E9

人用药品注册技术要求
国际协调会

ICH 三方协调指导原则

临床试验的统计学指导原则

ICH 指导委员会
1998 年 2 月 5 日
ICH 进程第四阶段推荐采纳

该指导原则由相应的 ICH 专家小组制定，按照 ICH 进程，已递交管理部门讨论。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。

目　录

临床试验的统计学指导原则

1. 引言

1.1 背景与目的

药品的有效性和安全性需要由临床试验来论证。临床试验需遵循ICH在1996年5月1日通过的“临床试验管理规范(Good Clinicd Practice，GCP)：联合指导原则”（ICH E6）。统计学在临床试验设计与分析中的作用要点在ICH指导中已阐明。由于统计学研究在临床试验中的不断发展，加以临床研究对药物注册及一般保健工作的重要作用，使得有关临床试验统计方面的简洁文件变得十分必要。本指导的目的在于协调欧洲、日本和美国在进行药品上市申请的临床试验时所应用的统计学方法的原则。

作为起点，本指导运用了CPMP(Committee for Proprietary Medicinal Products，医疗制品专卖委员会)在题为《医药制品申请上市许可的临床试验中的生物统计方法》(Biostatistical methodology in clinical trails in applications for marketing authorizations for medicinal products，1994年12月)中的指导意见，并参照了日本健康与福利部的《临床研究中的统计分析指导》(1992年3月)，及美国食品与药品管理局(Food and Drug Administration，FDA)的《新药申请中临床与统计部分的格式与

内容指导》(1988 年 7 月)的有关内容。涉及统计学原理与方法的有关论题也包含在ICH的其他指导中,特别是下列材料之中。

E1A：参与临床安全性评价试验的人群暴露程度

E2A：临床安全性数据管理：快速报告的定义与标准部分

E2B：临床安全性数据管理：传送个体病例安全性报告的数据要素

E2C：临床安全性数据的管理：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：药品注册所需的量效关系资料

E5：影响接受国外临床资料的种族因素

E6：临床试验管理规范

E7：特殊人群的研究：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：制订规章用的医学术语的标准化

M3：药物人体临床试验中的非临床安全性研究

本指导旨在为申办者所研究药物的整个临床发展阶段的临床试验中如何进行设计、实施、分析和评价提供指导。亦有助于在临床试验晚期阶段负责准备申请书或者评价有效性和安全性证据的科学工作者。

1.2　范围与说明

本指导专门论述统计学原理，不涉及具体的统计步骤或方法。保证原则得到正确实施的具体步骤是申办者的职责。本指导对临床试验中的资料综合亦作了讨论，但不作为重点。有关数据处理及临床试验监察活动的原则及步骤已在ICH指导的其他部分论述，此处不赘。

本指导应当对广泛的各学科的专业人员有用。然而，正如在“临床试验管理规范的ICH E6”中所指出的，所有与临床有关的统计工作的具体责任将由有相当资格的且富有经验的统计学专业人员负责。试验统计学专业人员（见词汇）在与其他临床试验专家合作时，其作用和职责为确保用于药物开发的临床试验中统计学原理的恰当应用。因此，试验统计学专业人员应在这方面受过良好的培训，并具有丰富的经验执行本指导中的原则。

在每一个为上市申请而做的临床试验中，所有有关设计、执行和拟采用的分析方法的主要细节等均应在试验开始前所写的试验方案中阐明。按照试验方案中各步骤执行的依从程度，以及对事先的计划进行主要分析，将有助于提高最终结果和试验结论的可信度。研究计划的制订及修改必须经负责人员的批准，包括试验统计学专业人员。参与研究的试验统计学专业人员要保证研究方案以及修订方案中清楚正确地覆盖所有相关的统计学问题，并使用恰当术语叙述。

本指导所阐述的原则首先主要用于临床试验后期，大多数是疗效的验证性试验。除有效性之外，验证性试验可以用安全性变量作为其主要变量(不良事件、临床实验室变量、心电图数据等)，或以药效、药代动力学变量为主要变量(如验证生物等效性的试验)。其次，有些确定性的结果要从各个研究的资料中综合而得，本指导中有些原则可用于这类情况。最后，尽管在早期的药物试验阶段，在本质上主要是探索性的临床试验，统计学原则也与这类临床试验有关。因此，本文件的内容应尽可能用到临床试验的所有各个阶段。

本指导中所阐述的很多原理涉及到使偏倚(bias)最小和使精度(precision)最高。偏倚一词在本指导中是指，临床试验中任何与设计、执行、结果的分析与解释等有关的因素，导致处理效应的估计值与真值偏离的系统性倾向。重要的是尽可能完全地验明导致偏倚的可能来源，以便事先采取措施限制这些偏倚。偏倚的存在将严重危及从临床试验中得出正确结论的能力。

有些偏倚源于试验的设计，例如，将病情较轻的病人系统地分配到一个组中的分配方案。还有些偏倚来源于临床试验的执行过程和资料的分析，例如，根据对病人结局的了解，规定违背试验方案标准及从分析中剔除病人，这些都可能是对处理效应的正确评价产生偏倚的来源。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就显得很重要。稳健性(robustness)概念是关于总体结论对于数据、假设和数据分析方法的各种局限性的敏感性。稳健性意味着对

不同假设条件下或者不同的分析方法进行的分析对所得处理效应和试验的主要结论没有实质性的影响。统计学上对处理效应、处理间的比较的不确定性统计度量的说明应包含偏倚对 P 值、可信区间或统计推断的影响。

由于在试验设计和分析时通常选用频率统计方法，因此在讨论假设检验和/或区间估计时，本指导主要使用频率法。这并不意味其他方法不可取，如理由充分，且所得结论是相当稳健的，则贝叶斯（Bayes）方法及其他方法亦可考虑。

2. 整个临床试验中需考虑的问题

2.1 试验内容

2.1.1 开发计划

新药临床试验的主要目标是寻找药物是否存在既安全又有效的用法与用量，在此范围内，风险利益比是可接受的，同时还要确定可能由该药得到好处的特定对象及使用适应证。

为满足这一总目标需要一个临床试验的流程，每一步均有特定的目标(参见 ICH E8)。这需要在临床研究计划中或一系列计划中阐明，这些计划中具有适当的决策点，并且需有灵活性，以便随认识的提高而对其进行修订。每一个上市申请均需清晰地描述其计划的主要内容，以及每次试验的作用。对整个试验所提供的证据的解释和评价包含了对每次试验提供的证据的综

合(见 7.2)。若能保证在每次试验中采用一些约定标准，如医学术语的标准化、主要测量的定义与时间表、处理偏离试验方案的方法等，将有助于对各次试验的综合。当多个试验中都论及同样的医学问题时，则统计学上的概括或后期综合分析(meta analysis，见词汇)将提供更丰富的信息。可能的话，应预先在计划中加以考虑，使相应的试验得到明确定义，指明设计的共同特点。可能影响共同计划中的试验的其他主要统计学问题(如果有的话)亦需在计划中陈述。

2.1.2 验证性试验

验证性试验(confirmatory trial)是一种事先提出假设，并对其进行检验的有对照的试验。通常，通过验证性试验，提供有效性及安全性的有力证据。在这类试验中，根据试验的主要目的,提出并事先定义假设，在试验完成后进行检验。在验证性试验中，准确估计研究所取得的效果大小和把这些效果与临床意义联系起来是同样重要的。

验证性试验主要是对所提出的假设提供坚实的依据,因此，坚持按试验方案及标准操作步骤进行试验尤为重要；一些不可避免的改变需要给予解释并提供书面材料，并检查由此所产生的影响。对每个这类试验方案的合理性，以及其他主要的统计方面如分析计划的主要特点等均需在试验方案中陈述。每个试验只能提出解决少量的问题。

寻找有力的证据支持所提出的假设，需要用验证性试验，以说明所开发的药物对临床是有益的。所以，验证性试验必须就提出的有关安全性及有效性的每一个关键性的临床问题给予充分的回答。另外，重要的是，把结果推论到所研究的病人的总体的基础要明白易懂，并给予解释，这也会影响到所需的中心和/或试验的数目或类型（例如专家或全科医师)。验证性试验的结果必须稳健。在某些情况下，单一的一个验证性试验所提供的依据就足够了。

2.1.3 探索性试验

验证性试验的必要性及其设计几乎总是基于一系列探索性试验的早期临床工作。像所有的临床试验一样，这些探索性研究也应有清晰和明确的目标。但与验证性研究相比，探索性试验的目的并不总是对预先提出的假设进行简单的检验。此外，探索性试验有时需要采用更为灵活可变的方法进行设计，以便根据逐渐积累的结果对试验进行适当的修改。其分析可能仅限于对数据进行探索性分析，可能要作一些假设检验，但假设是根据数据的特点而定的。虽然，这类试验对整个有效性验证有贡献，但不能作为证明有效性的正式依据。

单个试验往往同时具有探索和验证两方面。例如，在大多数验证性试验中，常对资料进行探索性分析，作为解释并支持

它们发现的基础，并为后续研究提出进一步的假设。试验方案中须明确区分探索和验证这两方面的内容。

2.2 试验范围

2.2.1 总体

在药物开发的早期阶段，临床试验研究对象的选择在很大程度上受到一种主观愿望的影响，这种愿望是希望最大可能地观察到期望的临床疗效，因此，研究对象往往是病人总体中很局限的、最容易显示疗效的一小部分。但在验证性试验阶段，试验对象须更具代表性。因此在这些试验中，在保持研究对象的同质性以便精确估计处理效应的条件下，尽可能地放宽纳入和排除标准。由于地理位置、研究时间，以及特定的研究者和医疗单位的医疗实践等因素的影响，没有一个单一的临床试验可望能完全代表将来的使用者。尽管如此，我们应尽可能减少上述因素的影响，并在对试验结果的解释中加以讨论。

2.2.2 主要变量与次要变量

主要变量[又称目标变量(target variable)、主要终点(primary endpoint)]是能够就试验的主要目的提供与临床最有关的且可信的证据的变量。通常主要变量只有一个。因大部分验证性试

验的主要目的是提供与有效性相关的强有力的科学证据，所以主要变量通常是一个有效性变量。安全性与耐受性也可以是主要变量，而且常常是一个重要考虑的内容。有关生活质量和卫生经济的测量值也是进一步可能考虑的主要变量。主要变量的选择应考虑相关研究领域已有的公认的准则和标准。建议使用在早期的研究中或在已发表的文献中报道的已累积有实践经验的可信且有效的变量。所选的主要变量要有充分的证据说明它对根据入选标准和排除标准规定的总体中的病人，能高效且可信地反映临床上相关的且重要的临床疗效。主要变量通常应当是用于样本量估计的变量(见 3.5)。

