arXiv:2303.01040v1 [stat.AP] 2 Mar 2023

组序贯方法简介：规划与多方面优化

预印本

Fraser I. LewisGSK
fraser.i.lewis@gsk.com

2023年3月3日

摘要

在规划关键性试验时，组序贯临床试验设计颇具吸引力，因为这种方法能够在试验成功时提前停止试验，同时也广受监管审评领域认可。与单阶段设计相比，规划组序贯试验时需要考虑并优化的可调整参数更多。本教程将简要概述组序贯方法，然后详细介绍一些关键的运行特征，以及如何估算和优化这些特征，并最终在确定最终研究设计时向决策者呈现。

关键词 临床试验 · 组序贯 · 教程

1 引言

在许多治疗领域，关键性临床试验通常采用经典的单阶段设计，即在整个试验期间不断积累患者数据，只有试验完成后才对数据揭盲并执行预先计划的统计分析。组序贯临床试验设计将这种单阶段设计扩展为多阶段设计，使得每个阶段都可以针对累积的中期数据开展预先计划的揭盲分析。这些中期分析可能允许因疗效或无效性而提前停止研究。因此，组序贯设计可以成为有效的风险缓解策略，用以应对结果意外不佳（因而因无效性而停止）或结果意外地非常积极的情况；例如，在期中分析数据时已有极具说服力的证据，从而因疗效而停止。尽管组序贯设计在文献中已十分成熟，但在关键性研究中规划和优化此类设计的实践经验并不普遍。本文是一篇基于案例研究的教程，介绍如何从基本原理出发设计和优化组序贯设计，而无需使用专业临床试验软件或程序库。本文还提供了补充材料，其中使用 R代码完成所需的各项计算。

我们聚焦于一个具体的应用场景，其中临床研究问题是：试验方案与标准治疗之间的治疗效应差异是多少？每位患者要么是应答者，要么是非应答者（二分类终点）；研究成功则定义为有充分证据拒绝试验方案相对于标准治疗的劣效性零假设。下文讨论的各项设计与计算步骤同样适用于其他情形，例如连续型终点和优效性检验，但由于检验统计量不同，具体技术细节也会有所差异。

组序贯方法非常适用于关键性试验，这也是本文隐含的重点。考虑到关键性试验所需的资源以及之后可能进行的监管审评，其风险管理方法可以说与早期分期研究大不相同。本文考察一次期中分析，但相关方法和计算自然可以扩展到多次期中分析。期中分析的次数，尤其是在关键性试验中，需要审慎考虑，因为反复在研究进行期间揭盲数据会带来额外的操作复杂性和相关风险。

在以下各节中，我们首先简要概述组序贯方法的部分关键文献，以及一些替代方法，然后概述所需的关键设计特征和决策。接着，我们详细介绍这些设计的运行特征，并通过若干示例展示所需的计算，说明不同设计参数的作用。在结果部分，我们将逐步分析案例研究，展示如何计算所需的各项数值结果，并说明选择最优组序贯设计所涉及的多方面考量。最后，我们将进行简短讨论。

1.1 组序贯方法简要文献综述

临床试验中的组序贯方法历史悠久。在 Christopher Jennison 和 Bruce Turnbull 撰写的该领域标准著作《组序贯方法：临床试验应用》问世之前的 20 年里，相关研究成果已相当丰富；而这本书问世至今也已超过 20 年[1], 。组序贯方法也见于国际人用药品注册技术协调会的指导文件，以及美国食品药品监督管理局的《药物和生物制品临床试验的适应性设计：行业指南》（2019年11月）。组序贯方法经过充分实践和验证，因此已不再被视为新颖或复杂的设计方法，通常无需事先与监管机构进行科学咨询。这通常与自适应设计形成对比：在研究实施期间，可能会进行由数据驱动的重大设计变更（包括中期非盲查看数据），从而给关键性试验设计带来额外风险和复杂性（例如，参见 Bauer[2] 和 Wassmer[3]）。标准的组序贯设计不涉及由数据驱动的适应性调整；相反，所有基于中期非盲查看数据而作出的停止或继续决定，均依据预先规定的标准，从而确保对 I类错误率进行严格控制。

与单阶段设计相比，在其他条件相同的情况下，组序贯设计需要更大的样本量才能达到相同的检验效能，因为重复分析不可避免地会带来统计代价。根据组序贯设计中所作的设计选择，与单阶段设计相比，样本量的增加可能不大，但这也需要权衡其他运行特征。关于不同组序贯设计与单阶段设计的相对效率，现有文献已有详尽论述，其中包括 Jennison 和 Turnbull[1]。本文重点讨论如何选择组序贯设计，以实现预期的理想运行特征。

对于关键性研究，存在一些组序贯方法的非自适应替代方案，这些方案会在中期对数据进行非盲查看；其中主要的竞争方案可能是 Mehta[4;5]提出的“有希望区域”方法。有希望区域设计基于一套规则，允许根据对数据进行非盲中期查看所得的结果，将研究的样本量增加一定幅度。这种样本量调整通常无法在组序贯设计中实现（但 Cui[6] 是一个例外）。顾名思义，如果非盲中期查看所得的中期结果落在某个特定区域——有希望区域——内，那么就可以在不对 I 类错误造成不利影响的情况下，将样本量增加一定数额。这种设计易于使用，因此对关键性研究颇具吸引力。已有多篇文章对组序贯设计与有希望区域设计进行了比较[7;8]；在期中进行样本量重新估计，在期望样本量和检验效能方面是否比组序贯设计更有效，仍是一个尚无定论且因具体情形而异的问题。

2 方法

本节大量借鉴 Jennison[1] 的内容，并聚焦于一种具体应用场景：由两个治疗组组成、以二元主要终点为指标，且旨在证明非劣效性的随机对照试验。以下各小节将简要介绍设计中的一个方面或一项值得关注的运行特征。在第3节中，我们将展示实际计算的数值示例；实现细节和完整的 R代码将在补充信息中提供，并说明在不同设计的各项运行特征之间进行权衡、以实现整体最优设计时所需作出的选择。

2.1 假设检验的表述

与任何单阶段研究一样，我们需要进行主要假设检验。组序贯设计的不同之处在于，随着我们多次查看数据，会对这一假设进行多次检验。因此，显而易见，控制I类错误率是组序贯设计的重要组成部分。对于我们的案例研究，零假设和备择假设定义如下：

$$H_0 : p_t - p_c - \delta \leq 0 \tag{1}$$

and

$$H_1 : p_t - p_c - \delta > 0; \tag{2}$$

其中， p_t 和 p_c 分别表示试验方案和标准治疗下的应答者比例， δ 为非劣效界值。有效治疗的应答者比例更高，因此随着 $p_t - p_c$ 增加，对备择假设的支持也会增强。

