附件

医疗器械临床试验设计指导原则

医疗器械临床试验是指在具备相应条件的临床试验机构中，对拟申请注册的医疗器械在正常使用条件下的安全有效性进行确认的过程。临床试验是以受试人群（样本）为观察对象，观察试验器械在正常使用条件下作用于人体的效应或对人体疾病、健康状态的评价能力，以推断试验器械在预期使用人群（总体）中的效应。由于医疗器械的固有特征，其试验设计有其自身特点。

本指导原则适用于产品组成、设计和性能已定型的医疗器械，包括治疗类产品、诊断类产品，不包括体外诊断试剂。

本指导原则是供申请人和审查人员使用的技术指导文件，不涉及注册审批等行政事项，亦不作为法规强制执行，如有能够满足法规要求的其他方法，也可以采用，但应提供详细的研究资料和验证资料。应在遵循相关法规的前提下使用本指导原则。

一、医疗器械临床试验目的

临床试验需设定明确、具体的试验目的。申请人可综合分析试验器械特征、非临床研究情况、已在中国境内上市（下文简称已上市）同类产品的临床数据等因素，设定临床试验目的。临床试验目的决定了临床试验各设计要素，包括主要评价指标、试验

设计类型、对照试验的比较类型等，进而影响临床试验样本量。不同情形下的临床试验目的举例如下：

（一）当通过临床试验确认试验器械在其预期用途下的安全有效性时，若更关注试验器械的疗效是否可满足临床使用的需要，其临床试验目的可设定为确认试验器械的有效性是否优于/等效于/非劣于已上市同类产品，同时确认试验器械的安全性。此时，临床试验的主要评价指标为有效性指标。

（二）当通过临床试验确认试验器械在其预期用途下的安全有效性时，若更关注试验器械的安全性是否可满足临床使用的需要，其临床试验目的可设定为确认试验器械的安全性是否优于/等效于/非劣于已上市同类产品，同时确认试验器械的有效性。此时，临床试验的主要评价指标为安全性指标，以乳房植入体为例，临床试验通常选择并发症发生率（如包膜挛缩率、植入体破裂率）作为主要评价指标。

（三）对于已上市产品增加适应症的情形，临床试验目的可设定为确认试验器械对新增适应症的安全有效性。例如，止血类产品在已批准适用范围（如普通外科、妇产科）的基础上，增加眼科、神经外科、泌尿外科使用的适应症。

（四）当已上市器械适用人群发生变化时，临床试验目的可设定为确认试验器械对新增适用人群的安全有效性。例如膜式氧合器产品，在原批准适用范围的基础上新增体重≤10kg 的适用人

群；又如治疗类呼吸机在已批准的适用于成人的基础上新增适用于儿童的适用范围。

（五）当已上市器械发生重大设计变更时，可根据变更涉及的范围设定试验目的。例如冠状动脉药物洗脱支架平台花纹设计发生改变时，临床试验目的可设定为确认变化部分对于产品安全有效性的影响。

（六）当已上市器械的使用环境或使用方法发生重大改变时，试验目的可设定为确认产品在特定使用环境和使用方法下的安全有效性。例如：已上市的植入式心脏起搏器通常不能兼容核磁共振检查，如申请兼容核磁共振检查，其临床试验目的可设置为对兼容核磁共振检查相关的安全有效性进行确认。

二、临床试验设计的基本类型和特点

（一）平行对照设计

随机、双盲、平行对照的临床试验设计可使临床试验影响因素在试验组和对照组间的分布趋于均衡，保证研究者、评价者和受试者均不知晓分组信息，避免了选择偏倚和评价偏倚，被认为可提供高等级的科学证据，通常被优先考虑。对于某些医疗器械，此种设计的可行性受到器械固有特征的挑战。

1.随机化

随机化是平行对照、配对设计、交叉设计等临床试验需要遵循的基本原则，指临床试验中每位受试者均有同等机会（如试验

组与对照组病例数为 1:1）或其他约定的概率（如试验组与对照组病例数为 n:1）被分配到试验组或对照组，不受研究者和/或受试者主观意愿的影响。随机化是为了保障试验组和对照组受试者在各种已知和未知的可能影响试验结果的基线变量上具有可比性。

非随机设计可能造成各种影响因素在组间分布不均衡，降低试验结果的可信度。一方面，协变量分析可能难以完全校正已知因素对结果的影响；另一方面，未知因素对试验结果产生的影响亦难以评价，因此，通常不推荐非随机设计。如果申请人有充分的理由认为必须采用非随机设计，需要详述必须采用该设计的理由和控制选择偏倚的具体措施。

2.盲法

如果分组信息被知晓，研究者可能在器械使用过程中选择性关注试验组，评价者在进行疗效与安全性评价时可能产生倾向性，受试者可能受到主观因素的影响。盲法是控制临床试验中因“知晓分组信息”而产生偏倚的重要措施之一，目的是达到临床试验中的各方人员对分组信息的不可知。根据设盲程度的不同，盲法可分为完整设盲、不完整设盲和不设盲。在完整设盲的临床试验中，受试者、研究者和评价者对分组信息均处于盲态。

