· 专家共识 ·

医学影像人工智能医疗器械临床试验统计学设计要点专家共识

主要执笔：郭秀花[1]　李　卫[2]

共识专家：夏结来[3]　贺　佳[4]　王　杨[2]　罗艳侠[1]　徐　涛[5]　阎小妍[6]　陶丽新[1]　赵延延[2]　尚美霞[6]　刘之光[7]

（代表中国医疗器械行业协会医学数据分析专业委员会）

【提　要】 围绕医学影像的人工智能医疗器械，介绍其临床试验的统计学设计要点，主要涉及临床试验设计类型确定、对照设置、比较类型选择、主要结局指标确定、样本量估计方法、随机化策略、盲法实施、缺失值处理及敏感性分析等内容。本共识旨在为从事该类医疗器械临床试验的相关人员提供统计学设计参考，以保证临床试验结果的科学性、可靠性及规范性。

【关键词】 医学影像　医疗器械　临床试验　统计学设计　专家共识

【中图分类号】 R195　　**【文献标识码】** A　　**DOI** 10.11783/j.issn.1002-3674.2024.03.031

医学影像（medical image）是医学影像设备产生的医学图像数据（如 X 射线、CT、MRI、超声、内窥镜、光学等图像）。医学影像是医疗诊断的工具之一，在临床疾病的诊断中发挥着重要的作用[1]。人工阅片是临床实践中使用医学影像进行诊断的主要方式，由于阅片量大、经验差异、阅读疲劳或视觉错觉等影响，常导致阅片结果不够准确[2]。以深度学习为核心的人工智能（artificial intelligence，AI）学习能力强大，不存在疲劳、错觉等问题，可明显提高医学影像的诊断效率和准确性。因此，基于医学影像的 AI 器械已被广泛应用在疾病筛查、决策及诊断等方面。

近年来，由于算法的进步、图形处理计算能力的增强和数据量的增加，基于深度学习的方法已广泛应用于超声、计算机断层扫描（computed tomography，CT）和磁共振成像（magnetic resonance imaging，MRI）等影像的自动识别，在医学成像领域取得了前所未有的成功[3-5]。有研究基于甲状腺超声图像开发的诊断桥本氏甲状腺炎的 HTNet 深度学习模型在准确性等方面的表现优于放射科医生[6]。还有研究结合深度学习和传统放射组学，根据癌症风险对肺结节进行分类，将得到的预测算法与 Herder 等人开发的肺癌诊断领域的经典模型集成，在测试集中对 Herder 评分为中低风险但实际为恶性结节的案例进行重新分类，其中 82% 的恶性结节被正确识别为高风险，该模型的应用可为及时干预提供依据[7]。目前，多个使用 AI 技术的医疗器械产品注册上市，如结肠息肉、肺结节、糖尿病视网膜病变等辅助识别筛查软件，该类产品能辅助临床医师提高诊断效能。

在医学影像 AI 器械开发过程中，确认最终产品的安全性和有效性，符合临床应用的要求是其重要环节。临床试验可以为医疗器械的安全性和有效性评价提供证据[8]。在医学影像 AI 器械的临床试验中，应遵循《医疗器械临床试验质量管理规范》，数据采集考虑采集设备、采集过程、数据脱敏等质控要求；建立数据采集操作规范；并考虑一定数量的灰区数据等，以确保 AI 器械在临床实践应用的适用性。同时，合理的临床试验统计学设计能够保证最大限度地控制偏倚、减少试验误差、提高试验质量。目前医学影像 AI 器械临床试验开展的越来越多，国家药品监督管理局医疗器械技术审评中心对此发布了《人工智能辅助检测医疗器械（软件）临床评价注册审查指导原则》[9]。由于该类 AI 器械统计学设计具有特殊性，本文在该指导原则基础上，针对医学影像 AI 器械临床试验统计学设计形成专家共识，旨在为临床试验相关人员提供统计学设计参考，以保证临床试验结果的科学性、可靠性及规范性。

临床试验设计类型及对照的设置

医学影像 AI 器械临床试验中研究对象是符合定义明确的入选/排除标准，前瞻性或回顾性收集的目标人群的影像样本。基于实时影像的 AI 器械考虑采用前瞻性设计，基于非实时影像的 AI 器械可考虑采用前瞻性或回顾性设计。研究设计阶段应考虑制定控制偏倚的策略，制定明确的入选/排除标准，连续收集目标人群的影像样本；基于目标疾病的流行病学特征，考虑阳性样本和阴性样本选取的合理性；制定临床参考标准、测量方法及标准等，以避免选择偏倚、临床参考标

1. 首都医科大学公共卫生学院；临床流行病学北京市重点实验室（100069）
2. 国家心血管病中心医学统计部
3. 空军军医大学卫生统计教研室
4. 海军军医大学卫生统计学教研室
5. 北京协和医学院基础学院
6. 北京大学第一医院
7. 首都医科大学附属北京安贞医院药物临床试验机构

