人工智能辅助检测医疗器械临床评价

注册审查指导原则

（征求意见稿）

本指导原则旨在指导注册申请人开展人工智能（artificial intelligence，AI）辅助检测类医疗器械临床评价的资料准备，同时为技术审评部门审评人工智能辅助检测类产品临床评价资料提供参考。

本指导原则进一步明确人工智能辅助检测类医疗器械临床评价的要求和适用情形。申请人可依据产品的具体特征确定其中内容是否适用，若不适用，需阐述理由并提供相应的科学依据，并依据产品的具体特征对临床评价资料的内容进行充实和细化。

本指导原则是供注册申请人和技术审评人员使用的指导性文件，但不包括审评审批所涉及的行政事项，亦不作为法规强制执行，应在遵循相关法规的前提下使用本指导原则。如果有能够满足相关法规要求的其他方法，也可以采用，但是需要提供详细的合理性论述和验证确认资料。

本指导原则是在现行法规和标准体系以及当前认知水平下制定，随着法规和标准的不断完善，以及科学技术的不断发展，相关内容也将适时进行调整。

一、适用范围

人工智能医疗器械从用途角度可分为辅助决策类和非辅助决策类。其中，辅助决策是指通过提供诊疗活动建议辅助医务人员进行临床决策，如通过异常识别、病灶性质判定、自动制定手术计划进行辅助分诊、辅助检测、辅助诊断、辅助治疗。

人工智能辅助检测产品，是指基于计算机人工智能算法，可包含模式识别和数据分析等功能，通过识别、标记、突出等方式提示医师关注可能的异常/病变区域，从而辅助临床医师做出相应诊疗决策的产品，分类编码为21-04-02，管理类别为第III类；产品还可同时包含非辅助决策功能，如结构化报告生成、前后片对比、正常组织的分割（如肺叶、肋骨等）、尺寸测量、CT值计算等临床功能和数据储存、传输等非临床功能等。人工智能辅助检测产品常见的有针对肺结节、乳腺结节、骨折、血管狭窄、结肠息肉等病变/异常的检出产品，本指导原则给出此类产品的通用要求，并以肺结节辅助检测和结肠息肉辅助检测产品为例（详见附录），阐述对人工智能辅助检测产品的临床试验中具体要素的考虑建议，同类型产品可参考相关适用部分。

本指导原则不适用于如下情形：1.可鉴别病变的性质（如良恶性）或疾病分期分型的人工智能辅助诊断类产品，2.预测疾病发生概率的产品，3.可同时辅助检测、鉴别诊断多种病变的多分类人工智能辅助检测产品（例如同时辅助检测并分类肺结节、条索、胸膜增厚、胸腔积液、肋骨骨折的产品）；4.人工智能辅助分诊、转诊产品，此类产品对患者是否患有目标疾病，从而对患者的分诊转诊提供辅助决策建议的产品，该类产品不给出具体病变情况，且无论辅助分诊结果为阴性、阳性，均需专业医师再一次对患者影像进行评阅，常见的有糖尿病视网膜病变辅助分诊、肺炎辅助分诊、脑出血辅助分诊等；5. 配合体外诊断试剂产品使用的人工智能辅助分析软件。上述产品可参照本指导原则中适用部分的要求。

二、临床试验设计

（一）临床试验目的

临床试验目的一般是评价申报产品在预期适用范围下使用时的辅助检测诊断学性能，亦可一并观察产品的可用性与安全性。

（二）临床试验基本设计类型

人工智能辅助检测类产品的临床意义在于提升医师的病变检测准确度，为充分评估产品的临床受益风险可接受性，此类产品一般需考虑开展对照试验，根据产品特征及临床诊疗实际，可以为平行对照、交叉自身对照或多阅片者多数据样本（multiple reader multiple case，MRMC）试验设计。

试验组一般为医师在软件的辅助下完成异常/病变的检测，对照组一般为临床医师独立的异常/病变的检测，比较二者的诊断准确率。

（三）研究对象

1．预期人群的影像学样本

预期人群的影像学样本是人工智能辅助检测产品临床试验的典型研究对象，影像学样本可以基于定义明确的入选和排除标准收集，可为临床已有数据（如临床诊疗中产生的真实世界数据）。考虑到AI与医师观察、操作的协同交互等因素，基于实时影像的辅助检测临床试验宜基于阅片者前瞻性检查影像样本作为试验研究对象，而非采用过往既有的检查录像。

