医疗器械临床试验设计指导原则

（征求意见稿）

二〇一七年十一月

目　录

一、适用范围

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或者验证的过程。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。本指导原则不适用于定制器械的临床试验设计，不适用于小样本可行性试验的设计。

二、临床试验的开展原则

在医疗器械设计开发过程中，确认最终产品符合临床使用的需要（包括安全性、有效性、适用范围/禁忌症、使用方法、使用注意事项等信息）是其重要环节。可采取多种方法实现确认过程，包括同品种产品的临床数据、模拟临床使用的功能试验（如利用离体动物组织模拟测试高频血管闭合设备的凝血功能，利用髋关节磨损试验机模拟测试髋关节假体的磨损性能等）、模型实验（如在人体下消化道模型中模拟插入电子下消化道内窥镜，以观察腔道内图像质量以及内窥镜的操控性能等）、动物实验（如将骨修复材料植入骨缺损动物模型中，观察其介导骨长入和自身降解特征的实验等）、体外诊断设备的比较研究试验以及临床试验等。按照保障受试者权益、保障实验动物福利的原则，上述确认方法的选择和开展顺序应恰当。

临床试验的目的是为临床评价提供临床数据，综合考虑

产品的非临床研究（如文献研究、性能研究、模拟临床使用的功能实验、模型实验、动物实验、体外诊断设备的比较研究试验等）数据，以评价产品的临床受益是否大于风险，产品的风险在现有技术水平上是否已得到合理控制，同时为临床医生和患者对器械使用的临床环境和方法提供重要信息。需考虑开展临床试验的产品包括但不限于下列情形：

（一）尚未在境内外批准上市、安全有效性未经医学证实的新产品；

（二）通过非临床研究难以确认产品临床使用的有效性和/或安全性。

（三）对于器械的工作原理、作用机理、适应证、临床疗效、不良反应等方面，目前尚不明确或存在争议的治疗类产品。例如植入式胃刺激器及电极导线，利用植入胃的电极与皮下部位的脉冲发生器对胃进行刺激以达到减肥或治疗糖尿病的作用,其工作原理、作用机理、适应证、临床疗效、不良反应等方面均不明确。例如含银盐敷料，在人体内的作用机理、不良反应尚无定论。

（四）器械的部分性能通过参与人体代谢的方式获得，或者器械在体内被吸收，且尚无公认的非临床研究方法可进行恰当模拟的情形。例如生物可吸收支架，在血管重建过程中可逐步被吸收，尚无公认的非临床研究方法可进行恰当模拟，故需要考虑开展临床试验。例如经抗生素浸渍的脑室导

管，其辅助抗菌或抗细菌定殖作用仅由体外实验证明，不能完全模拟人体颅内环境以及国内常见感染菌种等微生物大环境，因此，其辅助抗菌或抗细菌定殖作用、可能产生的不良事件需临床试验进行确认。例如腹腔、盆腔外科手术用防粘连产品，非临床研究不能模拟产品在人体内的使用效果和吸收效果，需考虑开展临床试验。

（五）对于产品设计和制造工艺复杂、仿制一致性难以确认的高风险医疗器械，例如粒子治疗设备、植入式心脏起搏器等，若申请人尚无同类产品在中国批准上市，需考虑开展临床试验。

三、医疗器械临床试验的设计原则和特点

临床试验是以受试人群（抽样）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，以推论试验器械在预期适用人群（总体）中的效应。由于医疗器械的固有特征，其试验设计亦有其自身特点。

（一）临床试验目的

临床试验应有明确的试验目的。临床试验目的决定了主要评价指标的选择、临床试验设计类型和比较类型，从而影响临床试验样本量。申办者可综合分析试验器械特征、非临床研究情况、同类产品上市情况和临床表现等因素，设定临床试验目的。将临床试验设置不同目的的情形举例如下：

1．当试验器械的安全性已通过非临床研究得以基本确

认，临床试验目的可设置为确认产品的有效性，同时观察产品的安全性。例如，MRI、CT 等影像类设备的试验目的可设置为评价产品的图像质量。例如，透析浓缩物通常为原料药或药用辅料进行简单物理混合而成，溶解后通过离子交换与人体作用，透析液不直接进入人体，对于成熟配方，其安全性已较为稳定。该产品临床试验目的为确认其有效性（主要评价指标为反映其有效性的复合指标），同时观察其安全性。

