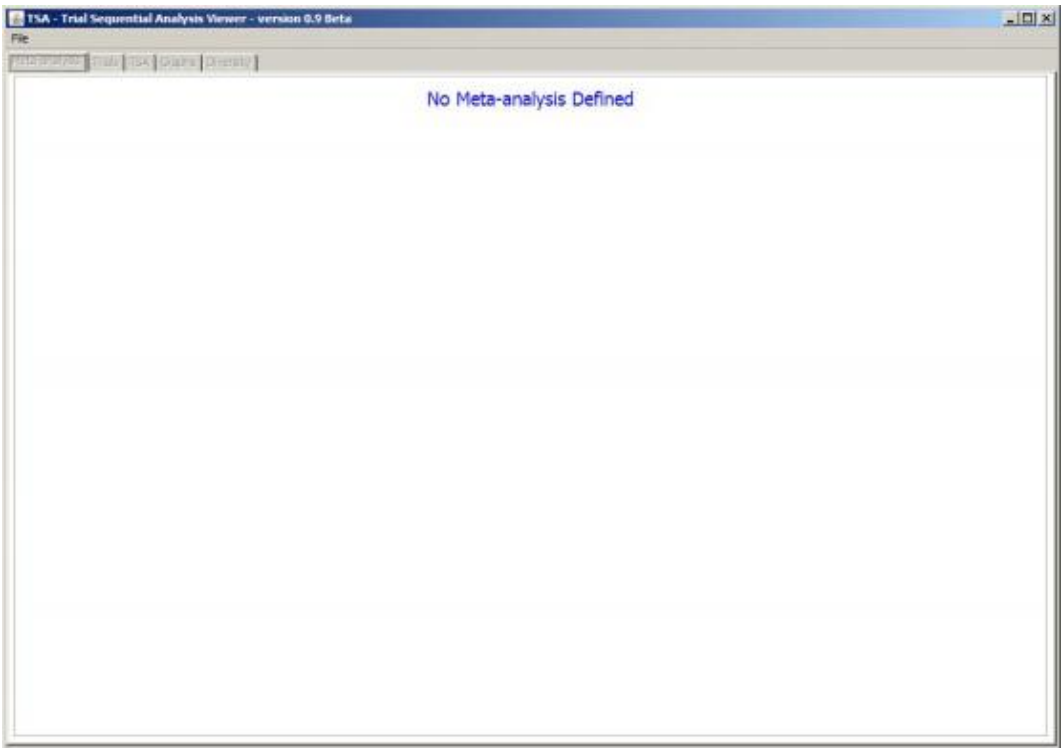


图 13 TSA 启动窗口

4.1.1 创建一个新的 Meta 分析

要创建一个新的 Meta 分析，需进入菜单栏，选择*文件*（*File*）>*新建 Meta 分析*（*New Meta-analysis*），会弹出一个对话框（图 14），在此您可以给当前的 Meta 分析命名，选择将进行 Meta 分析的数据类型（二分类变量或连续型变量），确定两个进行比较的干预措施，定义“阴性”或“阳性”结局指标类型并添加注释。单击*创建*（*Create*）按钮创建新的 Meta 分析。按*取消*（*Cancel*）按钮以取消此操作。如果您想要编辑 Meta 分析的名称，干预措施，或注释，需进入菜单栏，然后选择*文件*（*File*）>*编辑 Meta 分析*（*Edit meta-analysis*），此时

会再次弹出图 14 所示的对话框。

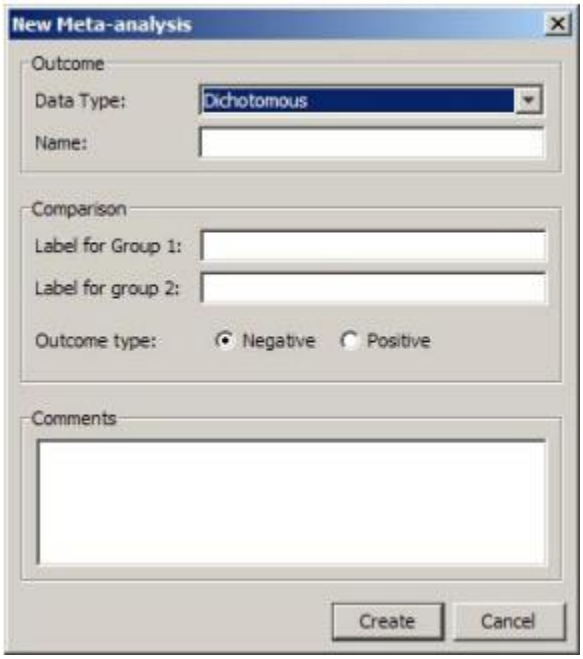


图 14 新的 Meta 分析创建对话框

注：对于二分类数据，阴性结局指标如：死亡，中风或癌症发病率;阳性结局指标如：生存率，病毒清除，或戒烟。对于连续型变量，阴性结果指平均效应增加代表负性事件（例如，抑郁评分的增加），而阳性结果是指平均值效应的增加是正性事件（如血小板计数）。 TSA 软件需要指定结局指标为阴性或阳性，以确定研究结果支持哪个组。

创建新的 Meta 分析后，左侧的开始窗口会出现多个选项。（这些选项将在 4.3 节中进行描述。自定义 Meta 分析设置。）

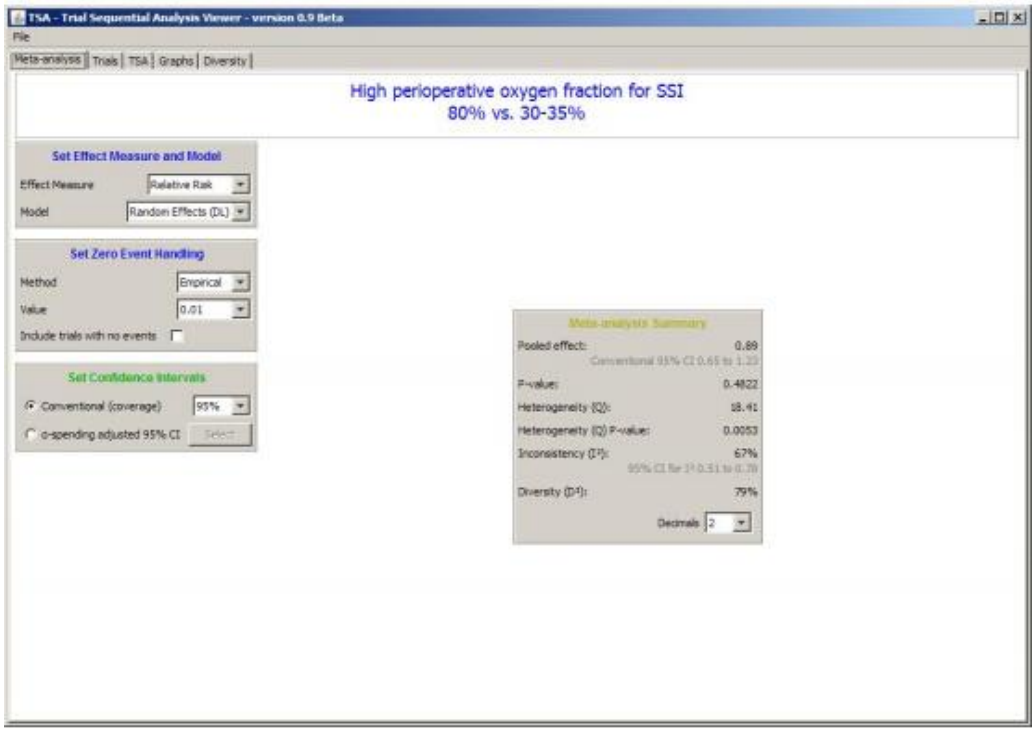


图 15 新建 Meta 分析并已输入数据的开始窗口

名为 Meta 分析摘要（Meta-analysis Summary）的对话框出现在窗口的中间。

4.1.2 保存和打开 TSA 文件

如果要保存工作，需找到菜单栏，然后选择*文件(File)>另存为...(Save as…)*。如果要对已经创建的 TSA 文件进行操作，需找到菜单栏，选择*文件（File）>打开（Open）*，选择您希望继续操作的 TSA 文件。

4.1.3 从 RevMan5 文件中导入 Meta 分析数据

需要使用另外一个单独的 RM5 转换软件导入在 Review Manager 5(*.RM5）中保存的的 Meta 分析数据（目前只适用于二分类结局指标），。该软件包含在之前下载的 TSA 安装包中（见第三章）。

图 16 从 Revman5 中导出分析数据窗口，选择希望导出的 Meta 分析数据（目前仅适用于二分类结局指标）为后缀名为 csv 的文件。本例转换的 RevMan 文件例子是一篇题为“聚乙二醇化干扰素α-2a 对比聚乙二醇化干扰素α-2b 治疗慢性丙型肝炎”的 Cochrane 综述[33]。

RM5 转换器可以读取逗号分隔的文件(*.CSV)。因此，第一步是把 RevMan 文件转换成以逗号分隔的文件。在 Revman5 中打开 RevMan 文件，在菜单栏，选择*文件（File）>导出(Export)>数据和分析(Data and analyses)*。此时会弹出一个包含树状结构的复选框（图 16）。检查要导出为以逗号分隔的文件的 Meta 分析数据，然后单击*下一步（Next）*按钮。在下面的屏幕中查看前三个复选框：Comparison Number（对比组数），Outcome Number（结局数），和 Subgroup Number（亚组数）。然后点击*完成（Finish）按钮*。

请注意，如果单击*下一步（Next）*按钮两次，您将看到一个选择字段分隔符（在数据中分离单元格）的选项。值得注意的是字段分隔符是一个逗号（默认值）。

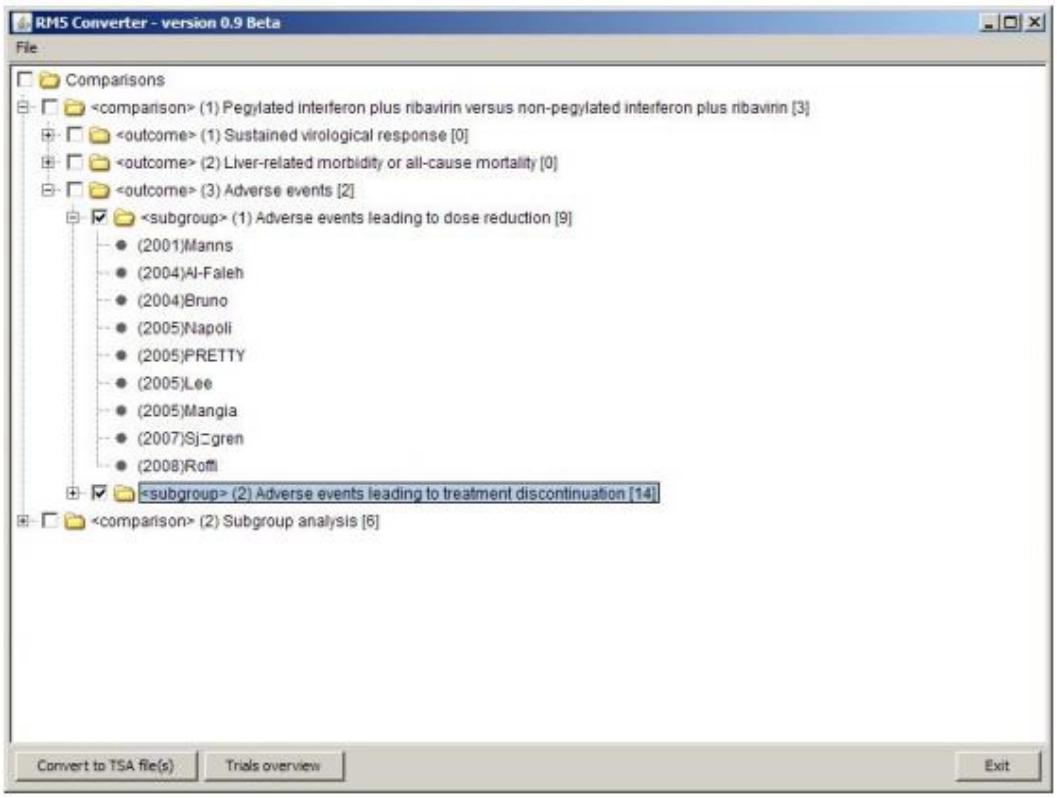
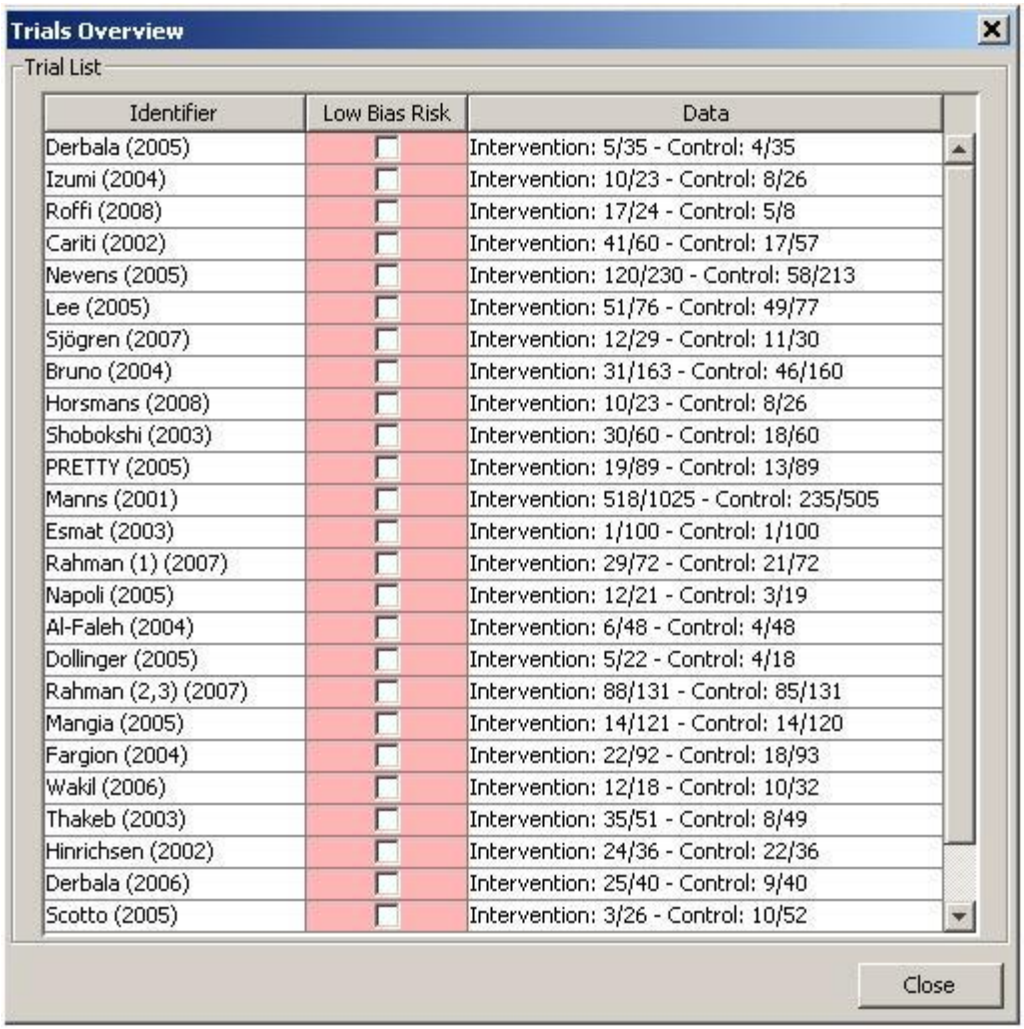


图 17 核对 RM5 转换器中的树状结构

数据导出成一个*.csv 文件后，通过双击图标打开 RM5 转换器，在菜单栏选择*文件（File）>打开（Open）*，随即包含树状结构的复选框将出现在应用程序窗口（图 17）。数据结构与在 Revman5 中一致。对于每个对比组，可以有多种结局指标，同时每个结局指标代表一个 Meta 分析。如果 Meta 分析中包含亚组分析，则各亚组将被列在每一个结果下。如果一个对比组被选中，该比较下的所有结果都将会自动进行检查。此外，如果一个结果被选中，该结果下的所有亚组都将会自动进行检查。每一个比较，结果，亚组下的所有试验均会自动包括在内。有时不必将所有给定的比较下的研究进行转换。也可以“取消选中”您不希望转换的对比组/结局和/或亚组。如果亚组中的试验中是唯一的，可以选择把这些亚组结果整理到一个分析框中。



Identifier	Low Bias Risk	Data
Derbala (2005)	☐	Intervention: 5/35 - Control: 4/35
Izumi (2004)	☐	Intervention: 10/23 - Control: 8/26
Roffi (2008)	☐	Intervention: 17/24 - Control: 5/8
Cariti (2002)	☐	Intervention: 41/60 - Control: 17/57
Nevens (2005)	☐	Intervention: 120/230 - Control: 58/213
Lee (2005)	☐	Intervention: 51/76 - Control: 49/77
Sjögren (2007)	☐	Intervention: 12/29 - Control: 11/30
Bruno (2004)	☐	Intervention: 31/163 - Control: 46/160
Horsmans (2008)	☐	Intervention: 10/23 - Control: 8/26
Shobokshi (2003)	☐	Intervention: 30/60 - Control: 18/60
PRETTY (2005)	☐	Intervention: 19/89 - Control: 13/89
Manns (2001)	☐	Intervention: 518/1025 - Control: 235/505
Esmat (2003)	☐	Intervention: 1/100 - Control: 1/100
Rahman (1) (2007)	☐	Intervention: 29/72 - Control: 21/72
Napoli (2005)	☐	Intervention: 12/21 - Control: 3/19
Al-Faleh (2004)	☐	Intervention: 6/48 - Control: 4/48
Dollinger (2005)	☐	Intervention: 5/22 - Control: 4/18
Rahman (2,3) (2007)	☐	Intervention: 88/131 - Control: 85/131
Mangia (2005)	☐	Intervention: 14/121 - Control: 14/120
Fargion (2004)	☐	Intervention: 22/92 - Control: 18/93
Wakil (2006)	☐	Intervention: 12/18 - Control: 10/32
Thakeb (2003)	☐	Intervention: 35/51 - Control: 8/49
Hinrichsen (2002)	☐	Intervention: 24/36 - Control: 22/36
Derbala (2006)	☐	Intervention: 25/40 - Control: 9/40
Scotto (2005)	☐	Intervention: 3/26 - Control: 10/52

图18 试验概述与偏倚风险复选框

单击*试验概述*（*Trials overview*）按钮，会弹出一个包含 csv 文件中所有试验的列表。每项试验都有一个相关复选框，提示此试验是否被指定为“低偏倚风险”（默认为高偏倚风险）。使用者可以在检查（或不检查）偏倚复选框之后对指定的风险偏倚作出修改。设定好风险偏倚级别之后单击 Close 按钮结束操作。