准、测量、回忆及操作等信息偏倚等。

在开展医学影像 AI 器械临床试验时，应结合器械实际情况选择合理的试验设计类型和对照类型。常用的设计方法包括随机平行对照设计、交叉自身对照设计[10]、多阅片者多病例设计（multiple reader multiple case，MRMC）[11-12]、自身配对设计、单组设计和适应性设计[13-14]等。

1. 随机平行对照设计

随机平行对照设计是临床试验中应用最广泛的设计方法之一，可应用于评价医学影像 AI 器械安全性和有效性的临床试验设计中[15]。考虑到 AI 器械的独特性，可考虑采用随机、平行对照、开放、评估者盲法临床试验设计。通过随机分组确保试验组和对照组间的各项影响因素分布均衡；评估者盲法确保评估者无法知晓分组信息，有效避免选择偏倚和评价偏倚，保证试验结果的准确性和可靠性。

在随机平行对照设计中，对照应采用目标疾病病灶诊断或检出的“金标准”或已上市的同类诊断器械，如肺结节辅助检测产品、骨折 CT 影像辅助检测产品等。在目标疾病病灶诊断或检出没有明确的“金标准”或标准诊断方法时，通常考虑采用临床认可的高年资临床专家组综合意见作为“临床参考标准”[9]。

2. 交叉自身对照设计

自身对照设计是医学影像 AI 器械试验的主要设计类型[16]。交叉设计是自身对照设计的一种特殊形式。交叉自身对照设计是指按照事先设定的试验次序，对同一受试者影像样本在不同试验阶段逐一实施各种诊断方法，以比较不同诊断方法间的诊断性能差异。交叉自身对照设计能有效地控制阅片顺序偏倚。

交叉自身对照设计时应考虑设置一定长度的洗脱期，以消除不同诊断方法残留效应（carryover effect）[17]的影响。在医学影像 AI 器械临床试验中，“交叉”一般具体指对同一受试者影像样本在不同试验阶段采用不同阅片方式进行诊断；设定一定长度的洗脱期（一般不少于 4 周）[9,11]是为了洗脱阅片者对影像样本的记忆，确保阅片者对同一受试者影像样本的两次诊断结果互不影响。例如在目标疾病病灶人群中，比较某一 AI 辅助医生诊断技术与常规医生人工阅片诊断方法的诊断性能，且检查顺序对统计分析没有影响，可考虑采用交叉设计，一般为最简单的两阶段交叉设计（如图 1 所示），该类设计实施时，通常会将受试者影像样本随机分为两组，其中一组在入组时先由医生人工独立阅片，经过一定时间的洗脱期后再由 AI 辅助医生阅片；另一组采用相反的顺序。此过程中受试者影像样本阅片的顺序随机且所有受试者影像样本都接受了两种阅片方法进行诊断。

在交叉自身对照设计中，对照应采用目标疾病病灶诊断或检出的“金标准”、已上市的同类诊断器械、临床认可的“临床参考标准”（如高年资临床专家组成的人工阅片）。

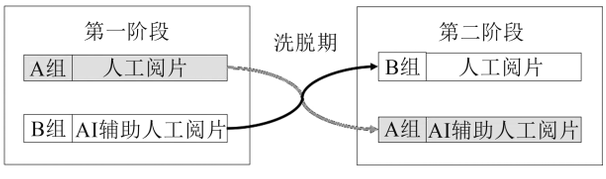
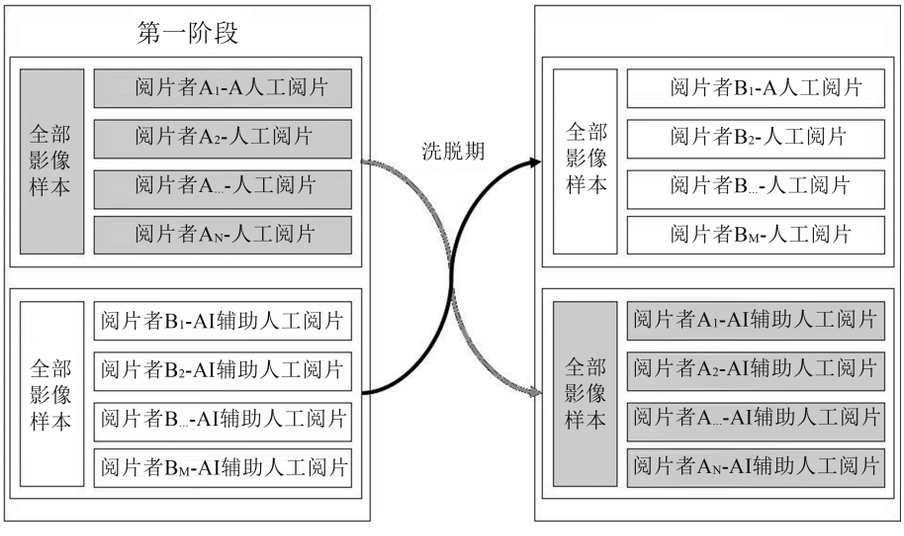


