药物临床试验样本量估计指导原则

（试行）

2024 年 12 月

目 录

药物临床试验样本量估计指导原则

（试行）

一、 概述

样本量估计，又称样本量确定，是药物临床试验设计的重要组成部分，也是确保研究具有合理性、准确性、可靠性、完整性和科学性的重要手段。通常，临床试验的样本要有充分的代表性，纳入的样本量必须足够大，以可靠地回答研究假设所针对的目标人群的临床问题。

对于药物临床试验，当样本量估计相关参数设置缺乏依据或依据不充分时，会为样本量估计带来不确定性，增大试验失败的风险等。另外，若试验中样本量调整的方法不当或操作不当则会导致总 I 类错误率（FWER）膨胀、破坏试验完整性及引入偏倚等方面的问题。

为指导申办者进行科学合理的样本量估计，本指导原则主要阐述基于统计假设进行样本量估计时的参数设置和样本量调整等的基本考虑。本指导原则适用于以注册为目的的确证性临床试验。

二、 样本量估计的主要考虑

对于临床试验所要回答的科学问题，统计学主要根据估计目标对应的研究假设提出统计假设，通过假设检验进

行推断。样本量估计一般依据主要估计目标，原则上应保证整体检验在控制 FWER 的同时并具有足够的检验效能。某些情况下，可能需要对关键次要估计目标进行样本量估计，此时同样需要考虑 FWER 的控制。不同的假设检验方法有不同的样本量估计方法。正确的样本量估计应基于与估计目标相一致的试验设计和恰当合理的统计分析方法。应在方案中描述计算样本量的方法，以及在计算时使用的相关参数及其依据。

样本量估计需考虑的因素众多，一般包括：①试验设计；②检验水准和检验效能；③统计分析方法；④预期治疗效应及变异等。

（一） 试验设计

试验设计是样本量估计时需考虑的重要因素，通常包括但不限于比较类型、设计类型、变量类型、伴发事件及其处理策略、多重性调整策略、组间分配比例及随机化方法等。

对于优效设计，试验组效应估计值应优于对照组，且应根据临床获益确定达到优效的判断标准，并据此明确统计学假设。在非劣效性和等效性试验中，阳性对照药疗效以及相应的非劣效界值和等效界值是样本量估计中的关键参数；在制定非劣效界值和等效界值时，应考虑所采用历

史研究的估计目标与当前研究的差异，非劣效界值的确定方法建议参考《药物临床试验非劣效设计指导原则》。

当单臂设计采用目标值对照时，目标值的设定应基于专业领域具有共识或认可程度较高的标准，当采用历史对照或基于历史数据确定目标值时，应选择可靠的历史数据。

对于时间-事件类型的变量，最终决定检验效能的主要是随访中观测到的结局事件数，因此结局事件发生率、入组速度、入组时长、随访时长、脱落率等因素也会进一步影响最终样本量。

样本量估计需要考虑估计目标中伴发事件及其处理策略对治疗效应及变异参数的影响。疗法策略将伴发事件作为治疗（处理）的一部分，其对样本量估计的影响主要体现在治疗（处理）中伴发事件对治疗效应和变异的影响。在治策略与复合变量策略，直接影响结局变量的定义，进而影响治疗效应和变异。假想策略设想一种没有发生伴发事件的情景，通常伴随着一定的假设，相应的假设会直接影响治疗效应和变异。主层策略体现在估计目标人群属性的定义中，该策略下主层人群的识别方法、主层人群所占比例及相应人群预期的治疗效应和变异会影响最终的样本量估计。

若存在多重性问题，可能会涉及调整检验水准、调整

检验效能及调整统计分析方法等方面，在样本量计算时应考虑这些调整。例如若设置了期中分析，则在样本量估计时需考虑 FWER 的控制；确证性亚组的样本量估计应结合多重性策略加以考虑。