2、当试验器械的有效性已得到基本证实，临床试验目的可设置为确认产品的安全性，同时观察产品的有效性。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，试验目的为确认产品的安全性，观察产品的有效性。

3．当已上市器械增加适应证时，临床试验目的可设置为确认试验器械对新增适应证的有效性。例如，止血类产品在已批准适用范围（如普通外科、妇产科等）的基础上，增加眼科、神经外科、泌尿外科使用的适应证。

4．当已上市器械使用人群发生变化时，临床试验目的可设置为确认试验器械对新增使用人群的有效性。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。例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。

5．当已上市器械发生重大设计变更时，可根据变更涉

及的范围设置试验目的。例如冠状动脉药物洗脱支架平台花纹设计发生改变时，临床试验目的可为评价变化部分对于产品安全性和有效性的影响；

6．当已上市器械的使用环境或使用方法发生重大改变时，试验目的可设置为对使用环境和使用方法的确认。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。

对于进入创新医疗器械特别审批程序的产品，建议申请人充分利用产品注册申请受理前沟通路径，就临床试验设计与医疗器械技术审评中心进行充分沟通。

（二）器械临床试验设计需考虑的特殊因素

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：

1．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。

2．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须的技能，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。

3．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。

4．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事更容易使用。

（三）临床试验设计的基本类型和特点

1.平行对照设计

随机、双盲、平行对照的临床试验设计可确保临床试验影响因素在试验组和对照组间的分布趋于相似，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，被认为可提供最高等级的科学证据，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 的临床试验设计）或其他约定的概率（如试验组与对照组病例数为 n：1 的临床试验设计）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。

部分医疗器械的临床试验采用非随机设计，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。即使通过协变量分析对已知影响因素进行校正，仍存在未知影响因素对试验结果产生影响的可能。非随机设计并不减少临床试验的受试者风险及申办方成本，从风险受益的角度，通常不推荐非随机设计。如果申办方有强烈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方法。

（2）盲法

由于知晓分组信息，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。部分试验未设置独立的评价者，研究者和评价者为同一人担任。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，盲法分为完整设盲、设盲不完整和非盲（开放）设计。在完整设盲的临床试验中，受试者、研究者、

评价者对分组信息均应处于盲态。例如用于四肢和脊柱非结构性植骨的骨填充材料，可通过试验设计实现对受试者和评价者设盲，当试验产品和对照产品具有相同的外观和规格时，可通过屏蔽包装和标签实现对研究者设盲，从而实现完整设盲。受试者、研究者和评价者中的一方不处于盲态时，为设盲不完整。开放性临床试验中，所有人员都可能知道处理信息。

在很多情形下，基于器械和相应治疗方式的固有特征，完整设盲是不可行的。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、研究者设盲。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。例如血管内金属支架，因支架具体结构、花纹不同，难以对研究者设盲。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，难以对评价者设盲，例如骨科内固定产品，其在 X 线、CT 影像学图片中完整显影，而临床试验主要评价指标通常包括影像学数据（如术后 24 周骨折部位正侧位 X 线片上骨折间隙模糊或消失，或者正侧位 X 线片上可见连续性骨痂越过骨折线），因此，该类产品临床试验难以对评价者设盲。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。

申办方需要对设盲不完整或开放性试验设计的理由进行论述，详述控制偏倚的具体方法（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操作偏倚等）。

（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案。选择阳性对照时，优先采用已上市同类产品。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案。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、对相应适应证的疗效已得到证实并被公认的产品。例如：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案（经尿道前列腺电汽化术）作为对照。在试验器械尚无相同或相似的已上市产品或标准治疗方案时，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，例如用于缓解疼痛的物理治疗类设备。

2.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

治疗的分配可进行随机设计。该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不良反应与试验器械或对照器械的相关性，且面部左右侧局部反应的互相影响需要进行排除。

对于诊断类产品，若试验目的是评价试验器械的诊断可靠性，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金标准方法来进行诊断。