在很多情况下，评价受试病人结局的方法并不是很简单的，需认真考虑确定。例如，将死亡率选作主要变量而无进一步的说明是不合适的；对死亡率的评价可以比较某时点尚存人数的比例，也可以比较在某时域内的生存时间的总的分布。另一个常见的例子是复发事件，疗效变量可以是简单的二分类变量(任何指定时域内的复发)，也可以是第一次复发的时间、复发率(在单位时间内的复发数)等等。在慢性病的疗效研究中，治疗时间中功能状态的评价又给选择主要变量提出了新的要求。这种评价可通过诸如比较开始和结束时机体的功能状态，比较整个试验期内所有观察结果计算得的斜率，比较超过或低于指定界值的病人的比例等方法来进行。由于主要变量将用于统计分析，因此，为避免因事后定义所引起的复杂性，在设计方案时精确定义在统计分析中将使用的主要变量至关重要。另外，对所选

的特定主要变量的临床相关性和有关的测定过程的有效性一般都应在设计方案中加以说明并证明其正确性。

主要变量及其选择理由均应在设计方案中加以说明。在揭盲后重新定义主要变量是不可接受的，因为由此所产生的偏倚很难判断。当根据主要研究目的所确定的临床疗效可由多种方法测定时，只要实际情况可行，在设计方案中，应根据测量方法的临床相关性、重要性、客观性和/或其他相关特性确定一种测定值作为主要变量。次要变量是与主要目的相关的支持性的变量，或与次要目的相关的疗效变量。在设计方案中对次要变量进行预先定义，对这些变量在解释试验结果时的作用及其相对重要性加以说明都是重要的。

次要变量是与主要目的有关的支持性测定，或者是与次要目的有关的疗效的测定。在方案中预先给予定义是很重要的，就像说明试验结果的相对重要性和作用的解释一样。次要变量的数目应当是很少的，并且应当是与试验中要回答的有限的问题有关的。

2.2.3 复合变量

如果从与主要目的有关的多种测定中不能选择一个单一的主要变量，另一种方法是应用预先确定的算法来结合或组合多个测定值，使其组成一个单一的或“复合的变量”。事实上，主要变量有时以多种临床变量相结合的复合变量的形式出现(如在关节病、精神障碍及其他疾病中的评分尺度)。该法虽涉及到

多重性问题，但不需对 I 类错误进行调整。将多种测定相结合的方法应在设计方案中加以指定，且应以相当于相关临床疗效的大小对联合变量的尺度进行说明。当复合变量被用作主要变量时，组成这个复合变量的每一个变量，当有临床意义并有效时，有时也进行单独分析；当量表评分被用作主要变量时，对含义的有效性(见词汇)，评定者内和评定者间的可靠性(见词汇)及查出疾病严重程度的变化的灵敏性等因素加以说明是重要的。

2.2.4 全局评价变量

在有些情况下，用全局评价变量(见词汇)来评价某个治疗的总体安全性、有效性和实用性。这种变量是客观变量与调查者对试验对象的状态或状态的改变程度总的印象的有机结合，它常常是一个有序的等级。总体有效性的全局评价方法已经在一些治疗研究领域建立，如神经科和精神科。全局评价变量一般都有一个主观成分。使用全局评价变量作为主要或次要变量，需要在试验方案中对尺度的以下方面进行详细说明：

(1)全局变量与试验主要目的的相关性；

(2)尺度的有效性和可靠性的基础；

(3)如何根据单一试验对象所搜集的试验数据，将其按全局评价变量归为尺度的一类；

(4)如何将有缺失数据的试验对象归为尺度的一类，或用其他方法评价。

如果研究者在用全局评价变量进行疗效评价时，对客观变量加以考虑，则这些客观变量应作为附加主要变量，或至少是重要的次要变量加以考虑。

全局有用性是综合了效应与危险因素后得出的，它也反映治疗医生的决策过程，医生在决定用药时必须权衡使用这些药物利弊。使用全局有效性评价方法的一个问题是，对具有不同有益作用和有害作用的两种药物合判为相同的疗效。例如，在判断一种治疗的全局有用性为等同或优于另一种治疗时掩盖了其无效或疗效很差而同时不良反应较少的事实。因而，不主张用全局有用性作为主要变量。如果全局有效性被用作主要变量，则对特定的有效性和安全性单独作为附加主要变量进行考虑是重要的。

2.2.5 多个主要变量

有时，会希望使用多个主要变量，每一个(或其中一部分)都足以包括治疗的有效范围。应当清楚地说明计划使用这种方法的证据。例如，应当明确是否影响任何一个变量，最少有几个变量，或全部变量，这被认为是达到试验目的所必须的。关于已定义的主要变量的主要假设或者感兴趣的假设与参数(如均数、百分数、分布)应详细说明，并对统计推论方法加以说明。由于多重性问题的可能，I 类错误的效用应加以解释(见 5.6)，也应该在设计方案中给出控制 I 类错误的方法。在评价对 I 类

错误的影响时，所提出的主要变量之间的相关程度也应加以考虑。如果试验的目的是显示所有指定的主要变量的效果，那就没有必要调整 I 类错误，但是，必须认真考虑对 II 类错误和所需样本大小的影响。

2.2.6 间接变量

如果不能通过观测实际临床效果来直接评价试验对象的临床效应，可以考虑间接变量(间接变量，见词汇)。被普遍接受的间接变量用于一些适应证，这时，间接变量被认为是临床疗效可靠的预测变量。在选用间接变量时，主要考虑两点。第一，它可以不是所感兴趣的临床结果的真正预测因子。例如，它可能是测定了与一个特定的药理学机制有关的处理效应，但不能提供处理作用范围与最终效应的全部信息，无论是阳性还是阴性。已有很多例子，原先认为处理对选用的间接变量有高度正效应，但最终结果却被证实试验对象的临床结果是有害的；与此相反，也有一些例子，处理有临床效应，但对选用的间接变量却没有可测量到的影响。第二，选用的间接变量并不能产生可直接与不良反应相权衡的定量的临床效果度量。尽管，如何证实间接变量有效的统计标准已经提出，但如何运用这些标准的经验还相当有限。事实上间接变量证据的强度取决于：(i)相互关系的生物学可能性；(ii)间接变量对临床结局预后判断价值的流行病学研究证据；(iii)从临床试验中获得的有关处理对间接变量影响程度与处理对临床结局影响程度相一致的证据。一个

产品的临床和间接变量之间的关系并不一定能适用于治疗同一种疾病的具有不同作用方式的另一个产品。

2.2.7 分类变量

二分类或其他连续的或顺序的变量的分类有时也是必要的。“成功”和“有反应”的尺度就是这种二分类变量的常见例子，它要求，例如，对一个连续变量以最低改善(相当于基线)百分率或等级变量等于或超过等级尺度中某一阈值(如“好”)，作为精确定义。舒张期血压下降到低于 90mmHg 就是一个常见的二分类。对有明显临床相关性的变量进行分类最有用，由于已知试验结果很容易对分类标准的选择产生偏倚，所以在设计方案中应对分类标准事先作出定义和明确的说明。分类通常意味着要丧失部分信息，由此导致分析的把握度降低，这应当在样本大小估算中说明。

2.3 避免偏倚的设计技巧

在临床试验中，避免偏倚的两个重要设计技巧是盲法和随机化，这些应是上市申请中所包含的临床对照试验的一般特点。大多数这样的试验采用双盲法，在这些双盲法试验中，根据适宜的随机化计划，事先将药物进行包装，在提供给试验中心的药物上标明试验对象号码和服用期，从而使参与试验的每一个

人都不知道哪一个对象服用哪一种药物，甚至不知道编码字母。这种方法大部分将在 2.3.1 和 2.3.2 中的大部分进行讨论，此外的情况在节尾讨论。

在设计方案中，应对旨在尽可能缩小试验进行过程中任何可损害统计分析满意程度的可预见的不正确情况的特定处理过程进行说明。这些不正确情况包括违反试验方案的各种情况、失访和缺失值。方案中应考虑到减少这些问题的频度和在数据分析中出现这些问题时如何处理的方法。

2.3.1 盲法

盲法(blinding)是为了控制在临床试验的过程中，以及对结果的解释时产生有意或无意的偏倚(bias)。这些偏倚来自由于对治疗的了解而对病例的搜集和安排、对病人的照顾、病人对治疗的态度、对终点(end point)的评价、对失访的处理、在分析中数据的剔除等的影响。其根本目的是，在有可能产生偏倚的时候防止知道采用的是何种处理。

双盲试验(double blind tdal)是，所有病人及所有参与治疗或临床评定的申办者及研究人员均不知道谁接受的是何种处理，包括挑选合格病人者、评价结局者或按照设计方案评价依从性者。这种盲法要持续整个试验实施过程，只有当数据整理到其质量能接受的水平时，方可对适当的人员揭盲。如确需要有不参与治疗和临床评价的人知道处理编码(treatment codex)（如生

物分析学家，参与严重不良事件报告的审视员等)，项目申办人必须制定适当的标准操作规程，以防处理编码不适当地扩散。在单盲试验中，研究者和/或其成员知道采用的是何种处理，但病人不知道，或者正相反。在开放性试验中，所有的人均知道采用的是何种处理。双盲是最优方法。需要试验中所采用的处理方法在用药前或用药时无法从外观、味道等方面识别出来，且在整个试验均保持盲法。

要做到双盲，会遇到很多困难：两种处理方法可能完全不同，如手术治疗和药物治疗；两种药物是两种不同的剂型，虽然用胶囊技术可使用两者无法分辨，但改变剂型可能会改变药代动力学或药效学的特性，因此，需要建立剂型的生物等效性；两种药物每天的用法不同，在这些情况下，要实现双盲法，就要采用“双模拟”(见词汇)技术。这一技术有时会使用药计划十分不寻常，以至于对病人的积极性和依从性产生负面影响。伦理上的困难也会干扰其应用，例如，必须进行无用的手术操作。无论如何，应当努力克服这些困难。

由于所采用的处理出现了效果，可能使双盲遭到部分损害。在这种情况下，不让研究者和有关申办人员知道某些检验结果(如某些临床实验室检验结果)，可以使盲法可以得到改善。在非盲法试验中，唯一的或特别的治疗效果可能无法不让病人知道，则可考虑采用类似的方法(见下文)，以使偏倚达到最小。

如果双盲不可行，则应考虑用单盲。在有些情况下，只有开放性试验才可行或符合伦理。单盲和非盲使试验更具灵活性，

但特别重要的是，研究者知道下一个病人接受哪种处理，不应影响对下一个进入研究的病人的纳入；纳入病人最好在知道随机化的处理之前。对这些试验，应考虑用中心化的随机化方法，如用电话通知指定随机化的治疗方法。在这两种情况下，进行临床评判的医务人员应不参与治疗，而且在评判过程中始终处于盲态。在单盲或非盲试验中，应尽最大努力使偏倚来源达到最小，主要变量应尽可能地客观。采用不同程度盲法的理由，以及通过其他方法使偏倚到达最小的步骤，均应在试验方案中说明。例如，申办人应当有适当的标准操作步骤以保证在数据进行分析前的数据库清理过程中，适当地限制对处理编码的接触。