组间样本量的分配比例是样本量估计中需考虑的参数，药物临床试验常采用平衡设计，即各组样本量相同。当出于伦理考虑或其他合理理由需要降低某组或某些组的样本量的分配比例时，可采用非平衡设计。样本量在各组的分配比例直接影响最终样本量估计，研究方案中须明确说明。采用分层随机化时可能需考虑各层比例与目标人群保持一致以及分层因素可能导致的某个或某些层受试者例数稀疏等问题。

（二） 检验水准及检验效能

检验水准和检验效能是样本量估计中考虑的基本参数，须在方案中明确。

设置合适的检验水准可以达到控制 FWER 的目的。对于确证性试验，FWER 通常要求控制在单侧 0.025、双侧 0.05 以内。当涉及多重性问题时，名义检验水准的设置可参考《药物临床试验多重性问题指导原则（试行）》。

对于检验效能，通常设定不低于 80%，当涉及多重性问题时，需考虑其对检验效能的影响。对于析因设计，当

研究目的包含交互作用的检验时，若基于检验主效应计算样本量，则交互作用的检验效能可能不足。对于多中心试验，样本量和检验效能的计算通常基于各中心的组间治疗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，同时试验的实施流程应该尽可能标准化。

（三） 统计分析方法

恰当合理的统计分析方法，是科学的样本量估计的基础和前提，样本量估计前须确保所选择的统计分析方法与研究设计相匹配。例如，统计分析方法应适合研究设计的设计类型、比较类型、随机化方法、变量类型等。方案中需明确给出样本量估计所基于的统计分析方法，且原则上应与主分析方法相一致，否则需有合理的理由认为所依据的方法能满足主要估计目标所需的样本量，不会导致样本量低估。方案中应明确样本量估计的具体计算方法、工具，当采用统计模拟估计样本量时，模拟的参数设置、模拟方法、种子数以及模拟代码等应在相关文件中详细描述并在与监管机构沟通时递交。

需要注意的是，每种统计分析方法均有其相对应的假设，在选择统计分析方法时需充分评估所选择方法违背其假设时的风险，以及违背假设对样本量估计可能产生的影

响。统计分析方法中常见的假设有关于数据分布的假设和关于统计分析模型的假设等。

1. 数据分布假设

统计分析常需要对数据的分布进行假设，例如，连续变量的数据服从正态分布假设，时间-事件变量的数据服从指数分布假设等。需评估偏离数据分布假设的可能性，一方面当偏离风险较大时应采用对分布假设更稳健的分析方法或不依赖于分布假设的分析方法，另一方面在参数设置时，也应考虑偏离分布假设带来的治疗效应高估或变异低估的可能性。当采用非参数或半参数方法分析时，样本量估计基于方便计算的考虑可能会依赖一定的参数假设，例如生存分析 log-rank 检验，样本量估计时可能会假设生存数据服从指数分布，在应用时需考虑偏离假设所带来的风险。

当数据存在相关性并影响治疗效应或变异估计时，若忽略相关性，则可能对样本量估计和分析带来影响，以及带来 FWER 膨胀的风险。因此在样本量估计时需评估数据间相关性，并在方案中明确描述其大小和依据；在统计分析时也应考虑相关性。

2. 统计分析模型假设

统计分析模型通常会基于一系列假设，需关注模型的适用性，对模型假设是否成立应进行预先判断并进行事后

验证。若统计分析模型的假设存在不成立的风险，建议在计算样本量时，将此风险纳入考虑。此外，协变量会影响治疗效应及变异的估计，样本量估计时应考虑是否调整协变量。关于协变量调整可参考《药物临床试验协变量校正指导原则》。另外，对于分层随机，若存在某些层的样本量较少等情况时，校正分层因素或采用分层分析可能存在降低检验效能的风险，在设计时应进行考虑。

（四） 预期治疗效应及变异

在基于给定的统计分析方法进行样本量估计时，需在原假设和备择假设分别成立的情况下对各参数进行合理设置，主要包括对影响预期治疗效应及变异的相关参数设置。

参数设置须有充分依据，一般应基于历史数据（前期研究结果或已发表的数据等），并注意其临床意义与合理性。不建议出于减少样本量的目的，设置激进的参数。当参数设置无依据（无历史数据）或参数设置依据不充分（历史数据过少）时，建议先开展探索性试验获得所需参数。