在很多情形下，基于器械及相应治疗方式的固有特征，完整设盲是不可行的。当试验器械与对照器械存在明显不同时，难以对研究者设盲，例如膝关节假体，试验产品和对照产品的外观可

能存在明显不同，且植入物上有肉眼可见的制造商激光标记；又如血管内金属支架，试验产品和对照产品的具体结构、花纹不同。此时，建议尽量对受试者设盲，即受试者不知晓其被分入试验组或对照组，并采用第三方盲法评价（如中心阅片室、中心实验室、评价委员会等）和盲态数据审核。当试验器械形态与对照器械存在明显不同且主要评价指标来自影像学数据时，难以对研究者、评价者设盲，例如生物可吸收支架，当对照产品为金属支架时，由于生物可吸收支架平台发生降解，评估晚期管腔丢失指标（该指标以影像学方式评价）时难以对评价者设盲。此时，建议尽量对受试者设盲，并采用盲态数据审核。上述由于器械的固有特征而不对研究者设盲、不对研究者和评价者设盲的情形，均为不完整设盲的临床试验设计。

当试验组治疗方式（含器械）与对照组存在明显差异时，难以对受试者、研究者、评价者设盲，只能采取不设盲的试验设计，如介入治疗和手术治疗进行比较时、器械治疗和药物治疗进行比较时。为最大程度地减少偏倚，可考虑采用以下方法：（1）在完成受试者筛选和入组前，受试者和研究者均不知晓分组信息（即分配隐藏）；（2）在伦理许可的前提下，受试者在完成治疗前，不知晓分组信息；（3）采用盲态数据审核。

申请人需要对采用不完整设盲或者不设盲试验设计的理由进行论述，详述控制偏倚的具体措施（如采用可客观判定的指标

以避免评价偏倚，采用标准操作规范以减小实施偏倚等)。

3.对照

对照包括阳性对照和安慰对照（如假处理对照、假手术对照等)。阳性对照需采用在拟定的临床试验条件下疗效肯定的已上市器械或公认的标准治疗方法。

对于治疗类产品，选择阳性对照时，优先采用疗效和安全性已得到临床公认的已上市同类产品。如因合理理由不能采用已上市同类产品，可选用尽可能相似的产品作为阳性对照，其次可考虑标准治疗方法。例如，人工颈椎间盘假体开展临床试验时，如因合理理由不能采用已上市同类产品，可选择临床广泛使用的、对相应适应症的疗效已得到证实并被公认的产品。又如，治疗良性前列腺增生的设备在没有同类产品上市的情形下，可采用良性前列腺增生症的标准治疗方法（经尿道前列腺电汽化术）作为对照。标准治疗方法包括多种情形，例如，对于部分临床上尚无有效治疗方法的疾病，其标准治疗方法可为对症支持治疗。在试验器械尚无相同或相似的已上市产品或相应的标准治疗方法时，若试验器械的疗效存在安慰效应，试验设计需考虑安慰对照，此时，尚需综合考虑伦理学因素。若已上市产品的疗效尚未得到临床公认，试验设计可根据具体情形，考虑标准治疗方法对照或安慰对照，申请人需充分论证对照的选取理由。例如用于缓解疼痛的物理治疗类设备。

对于诊断器械，对照需采用诊断金标准方法或已上市同类产品。

（二）配对设计

对于治疗类产品，常见的配对设计为同一受试对象的两个对应部位同时接受试验器械和对照治疗，试验器械和对照治疗的分配需考虑随机设计。配对设计主要适用于器械的局部效应评价，具有一定的局限性。例如，对于面部注射用交联透明质酸钠凝胶的临床试验，配对设计在保证受试者基线一致性上比平行对照设计具有优势，但试验中一旦发生系统性不良反应则难以确认其与试验器械或对照器械的相关性，且需要排除面部左右侧局部反应的互相影响。因此，申请人考虑进行配对设计时，需根据产品特征，综合考虑该设计类型的优势和局限性，恰当进行选择，并论述其合理性。

对于诊断器械，若试验目的是评价试验器械的诊断准确性，常见的配对设计为同一受试者/受试样品同时采用试验器械和诊断金标准方法或已上市同类器械来进行诊断。

（三）交叉设计

在交叉设计的临床试验中，每位受试者按照随机分配的排列顺序，先后不同阶段分别接受两种或两种以上的治疗/诊断。此类设计要求前一阶段的治疗/诊断对后一阶段的另一种治疗/诊断不产生残留效应，后一阶段开始前，受试者一般需回复到基线状态，可考虑在两个干预阶段之间安排合理的洗脱期。