3. 交叉设计

在交叉设计的临床试验中，每位受试者或每例样品按照预先确定的排列顺序，在不同时间接收两种或两种以上的治疗（或诊断测试）。当一种治疗（或诊断测试）的影响不延续到下一种治疗（或诊断测试）时，可能适合采用交叉设计。对于这种类型的设计，除非另有合理说明，否则顺序应是随机的。交叉设计可以用于治疗类和诊断类器械的研究。

4. 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，单组试验可能存在选择偏倚、评价偏倚等，应审慎选择。当器械技术比较成熟且对其适用疾病有较为深刻的了解时，或者当试验器械技术尚不成熟，设置对照不可

行时，方可考虑采用单组设计。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。

单组试验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。目标值是专业领域内公认的某医疗器械的有效性/安全性评价指标所应达到的最低标准。目标值通常为二分类（如有效/无效）的客观评价指标，包括平均值（靶值）和 95%置信区间下限（高优指标）/95%置信区间上限（低优指标），包括客观性能标准（Objective performance criteria，OPC）和性能目标（Performance goal，PG）两种。

由于没有设置对照组，单组目标值临床试验无法证明试验器械的优效、等效或非劣效，仅能证明试验器械的有效性/安全性达到专业领域内公认的最低标准。除了单组设计的限定条件外，考虑单组目标值设计时，还应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。

（1）与 OPC 比较

当试验器械技术比较成熟、对其适用疾病有较为深刻的了解且可获取该类器械充分的临床研究数据时，可考虑 OPC 单组试验。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。OPC 的构建需要

全面收集具有一定质量水平及相当数量病例数的临床研究数据，在受试者水平上进行科学的荟萃分析。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(95%置信区间下限)应至少为 90%,预期达标率(靶值)为 95%。例如，常规设计的髋关节假体，当临床试验采用单组目标值设计时,主要评价指标通常采用术后12个月Harris评分“优良率”，试验产品“优良率”的目标值（95%置信区间下限）应至少为 85%，预期优良率（靶值）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予以确认。

（2）与 PG 比较

当试验器械技术尚不成熟，且设置对照不可行（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，或者现有治疗方法因客观条件限制不具有可行性）时，可考虑采用 PG 单组试验设计。PG 是在现有治疗方法的临床研究数据的基础上分析得出。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，如果在试验数据中发现不正常的信号，需要对试验结果进行进一步探讨和论证。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。

5.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、实施偏倚、评价偏倚等问题，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参数，论证两组受试者的可比性，以控制选择偏倚。申办者可通过科学设计入组方式以保障两组受试者的可比性。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，以控制实施偏倚和评价偏倚。

（四）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，如适应证、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。排除标准主要考虑受试对象的同质性，对可能影响试验结果的因素予以排除，以精确评估试验器械的效应。

（五）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、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。指标类型包括定量指标（可测量的连续性指标，如血糖值）、定性指标（如有效和无效）、等级指标（如优、良、中、差）等。

1．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、能确切反映器械作用效应的指标。主要评价指标应尽量选择客观性强、易于量化、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。临床试验通常设立单一试验目的，主要评价指标通常只有一个。当一个主要评价指标不足以反映试验器械的作用效应时，可采用两个或多个主要评价指标。当有多个主要评价指标时，样本量估算需要考虑假设检验的多重性问题，对总Ⅰ类错误率和总Ⅱ类错误率的控制策略。以脑积水分流器（脑室-腹腔分流器）的非劣效平行对照试验为例，当临床试验同时采用三个主要评价指标（包括术后30天颅内压达标率、植入后1年的存留率、试验器械1年的存

留率不小于 90%）时，其样本量估算需同时考虑试验组术后30 天颅内压达标率非劣效于对照组，试验组 1 年的存留率非劣效于对照组；试验器械 1 年的存留率达到目标值要求三种情形。

临床试验的样本量是基于主要评价指标的相应假定后进行估算的。临床试验的结论亦基于主要评价指标的统计结果做出。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。

2．复合指标

当单一观察指标不足以作为主要评价指标时，通常采取的方法是按预先确定的计算方法，将多个评价指标组合构成一个复合指标。以冠状动脉药物洗脱支架为例，临床试验的主要评价指标之一为靶病变失败率。靶病变失败率包括心脏死亡、靶血管心肌梗死以及靶病变血运重建，是由反映产品安全性和有效性的指标组合而成的复合指标。以血液透析浓缩物为例，临床试验时可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，多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，申办者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。可对复合指标中有临床意义的单个指标进行单独的分析。

