平台试验：共同对照组对第一类错误和检验效能的影响

Quynh Nguyen[12] Katharina Hees[1] Benjamin Hofner[12]

[1] 生物统计学系，保罗·埃利希研究所，德国朗根[2] 医学信息学、生物统计学与流行病学系，埃朗根-纽伦堡弗里德里希-亚历山大大学（FAU），德国埃朗根，2023年2月10日

arXiv:2302.04713v1 [stat.ME] 9 Feb 2023

摘要

平台试验提供了一个框架，可在单项试验中研究多种干预措施，并可开放和关闭试验组。与单独对照组或每种治疗分别进行独立试验相比，在平台试验中使用共同对照组可提高效率。然而，由共同对照组导致的多重性调整需求，目前在研究人员、制药公司及监管机构之间仍存在争议。

我们通过一项模拟研究，考察了平台试验中共同对照组对第一类错误和检验效能的影响，并将其与采用单独对照组的平台试验所获得的结果进行比较。此外，我们还评估了在平台试验需要进行多重性调整的情况下，其对检验效能的影响。

在这两种研究设计中，若未进行多重性调整，家族错误率（FWER）相较于标准的双臂随机对照试验都会膨胀。采用共同对照时，族错误率膨胀幅度较小。在大多数情况下，经过多重性调整后，采用共同对照的平臺试验在样本量和检验效能方面仍有优势；但在某些情况下，采用共同对照的平台试验会失去效率增益。因此，我们进一步讨论是否有必要通过族定义或假设依赖关系进行调整。

关键词： 平台试验，I类错误，共同对照，检验效能

引言

在平台试验中，会评估多个治疗组，并且可以在正在进行的试验中，于不同时间点新增或撤除治疗。针对多个治疗组采用统一的主方案，可作为共同的物流、监管和方法学框架。在此前对主方案、篮式试验、伞式试验和平台试验开展的系统性文献综述中（Renfro 和 Sargent（2017年）、

Park 等人（2019年）或 Meyer 等人（2020年）），作者发现这类创新试验设计日益增多，尤其是在肿瘤学领域。此外，近年来，平台试验在评估 COVID-19 治疗方面也日益受到关注。平台试验可以灵活地新增或撤除治疗组，从而加快对新干预措施的评估，并将其纳入正在进行的试验。平台试验的例子包括 I-SPY2 试验（I-SPY 2：Quantum LeapHealthcare Collaborative，2022年）、STAMPEDE 试验（James 等人，2009年）和 RECOVERY 试验（RECOVERY：Nuffield 人口健康系，2022年）等。除了潜在的物流和运营优势外，采用共同对照也是许多平台试验的一项关键特征。与每个治疗组分别设置单独对照组相比，如果不进行多重性调整，采用共同对照可减少总样本量。

然而，研究人员目前对于平台试验中将多个治疗组与共同对照组进行比较时是否需要调整尚未达成共识（Molloy 等人，2022年；Wason 等人，2014年）。多组试验是否需要调整的争论由来已久，并随着主方案和平台试验的增多而愈演愈烈（Proschan 和 Follmann，1995年；Freidlin 等人，2008年；Stallard 等人，2019年；Parker 和 Weir，2020年）。主张进行调整的人认为，使用共同对照组会导致检验统计量之间存在依赖关系，因此在某些试验情境下需要进行调整（例如，在确证性情境中，或在评估同一治疗的不同剂量时）（Hung 和 Wang，2009年；Wason 等人，2016年）。对照组结果偶然偏低或偏高，可能导致更多假阳性或假阴性结果。尤其是在监管决策中，这种依赖关系可能产生影响，应当予以重视（Collignon 等人，2020年）。与此相反的观点认为，如果我们在独立试验中考察多种药物，就不会对这些试验进行调整，因此平台试验也不应进行调整。此外，Rothman（1990年）和 Perneger（1998年）均赞成不进行调整，并提出了一个哲学层面的问题：为控制家族错误率而纳入调整的检验应如何归入检验族，因为这似乎是一种任意选择。Hung 和 Wang（2009年）、Frane（2019年）等多位作者对此作了进一步讨论。作者强调了族定义的重要性，并指出，尽管这一界定在某种程度上具有主观性，但并非完全任意。Proschan 和 Waclawiw（2000年）根据临床试验中可能导致多重性问题的设计特点，以及假设之间的关联性、备择假设的性质等其他方面，针对可能需要进行调整的情形提出了指导意见（即，如果成功声明需要多个假设均达到显著性，对调整的要求会有所不同）。尽管 EMA 多重性问题指南（欧洲药品管理局，2017年）提到，家族错误率控制是确证性试验的先决条件，但该指南也被认为并未全面讨论涉及多个治疗组的所有情形。具体指导文件包括复杂临床试验启动与实施建议文件（欧洲药品管理局负责人（临床试验促进与协调组），2019年）、指南与 *FDA* 就药物和生物制品复杂创新试验设计进行沟通以及主方案：加速肿瘤学开发的高效临床试验设计策略（美国食品药品监督管理局，2020年、2022年），以及复杂临床试验——问答（欧洲药品管理局，2022年），这些文件均讨论了平台试验等复杂临床试验。

然而，其中并未深入讨论平台试验中是否、何时以及如何进行多重性调整。

因此，仅仅因为多个治疗组是在同一共同主方案下、使用共同对照进行评估，就认为平台试验中的多个治疗组检验总是需要进行多重性调整，这一点并不明确。仅仅使用共同对照，是否足以构成进行多重性调整的理由？事实上，与每个治疗组各自使用单独对照的平台试验相比，在平台试验中使用共同对照甚至会使家族错误率（FWER）更低。然而，Proschan 和 Follmann（1995年）、Howard 等人（2018年）以及 Bai 等人（2020年）等多位作者已经指出，尽管使用共同对照会降低 FWER，但对于 $k \geq 2$，至少出现 k 个假阳性结果的概率（ k-FWER ）与单独试验相比会更高。由于平台试验使用共同对照，相应的检验统计量彼此相关，因此会以相反的方式影响 $k \geq 2$ 的 FWER 和 k-FWER。这一发现为关于是否需要进行调整这一已经相当复杂的讨论增添了一个新的方面。

在一项模拟研究中，我们根据不同的I类错误和检验效能概念，评估采用共同对照的平 台试验的表现，并将其与每个治疗组均设有单独对照组的平 台试验进行比较；后者在不同对照组相互独立的假设下，就运行特征而言等同于针对每种治疗分别开展一系列双组随机对照试验。我们确定在未进行多重性调整时可能出现的误差膨胀，并评估在采用共同对照的平台试验中实施多重性调整后，为恢复检验效能所需增加的样本量。此外，我们还研究了与采用单独对照的平台试验，甚至与开展单独试验相比，在平台试验中进行调整是否仍能在样本量和检验效能方面带来益处。

本文结构如下。下一节将简要介绍一些符号以及若干误差和检验效能的定义。随后，我们介绍一项案例研究，该研究是开展模拟研究的动机，并为后文所述的模拟研究奠定基础。最后，我们总结并讨论研究结果。此外，我们还将在讨论部分进一步阐述由于多重性的不同来源以及假设依赖性而需要进行调整。

方法

检验统计量的记号与相关性

单独对照

首先，假设一个平台试验包含 m 种治疗，且每个治疗试验组均设有单独对照。令 X_i^j 表示患者 i的结局，其中 $i = 1, \ldots, n_j$，治疗 $j, j = 1, \ldots, m$，而 $X_i^{0j}, i = 1, \ldots, n_0^j$ 表示治疗 j对应的单独对照组中患者 i 的结局。假设结局服从正态分布，即 $X_i^j \sim \mathcal{N}(\delta_j, 1)$ 和 $X_i^{0j} \overset{d}{\sim} \mathcal{N}(0, 1)$（为简便起见，此处假设方差相等）

在 H_0 下，检验统计量 $\delta_j = 0\ \forall j$ 定义如下：

$$Z_j := \frac{\bar{X^j} - \bar{X^{0j}}}{\sqrt{\frac{1}{n_j} + \frac{1}{n_0^j}}} \sim \mathcal{N}(0,1),$$

其中 $\bar{X^j} := \frac{1}{n_j}\sum_{i=1}^{n_j} X_i^j$ 分别表示治疗_j 的单独对照组中，试验组 j 的患者的算术平均值。

共同对照组

现在，我们假设一个包含 m 种治疗和一个共同对照组的平台试验。同样，我们假设结局服从正态分布。现在，$X_i^0 \sim \mathcal{N}(0,1)$ $i = 1, ..., n_0$ 表示共同对照组中患者 i 的结局，n_0 表示共同对照组的总样本量。对于治疗组，我们可以继续使用上述符号。平台试验可以具有灵活性，即并非所有治疗都必须同时开始和结束招募。每种治疗都与共同对照进行比较，因此平台试验中又会进行 m 个假设检验。对于治疗组与共同对照的比较，仅使用对照组中同期入组的患者。因此，我们将上述 n_0^j 和 $\bar{X^{0j}}$ 的符号替换为以下表示。后者现在表示所有同期随机分配至试验组 j 的对照患者的平均值，而 $n_0^j := \#I_j$ 现在定义相应的对照患者数量。此外，我们将 I_j 定义为对照患者 $i = 1, \ldots, n_0$ 的子集，这些患者均同期随机分配至试验组 j，因此其数量为 $\#I_j = n_0^j$；并将 $n^{j,j'}_0$ 定义为同期随机分配至试验组 j 或试验组 j' 的对照患者数量，即 $n^{j,j'}_0 := \#(I_j \cup I_{j'})$。采用调整后的符号后，试验组 $j = 1, \ldots, m$ 与共同对照组进行比较的检验统计量，与单独对照组情形类似，定义为

$$Z_j^* := \frac{\bar{X^j} - \bar{X^{0j}}}{\sqrt{\frac{1}{n_j} + \frac{1}{n_0^j}}} \sim \mathcal{N}(0,1).$$

