国际人用药品注册技术要求协调会

ICH三方协调指导原则

临 床 试 验 的 统 计 学 原 则

E9

现行第四阶段版本

1998年2月5日

该指导原则由相应的 ICH 专家工作组制定，按照 ICH 进程，已通过药品监管机构讨论。在 ICH 进程第四阶段，最终草案被推荐给欧盟、日本和美国的管理机构采纳。

E9
文件历史

首次编码	历史	日期	新编码 **2005 年 11 月**
E9	在第二阶段由指导委员会批准并公开征求意见。	1997 年 1 月 16 日	E9

现行第四阶段版本

E9	由指导委员会批准的第四阶段，并建议三个 ICH 监管机构采纳。	1998 年 1 月 5 日	E9

临床试验的统计学原则

ICH三方协调指导原则

本指导原则经1998年2月5日的ICH 指导委员会会议上确认已达到ICH进程第四阶段，推荐ICH三方成员国的药品监管机构采纳本指导原则。

目录

临床试验的统计学原则

1.引言

1.1 背景与目的

医药产品的有效性和安全性需由临床试验来论证。所采用的临床试验需遵循 ICH 在 1996 年 5 月 1 日通过的“良好临床实践（GCP）：综合指南”（ICH E6）。ICH E6 已阐明统计学在临床试验设计和分析中不可或缺的作用。由于统计学研究在临床试验领域的不断发展，加之临床研究在药物审批流程及一般医疗保健中的重要作用，因此，有必要制订一份关于临床试验统计学问题的简明文件。本指南旨在协调在欧洲、日本和美国提交上市申请的临床试验所应用的统计学方法的原则。

作为起点，本指南使用了欧盟专利医药产品委员会（CPMP）在题为《用于申请医药产品上市许可的临床试验生物统计学方法》（1994 年 12 月）指南的意见，并参照了日本厚生省的《临床研究中的统计分析指南》（1992 年 3 月）和美国食品药品监督管理局的《新药申请中临床与统计部分的格式与内容指南》（1998 年 7 月）。其他 ICH 指南也包含一些与统计学原则和方法有关的主题，特别是下面所列的指南。本指南的各个部分会对包含相关内容的特定指南进行标注。

E1A：人群暴露程度对评价临床安全性的影响

E2A：临床安全性数据管理：快速报告的定义与标准

E2B：临床安全性数据管理：个例安全报告传输数据元素

E2C：临床安全性数据管理：上市药品的定期安全性更新报告

E3：临床研究报告的结构与内容

E4：支持药品注册的剂量反应信息

E5：国外临床数据可接受性的种族因素

E6：良好临床实践：综合指南

E7：特殊人群的支持性研究：老年医学

E8：临床试验的一般考虑

E10：临床试验中对照组的选择

M1：用于监管目的的医学术语标准化

M3：用于实施药物人体临床试验的非临床安全性研究

本指南旨在为申办方在整体临床研发背景下，对研究产品临床试验的设计、实施、分析和评价提供指导。本指南也将会帮助科学专家准备上市申请总结报告或者评价主要来自研发后期的临床试验的有效性和安全性证据。

1.2 范围与方向

本指南的重点是统计学原则，并不涉及具体统计步骤或方法的使用。确保这些原则得到正确实施的具体程序性步骤是申办方的职责。本指南对不同临床试验之间的数据整合亦作了讨论，但并不作为重点。其他 ICH 指南涵盖了与数据管

理及临床试验监查活动有关的原则和程序，此处不再赘述。

本指南对很多科学学科的人士都是有意义的。然而，正如 ICH E6 所述，我们假定所有与临床试验有关的统计工作的实际职责由训练有素且经验丰富的统计师承担。试验统计师（见词汇表）在与其他临床试验专家合作时，其作用和职责是确保在支持药物研发的临床试验中恰当地应用统计学原则。因此，试验统计师应同时具备足够的教育/训练和经验以贯彻本指南所阐明的原则。

对于每一个用于上市申请的临床试验，有关设计、实施和拟采用的统计分析的主要特征等重要细节需在研究方案中阐明。对方案中步骤的遵循程度和主要分析预先计划的程度，都将决定试验最终结果和结论的可信度。方案及后续修订应获得包括试验统计师在内的责任人员的批准。试验统计师应恰当使用技术术语，保证方案以及任何修订都能清楚准确地涵盖所有相关的统计问题。

本指南所述的原则主要与研发后期实施的临床试验有关，其中很多是有效性的确证性试验。除有效性外，确证性试验也可把安全性指标（如不良事件、临床实验室指标或心电图测量）、药效学或药代动力学指标（如确证性的生物等效性试验）作为主要指标。其次，有些确证性结果可能来源于不同试验的整合数据，本指南有些原则适用于这种情况。最后，虽然药物研发早期本质上以探索性临床试验为主，但统

计学原则也与这些临床试验有关。因此，本指南应尽可能地应用于临床研发的各个阶段。

本指南所描述的很多原则致力于最小化偏倚（见词汇表）和最大化精度。这里的术语“偏倚”是指与临床试验设计、实施、分析和结果解释有关的任何因素所导致的处理效应（见词汇表）的估计值与真实值偏离的系统性趋势。应尽可能地识别偏倚的潜在来源，以便采取措施限制这些偏倚。偏倚的存在可能严重削弱从临床试验中得出正确结论的能力。

有些偏倚源于试验设计，例如，在处理分配过程中将风险较低的受试者系统地分配到其中一个处理组。其他偏倚源于临床试验的实施和分析。例如，违背方案且基于对受试者结局的认识从分析中排除受试者是偏倚的可能来源，这可能影响处理效应的准确估计。偏倚常在不知不觉中发生，且难以直接测量，因而评价试验结果和主要结论的稳健性是重要的。稳健性是一个概念，是指整体结论对数据的各种限制、假设和数据分析方法的敏感性。稳健性意味着，当基于另一假设或分析方法进行分析时，试验的处理效应和主要结论不会受到实质性的影响。在对处理效应和处理间比较的不确定性的统计测量进行解释时，应考虑偏倚对 P 值、置信区间或推断的潜在影响。

由于临床试验设计和分析的主要方法基于频率派统计方法，因此在讨论假设检验和/或置信区间时，本指南主要使

用频率派方法（见词汇表）。这并不意味着其它方法不可取，如果理由充分且所得结论足够稳健，则贝叶斯方法（见词汇表）及其他方法亦可考虑。

2.总体临床研发的考虑

2.1 试验背景

2.1.1 研发计划

新药临床研发过程的广义目标是发现药物是否在某一剂量范围和用法上能够显示出既安全又有效，且其风险获益关系能够被接受。可能从药物获益的特定对象以及特定的适应症也需要被定义。

满足这些目标通常需要一系列循序渐进的临床试验，每一个临床试验有其特定目的（见 ICH E8），应该在一个或一系列临床计划中明确，这些计划应具有适当的决策点和随知识累积而进行修订的灵活性。上市申请应清晰地描述这些计划的主要内容和每个试验的作用。对整个试验项目证据的解释和评价需要综合单个试验的证据（见第 7.2 章节），为此应确保试验在一些特征上采用通用标准，如医学术语词典、主要测量的定义与时点、方案违背的处理，等等。当医学问题通过一个以上的试验来回答时，统计汇总、综述或 meta 分析（见词汇表）可能会有用。应尽量在计划中考虑到这一点，以便清晰地确定相关的试验，并且预先指定必要的设计方面的共同特征。应该在该计划中阐述可能会涉及整体计划中若

干试验的其他主要统计学问题（如果有的话）。

2.1.2 确证性试验

确证性试验是一种预先提出假设并进行评价的具有充分对照的试验。原则上确证性试验需要提供有效性或安全性的确凿证据。此类试验中，感兴趣的关键假设通常需预先定义，应能直接反映试验的主要目的，且在试验完成后得到检验。在确证性试验中，以适当的精度估计处理效应的大小，与把这些效应和临床意义联系起来同等重要。

确证性试验旨在提供确凿证据以支持主张，因此，按照方案及标准操作规程进行试验尤为重要。应该解释和书面记录不可避免的变化，并考察它们的影响。此类试验设计的合理性以及其它重要的统计方面，如计划分析的主要特征，均应写入方案。每个试验应仅解决有限的问题。

支持所主张的确凿证据要求确证性试验的结果证实研究产品具有临床获益。因此确证性试验应清晰明确地回答每一个与有效性或安全性主张有关的关键临床问题。另外，推论（见词汇表）到目标患者人群的基础得以理解和解释很重要，这也会影响到所需研究中心和/或试验的数量和参与人员（如专家或全科医师）。确证性试验的结果应当是稳健的。某些情况下，单一确证性试验所提供证据强度可能就足够了。

2.1.3 探索性试验

确证性试验的理论基础和设计几乎总是依赖于一系列

早期探索性临床研究工作。这些探索性研究和所有临床试验一样应有清晰和明确的目的，但与确证性试验相比，它们的目的并不总是对预先定义的假设进行简单检验。此外，探索性试验可能有时需要采用更灵活的方法进行设计，以便根据积累的结果更改设计。它们的分析可能仅限于数据探索，也可能进行假设检验，但假设的拟定可能依赖于数据。尽管这类试验可能对整体的相关证据有贡献，但不能作为证明有效性的正式依据。

任何试验可能同时具有确证性和探索性两个方面。例如，在大多数确证性试验中，也会对数据进行探索性分析，作为解释和支持研究发现、为后期研究提出进一步假设的基础。方案应明确区分进行确证试验和对数据做探索性分析的两种不同情况。

2.2 试验范围

2.2.1 人群

在药物研发的早期阶段，临床试验受试者的选择在很大程度上受到主观愿望的影响，即希望最大可能地观察到感兴趣的特定临床疗效，因此，研究对象往往是药物最终适用的患者总体中一个非常局限的亚组。但在开展确证性试验的时候，试验受试者应更能反映目标人群。因此，在保持足够的同质性以精确估计处理效应的同时，尽可能放宽目标人群的纳入和排除标准，这对确证性试验是有益的。由于地理位置、

实施时间、特定研究者和诊所的医疗实践等因素的影响，任何一个临床试验都不可能完全代表将来的用药者。尽管如此，应尽可能减少这些因素的影响，并在解释试验结果时充分讨论。

