药物临床试验多重性问题指导原则

（征求意见稿）

2020 年 8 月

目 录

药物临床试验多重性问题指导原则

一、概述

临床试验中普遍存在多重性问题，它是指在一项完整的研究中，需要经过不止一次统计推断（多重检验）对研究结论做出决策的相关问题。例如，多个终点（如主要终点和关键次要终点)、多组间比较、多阶段整体决策（如出于有效性决策为目的的期中分析)、纵向数据的多个时间点分析、亚组分析、分层分析、同一模型不同参数组合或不同数据集的分析、敏感性分析等。对于确证性临床试验，将总 I 类错误率α（FWER）控制在合理水平是统计学的基本准则。上述多重性问题有的可以导致 FWER 膨胀，有的则不会。对于前者，需要采用恰当的决策策略和分析方法将 FWER 控制在合理水平，这一过程称为多重性调整；对于后者，则无需多重性调整。因此，在制订临床试验方案和统计分析计划时，采用恰当的决策策略和分析方法以控制 FWER 是非常重要的。

本指导原则主要阐述常见的多重性问题和相应的解决策略，介绍常用的多重性调整的统计方法，旨在为确证性药物临床试验中如何控制 FWER 提供指导意见，所讨论的一般原则也适用于其它类型的临床研究。

二、多重检验中的 I 类错误、总 I 类错误率和 II 类错误

（一）I 类错误和总 I 类错误率

I 类错误是指原假设（或称无效假设）正确但检验结果拒绝了原假设的错误，相当于把实际上无效的药物经统计推断得出有效结论的错误，其概率需控制在某一水平，该水平称为检验水准，或称显著性水准。对于多重检验中某一假设检验的检验水准称之为名义检验水准，又称局部检验水准，用α_i表示。

总 I 类错误率是指在同一试验所关注的多个假设检验中，至少一个真的原假设被拒绝的概率，而不论多次检验中哪个或哪些原假设为真。如此定义的 FWER 得到控制时，称为强控制 FWER。在所有原假设都为真的条件下至少一个真的原假设被拒绝的概率，如此定义的 FWER 得到控制时，称为弱控制 FWER。弱控制只能得出整体性结论，而不支持其中单个假设检验的结论，故在确证性临床试验中的应用意义不大。本指导原则仅限于强控制 FWER 的应用问题。

（二）II 类错误

对于确证性临床试验，在 I 类错误得到有效控制的前提下，II 类错误的风险也需要注意。II 类错误是指原假设不正确，但检验结果未能拒绝原假设的错误，相当于把实际上有效的药物经统计推断得出无效结论的错误，其概率用β表示，相应地$1-\beta$称为检验效能。对于需要调整的多重检验，由于控制 FWER 降低了多重检验中每个独立检验的α_i，相应地也降低了检验效能。因此，当涉及多重性调整时，制定研究计

划应考虑控制 FWER 对检验效能的影响，例如通过适当增加样本量以保证足够的检验效能。

三、常见的多重性问题

临床试验中常见的多重性问题一般体现在多个终点、多组间比较、亚组分析、期中分析、纵向数据不同时间点的分析等方面。

（一）多个终点

1. 主要终点

主要终点是指与临床试验所关注的主要问题（主要目的）直接相关的、能够提供最具临床意义和令人信服的证据的终点，常用于主分析、样本量估计和评价试验是否达到主要目的。确证性临床试验中，单一主要终点较为常见，但某些情况下会涉及多个主要终点，对于多个主要终点的研究，通常有两类统计假设策略，即多个主要终点均要求显著和多个主要终点中至少有一个显著。

（1）多个主要终点均要求显著。即要求所有主要终点均显著时才认为研究药物有效（此种情况常称为共同主要终点）。例如，在一项治疗慢性梗阻性肺病（COPD）的 III 期临床试验中设置两个单独的主要疗效终点，第 1 秒用力呼气量（FEV1）和患者报告症状评分，决策规定两个主要终点均显著才可推断研究药物有效。在此情况下，不会导致 I 类错误膨胀，因为这种策略没有机会选择对研究药物最有利的某

个或某几个主要终点，只有一种可能得出药物有效的结论（即两个原假设都被拒绝）。但是，这会增大 II 类错误和降低检验效能。检验效能降低的程度与主要终点的个数和主要终点之间的相关性有关，个数越多、相关性越弱，检验效能降低的幅度越大。因此，对于多个主要终点均要求显著的情形，无需多重性调整，但应留意对检验效能的影响。

（2）多个主要终点中要求至少一个终点显著。即至少一个主要终点显著时就认为研究药物有效。例如，某一确证性临床试验旨在验证一种治疗烧伤伤口的药物，设置两个单独的主要终点：伤口闭合率和瘢痕形成，临床试验方案规定只要其中一个终点显著，或两个终点都显著，就可认为该药物整体临床有效。此种情况下需要多重性调整，因为得出药物有效的结论包括以下三种可能的情形：①伤口闭合率显著而瘢痕形成不显著；②伤口闭合率不显著而瘢痕形成显著；③伤口闭合率和瘢痕形成都显著。由于多个主要终点中至少有一个终点显著的组合策略不尽相同，多重性调整策略应视具体的统计假设而定。

2. 次要终点

临床试验的次要终点通常有多个，多数情况下它们用于提供药物对主要疗效终点疗效的支持作用。但在某种情况下，有些次要终点可能用于支持药品说明书声称的获益，一般被称为关键次要终点。此时，应将关键次要终点与主要终点共

同纳入 I 类错误控制。只有主要终点的检验认为整体显著后，才考虑关键次要终点的检验。

3. 复合终点

复合终点是指将多个临床相关结局合并为一个单一变量，如表示心血管事件的复合终点，只要发生心肌梗死、心力衰竭、冠心病猝死等其中的任一事件将被视为终点事件发生；或者将若干症状和体征的评分通过一定的方法合并为一个单一变量，如评价类风湿关节炎的 ACR20 量表。如果将某一复合终点作为单一主要终点，将不涉及多重性问题。但是，如果同时将复合终点中某一组成部分（如某一事件或构成量表的某一维度）用于支持药品说明书声称的获益，应将其定位于主要或关键次要终点，再根据上述定位对所涉及的主要或次要终点的多重性问题予以考虑。