对于所参考的历史数据，应充分评估当前拟开展研究的估计目标与历史数据相关研究估计目标之间的差异。当历史数据的相关研究与当前研究的估计目标相同或相近时，所提供的参数较为可靠；当差异在可接受范围时，建议基于合理假设设置参数；当差异较大时，建议进一步开展探

索性试验。由于人群、治疗（处理）或变量（终点）的定义会与伴发事件处理策略相关，因此应考虑伴发事件的处理策略对预期治疗效应及变异估计的影响，参数设置时，若假定的参数或参考的历史数据未能反映伴发事件的影响，则建议评估伴发事件的发生比例及其处理策略和可能带来的疗效损失，尽可能降低样本量低估的风险。此外，若当前拟开展研究与历史数据相关研究的统计分析方法之间存在差异，也需评估其对参数估计的影响。

样本量估计还应考虑缺失数据的影响，该影响可在治疗效应和变异的参数假设中考虑，也可采用对计算所得样本量增加一定比例等方法进行处理。

三、 样本量调整

样本量调整是指临床试验期间对初始设计的样本量所做的调整。样本量调整一般不建议减小样本量。应在方案中说明样本量调整的必要性和合理性，以及为保持盲态和试验完整性所采取的措施。

（一） 样本量调整的原则

只有合理的样本量调整才能达到提高试验效率的目的。应充分评估进行样本量调整的必要性、合理性、可行性，并确保试验的完整性，通常基于可行性及最小临床意义差别或可接受的最大方差设置最大可接受样本量。

（1）必要性

不应无根据地随意调整样本量，需充分评估样本量调整的必要性。当历史数据过少导致参数设置的依据不足时，可通过开展探索试验获得所需的数据；如果历史数据较充分可靠，相比样本量重新估计设计，固定样本量设计可在保证检验效能的情况下，具有节省期中分析成本、避免期中分析可能带来试验完整性破坏的风险、效率更高等优点。对于非劣效性和等效性试验，阳性对照药参数较明确，缺乏调整的必要性且样本量调整可能存在 I 类错误率膨胀的问题，建议谨慎考虑样本量调整。

（2）合理性

样本量调整应以控制 FWER 和保证试验完整性为前提，保证调整的合理性并根据调整方法制定正确的统计分析方法。与样本量计算的原则一致，调整后的样本量应避免有统计学意义但没有临床意义的情形。如果根据试验本身累积的数据进行样本量重新估计，建议选择合理的调整时间节点，不建议过早或过晚进行样本量调整。过早进行调整，可能会面临由于数据较少导致调整所基于的参数可靠性不足；过晚进行调整，可能面临入组结束等试验实施方面的问题。

（3）可行性

样本量调整需要考虑可行性。多次调整会增加设计和实施的复杂性，并存在引入操作偏倚的风险，一般不建议进行多次样本量调整；入组结束后的样本量调整会给临床试验实施带来挑战，因此样本量调整需要评估试验入组进度，以及数据收集和数据清理的进度和速度。另外，如果试验终点的观测需要随访的时间较长，还需要评估样本量调整对试验整体时间的影响。

（4）完整性

样本量调整应避免引入操作偏倚并保持试验的完整性。当根据试验累积数据进行样本量重新估计时应严格按照方案、统计分析计划和/或包含样本量调整的期中分析计划中预先设定的规则进行调整。

对于不使用试验分组信息且在双盲试验中进行的盲态样本量重新估计一般不会引入操作偏倚，可由申办者或数据监查委员会（DMC）及其独立统计团队完成；其他情况下的盲态样本量重新估计建议由 DMC 及其独立统计团队完成，以保持试验的完整性。