2.2.2 主要和次要指标

主要指标（又称“目标”指标，主要终点）应能够提供与试验主要目的直接相关的最具临床相关性和说服力的证据。通常应只设置一个主要指标。因大部分确证性试验的主要目的是提供与有效性相关的强有力的科学证据，所以主要指标通常是有效性指标。安全性/耐受性有时也可能是主要指标，且会一直是一种重要的考量。有关生活质量和卫生经济的指标是进一步的潜在主要指标。主要指标的选择应反映相关研究领域公认的准则和标准。建议使用在早期研究或发表文献中获得的具有实践经验的可靠且已验证的指标。在纳入和排除标准所描述的患者人群中，应该有充分的证据说明主要指标能够有效和可靠地度量临床相关的和重要的治疗获益。主要指标通常用于样本量估计（见第 3.5 章节）。

很多情况下，评价受试者结局的方法可能并不直接，应仔细定义。例如，将死亡率作为主要指标而无进一步说明是不够的，因为对死亡率的评价可以是比较某些固定时点的存活比例，也可以是比较在特定时域内生存时间的总体分布。另一个常见的例子是复发事件，处理效应的测量可以是简单

的二分类指标(特定时期内的任何复发)、首次复发的时间、复发率(观察的单位时间的事件数),等等。在评价慢性病的处理效应时,随时间变化的功能状态对选择主要指标提出了其他挑战。相应的方法有多种,例如,观察期开始和结束时所做评价的比较、由观察期所有评价求得的斜率的比较、超过或低于规定阈值的受试者比例的比较、基于重复测量数据方法的比较。为避免因事后定义所产生的多重性担忧,在方案中规定主要指标的精确定义至关重要,因为该定义将用于统计分析。另外,所选择的具体主要指标的临床相关性和相关测量过程的合理性通常需要在方案中阐明。

主要指标及其选择理由应在方案中详细说明。揭盲后重新定义主要指标通常是不可接受的,因为由此引入的偏倚很难评价。当根据主要目的确定的临床效应存在多种测量方法时,应根据临床相关性、重要性、客观性、和/或其它相关特性,在方案中选择其中一种切实可行的测量方法作为主要指标。

次要指标是与主要目的相关的支持性指标,或与次要目的相关的效应指标。在方案中预先定义次要指标,并说明它们的相对重要性以及在解释试验结果时的作用也很重要。次要指标的数量应有限制,且与试验要回答的有限问题相关。

2.2.3 复合指标

当与主要目的相关的多种测量方法中难以确定单一的

主要指标时，另一种有用的策略是按预先确定的计算方法将多个指标组合成一个单一或“复合”指标。主要指标有时以多种临床测量方法相组合的形式出现（如关节炎、精神疾病和其它疾病使用的量表），这虽涉及多重性问题，但无需调整 I 类错误。将多个指标组合的方法应在方案中详细说明，且应以临床获益的大小对结果进行解释。当复合指标被用作主要指标时，可以对复合指标中有临床意义的单个指标进行单独分析。当量表被用作主要指标时，阐明内容效度（见词汇表）、评价者内和评价者间信度（见词汇表）及检测疾病严重程度变化的反应度等尤其重要。

2.2.4 全局评价指标

在某些情况下，全局评价指标（见词汇表）用于评价某个处理的整体安全性、有效性和/或实用性。这种指标类型整合了客观指标和研究者对受试者的状态或状态变化的总体印象，它通常是一个有序分类量表。整体有效性的全局评价方法已经用于某些治疗领域，如神经病学和精神病学。

全局评价指标一般带有主观成分。使用全局评价指标作为主要或次要指标时，应该在方案中对量表的以下方面进行详细说明：

1）量表与试验主要目的的相关性；

2）量表的效度和信度基础；

3）如何根据所收集的数据将个体受试者归类于量表中

的特定类别;

4）如何将有缺失数据的受试者归类于量表中的特定类别，或用其他方法评价。

若研究者选取的全局评价指标中包含客观指标，则这些客观指标应作为附加的主要指标，或至少作为重要的次要指标。

全局实用性评价综合了获益与风险两方面因素，反映了经治医生的决策过程，即医生在做出使用产品的决策时，必须权衡获益与风险。全局实用性指标会产生这样的问题，即某些情况下会将获益和不良反应方面差别很大的两种产品判断为等效。例如，将一种治疗的全局实用性指标判断为等效于或优效于另一种治疗时，可能掩盖了其疗效甚微或无效但不良反应较少的事实。因此不建议将全局实用性指标作为主要指标。如果全局实用性指标被用作主要指标，则将特定的有效性和安全性结局分别作为附加的主要指标考虑是非常重要的。

2.2.5 多个主要指标

有时需要使用一个以上的主要指标，且每一个指标（或其中一个子集）都足以涵盖其治疗效果的范围。解释这类证据的既定方式应当详细说明，即应该说明对任一指标，或最少几个指标，或全部指标的影响是否被认为是达到试验目的所必需的。应该针对已定义的主要指标清楚地说明主要假设

或相关的假设与参数（如均数、百分数、分布），并清楚地叙述统计推断方法。因为存在潜在的多重性问题，所以应解释对I类错误的影响（见第5.6章节），也应在方案中给出控制I类错误的方法。在评价对I类错误的影响时，所提出的主要指标之间的相关程度也需要考虑。如果试验目的是证实所有主要指标的效果，则无需调整I类错误，但必须仔细考虑对II类错误和样本量的影响。

2.2.6 替代指标

当通过观察实际临床有效性直接评价受试者的临床获益不可行时，可以考虑间接标准（替代指标—见词汇表）。一些被认为可以预测临床获益的指标通常可作为替代指标。确定替代指标有两个主要关注点：第一，它可能不是相关临床结局的真正预测因子，例如，它可以测量与一个特定药理学机制有关的治疗活性，但不能提供治疗的作用范围与最终效果的全部信息，无论是阳性还是阴性。许多例证表明，治疗在替代指标显示出高度阳性效应，而最终被证明对受试者的临床结局是有害的。与此相反，也有一些例证显示，治疗的临床获益明确却未能在替代指标体现。第二，替代指标可能不会定量测量可直接权衡不良反应的临床获益。验证替代指标的统计学标准已经具备，但是使用它们的经验相对有限。在实践中，替代证据的强度取决于（1）替代关系的生物学合理性；（2）流行病学研究证明替代指标对临床结局的预后价

值;（3）临床试验证明替代指标的处理效应相当于临床结局的效应。一种产品的临床指标和替代指标之间的关系并不一定适用于治疗同一种疾病但具有不同作用方式的另一种产品。

2.2.7 分类指标

连续型或等级指标有时可能需要转化为二分类或其他分类指标。“成功”和“应答”的标准是二分类的常见例子。分类标准需明确规定，例如，连续型指标最小百分比的改善（相对于基线），或者有序等级量表中等于或高于某个阈值水平（如“良”）的按顺序分类。

舒张压降低于 90mmHg 是一个常见的二分类例子。当分类有明确的临床相关性时，它们是最有用的。众所周知，选择分类标准很容易使临床结果产生偏倚，因此在方案中应预先定义和特别说明分类标准。由于分类通常意味着信息丢失，因此在分析中会损失检验效能，样本量计算时需加以考虑。

2.3 避免偏倚的设计技术

临床试验中，避免偏倚的最重要的设计技术是盲法和随机化，它们为上市申请中大多数对照临床试验所常规采用。大多数此类试验采用双盲法，按照合适的随机化方案，对治疗药物进行预先包装并提供给试验中心，只标明受试者编号和疗程，从而使参与试验的任何人都不知道分配给任何特定受试者的具体治疗药物，甚至不知道编码字母。该方法会在

第 2.3.1 章节和第 2.3.2 章节中的大部分内容中进行介绍，例外情况会在最后考虑。

设计阶段应在方案中制定针对性措施，以使试验实施过程中可能损害分析的不规范操作最小化，从而减少偏倚。这里指的不规范操作包括各种类型的方案违背、退出和数据缺失。方案中应考虑一些方法，以减少出现这些问题的频率，以及解决在数据分析中出现的问题。

2.3.1 盲法

盲法或遮蔽是为了限制临床试验的实施和解释时所产生的有意或无意的偏倚，这些偏倚可能源于以下情况的影响：知晓受试者的招募和处理分组、受试者的后续治疗、受试者对治疗的态度、终点评价、退出的处理、从分析中剔除数据，等等。盲法的根本目标是防止知晓处理分组，直到所有产生偏倚的机会都消失。

在双盲试验中，所有受试者及参与受试者的治疗或临床评价的研究者和申办方人员，包括确定受试者资格、评价终点或评价方案依从性的任何人，均不知道受试者所接受的治疗。在整个试验实施过程中，这种盲态要始终保持，只有当数据被清理到可接受的质量水平时，才可对适当的人员揭盲。如果需要对不参与受试者的治疗或临床评价的申办方人员揭盲处理编码（如生物分析学家、稽查员、参与严重不良事件报告的人员），申办方应该制定严格的标准操作规程，以防

止处理编码的不当传播。在单盲试验中，研究者和/或他的成员知道处理分组信息，但受试者不知道，反之亦然。在开放试验中，所有的人都可能知道处理分组信息。双盲试验是最优方法，它要求试验所采用的处理在使用前或使用期间均无法被识别出来（如外观、味道等），且在整个试验期间均适当地保持盲态。

达到理想的双盲会有很多困难：有些处理可能具有完全不同的性质，例如，手术和药物治疗；两种药物可能具有不同的剂型，虽然使用胶囊可以令它们无法被区分，但改变剂型可能会改变药代动力学和/或药效学的特性，因此需要建立制剂的生物等效性；两种处理的每日用法可能不同。这些情况下，使用“双模拟”（见词汇表）技术是实现双盲条件的一种方法，该技术有时会强制实施一种非同寻常的使用方案，使得受试者的积极性和依从性受到负面影响。伦理上的困难也可能会干扰该技术的应用，例如手术过程的模拟。无论如何，应当努力克服这些困难。

某些临床试验的双盲性质可能由于明显的处理诱导效应而遭到部分破坏。这种情况下，使研究者和有关申办方人员对某些检验结果（如所选择的临床实验室测量）保持盲态，可以使盲法得到改善。使偏倚最小化的类似方法（见下文）应当在开放试验中考虑，例如独特的处理效应无法对患者设盲的试验。

如果双盲试验不可行，则应考虑用单盲方案。有些情况下，只有开放试验在实践上或伦理上是可行的。单盲和开放试验更具灵活性，但特别重要的是，研究者知道了下一个受试者的处理不应影响入组受试者的决定，即该决定应在知道随机化处理之前做出。对于这些试验，应考虑使用中央随机化方法，如采用电话随机化管理处理的分配。此外，应该由不参与治疗受试者并对处理保持盲态的医务人员进行临床评价。在单盲或开放试验中，应尽一切努力使各种已知的偏倚来源降到最低，并且应采用尽可能客观的主要指标。应在方案中解释所采用的盲态程度的原因，以及所采取的使偏倚最小化的措施。例如，申办方应当有严格的标准操作规程，以保证在清理数据库以供分析之前，适当限制对处理编码的获取。