4. 探索性终点

探索性终点可以是预先设定、也可以是非预先设定（例如数据驱动）的终点，一般包括预期发生频率很低而无法显示治疗效果的临床重要事件，或由于其它原因被认为不太可能显示效果但被纳入探索性假设的终点，其结果可能有助于设计未来新的临床试验。此类终点无需考虑多重性调整。

5. 安全性终点

如果安全性终点（事件）是确证性策略的一部分，即用于支持药品说明书声称的获益，则应事先确定，并将其与主

要疗效终点所涉及的多重性问题做同样处理。此时，安全性评价和有效性评价均应控制各自的 FWER。需注意，在临床试验的实践中，由于安全性事件具有很大的不确定性，有时难以事先规定主要安全性假设，因此，对于多个安全性终点（通常是严重的不良反应）的确证性策略可能会基于事后的多重性调整策略，此时应充分说明其合理性，并与监管机构达成共识。

（二）多组间比较

临床研究中多组间的比较颇为常见，如三臂设计、剂量-反应关系研究、联合用药和复方药的评价等。

1. 三臂设计

三臂设计多用于非劣效试验，安排的三个组分别是试验组、阳性对照组和安慰剂组。此时，统计假设应该考虑三种情形：①试验组与安慰剂组比较的优效性；②阳性对照组与安慰剂组比较的优效性；③试验组与阳性对照组比较的非劣效性（和可能的优效性）。对于这一多重性问题，如果三个假设检验的结果均显著才可认为试验药物有效，无需多重性调整；或者，基于一个比较弱的研究假设，即只要满足①即可认为试验药物有效；如果采用固定顺序策略，如检验顺序为①→②→③，此时也无需多重性调整。但需要注意，后者这种基于较弱的研究假设需得到监管机构的认可才可实施。其它的三臂设计如果不是遵循这一多重性检验策略，且不满

足所有检验结果均显著的话，需根据情况考虑是否需要多重性调整。

2. 剂量-反应关系

剂量-反应关系研究对于找到安全有效的治疗剂量或剂量范围至关重要。剂量探索的方法和目的在 II 期和 III 期试验中有所不同。

在 II 期试验中，剂量探索研究多用于估计剂量-反应关系，通常基于统计模型证明临床效应与剂量增加总体呈正相关关系，不需要对不同剂量组和安慰剂组之间进行比较，故无需控制 FWER。但是，如果剂量反应研究作为确证性策略的一部分，就需要控制 FWER。

在确证性临床试验中，剂量探索通常是基于假设检验进行多剂量组间的比较，旨在选择和确证试验药物在特定患者人群中推荐使用的一个或多个剂量水平，此时必须控制 FWER，如采用基于 p 值的多重检验，或基于参数方法的多重检验（如 Dunnett 检验）。

3. 联合用药和复方药

联合用药是指治疗用药同时使用两种或以上的药物，复方药是指治疗用药由两种或以上的药物组合而成。联合用药或复方药临床试验的目的主要是验证联合用药的获益-风险是否优于其中的单药，或复方药的获益-风险是否优于其组分药。

以两个单药的联合用药为例，试验设计至少会设置三个组，即联合用药组、单药 A 组和单药 B 组，后两组为阳性对照组。如果再增加一个安慰剂组，就是一个 2×2 的析因设计。无论是三组的设计还是四组的析因设计，其统计检验以推断联合用药组是否优于其它各组为主，这将不会导致 I 类错误膨胀，因为只有所有假设均显著的情况下方可证明联合治疗的疗效。

（三）纵向数据不同时间点的分析

纵向数据，即基于时间点的重复测量数据，是临床试验常见的数据类型。此类数据与时间点相关的分析分两种情况，一种是在不同时间点进行组间比较；另一种是比较处理组内不同时间点的效应。

假设研究设计只有一个主要终点且只涉及两个处理组（多于一个主要终点或多于两个处理组的多重性问题上文已述及），如果主要终点评价被定义为在多个时间点中的某一个时间点（如最后一个访视点）进行处理组间的比较，其它时间点的组间比较被视为次要终点评价，则不涉及多重性调整；如果主要终点评价被定义为在不止一个时间点进行处理组间的比较，若其所有相关时间点的组间比较达到显著才认为有效，就无需多重性调整，否则，就需要多重性调整。

对于比较处理组内不同时间点效应的情形，如果目的是通过时间点之间的比较确证最佳时间点的效应，即当时间效

应成为确证性策略的一部分时，就需要多重性调整；否则，无需多重性调整。

如果希望回避纵向数据的多重性调整问题，一种可能的解决方案是将不同时间点的效应转换为折线下的面积，例如治疗后不同时间点的疼痛 VAS 评分可以转化为折线下面积以代表治疗后总的疼痛评分，即把多个变量转化为一个变量，但相应地，在这种转换之后，每个时间点的组间比较就无法实施了。另一种可能的解决方案是对重复测量数据用单个模型分析，如重复测量方差分析或混合效应模型。

（四）亚组分析

亚组分析通常用于说明试验药物在某一特定亚组人群中的疗效、或者各亚组之间疗效的一致性。如果特定亚组的分析用于支持药品说明书声称的获益，则需要综合考虑总人群和亚组人群的多重性问题，同时还要注意保证亚组有足够的检验效能。反之，如果亚组分析不用于支持药品说明书声称的获益，则无需多重性调整。

（五）期中分析

针对有效性和/或无效性进行监查的期中分析，因为在研究过程中需要进行多次决策，所以 FWER 的控制显得尤为重要，多重性调整的策略和方法也复杂多样。在制定临床试验方案时，应仔细考虑并预先设定恰当的多重性调整策略和相应的统计方法。

（六）复杂设计

对于用于确证性目的的篮式设计、伞式设计、平台设计等涵盖多疾病领域、多种药物、跨研究的复杂设计，由于同时开展多个分题研究，涉及多重决策的问题。但由于这些分题研究多是独立的研究且回答特定的临床问题，如适用疾病、目标人群等，故一般无需多重性调整。

但是，对于复杂设计分题研究的目标人群有较大重叠时，以及对于多个分题研究使用同一个对照组时，是否需要多重性调整，应视具体情况而定。此时，建议申办方与监管机构进行充分沟通。