为了保证临床试验质量以及结果的可靠性，研究对象选取时，申请人需考虑如下措施：一是纳入数据样本独立于申报产品或前代产品开发所用数据集，如申报产品或前代产品的训练集、测试集。二是采用临床已有数据进行研究时，需基于明确且严格的入排标准和计划，收集过往某段时间内、特定医疗机构内患者影像学数据，避免主观挑选病例。三是考虑阳性数据样本中，目标疾病的疾病谱分布（如分型、分期）合理性，某些对辅助检测具有挑战性的分期、分型，必要时在临床试验中富集相关具有代表性的亚组。四是需避免在一项临床试验中同时入组同一患者的多组样本数据。五是临床已有数据收集时，需尽可能全面的收集与疾病相关的信息（适用的），具体包括但不限于：

（1）人口统计学信息(如年龄、性别);

（2）与辅助检测目标疾病相关的信息，如病史、疾病状态、分期、分型、病变大小、病变位置、器官特征(如乳腺成分)、伴随疾病等。

（3）确定为阳性/阴性病例的依据，如既往诊断结论，以及确定疾病状态、部位和程度的方法。

2.阅片者

由于阅片者表现的变异度及其与患者样本变异度和诊断方法（即AI辅助器械）之间的交互效应，一般情况下宜将阅片者列入研究对象。基于非实时影像的辅助检测产品（如肺结节/骨折/乳腺结节辅助检测等），采用MRMC设计可较好的控制阅片者偏倚，所需的样本量一般较少，申请人可优先考虑选择。采用MRMC设计时，根据预期的使用者情况，选取不同水平的多位医师作为阅片者，申请人需论述阅片者数量的合理性。

（四）评价指标

主要评价指标应结合产品设计特征和临床试验可行性进行综合选择，一般认为固有诊断准确性指标受样本患病率差异的影响较小，因此，在试验可行的前提下宜优先考虑此类指标作为主要评价指标。固有准确度指标包括敏感度、特异度、ROC或其衍生曲线等。

无论选择哪些指标作为主要评价指标，该类产品临床试验应当考虑整体的优效性设计，例如ROC或其衍生曲线下面积（Area Under Curve，AUC）的优效设计，或者目标疾病辅助检测的灵敏度优效和特异度非劣效，或者息肉/腺瘤初检检出率的优效性等。

（五）金标准构建

申请人应详述金标准的选择、构建方法及理由。可供选择的金标准构建方法包括：一是以临床已确认结果为金标准，即临床上结合患者影像学检查、病史、实验室检查（如活检）、长期随访结果等方法综合判定的临床诊断结果，基于前瞻数据或临床已有数据皆可；二是通过专家组对患者影像的阅片判定作为金标准。

对于人工智能辅助检测产品，通常可选择高年资医师组成的阅片专家组综合意见为金标准，并需要明确：1.专家数量；2.专家经验及专业水平；3.决策机制（如少数服从多数、背靠背第三人仲裁等）；4.专家决策时所依据的信息（如图像上是否有标记，是否还提供了病史或其他检查结果等）；5.判定所依据的临床准则（如临床指南、诊疗规范等）。

对于试验中对病灶的检出是否与金标准专家组意见一致，一般可考虑1.试验组/对照勾画病灶的中心在专家组勾画的病灶轮廓边界范围内；2.试验组/对照组勾画病灶与专家组勾画病灶的像素重合度高于一定比例；若采用其他判定方法，则需论述合理性。

若采用临床已确认结果作为金标准，则需明确1.已确认结果所依据的临床信息，包括检查类型及结果；2.各类检查的设备信息，包括影像检查的扫描条件等；3.已确认结果所依据的临床准则；4.得出已确认结果的医师情况，包括专家会诊，需明确医师资质；5.若还依据了随访数据，则还需明确随访的时间以及随访所做的检查类型及结果。