图 1 两阶段交叉设计

3. 多阅片者多病例（MRMC）设计

除上述临床试验设计方法外，美国食品药品监督管理局及中国国家药品监督管理局医疗器械技术审评中心建议医学影像辅助决策设备的临床性能评估可采用 MRMC 试验设计[11-12]。



注：阅片者随机分为A、B组，每组对应不同的阅片模式顺序

图 2 完全交叉的 MRMC 设计

MRMC 设计是一种适用于医学影像 AI 器械临床试验的新的设计类型。MRMC 设计作为评价不同影像诊断软件诊断性能的常用方法，可以根据用户的差异性，有效避免因读片医生、使用场景等不同而产生的临床评价过程中的读片偏倚，从而更加客观地评价试验结果[18]。采用此类设计，多名阅片者将评估多种阅片模式下的影像样本（如人工独立阅片与 AI 辅助人工阅片），通常应有至少 5 名阅片者[16]。

MRMC 设计通常采用完全交叉设计，即所有阅片者在每一种阅片模式下，对所有受试者影像样本给出对应的诊断结果（如图 2 所示）。在对影像样本进行随机后，一组阅片者先采用人工阅片，后采用 AI 辅助人工阅片的方式阅读所有影像样本；另一组阅片者先采用 AI 辅助人工阅片，后采用人工阅片的方式阅读所有影像样本。一般要求两组阅片人数相等，也可对阅片者进行随机分组。相比其他 MRMC 设计类型，完全交叉的 MRMC 设计具有最大统计检验效能的特点[18]，且研究结论能同时外推到阅片者和目标疾病总体人群中。为避免同一受试者影像样本多次阅片产生的偏倚，一般建议洗脱期至少为 4 周[11]。

4. 自身配对设计

在自身配对设计试验中，同一个受试者影像样本接受两种不同诊断方法进行诊断，来评估新的诊断技术的诊断性能及准确性。配对设计在保证受试者基线一致性方面上比平行对照设计具有优势，能够更好地控制混杂因素，消除不同受试者间的差异及影响。该试验设计存在一定的局限性，如阅片者顺序偏倚等信息偏倚。

5. 单组设计

在医学影像 AI 器械临床应用中，存在人工智能辅助分诊、转诊的需求。此类情况下可考虑单组设计。与传统医疗器械临床试验类似，单组设计存在非同期对照的固有偏倚，选择单组设计进行医学影像 AI 器械临床试验时应十分谨慎，详细内容可参考《单组目标值临床试验的统计学考虑》[19]。

6. 适应性设计

适应性设计是指按照预先设定的计划，在期中分析时使用试验期间累积的数据对试验做出相应修改的临床试验设计。适应性设计方法可参考《药物临床试验适应性设计指导原则（试行）》[20]。美国食品药品监督管理局发布的《医疗器械临床研究的适应性设计指导原则》讨论了在医疗器械临床试验中面临的适应性设计问题[21]。在医学影像 AI 器械的诊断试验中，适应性设计的主要目的是通过重新估计各种参数，特别是在样本量和分配比例上的动态调整，以优化研究设计、提高试验的效率和质量[22]。例如在自然人群的诊断试验中，通过适应性设计调整患病率，从而调整总体样本量（通常是在高估患病率的情况下）；在抽样设计中，根据试验期中分析结果或者试验外部证据，调整原先的灵敏度和特异度的参数，进而调整阳性、阴性受试者的例数等。此类设计虽然设计灵活，但设计要求高，需要有配套策略来保证试验的完整性。

比较类型选择

临床试验的比较类型分为优效、等效、非劣效。在医学影像 AI 器械临床试验设计中，优效和非劣效是最常用的两种比较类型。例如，在加载设计研究中，比较常规医生人工加载 AI 器械联合阅片和医生人工独立阅片两种阅片方法诊断效果，此种情况应考虑优效；预期以提高诊断时间效率为首要目标的 AI 器械，其临床试验设计可选择用户结合该器械联合决策与用户单独决策进行交叉自身对照设计，以灵敏度、特异度、时间效率作为共同主要评价指标，其中灵敏度、特异度可考虑采用非劣效或优效，时间效率指标应考虑采用优效。

优效的目标在于确证试验器械在有效性/安全性上优于对照器械，且其差异大于预先设定的优效界值（一般设为 0）。如果结合临床实际意义时，优效界值可设为大于 0。非劣效的目的是确证试验器械的有效性/安全性虽低于对照器械，但是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。无论优效还是非劣效试验，均应在试验设计阶段制定临床认可的界值并在方案中予以阐明。优效及非劣效性的具体假设参考《临床试验统计学》[23]。

主要结局指标确定

为评估医学影像 AI 器械的有效性和/或安全性，需要针对产品的目标人群和预期用途选择合适的研究终点（endpoints），且指标选择应该在试验开始前确定。

医学影像 AI 器械临床试验的评价指标多为诊断试验相关评价指标。首先明确器械测量结果的类型，定量结果主要用于测量图像中靶病灶的相关参数，如数量、长度、体积等；定性结果主要根据图像分析结果对疾病、诊断、预后进行分类，如是否患病、是否需要转诊、是否需要干预等。此外，还需要确定器械测量的结果是病例水平还是病灶水平。不同水平及不同结果性质各有适用的评价指标。