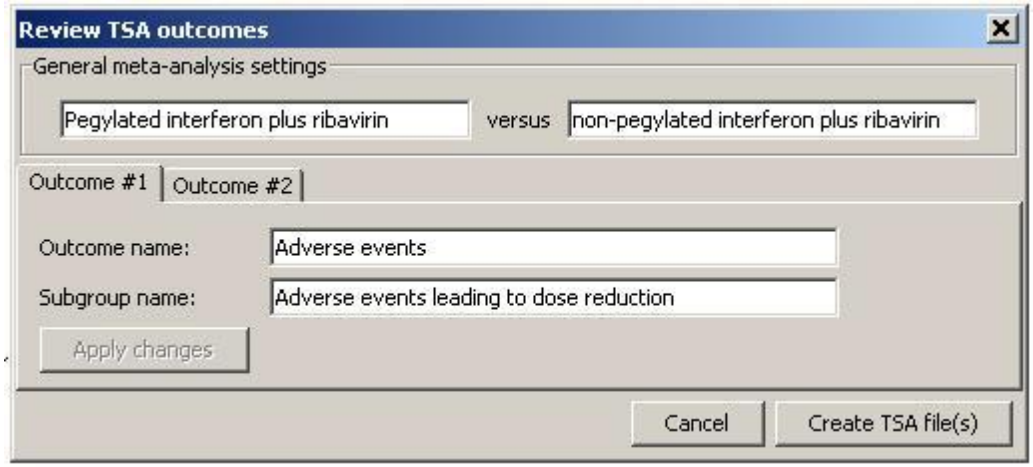


图19 浏览RM5转换器中的TSA结局指标

核对所有进行试验序贯分析时需要包含的试验对比组，结局指标和亚组分型信息，确认无误后，单击窗口底部*创建TSA文件（Create TSA file(s)）*按钮。弹出一个窗口，可供查看干预措施，结局指标和亚组分型的名称信息。浏览确认信息之后，点击*创建TSA文件（Create TSA file(s)）*按钮，并将你选择的Meta-分析另存为TSA文件于指定文件夹中。

4.2 添加、编辑和删除试验

在TSA程序菜单栏的下方有五个选项卡，分别为：*Meta分析*（*Meta-analysis*），*试验*（*Trials*），*TSA*，*图形*(*Graphs*)和*多样性*(*Diversity*)。如果需要对meta分析中的试验进行添加，编辑或删除操作时，首先要选择*试验*（*Trials*）选项卡（图20）。

图20 单击试验（Trials）选项卡对meta分析中的试验进行添加，编辑或删除操作

在窗口（试验（Trials）选项卡下）左侧会出现三部分内容，分别为：添加二分类变量/连续性变量性试验(*Add Dichotomous/Continuous Trial*)，编辑/删除试验(*Edit/Delete Trial*)，和忽略试验(*Ignore Trials*)。

4.2.1 添加试验

要添加新的试验，需要在*添加二分类变量/连续性变量性试验*(*Add Dichotomous/Continuous Trial*)区填写相关信息。不论meta分析选用的数据为何种类型，都需分别在“*研究*(*Study*) :”和“*年份*(*Year*):”栏中输入研究名称或标题（通常是研究标题缩写或第一作者姓氏）及研究发表年份字段。此外，可以对试验是否为低偏倚风险等级做出评价并勾选。

在分析二分类结局指标时，需要分别输入（试验）干预组和对照组中事件发生数和各组病人总数（图21）。

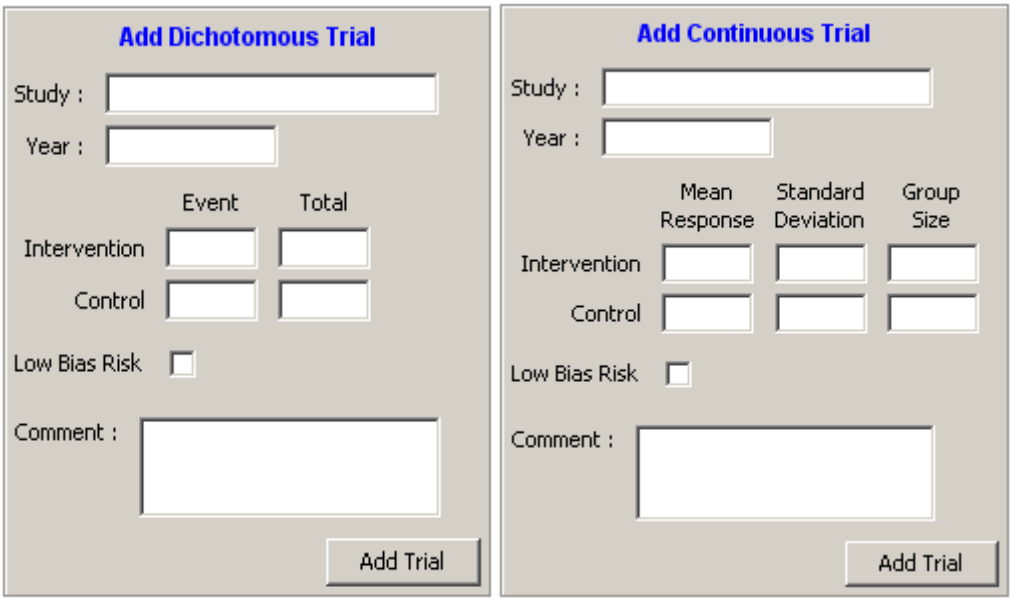


图21 二分类变量和连续型变量数据输入框（从左至右依次）

在分析连续型变量相关数据时，需要分别输入（试验）干预组和对照组的平均值，标准差和每组样本量(病人数)（图21）。可以（不必需）添加数据注释。

点击*添加（Add）*按钮可提交输入数据。

在窗口右侧共有四列，分别为研究(study)，偏倚风险(Bias risk)，忽略(Ignore)和数据(Data)。添加试验之后，这些试验的名单会出现在图22中。

第一列（从左侧起）内容为新添加试验的名称和出版年份，填写格式为“（年）名称'。第二列显示为各试验已指定的偏倚风险等级；试验的偏倚风险等级可以是'低'（绿色字母）或'高'（红色字母）。在第三列中您可以选择在进行meta分析时忽略一个或多个附加试验；点击忽略(Ignore)栏下复选框可忽略相应的试验数据。第四栏则显示各试验数据。二分类变量的填写格式为“干预组：事件发生数/样本总数。对照组：事件发生数/样本总数”。而连续型变量数据的格式是“干预组：平均效应/标准差/组样本量。对照组：平均效应/标准差/组样本量。”

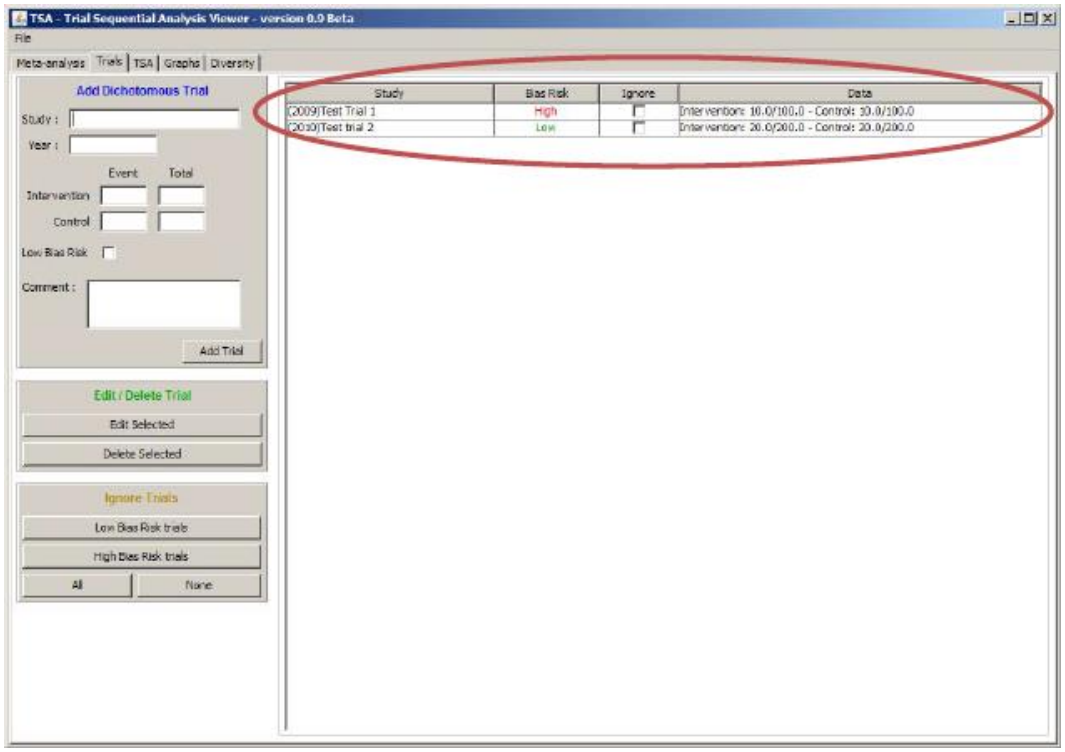


图22 红色椭圆标志内为已添加试验的列表

4.2.2 编辑和删除试验

首先选中需要编辑的试验行，然后点击*编辑/删除所选试验区（Edit/Delete Selected Trial area）*内的编辑试验（Edit Trial）按钮(图23)，即可对相应试验数据进行编辑，也可直接双击要编辑的试验行。

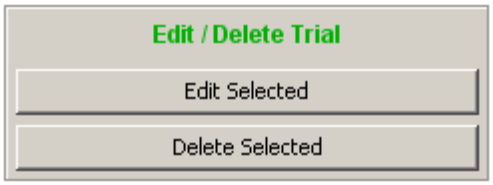


图23 编辑/删除所选试验区

编辑试验数据和添加新试验为同一个按钮，只不过此时按钮为*编辑试验（Edit Trial）*按钮而非*添加试验（Add Trial）*按钮。更改需要编辑字段内的数据后单击*编辑试验（Edit Trial）*按钮完成操作。

选中需要删除的试验行然后单击*删除试验（Delete Trial）*按钮，即可删除此试验。也可以在选中该试验行后按键盘上的<删除>（<Delete>）键来进行直

接删除。

4.3 自定义 Meta-分析设置

TSA程序提供多种选择进行Meta-分析，含多种效应估计方法，多种统计模型，零事件处理方法（针对二分类结局指标）和可信区间范围。这些选项均可*在试验（Trials）*选项卡左侧的*Meta-分析（Meta-analysis）*按钮内进行设置（图24）。

图24 单击Meta-分析（Meta-analysis）按钮设置效应估计类型，统计模型和零事件处理方法

窗口左侧为*效应估计和模型设置（Set Effect Measure and Model）*区，*零事件处理设置（Set Zero Event Handling）*区和*可信区间设置（Set Confidence Intervals）*区（图25-28）。窗口中间是Meta-分析概要（*Meta-analysis Summary*）区。

4.3.1 选择关联测量

TSA程序中效应估计方法与Revman 5文件中一致（见2.1.1：效应描述部分）。单击*效应估计和模型设置（Set Effect Measure and Model）*选项下的效应估计（Effect Measure）下拉按钮，则出现多种效应估计可选项，选择用于Meta-分析的效应估计项即可。

图25 在*效应估计（Effect Measure）*下拉框中选择要进行的效应估计项

4.3.2 选择统计模型

TSA程序提供了四种可用于合并meta-分析数据的统计模型，其中三种基于随机效应模型（见2.1.2）。单击*模型（Model）*下拉按钮后，在下拉框中会显示多种可选效应模型（图26，左侧图标志区），单击用于Meta-分析的模型完成设置。

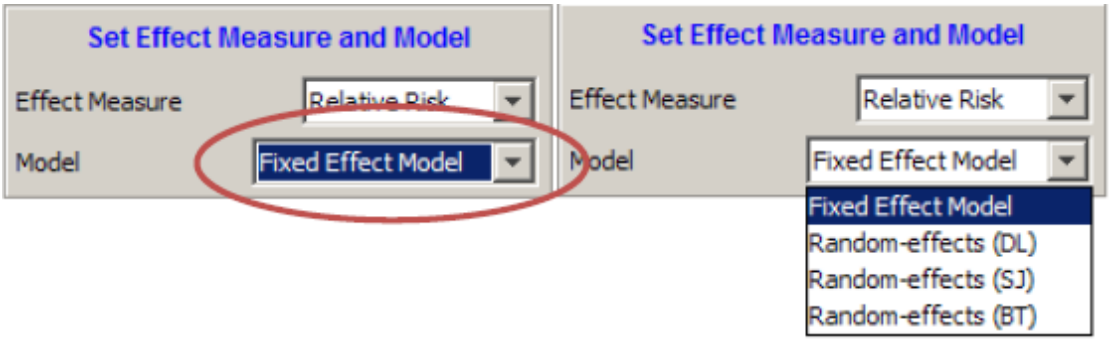


图26 在*模型（Model）*下拉框中选择所要进行的效应估计项

4.3.3 选择零事件数据处理方法

TSA程序提供三种零事件数据处理方法（见2.1.4）。单击*方法*（*Method*）下拉按钮会显示所有可用的连续性校正法，然后单击所要选择的方法即可完成设置（图27）。

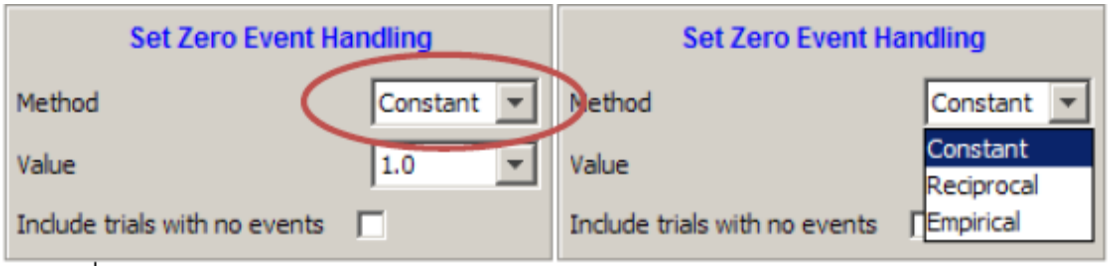


图27 在*方法*（*Method*）下拉框中选择需要的连续性校正法

有时需要设置连续性校正因子。TSA程序中，校正因子为两组校正因子总和（也称“校正因子值”）。例如，Revman中计算连续性校正因子总和的方法为1=0.5+0.5-因为两组事件数计算时均增加了0.5。单击*值*（*Value*）下拉按钮，选择矫正因子总和计算方法。当然，也可以选择将连续性矫正法用于零事件试验，选中含零事件试验（Include trials with no events）后的方框即可。

4.3.4 选择可信区间的类型

TSA程序提供多种可应用的可信区间类型（图28）。可信区间范围可设置为95%，99%，99.5%或99.9%。单击选中*可信区间设置*（*Set Confidence Intervals*）区中的“常规性（范围）（Conventional (coverage)”单选框，点击右侧下拉按钮选择所需区间范围。

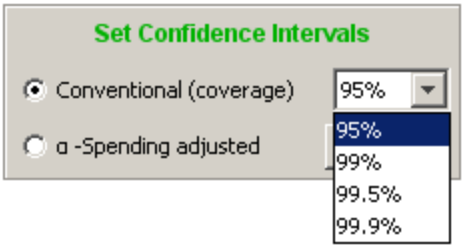


图28 选择常规可信区间范围

如果应用α-消耗函数构建并校正了显著性检验边界（见2.2.4和4.4.1），也可以选择α-消耗校正性可信区间（见 2.2.5）。先单击选中*可信区间设置*（*SetConfidence Intervals*）区中的“α-消耗校正性”（α-spending adjusted）按钮，然后单击“选择”（Select）按钮即可（图29）。

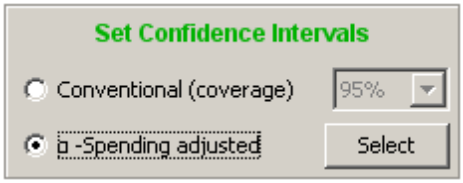


图29 选择α-消耗函数校正后的可信区间

显示屏中央弹出的窗口包含您填加的α-消耗边界列表。选择校正基于的α-消耗边界并单击选择（Select）按钮（图30）。注意，校正了可信区间范围的α-消耗函数的累积性可信区间范围与已选α-消耗函数设置的α水平相对应。

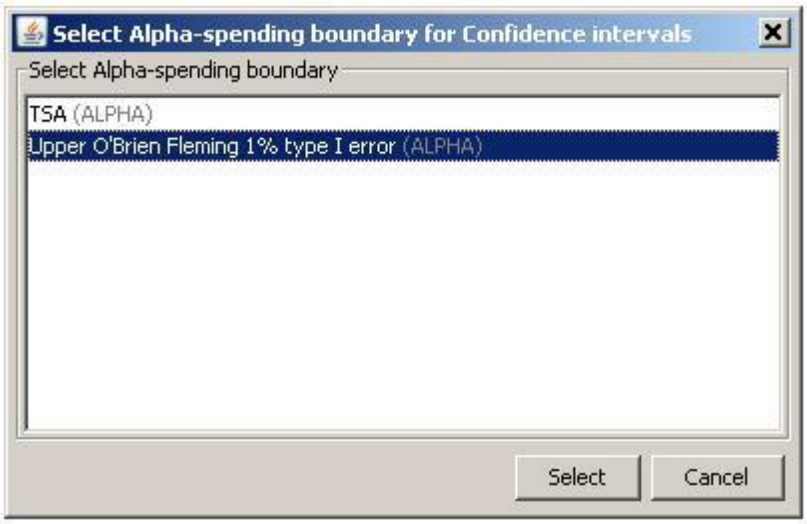


图30 选择基于α-消耗边界校正的可信区间

需要注意的是只有能够用于计算并没有被“忽略”的α-消耗边界才可以出现在列表中（见4.4.1节）。

4.4 校正后的显著性检验应用（应用 TSA）

TSA程序现提供两种显著性检验校正方法，分别为*O'Brien-Fleming α-消耗法*（详见2.2.4）和迭代对数法（详见2.2.6）。TSA程序中还包含*O'Brien-Fleming α-消耗法*和无效性检验相结合的方法（详见2.2.7）。使用上述方法，需单击*TSA*选项卡（位于试验（*Trials*）*选项卡右侧*），如图31所示。

图31 单击TSA选项卡选择显著性检验校正方法

4.4.1 添加显著性检验

该窗口的左上侧为*添加*（*Add*）区，包含*常规检验边界*（*Conventional Test Boundary*），α-*消耗边界*（*Alpha-spendingBoundaries*）和迭代对数法（*Law of the Iterated Logarithm*）按钮。单击其中任意按钮都会在TSA程序界面中弹出新的窗口。该窗口包含多个字段，可以用来界定设置需要的显著性检验类型。