（四）单组设计

单组试验的实质是将主要评价指标的试验结果与已有临床数据进行比较，以评价试验器械的有效性/安全性。与平行对照试验相比，单组试验的固有偏倚是非同期对照偏倚，由于时间上的不同步，可能引起选择偏倚、混杂偏倚、测量偏倚和评价偏倚等，应审慎选择。在开展单组试验时，需要对可能存在的偏倚进行全面分析和有效控制。

1.与目标值比较

与目标值比较的单组设计需事先指定主要评价指标有临床意义的目标值，通过考察单组临床试验主要评价指标的结果是否在指定的目标值范围内，从而评价试验器械有效性/安全性。当试验器械技术比较成熟且对其适用疾病有较为深刻的了解时，或者当设置对照在客观上不可行时（例如试验器械与现有治疗方法的风险受益过于悬殊，设置对照在伦理上不可行；又如现有治疗方法因客观条件限制不具有可行性等），方可考虑采用单组目标值设计。考虑单组目标值设计时，还需关注试验器械的适用人群、主要评价指标（如观察方法、随访时间、判定标准等）是否可被充分定义且相对稳定。为尽量弥补单组目标值设计的固有缺陷，需尽可能采用相对客观、可重复性强的评价指标作为主要评价指标，如死亡、失败等；不建议选择容易受主观因素影响、可重复性差的指标作为主要评价指标，如疼痛评分等。

目标值是专业领域内公认的某类医疗器械的有效性/安全性评价指标所应达到的最低标准，包括客观性能标准（Objective performance criteria，OPC）和性能目标（Performance goal，PG）两种。目标值通常为二分类（如有效/无效）指标，也可为定量指标，包括靶值和单侧置信区间界限（通常为 97.5%单侧置信区间界限）。目标值的构建通常需要全面收集具有一定质量水平及相当数量病例的临床研究数据，并进行科学分析（如 Meta 分析）。对临床试验结果进行统计分析时，需计算主要评价指标的点估计值和单侧置信区间界限值，并将其与目标值进行比较。

由于没有设置对照组，单组目标值设计的临床试验无法确证试验器械的优效、等效或非劣效，仅能确证试验器械的有效性/安全性达到专业领域内公认的最低标准。

（1）与 OPC 比较

OPC 是在既往临床研究数据的基础上分析得出，用于试验器械主要评价指标的比较和评价，经确认的 OPC 目前尚不多见。OPC 通常来源于权威医学组织、相关标准化组织、医疗器械审评机构发布的文件。例如一次性使用膜式氧合器，其临床试验可采用单组目标值设计，当主要评价指标采用《一次性使用膜式氧合器注册技术审查指导原则》中提及的复合指标“达标率”时，试验产品达标率的目标值应至少为 90%，预期达标率为 95%。又如，根据《髋关节假体系统注册技术审查指导原则》，

对于常规设计的髋关节假体，当临床试验采用单组目标值设计，主要评价指标采用术后 12 个月 Harris 评分“优良率”时，试验产品“优良率”的目标值应至少为 85%，预期优良率为 95%。随着器械技术和临床技能的提高，OPC 可能发生改变，需要对临床数据重新进行分析以确认。

（2）与 PG 比较

当有合理理由不能开展对照试验而必须考虑开展单组目标值设计时，若没有公开发表的 OPC，可考虑构建 PG。例如脱细胞角膜植片，适用于药物治疗无效需要进行板层角膜移植的感染性角膜炎患者。由于开展临床试验时市场上无同类产品，且与异体角膜移植对比存在角膜来源困难的问题，故采用 PG 单组设计进行临床试验，PG 来源于异体角膜移植既往临床研究数据，由相关权威的专业医学组织认可。与 OPC 相比，采用 PG 的单组设计的临床证据水平更低。PG 的实现/未实现不能立即得出试验成功/失败的结论，如果发现异常试验数据时，需要对试验结果进行进一步探讨和论证。

2.与历史研究对照

与历史研究对照的临床试验证据强度弱，可能存在选择偏倚、混杂偏倚等问题，应审慎选择。当采用某一历史研究作为对照时，需获取试验组和对照组每例受试者的基线数据，论证两组受试者的可比性，可采用倾向性评分来评估两组之间的可比性，以控制

选择偏倚。由于试验组和对照组不是同期开展，需要关注两组间干预方式和评价方式的一致性，以控制测量偏倚和评价偏倚。

三、受试对象

根据试验器械预期使用的目标人群，确定研究的总体。综合考虑对总体人群的代表性、临床试验的伦理学要求、受试者安全性等因素，制定受试者的选择标准，即入选和排除标准。入选标准主要考虑受试对象对总体人群的代表性，如适应症、疾病的分型、疾病的程度和阶段、使用具体部位、受试者年龄范围等因素。排除标准旨在尽可能规范受试者的同质性，将可能影响试验结果的混杂因素（如影响疗效评价的伴随治疗、伴随疾病等）予以排除，以达到评估试验器械效应的目的。