非盲态的样本量重新估计需由 DMC 及其独立统计团队完成。由于涉及非盲的数据和结果，期中分析的执行应是一个完全保密的过程。参与试验实施的所有人员（包括研究者、申办者项目组成员和申办者所雇佣的人员等）及受

试者应当对这些分析结果保持盲态，否则可能会导致招募患者的特征改变、患者依从性降低、入组速度降低及产生治疗组间比较的偏倚等方面的问题。研究者应仅被告知继续或终止试验的决定，或实施修订试验方案的决定。DMC向申办者提出的建议中也应避免提供可以反推疗效的相关内容，以保证试验的完整性。

单臂试验一般不建议进行样本量调整。单盲试验和开放试验的样本量重新估计，建议由 DMC 及其独立统计团队完成，并注意试验完整性以及盲态保持。

（二） 样本量调整的情形

常见样本量调整的情形包括基于外部数据的调整和基于内部数据的调整。

1. 基于外部数据的调整

指基于外部试验（非本试验）的数据修订初始设计的样本量估计相关参数（例如影响治疗效应和变异的参数），进而调整相应的样本量。此类调整应充分考虑外部数据的可靠性。

2. 基于内部数据的调整

又叫样本量重新估计，指依据方案预先设定的期中分析计划，利用本试验累积的数据重新计算样本量，以保证最终的统计检验能达到预先设定的目标或修改后的目标，

并同时能够控制 FWER。

应在方案中明确样本量重估的时间节点、样本量重估的决策规则、最大可接受样本量、样本量重估的方法、FWER 控制方法以及调整后各组点估计及区间估计的计算方法，组间差异的点估计和区间估计的计算方法以及组间比较时统计量及 P 值的计算方法。

建议在期中分析计划中明确样本量重估的具体决策规则和期中分析报告应包含的内容，并在相关文件中明确期中分析报告的接触权限，期中分析数据的传输。若期中分析计划涉及样本量重估中反推疗效的细节，建议对期中分析计划设置访问权限，避免非必要人员知晓相关内容，例如可限定仅撰写样本量重新估计计划和审批的统计师、DMC及其独立统计团队有访问权限。建议在DMC章程中明确 DMC 向申请人提出样本量重估时应遵循的程序。

相比盲态样本量重新估计，非盲态样本量重新估计可能引入的偏倚、FWER 膨胀、及破坏试验完整性的风险更高，在设计和实施时更需谨慎。是否采用非盲态样本量重新估计需要考虑多种因素。例如，若有比较可靠的前期数据，非盲态下样本量重新估计是否必要；采用非盲态样本量重估所付的代价（例如检验水准调整）与初始设计时略微放大样本量相比，是否获益；期中分析能否很快完成，是否

可能因为入组较快而导致没有充足时间用来调整试验；期中分析的时间节点和推断方法是否合理；现有数据能否支持进行计划内的期中分析等。因此，应根据试验本身的特点，仔细考虑各种因素，然后做出合适的决策。

（1）盲态样本量重新估计

对于盲态样本量重新估计，针对样本量重新估计的期中分析不对数据揭盲，不使用实际试验分组信息，或未做任何涉及组间疗效比较。通常是在假定初始设定的组间差异不变的情况下对其他参数（例如事件发生率、变异等）进行重新估计来调整样本量。期中分析时因不涉及组间疗效比较，一般不会导致 FWER 膨胀。

（2）非盲态样本量重新估计

对于非盲态样本量重新估计，针对样本量重新估计的期中分析使用试验分组信息，分析内容涉及组间疗效比较。通常根据试验累积数据以及分组信息，计算样本量的重要参数（例如预期治疗效应），然后对样本量进行重估，因期中分析涉及组间疗效比较，通常导致 FWER 膨胀，需要对 I 类错误率进行控制。