如上所述，使用共同对照会导致检验统计量之间存在依赖关系。更确切地说，针对 $j, j' = 1, \ldots, m$ 和 $j \neq j'$ 的两个检验统计量 Z_j^* 与 $Z_{j'}^*$ 之间的相关性为

$$\rho_{j,j'} = \frac{1}{\sqrt{\frac{1}{n_j} + \frac{1}{n_0^j}}\sqrt{\frac{1}{n_{j'}} + \frac{1}{n_0^{j'}}}} \frac{n_0^j + n_0^{j'} - n_0^{j,j'}}{n_0^j n_0^{j'}}, \tag{1}$$

且 Z_j 与 $Z_{j'}$ 的联合分布服从多元正态分布。关于公式 (1) 的证明，请参阅附录。如果所有治疗组的招募开始和结束时间均相同，即 $n_0^j = n_0^{j'} = n_0$，则公式 (1) 可简化为

$$\rho_{j,j'} = \frac{1}{\sqrt{\left(\frac{n_0}{n_j} + 1\right)\left(\frac{n_0}{n_{j'}} + 1\right)}}, \tag{2}$$

这与 Charles W. Dunnett（1955 年）已证明的结果一致。如果我们进一步假设治疗组和对照组的样本量相等（即 $n_0 = n_j = n_{j'}$），则两个检验统计量的相关性为 0。5。

为便于说明，我们来看一个图1所示的具体示例。假设将治疗组1、2和3与一个共同对照进行比较。在 n_{01} 名患者纳入对照组后，治疗组3加入该平台。每个条形表示各试验组招募期的开始和结束时间。治疗组和对照组的总样本量 n_1、 n_2、 n_3 和 n_0 可根据招募期的开始和结束时间进行划分，使得 $n_j = n_{j1} + n_{j2} + n_{j3}$，其中 n_{jt} 表示第 t期试验组 j 的样本量；若第 t期试验组 j 中未纳入受试者，则$n_{jt} = 0$ 。

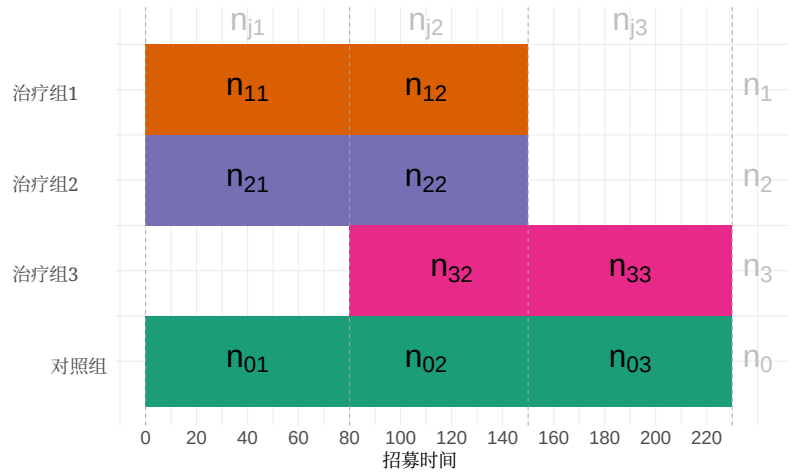


图 1: 平台试验示例：包含三个关注的治疗组和一个共同对照组 ol arm

如果现在每个治疗组只与同期随机分配的对照组患者进行比较 n_0^j，则治疗组2与对照组的检验统计量，以及治疗组3与对照组的检验统计量如下

$$Z_2^* = \frac{\bar{X}^2 - \bar{X}^{02}}{\sqrt{\frac{1}{n_{21}+n_{22}} + \frac{1}{n_{01}+n_{02}}}} \sim \mathcal{N}(0,1)$$

and

$$Z_3^* = \frac{\bar{X}^3 - \bar{X}^{03}}{\sqrt{\frac{1}{n_{32}+n_{33}} + \frac{1}{n_{02}+n_{03}}}} \sim \mathcal{N}(0,1),$$

其中， $\bar{X}^{02}$ 表示同期招募至治疗组2的$n_0^2 = n_{01} + n_{02}$ 对照组患者结局的算术平均值，而 $\bar{X}^{03}$ 表示同期招募至治疗组3的 $n_0^3 = n_{02} + n_{03}$ 对照组患者结局的算术平均值。可以看出，共同对照患者的数量现在由 n_{02}决定，因此相关性为 Z_2^*

以及 Z_3^*:

$$\rho_{2,3}^* = \frac{1}{\sqrt{\frac{1}{n_{21}+n_{22}} + \frac{1}{n_{01}+n_{02}}}\sqrt{\frac{1}{n_{32}+n_{33}} + \frac{1}{n_{02}+n_{03}}}} \frac{n_{02}}{(n_{01}+n_{02})(n_{02}+n_{03})}.$$

如果每个招募期的重叠患者数量相同，即$n_{21} = n_{01}$、 $n_{22} = n_{32} = n_{02}$ 和 $n_{33} = n_{03}$，那么相关性的公式可简化为

$$\rho_{2,3}^* = \frac{1}{2} \frac{n_{02}}{\sqrt{(n_{01}+n_{02})(n_{02}+n_{03})}}. \tag{3}$$

如果所有治疗组的招募开始和结束时间相同，且每组样本量相等（即 $n_{21} = n_{01} = n_{33} = n_{03} = 0$ 和 $n_{02} = n_{32} = n_{32}$），那么式（3）的结果为 0.5，与式（1）的结果一致。

采用共同对照的平 台试验总样本量为$N_{\text{CC}} = \sum_{j=1}^{m} n_j + n_0$。如果平台试验中每种治疗改为采用单独对照进行评估，并假设等比例分配，则总样本量为 $N_{IC} = \sum_{j=1}^{m} 2 \cdot n_j$。

表 1： 检验 m 假设时可能出现的结局

	H_0 真	H_0 假	总计
拒绝 H_0	V	S	R
不拒绝 H_0	U	T	$m-R$
总计	m_0	$m-m_0$	m

检验多个治疗组时的误差度量

经典的I类错误率定义为：在 H_0 为真的情况下，拒绝单个原假设 H_0 的概率。在检验一族 m 假设时，我们将重点关注以下错误率（假设采用表1中介绍的符号）：

- 家族错误率 (FWER)，其定义为至少出现一个假阳性发现的概率，即 FWER :$= P(V \geq 1)$;
- k族家族错误率 (k-FWER)，也称为假多重错误率，其定义为至少出现 k个假阳性发现的概率，即k–FWER :$= P(V \geq k)$，因此是 FWER 的推广;
- 以及 每族错误率(PFER)，即假阳性结果的期望数量，也就是 PFER :$= E(V)$。

有关错误率定义的更多细节，请参见 Dmitrienko 等（2009）或 Hochberg 和 Tamhane（1987）。
为了限制假阳性结果的数量，可以预先指定显著性水平或阈值 α，使所选错误率小于或等于

α。根据具体情况，控制某一种错误率可能比控制另一种更为重要。例如，对于确证性临床试验，以强意义控制家族错误率（即无论真实零假设的数量是多少）至关重要。如果所有零假设都成立，也称为全局零假设，则称家族错误率得到了弱意义上的控制。

如果以 α的显著性水平检验每个单独的假设，显然某些错误率可能会膨胀，例如，在全局零假设下，FWER $= 1 - (1 - \alpha)^m > \alpha$ 对于 $m > 2$。控制 m 假设这一族中所选错误率的方法包括但不限于闭合检验程序、Bonferroni或Dunnett校正。其中，后者利用了检验统计量之间的相关性。

平台试验中的效能定义

在进行假设检验时，除了I类错误外，还可能犯II类错误。互补事件的概率，即在备择假设为真时正确拒绝 H_0 的概率，也就是检测到特定治疗效应的概率，被称为试验的检验效能。下文将检验效能记为 $1 - \beta$，其中β 用于表示II类错误率。在检验多个假设时，对于特定的假设检验，该检验效能也称为**边际效能**。与错误率类似，在进行多个假设检验时，也可以定义以下检验效能概念（使用表1中的符号）：

- 析取效能，其定义为至少拒绝一个错误的 H_0 的概率，即 $P(S \geq 1)$;
- 以及合取效能，其定义为拒绝所有错误的 H_0 的概率，即 $P(S = m - m_0)$。

更多 有关不同检验效能概念的详情，请参见例如 Dmitrienko等（2009） .

案例研究

动机与设置

我们的模拟研究受到一项固定四臂平台试验案例研究的启发（包括三个治疗组和一个共同对照组）。该III期试验是作为EMA科学咨询程序的一部分提出的，因此其背景细节属于机密信息。每种治疗由不同的申办方提供，原计划分别与共同对照组进行比较，且未计划对三种治疗进行跨组分析。申办方认为，在共同主方案下评估这些治疗纯粹是出于后勤方面的原因。因此，申办方提议不对三个治疗组进行I类错误控制，并要求将每种治疗的结果视为相互独立，仿佛这是三项彼此独立的临床试验。由于每个治疗组都计划与同一个共同对照组进行比较，因此决策将不再相互独立，故对照组

治疗组间I类错误的控制（参照多重性指南（欧洲药品管理局，2017年））当时由EMA提出了要求。
申办方和监管机构提出的两种观点似乎都有道理，但显然相互矛盾。为了量化共同对照对I类错误率和整体决策的影响，我们开展了一项小型模拟研究。

按照提出的研究设计，我们假设每个试验组有 $n = 150$ 名患者，因此采用共同对照的平台试验的总样本量为 $N_{CC} = 4 \cdot 150 = 600$名患者。如果在平台试验中对各项治疗采用单独对照（或者类似地，在独立试验中进行），则需要 $N_{IC} = 3 \cdot 2 \cdot 150 = 900$名患者的总样本量才能达到相同的边际效能。在全局零假设下，我们假设每个治疗组和对照组的结局相互独立且服从均值为0、方差为1的正态分布。每项假设检验均采用双侧显著性水平 $\alpha = 0.05$。对于采用共同对照的平台试验，我们评估了以下三种情景下的错误率：不进行任何多重性调整、进行Bonferroni校正，以及进行Dunnett校正。对于采用单独对照的情况，则不考虑多重性调整。