四、评价指标

评价指标反映器械作用于受试对象而产生的各种效应，根据试验目的和器械的预期效应设定。在临床试验方案中应明确规定各评价指标的观察目的、定义、观察时间点、指标类型、测定方法、计算公式（如适用）、判定标准（适用于定性指标和等级指标）等，并明确规定主要评价指标和次要评价指标。指标类型通常包括定量指标（连续变量，如血糖值）、定性指标（如有效和无效）、等级指标（如优、良、中、差）等。对于诊断器械，临床试验评价指标通常包括定性检测的诊断准确性（灵敏度、特异性、预期值、似然比、ROC 曲线下面积等）或检测一致性（阳

性/阴性一致性、总一致性、KAPA 值等），以及定量检测回归分析的斜率、截距和相关系数等。

（一）主要评价指标和次要评价指标

主要评价指标是与试验目的有本质联系的、能确切反映器械疗效或安全性的指标。主要评价指标应尽量选择客观性强、可量化、重复性高的指标，应是专业领域普遍认可的指标，通常来源于已发布的相关标准或技术指南、公开发表的权威论著或专家共识等。临床试验的样本量基于主要评价指标的相应假设进行估算。临床试验的结论亦基于主要评价指标的统计分析结果做出。次要评价指标是与试验目的相关的辅助性指标。在方案中需说明其在解释结果时的作用及相对重要性。

一般情况下，主要评价指标仅为一个，用于评价产品的疗效或安全性。当一个主要评价指标不足以反映试验器械的疗效或安全性时，可采用两个或多个主要评价指标。以一次性使用脑积水分流器的临床试验为例，当参照《一次性使用脑积水分流器注册技术审查指导原则》进行方案设计时，同时采用两个主要评价指标，包括术后30天内颅内压的达标率、首次植入分流器后1年时分流器存留率。对于第二个主要评价指标（1年存留率），试验组与对照组间需进行组间比较，同时要求试验组1年存留率不小于90%。因此，该临床试验的样本量估算需同时考虑三重假设检验：（1）试验组术后30天颅内压达标率非劣效于对照组；（2）试验

组1年的存留率非劣效于对照组；(3) 试验器械1年的存留率达到目标值要求。上述三重假设检验都有统计学意义时，才可下推断结论。由于此时没有意图或机会选择最有利的某次假设检验结果，因此可设定每次检验的I类错误水平等于预先设定的α，无需进行多重性校正。对于同时采用多个主要评价指标的临床试验设计，当有可能选择最有利的某次假设检验结果进行结论推断时，样本量估算需要考虑假设检验的多重性问题，以及对总Ⅰ类错误率的控制策略。

（二）复合指标

按预先确定的计算方法，将多个评价指标组合构成一个指标称为复合指标。当单一观察指标不足以作为主要评价指标时，可采用复合指标作为主要评价指标。以冠状动脉药物洗脱支架的临床试验为例，主要评价指标之一为靶病变失败率。靶病变失败定义为心脏死亡、靶血管心肌梗死以及靶病变血运重建三种临床事件至少出现一种，即为复合指标。以血液透析浓缩物的临床试验为例，采用透析达标率作为主要评价指标，“达标”的定义为透析前后 K^+、Na^+、Ca^{2+}、Cl^-、CO_2CP (二氧化碳结合力）或 HCO_3^-、pH 值均达到预先设定的临床指标数值。复合指标可将客观测量指标和主观评价指标进行结合，形成综合评价指标。临床上采用的量表 (如生活质量量表、功能评分量表等) 也为复合指标的一种形式。需在试验方案中详细说明复合

指标中各组成指标的定义、测定方法、计算公式、判定标准、权重等。当采用量表作为复合指标时，尽可能采取专业领域普遍认可的量表。极少数需要采用自制量表的情形，申请人需提供自制量表效度、信度和反应度的研究资料，研究结果需证明自制量表的效度、信度和反应度可被接受。需考虑对复合指标中有临床意义的单个指标进行单独分析。

（三）替代指标

在直接评价临床获益不可行时，可采用替代指标进行间接观察。是否可采用替代指标作为临床试验的主要评价指标取决于：①替代指标与临床结果的生物学相关性；②替代指标对临床结果判断价值的流行病学证据；③从临床试验中获得的有关试验器械对替代指标的影响程度与试验器械对临床试验结果的影响程度相一致的证据。

（四）主观指标的第三方评价

部分评价指标由于没有客观评价方法而只能进行主观评价，临床试验若必需选择主观评价指标作为主要评价指标，建议成立独立的评价小组，由不参与临床试验的第三者/第三方进行指标评价，需在试验方案中明确第三者/第三方评价的评价规范。

五、比较类型和检验假设

（一）比较类型

临床试验的比较类型包括优效性检验、等效性检验、非劣效

性检验。采用安慰对照的临床试验，需进行优效性检验。采用疗效/安全性公认的已上市器械或标准治疗方法进行对照的临床试验，可根据试验目的选择优效性检验、等效性检验或非劣效性检验。

