文章编号：1671-7104(2019)05-0379-05

美国上市影像类人工智能辅助诊断软件临床评价的研究与思考

【作　　者】胡凯，甄辉，杨辉，夏建松，何蕊
浙江省医疗器械审评中心，杭州市，311121

【摘　　要】人工智能，作为当前信息技术的重要突破，正在越来越多的行业中得到重视并运用于实践。将人工智能应用到医疗领域的研究也逐渐成熟并有部分产品问世。美国已有相关产品得到FDA的审批，其主要是影像类软件产品。该文通过研究美国FDA审批的影像类人工智能辅助诊断软件的案例，集中分析代表性的产品的上市途径、临床评价方式、临床数据处理，归纳美国FDA对影像类人工智能辅助诊断软件的临床评价特点。最后对我国同类产品的临床评价可能遇到的问题进行思考，并提出相关的建议。

【关 键 词】辅助诊断软件；人工智能；临床评价；医疗器械
【中图分类号】F203
【文献标志码】A
doi: 10.3969/j.issn.1671-7104.2019.05.019

Study on the Clinical Evaluation of Image-based Artificial Intelligence Aided Diagnosis Software Approved in the United States

【 Writers 】HU Kai, ZHEN Hui, YANG Hui, XIA Jiansong, HE Rui
Zhejiang Medical Device Evaluation Center, Hangzhou, 311121

【 Abstract 】Artificial intelligence, as the breakthrough of current information technology, is gaining importance and being applied in more and more industries. Research on the application of artificial intelligence to the medical field has gradually matured and some products have come out. Relevant products in the United States have been approved by the FDA, mainly for image-based software products. In this paper, we studied the cases of image-based artificial intelligence aided diagnosis software approved by the US FDA, and analyzed the listing routes, clinical evaluation methods and clinical data processing of representative artificial intelligence products, and summarized the clinical evaluation characteristics of FDA for image-based artificial intelligence aided diagnosis software. Finally, problems that may be encountered in the clinical evaluation of similar products in China were considered, and relevant suggestions were put forward.

【Key words 】aided diagnosis software, artificial intelligence, clinical evaluation, medical device

0 引言

自美国国会通过“21世纪健康法案”以来，美国医疗产业界加大对人工智能产品的投入与应用。2019年2月，美国总统特朗普签署“美国人工智能倡议”行政令[1]，释放了产业扶持信号。我国十九大报告中明确指出“加快建设制造强国，加快发展先进制造业，推动互联网、大数据、人工智能和实体经济融合”[2]。2018年10月，中央政治局集体学习了人工智能产业发展[3]，也将医学作为人工智能重要的应用领域。可见，人工智能医疗器械产业已经受到全世界的重视，成为未来战略的一部分。如何服务好人工智能器械产业，引导其健康良性发展是医疗器械监管从业人员共同的重任。据报道[4]，2018年度FDA审批通过了十余款人工智能产品，文章列举的产品中有10款通过510(k)上市、4款通过DE NOVO上市[5]，预期用途包括影像辅助诊断、生理信号检测与诊断、决策支持等。其中，数量最为丰富的产品为影像辅助诊断产品，涵盖神经科、心胸科、眼科。影像类人工智能辅助诊断软件，由于其数据类型为“二维或三维图像”，相比于体外诊断领域的量化指标，其特征的提取较为抽象和复杂，对诊断医生的经验要求较高。本文将选取其中的若干个产品，探讨其临床评价方式并概括要点，探索国内同类产品在临床评价时的相关思路。

收稿日期：2019-03-20
作者简介：胡凯，E-mail: hukai666@126.com
通信作者：杨辉，E-mail: 15397081802@163.com

1 产品上市途径与临床评价

本文讨论了2018年FDA上市总结中明确提出，使用人工智能（AI）或深度学习的几款影像类人工智能辅助诊断软件：糖尿病视网膜病变辅助筛查软件IDx-DR、骨折辅助检测软件OsteoDetect、脑卒中辅助检查软件ContaCT、颅内出血辅助检查软件

AccipioIx、颅内出血辅助检查软件BriefCase、乳腺异常辅助诊断软件QuantX、冠状动脉钙化辅助评估软件HealthCCS、影像辅助分析软件ArterysMICA。其中IDx-DR、OsteoDetect、ContaCT、QuantX为低风险新产品上市途径De Novo（申请时无实质性等同上市产品），AccipioIx、BriefCase、HealthCCS、ArterysMICA上市途径为510(k)。