对于定性结果，在诊断试验有“金标准”时，主要评价指标优先考虑灵敏度和特异度。如一项用于特发性肺纤维化和慢性阻塞性肺病的深度学习诊断算法的临床试验（NCT05318599），使用了灵敏度、特异度作为主要评价指标[24]。同样地，一项基于人工智能的非酒精性脂肪性肝病的诊断临床试验（NCT04099147）也使用灵敏度、特异度作为主要评价指标[25]。其次可考虑受试者工作特征（receiver operating characteristic，

ROC)曲线下面积(area under curve, AUC),能够综合所有可能诊断界值下的灵敏度和特异度,从整体水平比较不同产品或不同阅片模式下的诊断效能。当诊断试验没有“金标准”时,可用 AI 器械诊断结果与“临床参考标准”的诊断结果之间的一致性来评估 AI 器械的性能,即评价两组诊断的阳性符合率、阴性符合率及总符合率[26]。除此以外,若医学影像 AI 器械用于检出病变情况(如结节)及其他目的时,检出率等也可作为主要评价指标[11]。

对于定量指标,其临床意义应由临床专家确定,如骨龄等。

在非诊断性试验的医学影像 AI 器械确证临床试验中,通常针对器械用途选取评价指标。如对于成像质量的评价可使用图像优良率,在成像设备辅助手术时,手术成功率等也可作为评价指标;除以上有效性评价指标之外,AI 器械安全性评价也极为重要,常用的安全性评价指标包括不良事件发生率、器械稳定性、器械缺陷等。

样本量估计方法

临床试验的样本量应结合研究假设,基于主要评价指标进行估计。ICH-E9[27]指出,临床试验的样本量必须足够大,以对研究假设所提出的相关问题提出可靠的回答,同时也不应过大而造成不必要的浪费。根据不同临床试验目的,可以确定合适的试验设计类型、对照形式、比较类型,同时选取恰当的主要评价指标及相应参数,以及有临床意义的界值,在假定的 I 类错误和 II 类错误下,采用相应的样本量估算公式,可估算出试验所需的样本量。

1. 常见临床研究设计的样本量估计

常见临床研究设计的样本量估计公式可参考医疗器械临床试验设计指导原则[8]。在医学影像 AI 器械的临床试验中,评价指标常为灵敏度和特异度,可基于灵敏度计算阳性组的样本量,基于特异度计算阴性组的样本量。若前瞻性收集受试者影像样本,在计算得到阳性组/阴性组样本量后,还需除以目标人群阳性率/阴性率方可得到每组的受试者数。

2. MRMC 研究设计样本量估计

MRMC 研究设计的主要评价指标常用灵敏度和特异度。MRMC 研究设计的临床试验中,样本量估计通常分为病例和阅片者数量两部分,影响样本量的参数可以概括为试验效应、变异和相关三部分。样本量计算的参数估计值需要事先假设并最好从预试验数据中获得,除试验设计类型和常规样本量计算参数以外,还需要明确的参数包括:①试验组与对照组诊断准确性的预期差值$\theta_\Delta(\theta_1-\theta_2)$;②比较组别与阅片者的二阶交互随机效应项所对应的方差分量$\sigma^2_{\tau R}$;③误差项对应的方差分量σ^2_ϵ;④同一阅片者在不同组别中给定判读结果的相关系数$r_1=\mathrm{Cov}_1/\sigma^2_\epsilon$;⑤相同组别下不同阅片者给定判读结果间的相关系数$r_2=\mathrm{Cov}_2/\sigma^2_\epsilon$;⑥不同阅片者在不同组别中给定判读结果的相关系数$r_3=\mathrm{Cov}_3/\sigma^2_\epsilon$;⑦入组病例中金标准判定的阳性和阴性病例的比例;⑧阅片者数量 r。其中 Cov_1、Cov_2 和 Cov_3 分别为同一阅片者不同组别、相同组别不同阅片者和不同阅片者不同组别情况下的诊断准确度协方差。

基于以上参数,假设所需病例数为 c,则在病例数 c 和阅片者数 r 的样本量组合条件下,检验出两种诊断方法灵敏度/特异度具有统计学差异的检验效能为[28]:

$$Power=Prb(F_{1,\widehat{df}_2;\Delta}>F_{1-\alpha;1,\widehat{df}_2})$$

此式中,Δ 和 df_2 分别为:

$$\Delta=\frac{\frac{r}{2}(\theta_1-\theta_0)^2}{\sigma^2_{\tau R}+\sigma^2_\epsilon-Cov_1+(r-1)(Cov_2-Cov_3)}$$

$$df_2=\frac{\{\sigma^2_{\tau R}+\sigma^2_\epsilon-Cov_1+[(r-1)(\widehat{Cov}_2-\widehat{Cov}_3)]\}^2}{\frac{\{\sigma^2_{\tau R}+\sigma^2_\epsilon-Cov_1-(\widehat{Cov}_2-\widehat{Cov}_3)\}^2}{(t-1)(r-1)}}$$