扩展至灵活设计

在灵活平台试验中，一个或多个治疗组可以在之后的时间点加入平台。在加入时，其他治疗组和对照组可能已经开始招募。例如，我们对案例研究进行了修改，使第三治疗组在较晚的时间点开始招募。表2描述了这样一种情形：在每个试验组已有80名患者完成招募后，治疗组3加入平台。治疗组3与其他治疗组共享的共同对照患者为70名，而治疗组1和治疗组2共享150名对照患者。请注意，在前两种治疗停止招募后，与第三治疗组同期招募的额外对照组患者为 $N^*_{CC} = N_{CC} + 80 = 600 + 80 = 680$，因此，采用共同对照的平台试验中的总样本量增加了。这表明，在预期每个时间点各单独活性治疗组与对照组之间均等比例分配的情景下，采用共同对照的灵活平台试验会导致样本量增加。针对这一情景，我们还开展了与上述类似的模拟，并将结果与采用单独对照的灵活平台试验进行了比较；在后者中，每个治疗组和对照组均由150名患者组成。

表2： 包含三种治疗和一个共同对照的灵活平台试验示例。平台中每个试验组已招募80名患者后，才开始招募治疗组3。

	n_{j1}	n_{j2}	n_{j3}	
治疗组1	80	70	0	150
治疗组2	80	70	0	150
治疗组3	0	70	80	150
共同对照	80	70	80	230
	240	280	160	680

结果

动机案例研究的模拟结果见表3。在有共同对照和无共同对照的两种研究设计中，如果不进行多重性调整，平台层面的家族错误率都会明显膨胀（即 > 0.05）。虽然采用共同对照时家族错误率较低（0.1251，而无共同对照时为0.1403），但与未进行调整的单独对照相比，k-家族错误率（$k \geq 2$）却有所增加。这可以通过平台试验中使用共同对照来解释，因为这会导致检验统计量相关：如果共同对照中的随机低值导致出现一个假阳性发现，则进一步出现假阳性结果的可能性会增加，并且高于采用单独对照的平台试验。该随机低值会同等影响所有试验组，因此至少出现 k 个假阳性结果的概率会增加。另一方面，随机出现的共同对照高值会降低至少出现一个假阳性发现的概率，这是由于每项检验都共享该对照。两种研究设计中的每族错误率保持不变，这主要是因为多个随机变量之和的期望值等于其期望值之和。尽管采用共同对照的平台试验更可能产生多个假阳性结果，平均而言，其假阳性结果数量与采用单独对照的平台试验相同（Proschan 和 Follmann，1995年；Howard 等人，2018年；Bai 等人，2020年）。

如预期，使用 Bonferroni 或 Dunnett校正时，家族错误率控制在 0.05。后者纳入了检验统计量之间的相关性，因此比 Bonferroni 更为宽松，也能更充分地利用显著性水平。因此，它会提高检验效能，正如我们将在下一节中进一步讨论的那样。

表3： 全局零假设下，采用共同对照的四组平台试验与采用单独对照的平台试验中的错误率。表中给出了采用共同对照的平台试验在未进行多重性控制和进行多重性控制时的结果

	共同对照			**单独对照**
	未校正	Bonferroni	Dunnett	未校正
家族错误率，P(V≥1)	0.1247	0.0436	0.0489	0.1400
2-族错误率，P(V≥2)	0.0207	0.0046	0.0056	0.0073
3-族错误率，P(V≥3)	0.0030	0.0005	0.0007	0.0001
每族错误率，E(V)	0.1485	0.0486	0.0552	0.1475

我们进一步考虑了灵活平台试验，并扩展了案例研究。表4显示了灵活平台试验中的错误率：当第三个试验组的申办方加入时，平台中每个试验组已招募80名受试者。在采用共同对照组和采用单独对照的两种情形下，如果未进行多重性调整，家族错误率（FWER）和 k-FWER都会膨胀。由于并非所有治疗都同时开始，使用共同对照组时，共同对照患者的数量会减少，因此，与采用共同对照组的固定平台相比，采用共同对照组的灵活平台试验中的FWER更高。对于k-FWER，观察到的情况则相反： $k \geq 2$ 在采用共同对照组的灵活平台试验中，其数值更低。随着共同对照组数量减少，检验统计量之间的相关性

会降低，因此，采用共同对照组的平台试验中的错误率会逐渐接近采用单独对照时的错误率。此外，我们发现，采用单独对照时，固定平台试验和灵活平台试验的错误率基本相同（忽略由模拟误差导致的小幅差异）。

表4： 在全局零假设下，四臂灵活平台试验中采用共同对照与采用单独对照的平台试验的错误率比较，此时每个试验组已有80名患者入组。列出了采用共同对照的灵活平台试验在未进行和进行多重性控制时的结果

	共同对照			单独对照
	未校正	Bonferroni	Dunnett	未校正
家族错误率，P(V≥1)	0.1360	0.0463	0.0495	0.1411
2-族错误率，P(V≥2)	0.0148	0.0029	0.0033	0.0073
3-族错误率，P(V≥3)	0.0010	0.0002	0.0002	0.0001
每族错误率，E(V)	0.1518	0.0493	0.0530	0.1486

模拟研究

基于该案例研究，我们扩展了模拟，进一步评估与不采用共同对照组的平台试验相比，共同对照对 I 类错误率、检验效能以及样本量的影响。为了评估其对不同运行特征的影响，在每次模拟中，我们改变一个或多个关注参数，同时保持其他所有参数不变。针对每种情景$M = 50.000$ 均进行了模拟。

对于采用共同对照的平台试验，除不进行调整外，我们还考察了两种多重性调整方法（Dunnett 和 Bonferroni）。对于采用单独对照的平台试验，由于数据不共享，因此未对各治疗组进行调整。因此，这些设置可视为每个治疗组分别开展单独临床试验，唯一的区别是采用了共同的随机化框架。

固定平台试验

对于固定平台试验，所有治疗组同时开始并结束招募，我们主要关注治疗组数量增加时误差和检验效能率的变化。对于每种情景，我们将结局模拟为

$$X_i^j \overset{\text{iid}}{\sim} \mathcal{N}(\delta_j, 1) \quad i = 1, \ldots, n_j,\ j = 1, \ldots, m, \tag{4}$$

$$X_i^{0j} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1) \quad i =, \ldots, n_0^j,\ j = 1, \ldots, m, \text{ and} \tag{5}$$

$$X_i^0 \overset{\text{iid}}{\sim} \mathcal{N}(0, 1) \quad i = 1, \ldots, n_0, \tag{6}$$

分别对应治疗组、各自的对照组和共同对照组。通过改变治疗组中的效应量 δ_j，我们得以模拟以下不同情形。在全局零假设下，我们假设 $\delta_j = 0\ \forall j$，即每种治疗均无效。

零假设情形

为了评估共同对照对相关错误率的影响，我们在全局零假设下模拟了采用共同对照和单独对照的平台试验。

我们将治疗组数量 m 设定在2至10之间。对于两种设置，我们假设每组有 $n_j = n_0^j = n_0 = n = 150$ 名患者，共有 $j = 1$、.....、m，因此，在采用共同对照的平台试验中，总样本量为 $N_{CC} = n(m+1)$；采用单独对照时则为 $N_{IC} = n(2 \cdot m)$。

单一有效情形

为了评估研究设计（单独对照或共同对照）、试验组数量以及多重性校正对所需样本量的影响，我们模拟了一个有效的治疗，同时对照组和其他治疗组均保持无效。在不失一般性的情况下，我们模拟了 $\delta_1 = 0.38$，假设治疗组1有效，而其余治疗组为 $\delta_j = 0$、$j = 2$、.....、m。效应量的选择与案例研究相匹配：在双臂试验中，每组需要 $n = 150$名患者，才能以约90%的检验效能和5%的双侧显著性水平检测到0.38的治疗效应。我们将治疗组数量设定为$m = 2$、.....、10。各情景下，治疗组和对照组的每组样本量 $n = n_j = n_0^j = n_0$，对于$j = 1$、.....、m，其设定使有效治疗组获得90%的边际效能。

在实践中，临床试验的样本量并非总能任意大幅增加。原因可能包括可用患者数量有限或预算限制。因此，我们进一步探讨了总样本量固定为 $N_{CC} = N_{IC} = 600$ 时的边际效能。随着治疗组数量变化，每个治疗组的样本量也会相应变化。在采用共同对照的 $n_j = n_0 = 600/\ m+1$ $n_j = n_0^j = 600/\ 2 \cdot m$ 平台试验（ ）以及采用单独对照的平台试验（ ）中，采用共同对照的试验中每组样本量在 200 至 54 之间变化，而采用单独对照的试验中每组样本量在 150 至 30 之间变化。

全部有效治疗情形

在固定样本量的情景（$N_{CC} = N_{IC} = 600$）中，我们还评估了随着治疗组数量增加时的析取效能和合取效能$m = 2$，.....10。对于这些模拟，我们假设所有治疗均有效，因此模拟了 $\delta_j = 0.38\ \forall j$。

灵活平台试验

此处，我们主要关注加入具有共同对照组的平台试验的时间如何影响I类错误和效能率，因此为求简洁，将模拟限定为 ≤ 3 个治疗组。我们将这些结果与采用单独对照组的灵活平台进行比较。对于每种情景，我们参照公式（4）至（6）模拟结局，并根据加入时间调整各试验组的样本量。值得注意的是，在采用单独对照组的试验中，平台采用固定还是灵活的平台设计并无影响，因为由一个治疗组及其对照组组成的子试验不会受到新试验组开放的影响，至少在理想情况下如此。但在实践中，新治疗可能会改变受试者加入试验的意愿，进而影响整体患者群体。不过，这不属于本研究的范围。