2.2 检验统计量定义

对于式（1）—（2）中的假设检验，可以使用多种不同的检验统计量。我们将采用最简单的一种：比较两个比例并纳入非劣效界值的标准 Wald 检验统计量。也可以采用基于得分的检验（例如，参见[9;10]）；将其纳入组序贯设计的步骤与 Wald 检验相同，但代数运算要复杂得多。纳入界值后，用于比较两个比例的标准 Wald 检验如下：

$$Z_i = (p_t - p_c - \delta)\sqrt{\left(\frac{p_t(1-p_t)}{n_i} + \frac{p_c(1-p_c)}{n_i}\right)^{-1}}; \tag{3}$$

$$Z_i = \theta\sqrt{I_i}; \tag{4}$$

and

$$Z_i \sim N(\theta\sqrt{I_i}, 1); \tag{5}$$

其中，n_i 表示每个治疗组的样本量，i 用于标记对数据进行查看的时点。在标准的单阶段设计中，我们有 $n_i = N$，即总样本量。请注意，式（3）中的Wald 检验统计量写成“自然估计量”$\theta = p_t - p_c - \delta$与信息量 I_i 的乘积，其中自然估计量直接作为零假设陈述中的特征，而后者定义为自然估计量方差的倒数（遵循 Jennison[1]）。在组序贯方法中，检验统计量通常以信息量而非方差或标准误为参数。这种术语也适用于数据的累积，因为随着试验累积更多数据，这等同于获得更多信息量；信息量是 n_i 的递增函数（如式3所示）。

2.3 联合分布的一般表述

在单阶段试验中，假设检验关注一个固定时点——数据的总体累积量——此时我们需要知道样本量为 N时检验统计量 Z 的概率分布。在组序贯设计中，我们需要了解随着数据不断累积，随机变量 Z_i 的表现。这是因为在多次查看数据的情况下，计算目标运行特征（例如I类错误和检验效能）时会涉及多个 Z_i。因此，我们需要了解任意一组索引 i对应的 Z_i的联合分布。本节末尾将对该联合分布作出正式定义。第3节展示了随着数据累积，Z 如何变化的模拟结果。

2.3.1 联合分布的定义

在多次数据查看中，Z_is 的联合分布是推进组序贯方法所需的关键分析结果。这与单阶段设计中需要在零假设和备择假设下，确定所选检验统计量的抽样分布直接类似。这里的区别在于，由于每次非盲查看数据时都有一个相关检验统计量，因此我们需要多元抽样分布。对于每个零假设和备择假设，都需要这个多元抽样分布。下文给出的联合分布表述源自 Jennison[1] p.49。

对于组序贯设　　计，在数据查看总次数为 K 时，我们有一个由检验统计量组成的向量 $(Z_1, \ldots, Z_K)$，其中 e

$$(i) \quad (Z_1, \ldots, Z_K) \text{ is multivariate normal,} \tag{6}$$

$$(ii) \quad E(Z_k) = \theta\sqrt{I_k}, \quad k = 1, \ldots, K, \tag{7}$$

and

$$(iii) \quad Cov(Z_{k_1}, Z_{k_2}) = \sqrt{I_{k_1}/I_{k_2}}, \quad 1 \le k_1 \le k_2 \le K. \tag{8}$$

这是一个一般性结果，其适用范围与最大似然估计本身相同（参见 Jennison[1] 第 70 页）。不过，这些当然是渐近结果；在试验早期信息量较低时，这种分布近似的稳健性可能要差得多。

我们当前的应用场景使用 Wald检验，但上述分布结果具有普遍性，这意味着采用更复杂的基于得分的检验，例如用于风险差的 Miettinen和Nurminen检验[9]，也适用于这一表述，并展现了组序贯方法的广泛适用性。

总结到目前为止，我们已经完成了以下工作：i）陈述了假设检验；ii）确定了所使用的检验统计量；以及 iii）确定了检验统计量的概率分布。现在，我们定义仅包含一次期中分析的组序贯设计的运行特征。研究设计的运行特征是一些概率事件；为了使研究得到良好的控制并成功完成，我们希望将这些事件控制在适当的高水平或低水平。这些事件基于公式（6）至（8）中的联合分布定义。我们仅考虑一次期中分析，但结果自然可以扩展到多次期中分析。

2.4 设计的运行特征

2.4.1 错误率与临界值

在任何研究设计中，I类错误率和II类错误率显然都至关重要，其中，在关键性研究中，对I类错误的控制尤其关键。这些错误率由临界值决定；更确切地说，在设计研究时，需要确定临界值，使错误率得到控制并达到预期水平。

在单阶段设计中，错误率经过设计，应在样本量为 N（即研究结束时）达到预期水平。在组序贯情境下，情况更为复杂，因为我们需要考虑（至少）两组相互依赖的错误。按照惯例，我们用 α 表示发生 I 类错误的概率，类似地，用 β 表示发生 II 类错误的概率。因此，在期中分析时，错误概率为 α_1 和 β_1；根据结果，期中分析也可能是最终分析。随后，在最终分析时（试验结束时），错误概率为 α_2 和 β_2。
试验期间，α_k 和 β_k 的单项数值以及总值 $\alpha = \sum_i^K \alpha_i$ 和 $\beta = \sum_i^K \beta_i$ 可能都需要满足特定的设计特征。例如，对于关键性试验中的单侧主要假设，确保 $\alpha < 0.025$ 通常至关重要。请注意，此处的 α 现在由两个子部分组成，即 α_1 和 α_2（两个独立事件），因为每次查看数据时都有可能发生 I 类错误。错误控制既可以针对总错误，也可以针对单项错误分量。如果希望进行不止一次中期分析，那么 α_i 和 β_i 项的数量也会相应增加。

具有一次期中分析的组序贯设计，其错误率可形式化为：

$$\alpha_1 = P(Z_1 > c_{12} \mid H_0); \tag{9}$$

$$\alpha_2 = P(Z_1 < c_{12} \cap Z_2 > c_2 \mid H_0); \tag{10}$$

$$\beta_1 = P(Z_1 < c_{11} \mid H_1); \tag{11}$$

and

$$\beta_2 = P(Z_1 \in (c_{11}, c_{12}) \cap Z_2 < c_2 \mid H_1); \tag{12}$$

其中，c_{12}，c_{11}，c_2 分别为期中因疗效（成功）而停止、期中因无效性（失败）而停止，以及研究结束时（此时可能成功或失败）的临界值。根据我们的应用情境假设陈述，对于 Z_i的较大正值，H_0 被拒绝；而 $c_{12} > c_{11}$ 分别为疗效和无效性停止界值。研究结束时的错误，即 α_2 和 β_2，均为联合事件，因为要在最终分析时发生错误，试验不能已在期中分析时停止。由于 Z_2 是已包含在 Z_1 中的数据累积，因此研究结束时的事件与期中事件并不独立。因此，计算 α_2 和 β_2 需要使用式（6）—（8）中定义的联合概率分布。