只有经治医师认为对某一受试者的治疗有必要知道其处理分配时，才应考虑对该受试者破盲。无论什么原因导致的任何有意或无意地破盲都应该在试验结束时给予报告和解释。处理分配的揭盲过程及时间都应该记录在案。

本文件中，数据的盲态审核（见词汇表）是指在试验完成（对最后一位受试者的最后一次观察）到揭盲之间的这段时间内对数据的检查。

2.3.2 随机化

在临床试验中，随机化将机会元素引入到受试者的处理

分配中。在试验数据的后续分析期间，它为定量评价与处理效应有关的证据提供了坚实的统计基础。它倾向于使各处理组的已知和未知的预后因素分布相似。与盲法结合，在受试者的选择和分配时，随机化有助于避免因处理分配的可预测性而可能出现的偏倚。

临床试验的随机化列表记录了施与受试者处理的随机分配，其最简单的方式是处理的序列表（或交叉试验中的处理序列），或按受试者编号对应的编码。有些试验，如具有筛选阶段的试验，可能使问题复杂一些，但是预先计划的受试者的处理分配或处理序列应是唯一的。不同的试验设计需要不同的程序来生成随机化列表。随机化列表应当有重现性（如果需要）。

虽然无限制条件的随机化是一种可接受的方法，但区组随机一般具有某些优势，它有助于增加处理组间的可比性，特别是当受试者特征可能随时间变化时，例如由于招募策略改变引起的变化。它还能更好地保证各处理组的样本量几乎相等。在交叉试验中，它提供了获得具有更高效率和更易于解释的平衡设计的方法。选择区组长度时需注意，既要足够短以限制可能的不平衡，又要足够长以避免对区组序列末尾的可预测性。区组长度通常应对研究者及其他有关人员保持盲态；使用两种或多种区组长度与每个区组随机选择长度，可达到同样目的。（理论上，在双盲试验中，可预测性并不重

要，但药物的药理作用可能提供猜测机会。）

对于多中心试验（见词汇表），应按中心进行随机化。提倡每个中心有一个单独的随机方案，即按中心分层或为每个中心分配若干完整的区组。更一般地，按照基线测量的重要预后因素（如疾病的严重程度、年龄、性别等）进行分层，可保障层内的平衡分配，这种方法在小型试验中潜在益处更大。分层因素一般不超过三个，否则实现平衡不仅困难，而且麻烦。应用动态分配程序（见下文）可能有助于同时在多个分层因素之间达到平衡，只要可以调整其余试验流程以适应这类方法。应当在后续的分析中对分层随机化的因素加以考虑。

进入试验的下一个随机化受试者，应该接受对应于随机化列表（如果随机化是分层的，则在相应的层中）中下一个号码的处理。只有当已经确认下一个受试者进入到试验的随机化阶段时，才能给受试者分配合适的号码和相关处理。具有增加可预测性的随机化细节，如区组长度，不应包含在试验方案中。随机化列表本身应该由申办方或独立方安全存档，以确保整个试验过程维持盲态。在试验期间获取随机化列表应该考虑在紧急情况下为任何受试者破盲的可能性。破盲应遵循的程序、必要的文件以及受试者后续的处理和评价均应在方案中写明。

动态分配也是一种选择，该方法根据当前已分配的处理

的平衡情况进行处理分配，对于分层试验，处理分配视受试者所属层内的平衡情况而定。应当避免确定性的动态分配程序，应当为每个处理分配纳入适当的随机化要素。应尽一切努力保持试验的双盲状态。例如，仅限于中央试验办公室知道处理编码，并由办公室通过电话联系来控制动态分配。这种方法允许对入选标准进行额外检查，并会建立试验入组的记录，这些信息对某些类型的多中心试验具有价值。随后会启用双盲试验的预包装和贴标签的药品供应系统，但它们的使用顺序不再是依次的。最好使用适当的计算机算法使中央试验办公室的人员对处理编码保持盲态。当考虑动态分配时，应该仔细评价物流的复杂性以及对分析的潜在影响。

3.试验设计的考虑

3.1 设计类型

3.1.1 平行组设计

对于确证性试验，最常见的临床试验设计是平行组设计，该设计将受试者随机分配到两组或多组中的一组，每组采用不同的处理。这些处理包括一个或多个剂量的研究产品，以及一个或多个对照处理，如安慰剂或/和阳性对照。该设计的假设比大多数其它设计简单，但与其它设计一样，可能会有使分析和解释复杂化的额外试验特征，如协变量、随时间的重复测量、设计因素之间的交互作用、方案违背、脱落（见词汇表）、退出等。

3.1.2 交叉设计

在交叉设计中，每个受试者被随机分到两个或多个处理序列，因此处理间的比较相当于自身对照。这种简单策略之所以有吸引力，主要因为它减少了满足检验效能所需的受试者，有时减少的程度相当可观。2×2 交叉设计是最简单的，该设计通常在先后两个处理周期中安排一个洗脱期，每个受试者以随机顺序在每个处理周期接受两个处理中的其中一个。最常见的扩展设计是 n 个周期和 n（>2）个处理，每个受试者先后接受所有 n 个处理。此类设计形式多样，例如，每个受试者接受 n（>2）个处理中的一个子集，或者对一个受试者重复给予处理。

交叉设计有很多问题可导致其结果无效，主要困难在于残留效应，即在后继处理周期内的前序处理的残余影响。使用相加模型时，不同的残留效应将使处理间的直接比较产生偏倚。对于 2×2 设计，统计上无法将残留效应从处理与周期的交互作用中区分开来，并且因为相应的对比是“受试者之间”，故检验这两个效应中任何一个都缺乏检验效能。这一问题在高阶设计中并不严重，但不能完全消除。

因此，使用交叉设计重要的是要避免残留效应，最好的办法是在充分了解疾病领域和新药的基础上有选择地和谨慎地使用该设计，诸如针对病情稳定的慢性病；治疗周期内可充分发挥药物的相关效应；洗脱期足够长以使药物效应完

全消退等。应该在试验前利用已有信息及数据确定是否可满足这些条件。

交叉试验还有一些需要密切注意的问题，其中，受试者失访导致的分析和解释的复杂化最值得关注。另外，残留效应的潜在作用导致后续处理周期所发生的不良事件很难判断是哪种处理所致。这些问题以及其它问题在 ICH E4 中已有阐述。交叉设计一般应严格限于预期仅有少数失访的试验。

采用 2×2 交叉设计验证相同药物的两种制剂的生物等效性甚为常用，往往令人满意，尤其是以健康志愿者为对象的试验，如果两个周期间的洗脱时间足够长，极不可能发生相关药代动力学指标的残留效应。不过，在分析期间基于获得的数据核实这一假设仍然非常重要，例如，通过在每个周期开始时未检测到药物来证实无残留效应。

3.1.3 析因设计

在析因设计中，通过使用不同的处理组合可以同时评价两个或多个处理。最简单的例子是 2×2 析因设计，受试者被随机分配到两个处理 A 和 B 的四种可能组合之一，即单独 A、单独 B、既有 A 又有 B、既无 A 又无 B。该设计多以检验 A 和 B 的交互作用为特定目的。如果基于检验主效应计算样本量，则交互作用统计检验的检验效能可能不足。当该设计被用于检验 A 和 B 的联合效应时，特别是如果两者可能被一起使用，这一考虑尤为重要。

析因设计的另一个重要用途是，建立同时使用处理 C 和 D 时的剂量-反应特征，特别是在先前试验中每种单一疗法的某个剂量的有效性已被证实的情况。设 C 的剂量数为 m（通常包括零剂量，即安慰剂），相似的 D 的剂量数为 n，整个设计由 m×n 个处理组构成，每个处理组为一种不同的 C 和 D 的剂量组合，则应用响应面的结果估计可以帮助确定临床使用的 C 和 D 剂量的恰当组合（见 ICH E4）。

某些情况下，如评价两种处理的有效性所需的受试者数量与单独评价任一种处理的有效性所需的受试者数量相同时，2×2 设计可能会更高效地利用受试者，这一策略已经被证实对非常大型的死亡率试验颇有价值。该方法的效率和可靠性取决于处理 A 和 B 之间不存在交互作用，使得 A 和 B 对主要有效性指标的主效应服从相加模型，因此，无论是否追加 B 的效应，A 的效应是确定的。对于交叉试验，应在试验前利用先前的信息和数据，这很可能会找到满足无交互作用的证据。

3.2 多中心试验

开展多中心试验主要有两个原因。首先，多中心试验是一种更加高效地评价新药的可接受的方法；某些情况下，为在合理的时间框架内获得足够的受试者以满足试验目的，它可能是唯一可行的方法。原则上，在临床研发的任何阶段均可开展这种性质的多中心试验。多中心试验可能有几个中心，

每个中心的受试者数量较大；也可能有很多中心，每个中心只有很少的受试者，比如罕见病研究。

其次，设计成多中心（和多个研究者）试验主要是为研究结果的后续推论提供更好的基础，因为从更广泛的人群中招募受试者和呈现更宽泛的使用药物的临床环境，从而呈现出更典型的未来用药场景。这种情况下，许多研究者的参与也可提供更宽泛的药物价值临床判断。此类试验在药物研发后期将成为确证性试验，可能有大量的研究者和中心参与。为增强可推论性(见词汇表)，多中心试验有时会在许多不同国家实施。

要想充分解释和外推多中心试验结论，所有中心实施研究方案的方式应该是明确的和相似的。样本量和检验效能的计算通常基于各中心的处理间差异是相同的无偏估计的假设，因此，制定共同研究方案并给予实施很重要。试验的实施流程应该尽可能标准化。通过研究者会议、试验前的人员培训和试验期间的严密监查，可以减少评价标准和方法的不一致性。良好设计的目的通常是实现每个中心内各处理组的受试者分布相同，而良好管理可以对该目的起到支持作用。应避免中心间的病例数相差太大以及个别中心病例数太少，这一考虑的好处会在后期探查中心间处理效应的异质性时显示出来，因为这样可以减少处理效应不同加权估计之间的差异。(这一点并不适用于所有中心病例数都非常少的试验，