只有在病人的治疗医生认为，为了病人的医疗必须了解病人所接受的处理时，才能对此病人揭盲。任何有意或无意地破盲，不管是什么理由，须在试验结束时报告并给予解释。揭盲的过程及时间亦需在报告中说明。

在这个文件中，数据的盲态核查（见词汇）是指，从试验全部结束（最后一个病人的最后一次观察）到破盲这段时间的数据核查。

2.3.2 随机化

随机化(randomization)为临床试验中的病例接受何种处理引入了一个仔细安排的机遇的要素。在后续的试验资料分析中，它提供了定量评价处理效应的坚实的统计基础，使各处理组的

预后因素、已知的和未知的分布趋于相似。与盲法合用，随机化有助于避免在病例的选择和分组时因处理分配的可预测性而导致可能的偏倚。

临床试验的随机化一览表就是用文件形式写出对病人的处理的随机安排。在最简单的情况下，它是处理（在交叉试验中是处理顺序）的序列表，或者是与病例号相应的编码。有些试验的随机化编码较复杂，如有预筛选的试验，但唯一的事先定义好的对病人的处理分配和处理顺序必须清楚。对于不同的试验，设计编制随机化一览表的过程亦不相同。随机化分配表必须有（如果需要）可以重新产生的能力。

虽然无限制条件的随机化是可接受的，但在区组随机地安排病人的方法更具优越性。这将有助于增加处理组间的可比性，特别当病例的某些特征随时间而变化，如接纳病人的策略的改变。它还更能保证各处理组的样本量几乎相等。在交叉试验中，它提供了一个较高效率且更易于解释的平衡设计的方法。在确定区组的大小时需注意，每个区组要尽可能地小，以防不均衡；又要足够大，以防对区组终末的可预测性。研究者及其他有关人员应对区组的大小保持盲态；用两种或以上的区组大小，每个区组的大小随机决定，可达到同样目的。(理论上，在双盲试验中，可预测性是无关紧要的，但药物显示出的药理上的反应常常给聪明的人提供了猜测机会。)

在多中心试验(见词汇)中，应按中心组织随机化过程。提倡为每一个中心建立一张单独的随机表，也即按中心分层，或

将某几个整的区组分到一个中心。一般，为了使各层趋于均衡，按照基线资料中的重要预后因素（如疾病的严重程度、年龄、性别等）进行分层，有时对促使层内的均衡安排也是很有价值的。很少需要多于两个或三个的分层因素，因为很难达到平衡且很麻烦。倘若试验的其余步骤可以调整以适应的话，应用动态分配(见后述)的方法有利于在一些分层因素中达到平衡。随机化时被分层的因素在以后分析中应加以说明。

下一个被随机化进入试验的病人总是按随机化分配计划（如果是分层随机化，则是相应的层）中的下一个数字接受相应的处理。下一个病例所接受的数字及相应的处理只应在确认该病例进入试验的随机化部分后才分配，使人容易预测的（如区组的长度等）随机化细节不应包含在试验方案中。随机化计划由申办者或一个独立的组织以确保整个试验按盲法进行的方式安全存档。在整个试验中，查阅随机化表必须考虑到在应急情况下对任何病例不得不揭盲的可能性。揭盲所采取的步骤、必要的文件、后续病例的治疗方法的评价均需在试验方案中写明。

动态分配是另一种随机化方法，即病人接受何种处理取决于当前各处理组的平衡情况，在分层试验中，取决于病人所属的层内的平衡情况。应当避免确定性的分配法，应对每一个处理的安排列入适当的随机化成分。必须尽一切努力确保试验是双盲的。例如，处理编码仅限于控制动态分配的中心试验办公室的有关人员才知道，一般通过电话告知。这也容许另外对入

组标准进行考核和入组，这对有些多中心试验是有价值的。然后可采用一般的双盲试验，事先将药物包装并编号，但不必依次选用。最好选用一个合适的计算机算法，不让试验中心办公室的人知道处理的编码。当考虑作用动态分配法时，操作上的复杂性，以及潜在的对分析的影响必须仔细评价。

3. 试验设计中需考虑的问题

3.1 试验类型

3.1.1 平行组设计

最常见的验证性临床试验采用平行组设计(parallel group design)，即将个体随机分配到两个或多个组中的一组，每组分别施以不同的处理。这些处理包括药品的一个或多个剂量、一个或多个对照，如安慰剂或/和阳性对照。该设计所基于的假定与其他大多数设计相比要简单得多。然而，与其他设计一样，其他一些特性会使分析及解释变得困难(例如协变量、重复多次测量、设计因素间的交互作用、违背研究计划、中途退出试验及失访)。

3.1.2 交叉设计

交叉设计(cross-over designs)中的每个个体随机按两个或多个处理的不同顺序安排，是一种自身比较的试验方法。这种

设计策略受到欢迎，主要是因为运用该法可减少观察例数，即达到指定的把握度所需的评定例数，有时减少的幅度还很大。在最简单的 2×2 交叉设计中，每个个体在相继的两个处理时期分别接受两种处理，两个处理时期之间常有一个洗脱期(washout period)。最常见的扩展是在 n(>2)个处理时期接受 n 个不同的处理。在这类设计中，每个个体都接受了 n(>2)个处理，这类设计有多种变异可分解，如每个对象所接受 n(n>2)种处理的一个子集或者处理在一个病人身上重复的设计。

交叉设计有很多问题足以使其结论无效。主要的是延滞作用(carryover)，即每个时期的处理在后继时期中的残余影响。在一个相加模型中，不等的延滞作用将使处理间的直接比较产生偏倚。在 2×2 设计中，从统计学上不能鉴别是延滞作用还是处理与时期的交互作用，且因为相应的对比是个体间的，故这两种效应的检验效能都不高。这一问题在高阶设计中不严重，但不能完全消除。

因此，在进行交叉设计时，最重要的是避免延滞作用。最好是在充分了解疾病与新药的有关知识的基础上有选择地精心设计。所研究的疾病应当是慢性病，且病人处在稳定期。药物的疗效需在处理期内完全发挥出来，洗脱期必须足够长，以使药物的作用完全消退。这些条件是否满足，要利用已有信息及资料在试验前确定。

在应用交叉试验时，还有一些问题需引起密切注意。最主要的是，当有病例失访时，分析和解释变得很复杂。另外，可

能的延滞作用，对在后续的处理时期出现的不良事件，难以判断是何种处理所致。这一问题及其他问题在 ICH E4 已有详细论述。交叉设计一般限于预期仅有少数失访的情形。

2×2 交叉设计的一个常用且满意的应用是，验证同一种药物的两种不同配方的生物等效性。在针对健康志愿者的特定应用中，如果洗脱期足够长，则对相应的药物动力学变量的延滞效应一般不大可能出现。然而，在分析时仍需在所得到的资料的基础上验证这一假设。例如，确认在每一个处理时期的开始无药物可检出。

3.1.3 析因设计

析因设计(factorial designs)是通过处理的不同组合，对两个或多个处理同时进行评价。最简单的是 2×2 析因设计，将研究对象随机分配到两个处理，如 A 和 B 的四种可能的组合之一，即只有 A，只有 B，A 及 B，既无 A 又无 B。在很多情况下，该设计主要用于检验 A 和 B 的交互作用。如果样本量是基于检验主效应计算的，则检验交互作用的统计把握度就会低。当应用该设计检验 A 和 B 的联合效应时，特别是欲将两者同时使用时，这一点需着重考虑。

析因设计的另一个重要应用是，当联合使用处理 C 和 D 时建立剂量反应特性，特别当各自单独使用时某种剂量的疗效已事先确定时。选择 C 的 m 个不同剂量(通常包括 0 剂量，即安慰剂)，D 的相同数目的 n 个剂量。全设计包括 n×m 个处理组，

每个处理组接受一种不同的 C 和 D 剂量的组合。反应曲线的估计结果将有助于在临床上确定 C 和 D 的最合适的剂量组合(见 ICH E4)。

有时，2×2 设计可以用来有效地使用临床试验病人，通过与评价单一处理所需的对象数相同的例数评价两种处理的效果。已经证实，这一策略对很高的死亡率的试验非常有效。这一方法的效率及有效性在于没有处理 A 和 B 的交互作用，因此，A 和 B 对主要效应变量的效应符合相加模型，故无论有无对 B 的附加效应，A 的效应实际是一样的。对于交叉设计，需根据先前的信息和资料在试验前给出能够满足这种条件的证据。

3.2 多中心试验

应用多中心试验(multicenter trial)主要有两个理由。首先，多中心试验是被大家所接受的高效的评价新药的方法；在某些情况下，这是在有限的时间内搜集研究所需的足够的试验例数的实际方法。原则上，这类多中心试验可用在临床试验的各个阶段。它可能只有少数几个中心，但每个中心均有较多的试验例数；而对稀有病例，它可能有很多中心，但每个中心都只有少数几个病例。

其次，将一个试验设计成多中心(多个研究者)试验，可为研究结果的推广与应用提供良好的依据。因为，可能搜集病例的范围广，用药的临床条件广泛，试验结果对将来的应用更具代表性。同时，因参与的研究人员较多，为新药疗效广泛的临

床评判提供了可能。这种试验在药物试验的后期将成为验证性试验，常包含很多研究者和试验中心。有时，为使新药的应用更具广义性(见词汇)，试验可在一些不同的国家进行。

如果想要多中心试验被有意义地说明和外推，则研究方案实施的方式必须是清晰的，且在各中心是一致的。更进一步，样本量与检验效能的计算均假设各中心的处理间差异是相同的量的无偏估计。制订一个共同的试验方案，并以此指导整个试验，这一点很重要。方法应尽可能完全标准化。通过召开研究者的会议、试验前对人员进行统一培训、试验过程中加强监视，可以减少评定标准和方案的不一致。完善的设计应使各中心的各处理组的受试者的分布相同，完善的管理要保持实现这一设计目标。如果以后发现有需要检验处理效应在各中心是否齐同一致，则避免各中心样本数相差悬殊以及个别中心的样本数太少是有益的。因为，这样减小了在估计处理效应时各不同权重估计值间的差别(这一点对所有中心样本数均很少时不适用，此时中心在分析中不考虑)。如果事先没注意这一点，且同时伴有关于结果齐性的怀疑，则严重时，会使多中心试验的价值降低到不能认为申办者的结论提供了令人信服的证据的程度。

在最简单的多中心试验中，每一个研究者将负责一个医院收集的病例，所以，中心是由研究者或医院唯一确定的。然而，在很多试验中，情况要复杂一些。一个研究者要负责几个医院的病例收集；一个研究者代表几个医院的一组临床医生(下属的研究者)，而每个医生负责从各自的或几个相关的医院收集病