在组序贯设计中，无效性决策通常被指定为非约束性，这正是上述所定义事件的情况。这可从式（10）中 α_2 的定义看出：只有在期中分析时 $Z_1 < c_{12}$ 的条件下，才可能在研究结束时发生I类错误。换言之，I类错误的定义忽略了期中的无效界值。约束性无效规则会将式（10）改为 $\alpha_2 = P(Z_1 \in (c_{11},\ c_{12}) \cap Z_2 > c_2 \mid H_0)$，这意味着如果 Z_1 低于无效界值，研究将停止；若研究在这种情况下仍继续，则总体I类错误将增大。非约束性无效性可以说是组序贯设计中更常见的假设。约束性或非约束性无效性不会影响II类错误事件。

2.4.2 错误消耗

上文提到，错误控制既可以针对总错误，也可以针对各个单项错误分量。如果先确定总I类错误，例如设为 $\alpha = 0.025$，那么按照组序贯设计的术语，我们可以决定在对数据进行的多次分析中如何“消耗”这 0.025 的错误率。同样，也可以对 II 类错误进行错误消耗。
在期中分析时决定消耗多少错误率，是制定最优组序贯设计的关键环节，因为这决定了研究在期中因疗效而停止的概率。在期中消耗总I类错误中较小的比例，意味着研究较不可能因疗效而停止；但如果确实停止，证据可能更有说服力。反之亦然。

文献中有许多不同方法，可在组序贯设计中分配I类和II类错误（见 Jennison[1] 第145页及其中的关键文献[11;12;13;14]）。严格来说，“错误消耗”指的是一组用于分配错误的方法，而其他经典方法，例如 Wang-Tstais[15], 提出的方法，同样会在各次期中分析中分配错误，但在方法学上有所不同。该方法还包含两个特例，分别对应可能是最常被提及的两种界值：较保守的 O’Brien & Fleming 界值和较不保守的 Pocock 界值。在我们的应用场景中，我们采用 Kim[12] 的错误消耗方法；该方法由 Jennison[1] 第148页推广，也是实践中常用的方法之一。

通过设计本身，错误消耗方法可以轻松地将额外的、未计划的期中分析纳入试验，同时仍能控制总体错误率。或许并不需要这种灵活性，但它可能会有帮助，而且这类方法的表现不逊于其他错误分配方法。为说明错误消耗的运作方式，以下表达式采用一个由参数 ρ 索引的错误消耗族（见[1] 第148页）。我们有

$$\alpha_1 = f_\alpha(I_1/I_2), \tag{13}$$

$$\alpha_2 = \alpha - f_\alpha(I_1/I_2) \qquad \text{where } f_\alpha(t) = t^{\rho'}\alpha, \tag{14}$$

$$\beta_1 = f_\beta(I_1/I_2), \tag{15}$$

and

$$\beta_2 = \beta - f_\beta(I_1/I_2) \qquad \text{where } f_\beta(t) = t^{\rho''}\beta. \tag{16}$$

参数 ρ' 和 ρ'' 分别决定总 I 类错误和总 II 类错误在中期分析与最终分析之间的分配比例。如果 $\rho' = 1$，疗效界值将具有类似 Pocock 的特性；如果 $\rho'' = 3$，无效界值将具有类似 O'Brien & Fleming 的特性。I_i 参数与前文相同，即各次分析时的信息率。

如公式 (13)–(16) 所示，将中期和研究结束时的各项错误相加后，部分项会相互抵消，从而确保总错误率符合设计要求。消耗函数 – $f_\alpha(t)$ 和 $f_\beta(t)$ – 将参数 t 设为每次中期分析时的信息量与最大信息量（即试验结束时的信息量）之比。采用这种方法，在控制错误率消耗并维持总体错误率目标的同时，新增一次未计划的中期分析也很容易处理。

将这些内容整合起来以适用于我们的情形，即使设计的错误要求与错误消耗方案相匹配，可得出以下需要求解的方程组：

$$P(Z_1 > c_{12} \mid H_0) = (I_1/I_2)^{\rho'}\alpha; \tag{17}$$

$$P(Z_1 < c_{12} \cap Z_2 > c_2 \mid H_0) = \alpha - (I_1/I_2)^{\rho'}\alpha; \tag{18}$$

$$P(Z_1 < c_{11} \mid H_1) = (I_1/I_2)^{\rho''}\beta; \tag{19}$$

and

$$P(Z_1 \in (c_{11}, c_{12}) \cap Z_2 < c_2 \mid H_1) = \beta - (I_1/I_2)^{\rho''}\beta. \tag{20}$$

方程组 (17)-(20) 包含了适用于我们情形的所有研究设计和抽样模型参数：

- 疗效参数 p_t，p_c 以及非劣效界值 δ
- 期中分析的位置，以总每组样本量 N 的比例 ψ 表示，而总每组样本量本身也是一个参数
- 总 I 类错误和总 II 类错误 α 分别为 β
- 错误消耗功效参数，ρ' 用于 I 类错误消耗，ρ'' 用于 II 类错误消耗
- 检验统计量停止界值的临界值，c_{11}，c_{12}，c_2

总计 12 个参数。我们有 4 个方程和 12 个未知量，因此需要预先固定其中 8 个，才能求解其余未知量。通过求解方程 (17)–(20)，可以确定满足给定目标误差所需的 N；或者，在样本量固定时求解检验效能（即将未知量 N 替换为未知量 β）。在结果部分，我们将展示一个如何求解该方程组以计算样本量和检验效能的示例。