以及分析时不考虑中心效应。)如果不采取这些预防措施，加之对结果同质性的质疑，会使多中心试验的价值降低，有时甚至严重到不能为申办方的主张提供令人信服的证据的地步。

最简单的多中心试验是每位研究者负责在一家医院招募受试者，所以，“中心”是由研究者或医院唯一确定的。可是，很多试验会更复杂一些，例如，一个研究者可能从几家医院招募受试者；一个研究者可能代表一个临床医生团队（参与研究者），他们或从一家医院所辖的几个诊所，或从几家相关的医院招募受试者。只要对统计模型中关于中心的定义有疑义，方案中的统计章节（见第 5.1 章节）就应在特定试验背景下明确定义该术语（例如，按研究者、场所或地区）。多数情况下，根据研究者定义中心较为可行，ICH E6 在这方面提供了相关指南。定义中心的目的是使影响主要指标测量的因素和处理的影响达到同质，以免因此引起质疑。任何将中心合并起来进行分析的规则应尽可能在方案中合理阐述并预先规定，但是，任何基于此方法的决策都应始终在盲态下做出，如盲态审核。

方案中应该描述处理效应的估计和检验的统计模型。主要处理效应估计可首先使用包含中心效应的模型，但不包含处理与中心的交互项。如果处理效应中心间是同质的，则在模型中常规地包含交互项会降低对主要效应的检验效率；如

果确实存在处理效应的异质性，则对处理效应的解释是有争议的。

某些试验，如大型的死亡率试验，每个中心只有很少受试者，设想中心对主要或次要指标有任何影响都是缺乏依据的，因为中心因素的影响不可能代表临床重要性。还有一些试验可能从一开始就会认识到每个中心有限的受试者使得统计模型中包含中心效应变得不切实际。这种情况下，模型中不应包含中心项，而且也没有必要按中心进行分层随机化。

对于每个中心都有充足的受试者的试验，如果发现阳性处理效应，通常应探索不同中心间处理效应的异质性，因为这可能影响结论的外推性。通过各中心结果的图示方法，或通过对中心与处理间交互作用的统计检验，可能会发现明显的异质性。对交互效应做统计检验时，需认识到其检验效能不高，因为试验是基于探测处理的主效应而设计的。

如果发现处理效应的异质性，则应当谨慎地加以解释，并应积极尝试从试验管理的其他特征或受试者特征方面来寻找原因。这样的原因通常会提示适当的进一步分析和解释。在缺乏原因的情况下，一旦证实处理效应的异质性，例如，通过明显的定量交互作用(见词汇表)，意味着处理效应可能需要另一种估计，比如给中心不同赋权以保障处理效应估计的稳健性。理解定性交互作用（见词汇表）的异质性甚至更为重要，当未能找到原因时，要想可靠地预测处理效应，可

能需要进一步开展临床试验。

以上针对多中心试验的讨论都是基于采用固定效应模型的。混合模型也可用于探索处理效应的异质性，它把中心效应和中心与处理间的交互效应看作是随机的，尤其适合于中心数量特别多的情况。

3.3 比较的类型

3.3.1 优效性试验

科学地讲，通过安慰剂对照试验显示优于安慰剂，或通过显示优于阳性对照处理，或显示剂量-反应关系，所得到的疗效是最可信的。此类试验被称为“优效性”试验(见词汇表)。本指南一般以优效性试验为假定，除非另有明确说明。

对于严重疾病，如果存在经优效性试验验证的有效的治疗方法，采用安慰剂对照试验可能被认为是有悖伦理的。这种情况下，应当科学地采用阳性对照。安慰剂对照和阳性对照的适用性应当不同试验给予不同考虑。

3.3.2 等效性或非劣效性的试验

某些情况下，研究产品与参照处理相比的目的并非为了显示优效性。此类试验根据其目的分为两大类，一类是“等效性”试验(见词汇表),另一类是“非劣效性”试验(见词汇表)。

生物等效性试验属于前一类。某些情况下，出于其他监管原因也进行临床等效性试验，例如，当化合物不被吸收并因此不存在于血液中时，验证仿制产品与已上市产品的临床

等效性。

很多阳性对照试验用于验证研究产品的有效性非劣效于阳性对照药，因此属于后一类。另一种可能是在试验中将研究药品的多个剂量与标准药品的推荐剂量或多个剂量进行比较。这种设计的目的是同时显示研究产品的剂量-反应关系，并将研究产品与阳性对照进行比较。

阳性对照等效性或非劣效性试验也可引入安慰剂对照，从而在一个试验中设定多个目标，例如，这种设计在验证优效于安慰剂的同时，还可以评价相对于阳性对照的有效性与安全性的相似程度。众所周知，采用不包含安慰剂或不设置新药多个剂量的阳性对照等效性（或非劣效性）试验会面临一些困难。与优效性试验相比，此类试验隐性缺乏内部效度，因此必须进行外部验证。等效性（或非劣效性）试验本质上并不保守，因此，在试验设计或实施中的许多缺陷倾向于使结果倾向等效的结论。由于这些原因，这些试验的设计特点应受到特别关注，它们的实施需要特别小心，例如，尽量减少违反入选标准、不依从、退出、失访、数据缺失和其它偏离方案的发生率，并使它们对后续分析的影响降至最低。

应谨慎选择阳性对照。恰当的阳性对照应该是一种被广泛使用的疗法，其针对相关适应症的疗效已在良好设计和良好记录的优效性试验中得到了量化确认，并且能够可靠地预期在将要实施的试验中显示出相似的疗效。为此，新试验应

该与以前实施且明确显示出临床相关疗效的优效性试验具有相同的重要设计特征（主要指标、阳性对照的剂量、入排标准等），且考虑与新试验相关的医学或统计学实践的进展。

在试验方案中，一个关键问题是要把证明等效性或非劣效性的意图清晰明确地表述出来。方案中应规定一个等效界值，该界值被视为临床可接受的最大差异，并且应当小于在阳性对照优效性试验中所观察到的差异。对于阳性对照等效性试验，需规定等效界值的上限和下限；而对于阳性对照非劣效性试验，仅需规定界值下限。等效界值的选择应具备临床的合理性。

统计分析通常采用置信区间方法（见第 5.5 章节）。对于等效性试验，应当使用双侧置信区间。如果置信区间完全落在等效界值之内，可推断为等效。在实操上，该法相当于双单侧检验方法，其（复合）无效假设是处理间差异在等效界值之外，（复合）备择假设是处理间差异在等效界值之内。由于两个无效假设无重叠，故 I 类错误可控。对于单侧假设检验，其无效假设是处理间差异（试验品减去对照品）等于或小于等效界值的下限，而备择假设是处理间差异大于等效界值下限。单侧或双侧检验的 I 类错误选择有所不同。样本量计算应当基于这些方法（见第 3.5 章节）。

在研究产品与阳性对照之间无差异的无效假设下，如果基于观察到无显著差异的检验结果，做出等效性或非劣效性

的结论是不合适的。

在选择分析数据集时也存在一些特殊问题。处理组或对照组退出或脱落的受试者都倾向于缺乏应答，因此使用全分析集(见词汇表)的结果证实等效性可能存在偏倚(见第 5.2.3 章节)。

3.3.3 剂量-反应关系的试验

新研究产品的剂量与应答如何相关，是一个在研发的所有阶段通过各种方法都可获得答案的问题（见 ICH E4）。剂量反应试验可服务于许多目的，相对重要的有：有效性的确证；剂量反应曲线的形状和位置的研究；适宜初始剂量的估计；个体剂量调整的最优策略确定；最大剂量的确定（超出该剂量不可能额外获益）。达到上述目的需要收集研究中各种剂量的数据，包括安慰剂（零剂量）。为此，需用到估计剂量反应关系的方法，包括统计检验以及同样重要的置信区间构建和图示方法。假设检验可能需要根据剂量的自然顺序或关于剂量-反应曲线的形状(如单调性)的特定问题做出调整。应当在方案中提供详细的统计分析计划。

3.4 成组序贯设计

采用成组序贯设计便于进行期中分析（见第 4.5 章节和词汇表）。成组序贯设计虽然不是用于期中分析的唯一可接受的设计类型，却是最常用的，因为在试验期间以周期性间隔评价不同分组的受试者的结局比在获得整个试验每一个

受试者数据后进行评价更为可行。在获得处理结局和受试者的处理分配（如揭盲，见第 4.5 章节）的信息之前，应充分说明统计方法。独立数据监查委员会（见词汇表）可对来源于成组序贯设计的数据实施审查或进行期中分析（见第 4.6 章节）。该设计不仅已被最广泛地、成功地应用于大型、长周期的以死亡率或主要非致死性结局为终点的试验，它在其它方面的应用也在增加。尤其是，人们已经认识到所有试验中都必须监查安全性，因此，为了出于安全原因提早终止试验而制定正式流程的必要性往往是需要考虑的。

3.5 样本量

临床试验的受试者例数应足够大，以对所提出的问题提供可靠答案。样本量通常由试验的主要目的确定，如果由其它要素确定，则应明确说明理由。例如，基于安全性问题或需要或者基于重要的次要目的确定的样本量可能比基于主要有效性问题确定的样本量需要更多的受试者(例如,见 ICH E1a)。

一般的样本量确定方法应考虑以下要素：主要指标、检验统计量、无效假设、所选剂量下的备择("工作")假设(所选受试者人群中在所选剂量下检测出或拒绝的处理间差异)、错误拒绝无效假设的概率（I 类错误）、错误地不拒绝无效假设的概率(II 类错误),以及应对退出和违背方案的处理方法。某些情况下，以事件率为评价检验效能的主要手段，此时需

要做出一些假设，以从所需的事件数推算出试验的最终样本量。

应在方案中给出计算样本量的方法，以及在计算中使用的任何估计量（如方差、均值、反应率、事件率、待检测的差异）。也应该给出这些估计的依据。研究这些假设的偏离对样本量估计的敏感性很重要，而根据偏离假设的合理范围给出对应的样本量范围则是一种方便可行的方法。在确证性研究中，假设通常应基于公开发表的数据或早期试验的结果。对于待检测的处理间差异，可依据在患者管理中对具有临床相关性的最小效应的判断，也可依据对新处理的预期效应的判断，相比之下后者的预期效应更大。通常 I 类错误概率设在 5%或者更小，或者由多重比较所需要的任何调整来决定；检验假设的事先合理性以及结果的预期影响可能会影响 I 类错误的精确选择。II 类错误的概率通常设在 10%到 20%之间，申办方通常愿意让该值尽可能低，尤其当试验难以或不可能重复时。某些情况下，采用与常规的 I 类和 II 类错误水平不同的值也可能被接受，甚至更可取。