例。一旦对统计模型中关于中心的定义有疑问，则方案中(见5.1 节)统计这一节需给出明确定义(例如，按研究者、场所或地区)。在大多数情况下，用调查者定义中心还是满意的，ICH 指导中 E6 提供了这方面的指导。如有疑问，则目的应是以使对主要变量测量值及处理效应有影响的重要因素达到齐同来定义中心。将各中心合起来分析的任何方法均需证明是合理的，并事先在试验方案中明确说明。但无论何时，应用该法必须对处理保持盲态，例如，在对资料作盲态核查时。

用于估计和检验处理效应的统计模型需在方案中阐述。首先，应当用一个考虑到中心间差异的模型研究主处理效应，但不应包含中心-处理的交互作用。如中心间处理效应是齐性的，则在模型中常规地包含交互作用项将降低主效应检验的效能。如中心间的处理效应是非齐性的，则处理效应的解释是有不同意见的。

在某些研究中，如每个中心均只有少数几个病例的大的死亡率试验，没有理由期望中心对主要变量及次要变量有什么影响，因为它们不像会表现出有临床重要性的影响。在其他一些研究中，可能一开始就意识到每个医院只有有限的病例数，在统计模型中包含中心的效应是行不通的。在这些情况下，模型中包含中心项是不合适的，而且，在这种情况下也没有必要按中心随机化。

如果每个中心均有足够数量的样本，而处理效应是阳性的，则将影响结论的广义性，一般均需检验各中心间处理效应的齐

性。对每个中心的结果用图示方法，或用分析方法，如交互作用的统计意义检验，可检验出明显的非齐性。当运用这样一种统计检验时，必须意识到，一般在一个为了检出主效应而设计的试验中，把握度是很低的。

如处理效应非齐性，解释时必须非常谨慎，应力图从试验的管理或病例的特征等其他方面找到解释。这种解释通常会提示合适的进一步的分析和解释。如果无法作出解释，例如，处理效应的非齐性，由明显的量的交互作用(见词汇)所证实，则意味着需要用不同中心予以不同权重的另一种处理效应的估计，使处理效应的估计具有稳健性。尤为重要的是，需了解任何用定性的交互作用(见词汇)显示的任何非齐性的基础，而找不到合理的解释，可能需进一步的临床试验，直到处理效应被可信地估计出来。

至今，我们对多中心试验的讨论基于固定效应模型的应用。混合模型也可被用于探索处理效应的非齐性。这些模型把中心和处理-中心效应看做是随机的，这在中心数大时是特别适当的。

3.3 比较的类型

3.3.1 显示优效性的设计

从科学上讲，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（active contral），或由剂量-反应关系证实效果是最可信的。这类试验称为优效性(superiority)试验(见词汇)，除非有特殊说明，一般情况下，本指导中所述均指优效性试验。

对于严重疾病来说，如果已经由优效性试验证实存在一个很有效的治疗方案，而采用安慰剂对照，就有悖伦理。此时，应考虑合乎科学地采用阳性对照。用安慰剂对照还是用阳性对照应根据各个试验而不同。

3.3.2 显示等效性或非劣效性的设计

在有些情况下，所研究的产品与某处理相比的目的并不是为了证实优效性。这类试验根据研究的目的可分为两大类：一是“等效性”试验（见词汇），二是“非劣效性”试验（见词汇）。

生物等效性试验属第一类。有时，进行临床等效性试验是为了其他法规上的理由。例如，非专利产品与市售产品相比，当混合物不被吸收，血液中无法检出时，需要验证临床等效性。

用很多阳性对照的试验，来说明一个被研究的产品不比阳性对照差，这属于第二类。另一种可能是试验药品的几个不同的剂量与推荐的剂量相比，或与标准药品的几个不同剂量相比。这种试验的目的是同时显示所研究的药物的剂量反应关系，并且以研究的药品与阳性对照进行比较。

阳性对照的等效或非劣效性试验若同时用安慰剂，可以一举多得。例如，是否比安慰剂有效，从而说明研究设计是合适的；同时评价与阳性对照药物的有效性与安全性的相似程度。在使用阳性对照的等效性(或非劣效性)试验中，不与安慰剂合用或不用多个剂量，有很多众所周知的困难。这与缺乏任何判断内部有效性的尺度(与优效性试验相比)有关，从而使得证实外部有效性成为必要条件。等效性试验(或非劣效性试验)本质上不是保守的，因此，设计或执行中的许多缺点使之偏向于得出等效性的结论。因此，这种试验的设计和实施需要特别小心。例如，特别重要的是要尽量减少违反入组标准、依从性不良、退出、失访、数据缺失和其他与设计方案偏离的事件，也要尽量减少在以后分析中的影响。

阳性对照药物要谨慎选择，一个合适的阳性对照，应当是普遍应用的，其相应指征的效果已经由设计良好并且有很好资料的优效性试验所确定和定量，并明确其可以在认真设计的阳性对照试验中表现出相似的效果。为此，新的试验必须与前述有效性试验具有同样的重要设计特点(主要变量、阳性对照的剂量、合格标准等)。在这种试验中，阳性对照药清楚地显示了临床疗效，充分考虑到了与新试验有关的临床和统计实践的进展。

验证等效性或非劣效性的试验设计方案中，包括表述清晰、语意明确是极其重要的。在计划中，必须指定一个等效界值，这个界值是临床上能接受的最大差别，并且应当小于阳性对照的优效性试验所观察到的差异。对阳性对照的等效性试验，需

指定上界和下界；对阳性对照的非劣效性试验只需要下界。等效界值的确定需要由临床上来认可。

统计分析常基于可信区间的应用(见 5.5)。对等效性试验应当用双侧可信区间，当可信区间完全落在等效界值之内，则推断为等效。在运作上，这一方法等价于同时进行两个单侧检验，以检验(复合的) 备择假设为处理效应的差别在等效界值之外，或者处理效应差别在等效界值之内。因为这两个无效假设是非联合的，用这种方法，I 类错误可以很好地被控制。对非劣效性试验应当用单侧区间。可信区间方法检验的无效假设为处理效应之差(被研究产品与对照之差)的等效下界相等，被择假设为处理效应之差大于等效下界的检验相同的单侧检验。 I 类错误的选择应当与使用单侧或双侧检验分开考虑。样本量的计算应当依据这些方法(见 3.5)进行。

基于观察结果得到研究产品与阳性对照没有差异的无效假设为无统计学意义而做出等效或非劣效的结论是不适当的。

无论是处理组还是对照组，病例的脱落就是倾向于反应的缺失。因而，对全分析集的分析将产生趋于显示等效性的偏差(见 5.2.3)。在选择分析集时也有讲究。

3.3.3 显示剂量-反应关系的试验

新的研究产品的剂量-反应关系是一个在开发的所有各期用多种方法(参见 ICH E4)都能得到答案的问题。剂量－反应关系的研究有多种目的，其中特别重要的有：确定是否有效；建

立剂量反应曲线的形态和位置；估计适宜的最初剂量；确定个别剂量调整的最优决策；确定最大剂量，超过这个剂量不会有更多好处。为这些目的，需要有一组不同剂量，如果合适的话，包括安慰剂(零剂量)的观察资料。为此，参数估计，包括可信区间的构建，以及建立图形的方法将与统计检验同样重要。所用的假设检验方法必须适应剂量的自然顺序，或关于剂量反应曲线形状(如单调性)的一些特殊问题。拟选用的统计方法需在实验方案中详细说明。

3.4 成组序贯设计

成组序贯设计是一种方便的进行期中分析的方法(见 4.5)。尽管成组序贯设计不是唯一的可用于期中分析的方法，但它应用得最广泛。因为，在试验中对某时间段中成组病例试验的结果进行评价比连续性对获得单个病人数据就进行评价更可行。统计学方法必须事先说明关于处理结果和病例所指定的处理（如破盲，见 4.5）信息的可获得性。一个独立的数据监视委员会（见词汇）将负责对成组序贯设计试验中的资料进行期中分析（见 4.6），该设计不但已被广泛、成功地应用于大型、长期的死亡率观察试验或非致命结果的观察，而且它在其他方面的应用亦越来越广泛。但特别是已经认识到在所有试验中必须监视其安全性，因而包括安全性的原因而提早终止试验等正规的程序，都应当考虑到。

3.5 样本量

临床试验中所需病例数必须足够多，以确保对所提出的问题给予一个可靠的回答。样本大小通常以试验中的主要目标来确定。如以其他基础来确定，则需讲清楚并且依据合理。例如，回答安全性问题或重要的次要目标所需样本量，要比回答主要目标所需样本量大(见 ICH-Ela)。

用通常的方法确定适宜的样本量，需要确定如下各项：主要变量、检验统计量、无效假设及所选剂量(包括对所选剂量和所选病例总体要检出或拒绝的处理差异的考虑)的被择工作假设、错误拒绝无效假设之概率(Ⅰ类错误)、错误地不拒绝无效假设之概率(Ⅱ类错误)，以及相应的对停止治疗、违背研究方案的处理方法。在有些情况下，事件率是估计把握度的主要变量。因此需要作一个假定，从所需的事件数外推试验最终所需的样本量。

计算样本量的具体方法，以及计算时所需的所有统计量之估计值（如方差、均值、反应率、事件率、待检差值），需在试验方案中给出，亦需给出这些估计值的依据。研究样本量对各种偏离假定的敏感性(sensitivity)是非常重要的。一个简单的方法是针对假定的一合理的偏离范围提出样本量的范围。在验证性研究中，样本量的确定主要依据出版物上的资料或预试验的结果来估算。要检出的处理差值可基于对在处理病人时与临床有关的最小效果的判断，或者基于新疗法的预期效果的判断，这比较大些。通常 I 类错误概率设在 5%或者更小，或者由多重

比较考虑所需要的调整来决定；精确的选择可能受到所检验的假设和所期望的效果的影响。Ⅱ类错误的概率通常在 10%～20%；申办者的意愿是使这一数字应尽可能地低，特别是试验很难或不可能重复时。也可能采用通常的 I、Ⅱ类错误数值外的数值，有些情况甚至更好。

样本量的计算要参照主要分析所需的病例数。如果这是“全分析集”，则所估计的样本量需要比符合方案集（见词汇）小。这是因为考虑到退出处理的病例或者依从性差的病例会冲淡处理的效应。关于变异的假定亦需加以修订。

等效性或非劣效性试验(见 3.3.2)的样本量的估计，通常应得到处理差异的可信区间的目标。这一可信区间显示处理效应的差异最多到一个临床上所能接收的最大差异。当等效性试验的检验效能是以真实差别为 0 估计时，如果差值不为 0，则往往低估达到这一把握度所需的样本大小；当非劣效性的试验的把握度是以 0 差值估计时，当试验药效低于阳性对照时，也会低估了达到这一把握度所需的样本大小。选择或确定“临床上可接受的差值”要按照对将来病人的意义来证明其合理性，且可能小于上文中确定差异存在的优效性试验中所确定的“临床上有关的”差值。