2.4.3 穿越概率

对于设有期中分析的设计，另一个参数或运行特征是该期中分析时因疗效（成功）而停止研究的概率。这个概率的合理下限因具体情况而异。例如，如果治疗效应存在很大的不确定性，那么期中分析的目的可能是获取相关信息，因此该参数没有目标值。或者，如果我们对可能疗效的先验信念认为，将停止概率作为一项明确的设计目标特征是合适的，那么也可以如此规划。该运行特征也称为穿越概率，其定义可根据上述事件直接得出：

$$\omega_1 = P(Z_1 > c_{12} \mid H_1) \tag{21}$$

and

$$\omega_2 = P(Z_1 \in (c_{11}, c_{12}) \cap Z_2 > c_2 \mid H_1) \tag{22}$$

其中，式（21）表示因疗效而在中期停止的概率，即在备择假设下进入原假设拒绝域的概率；式（22）表示中期未停止（无论是因为疗效还是无效性），随后在最终分析时因疗效而取得成功的概率。这两种情况的概率之和等于该设计的检验效能，即无论在中期还是在研究结束时取得成功的总体概率。也可以计算因无效性而提前停止的类似穿越概率。

2.4.4 最小差异

引入期中分析以便提前停止，可能是降低试验风险的一种务实做法；然而，鉴于试验（因成功）提前于计划停止，还必须提供有说服力的证据，证明治疗效应稳健。美国食品药品监督管理局关于自适应设计的指南[16]指出：“停止规则要求结果具有很强的说服力，既要体现在治疗效应估计值的大小上，也要体现在效应证据的强度上。”对此的一种解读是，该指南建议不要仅凭统计显著性就停止试验；此外，治疗效应大小也应具有临床说服力。

在组序贯研究中，量化期中分析时治疗效应大小的一种方法是询问：观察到的最小治疗效应需要达到多少，才能使检验统计量落入期中分析的 I 类错误拒绝域？我们将其称为最小差异，并且可以在期中分析和最终分析时计算该值。

根据定义检验统计量的表达式（公式 3），我们有

$$d_1 = \hat{p}_t - \hat{p}_c = \delta + c_{12}\sqrt{I_1}^{-1} \tag{23}$$

and

$$d_2 = \hat{p}_t - \hat{p}_c = \delta + c_2\sqrt{I_2}^{-1}. \tag{24}$$

其中，$\hat{}p$ 表示观察到的样本估计值。这些表达式表明，在其他条件相同的情况下，错误率消耗规则越保守（即临界值越大），最小差异也越大。我们还可以看出，随着样本量增大，最小差异会收敛于非劣效界值，这符合预期。

2.5 优化运行特征

在前面的章节中，我们已经说明，在所有其他研究设计参数（不包括临界值）均具有已知的固定值或假定值时，如何计算研究所需的典型运行特征，例如样本量和检验效能。我们还可以更进一步，将此前固定的一些参数设为未知量，从而也能够设定其他运行特征的目标。由此，我们进入多方面设计规划优化领域：希望找到一种设计，使其在满足检验效能和I类错误要求之外，同时满足多个目标运行特性。

回顾一下，在组序贯规划情境中需要解决的一般数值问题是：我们有一个方程组和一组未知量。只要问题没有参数过多，我们就可以灵活决定哪些参数固定，哪些参数需要估计。这可以视为数值优化，因为我们通过算法估计未知参数，以满足其他目标设计特征。从概念上讲，这只是对类似的单阶段研究设计情形的扩展：在该情形下，我们固定 N 并估计检验效能，或者反过来，同时保持对I类错误的控制。然而，在组序贯设计中，我们拥有更多参数和设计特征，并且还涉及多个相关检验统计量，因此在数值上比单阶段设计更加困难。

在组序贯情境中，一个具有实际意义的例子是：除了满足给定的总体最低检验效能外，设计还应达到中期因疗效而停止的越界概率下限，并且还应达到最小差异目标。为了设计一项满足这三个目标设计特征（此外还包括一个最高总体I类错误要求，该要求隐含为一项给定的必要条件）的研究，我们有以下方程组 (25)-(30)：

$$P(Z_1 > c_{12} \mid H_0) = (I_1/I_2)^{\rho'}\alpha, \tag{25}$$

$$P(Z_1 < c_{12} \cap Z_2 > c_2 \mid H_0) = \alpha - (I_1/I_2)^{\rho'}\alpha, \tag{26}$$

$$P(Z_1 < c_{11} \mid H_1) = (I_1/I_2)^{\rho''}\beta, \tag{27}$$

$$P(Z_1 \in (c_{11}, c_{12}) \cap Z_2 < c_2 \mid H_1) = \beta - (I_1/I_2)^{\rho''}\beta, \tag{28}$$

$$\omega_1 = P(Z_1 > c_{12} \mid H_1), \tag{29}$$

and

$$d_1 = \hat{p}_t - \hat{p}_c = \delta + c_{12}\sqrt{I_1}^{-1}. \tag{30}$$

这些方程与 (17)-(20) 中的方程类似，但现在我们增加了两个方程，分别对应越界概率和最小差异。

为达到这些目标运行特性，我们需要求解式（25）至（30）中的方程组。为此，我们需要释放一些可调整参数（将先前固定的参数视为未知量，并允许对其进行估计），其中两个

合适的候选参数是错误消耗方案（例如疗效方面的 ρ'），以及进行期中分析时的总样本量占比 ψ。这样我们就有了包含六个未知量的六个方程。我们将在结果部分展示一个实现方法的示例，以及相关结果和输出。

2.6 数值方法

为了求解（25）–（30）中的非线性方程组，或类似但更复杂的示例（例如增加期中分析），我们需要使用多维求根算法。开源库和平台提供了这类方法，通常实现了牛顿法的变体（例如参见 Moré[17]）。将多维求根算法用于非线性方程组时，通常只能保证局部收敛，这意味着数值搜索的初始值可能很重要，有助于避免收敛问题。组序贯方法中的非线性方程组还存在一项额外的复杂性：其中包含多维积分，且积分限是求根算法中的参数，这可能导致数值稳定性问题。

下一节的案例研究结果采用了一种混合数值方法：求根针对的方程组比（25）–（30）更简单，部分参数未纳入求根，而是在网格上固定。在该网格的每个点上，运行求根算法以求解其余未知参数。与直接尝试求解（25）–（30）相比，这种方法的结果数值稳定性更高，尽管计算成本有所增加。补充材料中提供了完整详情和 R代码。

3 结果

本节将通过数值示例说明，如何从基本原理出发，计算仅设一次期中分析、且主要假设检验和检验统计量如上文所述的组序贯设计的设计运行特征。生成本文所示结果所需的全部 R 代码均见补充材料；本节稍后将提供一段代码，展示其中的关键部分。

