药物临床试验多重性问题指导原则

（试行）

2020 年 12 月

目 录

药物临床试验多重性问题指导原则（试行）

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（如以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、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（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，需要采用恰当的策略与方法将 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，采用恰当的策略与方法控制 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决策策略，介绍常用的多重性调整方法和多重性分析方法，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，用α表示；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临床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。不论多次假设检验中哪个或哪些原假设为真，都能将 FWER 控制在α水平，称为强控制 FWER；在所有原假设都为真的条件下，将 FWER 控制在α水平，称为弱控制 FWER。弱控制 FWER 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。本指导原则所描述的“控制 FWER”均指强控制 FWER。

（二）II 类错误

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。对于需要调整的多重检验，由于控制 FWER 降低了多重检验中单个假设检验的 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，常用于主要分析、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，通常有两类研究假设，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，在一项治疗慢性阻塞性肺病的确证性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 FWER 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。

（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。此种情况下会导致 FWER 膨胀，因为得出药物有效的结论包括以下三种可能的组合：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不尽相同，是否会导致 FWER 膨胀应视具体的研究假设而定。

2. 次要终点

临床试验的次要终点通常有多个，多数情况下它们提供对主要终点的支持作用。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 FWER 控制。只有主要终点的假设检验认为整体显著后，才考虑关键次要终点的假设检验。

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难以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。此类终点不涉及多重性问题。

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，则应事先确定并考虑多重性问题。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，研究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。对于上述多重性问题，如果三个假设检验均显著才可认为试验药物有效，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（需得到监管机构的认可才可实施），或者采用固定顺序法，如假设检验顺序为①→②→③，此时不会导致 FWER 膨胀。其它的三臂设计如果不是遵循上述多重检验策略，且不满足所有假设检验均显著的话，需根据情况考虑是否会导致 FWER 膨胀。

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探索性试验和确证性试验中有所不同。

在探索性试验中，用剂量-反应关系进行剂量探索研究时，是否需要控制 FWER 由申办方自行决定。在确证性临床试验中，为了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，必须控制 FWER。

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，其假设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 FWER 膨胀，因为只有所有假设检验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。

以只有一个主要终点且只涉及两个处理组的研究设计为例，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，则不涉及多重性问题；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，则不会导致 FWER 膨胀，否则会导致膨胀。

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，就需要考虑多重性问题，否则无需考虑。

对于多于一个主要终点或多于两个处理组且涉及到纵向数据不同时间点分析的研究设计，其多重性问题更加复杂，需要综合考虑。

如果希望回避纵向数据的多重性问题，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目标亚组人群中的疗效、或者各亚组之间疗效的一致性。如果目标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的样本量有足够的检验效能。反之，如果亚组分析不用于支持药品说明书声称的获益，则无需考虑多重性问题。

（五）期中分析

针对有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，多重性问题复杂多样，所以控制 FWER 显得尤为重要。在制定临床试验方案时，应仔细考虑并预先设定恰当控制 FWER 的策略和方法。

（六）复杂设计

对于以确证性为目的的篮式设计、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，可能涉及多重性问题。但是，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，故一般不会导致 FWER 膨胀。

对于复杂设计分题研究的目标人群有较大重叠时，或者对于多个分题研究使用同一个对照组时，是否会导致 FWER 膨胀，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

针对临床试验中可能导致FWER膨胀的多重性问题，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、研究假设及其检验方法。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，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和序贯策略。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。根据研究目的和试验方案梳理出可能的多重性问题，可采用某一种策略或者多种策略组合，再根据所选策略或策略组合确定每一个检验假设所对应的统计分析方法和名义检验水准α_i的分配策略（如需要）。

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。

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。

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事先确定的顺序分阶段进行，其典型代表是以有效性为目的的期中分析。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。

（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 以达到将 FWER 控制在α水平的目的。名义检验水准 α_i 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假设检验的名义检验水准α_i之和等于α，即

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

各名义检验水准α_i可以相同（$\alpha_i=\alpha/m$），也可以不同，后者

往往在各个假设检验的重要性不同时使用。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 个主要终点的重要性相同，则每个假设检验的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 才被认为有显著性；如果 3 个主要终点的重要性不同，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才被认为有显著性。

（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名义检验水准α_i的互余的乘积等于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 $0.98\times0.975\times(1-\alpha_3)=0.95$，求得第 3 个终点的 α_3为 0.0057。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。需要注意，PAAS 法在满足多重检验呈独立或正相关时才能实现控制 FWER。

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。该法首

先计算出各假设检验的 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}，$\ldots H_{0m}$，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 $P_1\leq\alpha_1$，则拒绝 H_{01}，H_{A1} 成立，进入下一步假设检验。第 2 个假设检验的 $\alpha_2=\alpha/(m\text{-}1)$，将该假设检验的 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，并进入下一步假设检验。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_{0i}，$\ldots$，H_{0m}；否则，拒绝 H_{0i}（接受 H_{Ai}），并进入下一步假设检验；以此类推。