样本量应是主分析所需的受试者数量。如果这是“全分析集”，则效应大小的估计与符合方案集（见词汇表）相比，可能需要降低。这是因纳入了退出处理的或者依从性差的患者数据，而考虑稀释处理效应。相应地关于变异的假设可能也需要修改。

等效性或非劣效性试验（见第 3.3.2 章节）的样本量通常应基于获得处理间差异的置信区间的目的，该差异是指临床可接受的最大处理间差异。如果等效性试验的检验效能是在假设真实差异为 0 的条件下确定的，如果真实差异不为 0，则达到这一检验效能所需的样本量会被低估。如果非劣效性试验的检验效能是在假设 0 差异的条件下确定的，如果试验产品的效应低于对照，则达到这一检验效能所需的样本量会被低估。“临床可接受的”差异的选择需要合理说明它对将来患者的意义，并且可能小于上文提到的优效性试验旨在证明的“临床相关的”差异。

成组序贯试验不能预先确定确切的样本量，因为它依赖于机会作用以及所选择的终止试验的准则和真实的处理间差异。终止准则的设计应该考虑后续样本量的分布，通常表达为预期样本量和最大样本量。

当事件率低于预期或变异大于预期时，在不揭盲数据或不进行处理间比较的情况下，可使用样本量重新估计的方法（见第 4.4 章节）。

3.6 数据采集及处理

数据的收集和研究者向申办方传输数据可通过各种媒介进行，包括纸质病例报告表、远程现场监查系统、医疗计算机系统和电子传输。无论采用何种数据收集工具，所收集信息的形式和内容都应完全符合方案，并应在临床试验实施

前确定。应注重分析计划的实施所必须的数据，包括确认方案依从性或确定重要方案违背所需要的背景信息（如与服用剂量有关的时点评价）。“缺失值”应该与“0 值”或“特征缺失”区分开来。

从数据收集到数据库最终确定的过程应该按照 GCP 进行（见 ICH E6，第 5 章节）。具体来说，需要及时可靠的程序用于记录数据和纠正错误与遗漏，以确保交付高质量的数据库，并通过实施计划的分析达到试验目的。

4.试验实施的考虑

4.1 试验监查和期中分析

按照方案认真实施临床试验，对结果的可靠性具有重大影响（见 ICH E6）。仔细监查可以确保尽早发现困难，并将它们的发生和复发减至最小。

由制药企业资助的确证性临床试验，通常有两种截然不同的监查类型。一种关注试验质量的监督，另一种涉及破盲以进行处理间的比较（即期中分析）。两种试验监查，除人员职责不同外，还涉及不同类型试验数据和信息的获取，因此需用不同的规则控制潜在的统计和操作偏倚。

出于监督试验质量的目的，试验监查中所涉及的检查可能包括：是否遵循方案，累积数据是否可接受，计划的收集目标是否达到，设计假设是否合适，以及在试验中保留患者是否成功，等等（见第 4.2 至 4.4 章节）。这种类型的监查既

不需要获取比较处理效应的信息，也不需要对数据进行揭盲，因此对 I 类错误没有影响。出于这一目的对试验进行监查是申办方的职责（见 ICH E6），可由申办方或申办方选择的独立小组来进行。这种类型的监查周期一般是从选择试验现场开始，到收集和清理最后一位受试者的数据结束。

其他类型的试验监查（期中分析）涉及到比较处理结果的累积。期中分析需要揭盲（即破盲）获取处理组分配信息（实际的处理分配或者各组分配的标识）以及比较处理组的汇总信息。这需要在方案（或者首次分析之前的适当修订）中包含期中分析的统计计划，以防止某些类型的偏倚，见第 4.5 和 4.6 章节的讨论。

4.2 纳入与排除标准的更改

纳入与排除标准应按方案的规定保持恒定，贯穿受试者招募期。偶尔有些改变是允许的，例如，在长周期试验中，从试验外部或期中分析所获得的对医学知识新的认识，可能建议修改入组标准。监查人员发现违背入组标准情况经常发生，或者由于入组标准过严导致非常低的招募率，也都可能是修改入组标准的理由。修改入组标准应在不破盲的情况下进行，并通过方案修订进行描述，修订的方案应涵盖任何统计学方面的变动，如不同事件率所致的样本量调整，或者分析计划的修改，如根据修改的纳入/排除标准进行分层分析。

4.3 入组率

在受试者入组时间较长的试验中，应监查入组率，如果它明显低于预期水平，应该查明原因并采取补救措施，以确保试验的检验效能，并减轻对选择性入组和其他质量问题的担忧。这些考虑适用于多中心试验的各个中心。

4.4 样本量调整

在长周期试验中，通常有可能对原设计和样本量计算所依据的假设进行检查。如果试验设计的某些重要规定是根据初步的和/或不确定的信息做出的，这种检查尤其重要。对盲态数据进行期中检查可能会发现总应答的方差、事件率或生存状态不如预期。此时，可能需要通过适当修改假设来修正样本量，还应在方案修订和临床研究报告中说明其合理性并记录在案。应该解释为保持盲态所采取的措施及其对 I 类错误和置信区间宽度的影响（如果有）。只要可能，都应在方案中表述样本量再估计的潜在需要（见 3.5 章节）。

4.5 期中分析和提早终止试验

期中分析是指，在试验正式完成之前的任何时间，为比较处理组间的有效性或安全性而进行的任何分析。因为这些比较的次数、方法及结果影响试验的解释，因此所有期中分析都应当预先仔细计划并在方案中阐明。有些特殊情况，期中分析可能在试验开始后才发现有必要实施。对于这种情况，补充定义期中分析的方案修订应在分析数据揭盲之前。当期

中分析用于决定是否终止试验时，通常会采用成组序贯设计，该设计以统计监查计划作为准则（见第 3.4 章节）。对于这种期中分析，出现以下情况可以提早终止试验：研究处理的优效性已被证实；相关处理间差异已被证实是不可能的；发生了不可接受的不良反应。一般来说，与安全性监查相比，通过有效性监查来提早终止试验要求更多的证据，即边界更保守。当试验设计和监查目的涉及多个终点时，应考虑多重性问题。

方案中应描述期中分析计划，或至少描述一些相关的考虑，如是否使用灵活的 α 消耗函数方法，并在第一次期中分析前，在修订的方案中提供进一步的细节。终止试验的准则和特性应在方案或修订的方案中清晰阐述。其他重要指标的分析对提早终止的潜在影响也应考虑。如果试验设有数据监查委员会（见第 4.6 章节），上述材料应由其撰写或批准。偏离计划总有可能使试验结果失效。如果试验需要修正，任何统计方面的相应修改应尽早在方案修订中详细说明，特别是讨论这些修改对任何分析或推断的影响。在统计方面应始终确保控制总 I 类错误概率。

期中分析的执行应该是一个完全保密的过程，因为可能涉及非盲的数据和结果。参与试验实施的所有人员应当对这些分析结果保持盲态，因为他们对试验的态度可能会改变并导致招募患者的特征改变或产生处理间比较的偏倚。除了直

接参与执行期中分析的人员之外，这一原则可适用于所有研究人员和申办方所雇佣的人员。研究者应仅被告知继续或终止试验的决定，或实施修订试验程序的决定。

大部分支持研究产品有效性和安全性的临床试验应全部完成计划入组的样本量。只有出于伦理原因，或者出现检验效能不再可接受的情况，试验可提早终止。然而，人们都知道出于各种原因申办方的药物研发计划需要获取处理间比较的数据，如为其它试验制定计划；另外，仅有一部分试验会涉及到严重威胁生命的结局或死亡率的研究，出于伦理原因可能需要对入组病例的处理效应比较进行连续监查。无论是哪一种情况，为了应对可能引入的潜在统计偏倚和操作偏倚，应当在分析数据揭盲之前，在方案或修订方案中制定期中统计分析计划。

对于许多研究产品的临床试验，特别是那些具有重大公共卫生意义的临床试验，应将监查有效性和/或安全性结局比较的任务委托给外部独立团队，并清楚地描述其职责。通常将该团队称为独立数据监查委员会、数据和安全监查委员会或数据监查委员会。

当申办方充当监查有效性或安全性比较的角色并因此可以获取非盲的比较信息时，应特别注意保护试验的完整性，并适当地管理和限制信息共享。申办方应当确保并记录内部监查委员会遵守书面的标准操作规程，以及含有期中分析结

果记录的决策会议纪要被维护。

任何没有恰当计划的期中分析（不论有或没有提早终止试验的影响）都可能导致试验结果的缺陷，并可能降低所得结论的可靠性，因此，应该避免这些分析。如果实施非计划的期中分析，临床研究报告应该解释其必要性，交待破盲的程度，评价所引入偏倚的潜在程度和对结果解释的影响。

4.6 独立数据监查委员会（IDMC）的作用

（见 ICH E6 第 1.25 和 5.52 章节）

独立数据监查委员会可由申办方组建，每隔一段时间评价临床试验进展、安全性数据和关键有效性指标，并向申办方建议继续、修改或终止试验。该委员会应当有书面的操作规程，并保存所有会议记录，包括期中分析结果；当试验完成时，这些应可供审查。该委员会的独立性旨在控制重要的比较信息的分享，防止临床试验的完整性受到因获取试验信息而造成的不利影响。该委员会是独立于机构审查委员会或独立伦理委员会的实体，它的组成应包括通晓统计学等相关学科的临床试验科学家。

当独立数据监查委员会中有申办方代表时，在委员会的操作规程中应明确规定他们的作用（例如，他们是否能就关键问题进行投票）。由于这些申办方人员将会获得非盲信息，因此这些操作规程还应解决如何控制期中试验结果在申办方组织内散布。

5.数据分析的考虑

5.1 分析的预先确定

当设计一个临床试验时，数据的最终统计分析的主要特征应该在方案的统计章节进行描述。该章节应包括所提出的主要指标确证性分析的所有主要特征以及解决预期分析问题的方法。对于探索性试验，该章节可描述更一般性的原则和方向。

统计分析计划（见词汇表）可作为独立文件撰写，并在最终确定方案之后完成。该文件可以更加技术性地和详细地阐述方案所述的主要特征（见第 7.1 章节）。该计划可包括对主要和次要指标以及其他数据进行统计分析的详细程序。统计分析计划应经审核或根据数据盲态审核（见第 7.1 章节定义）结果更新后，在揭盲前最终确定。最终统计分析计划的确定及随后的揭盲应保留正式记录。