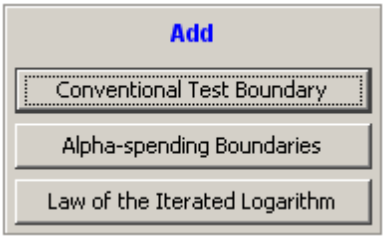


图32 单击任意按钮添加新的显著性检验

常规显著性边界

常规性（*Conventional*）选项允许添加Z曲线边界，此边界对应于某个用来检验一些最大I型错误风险即α的单一显著性检验。例如，双侧α=5%的单一显著性检验的常规边界为两条处于1.96和-1.96的水平线。单击*常规性*（*Conventional*）按钮，会弹出图33所示窗口。

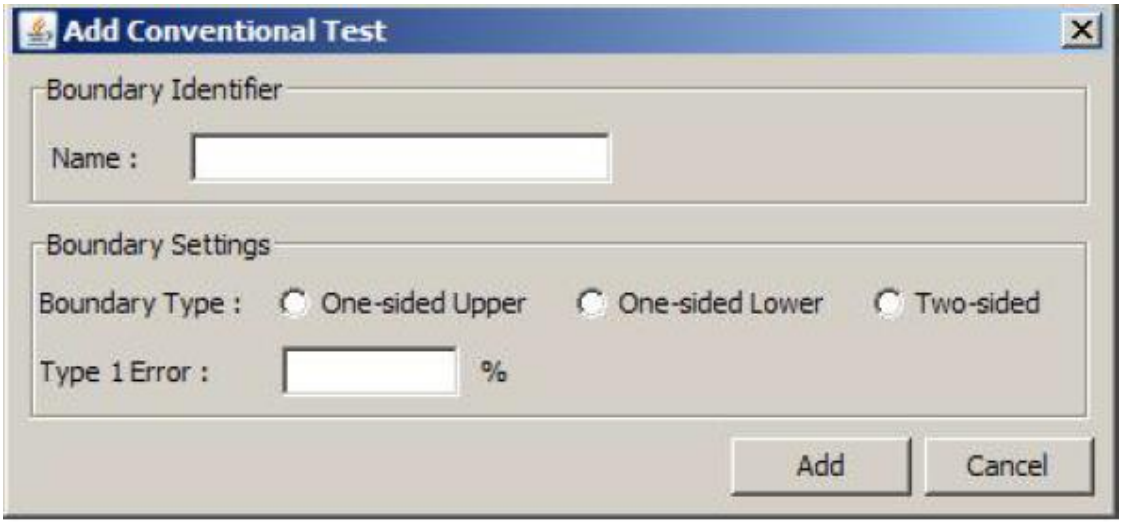


图33 点击常规检验界值按钮，将出现一个常规检验设置弹出式窗口。

需要对所要进行的常规检验进行命名（如" single test 5% threshold (5%临界值检验)"），并明确所进行的检验是双侧的（对称的）还是单侧的，允许犯 I 类错误（单个检验中）的总和是多少。对于单侧检验，单侧检验上限（*Upper* one-sided test）仅适用于干预措施的优效性检验，而单侧检验下限（*Low* one-sided test）仅适用于对照措施的优效性检验。对二分类结局指标的 meta 分析，应注意的是当结局指标为阳性指标而非阴性指标时（见 4.1.1 节），上限和下限检验的功能是相反的。当已对常规界值进行命名并明确设置后，可点击添加按钮来添加界值。

α-消耗界值

α界值选项允许采用α-消耗方法来添加Z曲线（Z-curve）的经校对的显著性界值，该方法已在2.2.4进行描述。由于α-消耗方法在某些必要的meta分析中因信息量不明时无法使用，而信息量的估算又必须同步界定。因此，在信息量估算方面，二分类变量meta分析的α界值的设置窗口与连续型变量meta分析是不同的。对于二分类变量meta分析，当点击α界值按钮时，将出现一个如图34的α界值设置窗口。

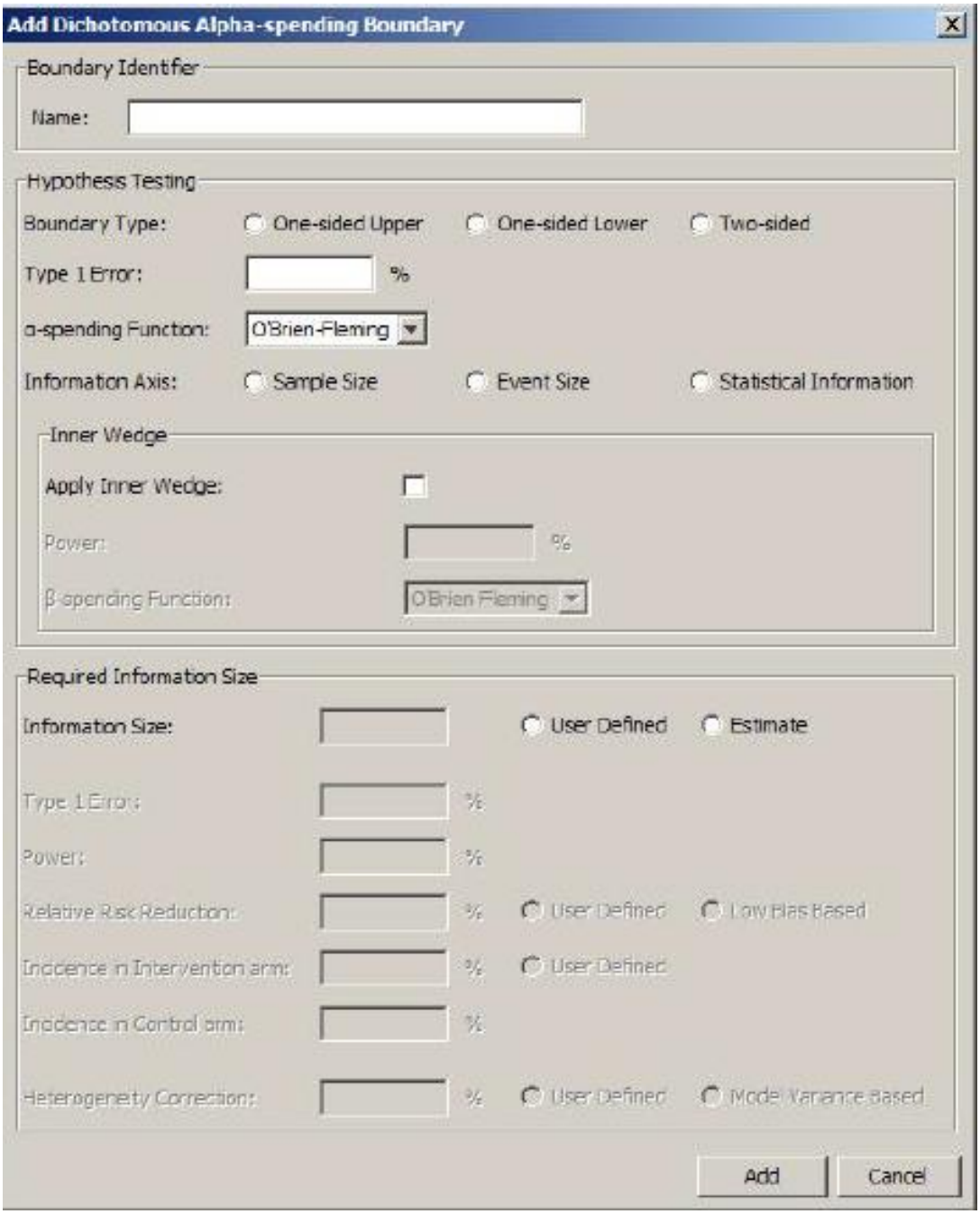


图34 对于二分类变量的meta分析，点击*α-消耗*按钮，将出现一个*α-消耗*检验设置窗口。

对于连续型变量的meta分析，当您点击α-*消耗*按钮时，将出现一个如图35的α-*消耗*界值设置窗口。

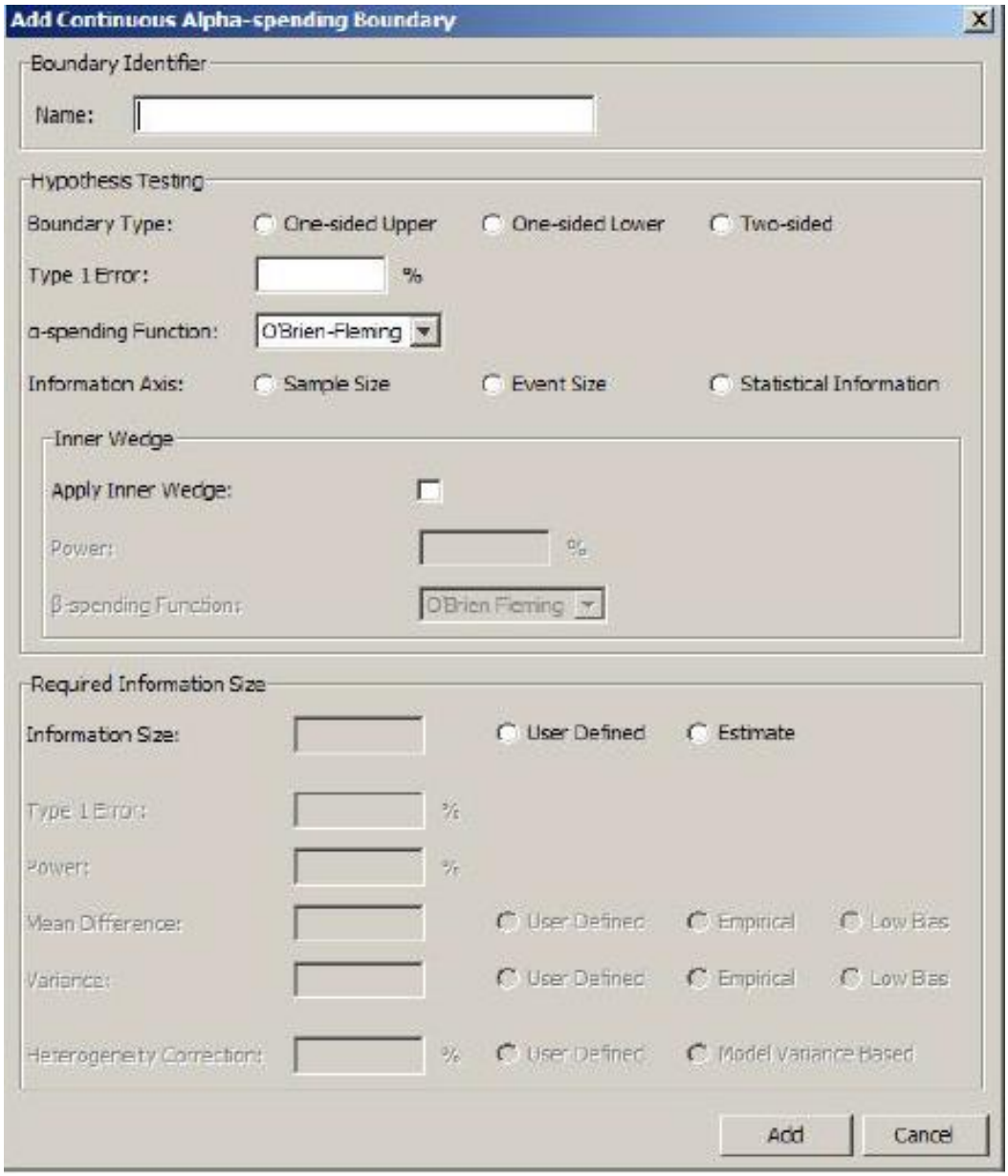


图35 对于连续型变量的meta分析，点击α-*消耗*按钮，将出现一个α-*消耗*检验设置窗口。

首先，要对进行的α-*消耗*检验进行命名（如 5%对称的 O'Brien-Fleming）。然后，需要明确要执行的是双侧还是单侧检验，所允许犯 I 类错误的总和是多少，及要执行哪种α-*消耗*（目前仅具备 O'Brien-Fleming 功能）。接下来需要确定是否希望将 meta 分析中的信息定义为累积病例数（样本量），累积事件数（事件量），或累积性统计学信息。此外，对于单侧检验，单侧检验上限（*Upper* one-sided test）仅适用于试验措施的的优效性检验，而单侧检验下限（*Low* one-sided test）仅适用于对照措施的有效性检验。对二分类变量的 meta 分析，应注意的是当结局指标为阳性指标而非阴性指标时（见 4.1.1），上限和下限的检验功能是相反的。

在"内插应用"复选框上打勾来检测无效性（如应用内插无效界值）。当输入信息量估算参数时，无效界值的 II 类错误（或把握度）将被自动设置（见下文）。目前，TSA 中唯一可用的β-消耗功能为 O'Brien-Fleming 功能。

需要输入必要的内容来进行信息量估算。可以选择将所需要的信息定义为任

意数字，这些数字可以独立于 TSA 软件获取。要递交用于计算信息量的值，请勾选“用户定义”按钮，并输入所需要的信息量值。也可以选择根据 2.2.1 部分描述的方法来估算所需要的信息量值。要使用 TSA 来估算所需要的信息量值，请勾选“估计”（Estimate）按钮。所需的信息量估计值将自动生成所累积的信息类型。举个例子，如果在信息轴（Information Axis）下面选择样本量，所需的信息量将自动生成 meta 分析所需的病例数。

信息量值的计算是自动基于所定义的α界值对应的最大I类错误，但需要在“把握度”输入框中输入设定的最大 II类错误/最小把握度（II类错误）。

当Meta分析存在异质性时，有两种方法可以用来校正所需要的信息量值。第一种是基于所选的随机效应模型的方差和固定效应模型的方差的估计比率（见2.2.1节）。要使用这个选项，需在“基于方差模型”按钮上打勾。值得注意的是，如果选择固定效应模型，校正因子将总等于1，也就是实际上并没有进行校正。

第二个选项是对预测的异质性进行猜测估计。当meta分析纳入的试验数量不足以可靠的估计校正因子时，可校正所需要的信息量以便得到先验最大或合理的预期异质性。

要进行信息量校正，需勾选“用户定义（User Defined）”按钮，在左侧的输入栏里输入最大预期的异质性。这里的异质性被定义为meta分析总方差的百分比（用于解释试验间的变异而非试验内的变异）。因此，当试验间的变异占总变异的50%左右时，用户定义的50%的校正值可产生一个需要的信息量值以保证一个可靠的推断。

对于二分类结局指标的meta分析，设置预期的事件率和干预效应，需在以下三个输入栏中任意两个内填入相应的数值：“相对危险度减低率（Relative risk reduction）”，“干预组事件发生率(Incidence in Intervention Group)”和“对照组事件发生率(Incidence in Control Group)”。如果纳入的部分试验被归为低风险偏倚研究，则可以将这些试验的合并meta分析值估计为预期的相对危险减低（anticipated relative risk reduction）。具体操作需要选择“基于低偏倚（Low-bias Based）”选项。

对于连续型变量的Meta分析，设置预期的均差和变异，需在均差和变异两个输入栏内填入相应的数值。如果纳入的部分试验为低风险偏倚的试验，可使用合并的Meta分析估计这些试验的预期均差和变异，同样选择“基于低偏倚(Low-bias Based)”选项。也可以选择使用合并纳入试验的估计值作为预期的变异（无论偏倚风险如何）。要使用所有的试验，选择“根据经验的(Empirical)”的选项。

确定了α*消耗边界*，定义假设检验设置以及信息量计算参数，然后点击“添加(Add)”按钮，以添加界值。

当添加*α消耗*界值后，需要明确之前进行的Meta分析进行了显著性检验。跳转至中期分析试图（interim looks）校正显著性检验清单右侧，勾选（或不勾选）已经进行显著性检验的试验。在图36，试验2,4和5被勾选，试验1和3未被勾选，这意味着执行了三个meta分析（包括显著性检验）：一个包括试验1和2，一个包括试验1到4，一个包括试验1到5。注意：清单中的最后一个试验应该被勾选，因为这意味着您将对所有纳入的试验进行显著性检验。

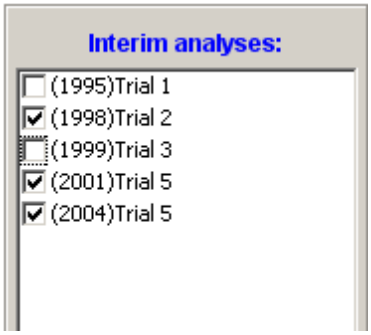


图36 当单击α界值按钮时将出现连续型变量meta分析的α界值设置窗口

在某些情况中，您可能希望勾选或不勾选先前进行过显著性检验的所有试验。在中期分析（Interim analyses）区域底部单击“取消选择（Select none）”按钮，以便取消选择所有试验，或者单击“全选（Select all）”按钮以选择所有试验（图37）。另外，可以点击中部按钮的逆转选项按钮，取消之前选定的所有试验，而选定之前未选定的所有试验。

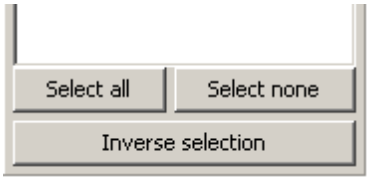


图37 勾选或取消先前显著性检验的所有试验

迭代对数法则惩罚Z曲线

迭代对数法则选项可以通过应用2.2.6节描述的方法对Z曲线进行惩罚，运行校正的显著性检验。当您单击迭代对数法则（Law of Iterated Logarithm）按钮，会出现如图38的窗口。

您需要给Z曲线的惩罚命名（如，5%对称LIL），定义试验的检验为双侧或单侧检验，总的I类错误概率，以及定义惩罚参数，1（见2.2.6部分和表2）。对于单侧检验，上限单侧检验将只测试试验干预措施的优效性，而下限将只测试对照措施的优效性。对于二分类数据的meta分析，必须注意的是当结局被定义为“阳性”结局指标而不是“阴性”结局指标时（见4.1.1），上限和下限的功能将会颠倒。

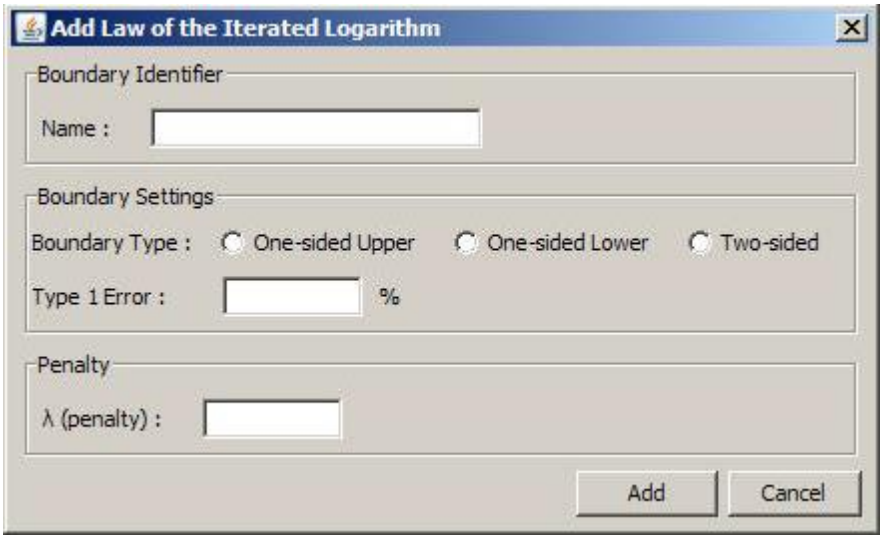


图38 当单击迭代对数法则按钮时出现迭代对数法则惩罚设置窗口

4.4.2 编辑和删除显著性检验

新加入的显著性检验将会出现在屏幕中部。每一个新增加的显著性检验都将如图39所示，显示在一行。

Identifier	Ignore	Type
TSA (ALPHA)	☐	Alpha-spending
Conventional 2-sided (CON)	☐	Conventional
LIL 5% (LIL)	☐	Law of Iterated Logarithm