常见的样本量调整决策规则有：①基于条件检验效能或试验成功率，当条件检验效能或试验成功率落在某个区间时对样本量进行重新估计，否则不对样本量进行调整；

②基于期中分析的组间治疗效应差异，当差异落在某个区间时对样本量进行重新估计，否则不对样本量进行调整。

与常用调整决策规则相对应的计算调整后样本量的方法通常有：①基于条件检验效能或试验成功率，使得根据调整后的样本量计算所得条件检验效能或试验成功率达到预先设定的要求；②基于期中分析得到的治疗效应与方案初始设计的治疗效应的比值调整样本量。

四、 其他

（一） 试验实施过程中的考虑

申办者应按照计划的样本量完成临床试验，除了出于伦理原因提前终止或出于检验效能不再可接受而放弃外，原则上不得随意终止试验。试验过程中不得随意增加或减少样本量，样本量调整也需在方案中明确并按照方案进行实施。

申办者应评估样本量重新估计中是否存在反推疗效、破坏试验完整性的风险，若存在须采取必要措施和方法加以防范。

（二） 基于贝叶斯方法的样本量估计

基于贝叶斯方法的样本量估计除先验分布的设置外，其余考虑因素与传统样本量估计考虑的因素一致。设计时应在全面严谨的模拟研究基础上，充分评估先验信息的合

理性，确定合理的拒绝原假设的决策规则，使方法满足控制 FWER 的要求。同时可使用其他合理的先验信息进行敏感性计算。建议就参数设置、FWER 控制及模型等与监管机构进行沟通。

（三） 基于其他目的的样本量估计

样本量一般基于主要估计目标来估计，如果由其他因素确定（例如，根据安全性评价或重要的次要估计目标确定样本量），应在方案中说明理由，并详细描述估计方法。最终所确定的样本量不能比基于主要估计目标所需样本量小。

（四） 样本量的敏感性计算

样本量估计时需考虑的因素众多，而历史数据通常相对有限，申办者一般需要进行各种偏离假设的计算，即敏感性计算，使各参数在一定合理范围内取值，或使参数取值来自某一分布，提供偏离假设的样本量范围，以在一定程度上降低样本量估计的不确定性。基于敏感性计算结果结合保守原则，用以指导确定合理的样本量，是一种谨慎和较稳妥的做法。

（五） 样本量重新估计与其他适应性设计的结合

当样本量重新估计与其他适应性设计相结合或与其他

多重性问题相结合时，由于设计的复杂性，尤其需要考虑较高的破坏试验完整性的风险等问题，通常需要严谨的理论及模拟确定调整的时间点和决策规则及其他运行特性，以证明这些规则满足样本量调整所需满足和达到的要求，例如控制FWER、达到检验效能、避免过早或者过晚（例如入组结束后）进行样本量调整等。需要慎重考虑是否有必要进行多重适应性调整，并在确认必要性后，严格遵守适应性设计的合理性、完整性和可行性的原则进行设计和执行，并与监管机构进行全面细致的沟通。

（六） 与监管机构的沟通

鼓励申办者在研究开始前，与监管机构就关键性临床试验中的样本量进行沟通，方案中应明确样本量计算时使用的所有参数及相应依据。建议申办者在沟通时对试验设计、检验水准及要达到的检验效能、统计分析方法、预期效应及变异参数设置和样本量计算方法等进行详细说明。如采用模拟的方法进行样本量估计，建议递交模拟代码和模拟方法的详细说明，包括但不限于模拟参数的假设及依据、模拟的种子数、模拟次数等。

涉及样本量调整的方案应在相关方案确定后尽快与监管机构沟通并提供调整依据，建议在沟通方案时一并递交（单独的）样本量调整计划，包括但不限于样本量计算或

模拟代码及结果， FWER 控制策略，调整后各组点估计及区间估计的具体计算过程，组间差异的点估计和区间估计的具体计算过程以及组间比较假设检验统计量及 P 值的具体计算过程等。