根据美国相关法律和规章要求，各上市途径中临床评价的要求不尽相同。对于高风险III类产品上市途径PMA，21 CFR 814.20部分明确要求提交包含人类数据的临床研究[6]；510(k)则基于“实质性等同”的评价方式，可通过非临床或临床测试数据来证明与对比产品的等同性（21 CFR 807.92）；而De Novo途径的诞生基于FD&C法案513(f)部份，产品风险尚未达到III类但尚无实质性等同产品，目前CFR暂未有具体的要求，仅有指南《De Novo Classification Process》对临床评价的要求作出建议，要求提供必要的临床和非临床数据。

但从FDA允许上市的若干影像类人工智能辅助诊断软件来看，除ArterysMICA均提供了与临床相关的数据，分别来自于美国境内外的临床机构。根据21 CFR 812/807.87等规定，对于取自临床研究的数据需符合患者利益及伦理审查机构等相关要求，前序需申请IDE。而对于使用回顾性临床数据进行性能测试，目前未有相关临床法规要求，但应保障数据的真实性。

基于这5个产品的临床评价情况，并结合其预期用途表述要点，分析如表1所示。

从上表可以观察，产品均为辅助诊断工具，并不替代专业医生最终的诊断结论；所输出的结果，不具有临床诊断意义，为物理学、图像几何学参数或基于这些参数所建立的评分指数、通知。

上表结合FDA公开发布文件，这8个II类影像类人工智能辅助诊断软件的临床评价具有以下特点：

（1）用于非紧急情况筛查的辅助诊断产品无同品种产品时，若预期用途同时涉及阴性和阳性判断提示，应进行临床研究；对于流程优化型产品，应在临床研究中考察使用者在有辅助和无辅助情况下的诊断结果差异；

（2）紧急情况下疾病筛查的辅助诊断产品若预期用仅涉及阳性判断提示，无论是否具有同品种，需临床数据性能测试，不进行临床研究；

（3）产品在有同类产品时均接受同品种比对。

根据FDA的CFR条款特点，对于影像类人工智能辅助诊断软件，部分II类产品除510(k)的一般控制外，可能以利用回顾性临床数据进行性能测试作为后续同类II类产品的特殊控制方式（Special Control）。

2 临床评价方案及过程的特点

根据前文，影像类人工智能辅助诊断软件的临床数据可来自于性能测试和临床研究两方面。

表 1人工智能辅助诊断软件临床评价信息
Tab.1 Information on the clinical evaluation of image-based artificial intelligence aided diagnosis software

产品名称	上市途径	临床评价方式	预期用途	输出结果
IDx-DR	De Novo	临床试验（前瞻性，软件识别结果和读者识别结果配对分析）	产品提供糖尿病视网膜病变自动筛查；由专业医生确认其结果	糖尿病视网膜病变自动筛查结果
ContaCT	De Novo	临床数据性能测试（回顾性，软件识别结果和已标注数据配对分析，软件辅助工作流和标准工作流的配对分析）	用于通知专业医生的并行工作流辅助工具，在紧急状况下标示疑似大血管闭塞，不替代诊断	阳性结果的通知，不改变原有图像的标记
AccipioIx	510(k)	实质性等同对比，临床数据性能测试（回顾性，软件识别结果和已标注数据配对分析，软件辅助工作流和标准工作流的配对分析）	在急症护理环境中，辅助分析颅内出血特征，调整可疑发现案例的工作流优先级，不替代临床决策	案例处理优先级标识
BriefCase	510(k)	实质性等同对比，临床数据性能测试（回顾性，多国中心，软件识别结果和已标注数据配对分析，软件辅助工作流和标准工作流的配对分析）	辅助医生优化流程，标示可疑颅内出血（下文简称ICH）阳性，调整查看的优先级，不用于最终诊断	可疑结果通知，改变优先级，不改变原有图像的标记
OsteoDetect	De Novo	临床数据性能测试（回顾性，软件识别结果和已标注数据配对分析），临床研究（回顾性，全交叉的多读者多案例，软件辅助读者的结果和无辅助读者的结果配对分析，以及软件辅助读者和无辅助读者各自与金标数据的配对分析）*注	帮助医生在阅片期间识别桡骨远端骨折，不作为最终诊断结论，可疑病例需医生确诊	疑似骨折边界的标注框，检测结果标签
QuantX	De Novo	临床数据性能测试（回顾性，软件识别结果和已标注数据配对分析），临床研究（回顾性，全交叉的多读者多案例，软件辅助读者的结果和无辅助读者的结果配对分析，以及软件辅助读者和无辅助读者各自与金标数据的配对分析）	分割和分析用户选择的感兴趣区域，辅助医生筛查乳房异常，不作为最终诊断	感兴趣区域的QI指数评分
HealthCCS	510(k)	实质性等同对比，临床数据性能测试（回顾性，多国中心，软件识别结果和已标注数据配对分析）	评估冠状动脉中的钙化斑块，供医生使用该信息进行进一步患者管理	Agatston等效风险评分
Arterys MICA	510(k)	实质性等同对比	肿瘤学工作流的工具，帮助医生确认病变，包括评估，量化，随访和记录任何此类病变，不产生任何诊断或潜在的发现	模块1：心脏功能量化包括每搏输出量等指标 模块2：Lung-RADS、LI-RADS数据报告
*注：临床研究所用“金标数据”是指回顾性病例中已形成的诊断结论，该结论的形成不仅依据研究针对的影像数据，而是由综合的临床数据所形成				