如果盲态审核建议修改方案中所述的主要特征，应记录在修订方案中。否则，根据盲态审核建议考虑更新统计分析计划就足够了。只有方案（包括修订方案）中预设的分析才被认为是确证性的。

在临床研究报告的统计章节中，应该清楚地描述所采用的统计方法，包括临床试验过程中何时做出的方法学决策（见 ICH E3）。

5.2 分析集

数据纳入主分析的受试者集应在方案的统计章节进行定义。另外，对试验程序（如导入期）启动的所有受试者进行文档记录可能是有用的。该受试者文档的内容取决于特定试验的详细特征，只要可能，至少应收集人口统计学和疾病状态的基线数据。

如果所有随机入组的受试者都满足全部入组标准，完全遵从所有试验程序且无失访，并能提供完整的数据记录，那么要纳入分析的受试者集是显而易见的。试验设计和实施的目标应该尽可能地接近这一理想状态，但实践中却难以达到这一状态。因此，方案的统计章节应该预先阐述可能影响受试者和分析数据的问题。方案还应该说明旨在减少研究实施中任何预期的且可能影响数据分析的不规则问题的程序，这些不规则问题包括各种类型的方案违背、退出和数据缺失。方案应考虑降低这些问题发生频率的方法以及如何解决数据分析中会发生的问题。在盲态审核期间，应确定针对方案违背分析方法可能的修订。最好是根据发生时间、原因及对试验结果的影响来确定任何重大方案违背。方案违背、数据缺失以及其它问题的发生频率和类型应记录在临床研究报告中，并描述它们对试验结果的潜在影响（见 ICH E3）。

关于分析集的确定应遵循以下原则：1）使偏倚减到最小；2）避免 I 类错误膨胀。

5.2.1 全分析集

意向性治疗（见词汇表）原则是指主分析应包括所有随机化受试者。遵循该原则需要完成所有随机化受试者的随访以获得研究结局。实践中这一理想状态很难达到。在本文件中，术语“全分析集”被用来描述尽可能完整的分析集，即尽可能接近包括所有随机化受试者的意向性治疗的理想状态的分析集。在分析中保持初始随机化对于防止偏倚以及为统计检验提供可靠基础是很重要的。全分析集的使用为许多临床试验提供了一种保守策略。许多情况下，它也可以提供处理效应的估计，这些估计更有可能反映了后续临床实践中观察到的效应。

一些有限的情况可能导致将随机化受试者从全分析集中排除，包括未能满足主要入组标准（入选标准违背），未服用过至少一次试验药物以及缺乏随机化后的任何数据。这些排除应是合理的。只有在以下情况下，未能满足入组标准的受试者可从分析中排除而不会引入偏倚:

（1）在随机化之前评判了入组标准;

（2）入选标准违背可以被完全客观地评价;

（3）所有受试者都接受相同的入选标准违背审查;（在开放试验中或者甚至在双盲试验中，如果在审查之前数据被揭盲，相同的审查就很难保证，所以要强调盲态审核的重要性。）

（4）排除所有确定为特定入组标准违背者。

某些情况下，从所有随机化受试者集中排除任何未服用试验药物的受试者可能是合理的。例如，是否开始治疗的决定并不受已知晓所分配治疗的影响，即使排除了这些患者，但意向性治疗原则仍得以遵守。其他情况下，可能需要从所有随机化受试者集中剔除任何随机化后无数据的受试者，除非来自这些特定排除的潜在偏倚或任何其它偏倚得到解决，否则任何分析都不是完整的。

当使用受试者全分析集时，随机化后发生的方案违背可能会对数据和结论产生影响，特别是如果它们的发生与处理分配相关时。大多数情况下把这些受试者的数据纳入分析是合适的，这符合意向性治疗原则。接受一次或多次剂量后退出治疗且以后未提供数据的受试者，或失访的受试者，导致了特殊问题的产生，因为不把这些受试者纳入全分析集中可能会破坏这个原则。这种背景下，受试者无论因任何原因失访，其已经获得的、或根据方案中规定的评价时间点随后收集到的主要指标测量数据，都是有价值的。在主要指标是死亡率或严重疾病发病率的研究中，后续数据的收集尤为重要。如何收集此类数据应在方案中描述。从末次观察值结转方法到复杂数学模型的填补技术可尝试用于替代缺失值。用于确保全分析集中每个受试者主要指标测量值可利用的其它方法，可能会要求做出关于受试者结局或更简单的结局（如成

功或失败）的一些假设。任何策略的使用都应在方案的统计章节中进行描述并说明合理性，并且所用的任何数学模型所依据的假设均应解释清楚。证实相应分析结果的稳健性也同样重要，特别是所考虑的策略本身可能会导致处理效应有偏估计的情况。

由于一些问题的不可预测性，有时把不规则问题应对方法的详细考虑推迟到试验结束对数据进行盲态审核时可能更可取，如果这样做则需要在方案中加以说明。

5.2.2 符合方案集

受试者的“符合方案”集，有时被称为“有效病例”、“有效性”样本或“可评价的受试者”样本，被定义为全分析集的受试者中对方案更具依从性的子集，并且以符合如下标准为特征：

（1）完成了对治疗方案的某个预先设定的最小暴露量；

（2）可以获得主要指标的测量值；

（3）无任何重大方案违背，包括入组标准违背。

在揭盲之前，应该按照适合于特定试验情况的方式完整定义并记录将受试者排除在符合方案集之外的确切原因。

使用符合方案集可能有最大的机会使新的治疗在分析中显示出额外的有效性，而且最紧密地反映方案中的科学模型。然而，相应的假设检验和处理效应估计可能保守也可能不保守，这取决于试验本身；对研究方案的依从性可能与处理和结局有关，它可能会导致偏倚甚至是严重的偏倚。

应充分识别和总结导致剔除受试者以生成符合方案集和其它方案违背的问题。相关的方案违背可能包括处理分配的错误、使用禁忌药物、依从性差、失访和数据缺失。从发生频率和发生时间方面评估各处理组间这些问题的模式是一种良好实践。

5.2.3 不同分析集的作用

一般说来，证明主要试验结果对选择不同受试者集具有不敏感性是有利的。在确证性试验中，计划对全分析集及符合方案集都进行分析通常是恰当的，这样可以明确地讨论和解释它们之间的任何差异。某些情况下，需要深入探讨用于分析的受试者集的选择对结论的敏感性。当全分析集和符合方案集得出实质上相同的结论时，会增加试验结果的可信度，但应注意，对于排除了大比例受试者的符合方案分析会给试验的整体正确性带来一些疑虑。

在优效性试验（试图验证研究产品更优）和等效性或非劣效性试验（试图验证研究产品具有可比性，见第 3.3.2 章节）中，全分析集和符合方案集发挥的作用不同。在优效性试验中，全分析集用于主分析（除了例外情况），因为它倾向于避免符合分析集所导致的对有效性的过度乐观估计，因为包含在全分析集中的非依从者一般会降低所估计的处理效应。然而，在等效性或非劣效性试验中，使用全分析集一般不保守，应非常仔细地考虑它的作用。

5.3 缺失值及离群值

缺失数据是临床试验中的一个潜在偏倚来源。因此，应尽一切努力满足方案对数据收集和管理的所有要求。然而，现实中几乎总会有一些缺失数据。虽然如此，只要缺失数据的处理方法合理，尤其是在方案中预先定义了这些方法，则试验可以被认为是可靠的。在盲态审核期间，可以更新统计分析计划，完善这些方法的定义。遗憾的是，没有可推荐的普遍适用的缺失数据处理方法。应该对缺失数据的处理方法做敏感性研究，特别是当缺失数据的比例较大时。

应采用类似的方法探索离群值的影响，它们的统计定义在某种程度上是主观的。只有从医学上和统计上都认为是合理的，把某一特定值明确地确定为异常值才最具说服力，而且医学方面通常会定义适当的操作程序。在方案或统计分析计划中预先设定的有关离群值的程序应当不倾向任何处理组。同样，在盲态审核期间可以有效地更新这方面的分析。如果在试验方案中未预先规定应对离群值的程序，则需要在对实际值做一次分析的同时，至少进行一次排除或减少离群值效应的分析，并讨论它们的结果之间的差异。

5.4 数据转换

最好在试验设计期间基于早期临床试验的类似数据，在分析前做出对关键指标进行转换的决定。应该在方案中对数据转换（如平方根转换、对数转换）进行详细说明，并叙述

基本原理，尤其是主要指标。在标准教材中可以找到进行数据转换的一般原则，可确保满足统计方法所依据的假设，而且在许多特定的临床领域已经形成了针对特定指标的惯例。是否以及如何对指标进行转换的决定应该受到对于刻度喜好的影响，以便于临床解释。

类似的考虑也适用于其他衍生指标，例如，自基线变化值、自基线变化百分比、重复测量的“曲线下面积”或两个不同指标的比值。应仔细考虑后续的临床解释，并在方案中说明衍生的合理性。与此密切相关的要点参见第 2.2.2 章节。

5.5 估计、置信区间及假设检验

为满足试验的主要目的，应该在方案的统计章节中详细说明待检验的假设和/或待估计的处理效应。用于完成这些任务的统计方法应当针对主要指标（以及优选的次要指标）进行描述，并明确所依据的统计模型。只要有可能，处理效应的估计应伴有置信区间，并确定其计算方法。应当说明使用基线数据以提高精度或以潜在基线差异校正估计值的任何意图，例如，使用协方差分析进行校正。

重要的是，要阐述清楚将使用单侧还是双侧统计检验，如果使用单侧检验一定要事先充分说明其合理性。如果认为假设检验不适用，那么应该给出获得统计结论的替代过程。关于单侧或双侧推断方法的问题是有争议的，在统计文献中可以找到各种各样的观点。在监管背景下，更可取的方法是

将单侧检验的 I 类错误设置为双侧检验中使用的传统 I 类错误的一半，这样就保持了与双侧置信区间的一致性。双侧置信区间通常适合于估计两种处理间差异的可能大小。

所选择的特定统计模型应当反映人们对待分析指标以及试验的统计设计在医学和统计方面的目前认识状态。应充分说明在分析中待拟合的所有效应（例如在方差模型分析中），并应解释根据初步结果对这些效应进行修改的方式（如果有）。同样的考虑也适用于在协方差分析中所拟合的协变量集合（见第 5.7 章节）。在选择统计方法时（如参数和非参数方法），应注意主要和次要指标的统计分布，其分析结果应包含处理效应量的统计估计值及置信区间（显著性检验除外）。