3.1 组序贯设计的可视化

在展示样本量和检验效能计算的数值结果之前，我们首先通过可视化来了解 Z_i 检验统计量随 i变化的行为。图1展示了三次模拟，说明在备择假设[1]，下，随着试验不断累积新的数据点，随机变量 Z_i 可能呈现怎样的变化；其中 $i = 1, \ldots, 800$，分别对应所示的三次实现 Z_i （这里稍微滥用了符号，i 表示累积样本量本身，而非像方程3中那样用于标记期中分析）。每次模拟使用相同的参数集（$p_t = 0.58, p_c = 0.6, \delta = -0.1$），但随机种子不同，因此各条轨迹之间的差异仅由随机性造成。在备择假设下，随着数据累积，各条轨迹的平均趋势将向上。竖线表示可能进行期中分析和最终分析的位置，横线则表示临界值截点，其中较高的一条对应期中分析的临界值。根据图中所示的临界值可见，模拟一在期中分析时低于无效界值，如此低的观测值可能导致研究因无效性而停止。相比之下，模拟二和模拟三在期中分析时均处于继续区域，并且在研究结束时均超过疗效（成功）界值，而模拟一则属于II类错误。

3.2 R代码片段

代码清单 1 给出了示例性 R 代码，演示如何计算组序贯设计中的各种 I 类和 II 类错误，以及越界概率和最小差异。还展示了一个在 R 中进行多维求根的示例。完整的 R 代码见补充材料。

3.3 确定样本量和检验效能

为计算适用于我们案例的样本量，我们将 (17)-(20) 编写为 R代码，并采用以下参数值：($p_t = 0.58$，$p_c = 0.6$，$\delta = -0.1$，$n_1 = \psi N$，$\psi = 0.6$，$\rho' = 2$，$\rho'' = 3$)，其中 n_1 是期中分析时每组的样本量，N 是每组总样本量。此外，我们设定 $\alpha = 0.025$ 和 $1 - \beta = 0.9$。采用与代码清单 1 类似的代码，得到结果 ($c_{11} = 0.548$，$c_{12} = 2.366$，$c_2 = 2.04$，$N = 831.6$)：疗效在期中分析时的越界概率为 0.58，最小差异为 -0.026，疗效的错误率消耗（期中分析时的 P 值界值）为 0.009。我们使用 Cytel 的 EAST 对这些结果进行了核验。

[1]从技术上讲，每对数据点在这里 i 为简化起见，指的是一对患者的累积，每个治疗组各一名；我们假设入组比例固定为 50/50

代码清单 1：用于错误率消耗和求根的 R 代码片段。参数名称与正文中的名称保持一致。

```
library(mnormt)li
brary(rootSolve)

## 计算第一类和第二类错误

## 在中期分析时的第一类错误 A1<–sadmvn(lower=c(c12,–Inf),upper=c(Inf,Inf),mean=meanH0(...),varcov=covz1z2(...))

## 研究结束时的第一类错误 A2<–sadmvn(lower=c(–Inf,c2),upper=c(c12,Inf),
    均值=零假设下的均值（.....），协方差=z1 和 z2 的协方差（.....））## 中期 II 类错误概率 B1<–sadmvn（下限=c（–无穷大，–无穷大），上限=c（c11，无穷大），均值=备择假设下的均值（.....），协方差=z1 和 z2 的协方差（.....））

## 研究结束时的 II 类错误概率 B2<–sadmvn（下限=c（c11，–无穷大），
上限=c（c12，c2），
                                均值=备择假设下的均值（.....），
                                协方差=z1 和 z2 的协方差（.....））

## 交叉概率 CR1<–sadmvn（下界=c（c12，–无穷），上界=c（无穷，无穷），均值=meanH1（...），
协方差=covz1z2
                                                    （...））

## 最小差异 DF1<–d+c12/（sqrt（I_k（）））# d=NI界值，I_k（）计算信息


## I类错误支出在中期 A1_S<–（（I_k（p_t，p_c，n_z1）/I_k（p_t，p_c，n_z2））
^rho1）* alpha

## 研究结束时的 I 类错误率 A2_S<–alpha –((I_k(p_t, p_c, n_z1) / I_k(p_t, p_c, n_z2))^
rho 1) * alpha

## 中期分析时的 II 类错误率 B1_S<–((I_k(p_t, p_c, n_z1) / I_k(p_t, p_c, n_z2))^
rho 2) * beta

## 研究结束时的 II 类错误率 B2_S<–beta –((I_k(p_t, p_c, n_z1) / I_k(p_t, p_c, n_z2))^
rho 2) * beta

## 用于求根结果 <–c(f1=A1–A1_S, f2=A2–A2_S, f3=B1–B1_S, f4=B2–B2_
S)

## 调用求根函数 –findc11, c12, c2, Nmultiroot(f=
myResultFunction, start=c(c11_init, c12_init, c2_init, N_init),
parms=myExtraPars)
```