成组序贯试验中的样本量在试验前无法确定，因为其值依赖于机遇的作用并结合所选择的试验结束规定及真实的处理分布，常常包含在期望的和最大的样本量之中。

当事件发生率比预期的低，或变异比期望的大，则无需揭盲的资料，亦无需对处理进行比较，即可重新估计样本量。

3.6 资料的搜集及处理

从研究者到申办者,数据的收集和传送可有通过多种媒体，包括病例记录表、远程监测系统、医学计算机系统和电子传输器。无论采用何种方式收集数据，资料的形式和内容必须与研究方案完全一致，且在临床试验前就要定下来。

从参数收集到数据库最终完成，应当集中在实施预先计划的分析所需的数据，包括确定对计划的依从性或识别违反方案的前后关系的信息(如有关服药的时间)来判定。缺失值(miss-ing value)需与“0 值”和空缺相区别。

从数据获得到数据库完成的过程应该按照 GCP(ICH E6，5)规定执行，尤其是及时可靠的数据记录、错误更正和补遗是建立高质量数据库，通过完成分析计划达到试验之目的所必需的。

4. 进行试验所需考虑的问题

4.1 试验监视和期中分析

按照试验方案认真进行临床试验，对结果的可靠性有着重要的影响(见 ICH E6)。认真进行监视能及早发现问题，并使问题的发生和再现达到最小。

在由制药企业发起的验证性临床试验中，有两种不同的监视方法。一种监视类型是监视整个试验的质量，另一种涉及揭盲以进行处理的比较（即期中分析）。这两种类型的监视方法，除人员职责不同外，所用的数据类型和信息也不同，因而用在控制可能的统计学和操作上偏倚的原则也不同。

为了了解试验的质量，对试验管理的监视应包括研究是否按计划进行、增加的数据的质量如何、是否达到了预期收集的数量目标、设计的假设是否合适，以及保持病人在试验中是否成功等（见 4.2~4.4）。这类监视既不需要比较处理效应的信息，也不要数据揭盲，所以对 I 类错误没有影响。为达到这一目的而对一个试验进行监视也就是试验申办者的职责（参见 ICH E6），它可由试验申办者或由试验申办者指定的独立小组完成。这种监视一般从研究地点的选定开始，到最后一位病人数据的收集和清理结束。

另一类试验监视(期中分析)涉及到处理结果的比较。期中分析需要对指定的处理组（实际处理的指定或组的指定标识）揭盲，以及对比处理组的小结信息，这需要方案（或初步分析前的修订方案）中包含期中分析的统计分析计划。

初步分析前应修订方案以防止某些偏倚。这将在 4.5 和 4.6 讨论。

4.2 入选标准与排除标准的更改

入选标准与排除标准在试验对象选择的全过程中应按试验方案中的定义保持不变。但有时对其作些修改也是适当的。例如，在周期较长的临床试验中，从本试验之外或是对本试验资料的期中分析中不断获得的医学知识就有可能提示对纳入标准进行修改。此外，标准的修改也可能来自监视人员的发现。他们在监视中发现常常不能按纳入标准选择对象或由于太严格的纳入标准导致入组率非常低。标准的修改不能破盲，对所作修改应在修订方案中写明，其内容包括任何统计的后果（如不同事件发生率导致样本量的调整、或分析方法的修订、如按修改的纳入标准或排除标准进行分层分析等）。

4.3 入组率

在病人入组时间长的试验中，必须对病人入组率进行监视。如入组率远低于试验方案中预定的水平，则需查明理由，并采取相应措施，适当放宽纳入标准和降低其他方面的质量，确保试验的把握度。在多中心试验中，这些考虑适用于每一个中心。

4.4 样本量的调整

在时间较长的临床试验中，常有机会对原设计及样本量的计算中所基于的假设进行检查。如试验计划是较为初步的或者建立在不确定的信息上的，这种调整就尤为重要。在不破盲的情况下，对数据进行期中检查，可发现总反应方差、事件率或生存经历与期望不符，则应适当地修订一般假设条件，重新计

算样本量，认证其正确性，并写入修订方案及临床研究的报告中。如果曾为控制 I 类错误及其可信区间宽度而采取某些措施以保持盲法及其后果，也需加以阐述。只要可能，在试验方案中要预计到是否有重新估计样本量的潜在需要(见 3.5)。

4.5 期中分析与提早终止试验

任何在正式完成临床试验前为了比较关于处理组间的安全性或有效性的分析称为期中分析。由于这些比较的次数、方法及结果将对试验结果的解释产生影响，所有期中分析必须预先计划并在试验方案中阐明。特殊情况可能导致在试验开始时并未确定期中分析。在这种情况下，则应在揭盲分析比较处理资料前完成描述期中分析方案的修改试验方案。如果一个期中分析是为决定是否终止试验而设计，则它常采用以统计学监视计划为指导原则的成组序贯设计(见 3.4)。如果所研究处理的有效性已很清楚，或相应的处理效应之差不可能达到，或出现了无法耐受的不良反应，这种期中分析的目的是及早终止试验。一般来说，用为监测有效性而设定的界限早期终止试验(即它们更为保守)比用为安全性考虑而设定的界限终止试验需更多的验证。当试验计划或监测目标中包含了多个的终点，则这些方面的复杂性也需要加以考虑。

试验方案中应当写明期中分析的日程，或至少有关安排其进行的考虑，例如，如果要用可变动的 α 消耗（alpha spending）函数方法，需在试验计划或第一次期中分析前的修订计划中写

明。试验的终止规则及其特性需在试验计划或修订计划中详细叙述。如果该试验有数据监视委员会，则这一材料需由数据监视委员会撰写或批准(见 4.6)。偏离计划常有使试验结果无效的可能性。如试验计划需要改变，则任何相应的统计方法的改变需尽早在修订计划中写明，特别应就由于计划改变而对任何分析或推断所产生的影响进行讨论。所选方法必须保证总的 I 类错误的概率得到控制。

期中分析的执行过程应是一个完全可信的过程，因为它可能包含了非盲态数据及结果。所有参与试验的人员，必须对这类分析的结果保持盲态。因为，提前知道结果可能导致他们对试验的态度改变，且引起新入组病人特征的改变或处理间比较的偏倚。这一原则可应用于除了直接参加实施期中分析的人员之外的所有研究人员和申办人员所雇佣的人员。研究者仅仅被告知是继续试验、暂停试验，或是对试验过程进行修订。

大部分验证安全性和有效性的临床试验，需完成计划的样本量。只有在遇到伦理方面的原因，或把握度不再是可以接受时，方可终止试验。尽管如此，大家已认识到，由于种种原因，在药物开发分析中由于各种理由也包括申办者可以查阅比较性处理的数据的需要，如要设计另外一个试验；另外，那些有可能出现严重威胁生命安全或死亡率的一些研究，出于伦理学考虑，对不断积累的比较性疗效结果进行连续性监视。在以上这些情况中，期中统计分析方案均必须在比较性治疗数据揭盲之

前列入试验设计方案或在修订的设计方案中，以避免可能产生的统计和操作偏倚。

对许多新药的临床试验，特别是与公众健康关系重大的新药，必须另外指定一个独立的小组负责监视关于安全性和有效性结果的比较，并明确其职责。这个组织常被称为独立数据监视委员会（Independent data monitoring committee，IDMC），或数据与安全性监视组（Data and safety monitoring board），或数据监视委员会（Data monitoring committee）。

当项目申办者担负起监视安全性和有效性比较的职责，并从而可获得非盲法比较信息时，需特别注意试验的完整性并且适当地管理和限制资料共享。项目申办者必须保证并书面记录内部监视委员会遵守了书面的标准操作程序，并保证保留该委员会包括期中分析结果记录在内的决策会议记录。

任何设计不良的期中分析(不管是否导致早期结束试验)可能使结果有误，所得结论缺乏可靠性。因此，应避免这种分析。如作了计划外的期中分析，在研究报告中必须解释其必要性，必需破盲的程度，提供可能导致偏倚的严重程度的判断，以及对结果解释的影响。

4.6 独立资料监视委员会(IDMC)的作用

（见 1.25 及 ICH E6 中 5.5.2）

数据监视委员会可由项目申办者组建，它的主要任务是不时对临床试验的进程、安全数据和主要疗效变量进行评价，建

议项目申办者是继续修订，还是终止试验。 IDMC 应当有书面的操作规程，并保留每次会议的记录，包括期中结果，在试验结束时可供查阅。 IDMC 的独立性旨在它既能控制重要的比较试验数据的共享，又能防止因接触试验信息而可能对整个临床试验的完整性所产生的影响。IDMC 是独立于社会评论机构(Institu-tional Review Board,IRB) 或 独 立 伦 理 委 员 会(Independent Ethics Committee,IEC)的机构，它的成员中应有通晓包括统计学等有关相应学科专业知识的临床试验科学家。

当有申办者的代表参与组成数据监视委员会时，这些代表的作用应在委员会的操作程序中加以明确规定(如在关键问题上是否具有投票权)。由于申办者的人员能够获得非盲信息，因此，在委员会的操作程序中还应说明如何控制期中试验结果在申办者组织内的散布。

5. 数据分析

5.1 预定的分析计划

在进行临床试验设计时，最终数据分析的统计方法的主要特征需在试验方案中的统计分析部分加以说明。这一部分需包括主要变量的验证性分析方法的所有主要特征，以及预期分析问题的处理方法。对于探索性试验，这一部分还可包含一些更一般的原则及思路。

统计分析计划（见词汇）可以在完成试验方案后单独成文，可包含设计方案中主要特征的更加技术性的详细细节（见 7.1）。这一计划可包括主要变量或次要变量及其他数据统计分析的详细执行步骤。在对试验资料进行盲态评阅（见 7.1 的定义）分析后，应对统计分析方案进行再检查和必要的修订，并应在揭盲之前完成。正式记录应当在统计分析计划完成和其后的揭盲前保持不接触。

如果盲态评阅建议更改试验方案中所述的主要特征，需以文件的形式写入修订方案中，否则，根据盲态评阅结果建议对统计分析计划作出修改即可。只有在试验方案(包括修订方案)中设想的，分析结果才可被认为是证实性的。

在临床研究报告的统计分析一节，必须详细写明统计方法，包括是什么时候作出关于临床试验过程方法学决策的（见 ICH E3）。

5.2 分析集

数据用于主要分析的病例集(Analysis Sets)需在分析方案的统计一节明确定义。另外，所有将要开始试验过程(例如，进入阶段)的病例的文件都是有用的，该文件的内容取决于特定试验的详细特点，但是，只要可能，应当收集人口学的及病情的基线资料。