优效性检验的目的是确证试验器械的疗效/安全性优于对照器械/标准治疗方法/安慰对照，且其差异大于预先设定的优效界值，即差异有临床实际意义。由于试验器械特征、对照和主要评价指标等因素的不同，部分优效性检验没有考虑优效性界值，申请人需论述不考虑优效性界值的理由。等效性检验的目的是确证试验器械的疗效/安全性与对照器械的差异不超过预先设定的等效区间，即差异在临床可接受的范围内。非劣效性检验的目的是确证试验器械的疗效/安全性如果低于对照器械，其差异小于预先设定的非劣效界值，即差异在临床可接受范围内。在优效性检验中，如果试验设计合理且执行良好，试验结果可直接确证试验器械的疗效/安全性。在等效性试验和非劣效性试验中，试验器械的疗效/安全性建立在对照器械预期疗效/安全性的基础上。

（二）界值

无论优效性试验、等效性试验或非劣效性试验，要从临床意义上确认试验器械的疗效/安全性，均需要在试验设计阶段制定界值并在方案中阐明。优效界值是指试验器械与对照器械之间的差异具有临床实际意义的最小值。等效或非劣效界值是指试验器

械与对照器械之间的差异不具有临床实际意义的最大值。优效界值、非劣效界值均为预先制定的一个数值，等效界值需要预先制定优侧、劣侧两个数值。

界值的制定主要考虑临床实际意义，需要被临床认可或接受。理论上，非劣效界值的确定可采用两步法，一是通过 Meta 分析估计对照器械减去安慰效应后的绝对效应或对照器械的相对效应 M1，二是结合临床具体情况，在考虑保留对照器械效应的适当比例 1-f 后，确定非劣效界值 M2（M2=f×M1）。f 越小，试验器械的效应越接近对照器械，一般情况下，f 的取值在 0～0.5 之间。制定等效界值时，可用类似的方法确定下限和上限。

（三）检验假设

试验方案需明确检验假设和假设检验方法，检验假设依据试验目的确定，假设检验方法依据试验设计类型和主要评价指标类型确定。附录 1 提供了部分试验设计和比较类型下的检验假设举例，供参考。

六、样本量估算

临床试验收集受试人群中的疗效/安全性数据，用统计分析将基于主要评价指标的试验结论推断到与受试人群具有相同特征的目标人群。为实现样本（受试人群）代替总体（目标人群）的目的，临床试验需要一定的受试者数量（样本量）。样本量大小与主要评价指标的变异度呈正相关，与主要评价指标的组间差

异呈负相关。

样本量一般以临床试验的主要评价指标进行估算。需在临床试验方案中说明样本量估算的相关要素及其确定依据、样本量的具体计算方法。附录 2 提供了样本量估算公式的样例，供参考。确定样本量的相关要素一般包括临床试验的设计类型和比较类型、主要评价指标的类型和定义、主要评价指标有临床实际意义的界值、主要评价指标的相关参数（如预期有效率、均值、标准差等)、Ⅰ类和Ⅱ类错误率以及预期的受试者脱落和方案违背的比例等。主要评价指标的相关参数根据已有临床数据和小样本可行性试验（如有）的结果来估算，需要在临床试验方案中明确这些估计值的确定依据。一般情况下，Ⅰ类错误概率 α 设定为双侧 0.05 或单侧 0.025，Ⅱ类错误概率 β 设定为不大于 0.2，预期受试者脱落和方案违背的比例不大于 0.2，申请人可根据产品特征和试验设计的具体情形采用不同的取值，需充分论证其合理性。

七、临床试验设计需考虑的其他因素

由于器械的固有特征可能影响其临床试验设计，在进行器械临床试验设计时，需对以下因素予以考虑：

（一）器械的工作原理

器械的工作原理和作用机理可能与产品性能/安全性评价方法、临床试验设计是否恰当相关。

（二）使用者技术水平和培训

部分器械可能需要对使用者进行技能培训后才能被安全有

效地使用，例如手术复杂的植入器械。在临床试验设计时，需考虑使用器械所必需的技能，研究者技能应能反映产品上市后在预期用途下的器械使用者的技能范围。

（三）学习曲线

部分器械使用方法新颖，存在一定的学习曲线。当临床试验过程中学习曲线明显时，试验方案中需考虑在学习曲线时间内收集的信息（例如明确定义哪些受试者是学习曲线时间段的一部分）以及在统计分析中报告这些结果。如果学习曲线陡峭，可能会影响产品说明书的相关内容和用户培训需求。

（四）人为因素

在器械设计开发过程中，对器械使用相关的人为因素的研究可能会指导器械的设计或使用说明书的制定，以使其更安全，更有效，或让受试者或医学专业人士更容易使用。

八、统计分析

（一）分析数据集的定义

意向性分析（Intention To Treat，简称ITT）原则是指主要分析应包括所有随机化的受试者，基于所有随机化受试者的分析集通常被称为ITT分析集。理论上需要对所有随机化受试者进行完整随访，但实际中很难实现。

