附件3

病理图像人工智能分析软件临床评价审评要点

本审评要点旨在指导注册申请人对病理图像人工智能分析软件临床评价注册申报资料的准备及撰写，同时也为技术审评部门提供参考。

本审评要点是对病理图像人工智能分析软件的一般要求，申请人应依据产品的具体特性确定其中内容是否适用。若不适用，需具体阐述理由并提供相应的科学依据，同时依据产品的具体特性对注册申报资料的内容进行充实和细化。

本审评要点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审评要点。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研究资料和验证资料。

本审评要点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

病理图像人工智能分析软件是指基于数字病理图像，采用深度学习等人工智能技术实现对病理图像的分割、检测等功能的医疗器械。数字病理图像包括经数据成像设备获得的显微镜下细胞或组织图像、全玻片数字扫描技术（Whole Slide Imaging, WSI）的病理图像等。产品在医疗机构和/或医学实验室使用，可辅助病理医师为疾病的诊断、预后、治疗等提供信息，不能作为临床诊断决策的唯一依据。

本审评要点基于全玻片数字扫描技术( whole slide imaging，WSI)的病理图像人工智能分析软件产品临床评价进行阐述，用于经数据成像设备获得的显微镜下细胞或组织图像分析的相关产品临床评价可以适当参考本审评要点。

二、临床评价路径

病理图像人工智能分析软件临床评价应符合《医疗器械临床评价技术指导原则》要求。影响该软件性能因素包括核心算法、算法训练等多个方面。针对两个不同的分析软件，即使采用的核心算法相同，但是，不同的分析软件在算法训练过程中使用的训练集不同，不同训练集的标注可能存在差异，因此，针对该类产品临床性能的确认应采用临床试验的方式，不适合采用同品种比对的方式。

三、临床试验

在开展临床试验之前，病理图像人工智能分析软件产品应完成产品开发，确保产品已经定型。基于该类产品的特点，产品在开发过程中需采用满足要求的数据对算法进行训练，在算法训练过程中，应保证训练集数据标注的质量以及训练集具有足够样本量，确保经过训练的软件在临床前研究中性能稳定。在软件开发过程中，有时存在算法调优的情况。需要注意的是，用于开展临床试验的产品必须为完成算法训练及调优的产品。产品临床试验启动后不能对算法进行训练或调优。产品应进行充分的临床前研究，以保证产品临床试验过程中，不会对受试者产生额外的伤害。

1.临床试验机构

考虑到病理图像判读容易在机构和阅片医师间产生差异，产品临床试验建议采用多中心临床进行，临床试验机构数量建议不少于3家。临床试验机构应获得国家药品监督管理局备案认可。临床试验机构应具有严格的质量管理体系，执行实验室内部日常质量控制。

临床试验建议选择具有相关学科优势的临床试验机构。临床试验中应根据试验设计设定不同资历的病理医师参与试验，病理医师资历划分应合理。

2.临床试验目的

产品临床试验应该能够充确认产品临床性能，证明产品能够实现其宣称的预期用途。病理图像人工智能分析软件的临床应用，应对病理医师的诊疗活动有所获益，如：提高病理医师诊断准确性，提高病理医师阅片重现性，缩短病理医师阅片时间等。临床试验应确认主要评价指标的优效性。

3.临床试验设计

根据产品特点和产品性能评价需要，病理图像人工智能分析软件临床试验可能包括不同的临床试验目的，有必要针对各个临床试验目的，分别进行科学的临床试验设计，包括选择适当的临床试验设计类型。本审评要点以提高病理医师诊断准确性或缩短阅片时间为例，进行临床试验设计阐述。

此类产品临床试验一般需考虑开展对照研究，对照研究有多种设计方法，根据产品特征及临床试验实际，可选择多阅片者多数据样本（multiple reader multiple case，MRMC）、随机交叉自身对照等试验设计。如临床试验采用其他设计，在阐明设计科学合理的前提下，也可用于评价产品临床性能。以下仅详细阐述多阅片者多数据样本（multiple reader multiple case，MRMC）、随机交叉自身对照两种设计。

3.1多阅片者多数据样本（multiple reader multiple case，MRMC）设计

如采用此种设计，整个临床试验由协调研究者组织，各临床试验机构阅片者整体进行分组，各临床试验机构入组的病例共同组成分析集。临床试验将参与的阅片者分为两组。两组病理医师分别在软件的辅助下及未使用软件的情况下对入组的病理图像（或病理切片）进行判读，判读结果以临床参考标准为准，分别评价软件辅助下及未使用软件的情况下病理医师诊断准确性。