四、常见的多重性调整的策略与方法

针对临床试验中普遍存在的多重性问题，所采用的多重性调整的策略与方法取决于试验的目的、设计、统计假设及其分析方法。申办方需在试验设计时对选用的多重性调整的策略与方法进行必要的评估，并在临床试验方案和统计分析计划中详述。

多重性调整的策略与方法可以从决策策略、调整方法和分析方法三个层面考虑。

（一）多重性问题的决策策略

临床试验的研究结论主要依据综合所有试验数据分析结果所做的推断，也是一个从局部决策到整体决策的过程。多重性问题的决策策略可分为平行策略（或称单步法）和序

贯策略（或称多步法）。除了从局部决策到整体决策的过程外，还有分阶段的整体决策，例如，出于有效性决策为目的的期中分析。

1. 平行策略

平行策略是指所包含的各个假设检验相互独立，平行进行，与检验顺序无关，就像一种并联关系，每个假设检验的推断结果不依赖于其它假设检验的推断结果。

图 1 是平行策略的示意图，H_{0i} 为第 i 个原假设（i=1，2，…，m），m 为假设检验的个数；H_A 为整体备择假设，即整个研究结论对应的假设，H_{Ai} 为第 i 个备择假设；α 为 FWER 水平，α_i 为第 i 个名义检验水准。平行策略有以下三种情形：

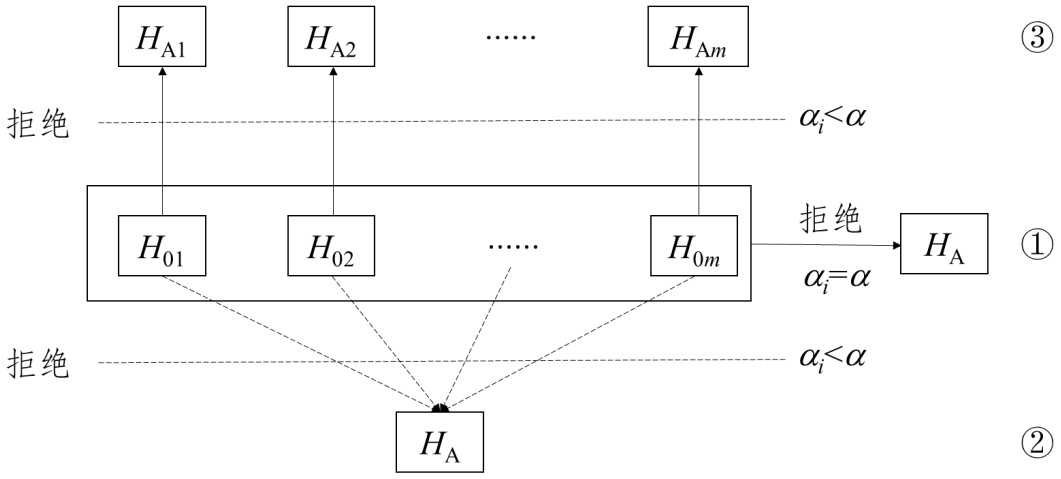


图 1　多重性问题的平行策略示意图

①如果所有假设检验均显著才被认为是阳性结论（如三臂设计的非劣效试验，多个主要终点等），即试验药物有效（图 1 右侧的备择假设 H_A 成立），则无需多重性调整，每个检验的名义水准与 FWER 水平相同（$\alpha_i=\alpha$）。

②如果其中至少一个假设检验结果显著就被认为是阳性结论但不包含①（图 1 下方的备择假设 H_A 成立），则需要多重性调整（$\alpha_i<\alpha$）。例如设有 3 个主要终点（O_1，O_2，O_3）的试验，如果采用 Bonferroni 法，每个终点的名义检验水准可以相同也可以不同，但其和为 FWER 水平，即 $\alpha_1+\alpha_2+\alpha_3=\alpha$。

③图 1 上部的 H_{A1}、H_{A2} 等代表局部决策，即在整体结论为阳性的前提下，并基于多重性调整（$\alpha_i<\alpha$）的检验结果，可进一步对某个或某几个备择假设是否成立做出独立的推断。仍以设有 3 个主要终点的试验为例，在采用策略②得出试验药物有效的整体结论后，局部决策有 6 种可能的组合，一个终点的假设检验结果显著有 3 种，3 个终点中任意两个终点的假设检验结果显著有 3 种。类似的例子还可见于剂量探索研究的确证性临床试验中，如设置 2 个或 3 个剂量组和一个安慰剂对照组，采用策略②，只要其中一个剂量组与安慰剂组比较显著就可整体以推断试验药物有效，并在此基础上进一步做出局部决策，即哪一个或几个剂量有效。