临床试验常用的分析数据集包括全分析集（Full Analysis Set，FAS)、符合方案集（Per Protocol Set，PPS）和安全性数据

集（Safety Set，SS）。需根据临床试验目的，遵循尽可能减少试验偏倚和防止Ⅰ类错误增加的原则，在临床试验方案中对上述数据集进行明确定义，规定不同数据集在有效性评价和安全性评价中的地位。全分析集为尽可能接近于包括所有随机化的受试者的分析集，通常应包括所有入组且使用过一次器械/接受过一次治疗的受试者，只有在非常有限的情形下才可剔除受试者，包括违反了重要的入组标准、入组后无任何观察数据的情形。符合方案集是全分析集的子集，包括已接受方案中规定的治疗、可获得主要评价指标的观察数据、对试验方案没有重大违背的受试者。若从全分析集和符合方案集中剔除受试者，一是需符合方案中的定义，二是需充分阐明剔除理由，需在盲态审核时阐明剔除理由。安全性数据集通常应包括所有入组且使用过一次器械/接受过一次治疗并进行过安全性评价的受试者。

需同时在全分析集、符合方案集中对试验结果进行统计分析。当二者结论一致时，可以增强试验结果的可信度。当二者结论不一致时，应对差异进行充分的讨论和解释。如果符合方案集中排除的受试者比例过大，或者因排除受试者导致试验结论的根本性变化（由全分析集中的试验失败变为符合方案集中的试验成功），将影响临床试验的可信度。

全分析集和符合方案集在优效性试验和等效性或非劣效性试验中所起作用不同。一般来说，在优效性试验中，应采用全分

析集作为主要分析集，因为它包含了依从性差的受试者而可能低估了疗效，基于全分析集的分析结果是保守的。符合方案集显示试验器械按规定方案使用的效果，与上市后的疗效比较，可能高估疗效。在等效性或非劣效性试验中，用全分析集所分析的结果并不一定保守。

（二）缺失值和离群值

缺失值（临床试验观察指标的数据缺失）是临床试验结果偏倚的潜在来源，在临床试验方案的制定和执行过程中应采取充分的措施尽量减少数据缺失。对于缺失值的处理方法，特别是主要评价指标的缺失值，需根据具体情形，在方案中遵循保守原则规定恰当的处理方法，如末次观察值结转（Last Observation Carried Forward, LOCF）、基线观察值结转（Baseline Observation Carried Forward, BOCF）等。必要时，可考虑采用不同的缺失值处理方法进行敏感性分析。

不建议在统计分析中直接排除有缺失数据的受试者，因为该处理方式可能破坏入组的随机性、破坏受试人群的代表性、降低研究的把握度、增加Ⅰ类错误率。

对于离群值的处理，需要同时从医学和统计学两方面考虑，尤其是医学专业知识的判断。离群值的处理应在盲态审核时进行，如果试验方案中未预先规定处理方法，在实际资料分析时，需要进行敏感性分析，即比较包括和不包括离群值的两种试验结

果，评估其对试验结果的影响。

（三）统计分析方法

1.统计描述

人口学指标、基线数据一般需选择合适的统计指标(如均数、标准差、中位数等）进行描述以比较组间的均衡性。

主要评价指标在进行统计推断时，需同时进行统计描述。值得注意的是，组间差异无统计学意义不能得出两组等效或非劣效的结论。

次要评价指标通常采用统计描述和差异检验进行统计分析。

2.假设检验和区间估计

在确定的检验水平（通常为双侧 0.05）下，按照方案计算假设检验的检验统计量及其相应的 P 值，做出统计推断，完成假设检验。对于非劣效性试验，若 $P \leq \alpha$，则无效假设被拒绝，可推断试验组非劣效于对照组。对于优效性试验，若 $P \leq \alpha$，则无效假设被拒绝，可推断试验组临床优效于对照组。对于等效性试验，若 $P_1 \leq \alpha$ 和 $P_2 \leq \alpha$ 同时成立，则两个无效假设同时被拒绝，推断试验组与对照组等效。

亦可通过构建主要评价指标组间差异置信区间的方法达到假设检验的目的，将置信区间的上限和/或下限与事先制定的界值进行比较，以做出临床试验结论。按照方案中确定的方法计算主要评价指标组间差异的 $(1-\alpha)$ 置信区间，α 通常选取双侧 0.05。

对于高优指标的非劣效性试验，若置信区间下限大于-Δ（非劣效界值），可做出临床非劣效结论。对于优效性试验，若置信区间下限大于Δ（优效界值），可做出临床优效结论。对于等效性试验，若置信区间的下限和上限在（-Δ，Δ）（等效界值的劣侧和优侧）范围内，可做出临床等效结论。