零假设情形

为了评估共同对照以及加入平台的时间对错误率的影响，我们使用了上述示例（表2），其中包含 $m = 3$ 个治疗，并假设第三治疗组稍后加入试验。在全局零假设下，我们假设每个治疗组与对照组比较均包含150名患者对150名患者，即所有治疗组均为 $n_1 = n_2 = n_3 = 150$，共同对照组为 $n_{01} + n_{02} = n_{02} + n_{03} = 150$ （样本量定义见图1）。我们将前两个治疗组和共同对照组的患者数量n_{j1} 从0变动至150名患者。这表示治疗组3加入平台试验的时间，其中极端情况包括：试验组3与试验组1和2并行进行（$n_{j1} = 0$），以及试验组3在试验组1和2完成招募后才开始运行，即独立进行（$n_{j1} = 150$）。我们将结果与采用独立对照组的设计进行了比较；该设计被模拟为灵活平台试验，其中包含 $n_j = n_0^j = 150,\ \ j = 1, 2, 3$。

单一有效情形

我们沿用了固定设计中单一有效情形的相同方法，并假设只有一种有效治疗，其对应的 $\delta_3 = 0.38$；其他治疗均无效（$\delta_j = 0, j = 1,2$）。我们还设定，有效治疗组在稍后加入平台，并采用共同对照。与上文类似，我们改变前两个治疗组和共同对照组的患者人数 n_{j1}，范围为 0 至 150 人。我们据此确定每组所需的样本量，使有效治疗组的边际检验效能达到 90%。我们将总样本量与采用单独对照的设计进行了比较。在后一种设计中，我们模拟了灵活平台试验，设置为 $n_j = n_0^j = n = 150,\ j = 1, \ldots, m$，使有效治疗组的边际检验效能达到 90%。

与之前一样，我们也考虑了资源受限的情景。显然，如果像上文一样，将灵活平台的总样本量固定为 N_{CC}，那么加入平台越晚的申办方，其检验效能损失就越大，因为该平台采用共同对照。第三个申办方加入试验越晚，第一阶段中无法使用的患者数量就越多 $n_{j1},\ j = 0, 1, 2$（见表2），而这些患者原本属于该申办方

。由于总样本量固定，这只会减少该申办方治疗组与对照组比较的比较层面样本量。因此，我们考虑了这样一种情景：第三个申办方拥有固定资源，可以将其用于采用单独对照的试验（或者在单独试验中忽略增加的管理费用），也可以用于采用共同对照的正在进行的平台试验。我们假设，新加入的治疗组申办方最多可纳入300名患者。在采用单独对照且等比例分配的试验中，这意味着每组有150名患者。在采用共同对照的灵活平台试验中，申办方可以利用共同对照的优势。表5展示了一个示例：第三个申办方加入采用共同对照的平台时，其他治疗组已各纳入90名患者。我们假设，第三个申办方同等分担第2期纳入的$n_{02} = 60$ 共同对照患者，即负责$60/3$名患者。因此，申办方3可在第2期和第3期将 $300 - 60/3$ 名患者分配给第三种治疗（n_{32} 和 n_{33}），并在第3期将n_{03}名患者分配给对照组，从而使治疗组与对照组比较中的患者人数相等。我们研究了这种预算驱动的样本量选择对申办方3边际效能的影响。在模拟中，当第三种治疗加入采用共同对照的平台试验时，我们将前两种治疗组和共同对照组中已纳入试验的患者人数 n_{j1} 从0变化到150。对于采用单独对照的设计，我们模拟了每组150名患者的灵活平台试验。

表5： 包含三种治疗和一个共同对照组的灵活平台试验示例。平台每个试验组已招募90名患者后，治疗组3开始招募。星号（*）表示申办方3招募的300名受试者。

	n_{j1}	n_{j2}	n_{j3}	
治疗组1	90	60	0	150
治疗组2	90	60	0	150
治疗组3	0	60^*	110^*	170
共同对照组	90	$40 + 20^*$	110^*	260
	270	240	220	730

全部有效治疗情形

我们在预算驱动情景下，针对 $m = 3$ 种治疗，按每项比较 300 名患者的样本量评估了两种检验效能概念。我们假设所有治疗均有效 $\delta_j = 0.38,\ \ j = 1,\ 2,\ 3$，并按照上文所述，将 n_{j1} 从 0 变动至 150。对采用单独对照组的灵活平台试验进行模拟时，每组纳入 150 名患者，共 150 对 150 名患者。

结果

固定平台试验

零假设情形：错误率比较

图2展示了在全局零假设下，随着治疗组数量增加，采用和不采用共同对照组的平台试验的家族错误率膨胀以及 2-族错误率和 3-族错误率的变化。在我们的案例研究中，对于 $j = 3$ 个治疗组，两种设计的家族错误率膨胀差异较小；但随着治疗组数量增加，差异也随之增大。2-族错误率和 3-族错误率则呈现相反的趋势。从数值上看，2-族错误率远低于家族错误率，但在采用共同对照的平台试验中超过 0.05（ $j = 5$ ），而采用单独对照组时也超过 0.05（ $j = 8$）。进行 Bonferroni 或 Dunnett 校正后，家族错误率得到控制，因此 k-族错误率也低于 0.05。正如案例研究中所讨论的，每族错误率在两种设计中相同（此处未展示结果）。

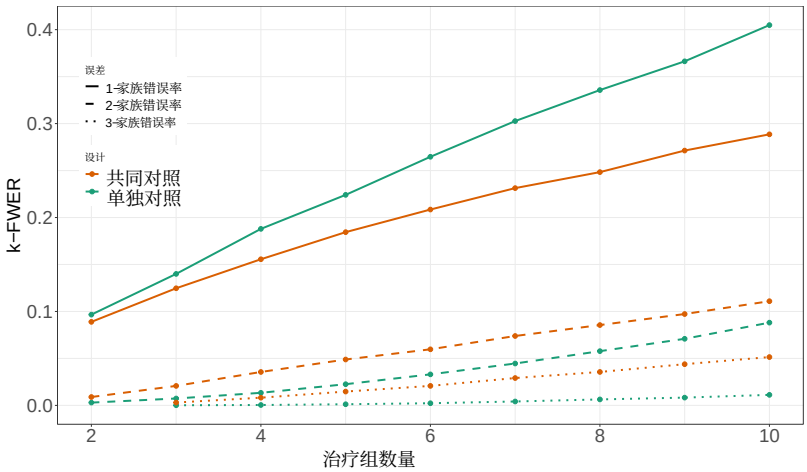


图2： k-家族错误率 ($k = 1$，2，3) 在一项包含 $j = 2$，……，10 个治疗组的平台试验中，治疗组与共同对照（绿色）及单独对照（橙色）进行比较，且未进行多重性调整。

单一有效情形：边际检验效能比较

要达到 90% 的边际检验效能，采用共同对照组时所需的总样本量始终显著低于采用单独对照的平行试验（图3，左）。正如预期，试验组数量增加时，总样本量也随之增加。同样，在总样本量固定时，采用共同对照组的边际检验效能始终高于采用单独对照时的边际检验效能（图3，右），并且会随着试验组数量增加而降低。采用共同对照组的平台试验具有更高的边际检验效能，这是总样本量固定的自然结果。共同对照组使每项比较的样本量更大。例如，在设有 3 个治疗组的平台试验中，采用共同对照组时每组分配 150 名患者，而采用单独对照时每组有 100 名患者。

如果预期对采用共同对照的平台试验进行多重性校正，就需要付出检验效能或样本量方面的代价。该

平台试验中采用共同对照时，总样本量会增加。尽管如此，为了获得例如90%的边际效能来检测特定治疗效应，采用单独对照所需的总样本量仍高于采用共同对照。图3（左）展示了试验组数量逐渐增加时的结果。正如预期，Dunnett校正比Boferroni校正的检验效能更高，即所需样本量更小。不过，获益相对有限。

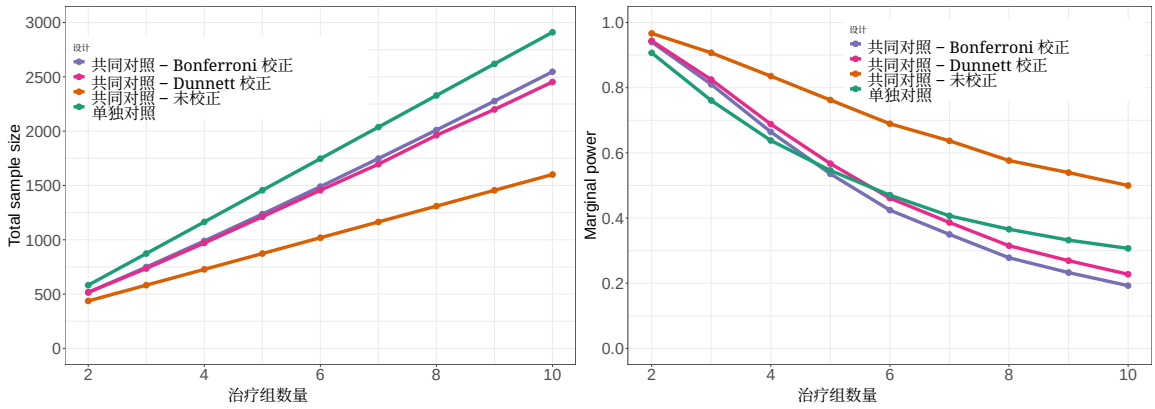


图3： 左图：在平台试验中，为以 90% 的边际效能检出特定治疗效应，且每个试验组的样本量相等时，采用共同对照与不采用共同对照所需的总样本量；治疗组数量为 $m = 2$、.....10。右图：总样本量固定为 $N_{CC} = N_{IC} = 600$ 位患者时，采用共同对照或单独对照的边际效能。

在总样本量固定而试验组数量不断增加的情况下，或者等价地，在每个试验组的样本量变得很小的情况下，多重性调整可能会变得过于严格。因此，与采用单独对照的试验相比，检验效能可能会降低（图3（右））。当治疗组数量达到 5 组或以上时，采用共同对照并进行 Bonferroni 校正的平台试验，其检验效能小于或等于采用单独对照的试验；当治疗组数量达到 6 组或以上时，采用 Dunnett 校正的检验效能也小于或等于采用单独对照的试验。Charles W. Dunnett（1955）指出，在采用共同对照的试验中，将等比例分配从 1:1:...:1 改为 1:1:...:$\sqrt{m}$，即向共同对照组分配 $\sqrt{m \cdot n}$ 名受试者，向治疗组分配 n 名受试者，可获得最优检验效能。采用这种最优分配后，对照组的患者数量增加，因此每项比较的受试者总数也随之增加。此外，检验统计量之间的相关性降低，因此当治疗组数量达到 8 组或以上时，临界点会向右移动（此处未展示结果）。这表明，在某些情况下，平台试验采用共同对照相较于为每个治疗组分别开展试验并采用单独对照，可能无法发挥共同对照所带来的优势；具体取决于治疗组数量和总样本量，尤其是在需要进行多重性控制时。