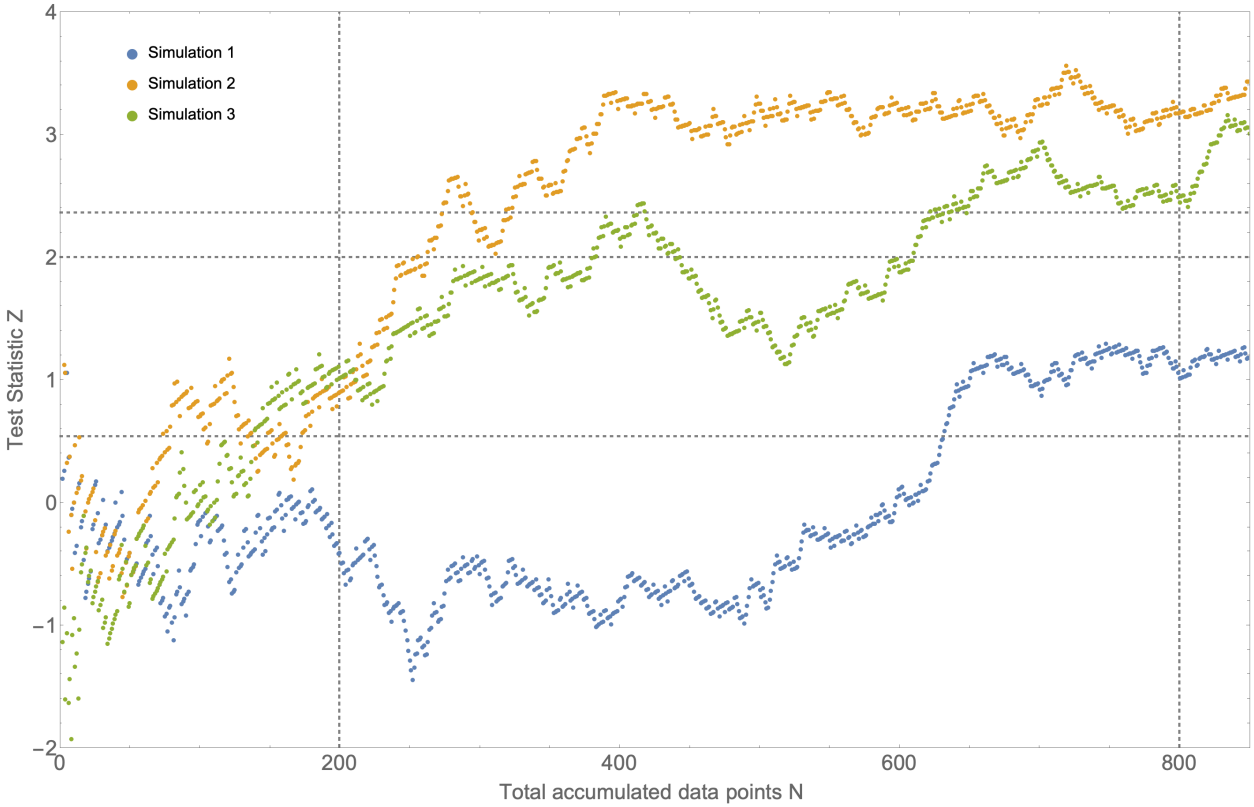


图1：随着试验不断累积新的数据（患者），检验统计量 Z_i（公式 (3)）的三种实现情况；所用参数见正文。竖线表示可能进行数据分析的位置；根据分析前希望累积的数据量，可将这些位置向左或向右移动。第一条竖线对应中期分析，第二条对应最终分析的位置。图中显示的是在 $n = 200$（每组）时进行期中分析，在 $n = 800$ 时进行最终分析。三条水平线表示检验统计量的临界值界限，可根据所需的运行特征上下移动。

要计算检验效能，我们使用几乎完全相同的代码，只需将参数 N 设为固定值，并将 β 作为待通过求根估计的参数。固定 $N = 831.6$ 后计算检验效能，便可得到与上文相同的结果。

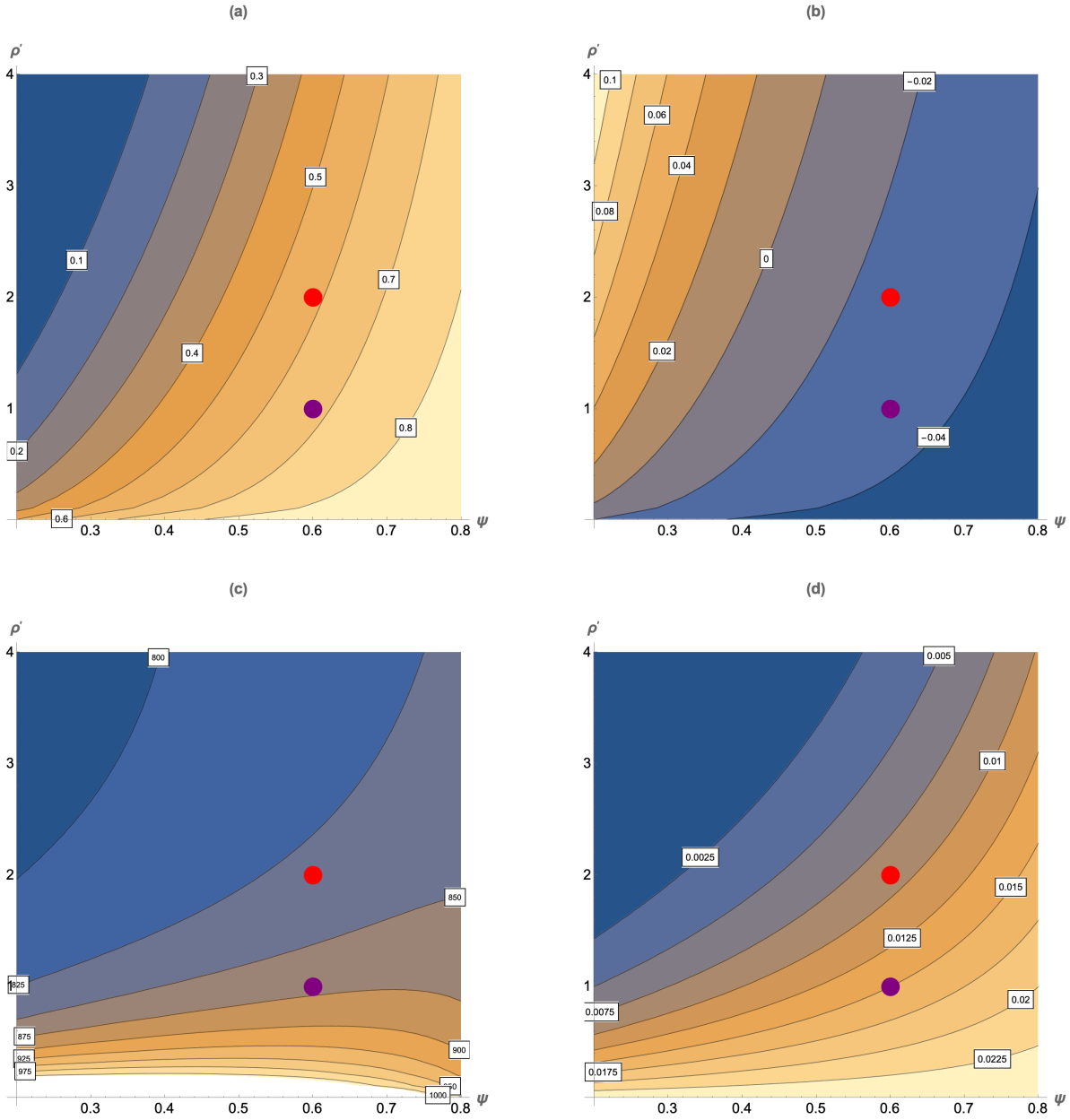
3.4 多方面示例——检验效能、越界概率和最小差异

在第2.5节中，我们详细介绍了设计研究时所需的表达式，以满足多个同时设定的目标运行特性，具体包括检验效能、中期分析时的成功（疗效越界概率）、最小差异和I类错误。在图2和图3中，我们展示了二维曲面上的求根结果，其中期中分析的位置（参数 ψ，即中期时总 N 的占比）以及疗效方面的中期错误率消耗（参数 ρ'）均有所变化。