试验过程分为两个阶段，第一组病理医师第一阶段在使用软件辅助的条件下完成所有病理图像（或病理切片）的判读；经一定的洗脱期后，第二阶段在不使用软件辅助的条件下完成所有病理图像（或病理切片）的判读。第二组医师第一阶段在不使用软件辅助的条件下完成所有病理图像（或病理切片）的判读；经一定的洗脱期后，第二阶段在使用软件辅助的条件下完成所有病理图像（或病理切片）的判读。该试验中医师使用软件辅助的判读结果作为试验组，不使用软件辅助判读的结果作为对照组，比较试验组与对照组同临床参考标准对比病理图像（或病理切片）判读的准确度。阅片过程中每名医师需分别在使用软件和不使用软件的条件下都完成所有入组病理图像（或病理切片）的阅片。

3.2随机交叉自身对照设计

该设计应在每家临床试验机构单独开展，临床试验协调研究者最终将各机构数据汇总统计。每家临床试验机构由资质满足要求的研究者参与。临床试验过程中研究者开展两个阶段工作，一是，研究者使用软件对病理图像（或病理切片）进行判读，二是，上述研究者未使用软件分别对入组的病理图像（或病理切片）进行判读。判读结果以临床参考标准为准，分别评价软件辅助下及未使用软件的情况下病理医师诊断准确性。

试验过程将入组的病理图像（或病理切片）随机平均分为两组，临床试验分为两个阶段，第一阶段病理医师在使用软件辅助的条件下完第一组病理图像（或病理切片）的判读，同时在不使用软件辅助的条件下完成第二组病理图像（或病理切片）的判读；经一定的洗脱期后，第二阶段在不使用软件辅助的条件下完成第一组病理图像（或病理切片）的判读，同时在使用软件辅助的条件下完第二组病理图像（或病理切片）的判读。该试验中医师使用软件辅助的判读结果作为试验组，不使用软件辅助判读的结果作为对照组，比较试验组与对照组同临床参考标准对比，病理图像（或病理切片）判读的准确度。阅片过程中每名医师需分别在使用软件和不使用软件的条件下都完成所有入组病例图像的阅片。

4.临床试验参考标准

临床参考标准是指现有条件下临床上可获得的能够用来确定受试者目标状态的最佳方法，通常来自临床和实验室的医学实践，包括：现有条件下公认的、可靠的、权威的疾病诊断标准（如组织病理学检查、影像学检查、病原体分离培养鉴定、长期随访所得的结论等），疾病诊疗指南中明确的疾病诊断方法，行业内专家共识推荐的或临床上公认的、合理的参考方法等。临床参考标准可能是一种方法，也可能是多种方法相结合。

如根据产品临床试验要求，临床试验过程中需要以病理阅片结果作为参考标准。一般而言，病理阅片的参考标准为多位具有资深经验的病理医师组成的病理医师专家组共同对病理切片进行判读的结果，专家组中病理医师的数量一般不少于3人。参考标准应明确病理结果判定的决策机制，特别是针对专家组中不同病理医师判读结果存在差异的情况下，应明确后续的处理措施及最终结果确定的依据。

5.临床试验入组人群

病理图像人工智能分析软件临床研究入组的受试者应为产品预期适用人群。受试者应具有代表性，如针对肿瘤辅助诊断的产品，入组的受试者应包括各种组织类型、不同的肿瘤分级、分期以及相关良性疾病、其他易混淆、干扰病例等。

为了更加科学的评价产品性能，临床试验应包含部分病理医师难以诊断的病例及阳性阈值临近范围内的病例。

根据产品评价要求，临床试验入组的病例应与算法训练及调优的病例不同。

6.临床评价指标

针对定性分析，临床试验评价指标特别是主要评价指标应根据临床试验目的及设计确定。临床试验主要评价指标为使用软件的病理医师与未使用软件的病理医师阅片与临床参考标准对比的阳性一致率、阴性一致率；使用软件的病理医师与未使用软件的病理医师阅片效率；软件独立阅片与临床参考标准的阳性一致率、阴性一致率等。

针对半定量/定量分析，临床试验应设置合理的评价指标如等级符合率、回归分析的回归系数、截距、相关系数和决定系数等。

临床试验应考虑整体的优效性评价，以定性分析为例，如：使用软件的病理医师与未使用软件的病理医师阅片同临床参考标准对比的阳性一致率或阴性一致率的有效性；或者确保使用软件的病理医师与未使用软件的病理医师阅片同临床参考标准对比的阳性一致率和阴性一致率的优效或非劣效前提下，阅片效率的优效性。

7.样本量估算

样本量一般根据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。

7.1确认软件提高阅片准确性的样本量估算

该类产品临床试验的主要评价指标为使用软件作为辅助的病理医师与未使用软件的病理医师在阅片准确性方面的差异。软件辅助病理医师阅片准确性与病理医师单独阅片相比，准确性具有显著提升，软件的临床意义才能被确认。因此，建议采用优效性模型或优效性模型与非劣效性模型结合进行样本量估算。