（2）Hochberg 法。Hochberg 法是一种基于 Simes 法的检验统计量逐步增大（P 值逐步减小）的多重调整方法。该法首先计算出各假设检验的 P 值，将各 P 值按从大到小排序，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_{01}，如果 $P_1\leq\alpha_1$（$=\alpha$），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_{01}，进入下一步假设检验。第 2 个假设检验的 $\alpha_2=\alpha/2$，将该假设检验的 P 值与 α_2 比较，若 $P_2\leq\alpha/2$，则停止检验余下的假设，除 H_{A1} 外，其余的备择假设均成立；否则，不拒绝 H_{02}，并进入下一步假设检

验。更一般地，在检验第 i 个原假设 H_{0i} 时，如果 $P_i \leq \alpha_i (=\alpha/\mathrm{i})$，则停止余下的检验，拒绝 H_{0i}，…，H_{0m}；如果 $P_i > \alpha_i$，则不拒绝 $H_{0\mathrm{i}}$ 并进入下一步假设检验；以此类推。需要注意，Hochberg 法在满足多重检验呈独立或正相关时才能实现控制 FWER。

（3）固定顺序法。固定顺序法是指按预先定义的顺序进行假设检验，每个假设检验的名义检验水准 α_i 与 α 相同，只有在上一个假设检验拒绝原假设时才进行到下一个假设检验，直到某一个假设检验不拒绝原假设为止，而最终的推断结论为该假设检验前面的显著性结论均被接受。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 α 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。

（4）回退法。回退法需事先根据固定顺序法对各假设检验排序，并确定每个假设检验的名义检验水准 α_i，然后依顺序进行假设检验。该法首先在 α_1 水平检验 H_{01}，如果不拒绝 H_{01}，则在 α_2 水平检验 H_{02}；如果拒绝 H_{01}，则在 $\alpha_1+\alpha_2$ 水平检验 H_{02}，余类推。例如，一项设有 2 个主要终点（O_1 和 O_2）的临床试验，采用回退法，对应 O_1 和 O_2 的名义检验水准分别是 $\alpha_1=0.04$ 和 $\alpha_2=0.01$，如果假设检验的 P 值分别是 $P_1=0.062$，$P_2=0.005$，则最终的决策结论为试验药物在 O_2 上有显著获益（$P_1=0.062>\alpha_1$，$P_2=0.005<\alpha_2$）；如果假设检验的 P 值分别

是 P_1=0.032，P_2=0.015，则最终的决策结论为试验药物在 O_1 和 O_2 上均有显著获益（P_1=0.032＜α_1，P_2=0.015＜α_1+α_2）。

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日历时间或累积数据占比相同，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，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。

当临床试验的多重性问题较为复杂时，可组合使用多种策略的多重性调整方法。需要注意的是，将多个多重性调整方法进行简单组合未必能控制 FWER。因此，在复杂情况下组合使用多个多重性调整方法时，为了确保能够控制 FWER，可考虑采用守门法或图示法等。

（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分析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、Mantel-Haenszel 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假设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假设检验的标准误。例如，定量结局变量基于方差分析的两两比较有 LSD 法、SNK 法等，多组与参照组的比较有 Dunnett 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）等。需注意的是，有些方法不一定能控制 FWER。对于在整体假设检验的基础上无法实现多重比较的统计分析方法，就需要采用局部假设检验（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

是要求满足多元正态分布，二是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，建议事先与监管机构充分沟通。

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。

五、其它考虑

（一）不需要多重性调整的情况

不需要多重性调整的情况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多组间比较（如非劣效试验的标准三臂设计），当所有假设检验均显著才被视为有效时;

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，当按固定顺序进行假设检验时，即第一步验证试验药物的疗效非劣于阳性对照药的假设，第一步原假设 H_0 被拒绝后，第二步验证试验药物的疗效优于阳性对照药的假设;

3. 针对多个主要终点，当且仅当所有终点的假设检验均显著时才被视为有效时；

4. 针对多个次要终点，当均不会用于在药品说明书中声称获益时；

5. 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；

6. 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；

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

8. 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）多重检验的参数估计问题

应根据多重性调整方法对相应的置信区间进行估计。多重性调整方法众多，有的方法较为简单但相对保守，易于进行区间估计，例如采用 Bonferroni 方法调整置信区间；有的方法较为复杂，可能难以做出相应的区间估计。

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。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，应与监管机构及时沟通。

六、参考文献

[1] 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.

[2] Alosh M, Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. Statistics in Medicine, 2014; 33 (4): 693-713.

[3] 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.