图2中，检验效能固定为 $1 - \beta = 0.9$，II类错误消耗固定为 $\rho'' = 3$，总体 I类错误为 $\alpha = 0.025$。疗效参数与图1中相同，即（$p_t = 0.58$，$p_c = 0.6$，$\delta = -0.1$）。四个面板分别展示了在参数约束条件下，不同运行特征如何随 ψ 和 ρ' 变化。图中突出显示了两个点：一个对应（ψ，ρ'）=（0。6，2），另一个对应（ψ，ρ'）=（0。6，1）；后者采用了更激进的错误率消耗规则（类似于 Pocock）。从图2的面板（a）至（d）可以看出，对于更激进的组合（ψ，ρ'），越界概率从 0.58 增至 0.68；最小差异从 -0.026 降至 -0.034；每组总样本量从 831.6 增至 869.4；中期 I类错误消耗从 0.009 增至 0.015。

图3与图2类似，但此处设计将每组总样本量固定为831.6（即图2中采用 $\psi = 0.6$ 和 $\rho' = 2$时对应90%检验效能的数值），而不是固定检验效能。图中突出显示了两个点：一个点的（ψ，ρ'）=(0.6，2.5)，另一个点的（ψ，ρ'）=(0.5，1)。后者采用了更激进的错误率消耗规则，同时也将期中分析提前至招募进度达到50%时。再次比较图中的各个面板，我们可以看到

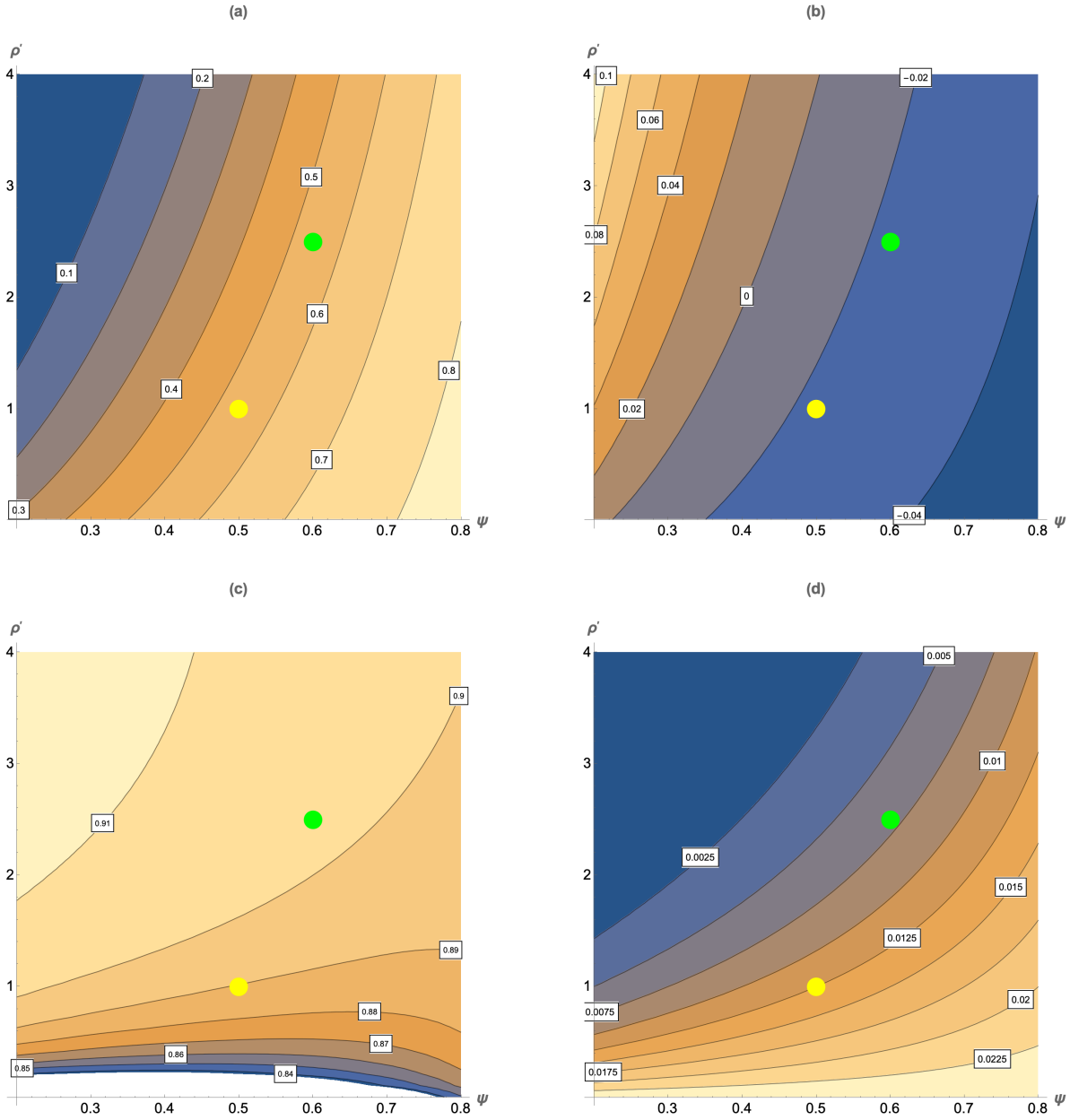
图2：多方面示例1——固定检验效能为90%，(a) 中期成功的越界概率，(b) 中期最小差异，(c) 每组总样本量，(d) 中期P值界值。有关高亮点的详细信息，请参阅正文。



更激进的消耗规则与更早开始招募在整体上对运行特征的影响大致相互抵消：越界概率从0.554略降至0.542；最小差异基本保持不变，为-0.023；总体检验效能从0.903略降至0.89；中期I类错误消耗（疗效的P值界值）则从0.007升至0.0125。

图2和图3展示了多方面设计规划的示例——我们尝试通过调整其他设计参数，同时优化多个设计参数，以实现各项参数之间的最佳权衡。这类似于采用单阶段设计进行常规研究规划：在固定I类错误的同时，权衡样本量与检验效能。组序贯设计的规划更为复杂，因为其参数和运行特征更多，而我们可能希望确保最终设计中的这些特征均达到至少最低限值。当前示例仅包含一次期中分析，也可以用类似方法分析增加期中分析次数的情况。

图3：多方面示例 2 - 固定 N=831.6，（a）中期成功的越界概率，（b）中期的最小差异，（c）每组总样本量，（d）中期的 P 值界值。高亮点的详情请见正文



这里展示的两个示例改变了中期分析的时点（ψ）和I类错误消耗（ρ'）。这只是众多可能方案中的一种；其他方案可以包括改变II类错误消耗，和/或将无效性越界概率作为一项明确的运行特征，同时改变疗效参数。哪些参数应保持固定（无论是预先设定，例如我们在此固定的假定疗效参数），哪些参数应在网格中变化，或通过求根确定，取决于具体问题。尽管有些参数更适合归入某一类，例如，临界值似乎更适合通过求根确定，而不是在网格中取值，因为我们通常不会直接选择临界值，它们是其他设计要求的副产品。