如果所有随机化进入临床试验的病例均符合入组标准，参与了试验的全过程且无失访，并提供了完整的数据记录，则该

例应包括在分析中是显然的。试验的设计与执行均应尽可能地到达这一目标。但实际上，是否能完全做到这一点是有疑问的。因此，试验方案中统计部分应有预见性地写明预期的问题，说明这些对用于分析的病例和数据有何影响。试验方案中还必须说明一些方法，以使研究执行过程中可能出现的不正确做法减到最少。这些会有损分析的满意性，包括各种类型的偏离试验方案、退出治疗及缺失值。试验方案应考虑到如何减少这类问题的出现频度，又要考虑到在分析资料时出现了这类问题的处理方法。在盲法核查时，应在可能有的补充文件中写明分析时对违反方案的处理方法。最好确认任何对试验方案的偏离发生的时间、原因及对结果的影响。偏离试验方案的频率、类型、缺失值，以及其他问题均需写入试验报告中，它们对试验结果可能产生的影响亦需在试验方案中加以论述(见 ICE-E3)。

关于分析集的决定，需遵循以下两个原则：1）使偏倚到达最小；2）控制 I 类错误的增加。

5.2.1 全分析集

意向性分析（intention-to-treat，见词汇）的基本原则是：主要分析应包括所有随机化的病例。依从这一原则需要完整地随访所有随机化对象的研究结果。但实际上这一想法难以达到，理由将在下面叙述。因此，在此文件中以“全分析集”用于描述分析集，是指尽可能完全并且尽可能接近包括所有随机化病

例的意向性分析的思想。在分析中保留初始随机化在防止偏倚并提供统计检验的安全基础方面是重要的。

在很多临床试验中，“全分析集”方法是保守的，在许多情况下它也会对治疗效果作出估计，这种估计更能反映以后实践中的情况。

有少数情况可能导致从“全分析集”中排除已随机化的病例，包括不满足主要入组标准（违反合格性），没有用过一次药以及在随机化后没有任何数据。这类排除都需证明其合理性。不符合入组标准的病例可以从分析中排除而不会引入偏倚的只有以下一些情况：

（1）入组标准判定在随机化之前。

（2）可以完全客观地作出有关违反合格性的检测。

（3）所有病例接受相同的违反合格性的检查(这在开放的研究中或即使在双盲研究中如果数据在强调盲态审核的重要性而在检查之前揭盲的情况下，可能是很困难的)。

（4）所有按特定的入组标准检出的违反行为都被排除掉了。

在某些情况下，从所有随机化病例集中除掉任何从未用过试验药的病人是合理的。尽管排除这些病例，例如，是否开始治疗的决定不会受到因了解病人的处理安排的影响时，仍然保持了意向性治疗的原则。在另外一些情况，有必要从所有随机化病人集中去除任何在随机化后没有数据的病人。除非由于这

些排除，或任何其他原因引起的可能偏倚都被述及，否则没有一个分析是完整的。

当用病人的“全分析集”时，在随机化之后违反方案可能对数据和结论有影响，尤其是当它发生与处理的指定有关时。从多方面考虑，将这些病人的数据包括在分析中，与意向性治疗一致，是恰当的。特殊问题在于，有关病人在接受一次或多次剂量后退出，而在此后不再有数据，以及由于其他原因而失访的病人，因为在“全分析集”中不包括这些病人，可能严重地削弱这种方法的基础。因而，当主要变量是在病人由于任何理由失访的情况下测定的，或是随后根据方案中的预期评定日程收集的则是有价值的；如主要变量是死亡率或严重的发病率时继续收集的资料，则特别重要。以这种方法意向性收集数据应在设计方案中写明。归因技术，从最后一次观察的结转(carrying forward)到应用复杂的数学模型的方法也可用于补偿缺失值。其他用于保证可对“全分析集”的每一个病例进行测量的方法可能需要关于病例的结果的某些假设或者结果的较简单的选择（如成功/失败）。任何这些策略的应用应当在设计方案的统计部分描述并证明其正确性，且所用任何数学模型所基于的假定应当清晰地说明。同样重要的是显示相应分析结果的强壮性，尤其是当讨论的策略能导致有偏倚的处理效应的估计时。

由于某些问题的不可预见性，有时将对试验中所出现的无规律性情况作出的详细考虑推迟到试验结束对试验数据盲态检查后则更好。如果循此做法，需在试验方案中加以说明。

5.2.2 符合方案集

病例的“符合方案”集，有时称为“有效病例”、“效验”样本或“可评价病例”样本。它定义了全分析集的一个子集。这些病人对方案更具依从性，并有符合如下准则的特征：

（1）完成某一个预定的处理规程的最小规定部分。

（2）测定主要变量的可能性。

（3）没有任何大的违反方案的地方，包括违反入组标准。

将病人排除在符合方案集之外的理由应当讲清楚，并以一种适合这一特定试验情况的方式，在破盲之前用文件写明。

应用符合方案集可能使新的治疗在分析中显示出附加效果的机会最大化，并且更密切地反应了作为方案的基础的科学模型。然而，相应的无效假设的检验和处理效应的估计依据试验不同而可能是保守的或不是保守的；由于虔诚地遵守方案而导致的偏倚，可能是严重的，与处理和结果有关。

导致排除病例产生符合方案集的问题，以及其他对方案的违反，应当被完全识别出来并加以总结。有关的违反方案可能包括处理指定的错误、使用了不许用的药物、依从性不好、失访和数据缺失。评价各处理组间关于频度和发生时间这种问题的模式是良好的做法。

5.2.3 不同分析集的作用

一般说来，显示选择不同的病例集进行分析对主要的试验结果不敏感是有优越性的。在验证性试验计划的同时对全分析集及符合方案集进行分析，一般来说是恰当的，由此可以对它们之间的任何差异进行清楚的讨论和解释。在有些情况下，最好能计划选择不同的分析集进行对结论的敏感性的探索。当全分析集和符合方案集得出实质上是相同的结论时，则试验结果的可信性增加了。然而有一点需注意，从符合方案集中排除较大比例的病例时，则对试验的总有效性会产生影响。

在优效性试验(为了显示研究产品的效果更好)和在等效或非优效性试验(为了显示研究产品具有可比性，见 3.3.2)中，全分析集和符合方案集起着不同的作用。在优效性试验中，全分析集用于主要分析（除了特殊情况），因为它倾向于避免由于符合分析集所致的效果的过于最优化估计。这是由于，在全分析集中包括了依从性不良者一般会减少估计的处理效应。然而，在一个等效性或非劣效性试验中，应用全分析集一般并不保守其作用应当非常仔细地考虑。

5.3 缺失值及离群值

缺失值是临床试验中的一个潜在的偏倚来源，因此，必须尽一切努力完成试验方案中所有有关搜集资料和数据管理的各项要求。然而，事实上任何试验几乎不可避免地总有缺失值。

不过，一个试验倘若处理缺失值的方法是敏感的，尤其那些方法在方案中已预先定义了，可以认为是有效的。在盲法核查时，在统计分析计划中更新这方面内容，可以改进方法的定义。遗憾的是，尚无一个通用的处理缺失值的方法可供推荐。研究者必须注意分析结果对处理缺失值方法的敏感性，特别当缺失值较多时。

应当用类似的方法探索离群值的影响。统计学上对离群值的定义在某种程度上讲是主观确定的。从医学和统计学上共同清晰地判断某一个特定数据是离群值更加可信，而医学上的判断常常确定适当的行动。任何在方案中或统计分析计划中设定的对离群值处理的步骤应当不会对任何一个处理组有偏向。同样，这方面的分析计划也常在资料的盲态核查时进行有用的更新。如果在试验方案中未预先指定处理离群值的方法，则用实际资料分析所得结果，以及去除或削弱离群值的影响后的至少一个分析结果均需给出，并对结果不一致之处加以讨论。

5.4 数据的变换

对关键变量是否要进行变换，最好根据以前的研究中类似资料的性质，在试验设计时就作出决定。拟采用的变换(如平方根、对数)及其原理需在试验方案中说明，特别是对主要变量。变换是为了确保资料满足统计分析方法所基于的假定，变换方法的选择原则在一般的教科书上均能找到，一些特定变量的常

用变换方法已在某些特定的临床领域得到成功应用。对一个变量是否采用变换，以及如何变换，常受到临床解释方法的影响。

导出的变量亦需作同样考虑，如从基线的改变量，从基线改变的百分数，重复测量“曲线下的面积”或两个不同变量之比值。后继的临床上的解释需仔细考虑，所选新变量导出方法需在试验方案说明其正确性。与此密切相关的一些问题已在 2.2.2 节作了讨论。

5.5 参数估计、可信区间及假设检验

试验方案中的统计部分应当说明要检验的假设和/或为了满足试验的主要目的而待估计的处理效应。为完成这些任务的主要变量(最好也有次要变量)的统计分析方法，以及所基于的统计模型需阐述清楚；如可能，处理效应的估计需同时有可信区间，并需说明其计算方法。如想要根据基线资料以提高估计精度，或对可能的基线差异估计值进行校正，如协方差分析，亦需在试验方案中写明。

明确说明所采用的假设检验是单侧的还是双侧的是非常重要的，特别是当要采用单侧检验时，需事先说明其是正确的。如果认为假设检验不合适，则需给出其他得到统计结论的方法。关于统计推断用单侧还是双侧是有争议的，在统计文献中可见到不同的观点。通常推荐在设定单侧检验的第 I 类错误时可以设为双侧检验中的一半，这就使得与通常适用于估计两种处理间差异的可能大小的双侧可信区间相一致。

所选择的统计模型应当能反映目前医学和统计学关于所分析的变量以及试验设计的知识。所有在分析中拟合的效应(例如在方差分析模型中的)应当全面地说明。而且，如果有的话，应当对由于初步结果而进行了修改的效应集加以说明。对在协方差分析(见 5.7)中拟合的协变量集也应作同样的考虑。在选择统计方法时，应注意到主要和次要变量的统计分布。在进行选择时（例如参数还是非参数方法），应当记住需要提供处理效应的大小及其可信区间（除了提供统计意义检验之外）。

主要变量的主要分析应当清晰地与主要或次要变量的附加分析区别开来。在方案的统计部分或者统计分析计划中，也应当概述除了主要和次要变量之外的数据总结和报告的方法。这应当包括各种试验，例如安全性数据中的分析达到一致性所采用的任何方法的参考文献。

建立模型方法与已知的药理学参数、病人对方案的依从性或其他基于生物学数据的了解相结合可以对实际的或可能的效果，特别是对于处理效应的估计，提供有价值的洞悉。这类模型所基于的假定都应当清晰地加以说明，而任何结论的局限性也应仔细地描述。