在假定c^*和r^*为与试验估计结果对应的入组病例和阅片者数量后,上述公式转换为:

$$\hat{\Delta}=\frac{\frac{r}{2}(\theta_1-\theta_0)^2}{\sigma^2_{\tau R}+\frac{c^*}{c}[\sigma^2_\epsilon-Cov_1+(r-1)(Cov_2-Cov_3)]}$$

$$\widehat{df}_2=\frac{\left\{\sigma^2_{\tau R}+\frac{c^*}{c}[\sigma^2_\epsilon-Cov_1+[(r-1)(\widehat{Cov}_2-\widehat{Cov}_3)]]\right\}^2}{\frac{\left\{\sigma^2_{\tau R}+\frac{c^*}{c}[\sigma^2_\epsilon-Cov_1-(\widehat{Cov}_2-\widehat{Cov}_3)]\right\}^2}{r-1}}$$

随机化策略

随机化是临床试验需要遵循的基本原则。随机化能够保证受试者有相同概率被分配到试验组或对照组,能够控制各种未知和已知因素的影响;随机化也是应用假设检验对研究资料进行比较分析的前提。随机化与盲法合用可以避免在分组时的可预测性而导致的可能偏倚。医学影像 AI 器械临床试验随机化策略通常包括影像样本的随机抽样、阅片者随机分组和阅片顺序的随机分配。

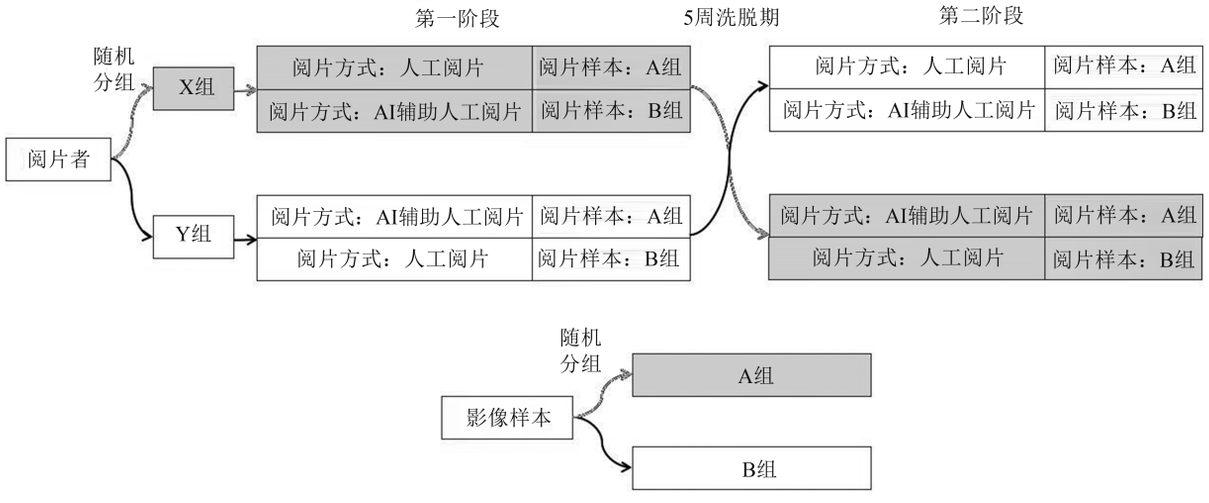
1. 常见的随机化方法

临床试验中要明确随机化方法,在试验开展前应明确随机的方式并制订随机表,以便临床试验实施[10]。常见的随机化分配方法可参考药物临床试验随机分配指导原则[29],在医学影像 AI 器械的临床试

验中，对阅片者进行随机分组及对阅片顺序进行随机一般采用简单随机化方法。

2. MRMC 设计中的随机化

在 MRMC 设计中，应当对阅片者、病例以及阅片模式进行随机化，以减少试验过程中的偏倚。为避免阅片顺序偏倚，常对阅片过程进行随机化，如对阅片医生的阅片顺序进行随机分配等。对于完全交叉的 MRMC 设计，随机化可分为三个步骤：首先将医学影像样本随机分配到不同阅片顺序的两组，同时将阅片者随机分为两组；其次对阅片者不同阅片模式的阅片顺序进行随机分配；最后在每个阶段每个阅片模式中，将相应的影像样本顺序随机分配给阅片者[30]。例如一项旨在比较在计算机辅助检测和诊断系统辅助下阅片者在解释乳腺病变超声图像方面表现的临床试验（NCT04551105），将影像样本随机分为两组（A 和 B），同时将阅片者随机分为两组（X 和 Y），X 组阅片者在第一阶段首先对 A 组进行人工阅片，然后对 B 组进行 AI 辅助人工阅片，在经历 5 周的洗脱期后，对 A 组进行 AI 辅助人工阅片，然后对 B 组进行人工阅片；而 Y 组则反之，在第一阶段首先对 B 组进行人工阅片，然后对 A 组进行 AI 辅助人工阅片。