对试验结果进行统计推断时，建议同时采用假设检验和区间估计方法。

3.基线分析

除试验器械及相应治疗方式外，主要评价指标常常受到受试者基线变量的影响，如疾病的分型和程度、主要评价指标的基线数据等。因此，在试验方案中应识别可能对主要评价指标有重要影响的基线变量，在统计分析中将其作为协变量，采用恰当的方法（如协方差分析方法等），对试验结果进行校正，以修正试验组和对照组间由于协变量不均衡而对试验结果产生的影响。协变量的确定依据以及相应的校正方法的选择理由应在临床试验方案中予以说明。对于没有在临床试验方案中规定的协变量，通常不进行校正，或仅将校正后的结果作为参考。

4.中心效应

在多个中心开展临床试验，可在较短时间内入选所需的病例数，且样本更具有代表性，结果更具有推广性，但对试验结果的影响因素更为复杂。

在多个中心开展临床试验，需要组织制定标准操作规程，组织对参与临床试验的所有研究者进行临床试验方案和试验用医疗器械使用和维护的培训，以确保在临床试验方案执行、试验器械使用方面的一致性。当主要评价指标易受主观影响时，建议采取相关措施（如对研究者开展培训后进行一致性评估，采用独立评价中心，选择背对背评价方式等）以保障评价标准的一致性。尽管采取了相关质量控制措施，在多中心临床试验中，仍可能出现因不同中心在受试者基线特征、临床实践（如手术技术、评价经验）等方面存在差异，导致不同中心间的效应不尽相同。当中心与处理组间可能存在交互作用时，需在临床试验方案中预先规定中心效应的分析策略。当中心数量较多且各中心病例数较少时，一般无需考虑中心效应。

在多个中心开展临床试验，各中心试验组和对照组病例数的比例需与总样本的比例基本相同。当中心数量较少时，建议按中心进行分层设计，使各中心试验组与对照组病例数的比例基本相同。

九、临床试验的偏倚和随机误差

临床试验设计需考虑偏倚和随机误差。偏倚是偏离真值的系统误差的简称，在试验设计、试验实施和数据分析过程中均可引入偏倚，偏倚可导致错误的试验结论。临床试验设计时应尽量避免或减少偏倚。

统计量的随机误差受临床试验样本量的影响。一方面，较大

的样本量可提供更多的数据，使器械性能/安全性评价的随机误差更小。另一方面，更大的样本量可能引入更大的偏倚，导致无临床意义的差异变得具有统计学意义。试验设计应该旨在使试验结果同时具有临床和统计学意义。

附录 1

检验假设举例

本附录中列举的检验假设和检验统计量，为特定试验类型、特定评价指标类型下的举例，有其适用范围和前提条件。

一、高优指标的两样本 t 检验

表 1 以高优指标的两样本 t 检验为例，列举了优效性试验、等效性试验、非劣效性试验的检验假设和检验统计量的计算公式。H_0 和 H_1 分别表示原假设和备择检验；T 和 C 分别表示试验组和对照组主要评价指标的参数 (如总体均数、总体率等)；$S_{\bar{d}}$为两组参数差值（T-C）的标准误；Δ表示界值，优效性界值用Δ表示，非劣效界值用-Δ表示，等效界值的优侧和劣侧分别用Δ和-Δ表示；$t/t_1/t_2$ 为检验统计量。

表 1 不同试验类型的检验假设和检验统计量

(以高优指标的两样本 t 检验为例)

试验类型	原假设	备择假设	检验统计量
非劣效性试验	$H_0: T-C \leq -\Delta$	$H_1: T-C > -\Delta$	$t=(T-C-(-\Delta)/S_{\bar{d}}$
优效性试验	$H_0: T-C \leq \Delta$	$H_1: T-C > \Delta$	$t=(T-C-\Delta)/S_{\bar{d}}$
等效性试验	$H_{01}: T-C \leq -\Delta$	$H_{11}: T-C > -\Delta$	$t_1=(T-C-(-\Delta)/S_{\bar{d}}$
	$H_{02}: T-C \geq \Delta$	$H_{12}: T-C < \Delta$	$t_2=(T-C-\Delta)/S_{\bar{d}}$

二、单组目标值试验的检验假设

π_0 为主要评价指标的目标值，π_1 为主要评价指标的总体率/均数。对于高优指标，检验假设为 $H_0:\pi_1 \leq \pi_0$，$H_1:\pi_1 > \pi_0$。对于低优指标，检验假设为 $H_0:\pi_1 \geq \pi_0$，$H_1:\pi_1 < \pi_0$。

附录 2

样本量估算公式举例

本附录中列举的样本量估算公式，为样本量估算公式举例，有其适用范围和前提条件。在实际的样本量估算中，需根据具体试验设计选择适用公式，包括本附录中未列举的公式。

一、平行对照设计样本量估算

以下公式中，n_T、n_C分别为试验组和对照组的样本量；$Z_{1-\alpha/2}$、$Z_{1-\beta}$为标准正态分布的分数位，当α=0.05时，$Z_{1-\alpha/2}$=1.96，当β=0.2时，$Z_{1-\beta}$=0.842；$(Z_{1-\alpha/2}+Z_{1-\beta})^2$=7.85