金标准独立于待研究的诊断方法可以减少偏倚，一是提供给专家组的待评阅数据需不包含AI标记结果；二是对于样本量非常大或影像评阅耗时长的研究，也可将包含和不包含AI辅助标记的结果同时提供给专家组，但两组结果是否采用AI辅助需对专家组设盲，由专家组自行开展综合评阅。

（六）样本量估算

样本量估算需综合考虑临床试验设计、主要评价指标和统计学要求。申请人需明确计算公式、相应参数及确定理由，以及所用的统计软件。

鉴于采用高于临床总体的样本患病率可以降低研究所需的样本量，节约时间和试验花费，且临床上对于一定样本量下的研究中，患病率是否影响灵敏度、特异度等固有诊断准确性指标尚无统一定论，且多数研究倾向于影响较小，增大受试人群阳性比例可能是合理的（例如可考虑采用阳性样本与阴性样本1:1的比例设定）。临床试验资料中可以提供样本患病率以及目标疾病的流行病学研究的患病率情况进行合理性论述，并确保临床研究设计中样本数据随机分配给阅片者进行评阅。

平行对照试验样本量计算可参考《医疗器械临床试验设计指导原则》中的相关内容。

若采用MRMC的试验设计，样本量计算需首先明确具体的分析方法，如Obuchowski-Rockette Analysis（OR分析方法）、Dorfman-Berbaum-Metz-Hillis Analysis（DBMH分析法），并进一步明确受试医师数量，显著性α、把握度1-β、预计效应值，优效/非劣效界值，其中预计效应值可通过预试验或调研并汇总分析目标病灶检出的诊断学研究文献获得。以DBMH分析法为例，样本量估算与交互随机效应值，检验的把握度、显著性，受试医师数量，优效界值，非劣效界值（如有）等要素有关。用于样本量估算的效应值（effect size）选择时，可通过预试验得到的误差和混合效应方差来估计，并考虑预试验样本量等情况，在预试验结果的基础上采用适度保守的估计。

所有应用了试验产品的受试医师和患者都将被纳入分析，对于主要指标，除给出主要评价指标（灵敏度、特异度、AUC）的点估计外，还将分别对其对应的95%置信区间进行估计，通过试验组与对照组的优效/非劣效比较判断本试验产品是否满足临床应用的需要。

（七）其他临床试验设计需关注的问题

1.临床试验培训

在试验前对阅片者开展必要的培训，可以有效降低试验的偏倚。除基本情况培训（试验流程、术语定义、数据样本阅片的评价标准等）以外，还需考虑案例培训和典型数据样本讲解等，且所用案例独立于试验研究数据样本；培训中宜设置阅片者培训结果测试及合格接受标准；建议临床试验中，对阅片者的培训与临床应用时的培训，在方法、时间、接受标准上尽量保持一致性；并考虑对临床试验阅片者资质、能力等要求与临床应用时使用者情况的匹配性。对专家组的培训时间和接受标准宜显著高于试验组。

2.影像样本评阅质量控制

一是由阅片者在独立盲法条件下对患者关于试验的影像检测结果进行解释。二是在选取试验阅片者时基于其专业能力和参与研究的可能性，充分保证阅片者对预期使用者的代表性。可考虑不同机构来源、不同年资/专业水平的医师。三是试验组与对照组使用相同软件界面，对照组医师软件界面无法开启人工智能辅助决策功能。四是将数据样本的临床诊断结果、金标准判定结果、其他诊断信息（如生化检测结果、后续治疗等）或基本信息（如年龄、病史等）对试验的阅片者隐藏。五是对照试验可考虑采用交叉阅片设计，交叉设计中可根据相关领域记忆曲线的研究设置合理的洗脱期。若采用多阅片者，最好每位阅片者按照不同的数据样本顺序进行评阅，有文献报道典型的洗脱期一般为4~6周。是否采用交叉设计可根据申报产品的临床应用方法、场景及适用范围等综合判定。