3．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。

4.指标裁定

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人员进行指标裁定，需在试验方案中明确指标裁定的规则。

（六）比较类型和检验假设

1.平行对照

（1）比较类型

平行对照的比较类型包括优效性比较、等效性比较、非劣效性比较。采用安慰对照的临床试验，需进行优效性比较。

采用疗效/安全性公认的已上市器械或已有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比较、等效性比较或非劣效性比较。优效性试验包括从统计学角度提出的优效和从临床意义上提出的优效两种，通常情况下，临床优效性试验具有临床实际意义。临床优效性试验的目的是验证试验器械的疗效/安全性优于对照器械或安慰对照或空白对照，且其差异大于预先制定的优效界值，即差异有临床实际意义。等效性试验的目的是验证试验器械的疗效/安全性与对照器械的差异小于预先制定的等效界值，即差异在临床上无实际意义。非劣效性检验的目的是验证试验器械的疗效/安全性如果低于对照器械，其差异小于预先制定的非劣效界值，即差异在临床可接受范围内。在优效性试验中，如果其设计合理且执行良好，试验结果可直接证明试验器械的疗效/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。

（2）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。在优效性试验中，界值指的是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。在等效性试验或非劣效性试验中，界值指的是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

效性界值、非劣效性界值均为预先制定的一个数值，等效性界值需要预先制定优侧、劣侧两个数值。

制定非劣效界值可采用两步法，一是通过荟萃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 M_2（$M_2=f\times M_1$，$0<f<1$）。f 越小，试验器械的效应越接近对照器械。制定等效界值时，可用类似的方法确定下限和上限。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。

（3）检验假设

表 1 列举了不同试验类型下检验假设和检验统计量的计算公式。H_0 和 H_1 分别表示无效检验和备选检验；T 和 C 分别表示试验组和对照组主要评价指标的均数或率；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。

表 1 不同试验类型的检验假设和检验统计量

试验类型	无效假设	备选假设	检验统计量
非劣效性试验	H_0：T-C≤-δ	H_1：T-C>-δ	$t=[T-C-(-\delta)]/S_d$
优效性试验	H_0：T-C≤δ	H_1：T-C>δ	$t=(T-C-\delta)/S_d$
等效性试验	H_{01}：T-C≤-δ	H_{11}：T-C>-δ	$t1=[T-C-(-\delta)]/S_d$
	H_{02}：T-C≥δ	H_{12}：T-C<δ	$t2=(T-C-\delta)/S_d$

2.单组试验

单组试验为样本率与已知总体率的比较研究，P_0为主要评价指标的目标值（95%置信区间下限）,P_1为主要评价指标的总体率（靶值）。对于高优指标,检验假设为H_0：$P_1 \leqslant P_0$，H_1：$P_1 > P_0$。对于低优指标,检验假设为H_0：$P_1 \geqslant P_0$，H_1：$P_1 < P_0$。

（七）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。为实现抽样（受试人群）代替整体（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。

样本量以试验的主要评价指标来确定。需在临床试验方案中说明确定样本量的相关要素和样本量的具体计算方法。确定样本量的相关要素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、对照器械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。对照器械主要评价指标的相关参数根据已有临床数据或探索性试验的结果来估算，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能大于 0.2。

1.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，$Z_{1-\beta}$=0.845；$(Z_{1-\alpha/2}+Z_{1-\beta})^2$=7.87

（1）优效性试验

当主要评价指标为事件发生率且不接近于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 为两组的预期率差，对于高优指标：$D=P_T-P_C$，对于低优指标：$D=P_C-P_T$；Δ为优效性界值，取正值。

当主要评价指标为定量指标时的样本量估算的公式为：

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

σ 为对照组预期标准差；D 为预期的两组均数之差，对于高优指标：$D=u_T-u_C$；对于低优指标：$D=u_C-u_T$。

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。

（2）等效性试验

当主要评价指标为事件发生率且不接近于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_C、P_T分别为对照组和试验组预期的样本率；D 为两组的预期率差，$D=P_T-P_C$；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。