应当清楚地区分主要指标的主分析与主要或次要指标的支持性分析。在方案的统计章节或统计分析计划中，除主要和次要指标外还应阐明数据的汇总和报告方式的大纲。为了在一系列试验中实现分析一致性的目的，例如对于安全数据，应当包括所采用方法的介绍。

对于已知的药理学参数、单个受试者的方案依从程度或其它生物学基础数据，整合这些信息的建模方法可以洞察实际或潜在有效性的价值，特别是对于处理效应的估计。应始终清晰地确定这些模型所依据的假设，并仔细描述任何结论的局限性。

5.6 显著性及置信水准的调整

当存在多重性时，用于临床试验数据分析常用的频率派方法可能需要对 I 类错误进行调整。多重性可能来源于多个主要指标（见第 2.2.2 章节）、处理的多重比较、随时间的多次评价和/或期中分析（见第 4.5 章节）。在可行的情况下，避免或减少多重性的方法有时更可取，例如，在多个指标中确定一个关键主要指标，在多重比较中选择一个关键的处理比较，对于重复测量使用汇总测量如“曲线下面积”等。在确证性分析中，除采取此类步骤，对多重性的其余任何解决办法也应当在方案中确定。应始终考虑多重性的调整，并应在分析计划中交待任何调整程序的细节，或者解释不必调整的理由。

5.7 亚组、交互作用及协变量

除处理之外，主要指标通常系统性地与其它影响因素相关。例如，它可能与年龄和性别等协变量相关，或者比如多中心试验中不同中心接受处理的受试者这样的特定亚组之间可能存在差异。有些情况下，对协变量影响的调整或者对亚组效应的调整是分析计划中不可缺少的部分，因此应在方案中阐明。应通过试验前的缜密考虑，确定这些协变量以及预期对主要指标有重要影响的因素，并考虑在分析中如何处理，以提高精度和补偿处理组之间的任何不平衡。如果使用一个或多个因素进行分层设计，那么在分析中应考虑这些因

素。当不确定调整的潜在价值时，通常建议主要关注未调整的分析，把调整分析作为支持性分析。应特别注意中心效应和主要指标基线值的作用。不建议在主分析中校正随机化后测量的协变量，因为它们可能受到处理的影响。

处理效应本身也可能随亚组或协变量而变化，例如，处理效应可能随年龄降低或者可能在特定诊断类别的受试者中更大。某些情况下，预期会产生交互作用或对交互作用有特别兴趣（如老年病学）时，亚组分析或者包含交互项的统计模型因此成为计划的确证性分析的一部分。然而，大多数情况下亚组分析和交互作用分析应当确定为探索性的，即探索所有处理效应的一致性。一般而言，应首先在所讨论的统计模型添加交互项进行分析，辅之以在相关受试者亚组内或者由协变量定义的层内进行额外的探索性分析。对于探索性分析，应谨慎解释其分析结果，仅仅基于探索性亚组分析的治疗有效性（或缺乏有效性）或安全性的任何结论都不太可能被接受。

5.8 数据的完整性与计算机软件的可靠性

分析结果的可信性取决于用于数据管理（数据录入、存储、验证、校正和检索）以及在统计上处理数据的方法和软件（内部和外部编写）的质量和可靠性。因此，数据管理活动应当基于全面和有效的标准操作规程。用于数据管理和统计分析的计算机软件应当是可靠的，并应提供适当的软件测

试过程的文件。

6.安全性与耐受性评价

6.1 评价的范围

在所有临床试验中，安全性和耐受性（见词汇表）的评价是一个重要方面。在早期阶段，这种评价主要是探索性的，并且只对毒性的直接表达敏感，而在后期阶段，可在更大样本量的受试者中更加全面地描述药物的安全性和耐受性特征。后期阶段的对照试验代表了以无偏的方式探索任何新的潜在不良反应的重要方法，即使这些试验在这方面通常缺乏检验效能。

某些试验可针对以安全性和耐受性的优效性或等效性（与其它药物或与研究药物的其它剂量相比）的特定主张为目的进行设计。这些特定主张需得到来自确证性试验的相关证据支持，就像相应的有效性主张需要证据支持一样。

6.2 指标选择与数据收集

在任何临床试验中，选择用于评价药物安全性和耐受性的方法和测量取决于许多因素，包括对与药物密切相关的不良反应的了解，来自非临床和早期临床研究的信息以及特定药物的药效/药代动力学特性的可能结果、给药方式、待研究的受试者类型，以及试验持续时间。有关临床化学和血液学、生命体征、临床不良事件（疾病、体征和症状）的实验室检查通常构成安全性和耐受性数据的主体。发生严重不良事件

以及因不良事件导致治疗终止对于注册是特别重要的（见 ICH E2A 和 ICH E3）。

此外，建议在整个临床试验规划中采用一致的方法来收集和评价数据，以便合并来自不同试验的数据。使用通用的不良事件词典尤为重要。该词典具有一种结构，提供了在三个不同层级上汇总不良事件数据的可能性，即系统-器官分类、首选术语和收录术语（见词汇表）。首选术语通常是汇总不良事件的层级，在数据的描述性展示中，可以汇集属于同一系统-器官分类的首选术语（见 ICH M1）。

6.3 待评价受试者集及数据展示

对于整体安全性和耐受性评价，待汇总的受试者集通常被定义为那些接受至少一个剂量研究药物的受试者。应尽可能全面地从这些受试者中收集安全性和耐受性指标，包括不良事件类型、严重程度、发病和持续时间（见 ICH E2B）。可能需要在特定的亚组人群，如女性、老年人（见 ICH E7）、严重疾病或那些有常见伴随治疗的人群，进行额外的安全性及耐受性评价。这些评价可能需要解决更加特殊的问题（见 ICH E3）。

在评价过程中需要注意所有安全性和耐受性指标，并且在方案中应阐明方法。所有不良事件都应报告，无论它们是否被认为与治疗有关。在评价中应当考虑研究人群中的所有可用数据。应当谨慎地定义测量值的单位和实验室指标的参

考范围，如果在同一试验中出现不同的单位或不同的参考范围(例如涉及一个以上的实验室),则测量值应当被适当标准化，以便统一评价。应预先确定毒性分级量表的使用，并说明合理性。

某种不良事件的发生率通常以经历事件的受试者数量与处于风险中的受试者数量之比来表示。然而，如何评价发生率并不总是显而易见的，例如，根据情况可考虑把暴露的受试者数量或暴露程度（用人年表示）作为分母。无论计算的目的是估计风险还是在处理组之间进行比较，重要的是要在方案中给出定义。如果计划进行长周期治疗，并预期有相当比例的退出治疗或死亡，这一点尤其重要。对于这些情况，应考虑生存分析方法，并计算累积不良事件率，以避免低估的危险。

对于存在大量体征和症状的背景噪声的情况（如精神病试验),在估计不同不良事件的风险时，应考虑对此进行解释的方法。一种方法是利用“治疗引发事件”（见词汇表）的概念，即只有当不良事件出现或相对于治疗前基线发生恶化时，才记录它们。

减少背景噪声影响的其他方法也许是合适的，如忽略轻度不良事件，或再次随访时观察到的事件才可计入分子。这些方法应在方案中解释并说明其合理性。

6.4 统计评价

安全性与耐受性的研究是一个多维问题。对于任何药物，虽然通常可以预见和监测到某些特定不良反应，但由于可能的不良反应范围非常大，新的和不可预见的反应总可能出现。此外，在违背方案之后经历的不良事件可能引入偏倚，如使用违禁药物。这个背景使药物安全性和耐受性的统计分析和评价变得困难，并且意味着来自确证性临床试验的结论性信息是一种例外而不是通例。

大多数试验中，应用数据的统计描述方法，辅以有助于解释的置信区间计算，是说明安全性和耐受性的最好方法。利用图示方法表达处理组间和受试者间不良事件的模式也有价值。

计算 P 值有时是有意义的，无论作为评价有关特定差异的辅助手段，还是作为“标记”符号以引起对大量安全性与耐受性指标所出现差异的进一步关注。这对于实验室数据尤其有用，否则可能难以适当地进行汇总。建议对实验室数据既要进行定量分析，如对处理组均数的评价，又要进行定性分析，如计算高于或低于某些阈值的比例。

如果使用假设检验，对多重性的统计调整以量化 I 类错误是合适的，但是 II 类错误通常更值得关注。如果未做多重性调整，应谨慎解释常规的统计显著性。

大多数试验中，与阳性对照药物或安慰剂相比，研究者

会试图确定未出现临床上不可接受的安全性及耐受性方面的差异。与有效性的非劣性或等效性评价一样，这种情况下使用置信区间比假设检验更可取，因为置信区间往往可以清楚地显示由低发生率所引起的精度变差。

6.5 综合性总结

在研究产品的开发过程中，特别是在上市申请时，通常会将不同试验的药物安全性与耐受性的特性进行汇总。然而，这样汇总是否可用取决于每一个具有高数据质量的、充分和控制良好的试验。

药物的总体可用性始终是风险与获益之间的平衡问题，单个试验中也可考虑这一观点，即使风险/获益评估通常在整个临床试验的总结阶段进行。（见第 7.2.2 章节）

有关安全性与耐受性报告的更多细节，见 ICH E3 第 12 章。

7.报告

7.1 评价与报告

如引言所述，临床研究报告的结构与内容是 ICH E3 的主题。该 ICH 指南充分地涵盖了统计工作报告并适当整合临床和其它资料，本章节因此相对简短。

如第 5 章节所述，在试验的计划阶段，分析的主要特征应在方案中确定。当试验结束而且数据经整理可供初步检查时，如第 5 章节提到的按计划进行盲态审核是有价值的。在

分析前盲态审核应当包括相关决定，例如，从分析集中排除受试者或数据，可能的数据转换的核查，离群值的定义，将近期其它研究中确定的重要协变量加入模型，参数或非参数方法的重新考虑，等等。这些决定应在报告中加以描述，而且应当与统计师获得处理编码之后做出的决定加以区别，因为盲态下的决定通常会减少产生偏倚的可能性。参与非盲期中分析的统计师或其他人员不应参与盲态审核或修订统计分析计划。数据中如果存在明显的处理诱导效应的可能，将会削弱盲态效果，此时，盲态审核需要特别谨慎。

许多更详细的报告内容和表格应在盲态审核时或盲态审核前完成，以便在实际分析时有一个包括各方面的完整计划，如受试者选择、数据选择与修改、数据汇总与列表、估计与假设检验等。一旦完成数据验证，应按照预先拟定的计划进行分析，越依从于这些计划，结果的可信度越高。应特别注意在方案、方案修订以及基于数据盲态审核更新的统计分析计划中所描述的计划分析与实际分析之间的任何差异。应对偏离计划的分析做出详细解释。