五、 参考文献

[1] 国家药品监督管理局药品审评中心. 药物临床试验多重性问题指导原则（试行）. 2020

[2] 国家药品监督管理局药品审评中心. 药物临床试验非劣效设计指导原则. 2020

[3] 国家药品监督管理局药品审评中心.药物临床试验盲法指导原则（试行）. 2022

[4] 国家药品监督管理局药品审评中心. 药物临床试验随机分配指导原则（试行）. 2022

[5] 国家药品监督管理局药品审评中心. 药物临床试验适应性设计指导原则（试行）. 2021

[6] 国家药品监督管理局药品审评中心. 药物临床试验数据管理与统计分析计划指导原则. 2021

[7] 国家药品监督管理局药品审评中心. 药物临床试验协变量校正指导原则（试行）. 2020

[8] 国家药品监督管理局药品审评中心. 药物真实世界研究设计与方案框架指导原则（试行）. 2023

[9] ICH E1a: The Extent of Population Exposure to Assess Clinical Safety: For Drugs Intended for Long-term Treatment of Non-Life-Threatening Conditions. 1995.

[10] ICH E8(R1): General Considerations for Clinical Trials. 2022.

[11] ICH E9: Statistical Principles for Clinical Trials. 1998.

[12] ICH E9(R1): Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials. 2019.

[13] U.S. Food and Drug Administration. Adaptive Designs for Clinical Trials of Drugs and Biologics. 2019.

附录 1 名词解释

I类错误（Type I Error）: 指原假设（或称无效假设）正确但检验结果拒绝了原假设的错误，相当于把实际上无效的药物经统计推断得出有效结论的错误。其概率需控制在某一水平，该水平称为检验水准，或称显著性水准，用α表示。

检验效能（Power）: 指原假设不成立时检验结果能拒绝原假设的概率。假设检验的把握度用1-β表示，β为II类错误率。

名义检验水准（Nominal Level）: 对于多重检验中某一假设检验的检验水准称之为名义检验水准，又称局部检验水准，用α_i表示。

总I类错误率（Familywise Error Rate，FWER）: 是指在同一临床试验所关注的多个假设检验中，至少一个真的原假设被拒绝的概率。其应控制在合理水平。

多重性问题（Multiplicity Issues）: 指在一项完整的临床试验中，需要经过不止一次统计推断（多重检验）对研究结论做出决策的相关问题。

估计目标（Estimand）: 对治疗效应的精确描述，反映了针对临床试验目的提出的临床问题。它在群体水平上汇总比较相同患者在不同治疗条件下的结局。

伴发事件（Intercurrent Event）: 治疗开始后发生的事件，可影响与临床问题相关的观测结果的解释或存在。在描述相关临床问题时，需明确伴发事件的处理策略，以便准确定义需要估计的治疗效应。

预期效应/差异（Treatment Effects of Interest，Treatment Difference to be Detected）: 研究旨在有一定检验效能去发现的关于分析指标的组间差异大小。

样本量重新估计（Sample Size Re-estimation）: 是指依据预先设定的期中分析计划，利用累积的试验数据重新计算样本量，以保证最终的统计检验能达到预先设定的目标或修改后的目标，并同时能够控制总I类错误率。

附录 2 中英文词汇对照

中文	英文
准确性	Accuracy
条件检验效能	Conditional Power
数据监查委员会	Data Monitoring Committee, DMC
估计目标	Estimand
事件发生率	Event Rates
总 I 类错误率	Familywise Error Rate, FWER
完整性	Integrity
伴发事件	Intercurrent Event
期中分析	Interim Analysis
名义检验水准	Nominal Level
最小临床意义差别	Minimum Clinically Important Difference
多重性	Multiplicity
试验成功率	Probability of Success
检验效能	Power
可靠性	Reliability
检验水准	Significance Level
样本量调整	Sample Size Adjustment
样本量重新估计	Sample Size Re-estimation
敏感性计算	Sensitivity calculations
统计分析计划	Statistical Analysis Plan, SAP
预期治疗效应/差异	Treatment Effects of Interest, Treatment Difference to be Detected

中文	英文
I 类错误	Type I Error
II 类错误	Type II Error
合理性	Validity