4 讨论

我们简要介绍了组序贯方法，并通过案例研究示例说明如何计算一些关键的运行特征。我们考虑的具体情形是一项以非劣效性为主要假设、终点为二分类终点的试验，但这些概念和通用方法广泛适用于其他类型的目标和终点。其他应用领域的示例可参见 Jennison 和 Turnbull[1]

从统计学角度来看，在研究设计中引入期中分析时，规划方面至少有两个重点。首先是任何期中分析的稳健性。如果计划在试验早期进行期中分析，那么检验统计量的抽样分布可能无法用标准解析形式（协方差为信息率之比的联合正态密度）很好地近似。实践中如何判断这一点可能颇具难度，模拟显然是一种选择。不过，至少对于关键性研究而言，通常样本量很可能足以确保渐近近似良好，例如患者人数达到数百人，而无需通过模拟加以验证。但也存在例外。

第二个重点是在组序贯设计中，就各种期望的运行特征开展利益相关方参与、教育和协调，并就其解读达成一致。这可以说是研究设计中最具挑战性的部分。试验设计中可能会纳入期中分析，因为人们强烈希望研究能够因成功而提前停止；而鉴于预先估计的治疗效应存在不确定性，组序贯设计可能是一个不错的选择。然而，此时有必要与临床负责人等关键利益相关方合作，确保他们清楚理解期中阶段组序贯运行特征的含义，并将其对风险的个人解读明确转化为例如目标穿越概率和最小差异。

使用专业软件，例如 Cytel 的 EAST 或 Solara，以及 NCSS 的 PASS，可以让试验统计学家轻松使用组序贯设计。这类软件包可能无法像定制代码那样提供充分的灵活性（例如我们案例研究示例中所示），但仍具备丰富的功能，适用于大多数使用场景。另一种选择是使用定制代码，其优势之一在于可以灵活搭配参数：我们有一个非线性方程组，可以根据所需的已知量与未知量组合进行求解。例如，前面我们求解出了一个设计，使其同时满足越界概率目标和最小差异目标。这种充分的灵活性固然有优势，但也伴随着使用临时代码通常会带来的风险，以及在药物临床试验质量管理规范（GCP）环境中进行质量保证所需承担的责任。尽管如此，正如上文所示，在 R 等平台上进行实际计算并不特别繁重。

此处呈现的所有结果均可使用补充材料中提供的代码和信息，以及 GitHub 上 <author 或 GSK 仓库中的内容复现：https:// >。

参考文献

[1] Christopher Jennison 和 Bruce W Turnbull。《组序贯方法及其在临床试验中的应用》。Chapman & Hall/CRC，博卡拉顿，2000年。ISBN 0849303168（碱性纸）。[2] Peter Bauer、Frank Bretz、Vladimir Dragalin、Franz König 和 Gernot Wassmer。确证性自适应设计的二十五年：机遇与陷阱。*Stat Med*，35(3):325–47，2016年2月。doi: 10.1002/sim.6472。[3] Gernot Wassmer 和 Werner Brannath。《临床试验中的组序贯与确证性自适应设计》。Springer药物统计学丛书。第1版，2016年版。ISBN 9783319325620。[4] Cyrus R Mehta 和 Stuart J Pocock。期中结果显示前景良好时自适应增加样本量：实用指南及示例。*Stat Med*，30(28):3267–84，2011年12月。doi: 10.1002/sim.4102。[5] Cyrus Mehta 和 Lingyun Liu。对自适应样本量重新估计的客观再评价：对《确证性自适应设计的二十五年》的评论。*Stat Med*，35(3):350–8，2016年2月。doi: 10.1002/sim.6614。[6] L Cui、H M Hung 和 S J Wang。组序贯临床试验中的样本量调整。*Biometrics*，55 (3):853–7，1999年9月。doi: 10.1111/j.0006-341x.1999.00853.x。[7] Christopher Jennison 和 Bruce W Turnbull。临床试验中的自适应样本量调整：先从小规模开始，再要求增加样本量？*StatMed*，34(29):3793–810，2015年12月。doi: 10.1002/sim.6575。[8] Julia M Edwards、Stephen J Walters、Cornelia Kunz 和 Steven A Julious。“前景区间”设计的系统综述。*Trials*，21(1):1000，2020年12月。doi: 10.1186/s13063-020-04931-w。[9] O Miettinen 和 M Nurminen。两个率的比较分析。*StatMed*，4(2):213–26，1985年。doi: 10.1002/ sim.4780040211。

[10] R G Newcombe。独立比例之差的区间估计：十一种方法的比较。*StatMed*，17(8):873–90，1998年4月。doi: 10.1002/(sici)1097-0258(19980430)17:8<873::aid-sim779>3. 0.co;2-i。[11] K. K. Gordon Lan 和 David L. DeMets。临床试验的离散序贯界限。*Biometrika*，70 (3):659–663，1983年12月。ISSN 0006-3444。doi: 10.1093/biomet/70.3.659。[12] Kyungmann Kim 和 David L. DeMets。基于第一类错误支出率函数的组序贯检验设计与分析。*Biometrika*，74(1):149–154，1987年3月。ISSN 0006-3444。doi: 10.1093/biomet/74.1. 149。[13] Irving K. Hwang、Weichung J. Shih 和 John S. De Cani。使用一族第一类错误概率支出函数的组序贯设计。医学统计学，9(12):1439–1445，1990年。doi: doi.org/10.1002/sim. 4780091207。[14] D L DeMets 和 K K Lan。期中分析：阿尔法支出函数方法。*Stat Med*，13(13-14):1341–52；讨论部分1353–6，1994年。doi: 10.1002/sim.4780131308。[15] Samuel K. Wang 和 Anastasios A. Tsiatis。组序贯试验的近似最优单参数界限。*Biometrics*，43(1):193–199，1987年。ISSN 0006341X, 15410420。[16] 美国食品药品监督管理局。药物和生物制品临床试验自适应设计行业指南。美国卫生与公众服务部、美国食品药品监督管理局，2019年2月。[17] Jorge J. Moré 和 Michel Y. Cosnard。非线性方程的数值解法。*ACM*数学软件汇刊，5(1): 64–85，1979年3月。ISSN 0098-3500。doi: 10.1145/355815.355820。URL https://doi.org/10.1145/355815. 355820。