5.6 I 类错误及可信水准的调整

当出现多重性(multiplicity)时，常用的分析临床试验资料的频率的方法需对 I 类错误进行调整。多重性可以由于以下情况而产生，例如多个主要变量(见 2.2.2)、处理的多重比较、不同

时期的多次评估和/或期中分析(见 4.6)。如果有可能，最好采用避免或减少多重性的方法，如确定一个关键的主要变量（多重变量）、选择关键的处理对比（多重比较）、运用综合变量如“曲线下面积”（重复测量时）。作了这样的处理后，在验证性分析中，如仍有多重性方面的问题，则需在试验方案中确定；必须考虑调整，调整的详细步骤，以及为何不必调整均需在分析计划中说明。

5.7 次级组、交互作用及协变量

主要变量常系统地与除处理因素以外的其他因素有关，例如，年龄、性别等与协变量有关，特定的次级组间，如在多中心试验中，不同中心治疗的病人可能有差异。在有些情况下，对协变量及对次级组效应的校正是分析计划中不可缺少的一部分，故亦需在试验方案中陈述。需在试验前深思熟虑地识别可能对主要变量有重要影响的协变量和因素，并且应当考虑如何对其进行分析以提高估计的精度，以及补偿处理组间不均衡所产生的影响。如果在设计中有一个或多个分层因素，在分析中应当包括这些因素。当一个校正的可能数值可疑时，建议将未经校正的分析结果作为主要依据，而将校正后的分析结果作为参考。特别要注意中心的作用及主要变量的基线值的作用。在随机化分组后测量的协变量值对主要分析作调整是不可取的，因为它们可能受处理的影响。

处理效应的大小会因次级组或协变量的不同而不同，例如，效应可能会随年龄的增加而减少，或对某一类病人较大。在有些情况下，这种交互作用能预期到或者特别感兴趣（如老年等），因此，次级组分析，或包含交互作用的统计模型，都属验证性分析计划的一部分。然而，在大多数情况下，次级组分析和交互作用分析是探索性的，并且应当清晰地认为是如此的；它们应当探索任何处理在不同情况下得出的效应是一样的。总之，这类分析首先应在所研究的统计模型中添加交互作用项，再加上对有关病人的次级组内或由协变量定义的层内的这病例作附加的探索性分析，加以补充。在作探索性分析时，对这种分析的解释必须十分审慎，任何仅基于次级组所作的探索性分析，任何关于有效（或无效），或是安全性的结论，均不宜被接受。

5.8 资料的完整性与计算机软件的正确性

资料分析的数值结果的可信程度，依赖于用于数据处理、数据输入、储存、核实、改错、检索和统计学处理中的方法和软件（内部和外部的）的质量和正确性。所以，数据处理须基于完善的、有效的标准操作程序。用于数据管理和统计分析的计算机软件必须可靠，并提供恰当的软件检验过程文件。

6. 安全性与耐受性评价

6.1 评价的范围

在所有的临床试验中，安全性及耐受性（见词汇）评价是非常重要的一个方面。在试验早期，这一评价主要是探索性的，且只对毒性明显的表现敏感，而在后期，由于样本较大，对于药物的安全性和耐受性的评价将更为全面。后期的对照试验，代表了一个重要的以无偏的方式探索任何新的潜在的不良反应的方法，尽管这类试验在这一方面的把握度较低。

为了说明在安全性与耐受性方面与其他药物或该药物的其他剂量的比较的优效性或等效性，可设计某些试验。这种申述需要相应的验证试验的支持，这与相应的有效性的申述要求是一样的。

6.2 变量的选择与资料搜集

在任何一个临床试验中，用于评价一种药物的安全性和耐受性的方法及度量准则依赖于一些因素，包括与之密切相关的药物的不良反应知识、非临床试验或早期临床试验的信息、该药物的药效学及药代动力学(pharmacodynamic/pharmacokinetic)特性、服药方式、所研究的病人类型，以及试验的期间等。实验室检验包括临床化学和血液学、生命指征(vital signs)及临床不良反应(疾病、体征和症状)的实验检查，通常构成了安全性与耐受性资料的主体部分。严重不良事件的发生，及因不良事件导致治疗终止对于注册特别重要(见 ICH E2A 及 ICH E3)。

此外，为便于对不同试验的资料进行合并，建议在整个临床试验中，资料的收集及评价所用的方法最好一致。使用一个

通用的不良事件的词典是特别重要的。该词典从三个不同的级别对不良事件的资料的概括给出可能性，即系统-器官级，推荐名词或包括名词(见词汇)。推荐名词是通常用于汇总不良事件所用的名词，然后，在数据的描述表达时让同一系统-器官级的推荐名词进行合并。

6.3 用于评价的病例集及数据的表达

评价总的安全性与耐受性时，用于汇总的病例集常定义为至少接受了一剂被研究药物。从这些病例中收集的安全性及耐受性变量应尽可能地全面，包括不良反应类型、严重程度、发生及持续时间(见 ICH E2B)。另外，在特定的次级人群，如女性、老年人(见 ICH E7)、危重病人，或接受了辅助药物治疗的人可能需要附加的安全性及耐受性的评价。这些评价需说明更特殊的问题(见 ICH E3)。

所有安全性及耐受性变量在评价中均需十分重视，所用的主要分析方法需在研究方案中指明。所有的不良事件均需报告，无论是否被认为与处理有关。在评价中，研究人群的所有可用资料均需说明。实验室变量的度量单位及参考值范围必须认真制订，如在同一试验，出现不同的单位及不同的参考值范围（如多个实验室参与研究），则需进行恰当的标准化，以便进行统一评价。毒性等级尺度也必须事先确定，并说明其正确性。

某不良事件的发生强度通常以出现不良事件的病例数与暴露病例数之比来表示。然而，发生强度并非总是十分清楚的。

例如，根据不同情况，可考虑用暴露病例数或暴露程度（用人年表示）作为分母。无论是用于估计危险度还是进行处理组间的比较，定义需在试验方案中写明，这一点是很重要的。尤其对于时间较长的治疗，估计会有较大的退出治疗的比例及死亡比例时，对这类情况，需考虑用生存分析，并计算累积不良反应率，以避免低估的危险。

当体征和症状存在较大的背景噪音（如精神病的试验）时，在估计不同不良事件的危险时需考虑对此进行说明的方法。有一种方法是运用“处理后出现的事件”(见词汇)的概念，只有当不良事件相对于治疗前的基线出现恶化时才被记录。

其他消除背景噪音的方法也可以选用，如忽略程度轻微的不良事件，或在重复随访观察到者方可计入分子。这些方法需在试验方案中解释并说明其正确性。

6.4 统计学评价

安全性与耐受性的研究是一个多方面的问题。虽然，某些不良反应通常可被预计到，且对所有药物都进行监测，但不良反应的可能范围很广，新的未预计到的不良反应总是有可能发生的。此外，当违背了试验方案，如使用了方案中禁用的药物，出现了不良事件，就可以产生偏倚。这一背景构成了药物安全性和耐受性评价有关统计上的困难，这意味着由验证性临床试验得到的结论性信息只是一种例外，而不是通例。

在大多数试验中，对安全性与耐受性最好用描述性统计方法对数据进行分析，并在有利于说明时辅以可信区间。对处理组间及病例间的不良事件的模式用图形式来表达也是有价值的。

计算 *P* 值有时也是很有用的，既有利于评价某一感兴趣的差别，又可作为一种"特殊标志"手段应用于大量安全性与耐受性变量，以显示其差别值得进一步注意。这对实验室资料特别有用，因为除此以外，很难给予恰当的汇总。建议实验室资料既要作定量分析，如估计处理的均数，又要作定性分析，计算高于或低于某一阈值的病例数。

如用假设检验，则在进行多重比较时需进行统计上的修正以控制 I 类错误，但通常更关注 II 类错误的大小。如未对多重比较作修正，则解释被认为统计学上有意义的结果时需特别小心。

在大多数研究中，观察者希望确定，与阳性对照药及安慰剂相比，安全性及耐受性未出现临床上不可接受的差别。对非劣效性或等效性评价，应用可信区间比用假设检验更佳，这样，因发生频数较低而造成的较大的不精确性可以清晰地表示出来。

6.5 综合总结

药物的安全性与耐受性通常是在药物的开发过程中连续地通过试验过程总结出来的，特别是进行上市申请时。然而，总

结的有用性依赖于适当的、严格控制的有高质量数据的个别试验。

药物的总的有用性总是一个权衡利弊的问题，即使对利与弊的评价总是对整个临床试验项目进行总结时才进行(见7.2.2)，但在单一试验中，这一观点亦应考虑到。

有关安全性与耐受性报告中所需的更详细的内容见 ICH E3 的第 12 节。

7. 研究报告

正如引言中所述，临床试验报告的格式与内容是 ICH E3 的内容。 ICH 全面地包括了统计工作的报告，亦适当结合了一些临床及其他材料。本节只作简单讨论。

如第 5 节所述，在试验的设计阶段，分析方法的主要特点必须在研究方案中确定。当试验结束后，数据已收集完整，则可作初步审查，正如第 5 节所描述的，对数据按计划好的分析进行盲法审查是很有价值的。这种对处理保持盲态的预分析审查应当包括关于以下一些问题的决定，例如从分析集中剔除个体或数据；考察可能的变量变换，定义离群值；将其他最新研究中确定的重要协变量增加到模型中去；重新考虑用参数方法还是用非参数方法。此时所作的决定需写入报告，并与统计学专业人员在知道处理编码后所作的决定相区别，因为在盲态所作的决定一般引入偏倚的可能性较小。参加非盲期中统计分析的统计学专业人员或其他人员不应当参加盲法审查或对统计分

析计划的修改。当处理所致的效应在数据中显示出来的可能性威胁到盲法时，盲法审查需要特别小心。

许多更详细的表达和列表方面应当在接近或正当盲法审查时最终完成以便在实际分析时整个计划的所有各方面已经存在，这些方面包括病例的筛选、数据的筛选与修正、资料的汇总与列表、参数估计及假设检验。一旦数据核查已完成，则应按预定的分析计划进行分析，越遵循分析计划，所得结论的可信度就越大。当实际分析有别于在试验方案中、修订方案中及对资料进行盲态审核时所确定的统计分析计划时，要特别注意，对于偏离计划的分析必须给予认真详细的解释。

凡进入临床试验的病例，不管是否参与统计分析，均需在研究报告中说明。所有排除在分析之外的理由均需写明，任何一个包含在全分析集但不包含在符合方案集中的病例亦需写明其排除符合方案集的理由。同样，所有参与分析集的病例，其所有重要变量的测量值均需说明其测量的时点。

所有病例或数据的丢失、退出处理及违背试验方案等情况对主要变量分析的影响必须认真考虑。病例的失访、退出治疗、或严重违背试验方案必须确认，并对其进行描述性分析，包括退出的理由，以及与处理及结果的关系。