对于②和③情形下的多重性调整可采用 Bonferroni 法或 Šidák 法。

2. 序贯策略

序贯策略是指按一定顺序对原假设进行检验，直到满足相关条件而停止检验，就像一种串联关系，根据设定条件前

一个假设检验的结果将决定是否进行后续的假设检验。序贯顺序分为固定顺序和非固定顺序两种方式，如下所述。

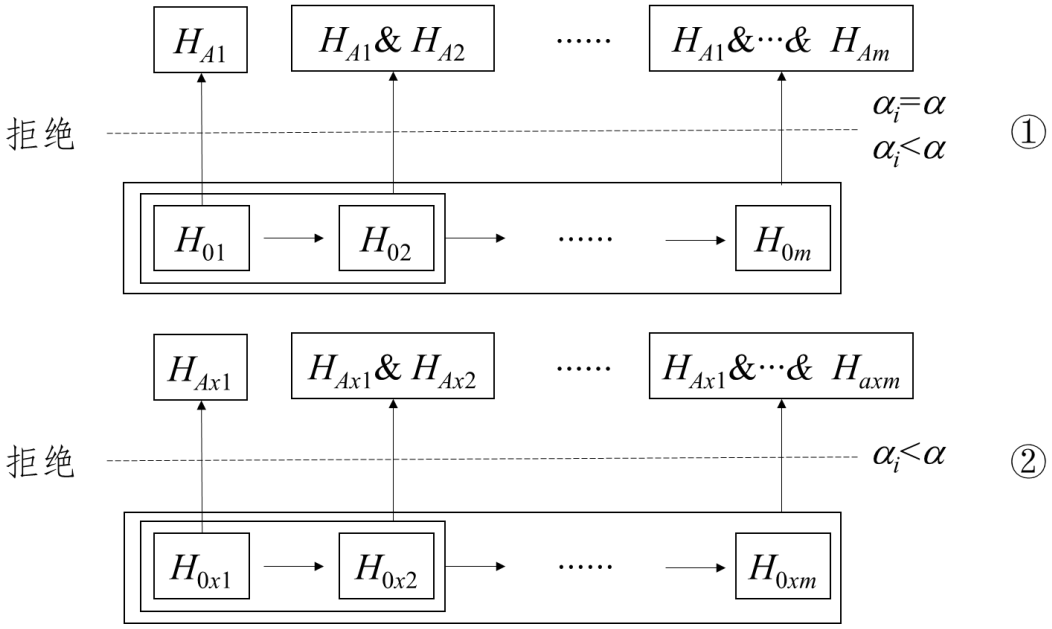


图 2　多重性问题的序贯策略示意图（②以向下法为例）

①固定顺序策略：见图 2 上半部分，假设检验的顺序需事先确定，分需要和不需要进行多重性调整两种情况。以不需要调整（$\alpha_i=\alpha$）为例，每一个假设检验的名义水准与 FWER 水平相同。假设检验以既定顺序依次进行，直到某一个假设检验不拒绝原假设（不显著）为止，而最终的推断结论为该假设前面的阳性检验结果均被接受。例如，按顺序有 3 个原假设分别是 H_{01}、H_{02} 和 H_{03}，若第 1 和第 2 个假设检验都在 0.05 水平拒绝了原假设，但第 3 个假设检验未能拒绝原假设 H_{03}，则备择假设 H_{A1} 和 H_{A2} 都成立，而 H_{A3} 不成立。

当固定顺序策略需要做多重性调整时，可参见回退法一节。

固定顺序策略也适用于对假设检验集合进行的检验，如守门法，即将所有原假设分成若干集合，并预先确定各集合假设检验顺序，按顺序对各集合进行检验。对于复杂的固定顺序策略，可借助图示方法直观展现决策规则。

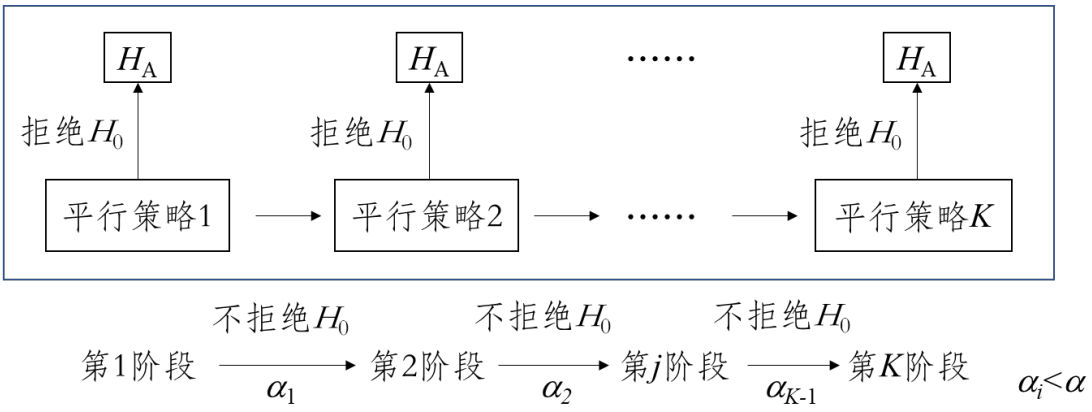
②非固定顺序策略：见图 2 下半部分，以向下法为例，假设检验的顺序按事先规定以检验统计量由大到小（p 值由小到大）排序，图 2 中下标“x”表示顺序位次在试验设计阶段无法确定，只能在事后求出检验统计量后才能确定，例如 H_{0x2} 的含义是在所有检验统计量中第 2 大的假设检验所对应的原假设。该策略需要做多重性调整（$\alpha_i<\alpha$），每个假设有各自的名义检验水准。假设检验以规定的顺序依次进行，直到某一个假设检验不拒绝原假设（不显著）为止，而最终的推断结论为该假设前面的阳性检验结果均被接受。

序贯策略中假设检验的顺序以及相应的多重性调整方法的不同对整体结论的影响也不同，这一点在设计阶段尤其要注意。序贯策略的检验效能通常优于平行策略，但其置信区间的计算较为复杂甚至难以估计。

3. 分阶段的整体决策策略

分阶段的整体决策策略是指将整体决策按照时间顺序分阶段进行，其典型代表是出于有效性为目的的期中分析，如图 3 所示。每个阶段都进行一次整体决策，确定试验因有效或无效提前终止还是继续。每一阶段的整体决策可以采用

多重性问题决策策略中的平行策略或序贯策略。多阶段决策需要多重性调整，即每个阶段都会消耗一定的α，各阶段的名义检验水准α_i 可以相同，也可以不同，视采用的α消耗策略而定。需要注意，在每个阶段的整体决策过程中，如果涉及到局部决策的多重性调整，则该阶段的名义检验水准α_i就是该阶段的总α水平。



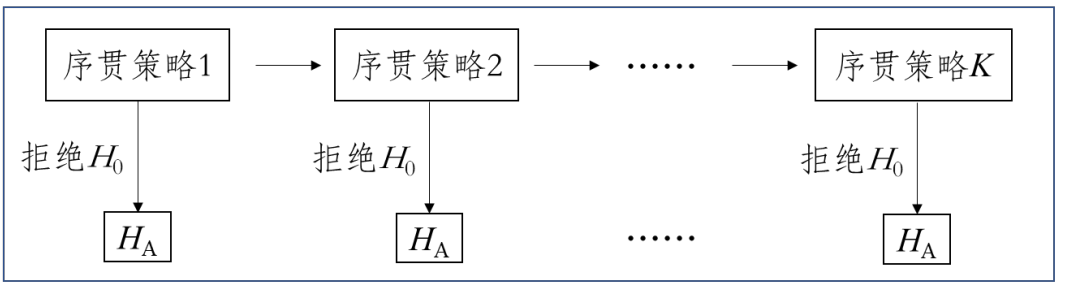


图 3　多重性问题的分阶段整体决策示意图

（二）多重性调整方法

多重性调整方法实质上是通过调整整体决策中每一个独立假设检验的名义检验水准 α_i 以达到控制 FWER 的目的。名义检验水准的确定方法可以根据多重性问题的决策策略选择。