三、产品中非辅助决策临床功能的评价

深度学习辅助决策医疗器械产品还可能包括结构化报告生成、前后片对比、正常组织的分割（如肺叶、肋骨等）、尺寸测量、CT值计算等非辅助决策类软件功能，这些功能可在临床试验中设置次要评价指标一并评价其安全有效，亦可根据《医疗器械临床评价技术指导原则》提交相应临床评价资料。若提交临床评价，上述功能的验证确认可以考虑测试集测试、压力测试、对抗测试、质量良好的数据库测试结果中的一种或几种作为临床评价的支持证据；若在临床试验中设置次要指标，指标一般采用临床上对功能准确性评价的金标准或学术上常用的方法，如分割的准确性考虑，与医师的分割结果的分割一致性Dice相似系数（Dice similarity coefficient）等；配准功能可考虑标志点配准误差(Fiducial Registration Error)、目标配准误差（target registration error）、标志点定位误差(Fiducial Localization Error)等。

四、与临床评价相关的说明书内容

说明书中一般需结合临床与非临床资料给出下列适用的信息：1）临床试验总结；2）适用范围；3）数据采集设备和数据采集过程相关要求等。临床评价总结通常考虑临床数据基本信息、评价指标与结果（必要时含亚组结果）。人工智能辅助检测类产品的适用范围需明确辅助检测适应症（如肺结节、骨折），所基于的影像类型（如胸部CT或结肠镜检查影像），产品其他主要功能（如影像的显示、处理、测量和分析），以及产品临床定位（不能单独用作临床诊疗决策依据）等。

建议申请人根据产品实际情况在说明中体现如下方面的警告注意事项。1)软件仅辅助医师进行检测，存在假阴/阳性可能，应由专业医师结合患者的病史、症状、体征、其他检查结果情况综合给出最终的病灶检出结论，核实是否需要进行一步诊疗的决策，并对临床诊断结果负责。2) 产品依据XXX版指南（例如《胸部CT肺结节数据标注与质量控制专家共识（2018）》）设计，如诊疗指南有所更新，使用者应充分评估指南差异可能带来的风险。3）明确产品临床试验中未对病灶边界分割的准确性进行评价，如依据软件的检测结果决定手术干预或穿刺活检，医师应当充分评估其风险。

# 附录1 肺结节人工智能辅助检测产品

临床评价考虑要素举例

# 一、试验设计基本类型

肺结节CT影像辅助检测产品采用MRMC设计、交叉自身对照设计或平行对照设计。平行对照设计的具体过程及方法参考《医疗器械临床试验设计指导原则》相关内容。

若采用MRMC的设计，试验过程具体为，将参与试验的临床医师随机划分为A、B两组。A组：医师第一阶段在使用AI辅助的条件下完成所有数据样本CT影像的肺结节检测；经一定的洗脱期后，第二阶段在不使用AI辅助的条件下完成所有数据样本CT影像的肺结节检测。B组：医师第一阶段在不使用AI辅助的条件下完成所有数据样本CT影像的肺结节检测；经一定的洗脱期后，第二阶段在使用AI辅助的条件下完成所有数据样本CT影像的肺结节检测。该试验中医师使用AI辅助的作为试验组，不使用AI辅助的作为对照组，对比试验组与对照组肺结节检测的诊断准确度指标，阅片过程中每名医师需分别在使用AI和不使用AI的条件下都完成所有入组影像数据样本的阅片。

# 二、评价指标

此类产品临床试验通常选择检出的灵敏度（以病灶为单位）和特异度(以患者为单位)等组成的复合指标，亦可在此基础上添加受试者工作特征（receiver operating Characteristic, ROC）曲线或其衍生曲线组成复合指标，如因变量自由的受试者工作特征（free-response receiver operating Characteristic, FROC）曲线、定位受试者工作特征（location-specific receiver operating characteristic, LROC）曲线等。使用ROC曲线或其衍生曲线时，可考虑在试验组医师阅片阶段，由阅片医师标记对目标疾病判定的把握度（可考虑百分数评分），从而构建ROC曲线，并需考虑在培训阶段明确医师百分数评分的标准和方法。

次要指标可包括AI辅助医师检出目标疾病的阳性预测值/阴性预测值，似然比，Kappa系数，软件独立检出目标疾病的灵敏度（以病灶为单位），软件独立检出目标疾病的灵敏度和特异度（以患者为单位），阅片时间，医生诊断信心评价，软件功能易用性和稳定性（可采用主观感受评价，如李克特量表等）。若同时观察量化分析等非辅助决策临床功能的性能表现，亦可设立相应次要评价指标。