2.1 人工诊断/标注的单个病例应由多位专家或专业医生参与

其中IDx-DR的前瞻性临床研究，OsteoDetect、HealthCCS的临床数据性能测试使用3位读者/专家采取多数表决法获得结果，与软件结果进行对比；ContaCT的临床数据性能测试采用“双人复核+第三人仲裁”法获取结果，与软件结果对比；OsteoDetect、QuantX的临床研究则进行了全交叉研究，每个医生对每个病例进行阅读。

2.2 主要评价指标重点关注灵敏度、特异性及AUC

从已上市产品的输出结果分析，目标症状均有阴阳性的二分类特征。绝大多数对照判断均为单一的“阴/阳—是/否”维度；而OsteoDetect是性能测试使用框柱法，软件结果表现除“是/否”外，还具有空间位置特异性，因此需特别注意假阳性（专家标注阳性结果为所有专家标注框边界的并集，并集结果为0像素则是阴性）的两类表现方式：软件输出框结果非空，专家标注结果非空，但两者无交集；软件输出框结果非空，专家标注结果为空。

对于二分类结果，产品研究可采用混淆矩阵来分析灵敏度和特异性。部分产品的临床评价终点公开数据如表2。

表2 部分产品临床评价终点
Tab.2 The clinical endpoints of some products

产品名称	灵敏度(%)	特异性(%)
IDx-DR	87.0	90.0
ContaCT	87.8	89.6
AccipioIx	92.0	86.0
BriefCase	93.6	92.3
OsteoDetect	92.1	90.2
QuantX	84.8	51.7

此外，全交叉的多读者方案临床研究的产品还关注了ROC曲线下的面积，进行了两组对照的AUC面积差的假设检验。在临床试验过程，得到了纯人工组（无软件辅助）的ROC曲线，以及辅助组的ROC曲线，验证两者的曲线下面积差值的假设检验。

2.3 次要评价指标根据预期用途的应用场景设置

由于不同产品在临床决策的作用不同，所针对适应症的缓急等特征不同，因此在次要评价指标的设置具有比较大的差异性。

（1）非紧急情况使用的辅助诊断产品考虑检测结果的重复性。应考虑同一素材其结果在不同时间和地点的测量值的重复性。IDx-DR软件进行了一项临床子研究，对24名受试者进行了重复试验。其中12名受试者人工判读对照组结果阴性，12名受试者为阳性。每名受试者由3名不同的操作员在两台不同的Topcon眼底摄像机上成像。每人接受完整的IDx-DR判读10次产生10组图像，共240组图像。结果一名受试者的5张图像不能被分析，其余235张（97.9%）图像可被分析。对于24名受试者中的23名，每人所有经IDx-DR软件输出结果相同。因此，IDx-DR的输出结果重复性（99.6%）好，且对人员和设备不敏感。

（2）对于结果呈现为框注感兴趣区域[7]的辅助诊断产品，考虑其结果的位置精确程度。OsteoDetect的性能测试设置了中心位置对比，软件预测边界框的图心与参考标准边界框的图心之间的平均像素距离是33.52（标准差为30.03）。图像的平均大小为1 663像素 × 1 109像素（面积为1 844 267个像素），以及参考标准边界框的平均面积为30 164个像素，软件的预测边界框的平均面积为34 924个像素，中心差远小于框的长宽尺度。可见，OsteoDetect通常会在桡骨远端骨折部位检出点附近画出边界框。

（3）对于紧急情况下使用的辅助诊断产品，时间与诊疗效果可能相关的适应症，应考察其在工作流中的时间指标。两款用于检测脑部血管状态的软件，性能测试均对真阳性病例统计观察了时间指标，软件检测出阳性结果的时间和按照临床的标准流程判断出阳性结果的时间进行了对比。$t_{软件}/t_{标准}$的值为51.4 min、68.1 min，具有统计学意义。应特别注意预期用途中未宣称对阴性结果优化工作流程，所以不需考察阴性结果的时间指标。