以定性产品为例，临床试验应根据阳性一致率估算阳性病例数量，根据阴性一致率估算阴性病例数量。临床试验可针对阳性一致率、阴性一致率均采用优效性模型进行样本量估算；也可选择其中一个指标采用优效性模型进行样本量估算，另一个指标采用非劣效性模型进行样本量估算。

7.2确认软件缩短阅片时间的样本量估算

该类产品临床试验的主要评价指标为使用软件作为辅助的病理医师与未使用软件的病理医师在阅片时间等方面的差异。软件辅助病理医师阅片时间与病理医师单独阅片相比，阅片效率具有显著提升，同时，软件辅助病理医师阅片与病理医师单独阅片相比，阅片准确性应无显著差异，软件的临床意义才能被确认。因此，建议针对阅片时间的评价采用优效性模型进行样本量估算，同时针对阳性一致率和阴性一致率采用非劣效性模型或优效性模型进行样本量估算，临床试验样本量应能够同时满足上述要求。

临床试验样本量估算还应考虑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。针对确认软件缩短阅片时间的样本量估算，应确保总研究的Ⅱ类错误概率β设定不大于0.2。

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例检出的诊断学研究文献获得。

8.统计分析

临床试验应采用合理的方式进行统计分析。以定性产品为例，临床试验中针对软件作为辅助的病理医师、未使用软件的病理医师、分析软件阅片的准确性进行统计分析，应分别以临床参考标准为对照，以2×2表的形式统计阅片结果的阳性一致率、阴性一致率及总一致率。同时针对二者的阳性一致率、阴性一致率及总一致率分别按照要求进行假设检验或区间分析。

确认软件缩短阅片时间的临床试验还应针对阅片效率进行统计分析。

9.偏倚控制

为了控制临床试验的偏倚，研究过程中应针对病例/切片应进行设盲，使研究者在试验过程中不知晓受试者疾病诊断或其他相关检测等信息，从而避免引入偏倚。

临床试验涉及同一个病理医师两次阅片的情况，病理医师在进行下一次阅片时应保留一段时间的记忆清除期（脱敏期），此段时间一般不少于两周。每次阅片时建议人为纳入一定数量的不同的干扰病例。干扰病例应同时具有阳性和阴性病例，将干扰病例混入分析组中，使研究者同时针对分析组中的病例及干扰病例进行操作，但干扰病例结果不纳入统计分析。

针对适用于显微镜下组织或细胞图像的产品，临床试验过程中应注意，临床试验开展过程应与产品临床应用过程相符，即同一研究者参与显微镜下图像的选择与结果判读全过程。

10.质量控制

临床试验开始前，应进行相关培训，以确保研究者熟悉并掌握相关试验方法，最大限度控制试验误差。整个试验过程都应处于有效的质量控制下，最大限度保证试验数据的准确性及可重复性。

临床试验应严格按照产品说明书进行操作，应根据相关要求进行必要的HE染色，应设置合理对照。

四、关于配套使用产品的要求

产品临床试验应明确配合使用的检测试剂及图像采集设备。产品配合使用的检测试剂及图像采集设备应固定，且与产品临床前研究一致。

临床试验过程中软件运行环境如：运行所需的硬件配置、软件环境和网络条件应与产品临床前研究一致。其中硬件配置包括处理器、存储器和外设器件，软件环境包括系统软件、支持软件和安全软件，网络条件包括网络架构（BS、CS）、网络类型（广域网、局域网、个域网）和带宽。

五、产品适用范围

病理图像人工智能分析软件仅在医疗机构与配套的检测检测试剂、病理图像扫描仪配合使用。适用于病理检测过程中对病理图像的辅助识别，辅助病理人员识别出目标切片。本产品不单独用于病理诊断，不单独作为临床诊疗决策依据，所有经软件判读的样本最终结果需要医师确认。该产品经培训合格的医师使用。图像采集由指定型号的显微镜或图像采集设备，并经培训合格的人员对图像质量进行认可。

六、参考文献

[1] 原国家食品药品监督管理局.医疗器械临床试验设计指导原则:食品药品监管总局通告2018年第6号[Z].(2018-01-04)[2021-05-01].https://www.cmde.org.cn/CL0058/6

882.html.

[2] Xiao-Hua Zhou，Nancy A.Obuchowski，Donna K.McClish. 诊断医学中的统计学方法（第二版） [M].北京:高等教育出版社，2016。

[3] 李卫.医疗器械临床试验统计方法（第二版） [M].北京:科学出版社，2016。