图 39 添加的显著性检验列表。

要编辑一个显著性检验，首先选择需要编辑检验的相应行，然后在“*Edit（编辑）*”区域单击“Edit selected（编辑选中行）”按钮（图40）。或者直接双击待编辑的相应行。

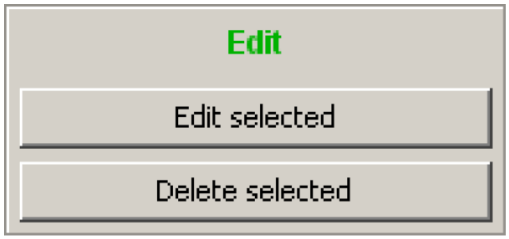


图40 编辑/删除选定的显著性检验。

即弹出显著性检验的设置窗口。您可以对检验进行编辑然后在弹出窗口的右下角单击“Apply changes（应用更改）”按钮。

如果您希望删除一个显著性检验，选择需删除检验所在的行，然后在*Edit（编辑）*区域按住“*Delete Selected（删除所选项）*”按钮，或者选择希望删除检验所在的行，然后在键盘上按<Delete（删除）>按钮。

4.4.3 添加和加载显著性检验模板

在TSA表左下角的*Templates（模板）*区域提供了保存完成构建的显著性检验以及加载之前构建的显著性检验(图41)的按钮。如果希望在其他的meta分析中再次使用相同的显著性检验，可以将该显著性检验保存在模版中并在需要应用时在软件中加载这个显著性检验。将构建的显著性检验保存为模版，请选择希望保存的检验相应行，然后按“Save astemplate（保存为模板）”按钮。要加载之前保存的模版，首先单击“Manage templates（管理模板）”按钮。

图41 *Templates*（模版）区域可以下载和保存(增加)构建的显著性检验。

弹出窗口将出现在TSA程序窗口中部，可用的模版列表显示在左侧。

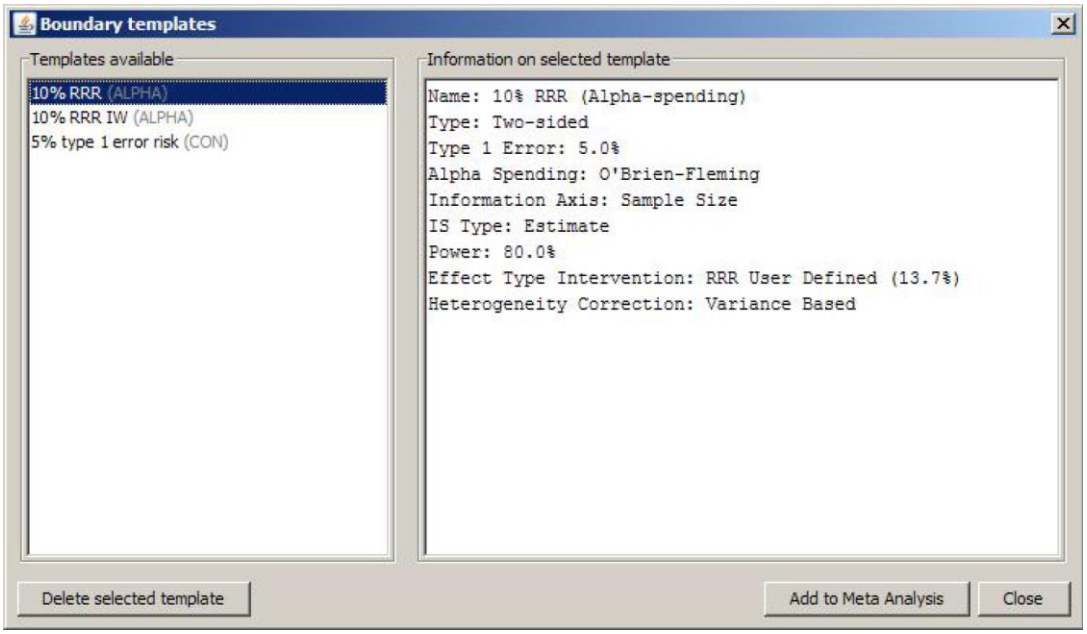


图42 *Templates*（模版）窗口。显著性检验“10% RRR”已经被选中，这个检验的设置展示在“Information on selected boundary（所选模版的信息）”下。

在右边单击一个模版标题展示可以使用的显著性检验设置。要为meta分析下载一个显著性检验的模版，选择希望下载的模版然后单击“Add to Meta-analysis（添加Meta分析）”按钮。如果要永久地删除一个可使用的模版，选择要删除的模版然后单击“Deleteselected Template（删除所选模版）”按钮。

4.4.4 显著性检验的运算

一旦已经加入所有希望使用的显著性检验，则需要 TSA 程序进行必要的运算。

为了实现运算，在 *Edit*（编辑）区域下的 *Calculations* （运算）区域中单击“Perform calculations（进行运算）”按钮。依据加入的显著性检验的数量多少，TSA 程序可能会花几秒钟完成以计算。如果中期分析数量较多，进行O’Brien-Fleming（奥布莱恩-弗莱明）(α-消耗)类型界值计算时可能需要 5-10 秒。

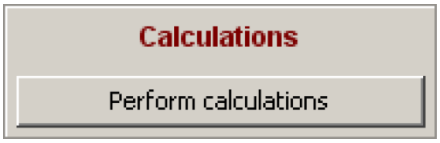
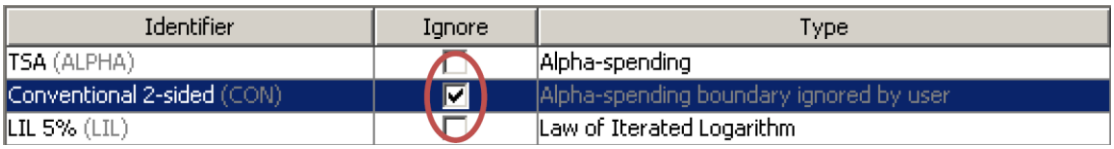


图43 进行运算的按钮

在一些情况下，在两个中期分析中增加相对小的信息，α-消耗边界的数值计算（极小的尾概率的数值积分）会发生错误。例如，如果需要的信息大小是20000例患者，中期分析在每一个试验之后运算，增加一个新的40名患者的试验仅增加计算信息部分的0.2%。为避免计算错误，TSA的程序会自动删除（未检查）对应于一个信息部分增量的1％或更小的中期分析。这种情况发生时，在TSA程序中间会自动弹出窗口显示中期分析已经被删除（图44）。但这些试验的数据仍然存在于TSA meta分析及累计的Z值中。

图44 中期分析已经被删除试验的弹出窗口。但这些试验的数据仍然存在于TSA meta分析及累计的Z值中。

如果已经添加了至少一个显著性检验并且不希望对所有这些进行运算，则可以选择忽略显著性检验。要忽略一个显著性检验，将希望忽略的显著性检验相应行的中间栏检查框打钩。



Identifier	Ignore	Type
TSA (ALPHA)	☐	Alpha-spending
Conventional 2-sided (CON)	☑	Alpha-spending boundary ignored by user
LIL 5% (LIL)	☐	Law of Iterated Logarithm

图45 忽略一个显著性检验的例子（忽略“Conventional 2-sided（传统双面）”）

累计Z曲线和显著性边界（α-消耗函数）可以使用样本大小、事件大小或统计信息三个变量中的一个显示。不同规模定义的显著性检验不能同时在一个表中显示，所以需要为所有分析选择其中的一个变量。在计算区域下的信息轴区域中单击合适的单选按钮。

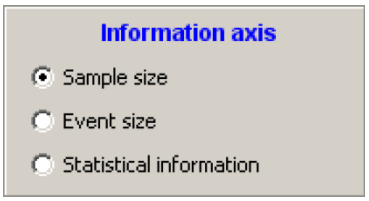


图46 显示了累积性显著性检验选择的信息规模的单选按钮

在任何添加的α-消耗边界中尚未选定的三个部分（样本大小、事件大小或统计信息）中的一个或两个会自动在信息轴区域显示为灰色。

4.5 TSA 图示选项

TSA程序的图允许显示Z-曲线和构造的显著性检验有关的证据（即，病例累积数、事件或统计信息）的强度。TSA程序提供了一些图形编辑选项以方便制作科研论文所需的图形。

要进入图形选项，单击图形选项卡（在TSA选项卡的右侧），如图47。

图47 单击图表选项卡查看Z曲线和依据证据强度构造的显著性检验结果

在TSA程序窗口的左侧您会发现测试和边界布局区域，设置图形布局，以及两个打印选项（“打印当前图形”和“生成TSA报告”）。这些区域的右侧，则会显示Z-曲线和构造的显著性检验图形。

在测试和边界布局区域中，有一些图形编辑选项，这些选项允许更改Z-曲线的呈现和构造的显著性检验（边界）。构造的显著性检验和Z-曲线将被列在白色区域;见图49。要改变其中之一的呈现，首先选择列表中的某一测试（曲线）然

后根据作者的需要进行编辑。

在TSA程序中，可以选择编辑颜色、线条类型、在每次试验或中期分析中显示的图表类型和大小，以及试验相关曲线或一个试验的大小和字体。也可以选择从图形中隐藏曲线/试验。

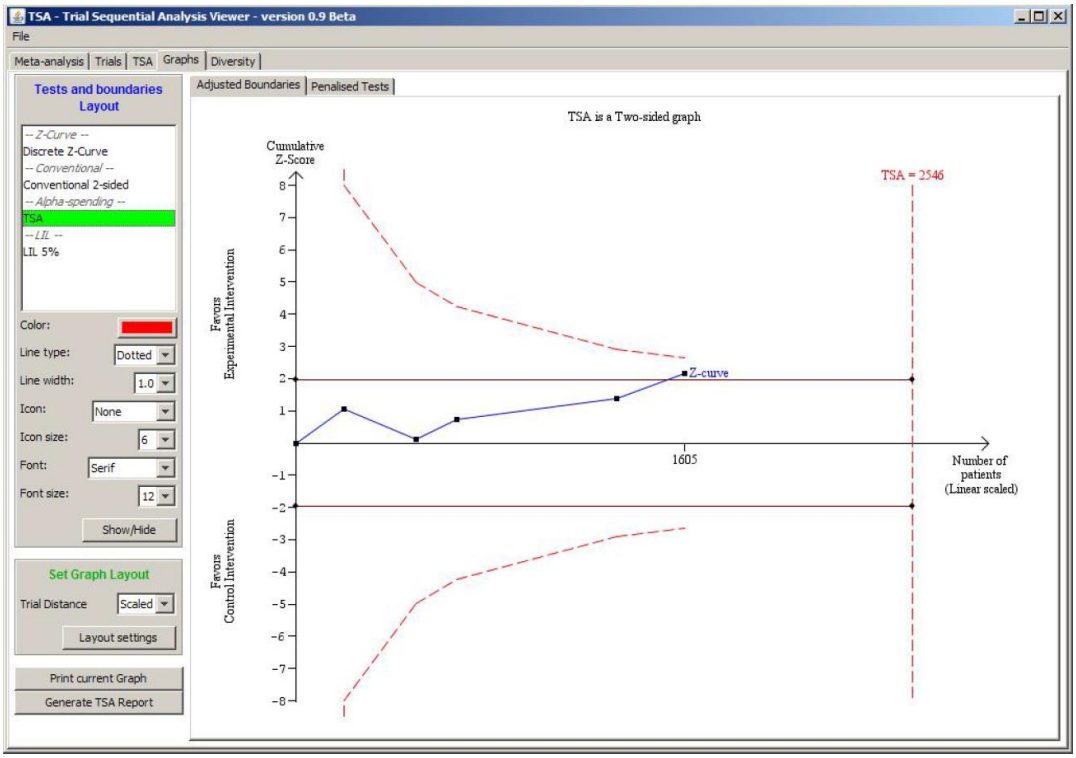


图48 图形窗口

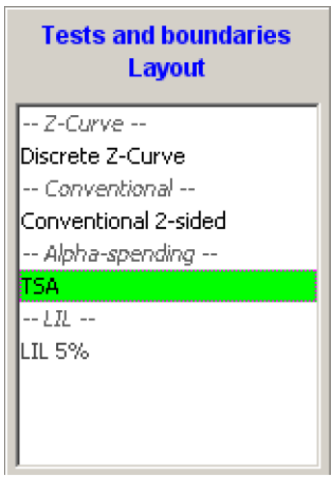


图49 显著性检验和Z-曲线列表区域

在设置图形布局区域，有一些改变一般图形显示的选项。单击“Layout settings（布局设置）”按钮会弹出一个窗口（图50），在此窗口可校正x轴和y轴宽度的选择、坐标字体大小或图形中固定的文本组件上的字体和大小。

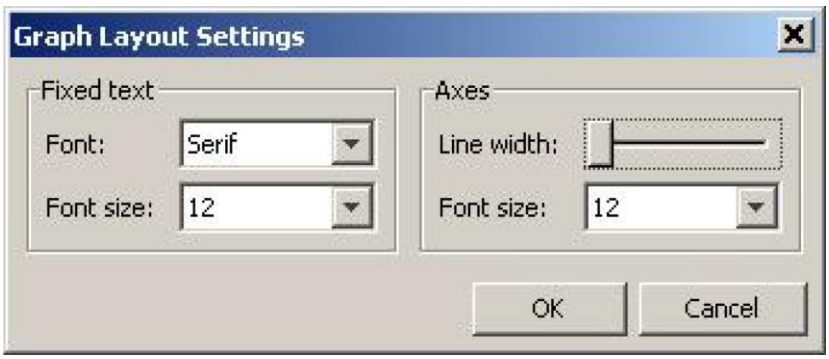


图50 总体图形布局设置可以校正固定文本组件的字体和字体大小，与x轴和y轴的宽度和坐标的字体大小。

在信息轴上，边界间的距离和Z-曲线间的距离通常用信息的相对增加来显示。TSA程序会按比例缩小的方式显示这些距离。

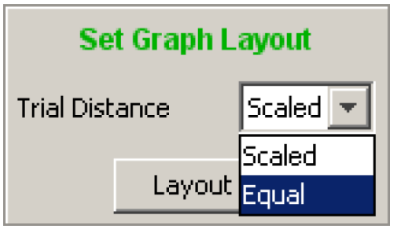


图51 在x轴上为选择试验间“相等”距离

然而在某些情况下，其他布局或许可以为视觉的解释提供更好的基础。在 Pogue 和 Yusuf 的论文中，TSA 程序也可以在论文中提供布局格式，这种格式是在信息轴上显示相同距离的试验和在 X 轴下 45 度角显示试验标题。选择这种布局格式，在“Set Graph Layout （设置图形布局）”区域的下拉框中单击“Trial Distance（试用距离）”，然后选择“Equal（相等的）”（图 51）。

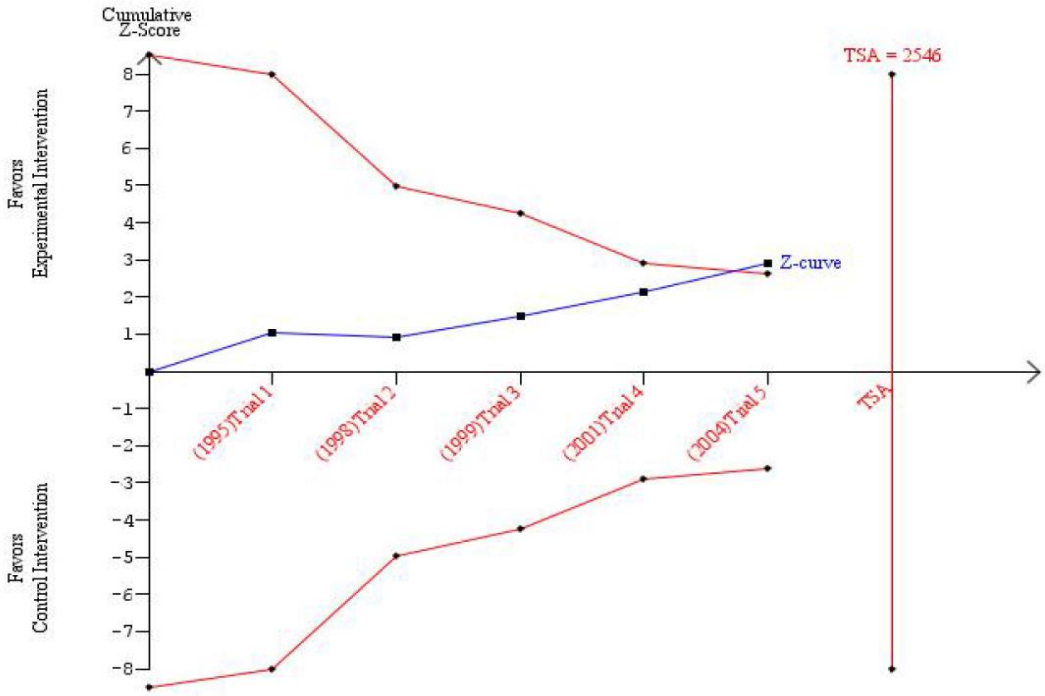


图52 选择信息-轴测量/显示格式.