*：A、B组的影像样本随机分配给X、Y组的阅片者

图 3　临床试验 NCT04551105 中的随机化图示

盲法实施

盲法是临床试验过程中控制研究者选择偏倚、信息偏倚等偏倚的重要措施，医疗器械临床试验盲法的实施需要根据具体产品特性、适应症以及伦理批准情况而定[31]。理想情况下，要求临床试验都尽可能做到完整设盲，即受试者、研究者和评价者都不知道分组信息。而医学影像 AI 器械临床试验中，有时难以实现受试者或研究者的设盲，此时应采取评估者盲法，例如盲态中心评估（blinded independent central review，BICR）及盲态独立评审委员会（blinded independent review committee，BIRC）评估等。在按照临床试验方案采集影像资料及必要的临床试验数据后，对标识信息进行遮盖隐藏后，提交给第三方中心进行盲态阅片。

双盲是指研究对象和研究者都不了解试验分组情况，而是由研究设计者来安排和控制全部试验，避免研究对象和研究者的主观因素带来的偏倚。双盲试验在临床新药研发中应用非常广泛，但是 AI 器械领域却由于双盲设计困难等诸多原因导致双盲随机临床试验较少。现有非盲法研究的一个主要局限性是引入了信息偏倚，即使用 AI 辅助检测系统的影像医师可能会处于因竞争精神而更加专注或因依赖 AI 系统而放松等非正常诊断时的状态。而 Wision 肠道癌前病变检测产品 EndoScreener 的临床试验[32]作为全球首个关于 AI 的双盲随机对照试验，参照新药研发中双盲试验的安慰剂对照组，设计了“伪装 AI 系统”及“引入第二观察者”以便对操作医师设盲，所有符合条件的受试者被随机分配到 CADe 系统辅助的试验组和伪装 AI 系统的对照组，内镜医生被告知在 CADe 系统的帮助下进行所有结肠镜检查程序，而并不知道使用了伪装 AI 系统。在这项研究中，实现了对研究对象和研究者的分组信息隐藏，减少了可能存在的偏倚，为全球其他 AI 辅助诊断领域的临床验证方法提供了双盲设计的参考。

缺失值处理及敏感性分析

医学影像 AI 器械临床试验，多因回顾性影像收集、失访、影像质量不佳、技术操作失误等原因导致缺失数据的产生，从而导致研究结果偏倚。临床试验中，应尽可能预防数据缺失，特别是主要评价指标的缺失。

针对主要评价指标的缺失,应事先在研究方案和/或统计分析计划书中对缺失值的处理方法予以明确[19]。缺失值处理常用的方法有 4 种,包括完整观测法(complete case, CC)、填补法(imputation, IM)、基于极大似然(maximum likelihood, ML)和基于非随机缺失机制(missing mot at random, MNAR)。对于医学影像 AI 器械临床试验缺失值处理一般常采用完整观测法,此方法适用于完全随机缺失(missing completely at random, MCAR)机制。参考 ICH-E9[27] 和 ICH E9 (R1)[33],对于主要评价指标缺失值,将采用针对性的方法处理缺失数据并进行敏感性分析,以评估缺失数据对研究结果的影响,保证结果的稳健性和可靠性。此外,若存在多个不同临床试验中心或不同人群来源的影像数据,可以考虑进行分层分析或亚组分析。

结 语

良好设计的医疗器械临床试验,不仅是法规的要求,也是临床试验是否能顺利实施并提供可靠研究结果的关键。医学影像 AI 器械应用越来越广泛,在进行此类临床试验设计时应考虑其特殊性,而不能与常规医疗器械临床试验设计方法一概而论。本共识聚焦医学影像 AI 器械临床试验的统计学设计要点,涵盖了该类临床试验设计类型和对照选择、试验比较类型、主要结局指标、样本量估计、随机化和盲法的确定方法和依据,同时强调了在此类临床试验设计中应事先考虑缺失值处理及敏感性分析,以期获得客观公正的研究结果,为基于医学影像开展 AI 器械临床试验的统计学设计提供参考。

参 考 文 献

[1] 伍康振, 温福林. 医学影像技术在医学影像诊断中的应用分析. 中国医疗器械信息, 2020, 26(18): 96-97.

[2] Alexander RG, Yazdanie F, Waite S, et al. Visual illusions in radiology: untrue perceptions in medical images and their implications for diagnostic accuracy. Frontiers in Neuroscience, 2021, 15: 629469.

[3] Gao R, Zhao S, Aishanjiang K, et al. Deep learning for differential diagnosis of malignant hepatic tumors based on multi-phase contrast-enhanced CT and clinical data. Journal of Hematology & Oncology, 2021, 14(1): 154.