1. 平行策略的多重性调整方法

（1）Bonferroni 法。Bonferroni 法的基本思想是各个独立检验的名义水准之和等于 FWER 水平α，即

$$\alpha_1+\alpha_2+\ldots+\alpha_i\ldots+\alpha_m=\alpha$$

各名义水准可以相同（$\alpha_i=\alpha/m$），也可以不同，后者往往在各个检验假设的优先顺序时使用。例如，某临床试验设有 3 个主要终点，需要进行 3 次假设检验，设定α=0.05。如果 3 个主要终点的优先顺序相同，则每个检验的α_i 相同，均为 0.0167（=0.05/3），则每个假设检验的 p 值小于 0.0167 才被认为该检验显著；如果 3 个主要终点的优先顺序不同，如设置α_1、α_2 和α_3 分别为 0.030、0.015 和 0.005，则每个假设检验的 p 值小于所对应的α_i 才被认为该检验显著。该法较为保守，各检验统计量正相关程度越高越保守。尽管如此，由于该法简单，其应用最为广泛，而且其思想为许多方法所借鉴，如后述的 Holm 法、Hochberg 法、回退法等。

（2）前瞻性 α 分配法。前瞻性 α 分配法（PAAS）与 Bonferroni 法思想相近，可理解为各个假设检验的互余的乘积等于 FWER 水平α的互余，即

$$(1-\alpha_1)(1-\alpha_2)\ldots(1-\alpha_i)\ldots(1-\alpha_m)=(1-\alpha)$$

各 α_i 可以相同也可以不同，若相同，则可根据 Šidák 法求得

$$\alpha_i=1-(1-\alpha)^{1/m}$$

例如，一个有 3 个终点的临床试验，其中两个终点被指定分配了 α_i 值，α_1=0.02、α_2=0.025，若设 α 为 0.05，则根据上式

有 0.98×0.975×(1-α_3)=0.95，求得第 3 个终点的 α_3 为 0.0057。如果采用 Bonferroni 法，则第 3 个终点的 α 值为 0.005。可见 PAAS 法分配的 α_3 要高于 Bonferroi 法。如果 3 个原假设的 α_i 等权重分配，则基于 Šidák 法求得 α_i 为 0.01695，略高于 Bonferroni 法分配的 0.0167。因此，PAAS 法较 Bonferroni 法可略微增加检验效能。

2. 序贯策略的多重性调整方法

（1）Holm 法。Holm 法是一种基于 Bonferroni 法的检验统计量逐步减小（p 值逐步增大）的多重调整方法，又称向下法。该法首先计算出各检验假设的 p 值后，将各 p 值按从小到大排序，记为 $p_1<p_2<\ldots<p_m$，其相对应的原假设为 H_{01}，H_{02}，…H_{0m}，然后按照 p 值从小到大顺序依次与相对应的 α_i 进行比较，依次检验 H_{0i}，$1\leq i\leq m$。第一步从最小的 p 值开始，检验原假设 H_{01}，如果 $p_1>\alpha_1$（$=\alpha/m$），则不拒绝原假设 H_{01}，并停止检验所有剩余的假设；如果 $p_1<\alpha_1$，则拒绝 H_{01}，H_{A1} 成立，进入下一个检验。第 2 个检验的名义水准 $\alpha_2=\alpha/(m-1)$，将该检验的 p 值与 α_2 比较，若 $p_2>\alpha_2$，则停止检验余下的假设；否则，H_{A2} 成立，并进入下一个检验。更一般地，在检验第 i 个原假设 H_{0i} 时，如果 $p_i>\alpha_i$（$=\alpha/(m-i+1)$），则停止检验并接受 H_{0k}，…，H_{0m}；否则，拒绝 H_{0i}（接受 H_{Ai}），并进入下一个检验。

（2）Hochberg 法。Hochberg 法是一种基于 Bonferroni

法的检验统计量逐步增大（p 值逐步减小）的多重调整方法，又称向上法。该法首先计算出各检验假设的 p 值，将各 p 值按从小到大排序，记为 $p_1<p_2<\ldots<p_m$，然后按照 p 值从大到小顺序依次与相对应的 α_i 进行比较。第一步从最大的 p 值开始，检验原假设 H_{0m}，如果 $p_m<\alpha$，则拒绝所有原假设，并停止检验，所有的备择假设 H_{Ai} 成立；否则不拒绝 H_{0m}，进入下一步检验。第 2 个检验的名义水准 $\alpha_{m-1}=\alpha/2$，将该检验的 p 值与 α_{m-1} 比较，若 $p_{m-1}<\alpha/2$，则停止检验余下的假设，除 H_{Am} 外，其余的备择假设均成立；否则，不拒绝 $H_{0(m-1)}$，并进入下一个检验。第 3 个检验的名义水准 $\alpha_{m-2}=\alpha/3$ 将该检验的 p 值与 α_{m-2} 比较，若 $p_{m-2}<\alpha/3$，则停止检验余下的假设，除 H_{Am} 和 $H_{A(m-1)}$外，其余的备择假设均成立；否则，不拒绝 $H_{0(m-2)}$，并进入下一个检验。余类推。需要注意，Hochberg 法在满足终点变量独立或检验统计量正相关条件才能实现 FWER 强控制。

（3）回退法。回退法是固定顺序策略中的一种多重性调整方法。对于固定顺序策略不做多重性调整的情况，由于固定顺序的限制，一旦前一个检验结果不显著，后续的其它检验将终止，这种策略可能失去发现有意义的研究假设的机会。例如，一项设有 2 个主要终点的临床试验，采用固定顺序策略（$O_1\rightarrow O_2$），α 为 0.05。如果两个终点的检验结果分别是 $p_1=0.062$，$p_2=0.005$，那么决策的结论是两个终点均无