# 三、样本量估算及统计学分析

以MRMC设计为例，说明样本量估算考虑要素，若主要评价指标为AUC、结节水平灵敏度、患者水平特异度。试验中，检验的把握度取80%。对于主要评价指标，优效性与非劣效统计分析将在单侧0.025的显著性水平下进行。

1.基于AUC的样本量计算，预试验或文献资料确认效应值取0.05，受试医师设定为12名，AUC优效界值取0，则总样本量不得低于87例，阴性、阳性患者比为1:1。

2.基于结节水平灵敏度，效应这去0.069，受试医师设定为12名，优效界值取0，需阳性样本至少116个结节，假定平均每个阳性患者存在2个结节，则至少需要58例阳性数据样本。

3.基于患者水平的特异度，效应值取-0.055，非劣效值取0.1（参考Riverain ClearRead CT 的特异度95%区间下限），在当受试医师为12名时，非劣效验证（非劣效界值取0.1）需要阴性样本达到至少157例。

综合以上各指标样本量估算值，本试验样本量最低为215例，阳性数据样本58例，阴性数据样本157例。考虑到样本存在剔除和脱落的可能，当预期脱落率为5%，剔除率为5%时，阳性样本需要达到至少62例，阴性样本需要达到至少166例，共计228例。

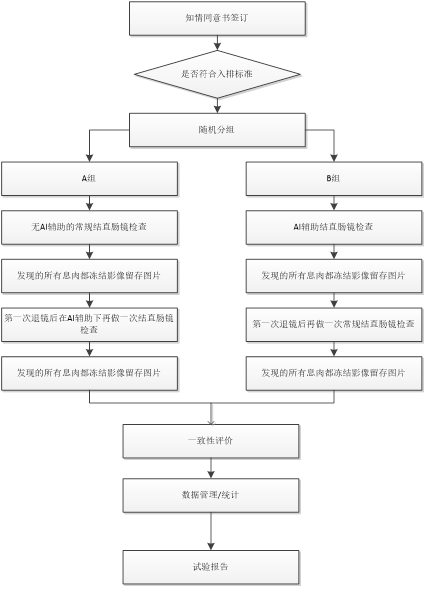
# 附录2. 结肠息肉人工智能辅助检测产品

临床评价考虑要素举例

# 一、试验设计基本类型

基于结肠镜等实时影像结果的辅助检测可考虑平行对照设计，具体设计及试验程序（以序贯的平行对照为例）如下：

入组的受试者随机分配到A组和B组， A组：常规结直肠镜检查先进行组：先接受无 AI 辅助的常规结直肠镜检查，对检查过程中发现的所有息肉都冻结影像留存图片。第一次退镜后在AI 辅助下再做一次结直肠镜检查，对检查过程中发现的所有息肉都冻结影像留存图片。B 组：AI 辅助检查先进行组：先接受 AI 辅助的结直肠镜检查，对检查过程中发现的所有息肉都冻结影像留存图片。第一次退镜后再进行一次常规结直肠镜检查，对检查过程中发现的所有息肉都冻结影像留存图片。分别以各组两次检出的肠息肉作为基准，比较第一次检查时，有AI辅助检测和无AI辅助检测的病变检出能力。



# 二、评价指标

此类产品临床试验因不完美金标准等问题，可考虑选取息肉/腺瘤的初检检出率/漏诊率或息肉检出的敏感性与特异性等作为主要评价指标。相关指标需以息肉/腺瘤水平而非受试者水平进行计算，例如初检息肉检出率=初检发现的息肉数量÷两次检测到的息肉总数，腺瘤漏诊率=第二次检测到的腺瘤数÷两次检测到的腺瘤总数。

次要评价指标可考虑息肉/腺瘤检出率（受试者水平），平均息肉/腺瘤检出数，软件独立评估性能（即软件独立的息肉/腺瘤检出率、息肉检测准确性等）：结直肠镜操作时间（包括各次进镜时间、退镜时间）、软件易用性评价、稳定性评价、不良事件情况等。若同时观察图像传输、储存等非辅助决策临床功能的性能表现，亦可设立相应次要评价指标。