（一）优效性试验

当试验组和对照组按照 1:1 随机化分组，主要评价指标为事件发生率，其方差齐且不接近于 0%或 100%时，其样本量估算公式为：

$$n_T = n_C = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 [P_C(1-P_C) + P_T(1-P_T)]}{(|D| - \Delta)^2}$$

P_T、P_C 分别为试验组和对照组预期事件发生率；$|D|$为两组预期率差的绝对值，$|D| = |P_T - P_C|$；Δ 为优效性界值，取正值。

当试验组和对照组按照 1:1 随机化分组，主要评价指标为定量指标且方差齐时，其样本量估算公式为：

$$n_T = n_C = \frac{2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \sigma^2}{(|D| - \Delta)^2}$$

σ 为对照组预期标准差；$|D|$为预期的两组均数之差的绝对值，$|D| = |u_T - u_C|$；Δ 为优效性界值，取正值。

使用该公式计算样本量为 Z 值计算的结果，小样本时宜使用 t 值迭代，或总例数增加 2—3 例。

（二）等效性试验

当试验组和对照组按照 1:1 随机化分组，主要评价指标为事件发生率，其方差齐且不接近于 0%或 100%时，其样本量估算公式为：

$$n_T = n_C = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 [P_C(1 - P_C) + P_T(1 - P_T)]}{(\Delta - |D|)^2}$$

P_T、P_C 分别为试验组和对照组预期事件发生率；$|D|$为两组预期率差的绝对值，$|D| = |P_T - P_C|$；Δ为等效界值（适用于劣侧界值与优侧界值相等的情形），取正值。

当试验组和对照组按照 1:1 随机化分组，主要评价指标为定量指标且方差齐时，其样本量估算公式为：

$$n_T = n_C = \frac{2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \sigma^2}{(\Delta - |D|)^2}$$

σ 为对照组预期标准差；$|D|$为预期的两组均数之差的绝对值，$|D| = |u_T - u_C|$；Δ为等效界值（适用于劣侧界值与优侧界

值相等的情形），取正值。

使用该公式计算样本量为 Z 值计算的结果，小样本时宜使用 t 值迭代，或总例数增加 2—3 例。

（三）非劣效试验

当试验组和对照组按照 1:1 随机化分组，主要评价指标为预期事件发生率，其方差齐且不接近于 0%或 100%时，其样本量估算公式为：

$$n_T = n_C = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 [P_C(1-P_C) + P_T(1-P_T)]}{(|D| - \Delta)^2}$$

P_T、P_C 分别为试验组和对照组预期事件发生率；$|D|$为两组预期率差的绝对值，$|D| = |P_T - P_C|$，Δ为非劣效界值，取负值。

当试验组和对照组按照 1:1 随机化分组，主要评价指标为定量指标且方差齐时，其样本量估算公式为：

$$n_T = n_C = \frac{2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \sigma^2}{(|D| - \Delta)^2}$$

σ 为对照组预期标准差；$|D|$为预期的两组均数之差的绝对值，$|D| = |u_T - u_C|$；Δ为非劣效界值，取负值。

使用该公式计算样本量为 Z 值计算的结果，小样本时宜使用 t 值迭代，或总例数增加 2—3 例。

二、单组目标值试验的样本量估算

以下公式中，n为试验组样本量；$Z_{1-\alpha/2}$、$Z_{1-\beta}$为标准正态分布

的分数位，当α=0.05时，$Z_{1-\alpha/2}$=1.96，当β=0.2时，$Z_{1-\beta}$=0.842。

当主要评价指标为事件发生率，统计发生率的研究周期相同，且发生率不接近于 0%或 100%时，其样本量估算公式为：

$$n=\frac{\left[Z_{1-\alpha/2}\sqrt{P_0(1-P_0)}+Z_{1-\beta}\sqrt{P_T(1-P_T)}\right]^2}{(P_T-P_0)^2}$$

P_T 为试验组预期事件发生率，P_0 为目标值。

三、诊断试验的样本量估算

以抽样调查设计的诊断试验为例，其评价指标为灵敏度和特异度，用灵敏度计算阳性组的样本量，用特异度计算阴性组的样本量。

阳性组/阴性组样本量的估算公式为：

$$n=\frac{Z_{1-\alpha/2}^2P(1-P)}{\Delta^2}$$

公式中 n 为阳性组/阴性组样本量，$Z_{1-\alpha/2}$ 为标准正态分布的分位数，P 为灵敏度或特异度的预期值，Δ 为 P 的允许误差大小，一般取 P 的 95%置信区间宽度的一半，常用的取值为 0.05—0.10。