效，因为第 1 个检验的结果不显著，未能进行到第 2 个检验，丧失了发现对第 2 个终点获益的机会。回退法需事先根据固定顺序策略对各假设排序，并采用 Bonferroni 法确定每个检验的 α_i，然后依顺序进行检验。该法首先在 α_1 水平检验 H_{01}，如果拒绝 H_{01}，则在 $\alpha_1+\alpha_2$ 水平检验 H_{02}；如果不拒绝 H_{01}，则在 α_2 水平检验 H_{02}，余类推。该法具有两个特点，一是在前一个原假设未被拒绝时，仍可继续后续的检验，例如上例，采用回退法，对应 O_1 和 O_2 的名义水准分别是 0.04 和 0.01，最终的决策结论为试验药物对第 2 个主要终点 O_2 有显著获益；二是如果前一个检验显著，其对应的 α_i 可以叠加到下一个检验的名义水准，体现了 α_i 的传递思想。例如，假设对应 O_1 和 O_2 的名义水准分别是 0.04 和 0.01，如果对 O_1 的假设检验显著（α_1=0.04），则对 O_2 的检验水准为 0.05（=0.01+0.04），即把前一次检验显著的名义水准传递给了下一次检验。对于固定顺序策略是否采用多重性调整各有利弊，需权衡之。

3. 期中分析常见的 α 分割方法

期中分析较经典的 α 分割方法有 Pocock 法、O’Brien-Fleming 法和 Haybittle-Peto 法。这三种分割方法的一个共同前提是每一次期中分析的间隔和样本量相同，只是每次假设检验 α_i 的分配有不同侧重。更为灵活的 α 分割方法则是 α 消耗函数，如 Lan-DeMets α 消耗函数，该方法是上述经典方法的扩展，它不要求期中分析间隔样本量相等，在设

定期中分析时间点上更为灵活。例如，一项评价免疫靶点抑制剂抗肿瘤药物的确证性临床试验，主要评价指标为全因死亡，拟进行一次期中分析，可基于有效性早期终止试验。考虑到免疫靶点抑制剂起效时间可能存在延迟，因此计划在研究相对较晚的时间点，即观察到 75%的死亡事件时，开展期中分析。采用近似 O'Brien Fleming 边界的 Lan-Demets α 消耗函数，且要求双侧 FWER 控制在 0.05，则期中分析和最终分析的双侧名义检验水准分别为 0.019 和 0.044。

（三）多重性分析方法

对于需要解决的多重性问题，多数是基于具体的统计检验方法结合多重性调整方法来实现的。例如，对于不同数据类型的多个终点（如定量、定性、生存时间），组间比较会用到不同的统计分析方法（如协方差分析、M-H χ^2 检验、Kaplan-Meier 检验），与此同时，还要依靠多个终点的多重性调整方法（如 Bonferroni 法等）来确定每个假设检验的检验水准 α_i，然后才能做出决策结论。

对于单一终点变量、同一研究阶段的多组比较，有些统计分析方法是在整体检验的基础上解决多重比较的问题，其根本思想是两两比较所涉及的标准误是整体检验的标准误，由此达到控制 FWER 的目的。例如，定量结局变量基于方差分析的两两比较有 LSD 法、SNK（Student-Neuman-Keuls）法、Scheffe 法、Tukey 法、Levy 法、Ryan 法、Duncan 法，

等等，多组与参照组的比较有 Dunnett 法、Dunnett-SNK 法、Dunnett-Levy 法等；定性结局变量的多重比较可通过变量变换（如反正弦变换）成为定量变量，然后采用上述定量变量的分析方法；生存时间结局变量基于 Kaplan-Meier 法的 log rank 检验(Mantel-Cox 法)、Breslow 法(扩展 Wilcoxon 法)、Tarone-Ware 法等。上述方法可通过专业统计软件实现。对于在整体检验的基础上无法实现多重比较的统计分析方法，就需要采用局部检验（两两比较）结合α分配的方法（如 Bonferroni 法等)。

多变量的参数方法（如多元方差分析）是解决多重性问题的手段之一，特别是对于多终点的情况，但是此类方法一则要求满足多元正态分布，二则分析结果的解释往往不直观，限制了其应用。

重复抽样（如 bootstrap 法和 permutation 法）也是解决多重性问题的手段之一，此类方法的优点是在控制 FWER 的同时还能保证较高的检验效能；其不足之处在于它所基于的经验分布难以验证从而导致估计的准确性不足，此外它更依赖于大样本。因此，该类方法在临床试验中少有实践，需慎重使用。

由于解决多重性问题的统计分析方法众多，每种方法都有其优势与不足，申办方需要在临床试验方案或统计分析计划中事先规定针对多重性问题所采用的统计分析方法。

（四）多重性问题的基本解决思路

临床试验的多重性问题较为普遍而且复杂，解决这一问题的基本思路如图 4 所示。首先，根据研究目的和试验方案，梳理出可能的多重性问题。其次，判断哪些多重性问题需要多重性调整，哪些不需要。之后，进入多重性调整过程。先判断是做一次整体决策还是分阶段做若干次整体决策（如基于有效性决策的期中分析），对某一个整体决策而言可采用平行策略、序贯策略或平行+序贯策略，最后根据所选策略确定每一个检验假设（局部决策）所对应的统计分析方法和名义水准 α_i 的分配策略（如需要）。