# 三、样本量估算

1.以序贯的平行对照试验为例，若采用腺瘤漏诊率作为评价指标，样本量的计算基于腺瘤漏检率的主要结果（每个病灶），假设如下： AI 先进行组别的腺瘤漏诊率为 10%，标准检查先进行组别的腺瘤漏诊率为 30%，优效界值设定为0，在 80％统计功效的条件下， 双侧α值设为 0.05。主要评价指标为初检息肉检出率，其样本量估算公式为：

假设每位患者的平均腺瘤数为 0.6，考虑10%的脱落率，需入组220名受试者，AI 先进行组110名，标准检查先进行组110名。

2.若采用初检息肉检出率作为主要评价指标，A 组： 常规结直肠镜检查先进行组； B 组： AI 辅助检查先进行组。PC、 PT 分别为 A 组和 B 组预期初检息肉检出率，其中 PC=70%、 PT=80%；|D|为两组预期率差的绝对值， |D| = |PC - PT|； Δ 为优效性界值，取 0。主要评价指标为初检息肉检出率，其样本量估算公式为：

假设每位患者的平均息肉数为 2.5， 需约 234 例，脱落率为 20%，最终样本量为 296 例， A组和 B 组各 148 例。

# 五、参考文献

[1] 中华医学会呼吸病学分会肺癌学组， 中国肺癌防治联盟专家组，肺结节诊治中国专家共识（2018版）[J]. 中华结核和呼吸杂志， 2018,41(10): 763-771.

[2] 国家药品监督管理局医疗器械技术审评中心. 总局关于发布医疗器械临床试验设计指导原则的通告[EB/OL].(2018-01-04) [2021-05-01]. https://www.cmde.org.cn/CL0058/6882.html

[3]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[EB/OL]. (2019-07-03)[2021-02-04]. https://www.cmde.org.cn/CL0050/19360.html

[4] Guidance for Industry and Food and Drug Administration Staff Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions（FDA，2012.07.03）

[5] Lung Cancer Screening CT ProtocolsVersion5.1（AAPM，2019.9.13）

[6] FDA. Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Approval (PMA) and Premarket Notification [510(k)] Submissions - Guidance for Industry and FDA Staff，(2012.11)[2021-4-15]. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-performance-assessment-considerations-computer-assisted-detection-devices-applied-radiology

[6] Sica, G.T., 2006. Bias in research studies. Radiology, 238(3), pp.780-789.

[7] Xiao-Hua Zhou, Nacy A. Obuchowski, Donna K. McClish. Statistical Methods in Diagnostic Medicine [M]. Hoboken: JohnWiley&Sons, 2011.

[8] 国家卫生健康委员会. 《人工智能辅助诊断技术临床应用质量控制指标（2017年版）》.国家卫生计生委办公厅关于印发造血干细胞移植技术管理规范（2017年版）等15个“限制临床应用”医疗技术管理规范和质量控制指标的通知[EB/OL].（2017-02-17）[2021-4-27] http://www.nhc.gov.cn/yzygj/s3585/201702/e1b8e0c9b7c841d49c1895ecd475d957.shtml

[9] Brandon D. Gallas.et al. Evaluating Imaging and Computer-aided Detection and Diagnosis Devices at the FDA [J]. Acad Radiol. 2012 Apr; 19(4): 463–477. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5557046/

[10] 萧毅, 刘士远. 肺结节影像人工智能技术现状与思考[J]. 肿瘤影像学, 2018, 27: 249-252.

[11] Rees, Colin J., et al. “UK key performance indicators and quality assurance standards for colonoscopy.” Gut 65.12 (2016): 1923-1929.

[12] Misawa, Masashi, et al. "Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video)." Gastrointestinal endoscopy 93.4 (2021): 960-967.

[13] 国家消化系统疾病临床医学研究中心（上海）， 中华医学会消化内镜学分会， 中华医学会健康管理学分会等. 中国早期结直肠癌筛查流程专家共识意见（2019）. 中华医学会杂志，2019,99（38）：2961-2970.

[14] Rompianesi, Gianluca, et al. "Artificial intelligence in the diagnosis and management of colorectal cancer liver metastases." World Journal of Gastroenterology 28.1 (2022): 108.