[4] Wood DA, Kafiabadi S, Busaidi AA, et al. Deep learning models for triaging hospital head MRI examinations. Medical Image Analysis, 2022, 78: 102391.

[5] Mu W, Jiang L, Shi Y, et al. Non-invasive measurement of PD-l1 status and prediction of immunotherapy response using deep learning of PET/CT images. Journal for Immunotherapy of Cancer, 2021, 9(6): e002118.

[6] Zhang Q, Zhang S, Pan Y, et al. Deep learning to diagnose hashimoto's thyroiditis from sonographic images. Nature Communications, 2022, 13: 3759.

[7] Hunter B, Chen M, Ratnakumar P, et al. A radiomics-based decision support tool improves lung cancer diagnosis in combination with the herder score in large lung nodules. EBioMedicine, 2022, 86: 104344.

[8] 食品药品监管总局. 食品药品监管总局关于发布医疗器械临床试验设计指导原则的通告(2018 年第 6 号)(2018-01-04)[2022-10-12]. https://www.nmpa.gov.cn/ylqx/ylqxggtg/ylqxzhdyz/20180108183301635.html?type=pc&m=.

[9] 国家药品监督管理局医疗器械技术审评中心. 国家药监局器审中心关于发布人工智能辅助检测医疗器械(软件)临床评价注册审查指导原则的通告(2023 年第 38 号)(2023)[2023-11-28]. https://www.cmde.org.cn/xwdt/shpgzgg/gztg/20231107153309174.html.

[10] 吴建元, 蔡君龙, 陈博, 等. 医疗器械临床试验方案设计要点探讨. 医疗卫生装备, 2020, 41(2): 65-69+86.

[11] US Food and Drug Administration. Clinical performance assessment: considerations for computer-assisted detection devices applied to radiology images and radiology device data in premarket notification (510(k)) submissions (2022) [2022-10-11]. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-performance-assessment-considerations-computer-assisted-detection-devices-applied-radiology.

[12] 国家药品监督管理局医疗器械技术审评中心. 关于发布深度学习辅助决策医疗器械软件审评要点的通告(2019 年第 7 号)(2019-07-03)[2022-10-11]. https://www.cmde.org.cn//xwdt/shpgzgg/gztg/20190703141714991.html.

[13] Pallmann P, Bedding AW, Choodari-Oskooei B, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Medicine, 2018, 16(1): 29.

[14] Huang Z, Samuelson F, Tcheuko L, et al. Adaptive designs in multi-reader multi-case clinical trials of imaging devices. Statistical Methods in Medical Research, 2020, 29(6): 1592-1611.

[15] Park SH, Han K, Jang HY, et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. Radiology, 2023, 306(1): 20-31.

[16] Obuchowski NA, Bullen J. Multireader diagnostic accuracy imaging studies: fundamentals of design and analysis. Radiology, 2022, 303(1): 26-34.

[17] Wellek S, Blettner M. On the proper use of the crossover design in clinical trials. Deutsches Ärzteblatt International, 2012, 109(15): 276-281.

[18] 尚美霞, 阎小妍, 姚晨, 等. 深度学习辅助医疗产品开展确证性临床试验的多阅片者设计和方法学考虑. 中国食品药品监管, 2021(07): 100-105.

[19] 李卫, 赵耐青. 单组目标值临床试验的统计学考虑. 中国卫生统计, 2017, 34(3): 505-508.

[20] 国家药品监督管理局药品审评中心. 国家药监局药审中心关于发布《药物临床试验适应性设计指导原则(试行)》的通告(2021 年第 6 号)(2021)[2023-11-10]. https://www.cde.org.cn/main/news/viewInfoCommon/bc2b326bd49bac7437368272be6ec00d.

[21] US Food and Drug Administration. Adaptive designs for medical device clinical studies(2016). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-designs-medical-device-clinical-studies.

[22] Zapf A, Stark M, Gerke O, et al. Adaptive trial designs in diagnostic accuracy research. Statistics in Medicine, 2020, 39(5): 591-

601.

[23] 陈峰,夏结来主编. 临床试验统计学. 北京:人民卫生出版社, 2018.

[24] Pediatric Clinical Research Platform. Deep learning diagnostic and risk-stratification for idiopathic pulmonary fibrosis and chronic obstructive pulmonary disease in digital lung auscultations: NCT05318599. clinicaltrials. gov, 2022. https://clinicaltrials. gov/ct2/show/NCT05318599.

[25] Instituto de Investigación Marqués de Valdecilla. Diagnosis and characterization of non-alcoholic fatty liver disease based on artificial intelligence.: NCT04099147. clinicaltrials.gov, 2019. https://clinicaltrials.gov/ct2/show/NCT04099147.

[26] Theel ES, Hilgart H, Breen-Lyles M, et al. Comparison of the quantiferon-tb gold plus and quantiferon-tb gold in-tube interferon gamma release assays in patients at risk for tuberculosis and in health care workers. Journal of Clinical Microbiology, 2018, 56(7): e00614-18.