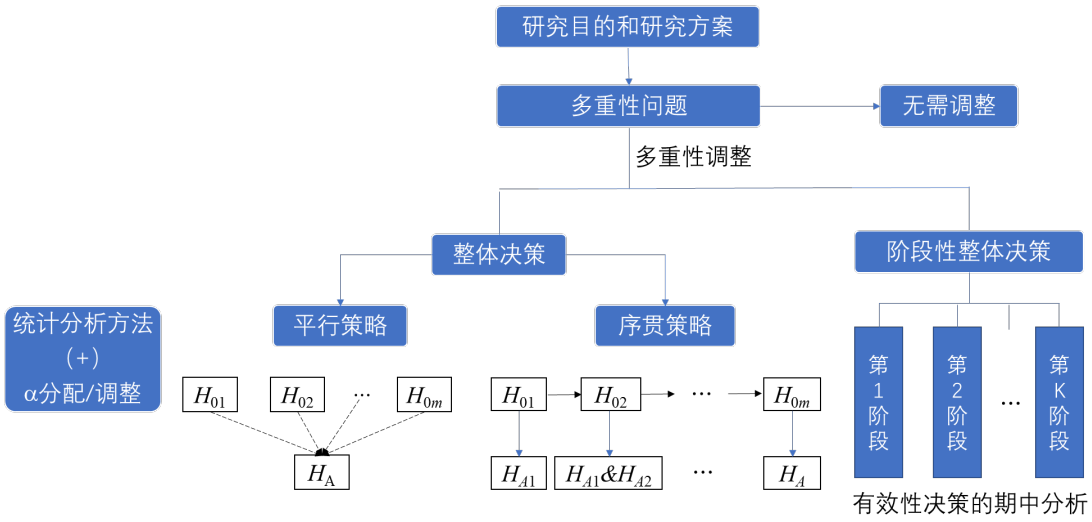


图 4　多重性问题的基本解决思路

五、其它考虑

（一）不需要调整的多重性问题

不需要调整的多重性问题包括但不限于以下情形（均不包含有效性的期中分析）：

1. 针对单一主要终点的非劣效试验的标准三臂设计，所有假设检验结果均显著才被视为有效；

2. 针对单一主要终点，研究假设为试验药物的疗效至少非劣于阳性对照药，即检验假设为固定顺序，第一步验证试验药物的疗效非劣于阳性对照药的假设，第二步验证试验药物的疗效优于阳性对照药的假设（在第一步假设被拒绝后），每一步的检验水准与 FWER 水平相同；

3. 针对多个主要终点，当且仅当所有终点的假设检验结果均显著才被视为有效；

4. 针对多个均不以说明书声称的获益为目的的次要终点；

5. 有效性和安全性评价应分别独立控制 FWER，两者间无需调整；

6. 对于篮式设计、伞式设计、平台设计等跨研究的复杂设计，如果分题研究多是独立的研究且回答各自的临床问题，如适用疾病、目标人群等；

7. 在统计分析过程中，对同一主要终点指标，可能会对不同的分析数据集进行分析，只要事先定义以哪个分析数据集为主要结论依据；

8. 采用不同的统计模型或同一模型采用不同的参数设置，只要事先定义主分析模型；

9. 根据不同的假设进行敏感性分析，例如采用不同的缺

失数据估计方法填补后的分析，对离群值采用不同处理后的分析等。

（二）多重性检验的参数估计问题

多重性调整的假设检验方法众多，有的方法较为复杂，可能难以做出相应的区间估计，此时应该考虑采用较为简单但是相对保守的方法进行区间估计，例如采用 Bonferroni 方法调整置信区间。

多重性调整还有可能带来点估计的选择性偏倚。例如，在含有多个剂量组的确证性临床试验中，如果多重性问题的决策策略选择了在药物说明书中标示与安慰剂差异最大的剂量组的效应量，则有可能高估药物的疗效。类似的选择性偏倚也会因亚组的选择而产生。因此，有必要评估多重性调整可能带来的选择性偏倚。

（三）与监管机构的沟通

在临床试验方案和统计分析计划中应事先明确多重性问题和多重性调整的策略和方法。对于复杂的多重性问题，是否需要多重性调整以及如何调整，现有的策略和方法可能面临挑战，因此鼓励申办方在确证性临床试验设计阶段积极与监管机构沟通，以求双方能够达成共识。在试验过程中，如果因为更改多重性调整策略和方法而使临床试验方案做出重大调整，应与监管机构充分沟通，在征得同意的情况下对方案进行修改和备案。

六、参考文献

1. CDE. 非劣效设计临床试验指导原则
2. CDE. 临床试验数据监查委员会指导原则（征求意见稿）
3. CDE. 药物临床试验适应性设计指导原则（征求意见稿）
4. CDE. 药物临床试验的富集策略与设计指导原则（征求意见稿）
5. CDE. 药物临床试验亚组分析的指导原则（征求意见稿）
6. ICH E9（临床试验的统计学指导原则）
7. ICH E8（临床研究的一般注意事项）
8. ICH E17（多地区临床试验计划与设计总体原则）
9. 钱俊，陈平雁. Bootstrap和Permutation方法在样本率多重比较中的应用. 中国医院统计，2008；15（1）：43-45.
10. 钱俊，陈平雁. 多个样本率的多重比较. 中国卫生统计，2008；25（2）：206-212.
11. 钱俊，陈平雁. 样本率多重比较方法的模拟研究. 中国卫生统计，2009；26（2）：131-134.
12. Bretz F, Tamhane AC, Pinheiro J, et al. Multiple Testing in Dose-Response Problem, Chapter 3 of Multiplicity Testing Problem in Pharmaceutical Statistics. CRC Press, 2010.
13. Chen J, Luo JF, Liu K, et al. On power and sample size computation for multiple testing procedures. Computational Statistics and Data Analysis, 2011; 55: 110-122.

14. Collignon O, Christian Gartner C, Haidich AB, et al. Current statistical considerations and regulatory perspectives on the planning of confirmatory basket umbrella and platform trial. Clinical Pharmacology & Therapeutics, 2020; doi:10.1002/cpt.1804.
15. Dmitrienko A, Tamhane AC, Bretz F, et al. Multiple Testing Methodology, Chapter 2 of Multiplicity Testing Problem in Pharmaceutical Statistics. CRC Press, 2010.
16. Dmitrienko A, Tamhane AC, Bretz F, et al. Gatekeeping Procedures in Clinical Trials, Chapter 5 of Multiplicity Testing Problem in Pharmaceutical Statistics. CRC Press, 2010.
17. EMA. Guidance on Multiplicity Issues in Clinical Trials.
18. FDA. Multiple Endpoints in Clinical Trials –Guidance for the Industry.
19. Hochberg Y, Tamhane A. Multiplicity Comparison Procedure. New York: Wiley, 1987.
20. Huque MF, Rohmel J. Multiplicity Problem in Clinical Trials, Chapter 1 of Multiplicity Testing Problem in Pharmaceutical Statistics. CRC Press, 2010.
21. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika, 1983; 70: 659-663.