基于α-消耗函数校正的显著性检验校正的是Z曲线的阈值。而基于迭代对数法校正的显著性检验结果则必须结合单个试验的显著性检验结果才可进行合理解释。因此，将这两种结果在一张图中显示是没有意义的。TSA程序为基于α-消耗函数校正的显著性检验，和基于迭代对数法则校正的显著性检验提供了不同的图表。查看α-消耗函数校正的显著性检验图示，选择图表上方的*Adjusted Boundaries*（*校正界值*）选项卡；查看迭代对数法则校正的显著性检验图示，则选择图表上方的*Penalised Tests*（*缺陷检验*）选项卡（如图53）。

图53 查看界值检验或缺陷检验表

4.6 探索试验间的差异性

TSA程序还提供一种通过3个随机效应模型（DL、SL和BT），探索差异性估计及权重比较的选项。这些选项具体可见*Diversity*（*差异性*）选项卡（如图54）。

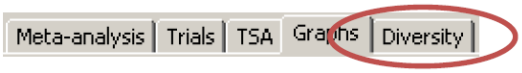
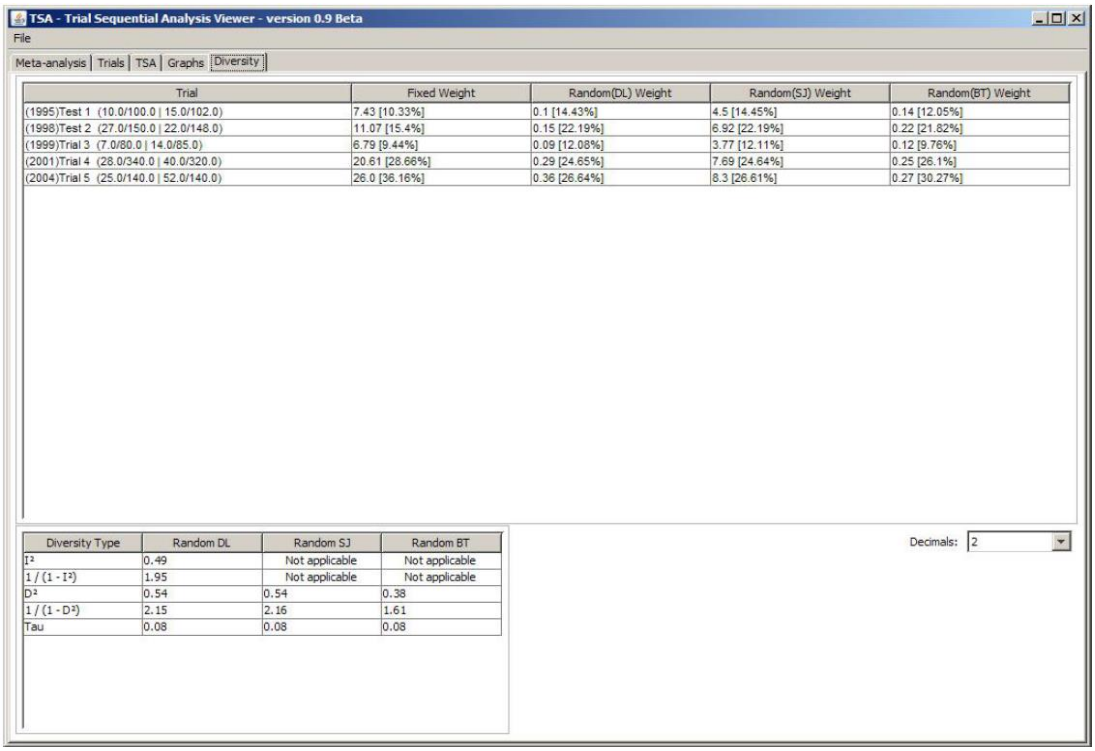


图54 单击*Diversity*（*差异性*）选项卡，通过随机效应模型探索差异性估计并比较权重。

单击*Diversity*选项卡后，屏幕将会如图55所示。屏幕上方显示试验及其显著性检验结果、权重、权重百分比；屏幕左下角显示的是3个随机效应模型各自的I^2异质性估计值及其相应校正异质性$1/(1-I^2)$、D^2异质性估计值及其相应校正异质性$1/(1-D^2)$，与试验间方差估计值（τ^2）。而异质性估计值仅显示于DL模型中。值得注意的是，DL与BT模型中试验间方差估计值是相同的（参见2.1.3节）。在右下角中，有一个选项可以选择全部数据显示的小数位数。单击下拉窗口，即可选择小数位数。



Trial	Fixed Weight	Random(DL) Weight	Random(SJ) Weight	Random(BT) Weight
(1995)Test 1 (10.0/100.0 | 15.0/102.0)	7.43 [10.33%]	0.1 [14.43%]	4.5 [14.45%]	0.14 [12.05%]
(1998)Test 2 (27.0/150.0 | 22.0/148.0)	11.07 [15.4%]	0.15 [22.19%]	6.92 [22.19%]	0.22 [21.82%]
(1999)Trial 3 (7.0/80.0 | 14.0/85.0)	6.79 [9.44%]	0.09 [12.08%]	3.77 [12.11%]	0.12 [9.76%]
(2001)Trial 4 (28.0/340.0 | 40.0/320.0)	20.61 [28.66%]	0.29 [24.65%]	7.69 [24.64%]	0.25 [26.1%]
(2004)Trial 5 (25.0/140.0 | 52.0/140.0)	26.0 [36.16%]	0.36 [26.64%]	8.3 [26.61%]	0.27 [30.27%]

Diversity Type	Random DL	Random SJ	Random BT
I^2	0.49	Not applicable	Not applicable
$1 / (1 - I^2)$	1.95	Not applicable	Not applicable
D^2	0.54	0.54	0.38
$1 / (1 - D^2)$	2.15	2.16	1.61
Tau	0.08	0.08	0.08

图55 差异性选项卡

5 TSA 应用实例

5.1 数据集

我们以几项已发表系统综述的数据为例，来说明TSA的应用方法。本章出现的一些TSA分析和软件应用是在原始出版物的基础上修改和添加而来。

5.2 避免假阳性

在本例中，我们使用的综述数据，是比较接受医院联络且至少1月随访的患者和未接受干预措施的患者的戒烟率[64]。该综述中，干预措施和随访时间在纳入的研究之间有很明显差异。因此，作者使用了以下对干预强度的分类方法[64]。

1. 仅在医院联络一次，持续时长≤15分钟，无随访支持。
2. 一次或以上医院联络，时长>15分钟，无随访支持。
3. 医院联络次数不限，随访≤1月。
4. 医院联络次数不限，随访>1月。

干预强度为第 3 个类别的 meta 分析纳入了 6 项试验，包括 4476 名患者和 628 个事件。固定效应模型计算得合并相对危险度为 1.05（95% CI 0.91~1.21）（比值比的 meta 分析结果与此相似）。异质性（inconsistency）的估计值(I^2)为 9%，变异性（estimateddiversity）的估计值(D^2)为 10%。每当一项新的试验发表，我们通过对累积数据的重复传统 meta 分析，进行回顾性的试验序贯分析。第一个发表试验的相对危险度为 1.47（95%CI 1.05~2.05）。加入第二个试验后，合并的相对危险度为 1.33（95%CI 1.02~1.75）。因此，在进行最初的两项试验后，接受第 3 类干预强度（见上）的干预组与对照组之间进行的 meta 分析结果显示，组间差异似乎具有了统计学意义。

我们对这些数据进行了TSA分析。计算证实或拒绝20%的相对获益增量（戒烟是获益的结局）所需的信息量（IS）。假设对照组事件比例为14%，这大致为对照组事件比例的中位数和平均数。设定I型错误为5％，II型错误为20％。不进行异质性校正。通过这些设置，我们计算出所需的信息量为5218名患者。由于meta分析纳入患者的数量未超过所需信息量，我们也应用无效性界值以探索能确证的“阴性”结论。

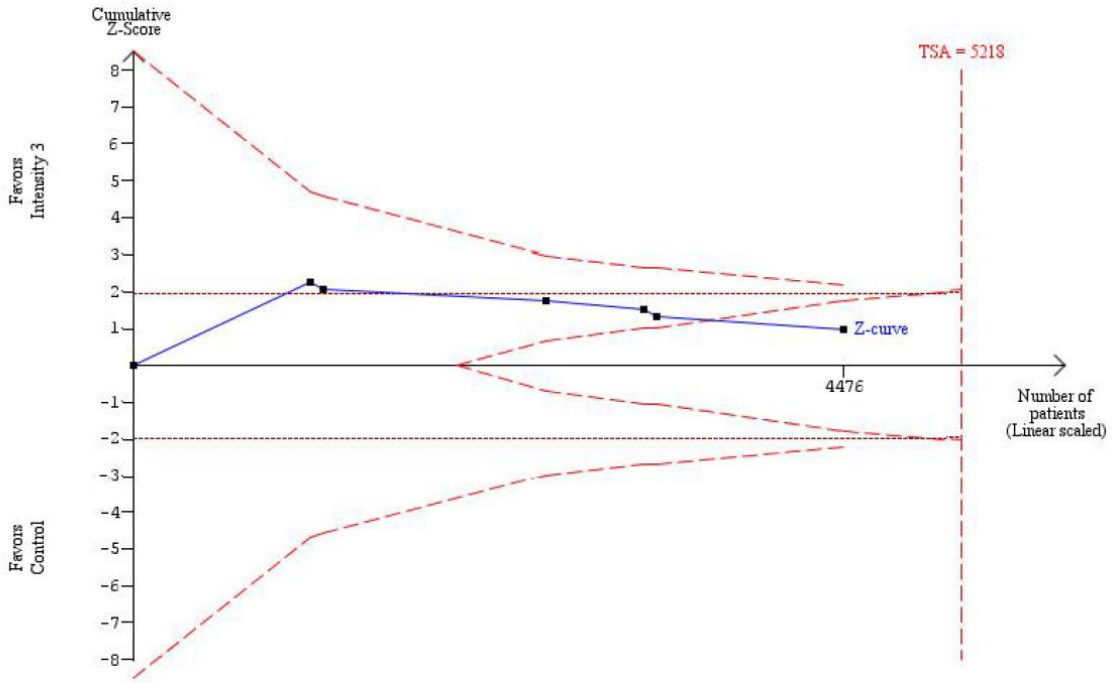


图56 当对照组戒烟比例为14%，α值为5%，β值为20%时，证实或拒绝戒烟率的20%相对增量获益所需的信息量为5218名患者（垂直红线）。红色虚线表示试验序贯监测界值及无效性界值。蓝色实线为累积Z曲线。（Cumulative Z-score：累计Z分值，Number of patients(Linear scaled)：患者数（线性），Z-curve：Z-曲线，Favour Intensity 3：倾向第三类强度，Favour Control：倾向对照组）

试验序贯分析的结果见图56。纳入第一和第二个试验后，使用传统分析技术时，累积Z统计量超过1.96，这相当于统计学显著性的名义阈值。从第三个试验开始，meta分析名义上不再具有统计学意义。纳入最后一个试验时，meta分析值穿过至无效性界值之下，表明第3类干预强度的戒烟率效果在80%的把握度下并未超过20%相对增量。也就是说，在当前设定的置信度和效应量下，这种干预措施无效。

5.3 阳性结果的确证

我们使用非体外循环和体外循环冠状动脉旁路移植术(CABG)[65]的系统综述数据，来演示TSA在确证阳性结果的应用方法。

在本例中，每当一项新的试验加入meta分析，即重复进行显著性检验，累积Z曲线的校正显著性界值是在此假设下构造而成。由于过去十年中，非体外循环与体外循环的争论被给予相当多的关注，这种假设似乎是合理的。我们使用了基于奥布莱恩-弗莱明型α-消耗函数的监测界值，该方法对多重显著性检验的次数相对不敏感（见2.2.3节）。

在所考虑的meta分析数据集中，有些年份不止一项试验发表。对于这些年份，我们根据第一作者姓氏按照字母排序，来分析这些试验。

5.3.1 “答案是”的确证

为了说明维护“答案是”的 TSA 应用，在体外循环对照非体外循环 meta-分析中我们使用了心房纤维性颤动作为结局。共计 3634 名患者（包括 2 个零事件试验）的 30 项试验报告了房颤发生情况[65]。根据显著性检验传统标准，在降

低房颤发生方面，非体外循环 CABG 显著优于体外循环 CABG(RR 值为 0.69，95% CI 为 0.57~0.83)（图 57）。异质性估计值 I^2 为 47.3%，差异性估计值 D^2 为 49.0%。

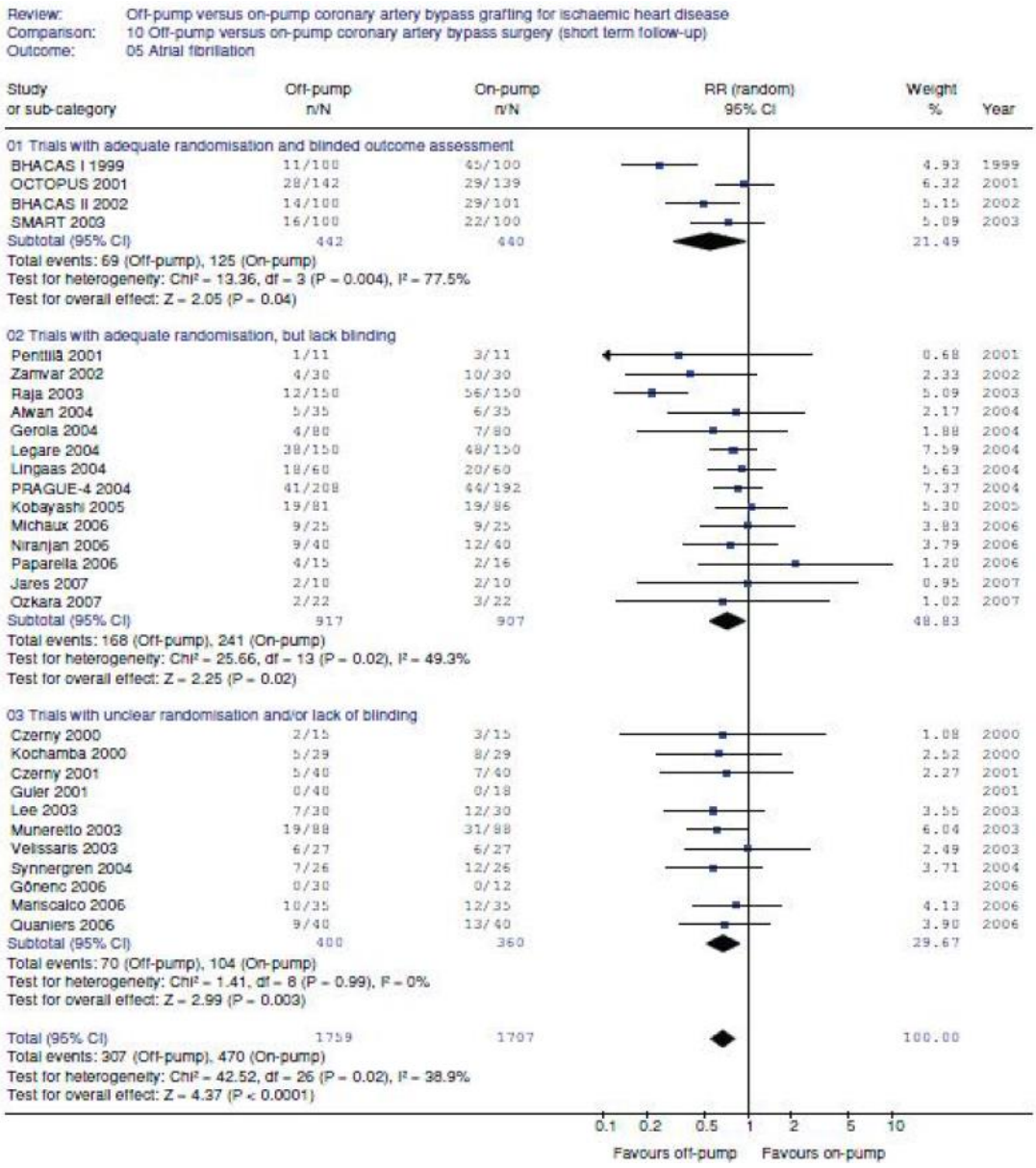


图57非体外循环对照体外循环CABG房颤疗效的森林图

在低偏倚风险试验（1050名患者）的meta分析中，疗效并不显著（RR值为0.63，95% CI为0.35~1.13），异质性估计值I^2为77%，差异性估计值D^2为79.0%。

对房颤的试验序贯分析

我们为本例计算两个信息量。第一，α值为1%，β值为10%时，我们计算出证实或拒绝先验预期干预疗效的相对危险度降低20%所需的信息量为7150名患者。之所以选择20%预期干预疗效值，是因其被认为代表了此临床情境下的合理干预疗效。第二，我们计算出来源于综述中纳入的低偏倚风险试验（36.9%）的相对危险度降低meta分析估计所需信息量为1964名患者。

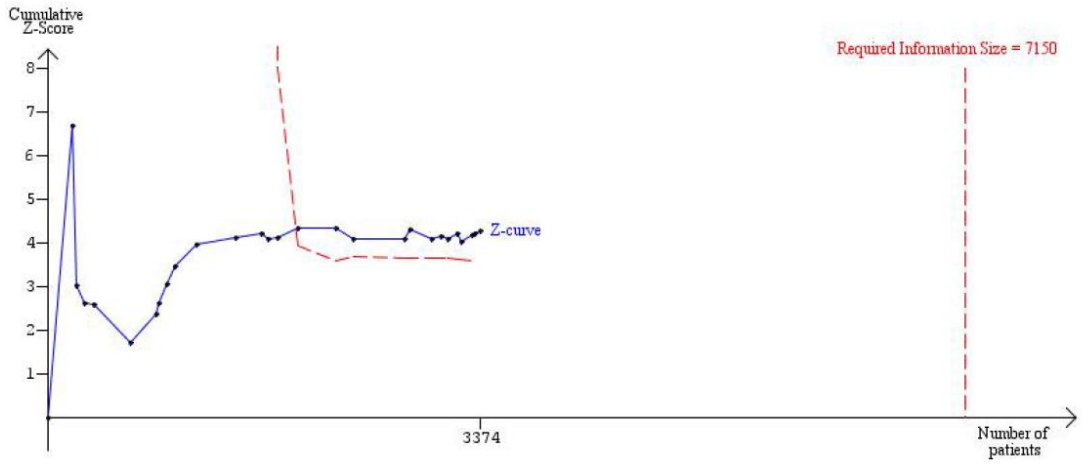


图58 证实或拒绝房颤（对照组比例为27.6%，α值为1%，β值为10%）相对危险度降低20%（先验估计值），校正异质性所需信息量为7150名患者。向内倾斜的红色虚线至左侧表示最初的14项试验的试验序贯监测界值被截断，蓝色实线是累积Z曲线。（Cumulative Z-Score：累计Z分值，Number of patients：患者数，Required Information Size：所需信息量）

所有信息量的计算均基于确保I型错误的上限为1%及II型错误的上限为10%（即90%把握度）。所有信息量均经过异质性校正，采用差异性估计值D^2。信息量均基于体外循环事件比例为27.6%（本对照组中事件比例的中位数）的假定。

累积 Z 曲线同由两个信息量估计值的监测界值相交（图 58 和图 59），从而确证了非体外循环 CABG 在降低房颤方面优于体外循环 CABG。

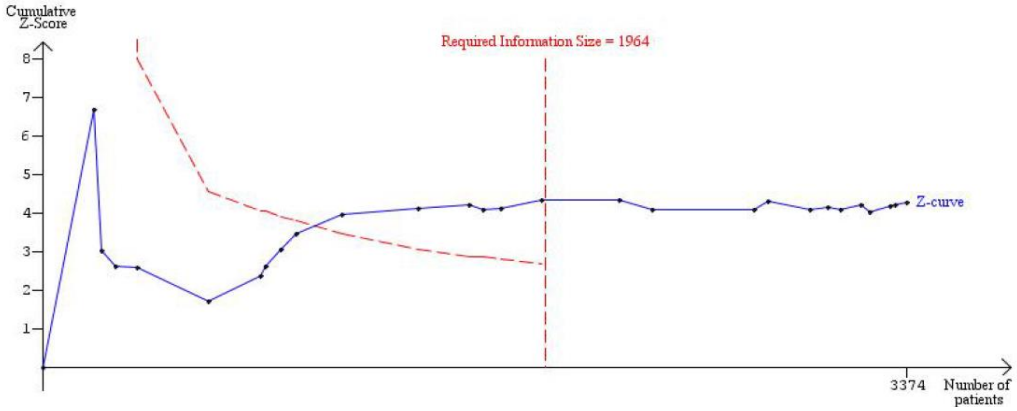


图59证实或拒绝房颤（对照组比例为27.6%，α值为1%，β值为10%）相对危险度降低36.9%（低偏倚风险估计）的校正异质性所需信息量为1964名患者（垂直红色虚线）。向内倾斜的红色虚线向左弥补被截断的最初4项试验的试验序贯监测界值。蓝色实线是累积Z曲线。（Cumulative Z-Score：累计Z分值，Number of patients：患者数，Required Information Size：所需信息量）