当效应指标为定量指标时的样本量估算的公式为：

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

其中 n 为每组的样本量；σ 为标准差；D 为两组预期的均数之差 $D=u_T-u_C$；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。

（3）非劣效试验

当主要评价指标为事件发生率且不接近于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_C、P_T分别为对照组和试验组预期的样本率；D 为两组的预期率差，$D=P_T-P_C$，如果指标为低优指标，则 $D=P_C-P_T$；Δ 为非劣效界值，界值Δ 取负值。

当效应指标为定量指标时的样本量估算的公式为：

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

σ 为标准差；D 为预期检测到的两组均数之差 $D=u_T-u_C$；Δ 为等效界值，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。

2.单组试验的样本量估算

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 $\alpha=0.05$时，$Z_{1-\alpha/2}=1.96$，当 $\beta=0.2$时，$Z_{1-\beta}=0.845$。

当主要评价指标为事件发生率且不接近于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为试验组预期的样本率，P_0为目标值。

3.诊断试验的样本量估算

诊断试验的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。

单个诊断试验样本含量的估算公式为：

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

公式中n为所需样本量,$Z_{\alpha/2}$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0.05-0.1。

（八）统计分析

1.分析数据集的定义

临床试验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。全分析集通常应包括所有入组且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，对于设盲的临床试验设计，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。当二者数据一致时，可以增强试验结果的可信度。当二者数据不一致时，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。

2．缺失值的填补

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避免数据缺失。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（LOCF）、基线观察值结转（BOF）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。

3.统计分析方法

①统计描述

人口学指标、基线数据和次要评价指标的数据，通常采用统计描述进行统计分析，如均数、标准差、中位数、t 检验、方差分析等。

主要评价指标在进行假设检验和置信区间分析前，亦先

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。

②假设检验和置信区间

在确定的检验水平（通常为双侧 0.05）下，计算表 1 中的检验统计量,查相应的界值表得 P 值,即可做出统计推断，完成假设检验。对于非劣效试验，若 $P \leqslant \alpha$，则无效假设被拒绝，可推论试验组非劣效于对照组。对于优效性试验，若 $P \leqslant \alpha$，则无效假设被拒绝，可推论试验组临床优效于对照组。对于等效性试验，若 $P_1 \leqslant \alpha$ 和 $P_2 \leqslant \alpha$ 同时成立，则两个无效假设同时被拒绝，前者可推论试验组不比对照组差，后者可推论试验组不比对照组好，综合推断试验组与对照组等效。

通过构建主要评价指标组间差异的置信区间，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。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。对于非劣效性临床试验，若置信区间下限大于 $-\delta$（非劣效界值），可得出非劣效结论。对于优效性试验，若置信区间下限大于 δ（优效界值），可作出临床优效结论。对于等效性试验，若置信区间的下限和上限在（$-\delta$，δ）（等效界值的劣侧和优侧）范围内，可得出临床等效结论。

对试验结果进行分析时，建议同时采用假设检验和区间

分析，以进行统计推断，得出试验结论。

④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数据的影响，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数据，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），对试验结果进行校正，以补偿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。

⑤中心效应

根据《医疗器械临床试验质量管理规范》（国家食品药品监督管理总局 中华人民共和国国家卫生和计划生育委员会令第 25 号),多中心临床试验是指按照同一临床试验方案，在三个以上（含三个）临床试验机构实施的临床试验。多中心临床试验可在较短时间内入选所需的病例数，且入选病例范围广，临床试验的结果更具代表性，但对试验结果的影响因素更为复杂。

多中心临床试验要求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试验，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，每个中心的样本数均较少时，一般无需考虑中心效应。

对于在两个临床试验机构开展的临床试验，如果遵循《医疗器械临床试验质量管理规范》中对于多中心临床试验的相关要求执行，临床试验按照同一试验方案在两个临床试验机构同期进行，保障临床试验方案执行和器械使用的一致性，可按照多中心临床试验的统计分析原则和方法进行统计分析。

对于在两家及以上临床试验机构开展的临床试验，各中心试验组和对照组病例数的比例原则上应与总样本的比例大致相同。

（九）临床试验的偏倚和抽样误差

临床试验设计需考虑偏倚和抽样误差。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试验偏倚。

抽样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，对器械性能/安全性评价的抽样误差更小。另一方面，更大的样本量可能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。