全部有效治疗情形：析取功效与合取功效的比较 r

与一类误差一样，在多重检验情形下，也可能需要考察平台试验的“总体”检验效能，即析取效能和合取效能。图4展示了在有共同对照和无共同对照的平台试验中，至少检测出一种有效药物的概率（析取效能；虚线），以及检测出所有有效药物的概率（合取

效能；实线）。对于采用单独对照的试验，检测出至少一种有效药物的概率更高；而对于采用共同对照的试验，检测出所有有效药物的概率更高。与对家族错误率和 k-家族错误率的解释类似，在平台试验中使用共同对照会影响这两种检验效能概念。如果已经检测出一种有效药物，那么检测出其余所有有效药物的可能性就会增加。另一方面，如果未能检测出某种特定的有效药物，那么其他有效药物也未能被检测出的可能性就更大。原因仍在于共同对照，它可能出现随机偏高或偏低。当采用Bonferroni 或 Dunnett校正时，与采用共同对照但不进行校正的设计相比，析取效能和合取效能都会降低（此处未展示结果）。尽管采用校正和共同对照的设计，其合取效能仍高于采用单独对照的设计，但其析取效能更低，因为未经校正的平台试验所得到的检验效能本就低于采用单独对照的设计。

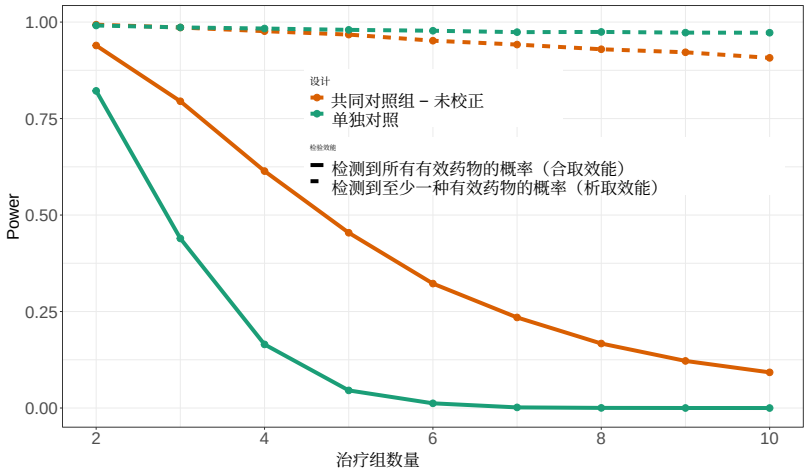


图 4： 固定样本量为 $N_{CC} = N_{IC} = 600$ 患者时的析取效能（检测到至少一种有效药物的概率）和合取效能（检测到所有有效药物的概率）。

灵活平台试验

零假设情形：错误率比较

图5展示了第三治疗组加入采用共同对照的固定平台试验后，随着平台上患者人数增加，家族错误率和2-族错误率的变化（共同对照的样本量为 n_{01}；样本量的定义见图1）。治疗组3加入平台越晚，该治疗组与其他治疗组之间的共同对照患者就越少。因此，相关性降低，采用共同对照的平台试验中的家族错误率便逐渐趋近于采用单独对照时得到的家族错误率。2-族错误率和3-族错误率也呈现相同趋势。最左侧的情景（治疗组3加入平台时，前两个治疗组和共同对照组均未招募患者）代表固定平台；最右侧的情景（治疗组3加入平台时，每个试验组已招募150名患者）则表示第三治疗组与另外两个治疗组之间没有重叠的对照组患者——类似于独立设定。

在这种情况下，图示情景中的效应并不十分显著，这一点

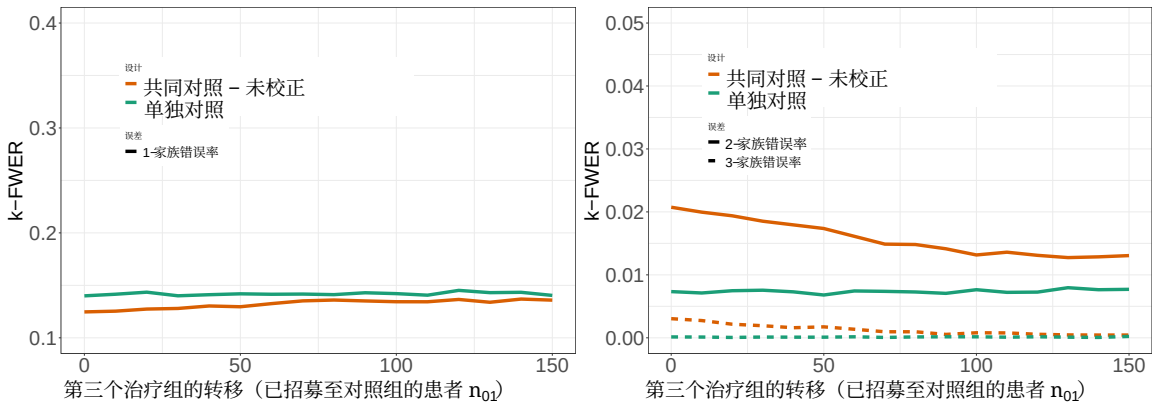


图5： 在包含3种治疗和一个共同对照组或单独对照组的灵活平台试验中，1-家族错误率（左图）、2-族错误率和3-族错误率（右图）。横轴表示第三种治疗进入平台时，共同对照组中的试验患者人数（组规模为 n_{01}）。

这是由于平台中的治疗组相对较少，因此，与共同对照组相比，采用单独对照组时家族错误率的增幅也相对较小。此情况下，家族错误率、2-族错误率和3-族错误率并未完全重合，是因为前两种治疗之间仍存在相关性，因此，采用共同对照组的平台试验整体上家族错误率有所降低，而k-家族错误率有所升高（此处未展示3-族错误率的结果）。采用Bonferroni或Dunnett校正时，家族错误率受到控制，因此 k-家族错误率也低于0.05。

单一有效情形：边际效能比较

图 6（左）显示了第三治疗组加入平台试验时，平台上已有患者人数不断增加的情况下，共同对照平台试验与采用单独对照的平台试验所需的总样本量。对于采用单独对照的平台试验，要达到 90% 的边际效能，所需总样本量保持不变。对于采用共同对照组的平台试验，所需总样本量会随着平台上已有患者人数的增加而增加。由于比较时仅使用同期入组的对照组患者，共同对照平台试验中的对照组样本量以及总样本量会自动增加 n_{01}，即非同期对照组患者的数量。纳入多重性调整后，增幅会更大，并随着 n_{01} 的增加而趋近于采用单独对照时所需的样本量。

图 6（右）显示了当平台试验中已有两个治疗组与一个共同对照组开展试验时，治疗组3的招募开始时间以 n_{01}为幅度逐渐推迟所对应的边际效能。如表5所述，假设治疗组3的申办方总共可以招募300名患者，并在稍晚时间加入平台。只要该申办方能够使用任何共同对照患者，新活性治疗组就可以获得 > 150名患者的样本量，并配有共同对照。在采用单独对照的试验中，样本量在治疗组和对照组之间平均分配，即各150名。在采用共同对照的平台试验中，随着

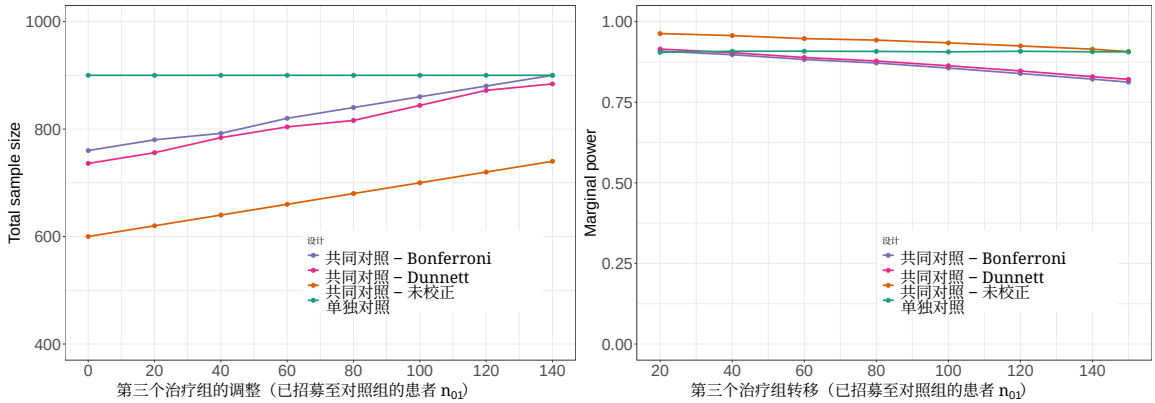


图 6： 左：为获得 90% 的边际效能，包含 3 个治疗组且有无共同对照组的平台试验的总样本量。右：比较样本量固定为 300 例患者时的边际效能。随着治疗组 3 加入带有共同对照的平台试验的比例增加，共同对照的使用量减少。

第三治疗组的招募开始时间逐渐推迟，共同对照患者的数量会减少。随着推迟幅度增加，可从共同对照中使用的患者越来越少，比较层面的样本量逐渐趋近于采用单独对照时的样本量。在极端情况下，当治疗组3加入平台时，其他治疗组已经招募了150名对照患者，则无法使用任何共同对照患者。在这种情况下，获得的比较层面的样本量与采用单独对照时相同，即治疗组和对照组各有150名患者。