5.3.2 避免过早的高估

CABG治疗心房纤颤的同一例子中，也可用来说明常规meta分析过程中，如何过早出现疗效高估。根据常规标准(p<0.05)，纳入第一项试验后，房颤的meta

分析具有了统计学意义。除一项外，所有累积性meta分析随后的P值均小于0.05。事实上，大多数随后的P值均小于0.01。经验证据表明，当仅有有限数量的事件数与患者数被累积时，即使具有统计学意义，合并疗效估计值仍不稳定[4;5;9;29]。坚持确定一项meta分析是否超过其所需信息量，以可能确保可靠的合并估计值[1;2;4;6;19;23]。

表 3 显示了本例中每年年末治疗效果随时间的演变。最初两年中，在 P 值小于 0.01（常规 99%可信区间排除 1.00）的支持下，合并相对危险度被严重高估。随后三年中，合并相对危险度被绝对危险度高估了至少 10%。2003 年，meta 分析超过了基于低偏倚风险估计的信息量计算的监测界值。

2004年，meta分析中的样本超过了所需信息量。同样在2004年，基于20%的先验相对危险度降低（*priori relative risk reduction*），meta分析跨越了监测界值。传统和校正后的置信区间在2002年和2004年之间收敛。

	总数量			综合	99%置信区间	
年份	试验	事件	病例	效应	原始	校正
1999	1	55	200	0.24	0.14至0.42	0.03至7.74
2000	3	74	288	0.39	0.15至0.99	0.02至7.18
2001	5	143	649	0.57	0.24至1.34	0.12至2.87
2002	8	204	932	0.52	0.30至0.90	0.22至1.21
2003[a]	10	285	1168	0.55	0.37至0.81	0.35至0.85
2003[b]	13	391	1722	0.53	0.35至0.79	0.34至0.83
2004	19	641	2832	0.61	0.46至0.82	-
2005	20	679	2999	0.63	0.49至0.85	-
2006	25	768	3310	0.67	0.53至0.86	-
2007	27	775	3372	0.67	0.53至0.86	-

[a]第一次跨界, [b] 年终

表3显示了综合效应的变化（相对危险度估计），按照试验、事件和患者的累积数量，于每年年末列出了合并效应的传统和校正后的99%置信区间。校正后的99%置信区间是基于与所需信息量有关（1964名患者）的α-消耗函数，该α-消耗函数是基于低偏倚风险试验提示的相对危险度估计确定的。

这个例子阐释了为什么基于相对小样本事件和病例数（本例中指事件数低于100且病例数低于300）的合并估计值不可信。除非累计达到至少100例事件数和1000例病例数，否则此meta分析中的点估计值不具备充分的可信度。校正后的置信区间用于避免meta分析在早期阶段得出虚假推论，同时在合理范围之内接近原始的置信区间，使累积病例数量接近所需的meta分析信息量。

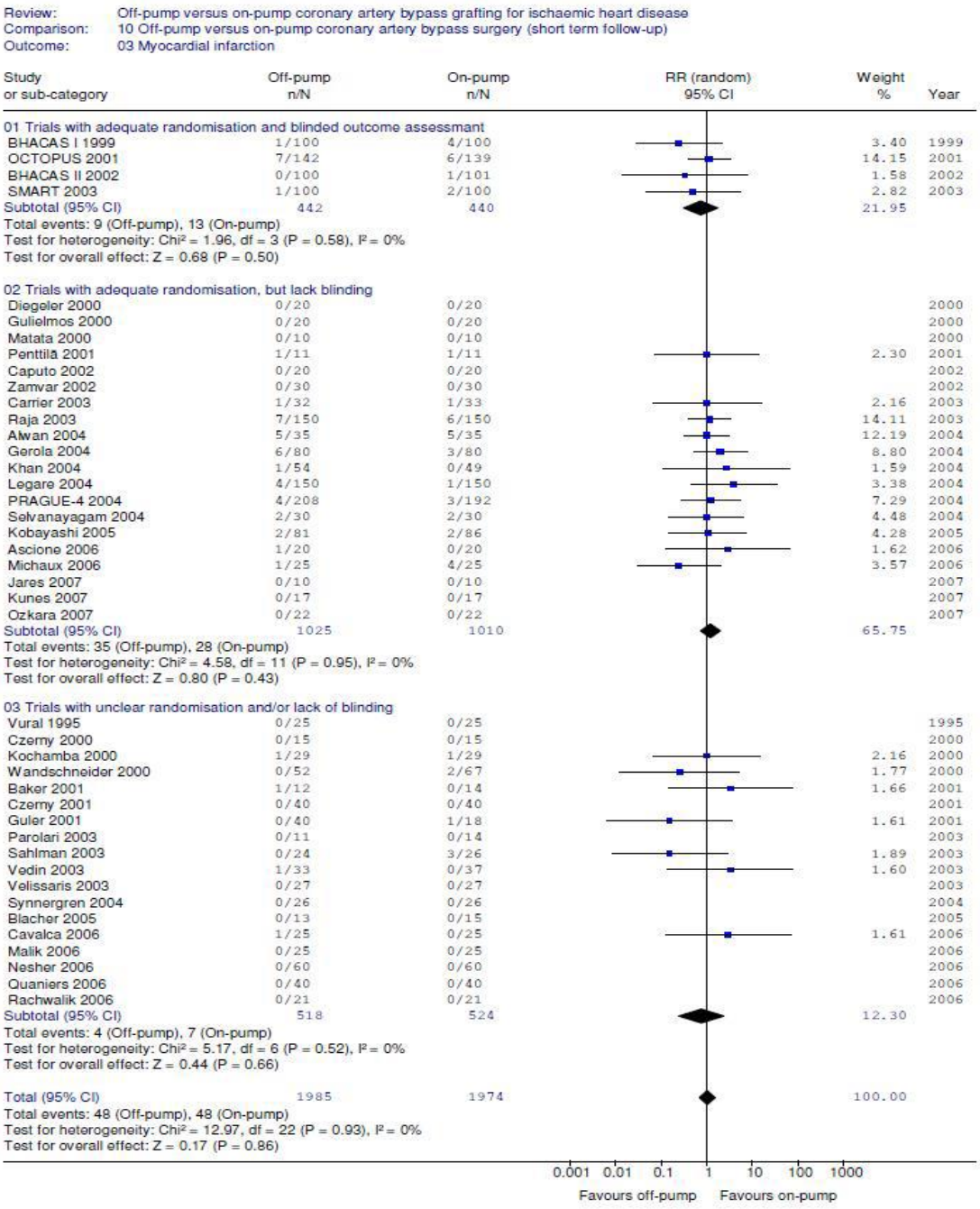


图60 非体外循环与体外循环冠状动脉旁路移植术对心肌梗塞作用的森林图

5.4 无效性检验

非体外循环与体外循环冠状动脉旁路移植术的meta分析的例子同样可以用于演示无效性检验，这次选用的结局是心肌梗死。44个试验报告了心肌梗死的发生，共纳入病人4303例，结果显示体外循环和非体外循环手术无显著性差异（RR 1.06; 95% CI 0.72-1.54）(图 60)，该结果无偏倚风险。异质性（I^2 =0%）无统计学意义。19个试验（909名病人）为零事件试验。对零事件试验进行连续性校正后，结果仍无显著的变化（RR 1.05; 95% CI 0.74-1.48）。

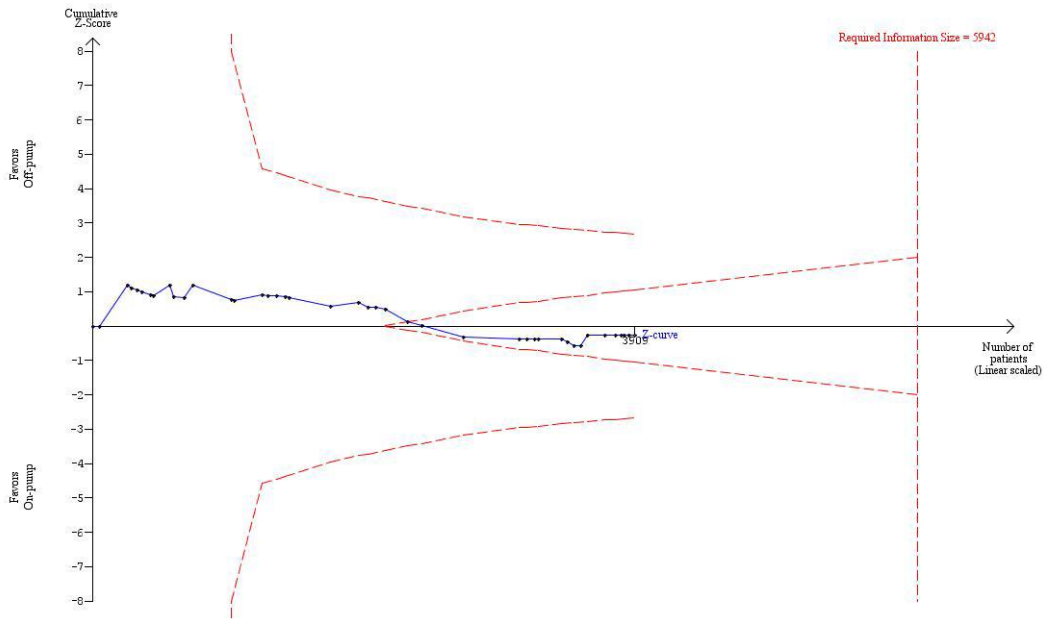


图61 异质性校正后所需的信息量为5942个病人（垂直的红线），以证明或拒绝33%心肌梗死（体外循环组的心肌梗塞发生率为3.9%，alpha值为5%，beta值为20%）的相对危险度降低（先验估计）。左侧向内倾斜的红色虚线构成了试验的连续监测界值。右侧向外倾斜的红色虚线构成了无效区域。蓝色实线是累积Z曲线。（Cumulative Z-score：累计Z分值，Number of patients(Linear scaled)：患者数（线性），Z-curve：Z-曲线，Favour Off-pump：倾向非体外循环，Favour On-pump：倾向体外循环，Required Information Size：所需信息量）

我们估算了证实或拒绝一个预期的干预作用达到33%相对风险降低的先验检验所需的信息量。在这种临床情境下，选择33%的值被认为可以代表一个合理的干预作用。相对于房颤的信息量计算，我们设定了最高5%的I类错误和最高20%的II类错误（80%把握度）。我们使用了体外循环组的心肌梗死事件发生比率的中位数（排除零事件发生的试验）作为对照组的时间发生比例（3.9%）。基于这些假设计算得出所需信息量为5942。累积Z曲线跨越了无效界限（图61），我们因此可以做出推论，非体外循环与体外循环冠状动脉旁路移植术的增效作用均不超过33%。当然，这个结果存在20%的风险，可能是一个假阴性的结果（II类错误为20%）。

5.5 估算一个新的临床试验的样本量

当meta分析结果既未跨越监测界值也未跨越无效限时，可以粗略估计出在下一个试验中需要将多少例病人进行随机，才能使得meta分析结果跨越两条界限中的任意一条。最近的一篇方法学论文采用一项异烟肼预防HIV阳性患者感染结核病的meta分析阐明了这种方法[25]。此meta分析共纳入9项试验，包括病人2911例，共发生131例事件，合并后的相对危险度为0.74（95% CI 0.53，1.04）。异质性和差异性的估计值均为0%。

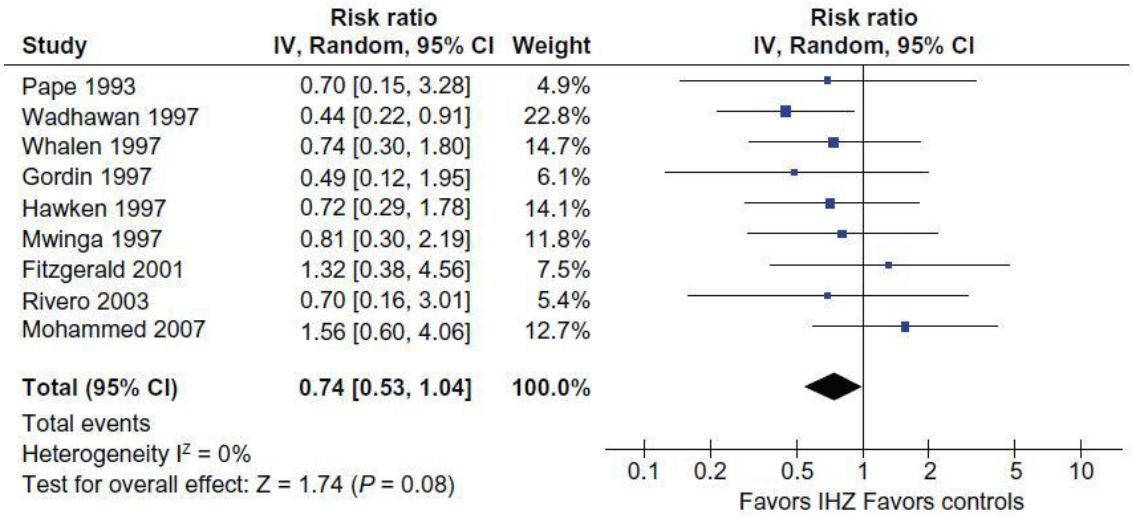


图62 比较异烟肼与对照组在纯蛋白衍生物阴性艾滋病病毒感染者中预防肺结核森林图。

在α为5%和β为20%（80%把握度）的条件下，我们估算了观测25%的肺结核的相对危险度降低所需的信息量。所需的信息量是基于假设对照组有5%的事件发生率(近似于各项试验的比率的中位数)。我们亦假定存在20%差异性(D^2)，并对得出的信息量做了异质性校正，得出所需信息量为10508例。根据设置的显著性水平和所需信息量，随后确定了统计学监测界值和无效限。

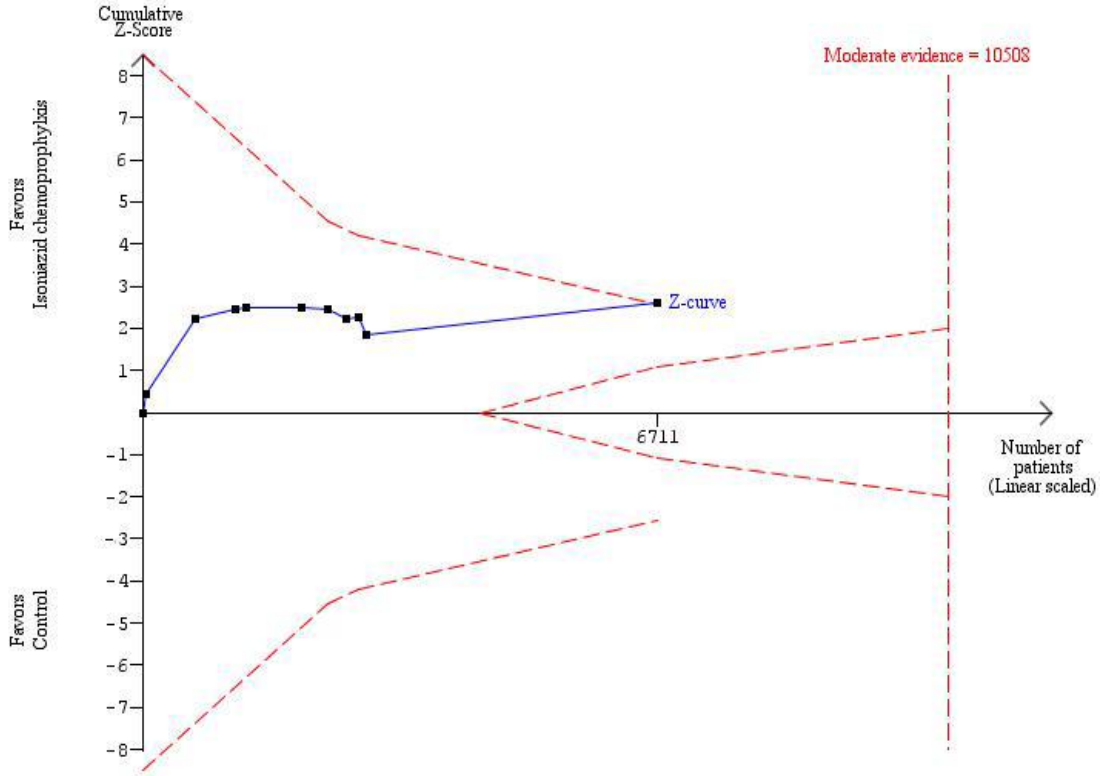


图63 前瞻性试验序贯分析了异烟肼对比对照组对肺结核的预防作用。左侧向内倾斜的红色虚线为试验的连续监测界值。右侧向外倾斜的红色虚线构成了无效区域。蓝色实线是累积Z曲线。Z曲线上的最后一段线代表了一个假想的试验，使meta分析得出异烟肼对肺结核有预防作用的结论。（Cumulative Z-score：累计Z分值，Number of patients(Linear scaled)：患者数（线性），Z-curve：Z-曲线，Favour Isoniazid

chemoprophylaxis：倾向异烟肼，Favour Control：倾向对照，Moderate evidence：中等证据）

为了估计今后的临床试验中需要随机多少例病人才能得出meta分析的结论，我们在一个虚构的前瞻试验中估算了病人的数量，使得累积Z曲线跨越监测界值或无效限。如果未来的试验要使得meta分析的结果是阳性的结论，我们需要假设试验与考虑信息量时相同的对照组事件发生比例和干预效应。也就是说，我们假设一个试验的对照组有5%的事件发生比例并有25%的相对危险度降低（即试验的干预组的事件发生比例为3.75%）。