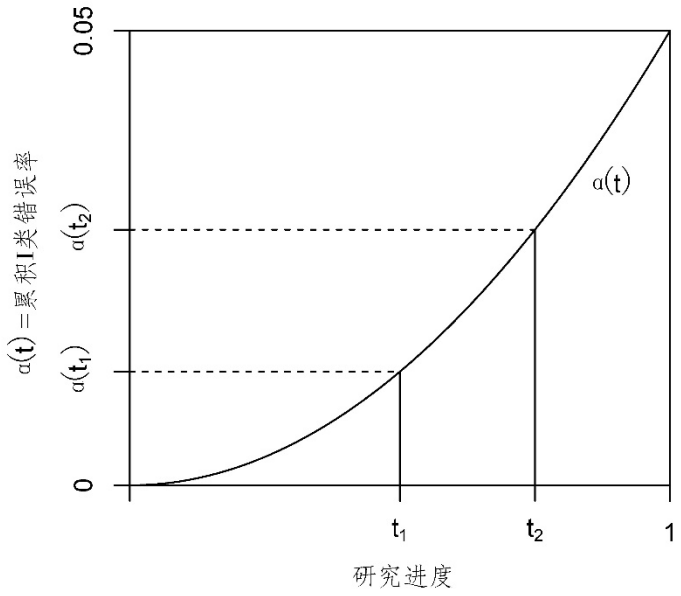
22. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics, 1979; 35: 549-556.
23. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observations of each patient, I. Introduction and design. British Journal of cancer, 1976; 34: 585-612.
24. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biornetrika, 1997; 64:191-199.
25. Sen. Some remark on Simes-type multiple tests of significance. Journal of statistical Planning and Inference, 1991; 82:139-145.
26. Wang DL, Li YH, Wang X, et al. Overview of multiple testing methodology and recent development in clinical trials. Contemporary Clinical Trials, 2015; 45: 13-20.

附录 1：词汇表

I 类错误（Type I Error）：指原假设（或称无效假设）正确但检验结果拒绝了原假设的错误，相当于把实际上无效的药物经统计推断得出有效结论的错误，其概率需控制在某一水平，该水平称为检验水准，或称显著性水准，习惯用α表示。

II 类错误（Type II Error）：指原假设不正确，但检验结果未能拒绝原假设的错误，相当于把实际上有效的药物经统计推断得出无效结论的错误。

***α*消耗函数（*α* Spending Function）：**当某个临床研究分若干阶段进行整体决策时（如基于有效性或无效性所做的期中分析），每个阶段都要消耗一定的α，随着研究进展，研究所完成的比例（如1/3、1/2、60%等）与累积的I类错误率呈现某种函数关系，如下图所示。



Bonferroni 法（Bonferroni Method）：Bonferroni 法的基本

思想是各个独立检验的名义水准之和等于 FWER 水平α，即

$$\alpha_1+\alpha_2+\ldots+\alpha_i\ldots+\alpha_m=\alpha$$（m 是独立检验的个数）

各名义水准可以相同（$\alpha_i=\alpha/m$），也可以不同，后者往往在各个检验假设的优先顺序时使用。

多重性问题（Multiplicity Issues）：指在一项完整的研究中，需要经过不止一次统计推断（多重检验）对研究结论做出决策的相关问题。

多重性调整（Multiplicity Adjustment）：采用恰当的决策策略和分析方法将FWER控制在合理水平的过程。

复合终点（Composite Endpoint）：是指将多个临床相关结局合并为一个单一变量，如表示心血管事件的复合终点，只要发生心肌梗死、心力衰竭、冠心病猝死等其中的任一事件将被视为终点事件发生；或者将若干症状和体征的评分通过一定的方法合并为一个单一变量，如评价类风湿关节炎的ACR20量表。

关键次要终点（Key Secondary Endpoint）:次要终点指标中用于支持药品说明书声称的获益的指标，其通常与次要研究目的联系在一起。

联合用药（Drug Combination）：指治疗用药至少使用了两种或以上的药物。

复方药（Compound Medicine）：指治疗用药由两种或以上的药物组合而成。

名义检验水准（Nominal Level）：对于多重检验中某一假设检验的检验水准称之为名义检验水准，又称局部检验水准，用α_i表示。

平行策略（Parallel Strategy）：又称单步法，是指所包含的各个假设检验相互独立，平行进行，与检验顺序无关，就像一种并联关系，每个假设检验的推断结果不依赖于其它假设检验的推断结果。

序贯策略（Sequential Strategy）：又称多步法，是指按一定顺序对原假设进行检验，直到满足相关条件而停止检验，就像一种串联关系，前一个假设检验的结果根据设定条件将决定是否进行后续的假设检验。

总 I 类错误率（Familywise Error Rate，FWER）：是指在同一试验所关注的多个假设检验中，至少一个真的原假设被拒绝的概率，而不论多次检验中哪个或哪些原假设为真。

主要终点（Primary Endpoint）：是指与临床试验所关注的主要问题（主要目的）直接相关的、能够提供最具临床意义和令人信服的证据的终点，常用于主要分析、样本量估计和评价试验是否达到主要目的。

618

附录 2：中英文对照表

中文	英文
α 分配	α Allocation
α 消耗	α Spending
α 消耗函数	α Spending Function
Bonferroni 法	Bonferroni Method
I 类错误	Type I Error
II 类错误	Type II Error
成组序贯分析	Group Sequential Analysis
单步法	Single-step Procedures
多步法	Multi-step Procedures
多个终点	Multiple Endpoints
多重性	Multiplicity
多重性调整	Multiplicity Adjustment
多重性问题	Multiplicity Issue
分题研究	Substudies
固定顺序检验法	Fixed-sequential Procedure
关键次要终点	Key Secondary Endpoint
回退法	Fallback Method
剂量-反应关系	Dose-response Relationship
假设检验	Hypothesis Test
检验效能	Power

中文	英文
篮式设计	Basket Design
联合用药	Drug Combination
名义水准	Nominal Level
偏倚	Bias
平行策略	Parallel Strategy
平台设计	Platform Design
前瞻性 α 分配法	Prospective Alpha Allocation Scheme, PAAS
伞式设计	Umbrella Design
守门法	Gatekeeping
显著性水准	Significance Level
序贯策略	Sequential Strategy
序贯设计	Sequential Design
主要终点	Primary Endpoint
总 I 类错误率	Familywise Error Rate, FWER
纵向数据	Longitudinal Data

619