对于采用共同对照并进行多重性调整的平台试验，可获得与固定平台中类似的观察结果。当招募仅发生小幅转移（n_01）时，与采用单独对照的试验相比，采用共同对照并进行调整的试验具有更高的边际效能。然而，随着转移程度增加，共同对照的数量减少，导致比较层面的样本量减少，进而使检验效能降低；同时，调整会进一步降低检验效能。当带有共同对照的平台试验中每个治疗组已招募超过 40 例患者时，进行多重性调整后的边际效能低于采用单独对照的平台试验。在这种情况下，很容易证明，治疗组 3 和同期对照组的样本量均为 187。尽管这一样本量高于采用单独对照时的样本量（150 vs. 150），但多重性调整抵消了这一优势。因此，对于申办方而言，如果预先计划进行多重性调整，加入正在进行的带有共同对照的平台试验将是不利的。

全部有效治疗情形：析取效能与合取效能的比较

图7展示了在全部有效治疗情形下，包含三个治疗组的灵活平台试验中析取效能（检测出至少一种有效药物的概率）和合取效能（检测出所有有效药物的概率）。横轴显示平台试验中第三个治疗组的启动时间（试验规模为 n_{01}）。随着该治疗组延迟时间的增加，共同对照组的数量减少，因此可以看出，采用共同对照的试验的合取效能逐渐趋近于采用单独对照的试验的检验效能，而析取效能仅表现出极小程度的趋近

。当进行Bonferroni或Dunnett校正时，与未经校正且采用共同对照的设计相比，析取效能和合取效能均有所降低（此处未展示结果）。

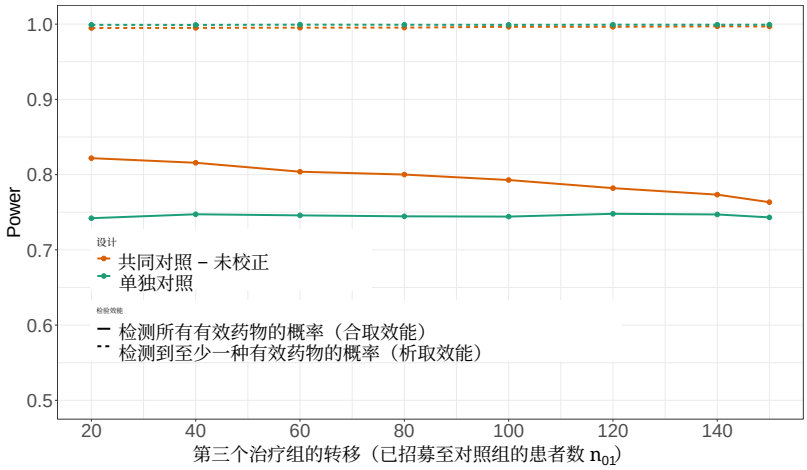


图7： 固定样本量下的析取效能（检测到至少一种有效药物的概率）和合取效能（检测到所有有效药物的概率）

讨论

多重性调整对样本量和检验效能的影响

我们已经表明，在许多情况下，与采用单独对照的试验相比，采用共同对照的平台试验在样本量和检验效能方面具有优势。对于给定的目标边际效能，采用共同对照可大幅减少所需的总样本量。对于固定的总样本量，采用共同对照可提高平台试验的边际效能。

显然，无论是否采用共同对照，在未经多重性调整的情况下，平台试验的家族错误率都会随着治疗组数量增加而升高。事实上，由于检验统计量之间存在相关性，与采用单独对照的试验相比，采用共同对照的平台试验中家族错误率的膨胀甚至更小。随着相关性减弱，其家族错误率会趋近于采用单独对照的平台试验的家族错误率。例如，增加分配给对照组的患者数量，或延迟一个或多个治疗组的招募，都可以降低相关性。另一方面，k-家族错误率则呈现相反的变化。由于采用共同对照，在采用共同对照的平台试验中，出现两个或更多假阳性结果的概率更高。不过，增幅要小得多。

我们进一步研究了平台试验中采用共同对照时，多重性调整如何影响样本量或检验效能。在许多情况下，与采用单独对照相比，在采用共同对照的设计中进行多重性控制，仍可提高边际效能或减少样本量。然而，如果

平台试验中每项治疗组与对照组比较所用的样本量较小，那么共同对照的优势可能会被多重性调整带来的“惩罚”抵消。当然，这也取决于平台试验中的治疗组数量，因为这会直接影响第一类错误调整。治疗组数量较多，且每项治疗组与对照组比较的样本量较小（尽管仍高于单独试验中的样本量），可能导致其边际效能低于在总体样本量相同的情况下采用单独对照所能达到的效能。这是因为调整带来的“惩罚”较大，同时样本量相对较小。在这种情况下，采用单独对照开展试验，而不是加入采用共同对照的平台试验，或许更为可取。因此，尤其是在可实现的总体样本量较小的情况下，例如罕见病或儿科研究中，而精简临床开发又可能至关重要时，采用共同对照的平台试验中的多重性调整会抵消其节省样本量的优势。因此，问题仍然是：仅仅采用共同对照是否足以成为要求进行多重性调整的理由——因为平台试验采用共同对照时，可能会因这一“惩罚”而失去其优势和益处——还是说，应由其他考量来决定是否要求对整个平台试验进行多重性控制。

当前监管指南

对于平台试验而言，监管指导文件中有关假设族或假设间依赖关系的理解，可能无法完全从传统临床试验直接推广到更复杂的情境。目前，关于多重性问题的指南，例如多重性指南（欧洲药品管理局，2017年）或多重终点指南（美国食品药品监督管理局，2017年），要求对确证性声明控制研究层面的第一类错误率——强调的是研究层面的控制。由于这些指南并未详尽讨论所有问题，而且显然并非针对复杂的平台试验撰写，因此尚不清楚对于有共同对照组和没有共同对照组的平台试验，有何具体要求。如果没有进一步澄清，可能会导致过于保守甚至无法实施的调整：原因可能是治疗组数量过多，也可能是试验开始时无法确定需要进行调整的试验组数量。针对复杂临床试验的欧洲具体指导文件（欧洲药品管理局负责人（临床试验促进与协调组），2019年；欧洲药品管理局，2022年）提到了多重治疗调整的要求，但未进一步说明何时需要进行调整，或在实践中如何实现。类似地，美国食品药品监督管理局关于复杂临床试验的指南（美国食品药品监督管理局，2020年）要求在适用情况下评估“第一类错误概率控制和检验效能”，但未说明何时适用。美国食品药品监督管理局关于主方案的指南没有讨论第一类错误，而是引用了适应性设计指南；该指南将第一类错误表述为试验中得出错误结论的概率（美国食品药品监督管理局，2022年）。2021年5月，美国食品药品监督管理局发布了关于新冠肺炎药物开发主方案的指南，并澄清，在平台试验或伞式试验中进行多重比较无需进行多重性调整。指南还承认，可能会出现多个错误发现（美国食品药品监督管理局，2021年）。尚不清楚针对新冠肺炎药物以外治疗的平台试验是否可以采用

相同的方法。尽管 Collignon 等（2020）并非官方监管指导文件，但其中分享了作者（其中数位现任或曾任欧洲药品管理局生物统计工作组成员）对确证性篮式、伞式和平台试验的统计与监管考量的个人观点。作者认为，使用共同对照组并不一定需要进行多重性控制，尽管相应的监管决策之间存在相关性。此外，作者还详细阐述了若干可能需要或不需要进行调整的情形。例如，若各项假设彼此无关，因为它们针对的是不同疗法，作者认为通常无需调整。然而，如果评估的是同一治疗的多个剂量，他们则认为需要进行调整。

检验族的定义

在规划研究时，通常会考虑进行多重性调整，以控制家族错误率。这自然会引出一个问题：应确定一个适当或合理的检验族，并对其进行控制。在定义检验族时，采用共同对照的平台试验有两种极端情况：（1）将每项治疗组与对照组比较分别定义为一个独立的族，此时无需进行多重性调整；（2）将平台上的所有治疗组与对照组比较定义为一个族，此时需要控制平台上所有试验组的多重性。Frane（2019）和 Hung 和 Wang（2009）强调，检验族的定义在一定程度上具有主观性，但并非完全任意。平台试验中的检验族可依据多种标准来定义，包括但不限于

- 由于使用共同对照，检验统计量之间存在依赖关系，
- 假设之间可能存在的依赖关系（例如，由于使用相同或相似的治疗），
- 可行性方面（例如持续性试验，以及罕见病或儿科试验等特殊试验设置）

或上述多个方面的相互作用。前两个方面侧重于依赖关系／相关性，而后者描述的是操作性／可行性标准。显然，这些方面并非纯粹属于统计学范畴，因此，确定检验族应当涉及跨职能领域的专家，包括但不限于统计学家、临床医生以及非临床同事。因此，“何时进行调整”这一问题最终基本上可归结为：如何界定一个族。因此，这实际上只是将决策“推迟”了。接下来，我们将进一步阐述上述标准。我们的目的不是提供具体的决策规则，而是强调在决定是否以及如何进行调整时，可以单独或结合考虑的不同方面。

统计依赖性

在平台试验中使用共同对照组会导致检验统计量之间存在依赖关系。共同对照组对错误率和检验效能的影响已在

本文中得到广泛讨论。Rothman（1990）和 Perneger（1998）等人认为，如果我们开展独立试验，就不需要进行多重性调整。从推断的角度来看，开展独立试验可视为与采用单独对照的平台试验相当。在后一种情况下，可以轻松采用两步随机化：先将受试者随机分配至子试验，再随机分配至该试验的治疗组或对照组，而不采用一步随机化。然而，在这两种情况下，子试验内的比较原则上都与独立试验内的比较相同。平台试验只会为各子试验增加一层随机化（因此，理论上允许在子试验之间进行随机比较）。不过，这不会直接影响子试验本身的I类错误，因此也就不需要对所有子试验进行多重性控制。正如我们所见，采用共同对照组的平台试验，其家族错误率甚至低于采用单独对照的平台试验。从这个角度来看，仅仅因为设置了共同对照组，就要求对平台试验中的所有试验组控制I类错误，似乎有悖直觉。