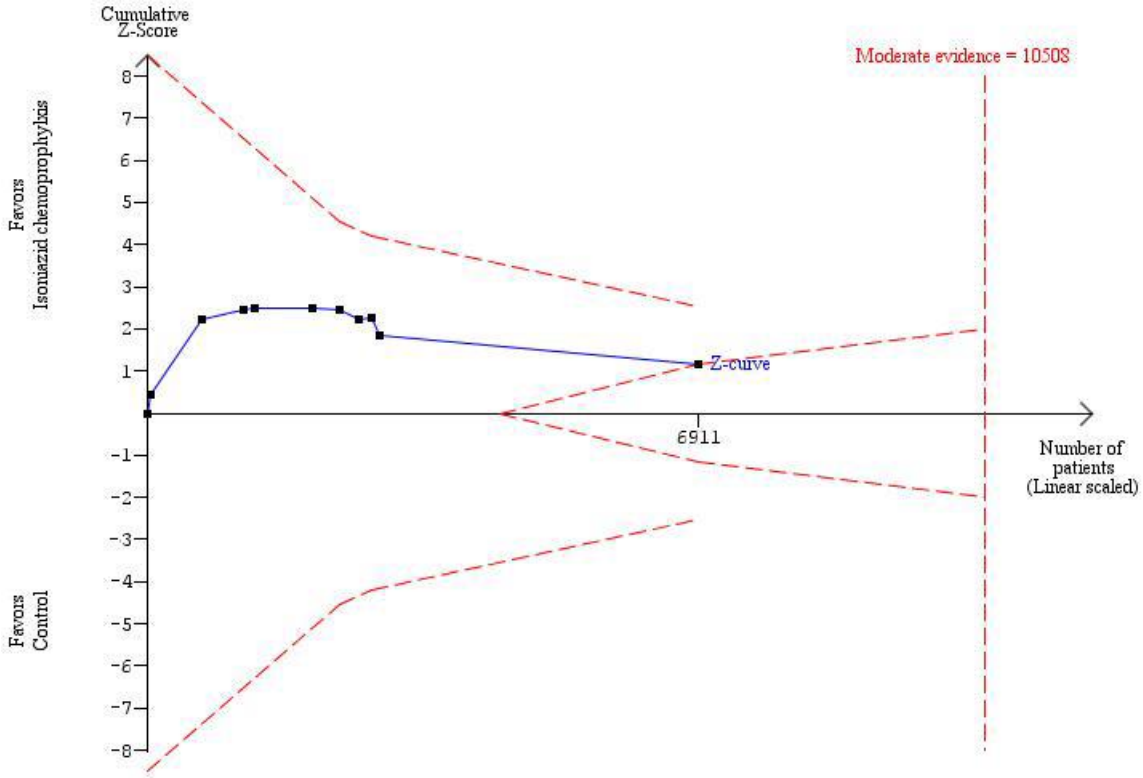


图64 异烟肼和对照组比较，预防肺结核的前瞻性试验序贯分析。左侧向内倾斜的红色虚线构成了试验的连续监测界值。右侧向外倾斜的红色虚线构成了无效区域。蓝色实线是累积Z曲线。Z曲线上的最后一段线代表了一个假想的试验，使meta分析得出异烟肼对肺结核没有预防作用的结论。（Cumulative Z-score：累计Z分值，Number of patients(Linear scaled)：患者数（线性），Z-curve：Z-曲线，Favour Isoniazid chemoprophylaxis：倾向异烟肼，Favour Control：倾向对照，Moderate evidence：中等证据）

如果未来的临床试验要使得meta分析的结果为阴性的结论，我们假设干预组事件比例也将是5%（即，无效）。要得出确证性的阳性结论的meta分析约需要3800名病人（每个干预组1900名病人）（图63），要得出确证性的阴性结论的meta分析约需要4000名病人（每个干预组2000名病人）（图64）。

5.6 其它已发表的试验序贯分析的应用实例

本手册的作者已经撰写了数个系统综述，其中，每个系统综述至少有一个meta分析应用了序贯分析方法[14;24;63;65-74]。表4是对这些已发表文献的简要汇总（按发表年份排序）

表4 已发表的应用试验序贯分析的meta分析的概览

第一作者	期刊 (年份)	Meta分析
Bangalore[75]	*BMJ* *(2011)*	比较血管紧张素受体阻滞剂（ARB）和对照组在如下方面 i) 非致命的心肌梗死 ii) 全因死亡率 iii) 心血管死亡率 iv) 心绞痛 v) 中风 vi) 心衰 vii) 新发糖尿病
Bangalore[76]	*Archives of Neurology* *(2011)*	颈动脉支架置入术（CAS）对比颈动脉内膜切除手术在如下方面 i) 死亡，心肌梗死或中风 ii) 围手术期死亡或中风 iii) 围手术期中风
Bangalore[77]	*Lancet Oncology* *(2011)*	i) 比较血管紧张素受体阻滞剂和对照组：在癌症风险和癌症相关死亡方面的作用 ii) 比较血管紧张素转换酶抑制剂和对照组：在癌症风险和癌症相关死亡方面的作用 iii) 比较β受体阻断剂和对照组：在癌症风险和癌症相关死亡方面的作用 iv) 比较钙通道阻滞剂和对照组：在癌症风险和癌症相关死亡方面的作用 v) 比较利尿剂和对照组：在癌症风险和癌症相关死亡方面的作用
Afshari A[24]	*The Cochrane Library (2010)*	i) 比较一氧化氮吸入和对照组治疗急性呼吸窘迫综合征 ii) 比较一氧化氮吸入和对照组治疗肺损伤
Awad T[63]	*Hepatology* *(2010)*	比较聚乙二醇化干扰素 α-2a 和聚乙二醇化干扰素 α--2b 治疗丙型肝炎
Brok J[66]	*J Alim Pharm &Ther* *(2010)*	比较病毒唑加干扰素和单独干扰素治疗丙型肝炎
Nielsen N[70]	*Int J Cardiol* *(2010)*	比较降低体温和对照组在心搏停止后的作用
Tarnow-MordiWO[72]	*Pediatrics* *(2010)*	i) 比较益生菌和对照组在降低新生儿死亡率中的作用 ii)比较益生菌和对照组在降低新生儿坏死性小肠结肠炎中的作用
Knorr U[69]	*Psychoneuroendocrinology (2010)*	比较唾液皮质醇在抑郁病人和对照组人群中的作用
Bangalore S[14]	*The Lancet* *(2009)*	i)比较围手术期β受体阻滞剂和安慰剂对死亡率的影响 ii)比较围手术期β受体阻滞剂和安慰剂对心肌梗死的影

		响
Brok J[67]	*The CochraneLibrary* (2009)	比较病毒唑单一治疗和安慰剂对丙型肝炎的影响
Whitfield K[73]	*The Cochrane Library* (2009)	比较己酮可可碱和安慰剂对酒精性肝炎的影响
Moller CH[65]	*Europ Hearj J*	i)比较非体外循环和体外循环冠状动脉旁路移植术对心房颤动的影响 ii)比较非体外循环和体外循环冠状动脉旁路移植术对心肌梗死的影响
Ghandy GY[68]	*Mayo Clin Proc (2008)*	i)比较围手术期胰岛素注射和对照组对死亡率的影响 ii)比较围手术期胰岛素注射和对照组对发病率的影响
Rambaldi A[71]	*J Alim Pharm &Ther (2008)*	比较糖皮质激素和对照组对酒精性肝炎的影响
Whitlock R[74]	*Europ Heart J (2008)*	比较预防性使用类固醇和对照组对接受心肺分流术病人的影响
Afshari A[24]	*BMJ (2007)*	比较抗凝血酶III和对照组对降低心脏的・・・・・・

6 附录

6.1 二分类变量和连续性变量 Meta 分析中的效应量估计

各个效应量估计的标准误计算方法与在 ReviewManager v.5.[27] 软件里的方法相似。

对于每一项试验，我们用 e_A 和 e_B 表示两个干预组内被观测到的事件（比如死亡）数，用 n_A 和 n_B 表示两个干预组的总人数。

危险度差（risk differences，RD）、相对危险度（relative risks，RR）和比值比（odds ratios，OR）的标准误计算公式如下：

$$se(RD) = \sqrt{\frac{e_A(1-e_A)}{n_A^3} + \frac{e_B(1-e_B)}{n_B^3}}$$

$$se(RR) = \sqrt{\frac{1}{e_A} + \frac{1}{e_B} - \frac{1}{n_A} - \frac{1}{n_B}}$$

$$se(OR) = \sqrt{\frac{1}{e_A} + \frac{1}{e_B} + \frac{1}{(1-e_A)} + \frac{1}{(1-e_B)}}$$

皮托比值比（Peto’s oddsratio）的标准误计算公式为：

$$se(OR) = \sqrt{1/v}$$

其中，

$$v = \frac{\left(n_A n_B (e_A + e_B)\left((1 - e_A) + (1 - e_B)\right)\right)}{(n_A + n_B)^2 (n_A + n_B - 1)}$$

6.2 随机效应的处理

6.2.1 Biggerstaff-Tweedie 法的公式

用 $f_{DL}(t)$表示 τ^2 的 DL 估计的概率密度函数，用 $F_{DL}(t)$表示对应累积分布函数。定义试验权重为一个关于 t 的函数 $w_i(t)=(\sigma_i^2+t)^{-1}$，使用对 τ_{DL}^2 估计值的求得的分布，所谓的试验权重的频率-Bayes 的估计量可从以下公式获得：

$$\begin{aligned} w_i^*(\tau^2) &= E[w_i^*(\tau_{DL}^2)] \\ &= F\left(t \cdot 1_{(0.\infty)}(t)\right) w_i(0) + \int_0^\infty w_i^*(t) f_{DL}(t) dt \end{aligned}$$

由此得出总人口干预效应的综合估计量：

$$\mu_{BT} = \frac{\sum_i w_i^* Y_i}{\sum_i w_i^*}$$

和方差：

$$V_{ar}(\mu_{BT}) = \frac{1}{(\sum_i w_i^*(\tau_{DL}^2))^2} \left[\sum_i \left(w_i^*(\tau_{DL}^2)\right)(s_i^2 + \tau_{DL}^2)\right]$$

从而确保总效应估计量的方差是基于试验间方差估计相关的不确定性进行校正的。

6.3 试验序贯分析

6.3.1 多重显著性检验夸大 I 型错误

根据基础概率理论，当数据一段时间内被检验两次，且 5%的α值（α=0.05）被用作两次检验的阈值（或 Z 值为 1.96）时，在零假设（null hypothesis）下两种干预措施有统计学意义的概率是：

$$\Pr(H_{0\ \text{rejected}}) = \Pr(|Z_1| \geq 1.96 \text{ or } |Z_2| \geq 1.96)$$
$$= \Pr(|Z_1| \geq 1.96) \cdot \Pr(|Z_2| \geq 1.96 | |Z_1| < 1.96)$$
$$= \left(1 - \Pr(|Z_1| < 1.96)\right) \cdot \left(1 - \Pr(|Z_2| < 1.96 | |Z_1| < 1.96)\right)$$
$$= 1 - \Pr(|Z_1| < 1.96) - \Pr(|Z_2| < 1.96 | |Z_1| < 1.96) + \Pr(|Z_1| < 1.96) \cdot \Pr(|Z_2| < 1.96 | |Z_1| < 1.96)$$
$$= 1 - \Pr(|Z_1| < 1.96) - \Pr(|Z_2| < 1.96 | |Z_1| < 1.96) + \Pr(|Z_1| < 1.96\ or\ |Z_2| < 1.96)$$
$$= 0.05 - \Pr(|Z_2| < 1.96 | |Z_1| < 1.96) + \Pr(|Z_1| < 1.96\ or\ |Z_2| < 1.96)$$
$$> 0.05$$

其中，不等式由以下直接推知

$$\Pr(|Z_2| < 1.96\ or\ |Z_1| < 1.96) > Pr(|Z_2| < 1.96 | |Z_1| < 1.96)$$

以上对于任何α值和重复任何次数的显著性检验均适用。

6.3.2 未应用 TSA 软件的其它备选方法

很多方法适用于随机临床试验的多重显著性检验——其中一些也可应用于 meta-分析。[30]TSA 软件中的统计方法属于围绕标准化 Z-统计量（或至少一项校正）的监测进行构建的方法范畴。其它 meta-分析中较引人关注的序贯方法则依据监测其它统计量而构建。

有效分数的序贯分析（监测）或称 meta-析效应的似然分数统计量，最近受到关注。[78-81] 在标准 meta 分析中，设置每一项试验的有效分数，只是通过计算估计的试验治疗效应乘以其方差的值，而 meta 分析有效分数则是试验有效分数的总和。在有效分数的序贯分析中，一般信息作为统计信息（比如 Fischer 信息）进行测量。有效分数以 y 轴表示，统计学信息以 x 轴表示，并通过一些界值进行监测。正如基于界值的 α-消耗和 β-消耗一样，监测有效分数的序贯方法得出优效、劣效和无效的界值。图 65 显示的是这些界值的例子。

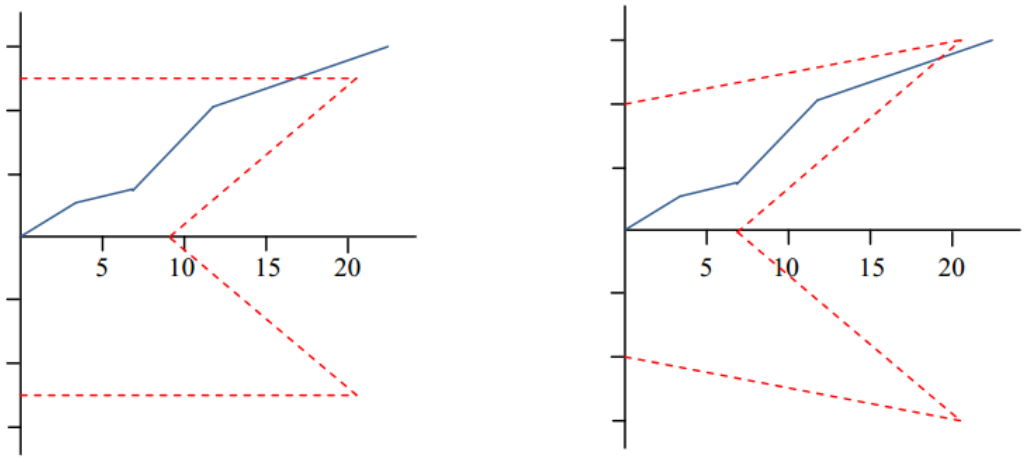


图 65 由有效分数的序贯 meta-分析产生的两种类型监测界值的图示。左图所示为 O'Brien-Flemingα-消耗显著性界值和 O'Brien-Fleming β-消耗无效性界值的情形；右图所示通常所知的 Whitehead 三角界值的情形。后者用于最小化统计误差的总体风险。（即，同时考虑 I 型和 II 型错误）

就像不同的 α-消耗函数产生不同类型的校正显著性界值（同理 β-消耗函数产生不同类型的无效性界值），三角检验（triangular test）可以用于构建不同类型的界值 [45;80]。例如，在累计统计信息的特殊情况下，三角检验产生的界值等于 O'Brien-Fleming 界值（图 65 中的左图）。

最近有研究通过模拟对 O'Brien-Fleming 型有效分数序贯界值进行了经验性探索 [80]。van der Tweel 和 Bollen 的一项研究在 6 个 meta 分析中比较了 O'Brien-Fleming 显著性界值（应用于 TSA 软件）和 O'Brien-Fleming 型有效分数序贯界值 [80]。作为演示例子，这 6 个 meta 分析一开始（且随机地）就被选入本书方法部分，这部分提供了试验序贯分析的信息量异质性校正（information sizeheterogeneity correction），在本手册 2.2.1.部分有描述 [6]。Tweel 和 Bollen 发现两种方法在显著性检验中是完全相同的。Higgins 等人的一项模拟研究研究了 I 型错误和校正可信区间范围，此范围与一些随机效应模型方法下的 O'Brien-Fleming 型有效分数序贯界值有相关 [79]。他们发现在这种设计下，通常的 DerSimonian-Laird 随机效应模型（DerSimonian-Lairdrandom-effects model）和 Biggerstaff-Tweedie 方法（Biggerstaff-Tweedie approach）不能产出令人满意的结果，但是半 Bayesian 方法（semi-Bayesian approach）却可以，这种方法使用试验间方差的 γ 分布，能够提供有效的信息。另一个有效分数序贯界值的例子是 Whitehead 提出的三角检验 [45;81]。图 65 中的右图显示这种方法产生的界值。在统计学意义上构建三角检验界值是为了使出现误差的可能风险降到最低？（无论是 I 型错误还是 II 型错误）[30;45]。同时强调最小化的两种误差，使这种方法在 I 型错误的风险之上，更重视误差的总风险。在医药研究领域中，通常理论不支持这种同时最小化两种错误的平衡，预防 α 错误总是被认为更加重要。

模拟研究探索了 Whitehead 三角检验应用于 meta 分析的表现，在模拟研究中，这种方法在异质性 meta 分析中对最大 I 型错误的控制表现很差 [81]。此研究结果显示，meta 分析数据集的异质性（heterogeneous）越大，三角检验表现出对 I 型错误的控制越差 [81]。迄今为止，文献中包含 1 例 Whitehead 三角检验应用于 meta 分析的例子，比较早产儿高频通气后和常规机械通气后死亡或慢性肺部疾病发生的情况 [78]。在这个例子中，累计分数统计值超过了无效性界值 [78]， meta-分析显示两种干预措施间没有差异。

随机缩减（Stochastic curtailment）是用于控制假阳性和假阴性风险的另一种方法 [1;2]。当应用于 meta 分析时，这种方法集中在预测一旦 meta 分析超过其所要求的信息量时的结果 [1;2]。更确切地说，随机缩减用于计算在超过所需信息量之前，数据实时趋势发生逆转的可能性。当这种逆转的概率足够小，才可能认为该 meta 分析有说服力。两个条件概率可以计算出来。首先，如果数据的当前趋势提示试验干预措施有效，当 meta 分析超过所需信息量时，随机缩减可能被用于计算拒绝无效假说的概率。如果这个条件概率足够高，可以认为这个 meta 分析有说服力。同样，如果当前数据没有提示有效的趋势，随机缩减可以用来计算 meta 分析超过所需信息量时不拒绝无效假说的概率。如果该条件概率足够高，可以认为这个 meta 分析有说服力。随机缩减作为一个有价值的辅助工具，可以用于辅助从正式显著性检验方法中做出的决策。然而因为大多数 meta 分析易受某些时间-趋势偏倚的支配，逆转趋势的条件概率也容易受到偏倚的影响。

7 缩略语和统计符号列表

以下的章节中包括本手册使用的缩略语、符号和术语。某些情况下，这些定义会不同于其它资料的定义。我们的目的是向读者提供在本手册如何使用这些术语的指导。

7.1 常用缩略语

AF: Adjustment Factor 校正因子
BT：Biggersaff-Tweedie
CI: ConfidenceInterval 可信区间
D^2: Diversity 差异性
DL: DerSimonian-Laird
I^2- Inconsistency 非一致性
IF: Information Fraction 信息分数
IS: Information Size 信息量
JRE: Java Runtime Environment Java 运行环境
MD: Mean Difference 平均差
OIS: Optimal InformationSize 最优信息量
OR: Odds Ratio 比值比
RCT: Randomised Controlled Trial 随机对照试验
RD: Risk Difference 危险度差
RR: Relative Risk 相对危险度
RRR: Relative Risk Reduction 相对危险度降低
SJ: Sidik-Jonkman
SMD: StandardisedMean Difference 标准化均差
TSA: Trial Sequential Analysis 试验序贯分析