[4] Bretz F, Maurer W, Brannath W, et.al. A graphical approach to sequentially rejective multiple test procedures. Statistics in Medicine, 2009; 28 (4): 586-604.

[5] 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 (1): 110-122.

[6] Collignon O, 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; 107 (5): 1059-1067.

[7] Dmitrienko A, Tamhane AC, Bretz F, et al. Multiple Testing Methodology, Chapter 2 of Multiplicity Testing Problem in Pharmaceutical Statistics. CRC Press, 2010.

[8] 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.

[9] Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 1955; 50 (272): 1096-1121.

[10] European Medicines Agency. Guidance on Multiplicity Issues in Clinical Trials.

[11] Freidlin B, Korn EL, Gray R, et.al. Multi-arm clinical trials of new agents: some design considerations. Clinical Cancer Research, 2008; 14 (14): 4368-4371.

[12] Hochberg Y, Tamhane A. Multiplicity Comparison Procedure. New York: Wiley, 1987.

[13] Howard DR, Brown JM, Todd S, et.al. Recommendations on multiple testing adjustment in multi-arm trials with a shared control group. Statistical Methods in Medical Research, 2018; 27 (5): 1513-1530.

[14] Huque MF, Rohmel J. Multiplicity Problem in Clinical Trials, Chapter 1 of Multiplicity Testing Problem in Pharmaceutical Statistics. CRC Press, 2010.

[15] International Conference on Harmonization (ICH). E9 guideline “Statistical Principles for Clinical Trials”.

[16] International Conference on Harmonization (ICH). E8 guideline “General Considerations for Clinical Trials”.

[17] International Conference on Harmonization (ICH). E17 guideline “General Principles for Planning And Design Of Multi-Regional Clinical Trials”.

[18] Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika, 1983; 70 (3) :659-663.

[19] O’Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics, 1979; 35 (3): 549-556.

[20] 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 (6): 585-612.

[21] Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika, 1977; 64 (2): 191-199.

[22] Sen PK. Some remark on Simes-type multiple tests of significance. Journal of statistical Planning and Inference, 1999; 82 (1-2): 139-145.

[23] U.S. Food and Drug Administration. Multiple Endpoints in Clinical Trials – Guidance for the Industry.

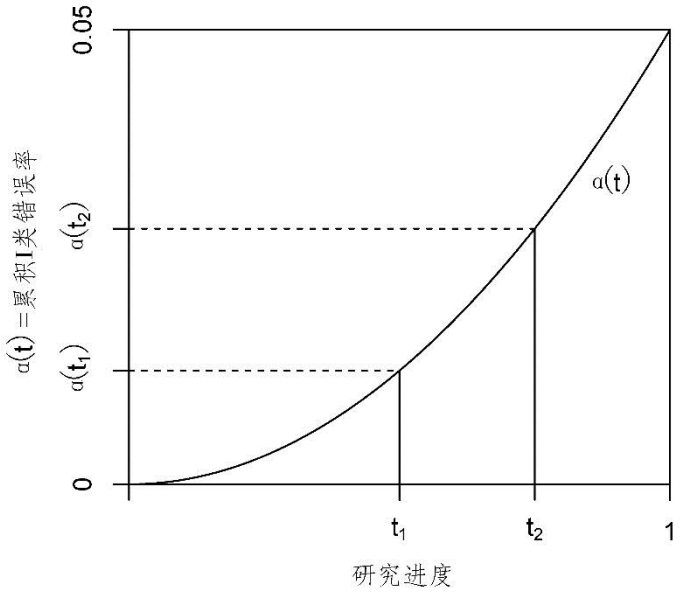
[24] Wang DL, Li YH, Wang X, et al. Overview of multiple testing methodology and recent development in clinical trials. Contemporary Clinical Trials, 2015; 45 (Pt A): 13-20.

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，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、3/5等）与累积的I类错误率呈现某种函数关系，如下图所示。



多重性问题（Multiplicity Issues）： 指在一项完整的临床研

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）：采用恰当的策略与方法将总I类错误率控制在合理水平的过程。

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。

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。

总 I 类错误率（Familywise Error Rate，FWER）: 是指在同一临床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。其应控制在合理水平。

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。

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
I 类错误	Type I Error
II 类错误	Type II Error
多重性	Multiplicity
多重性调整	Multiplicity Adjustment
多重性问题	Multiplicity Issue
多个终点	Multiple Endpoints
分题研究	Substudies
关键次要终点	Key Secondary Endpoint
回退法	Fallback Method
剂量-反应关系	Dose-response Relationship
名义检验水准	Nominal Level
前瞻性 α 分配法	Prospective Alpha Allocation Scheme, PAAS
守门法	Gatekeeping Procedure
图示法	Graphical Approach
显著性水准	Significance Level
总 I 类错误率	Familywise Error Rate, FWER