[27] US Food and Drug Administration. E9 statistical principles for clinical trials. FDA, 1998. https://www. fda. gov/regulatory-information/search-fda-guidance-documents/e9-statistical-principles-clinical-trials.

[28] Hillis SL, Obuchowski NA, Berbaum KS. Power estimation for multireader roc methods an updated and unified approach. Academic Radiology, 2011, 18(2): 129-142.

[29] 国家药品监督管理局医疗器械技术审评中心. 国家药监局药审中心关于发布《药物临床试验随机分配指导原则(试行)》的通告(2022 年第 5 号), 2022. https://www. cde. org. cn/main/news/viewInfoCommon/402c511b46bfa8c472fd6aad6e164557.

[30] Ba W, Wu H, Chen WW, et al. Convolutional neural network assistance significantly improves dermatologists' diagnosis of cutaneous tumours using clinical images. European Journal of Cancer, 2022, 169: 156-165.

[31] 吴建元, 黄志民, 蔡君龙, 等. 激光类医疗器械临床试验方案设计探讨. 中国医疗器械杂志, 2020, 44(2): 158-162.

[32] Wang P, Liu X, Berzin TM, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (cade-db trial): a double-blind randomised study. The Lancet: Gastroenterology & Hepatology, 2020, 5(4): 343-351.

[33] 国家药品监督管理局药品审评中心. 关于公开征求 ICH《E9(R1): 临床试验中的估计目标与敏感性分析》中文翻译稿意见的通知(2017). https://www. cde. org. cn/main/news/viewInfoCommon/b62a86e3b88b7ec7f3114798a0b5bc4f.

(责任编辑:邓　妍)

(上接第 473 页)

[21] Wang X, Suttner L, Jemielita T, et al. Propensity score-integrated Bayesian prior approaches for augmented control designs: a simulation study. J Biopharm Stat, 2021: 1-21.

[22] Yue LQ, Campbell G, Lu N, et al. Utilizing national and international registries to enhance pre-market medical device regulatory evaluation. J Biopharm Stat, 2016, 26(6): 1136-1145.

[23] Yue LQ, Lu N, Xu Y. Designing premarket observational comparative studies using existing data as controls: challenges and opportunities. J Biopharm Stat, 2014, 24(5): 994-1010.

[24] Li H, Yue LQ. Propensity score-based methods for causal inference and external data leveraging in regulatory settings: From basic ideas to implementation. Pharm Stat, 2023.

[25] Lin J, Gamalo-siebers M, Tiwari R. Propensity-score-based priors for Bayesian augmented control design. Pharm Stat, 2019, 18(2): 223-238.

[26] CSCO 生物统计学专家委员会 RWS 方法学组. 倾向性评分方法及其规范化应用的统计学共识. 中国卫生统计, 2020, 37(6): 952-958.

[27] Harton J, Segal B, Mamtani R, et al. Combining Real-World and Randomized Control Trial Data Using Data-Adaptive Weighting via the On-Trial Score. Statistics in biopharmaceutical research, 2022: 1-13.

[28] Wang C, Li H, Chen WC, et al. Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. J Biopharm Stat, 2019, 29(5): 731-748.

[29] Baron E, Zhu J, Tang RS, et al. Bayesian Divide-and-Conquer Propensity Score Based Approaches for Leveraging Real World Data in Single Arm Clinical Trials. J Biopharm Stat, 2022: 1-15.

[30] Liu MZ, Bunn V, Hupf B, et al. Propensity-score-based meta-analytic predictive prior for incorporating real-world and historical data. Statistics in medicine, 2021, 40(22): 4794-4808.

[31] Zhu AY, Roy D, Zhu Z, et al. Propensity score stratified MAP prior and posterior inference for incorporating information across multiple potentially heterogeneous data sources. J Biopharm Stat, 2023: 1-15.

[32] Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. BMJ(Clinical research ed), 2019, 367: l5657.

[33] Sachdeva A, Tiwari RC, Guha S. A novel approach to augment single-arm clinical studies with real-world data. J Biopharm Stat, 2022, 32(1): 141-157.

[34] Su L, Chen X, Zhang J, et al. Comparative Study of Bayesian Information Borrowing Methods in Oncology Clinical Trials. JCO precision oncology, 2022, 6: e2100394.

[35] Yu G, Bian Y, Gamalo M. Power priors with entropy balancing weights in data augmentation of partially controlled randomized trials. J Biopharm Stat, 2022, 32(1): 4-20.

[36] Jiang L, Nie L, Yuan Y. Elastic priors to dynamically borrow information from historical data in clinical trials. Biometrics, 2021.

[37] FDA. Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products. (202302) [2023.04.01]. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-design-and-conduct-externally-controlled-trials-drug-and-biological-products.

[38] Shan M, Faries D, Dang A, et al. A Simulation-Based Evaluation of Statistical Methods for Hybrid Real-World Control Arms in Clinical Trials. Statistics in biosciences, 2022, 14(2): 259-284.

(责任编辑:邓　妍)