假设之间的依赖性

如果我们在平台试验中考察同一种药物的不同剂量，或准相同化合物（差异极小，但作用机制相同），那么还会引入另一种依赖关系，即假设之间的依赖关系，以及由此产生的决策之间的依赖关系：某个剂量或某种治疗是否被判定为有效。这种依赖关系需要结合试验背景，根据生物学、化学和临床考量加以讨论。同一种治疗的多个剂量会导致相关假设，因为某一剂量有效这一事件，取决于另一剂量有效这一事件。因此，在一项试验中检验同一种药物的两个剂量，会在实际意义上增加该药物的I类错误率，因为我们有多次机会（错误地）得出该药物有效的结论。对于作用方式相同或相似的治疗是否也会导致相关假设，或者是否应始终将同一申办方提供的治疗视为存在依赖关系（因为这会增加申办方获得一种（错误地）被判定为有效的治疗的机会），则要达成共识困难得多。

对于独立假设的判定，应根据上述跨职能讨论提供论证。例如，可以通过说明不同的作用机制或不同的生物学靶点来论证其在生物学上的独立性。但仅仅声称每项治疗组与对照组比较都会形成独立申报或标签声明，通常是不够的。研究问题的独立性需要得到充分论证。如果在正在进行的试验中，某种治疗被判定为优效，这是否会影响继续或停止另一种治疗的决策？如果能够清楚表明这些假设彼此独立，则很可能无需进行调整。

可行性方面

除了讨论适当的族定义所涉及的不同独立性方面之外，可行性也是需要考虑的重要方面之一。尤其是在

持续性试验中，对所有曾在某个时间点进入试验的治疗（既包括既往治疗组，也包括未来治疗组）进行调整可能过于保守，甚至无法实现；即使在其他治疗已停止招募数年之后也是如此。例如，I-SPY 2 是一项针对乳腺癌治疗的持续性平台试验，涉及超过20种治疗（包括正在进行的治疗和已经完成的治疗），时间跨度超过10年。然而，并非所有治疗都曾同时处于平台中——在 I-SPY 2 药物时间线中，同时存在的治疗组最多约为5个（I-SPY 2：Quantum Leap Health-care Collaborative, 2022年）。究竟应针对5种、20种治疗，还是预先定义的最大治疗数量进行调整，将对多重性调整产生巨大影响，进而影响加入平台的决策。这种调整不仅可能过于保守或难以实施，还可能引发伦理方面的质疑。有人可能会认为，如果不同治疗之间没有共同的对照数据，那么将族定义限制为当前同时招募、因而共享对照组数据的治疗，可能是合理的。然而，如上所述，是否需要调整不仅取决于共同对照组本身，还取决于其他依赖关系。如果这些非同期治疗采用的是同一种药物但剂量不同，则很可能需要进行调整。

此外，在罕见病、儿科环境或探索性试验等特殊情况下，还可以根据可行性和平台试验目标等其他方面讨论不同的要求。即使罕见病或儿科人群的平台试验可能需要进行多重性控制，也必须同时考虑可行性问题。通常，探索性试验对多重性控制的要求较宽松，因为此类试验的结局通常不会用于确证性声明。然而，在某些情况下，异常结果仍可能导致申报；若未进行调整，则可能引发对结果解读的担忧。

最后需要提醒的是，与采用单独对照相比，采用共同对照本身就能减少总样本量，而且在许多情况下，即使进行了多重性控制，总样本量仍会更小。若采用共同对照但不进行多重性控制，所获得的额外“收益”应与家族错误率增加的风险权衡。

总结与局限性

在我们的模拟中，我们仅考虑了这样一种简单情形：在所有治疗组停止招募后才进行主要分析，并且已知用于调整的治疗组总数。在我们的模拟中，如果有效治疗组是已经在试验中的治疗组之一，那么加入试验的时间不会影响检验效能。对于已经在平台上的治疗组，治疗组-对照组样本量不会受到第三个治疗组延迟加入的影响，并且由于已知用于调整的治疗组数量，多重性控制已确定。因此，我们仅考虑有效治疗组晚些加入灵活平台的情况，并允许在检验效能分析中增加该治疗组及其同期对照组的成对样本量。在实践中，这些简单假设通常并不成立。尤其是在永久性

平台试验中，试验开始时可能无法获得治疗组总数，而且人们不会等到所有治疗组都结束后再进行分析。等待分析会带来商业和伦理问题；另一方面，这样做可以避免对正在进行的研究造成揭盲或偏倚。在实践中，当新的治疗加入试验时，多重性调整会随之改变。文献中已经提出了多种适用于此类情况的调整方法，例如错误率在线控制（Tian 和 Ramdas，2021）。这些方法允许根据先前的假设检验进行调整，而不要求知道未来的假设检验。由于每次假设检验都取决于先前的拒绝结果，加入平台试验的时间决定了可用于检验的 alpha 剩余量。因此，我们在模拟中重点考察了传统调整方法。错误率在线控制因而不在本文讨论范围内。尽管采用共同对照时家族错误率较小，但与单独对照相比， k-家族错误率（$k \geq 2$）更高。如果不存在可接受的家族错误率多重性控制方案，则可考虑控制 k-家族错误率作为替代方案。这将确保采用共同对照获得的错误多重率不会高于采用单独对照获得的错误多重率。Lehmann 和 Romano（2005）以及 Keselman 等（2011）提出了控制k-家族错误率的方法。未来的研究可以评估 k-家族错误率控制对运行特征的影响。

我们的模拟旨在促进有关在设有共同对照的平行试验中是否需要进行多重性调整及其影响的讨论，从而为讨论是否普遍需要共同对照提供高层次的结果。除了平台试验在操作层面可能具有的优势以及采用共同对照的好处之外，在许多情况下，与采用单独对照或开展单独试验相比，多重性调整仍可降低总样本量。总体而言，目前不仅研究人员尚未达成共识，监管机构也尚未提供进一步指导。如上所述，仅凭是否采用共同对照并不能决定是否需要进行调整。因此，我们建议基于多个概念，对如何确定适当的族提供依据，整体目标是避免错误地得出药物有效的结论，并以有意义的方式考量其后果。对于每一项新的平台试验及其治疗组，都需要逐案展开讨论，并提供基于统计学、生物学和化学论据的理由。

附录

公式（1）的证明：设

$$Z_j^* := \frac{\bar{X^j} - \bar{X^{0j}}}{\sqrt{\frac{1}{n_j} + \frac{1}{n_0^j}}} \sim \mathcal{N}(0,1),$$

为灵活平台试验中治疗 j 与共同对照组比较的检验统计量。此外，我们有

$$I_j := \Big\{ i = 1, \ldots, n_0 : \text{ Patient } i \text{ in control arm, concu. randomized to arm } j \Big\},$$
$$n_0^j := \#I_j, j = 1, \ldots, m \text{ and } n_0^{j,j'} := \#\left(I_j \cup I_{j'}\right) \text{ for } j, j' = 1, \ldots, m.$$

对于 $\rho_{j,\ j'}$ 关于 j，$j' = 1$、.....、m 和 $j \neq j'$ 的检验统计量 Z_j^* 与 $Z_{j'}^*$ 之间的相关性，通过标准计算可得

$$\begin{aligned}\rho_{j,j'} = \ & \operatorname{Corr}(Z_j^*, Z_{j'}^*) = \frac{\operatorname{Cov}(Z_j^*, Z_{j'}^*)}{\sqrt{\operatorname{Var}(Z_j^*)\ \operatorname{Var}(Z_{j'}^*)}} = \operatorname{Cov}(Z_j^*, Z_{j'}^*) \\ & = \frac{1}{\sqrt{\frac{1}{n_j} + \frac{1}{n_0^j}}} \frac{1}{\sqrt{\frac{1}{n_{j'}} + \frac{1}{n_0^{j'}}}} \cdot \operatorname{Cov}(\bar{X}_j - \bar{X}^{0j}, \bar{X}_{j'} - \bar{X}^{0j'}).\end{aligned}$$

由于第二个因子可以改写为

$$
\begin{aligned}
&\mathrm{Cov}(\bar{X}_j - \bar{X}^{0j}, \bar{X}_{j'} - \bar{X}^{0j'}) \\
&= \frac{1}{2}\Big[\mathrm{Var}(\bar{X}_j - \bar{X}^{0j} - \bar{X}_{j'} - \bar{X}^{0j'}) - \mathrm{Var}(\bar{X}_j - \bar{X}^{0j}) - \mathrm{Var}(\bar{X}_{j'} - \bar{X}^{0j'})\Big] \\
&= \frac{1}{2}\Big[\mathrm{Var}(\bar{X}^{0j} + \bar{X}^{0j'}) - \mathrm{Var}(\bar{X}^{0j}) - \mathrm{Var}(\bar{X}^{0j'})\Big] \\
&= \frac{1}{2}\left[\mathrm{Var}\left(\frac{1}{\#I_j}\sum_{i\in I_j} X_i^0 + \frac{1}{\#I_{j'}}\sum_{i\in I_{j'}} X_i^0\right) - \frac{1}{n_0^j} - \frac{1}{n_0^{j'}}\right] \\
&= \frac{1}{2}\left[\mathrm{Var}\left(\frac{1}{n_0^j}\Big(\sum_{i\in I_j\setminus\{I_j\cap I_{j'}\}} X_i^0 + \sum_{i\in\{I_j\cap I_{j'}\}} X_i^0\Big)\right.\right. \\
&\qquad \left.\left. + \frac{1}{n_0^{j'}}\Big(\sum_{i\in I_{j'}\setminus\{I_j\cap I_{j'}\}} X_i^0 + \sum_{i\in\{I_j\cap I_{j'}\}} X_i^0\Big)\Big) - \frac{1}{n_0^j} - \frac{1}{n_0^{j'}}\right] \\
&= \frac{1}{2}\left[\frac{n_0^{j,j'} - n_0^{j'}}{(n_0^j)^2} + \frac{n_0^{j,j'} - n_0^j}{(n_0^{j'})^2} + \left(\frac{1}{n_0^j} + \frac{1}{n_0^{j'}}\right)^2 (n_0^j + n_0^{j'} - n_0^{j,j'}) - \frac{1}{n_0^j} - \frac{1}{n_0^{j'}}\right] \\
&= \frac{1}{2}\left[\frac{n_0^{j,j'} - n_0^{j'}}{(n_0^j)^2} + \frac{n_0^{j,j'} - n_0^j}{(n_0^{j'})^2} + \left(\frac{1}{(n_0^j)^2} + \frac{2}{n_0^j n_0^{j'}} + \frac{1}{(n_0^{j'})^2}\right)\cdot (n_0^j + n_0^{j'} - n_0^{j,j'})\right. \\
&\qquad \left. - \frac{1}{n_0^j} - \frac{1}{n_0^{j'}}\right] \\
&= \frac{1}{2}\left[\frac{n_0^{j,j'} - n_0^{j'}}{(n_0^j)^2} + \frac{n_0^{j,j'} - n_0^j}{(n_0^{j'})^2} + \frac{n_0^j + n_0^{j'} - n_0^{j,j'}}{(n_0^j)^2} + \frac{n_0^j + n_0^{j'} - n_0^{j,j'}}{(n_0^{j'})^2} + \frac{2(n_0^j + n_0^{j'} - n_0^{j,j'})}{n_0^j n_0^{j'}}\right. \\
&\qquad \left. - \frac{1}{n_0^j} - \frac{1}{n_0^{j'}}\right] \\
&= \frac{n_0^j + n_0^{j'} - n_0^{j,j'}}{n_0^j n_0^{j'}},
\end{aligned}
$$