进入试验的所有受试者，无论是否纳入分析，都应在报告中说明。排除在分析之外的所有原因都应记录，还应记录受试者被纳入全分析集但未被纳入符合方案集的原因。类似地，对于纳入分析集的所有受试者，所有重要指标的测量值在所有相关时间点都应该进行说明。

应仔细考虑受试者或数据的所有缺失、退出治疗和重要方案违背对主要指标的主分析的影响。应确定失访、退出治疗或严重方案违背的受试者，并对他们进行描述性分析，包括他们缺失的原因及其与处理和结局的关系。

描述性统计是报告不可缺少的部分。合适的表格和/或图示应清楚地说明主要和次要指标、关键预后指标和人口统计学指标的重要特征。应特别仔细地描述与试验目的有关的主分析的结果。当报告显著性检验结果时，应报告精确的 P 值（如“P=0.034”）而不是参考临界值。

尽管临床试验分析的主要目标是回答其主要目的提出的问题，但在非盲分析过程中，基于观察数据的新问题很可能会出现，随之可能需要额外的或许复杂的统计分析。报告中应严格区分这种额外工作与方案中计划的工作。

对于计划分析中未被预先定义为协变量但仍然具有某些预后重要性的基线测量，机会作用可能会导致它们在处理组间出现无法预料的不均衡。最好的解决办法是，证明针对这些不均衡进行校正的补充分析得出了与计划分析基本相同的结论。否则，应讨论这种不均衡对结论的影响。

一般而言，应少用计划外分析。当认为处理效应可能随某个或某些其他因素而变化时，会用到计划外分析，比如会尝试确定特别获益的受试者亚组。众所周知，计划外亚组分析有过度解释的潜在风险（见第 5.7 章节），应谨慎避免。虽

然当受试者亚组中未显示出获益或具有不良反应时会出现类似的解释问题，但应该恰当地评价这些可能性并予以报告。

最后，应根据临床试验结果的分析、解释及展示做出统计判断。为此，试验统计师应是负责临床研究报告的小组成员之一，还应批准临床报告。

7.2 临床数据库的总结

上市申请需要对所有报告临床试验的安全性和有效性证据进行全面总结和综合（欧盟的专家报告、美国的综合总结报告、日本的概要），在适当的时候还可能伴随结果的统计汇总。

总结中有一些特定的统计关注的领域：描述在临床试验项目过程中受试人群的人口统计学和临床特征；通过考虑相关（通常有对照组）试验的结果并强调它们相互印证或矛盾的程度来解决有效性的关键问题；对于其结果有助于上市申请的所有试验，总结从它们的合并数据库中可获得的安全信息，并确定潜在的安全问题。在设计临床项目中，应认真关注测量的统一定义和收集，这将有助于随后一系列试验的解释，特别是如果不同试验之间的测量可能被合并时。应该选择和使用可记录用药细节、病史和不良事件的通用词典。对主要和次要指标采用通用定义几乎总是有价值的，这对 meta 分析极为重要。关键有效性指标的测量方式、相对于随机化/入组的评价时机、方案违背和偏离的应对以及可能的预后因

素定义都应该保持一致，除非有合适的理由不这么做。

应当详细描述用于不同试验之间数据合并的任何统计程序。应注意与试验选择有关的偏倚的可能性、试验结果的同质性、以及各种变异来源的恰当建模。应探索结论对假设和选择的敏感性。

7.2.1 有效性数据

单个临床试验的样本量应该总是大到足以满足其目的的程度。通过总结一系列解决基本相同的关键有效性问题的临床试验，也可以获得额外的有价值的信息。为了便于比较，应该以相同的形式，通常是关注于估计值和置信限的表格和图形，呈现一系列试验的主要结果。使用 meta 分析技术来合并这些估计值常常是一个有用的补充，因为它允许对处理效应量生成更精确的总体估计，并提供完整而简明的试验结果总结。在一些特殊情况下，meta 分析方法也可能是通过整体假设检验提供充分的有效性整体证据的最适当方式，或者唯一方式。当用于此目的时，meta 分析应该有它自己的前瞻性书面方案。

7.2.2 安全性数据

在总结安全性数据时，重要的是要彻底检查安全性数据库，以寻找潜在毒性的任何迹象，并通过寻找相关的支持性观察模式来跟踪这些迹象。将人暴露于药物的所有安全数据进行合并，能提供重要的信息来源，因为较大的样本量能提

供发现更罕见不良事件的最佳机会，并且可能提供估计罕见不良事件近似发生率的最佳机会。然而，由于缺乏对照组，难以评价来自该数据库的发生率数据，来自对照试验的数据在克服这种困难方面特别有价值。应合并具有相同对照组（安慰剂或特定阳性对照）的研究的结果，并分别展示每个提供充足数据的对照组的结果。

所有通过数据探索发现的潜在毒性的迹象都应报告。评价这些潜在不良反应的现实情况应考虑到由于多次比较而产生的多重性问题。还应适当地使用生存分析方法进行评价，以探索不良事件的发生率与暴露时间和/或随访时间的潜在关系。应适当地量化确定的不良反应的风险，以便正确评价风险/获益关系。

词汇表

贝叶斯方法

是指为某些参数（如处理效应）提供后验概率分布的数据分析方法。后验概率分布由该参数的观测数据和先验概率分布衍生而来，被用作统计推断的基础。

偏倚（统计的和操作的）

是指与临床试验的设计、实施、分析和结果评价有关的任何因素导致的处理效应估计值偏离其真实值的系统趋势。由实施偏离所引入的偏倚称为“操作”偏倚，而上述其他来源的偏倚称为“统计”偏倚。

盲态审核

是指在试验完成（最后一位受试者的最后一次观察）到揭盲这段时间内对数据的检查和评价，旨在最终确定分析计划。

内容效度

是指一个指标（如量表）测量它所预期测量的内容的程度。

双模拟

是指在临床试验中当两种处理不能做到完全相同时，使处理实施仍能保持盲态的一种技术。先准备处理 A（阳性药和不能区分的安慰剂）和处理 B（阳性药和不能区分的安慰剂），然后受试者接受两套处理：A（阳性药）和 B（安慰剂），

或者 A（安慰剂）和 B（阳性药）。

脱落

是指临床试验的受试者由于任何原因不能继续按研究方案进行到所要求的最后一次随访。

等效性试验

是指主要目的为证实两种或多种处理的应答差别无重要临床意义的试验。通常以真实的处理间差异落在临床上可接受的等效性界值上下限之间来证实等效性。

频率派方法

是指在假设重现相同实验情境时，用某些结局的发生频率做出解释的统计方法，例如显著性检验和置信区间。

全分析集

是指尽可能接近符合意向性治疗原则的理想的受试者集。该数据集是从所有随机化的受试者中以最少的和合理的方法排除受试者后得到的。

可推论性，推论

是指将临床试验的发现从参与试验的受试者可靠地外推到更广泛的患者人群和临床环境的程度。

全局评价指标

是指将客观指标和研究者对受试者的状态或状态变化的总体印象综合起来所设定的一个单一指标，通常是一个有序分类量表。

独立数据监查委员会（数据和安全监查委员会、监查委员会、数据监查委员会）

独立数据监查委员会由申办方设立，职责是定期评价临床试验进度、安全性数据以及关键有效性终点，并向申办方建议是否继续、修改或终止试验。

意向性治疗原则

是指基于受试者的治疗意向（即计划的治疗方案）而不是实际给予的治疗进行评价的原则，该原则可以对治疗策略的效应做出最佳评价。它的结果是，分配到每一个处理组的受试者即应作为该组的成员被随访、评价和分析，无论他们是否依从于所计划的治疗过程。

交互作用（定性和定量）

是指处理间的比较（如研究产品与对照之间的差异）依赖于另一因素（如中心）的情况。定量交互作用是指该因素的不同水平之间在量的比较上有差异，而定性交互作用是指比较结果至少在该因素某一水平上显示方向不同。

评价者间信度

是指不同评价者在不同场合使用评价工具时产生相同结果的可靠程度。

评价者内信度

是指同一评价者在不同场合使用评价工具时产生相同结果的可靠程度。

期中分析

是指正式完成临床试验前，比较处理组间的有效性或安全性所做的任何分析。

Meta 分析

是指来源于针对同一个问题的两个或多个试验的量化证据的规范评价，常见的方法是将各试验的汇总统计量进行统计合并，有时也采用原始数据的统计合并方法。

多中心试验

是指多个研究者在多个场所按同一个方案实施的临床试验。

非劣效性试验

是指主要目的为验证研究产品的应答在临床上不劣于对照（阳性药或安慰剂对照）的试验。

首选术语和收录术语

在分层级医学词典中，例如 MedDRA，收录术语是词典术语的最低层级，以研究者的描述进行编码。首选术语是收录术语的分组层级，通常用于报告发生率。例如，研究者写的是“左臂疼痛”，收录术语编码为：“关节疼痛”，在首选术语层级上报告为“关节痛”。

符合方案集（有效病例，有效性样本，可评价的受试者样本）

是指由充分依从于方案的受试者子集所产生的数据集，

以确保这些数据按照所依据的科学模型可能展现出处理效应。依从性包括以下一些考虑：暴露于处理、可获得测量值以及无重大方案违背等。

安全性和耐受性

医疗产品的安全性是指受试者的医学风险，通常在临床试验中由实验室检查(包括临床生化和血液学)、生命体征、临床不良事件(疾病、体征和症状)，以及其他特殊的安全性检查（如心电图、眼科检查）等来评价。医疗产品的耐受性是指受试者能耐受明显不良反应的程度。

统计分析计划

是指更技术性地和更详细地阐述方案中描述的分析要点的文件，包括对主要和次要指标及其他数据进行统计分析的详细程序。

优效性试验

是指主要目的为显示研究产品的应答优于对照（阳性药或安慰剂对照）的试验。

替代指标

是指在直接测量临床效应不可行或不实际的情况下，用于间接测量临床效应的指标。

处理效应

是指在临床试验中归因于处理的效应。在大多数临床试验中，感兴趣的处理效应通过两个或多个处理间的比较体现。

治疗引发事件

是指出现在治疗期间的、但在治疗前未曾发生或比治疗前明显恶化的事件。

试验统计师

是指同时具备丰富的教育/训练和经验，可以实施本指南中的原则并负责临床试验统计方面的统计师。