描述性分析是研究报告中必不可少的部分。应当用图或表的形式清晰地表示主要变量、次要变量、主要预后及人口学变量的重要特征。与试验目的相关的主要分析的结果应当是研究报告中特别仔细描述的内容。在报告统计学意义检验的结果时，

应当报告精确的 *P* 值(如 P=0.034)，而不是列出唯一的参照临界值。

尽管临床试验分析的主要目的应当是回答总目标中提出的问题，但在非盲态分析时基于观察数据又会出现一些新的问题，这时就需要用其他的或更复杂的统计分析方法来处理。在研究报告中，这部分的工作必须与方案中计划分析的内容严格区分开来。

由于机遇的作用，可能导致对处理组间基线测定项目的未预见的不均衡项在计划的分析方案中没有被预先定义为协变量，但它对预后具有一定的重要意义。处理这种不平衡的最佳方法是用一种附加的统计分析，说明在考虑这种不平衡因素后可以得出与原计划的统计分析方案相一致的结论。如果经过如上处理不能得出相一致的结论，则需对这种不平衡对结论的影响加以讨论。

一般而言，计划外的分析应尽量少用。如果认为治疗效果有可能由于其他某个或某些因素的改变而不同时，常需进行这种分析。这时可能是企图识别效果特别好的试验对象的某一亚组。对于计划外亚组分析结果过度解释的潜在危险是众所周知的（参见 5.7），应设法小心地避免。虽然，当一个处理无效或该处理对亚组试验对象具有副作用也会出现类似解释的问题，但我们应对其可能性作出适当评价并加以报告。

最后的统计学判定对临床试验结果的分析、解释及表达有关。为此，试验统计学专业人员应当是临床试验报告负责人员之一，并批准最终报告。

7.2 临床数据库的总结

在进行药品上市申请时，需要所有报告和临床试验的有关安全性和有效性全面小结和证据的综合材料(欧盟的专家报告，美国的整体小结和日本的概要)。在适当的时候需附有统计学的综合结果。

在小结中，应包括如下专门的统计学分析内容：参与临床试验过程中治疗人群的人口学特征和临床表现的描述；根据有关的（一般是有对照的）试验结果回答较关键问题，且着重说明其一致和不一致的程度；总结所有试验的综合数据库中所有的安全性信息，其结果对于上市申请有作用并可验明可能的安全性问题。在临床试验计划设计时，必须注重变量的定义及测量值收集的一致性，这将有利于随后的系列试验结果的解释，特别是当几个试验进行联合时。必须选用一个通用的记录用药详情、病史及不良事件的通用词典，对主要变量与次要变量采用公认的定义往往是有益的，而且是后继综合分析的基础。测量关键有效性变量的方式、安排对随机化/进入试验评价的时间、处理对违反或偏离试验方案者以及可能对预后因素的定义，都必须保持一致，除非有充分理由不这么做。

任何用于不同试验间数据联合的统计方法均需详细描述，对因试验的选择而可能导致的偏倚、对它们结果的齐性，以及对不同的变异来源建立恰当的模型都必须予以十分注意。结论对假定及所作的选择的敏感性必须进行探索。

l7.2.1　有效性资料

每一个临床试验的样本量都必须足够大，以确保达到预期的目的。通过对本质上是说明相同的关键的效应问题的一系列临床试验结果的总结，也可能得到附加的有价值的信息。这一系列试验的主要结果应当以统一的格式表达，以便于比较，一般是用表格或图形的方式表达，主要是估计值和可信区间。用后继综合分析技术对这些参数进行综合就是一个很好的方法，因为该法可为所产生的处理效应的大小提供一个更加精确的总的估计，为试验结果提供一个完整而简洁的总结。在某些特殊情况下，后继综合分析也可能是最合适或唯一的方法，它通过总的假设检验提供充分的总的有效证据。当为此目的应用后继综合分析时，应该有其自己写好的方案。

7.2.2　安全性资料

在总结安全性数据时，要彻底检查安全性数据库中任何可能的中毒迹象，并且以寻找观察值的有关联的提供证据的模式来随访任何迹象，这是重要的。将所有服用新药的人群的安全

性资料联合起来分析，可提供信息的重要来源，因为大样本为检出发生率较低的不良事件提供机会，也许还可估计出近似发生率。但因为没有对照组，很难对由这一数据中得到的发生率进行评价，对照试验的资料对克服这种困难就显得特别有价值。用共同对照组(安慰剂或指定的阳性对照)的研究结果，应当进行综合，并对每一有足够数据的比较组分别给出研究结果。

探索资料时发现的任何潜在的中毒迹象均需报告。对这些潜在的不良反应的真实性的评价需考虑因大量的比较而产生的复杂性问题。在评价时也可适当运用生存分析方法，求得不良事件的发生率与服药时间和/或随访时间之间的潜在关系。与已识别的不良反应事件相联系的危险性必须适当量化，以便权衡利弊关系。

词汇

贝叶斯方法(Bayesian Appmaches)

数据分析的方法，由观察数据及参数的先验概率分布导出某些参数(如处理效应)的后验概率分布。

偏倚(统计的和操作上的)(Bias Statistical & Operational)与设计、执行、分析和评价临床试验结果有关的任何因素的系统倾向使操作效应的估计值偏离其真值。由于执行不正确造成的偏倚称为“操作”偏倚。上面所列出的偏倚的其他原因称为“统计学的”。

盲态审核（Blind Review）

在试验完成（最后一例病人的最后一次观察）与揭盲之间对数据进行核对和评价，以便把计划的分析最后定下来。

含义的有效性（Content Validity）

一个变量（如等级量表）度量了其应该度量的大小的程度。

双模拟（Double-Dummy）

在临床试验中当两种处理不能做到一样时，使应用制品时仍保持盲态的一种技术，如为处理 A（有效药和不能区别的安慰剂）及处理 B（有效药和不能区别的安慰剂）制备制品。病人接受两套处理：或者是 A（有效药）及 B（安慰剂），或者是 A（安慰剂）和 B（有效药）。

脱落（Drop out）

临床试验中的病人由于任何原因不能继续进行试验到按试验方案要求他/她的最后一次随访。

等效性试验（Equivalence Trial）

一个试验的主要目的是要显示两种或多种处理的反应差别大小在临床上并无重要性。这通常以显示真正的处理差异是在临床上可以接受的等效性的上下界之间。

频率法（Frequentist Methods）

统计方法，如统计意义检验和可信区间，可以用同一试验情况下假设的重复实现时某一结果出现的频率来说明。

全分析集（Full Analysis Set）

尽可能接近按意向性治疗原则的理想的病例集。由所有随机化的病人中以最少的和合理的方法排除病例得出。

广义性（Generalisability, Generalisation）

一个临床试验的结果可以被可信地由参加试验的病人外推到广大的病人群体和广大范围的临床环境的程度。

全局评定变量（Global Assessment Variable)

单一变量，通常是把客观变量和研究者对病人的状况或者状态的改变情况结合起来的顺序分类等级尺度。

独立数据监视委员会（数据和安全监视组，监视委员会，数据监视委员会）（Independent Data Monitoring Committee-IDMC,Data and Safty Monitoring Board,Monitoring Committee, Data Monitoring Committee）

一个独立的数据监视委员会可以是由申办者建立的经常评定临床试验的进度、安全性数据以及关键性效果的结果，并且向申办者提出建议是否继续、修改或停止试验。

意向性治疗原则(Intention To Treat Principle)

一种认为处理策略以想要治疗病人(即计划好的治疗进程)，而不是基于实际给予的治疗为基础进行评价，可以对效果作出最好的评定原则。其结果是分到一个处理组的病人即应作为该组的成员被随访、评价和分析，而不管他们是否依从计划的处理过程。

交互作用(定性的和定量的)(Interaction，Qualitative & Quantitative)

一种处理的对比(例如研究产品与对照之间的差异)依赖于另外一个因素(如中心)的情况。定量的交互作用是指对比差异

的大小在因素的不同水平时不同，而定性交互作用时对比差异的方向至少在因素的一个水平上不同。

平定者间的可靠性(Inter-Rater Reliability)

不同评定者在不同情况下产生相同结果的性能。

评定者内的可靠性(Intre-Rater Reliability)

同一评定者在不同情况下产生相同结果的性能。

期中分析(Interim Analysis)

在正式结束试验之前在任何时期为了比较效果或安全性的任何分析。

后期综合分析（Meta Analysis）

对同一个问题的两个或更多的试验的定量证据进行正式的评价。这常是从各试验的小结统计资料进行统计合作，但此名词有时也用于对原始数据的合并。

多中心试验（Multicentre Trial）

按单一试验方案在多个地点进行的临床试验。因而，由多个研究者进行。

非劣效性试验（Non-Inferiority Trial）

主要目的是显示研究产品的反应在临床上不劣于比较制剂（阳性或安慰剂对照）的试验。

推荐和包括名词（Preferred and Included Terms）

在一个分层次的医学词典中，例如 MedDRA，包括名词是最低级别的词典名词，以研究者的描述进行编码。推荐名词是对包括名词进行并组的级别，用于报告发生频率。例如，研究

者写的是“左臂疼痛”，包括名词编码为“关节疼痛”，在推荐名词级别可报告为“关节痛”。

符合方案集（有效病例，有效样本，可评价病例样本）（Perprotocol Set，Valid Cases，Efficacy Sample，Evaluable Subjects Sample）

由充分依从于方案以保证这些数据会按所基于的科学模型而表现出治疗效果的病例子集所产生的数据集。依从性包括以下一些考虑，如暴露于处理、可以测定以及没有对方案大的违反等。

安全性和耐受性（Safety and Tolerability）

医学产品的安全性涉及到病人的医疗风险，通常在临床试验中由实验室检查（包括临床生化与血液学）、生命体征、临床不良事件（疾病、体征和症状），以及其他专门的安全性检查（例如心电图、眼科检查）等来评定。医学产品的耐受性代表了病人能忍受明显的不良反应的程度。

统计分析计划（Statistical Analysis Plan）

统计分析计划是包括比方案中描述的主要分析特征更加技术性和更多详细细节的文件，并且包括了对主要和次要变量及其他数据进行统计分析的详细过程。

优效性试验（Superiority Trial）

主要目的是显示研究产品的反应优于对比制剂(阳性或安慰剂对照)的试验。

间接变量（Surrogate variable)

在直接测定临床效果不可能或不实际时，提供效果间接测定的变量。

处理效应（Treatment Effect）

在临床试验中归因于处理的效果。在大多数临床试验中感兴趣的处理效应是两个或多个处理的比较（或对比）。

处理后出现的事件（Treatment Emergent）

在治疗时出现的，而在治疗前没有的或比治疗前状况更坏的事件。

试验统计学专业人员（Trial Statistician）

经过教育、培训并且有经验足以贯彻本指导中的原则并且负责试验的统计方面的统计学专业人员。