7.2 统计符号

7.2.1 小写字母符号

c – The statistical significancethresholdwith respect |Z|
|Z|的统计显著性阈值
c_i– The adjustedthresholdfor Z_i under repeatedtesting
重复检验下 Z_i 的校正阈值
e_X– The number of events in intervention groupX
干预组 X 的事件数

$f_{DL}(t)$– The probability distributionfor the DerSimonian-Laird estimator
DerSimonian-Laird 估计量的概率分布
k– The number of trials in ameta-analysis
meta-分析中的试验数
m_X–Themeanresponse in intervention groupX
干预组 X 的平均响应
n_X–The number of patients in intervention groupX
干预组 X 的患者数
sd_X–The standard deviation in intervention groupX
干预组 X 的标准差
v–Variance estimate
方差估计量
v_F–The variance in afixed-effectmodel
固定效应模型的方差
v_R–The variance in arandom-effectsmodel
随机效应模型的方差
w_i–The weight assignedto the i-thtrial in afixed-effect model
固定效应模型中分配到第 i 个试验的权重
w_i^*–The weight assignedto the i-thtrial in arandom-effectsmodel
随机效应模型中分配到第 i 个试验的权重
$w_i(t)$– The i-thtrial weight as afunction of the between-trial variance
第 i 个试验权重为试验间方差函数

7.2.2 大写字母符号

AF–The heterogeneity adjustmentfactor
异质性校正因子
C–The sum of the continuity correctionsfor two groups
两组连续性校正值之和
CF_X–The continuity correctionfor intervention groupX
干预组 X 的连续性校正值
D^2–The diversitymeasure usedto quantify heterogeneity
量化异质性的差异性测量值
E(X)– The expectation of X
X 的期望
H–A conceptualmeasure of D^2
D^2 的概念度量
H0–The null hypothesis
零假设
I^2–The inconsistencymeasure usedto quantify heterogeneity
量化异质性的不一致性测量值
I_j –The cumulative statistical information after the j-th
第 j 个后的累计统计信息
IF_i–The cumulative informationfraction after the i-th trial
第 i 个试验后的累计信息分数

$IS_{Patients}$–Therequired number of patients in ameta-analysis
meta 分析中所需的患者数

IS_{Events}–Therequired number of events in ameta-analysis
meta 分析中所需的事件数

$IS_{Statistical}$–Therequired statistical information in ameta-analysis
meta 分析中所需的统计信息

IS_{Fixed}–Therequired information sizefor afixed-effect model
固定效应模型所需的信息量

IS_{Random}–Therequired information sizefor arandom-effects model
随机效应模型所需的信息量

OR_i–The oddsratio estimate of the i-thtrial
第 i 个试验的比值比估计量

P–ThetestP-value derivedfrom Z
由 Z 值得出的检验 P 值

P_X–The event rate in intervention groupX
干预组 X 的事件发生率

P^*–The average event rates of thetwotreatment groups
2 个治疗组的平均事件发生率

Pr(X)–The probabilitythat some event X occurs
某事件 X 发生的概率

Pr(X|Y)–The probabilitythat some event X giventhe event Y occurred
事件 Y 的情况下，发生事件 X 的条件概率

Q–The Cochran homogeneitytest statistic
Cochran 同质性检验统计量
R–Therandomizationratio
随机比例
RD_i–Therisk difference estimate ofthe i-th trial
第 i 个试验的危险度差估计量
RR_i–Therelativerisk estimate of the i-th trial
第 i 个试验的相对危险度估计量
S_r–The sum of trial weightsto the r-th power
试验权重 r 次方的总和
SE(X)–The standard error of X
X 的标准误
Var(X)–The variance of X
X 的方差
Z–Thetest statisticfor whether there exists an intervention effect
用于检验干预效应的统计量
Z_i–The Z-valuefrom themeta-analysis includingthefirst i trials
最初 i 个试验的 meta-分析得出的 Z 值
$Z_{1-\alpha/2}$–The (1-α/2)-th percentile of the standard normal distribution
标准正态分布的第 1-α/2 个百分位数
$Z_{1-\beta}$–The (1-β)-th percentile of the standard normal distribution
标准正态分布的第 1-β 个百分位数
Y_i–The observed intervention effect inthe i-thtrial
第 i 个试验中观察到的干预效应

7.2.3 希腊字母符号

α–Themaximum risk of type I error
I 型错误的最大风险
α(t)–The cumulativetypeI error risk as afunction of time
累计 I 型错误风险的时间函数
β–Themaximum risk of type II error
II 型错误的最大风险
β(t)–The cumulativetypeII errorrisk as afunction of time
累计 II 型错误风险的时间函数
δ–The a priori estimate of an anticipated intervention effect
期望干预效应的先验估计
δ_F–The anticipated intervention effect in afixed-effect model
固定效应模型中的期望干预效应
δ_R –The anticipated intervention effect in arandom-effectsmodel
随机效应模型中的期望干预效应
λ–A constant to ensure control of α when penalising Z
当惩罚 Z 值时，用于控制 α 的常数
μ_i–The underlying ‘true’ intervention effect ofthe i-thtrial
第 i 个试验的潜在“真实”干预效应
μ–The overall ‘true’ intervention effect

总体的“真实”干预效应

$\hat{\mu}$–The pooled intervention effect

合并的干预效应

σ^2–The variance ofδ

δ 的方差

α_i^2–The variance of Y_i

Y_i 的方差

τ^2–The between-trial variance

试验间方差

τ_{DL}^2 –The DerSimonian-Laird estimatefor the between-trial variance

试验间方差的 DerSimonian-Laird 估计量

τ_{SJ}^2 –The Sidik-Jonkman estimatefor the between-trial variance

试验间方差的 Sidik-Jonkman 估计量

$\hat{\theta}$–The pooled oddsratio of excluding zero-event trials

排除零事件试验后的合并比值比

Φ–The cumulative standard normal probability distributionfunction

累计标准正态概率分布函数

参考文献

(1) Pogue J, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled Clinical Trials*1997;18:580-593.
(2) Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. *Lancet* 1998;351:47-52.(3) Sterne JA, Davey SG. Sifting the evidence - what's wrong with significance tests? *BritishMedical Journal* 2001;322:226-231.
(4) Thorlund K, Devereaux PJ, Wetterslev J et al. Can trial sequential monitoring boundariesreduce spurious inferences from meta-analyses? *International Journal of Epidemiology*2009;38:276-286.
(5) Trikalinos TA, Churchill R, Ferri M et al. Effect sizes in cumulative meta-analyses ofmental health randomized trials evolved over time. *Journal of Clinical Epidemiology*2004;57:1124-1130.
(6) Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish whenfirm evidence is reached in cumulative meta-analysis. *Journal of Clinical Epidemiology*2008;61:64-75.
(7) Hu M, Cappeleri J, Lan KK. Applying the law of the iterated logarithm to control type Ierror in cumulative meta-analysis of binary outcomes. *Clinical Trials* 2007;4:329-340.
(8) Lan KK, Hu M, Cappelieri J. Applying the law of the iterated logarithm to cumulativemeta-analysis of a continuous endpoint. *Statistica Sinica* 2003;13:1135-1145.
(9) Ioannidis J, Lau J. Evolution of treatment effects over time: empirical insight from recursive cumulative metaanalyses. *Proc Natl Acad Sci U S A* 2001;98:831-836.
(10) Borm GF, Donders AR. Updating meta-analyses leads to larger type I errors than publication bias. *J Clin Epidemiol* 2009;62:825-830.
(11) Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficientinformation size and potentially false positive results in many meta-analyses. *Journal ofClinical Epidemiology* 2008;61:763-769.
(12) Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may beinconclusive - Trial sequential analysis adjustment of random error risk due to repetitivetesting of accumulating data in apparently conclusive neonatal meta-analyses. *InternationalJournal of Epidemiology* 2009;38:287-298.
(13) Higgins JPT, Green S. *Cochrane Handbook for systematic reviews of interventions, version5.0.0*. John Wiley & Sons, 2009.
(14) Bangalore S, Wetterslev J, Pranesh S, Sawhney S, Gluud C, Messerli FH. Perioperative betablockers in patients having non-cardiac surgery: a meta-analysis. *Lancet* 2008;372:1962-1976.
(15) Devereaux PJ, Beattie WS, Choi PT et al. How strong is the evidence for the use ofperioperative beta blockers in non-cardiac surgery? Systematic review and meta-analysis ofrandomised controlled trials. *BMJ* 2005;331:313-321.
(16) DeMets D, Lan KK. Interim analysis: the alpha spending function approach. *Statistics inMedicine* 1994;12:1341-1352.
(17) Lan KK, DeMets D. Discrete sequential monitoring boundaries for clinical trials.*Biometrika* 1983;659-663.
(18) Pocock S. Interrim analyses for randomized clinical trials: the group sequential approach.*Biometrics* 1982;38:153-162.
(19) GRADE Working Group. Grading quality of evidence and strength of recommendations inclinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and gradingquality of evidence

about interventions. *Allergy* 2009;64:669-677.
(20) Guyatt G, Mills E. In the era of systematic reviews, does the size of an individual trial stillmatter. *PLoS Medicine* 2008;5:e4.
(21) Sutton AJ, Cooper NJ, Jones DR. Evidence synthesis as the key to more coherent endefficient research. *BMC Medical Research Methodology* 2009;9.
(22) Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations forimproving the quality of reports of parallel-group randomised trials. *Lancet* 2001;357:1191-1194.
(23) Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size byquantifying diversity in a random-effects meta-analysis. *BMC Medical ResearchMethodology* 2009;9.
(24) Afshari A, Wetterslev J, Brok J, Moller A. Antithrombin III in critically ill patients:systematic review with meta-analysis and trial sequential analysis. *BMJ* 2007;335:1248-1251.
(25) Thorlund K, Anema A, Mills E. Interpreting meta-analysis according to the adequacy ofsample size. An example using isoniazid chemoprophylaxis for tuberculosis in purifiedprotein derivative negative HIV-infected individuals. *Clinical Epidemiology* 2010;2:57-66.
(26) Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance,not choice. *Lancet* 2002;359:519.
(27) Review Manager (RevMan) [Computer program]. Version 5.1. Copenhagen: The NordicCochrane Centre, The Cochrane Collaboration, 2011.
(28) Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reportingcharacteristics of systematic reviews. *PLoS Medicine* 2010;4:e78.
(29) Gehr B, Weiss C, Porzsolt F. The fading of reported effectiveness. A meta-analysis ofrandomised controlled trials. *BMC Medical Research Methodology* 2006;6:25.
(30) Jennison C, Turnbull B. *Group sequential methods with applications to clinical trials*.Chapman&Hall/CRC Press, 2000.
(31) Berkey C, Mosteller F, Lau J. Uncertainty of the time of first sifnificance in random-effectscumulative meta-analysis. *Controlled Clinical Trials* 1996;17:357-371.
(32) DerSimonian L, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials*1986;7:177-188.
(33) Sidik K, Jonkman J. Simple heterogeneity variance estimation for meta-analysis. *Journal ofRoyal Statistical Society(C)* 2005;54:367-384.
(34) Sidik K, Jonkman J. A comparison of heterogeneity variance estimators in combiningresults of studies. *Statistics in Medicine* 2007;26:1964-1981.
(35) Brockwell S, Gordon I. A comparison of statitical methods for meta-analysis. *Statistics inMedicine* 2001;20:825-840.
(36) Brockwell S, Gordon IR. A simple method for inference on an overall effect in metaanalysis. *Stat Med* 2007;26:4531-4543.
(37) Sanchez-Meca J, Martin-Martinez, F. Confidence intervals for the overall effect size inrandom-effects meta-analysis. *Psychological Methods* 2008;13:31-48.
(38) Sidik K, Jonkman J. Robust variance estimation for random-effects meta-analysis. *CompStat Data An* 2006;50:3681-3701.
(39) Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the randomeffects odel. *J Educa Behav Stat* 2005;30:261-293.
(40) Makambi KH. The effect of the heterogeneity variance estimator on some tests of efficacy.*J Biopharm Stat* 2004;14:439-449.

(41) Biggerstaff B, Tweedie R. Incorporating variability in estimates of heterogeneity in the random-effects model in meta-analysis. *Statistics in Medicine* 2009;16:753-768.
(42) Sweeting M, Sutton A, Lambert P. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis. *Statistics in Medicine* 2004;23:1351-1375.
(43) Armitage P. Sequential analysis in therapeutic trials. *Annual Review of Medicine* 1969;20:425-430.
(44) Pocock S. Group sequential methods in the design and analyss of clinical trials. *Biometrika* 1977;191-199.
(45) Whitehead J. *The design and analysis of sequential clinical trials*. John Wiley & Sons, 2000.
(46) Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine* 1991;10:1665-1677.
(47) Higgins JPT, Thompson S. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002;21:1539-1558.
(48) Federov V, Jones B. The design of multicentre trials. *Statistical Methods in Medical Research* 2005;14:205-248.
(49) Ioaniddis J, Patsopoulos N, Evangelou E. Uncertainty in heterogeneity estimates in metaanalysis. *British Medical Journal* 2007;335:914-916.
(50) Jackson D. The implications of publication bias for meta-analysis' other parameter. *Statistics in Medicine* 2006;25:2911-2921.
(51) Chan AW, Hrobjartsson A, Haahr M, Gotzsche P, Altman D. Empirical evidence for selective reporting of outcomes in randomized trials. comparison of protocols to published articles. *Journal of American Medical Association* 2004;291:2457-2465.
(52) Chan AW, Altman D. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Association Journal* 2004;171:735-740.
(53) Dwan K, Alman D, Arnaiz J et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS Medicine* 2008;3:e3081.
(54) Hrobjartsson A, Chan AW, Haahr M, Gotzsche P, Alman D. Selective reporting of positive outcomes in randomised trials - secondary publication. A comparison of protocols with published reports. *Ugeskr Laeger* 2005;167:3189-3191.
(55) Kjaergaard L, Villumsen J, Gluud C. Reported methodological quality and discrepancies between small and large randomized trials in meta-analyses. *Annals of Internal Medicine* 2001;135:982-989.
(56) Moher D, Pham B, Jones A et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;352:609-613.
(57) Schulz K, Chalmers I, Hayes R, Altman D. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of American Medical Association* 1995;273:408-412.
(58) Wood L, Egger M, Gluud LL et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemioogical study. *British Medical Journal* 2008;336:601-605.
(59) Guyatt GH, Oxman AD, Vist GE et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924-926.
(60) Reboussin DM, DeMets DL, Kyungmann K, Lan KKG. Programs for Computing Group Sequential Boundaries Using the Lan-DeMets Method, v.2. 2009.
(61) O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979;35:549-556.

(62) Pocock SJ. When to stop a clinical trial. *British Medical Journal* 1992;305:235-240.
(63) Awad T, Thorlund K, Hauser G, Stimac D, Mabrouk M, Gluud C. Peginterferon alpha-2a is associated with higher sustained virological response than peginterferon alfa-2b in chronic hepatitis C: Systematic review of randomized trials. *Hepatology* 2010;51:1176-1184.
(64) Rigotti N, Munafo MR, Stead LF. Interventions for smoking cessation in hospitalized patients. *Cochrane Database of Systematic Reviews* 2007.
(65) Moller CH, Penninga L, Wetterslev J, Steinbruchel DA, Gluud C. Clinical outcomes in randomized trials of off- vs. on-pump coronary artery bypass surgery: systematic review with meta-analyses and trial sequential analyses. *Eur Heart J* 2008;29:2601-2616.
(66) Brok J, Gluud LL, Gluud C. Meta-analysis: ribavirin plus interferon versus interferon montherapy for chonic hepatitis C - an updated Cochrane review. *Alimentary Pharmacology and Therapeutics* 2010;32:840-850.
(67) Brok J, Gluud C. Ribavirin monotherapy for hepatitis C. *Cochrane Database of Systematic Reviews* 2010;Issue 1.
(68) Ghandi GYMM, Flynn DN, et al. Effect of perioperative insulin infusion on surgical morbidity and mortality: systematic review and meta-analysis of randomized trials. *Mayo Clinique Procedings* 2008;83:418-430.
(69) Knorr U, Vinberg M, Kessing LV, Wetterslev J. Salivary cortisol in depressed patients versus control persons: a systematic review and meta0analysis. *Psychoneuroendocrinology* 2010; 35(9):1275-86.
(70) Nielsen N, Friberg H, Gluud C, Herlitz J, Wetterslev J. Hypothermia after cardiac arrest should be further evaluated - a systematic review of randomised trials with meta-analysis and trial sequential analysis. *International Journal of Cardiology* 2010.
(71) Rambaldi A, Saconato HH, Christensen E, Thorlund K, Wetterslev J, Gluud C. Systematic review of glucocorticosteroids for alcoholic hepatitis - a Cochrane Hepato-Biliary Group systematic review with meta-analysis and trial sequential analysis of randomised clinical trials. *Alimentary Pharmacology and Therapeutics* 2008;27:1167-1178.
(72) Tarnow-Mordi WO, Wilkinson D, Trivedi D, Brok J. Probiotics reduce all-cause mortality and necrotizing entercolitis: it is time to change practice. *Pediatrics* 2010;125:1068-1070.
(73) Whitfield K, Rambaldi A, Wetterslev J, Gluud C. Pentoxifylline for alcoholic hepatitis C. *Cochrane Database of Systematic Reviews* 2009.
(74) Whitlock R, Chan S, Devereaux PJ. Clinical benefit of steroid use in patients undergoing cardiopulmonary bypass: a meta-analysis of randomized trials. *European Heart Journal* 2008;29:2592-2600.
(75) Bangalore S, Kumar S, Wetterslev J, Messerli FH. Angiotensin receptor blockers and risk of myocardial infarction: meta-analyses and trial sequential analyses of 147 020 patients from randomised trials. *BMJ* 2011;342:d2234.
(76) Bangalore S, Kumar S, Wetterslev J et al. Carotid artery stenting vs carotid endarterectomy: meta-analysis and diversity-adjusted trial sequential analysis of randomized trials. *Arch Neurol* 2011;68:172-184.
(77) Bangalore S, Kumar S, Kjeldsen SE et al. Antihypertensive drugs and risk of cancer: network meta-analyses and trial sequential analyses of 324,168 participants from randomised trials. *Lancet Oncol* 2011;12:65-82.
(78) Bollen C, Uiterwaal C, Vught A, van der Tweel I. Sequential meta-analysis of past clinical trials to

determine the use of a new trial. *Epidemiology* 2006;17:644-649.

(79) Higgins JPT, Whitehead A, Simmonds M. Sequential methods for random-effects metaanalysis. *Stat Med* 2011;30:903-921.

(80) van der Tweel I, Bollen C. Sequential meta-analysis: an efficient decision making tool. *Clinical Trials* 2010;7:136-146.

(81) Whitehead A. A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Statistics in Medicine* 1997;16:2901-2913.