2.4 其他考量因素

2.4.1 最坏情况考虑

由于个体差异的存在，有较小概率人工判读没有结果。这种情况下，为最大限度地保证产品的安全性，将人工判读无结果修正为人工判读阳性。例如，IDx-DR在892名中的73名不可分析的受试者图像中，有35名（4%）受试者无法通过人工评分（FPRC）[8]。在最坏的情况下，假设这35名受试者均患有糖尿病视网膜病变，则这73例病例灵敏度和特异性分别为80.7%和89.8%，仍符合总体的临床评价指标。

2.4.2 阴阳性病例数量

由于部分适应症可能存在人群发病率较低的情况，导致特异性虚高。为防止这一现象的发生，应适当调整阴阳性病例的比例：

（1）开展回顾性的临床研究，可选择已有的病例数据开展，尽量控制阴阳性病例的数量相等。但应注

意选择过程对试验操作双方“双盲”。对人工判读/标注者盲，病例入组操作人员和判读操作人员不能相同；对软件“盲”是指，入组的病例不能在前期已作为软件的基础训练病例（如OsteoDetect）。

（2）前瞻性实验人为富集阳性病例。为防止阳性病例收集过慢，人为设置条件加快阳性病例入组，但应从统计学角度观察和排除该条件对结果的影响，同时尽可能防止阳性病例的漏判。如IDx-DR软件依据糖化血红蛋白水平（HbA1C）来收集阳性病例，并通过逻辑回归的方式排除了这一影响；同时对糖尿病视网膜病变，阳性的判断综合考虑眼底相机数据、OCT数据、评分法的结果[9]。

2.4.3 真实世界数据

通过真实世界已有相关研究，来辅助临床性能测试或临床研究的指标来衍生推论，从而证明产品的临床价值。如已有足够文献指出，神经血管专家在LVO患者的管理中起着关键作用，并且神经血管专家的早期介入明显有益于LVO患者。因此，通过软件检测发现LVO阳性的平均时间少于标准流程操作所用平均时间，可证明产品有利于LVO患者。

2.4.4 数据格式

预期用途中描述的硬件适配的不同可能造成输出文件的不同，包括文件格式和因操作造成的图像分辨率、图像层间距等的不同。对于文件格式，可尽可能采用统一标准处理，如DICOM。

2.4.5 可用性因素

部分辅助诊断产品的模块和功能较为复杂多样，使用者学习曲线较为平缓。在实际操作过程中可加强对使用者的培训，防止数据不必要的脱落。也可通过开展子研究来排除可用性因素干扰。

3 我国影像类人工智能辅助诊断软件临床评价的思考和建议

对比上述美国影像类人工智能辅助诊断软件的上市途径和临床评价特点，我国从业人员在临床评价时存在诸多困难，如缺乏对临床影像标注的标准、同类产品临床对比数据较难获取、尚未有统一的产品标准、产品性能泛化能力弱等。结合以上情况，提出下文建议。

3.1 鼓励公开部分临床指标评价终点

FDA对几款影像类人工智能辅助诊断软件公开了临床评价信息，企业在申请文件中参照或对比同类产品的临床或者性能测试数据，而这些数据都可在公开数据库中获取。在中国，除了相关方主动公开（如发表文献、审评报告公开等），较难获取临床数据和真实世界数据[10]。这直接增加了企业在同品种比对时获取数据的时间和经济成本。对此，一方面对于企业公开相关信息给予支持和鼓励；另一方面，可参考美国相关机构，将信息公开制度化。如FDA通过510(k)的summary文件明确了几款影像类人工智能产品的部分评价终点数据（灵敏度、特异性、一致性）；NIH（美国国立卫生研究院）则将一些临床试验予以公开和公布[11]。公开部分临床评价数据有利于降低同品种对比数据获取难度，缩短低风险辅助诊断产品的上市时间。

3.2 推动产业整合资源建立公开的验证集数据库

在研发初步完成后，高效的产品研发检测给后续的临床评价做好铺垫。“临床数据性能测试+临床研究”的模式可以理解为“验证+测试”，所使用数据分别为“验证集”与“测试集”[12]。对于人工智能辅助诊断软件产品，通过高效的验证集来调整优化产品的参数，为临床研究提供最佳的产品性能状态。对于后续进行临床研究的产品，可考虑使用“公开数据+非公开数据”进行前期的临床数据性能测试。美国已建立部分开放资源，如NIH的CT图像开放数据集DeepLesion[13]，美国国家癌症研究所（National Cancer Institute）的胸部影像数据集LIDC-IDRI[14]，斯坦福大学的上肢肌肉骨骼X光片数据集MURA[15]等等。但也应注意，在产品的训练时如使用了公开数据集训练，则应避免验证集使用同一个数据集而造成结果偏倚。