我们总共收到

$$
\rho_{j,j'} = \frac{1}{\sqrt{\frac{1}{n_j} + \frac{1}{n_0^j}}\sqrt{\frac{1}{n_{j'}} + \frac{1}{n_0^{j'}}}}\frac{n_0^j + n_0^{j'} - n_0^{j,j'}}{n_0^j n_0^{j'}}.
$$

致谢

本研究部分满足了获得人类生物学博士学位的要求。本文所表达的观点仅代表作者个人意见，不应被理解为代表作者所隶属的监管机构或组织的立场，也不应被引用为这些机构或组织的立场。作者还要感谢 Hue Kästel 在准备附录时提供的帮助。

资金支持

EU-PEARL（EUPatient-cEntric clinicAl tRial pLatforms）项目获得创新药物倡议（IMI）2联合事业（JU）的资助，资助协议编号为853966。该联合事业获得欧盟地平线2020研究与创新计划、欧洲制药工业协会联合会、儿童肿瘤基金会、全球结核病药物开发联盟（非营利组织）以及Springworks Therapeutics Inc.的支持。本出版物仅反映作者的观点。IMI、欧盟、欧洲制药工业协会联合会及任何关联合作伙伴均不对本文所载信息可能被如何使用承担责任。

利益冲突

作者声明，他们不存在任何利益冲突。

补充材料

模拟代码见补充材料。

参考文献

Bai, X., Deng, Q., Liu, D.：平台试验中使用共享对照组的多重性问题。《生物制药统计学杂志》30(6), 1077–1090 (2020年)。doi:10.1080/10543406.2020.1821703

Charles W. Dunnett：一种用于将多个治疗组与对照组进行比较的多重比较程序。《美国统计学会杂志》50(272),1096–1121 (1955)。doi:10.1080/01621459.1955.10501294

Collignon,O., Gartner, C., Haidich, A.-B., James Hemmings, R., Hofner, B., Pétavy, F., Posch, M., Rantell, K., Roes, K., Schiel, A.: 关于确证性篮式、伞式和平台试验规划的当前统计考量与监管视角。《临床药理学与治疗学》107(5), 1059–1067 (2020). doi:10.1002/cpt.1804

Dmitrienko, A., Tamhane, A.C., Bretz, F.:《药物统计学中的多重检验问题》。Chapman and Hall/CRC，纽约 (2009) 欧洲药品管理局：临床试验中多重性问题指南草案。EMA/CHMP/44762/2017 (2017) 欧洲药品管理局：《复杂临床试验——问答》。EMA/298712/2022 (2022) Frane, A.V.: 对多重性调整的错误反对仍然是个问题。《现代应用统计方法杂志》18(2), 266–267 (2019). doi:10.22237/jmasm/1556669400 Freidlin, B., Korn, E.L., Gray, R., Martin, A.: 新药物的多臂临床试验：一些设计考量。《临床癌症研究》：美国癌症研究协会官方期刊14(14), 4368–4371 (2008). doi:10.1158/1078- 0432.CCR-08-0325 欧洲药品管理局负责人（临床试验促进与协调组）：《复杂临床试验启动与实施建议文件》(2019) Hochberg, Y., Tamhane, A.C.:《多重比较程序》。John Wiley & Sons, Inc., ??? (1987) Howard, D.R., Brown, J.M., Todd, S., Gregory, W.M.: 关于多臂试验中共享对照组的多重检验调整建议。《医学研究统计方法》27(5), 1513–1530 (2018). doi:10.1177/0962280216664759 Hung, H.M.J., Wang, S.-J.: 监管申报中的一些有争议的多重检验问题。《生物制药统计学杂志》19(1), 1–111241 (2009). doi:10.1080/10543400802541693 I-SPY 2：Quantum Leap Healthcare Collaborative：《I-SPY 2 药物时间线》(2022). https://www.ispytrials.org/results/past-agents 访问日期：2022年6月27日 James, N.D., Sydes, M.R., Clarke, N.W., Mason, M.D., Dearnaley, D.P., Anderson, J., Popert, R.J., Sanders, K., Morgan, R.C., Stansfeld, J., Dwyer, J., Masters, J., Parmar, M.K.B.: 晚期或转移性前列腺癌的全身治疗（STAMPede）：一项多臂多阶段随机对照试验。《BJU国际》103(4), 464–469 (2009). doi:10.1111/j.1464-410X.2008.08034.x Keselman, H.J., Miller, C.W., Holland, B.: 多项显著性检验：控制I类错误的新方法。《心理学方法》16(4), 420–431 (2011). doi:10.1037/a0025810 Lehmann, E.L., Romano, J.P.: 家族错误率的推广。《统计学年鉴》33(3), 1138–1154 (2005). 访问日期：2022-10-06

Meyer,E.L., Mesenbrink, P., Dunger-Baldauf, C., Fülle, H.-J., Glimm, E., Li, Y., Posch, M., König, F.：主方案临床试验设计的演变：系统性文献综述。《临床治疗学》42(7), 1330–1360 (2020年)。doi:10.1016/j.clinthera.2020.05.010

Molloy, S.F., White, I.R., Nunn, A.J., Hayes, R., Wang, D., Harrison, T.S.: 共享对照组的平行组多臂试验中的多重性调整：亟需明确的指导。《当代临床试验》113, 106656（2022年）。doi:10.1016/j.cct.2021.106656 Park, J.J.H., Siden, E., Zoratti, M.J., Dron, L., Harari, O., Singer, J., Lester, R.T., Thorlund, K., Mills, E.J.: 对篮式试验、伞式试验和平台试验的系统综述：主方案的全景分析。《Trials》20(1)，572（2019年）。doi:10.1186/s13063-019-3664-1 Parker, R.A., Weir, C.J.: 不同治疗的多组试验中不对多重检验进行调整：理由与依据。《临床试验》（伦敦，英格兰）17(5)，562–566（2020年）。doi:10.1177/1740774520941419 Perneger, T.V.: Bonferroni校正有什么问题？《英国医学杂志》（临床研究版）316(7139)，1236–1238（1998年）。doi:10.1136/bmj.316.7139.1236 Proschan, M.A., Waclawiw, M.A.: 临床试验中多重性调整的实用指南。《对照临床试验》21(6)，527–539（2000年）。doi:10.1016/s0197-2456(00)00106-9 Proschan, M.A., Follmann, D.A.: 在单项实验中与对照组进行多重比较和分别进行实验：为什么我们的看法不同？《美国统计学家》49(2)，144–149（1995年）。访问日期：2022-06-24 RECOVERY：Nuffield 人口健康系：RECOVERY：COVID-9疗法随机评估（2022年）。https://www.recoverytrial.net/ 访问日期：2022年6月27日 Renfro, L.A., Sargent, D.J.: 临床研究中的统计学争议：篮式试验、伞式试验及其他主方案：综述与实例。《肿瘤学年鉴：欧洲肿瘤内科学会官方期刊》28(1)，34–43（2017年）。doi:10.1093/annonc/mdw413 Rothman, K.J.: 多重比较无需进行调整。《流行病学》（马萨诸塞州剑桥）1(1)，43–46（1990年） Stallard, N., Todd, S., Parashar, D., Kimani, P.K., Renfro, L.A.: 关于主方案确证性临床试验中进行多重性调整的必要性。《肿瘤学年鉴：欧洲肿瘤内科学会官方期刊》30(4)，506–509（2019年）。doi:10.1093/annonc/mdz038 Tian, J., Ramdas, A.: 家族错误率的在线控制。《医学研究中的统计方法》30(4)，976–993（2021年）。doi:10.1177/0962280220983381 美国食品药品监督管理局：行业指南：临床试验中的多重终点。FDA-2019-D-4460（2017年）

美国食品药品监督管理局：行业指南：与 FDA 就药物和生物制品复杂创新试验设计进行沟通。FDA-2019-D-3679（2020年）美国食品药品监督管理局：行业指南：COVID-19：主方案——评估用于治疗或预防的药物和生物制品。FDA-2021-D-0409（2021年）美国食品药品监督管理局：行业指南：主方案：加速肿瘤学开发的高效临床试验设计策略。FDA-2018-D-3292（2022年）Wason, J.M.S., Stecher, L., Mander, A.P.: 多组试验中如何进行多重检验校正：是否有必要，又是否付诸实践？《Trials》15(1)，364（2014年）。doi:10.1186/1745-6215-15-364 Wason, J., Magirr, D., Law, M., Jaki, T.: 关于多臂多阶段试验的一些建议。《医学研究中的统计方法》25(2)，716–727（2016年）。doi:10.1177/0962280212465498