3.3 建立提前介入评价的程序

软件类产品的研发过程较为复杂，特别是人工智能产品训练集、验证集、测试集等各个环节的优化，需要大量的前期工作基础。而等到产品上市审批时，在短时间对智力密集型成果进行评价有较大难度。目前美国FDA对产品提前介入形式有“早期可行性研究（EFS）”[16]和“预认证（pre-cert）”[17]。EFS在提交临床研究（IDE）之前，企业预提交器械概念描述、临床背景和基本原理，目标在于与FDA就基于风险分析、非临床测试和临床风险缓解策略支持研究启动所需的信息达成共识。而对于数字软件产品，FDA于2017年启动了“pre-cert”计划，并公布了一批参与该计划的企业。2019年1月公布了该计划最新的1.0版本，以简化版De Novo途径对相关厂家的产品

进行提前介入。该计划的企业需先满足质量体系法规（QSR）的要求。国内可参考美国的这一模式，提前介入人工智能产品的评价，加大对人工智能产品的支持力度。

3.4 探索临床动态再评价体系

由于人工智能产品自身特性，产品会不断完善、更新和迭代。最直观的如产品训练集发生变化，可能导致对于同一样本前后两次处理结果不相同，并且不能完全保证这种变化是有利于提高产品的性能。目前尚无统一模式对这一变化进行量化评价。可通过建立完善临床再评价体系，利用包含独立非公开的标准测试数据集合在内的工具，定期对产品的性能进行综合的临床评价和“校准”[18]，保证产品的可靠性。同时，参照再评价体系，成立合规的第三方再评价中心，客观上促进数据的标准化和网络资源的数据安全，并促进影像类人工智能辅助诊断软件质量的提升，引导产业健康务实发展。

参考文献

[1] White House Fact Sheet: President Donald J. Trump is accelerating America’s leadership in Artificial Intelligence[R/OL]. (2019-02-11)[2019-03-01]. https://www.energy.gov/articles/white-house-fact-sheet-president-donald-j-trump-accelerating-america-s-leadership.
[2] 习近平. 决胜全面建成小康社会夺取新时代中国特色社会主义伟大胜利——在中国共产党第十九次全国代表大会上的报告[Z]. (2017-10-27)[2019-03-01]. http://www.gov.cn/zhuanti/2017-10/27/content_5234876.htm.
[3] 抓住新一代人工智能发展的重大机遇[Z]. (2018-11-02)[2019-03-01]. http://www.gov.cn/xinwen/2018-11/02/content_5336686.htm.
[4] TOPOL E J. High-performance medicine: the convergence of human and artificial intelligence[J]. Nature Med, 2019, 25(1): 44-56.
[5] Medical Device Databases, US Food and Drug Administration[EB/OL]. [2019-03-01]. https://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/Databases/default.htm.
[6] CFR - Code of Federal Regulations Title 21[EB/OL]. [2019-03-01]. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm.
[7] 赵治翔, 白万民, 刘白林. 图像感兴趣区域提取的研究[J]. 信息系统工程, 2017(11): 156-156.
[8] MARKUS R, ELLEDGE J, SIMADER C, et al. Evaluation of optical coherence tomography findings in age-related macular degeneration: a reproducibility study of two independent reading centres[J]. Br J Ophthalmol, 2011, 95 (3): 381-385.
[9] GRÉGOIRE G. Logisticregression[C]. Statistics for Astrophysics Methods and Applications of the Regression, 2013: 89-120.
[10] Use of real-world evidence to support regulatory decision-making for medical devices[Z]. (2017-08-31)[2019-03-01].
[11] ClinicalTrials.gov is a database of privately and publicly funded clinical studies conducted around the world[Z]. [2019-03-01]. https://www.clinicaltrials.gov/.
[12] 王浩, 孟祥峰, 李澍, 等. 数据集在人工智能医疗器械质控中的角色与要求[J]. 中国医疗器械杂志, 2019, 43(1): 54-57.
[13] KE Y, WANG X S, LU L, et al. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning [J]. J Med Imaging, 2018(5): 036501
[14] RAFAEL W, MARTIN B, EKTA D, et al. Agreement of CAD features with expert observer ratings for characterization of pulmonary nodules in CT using the LIDC-IDRI database[C]. Medical Imaging 2009: Computer-Aided Diagnosis, 2009, 20092512133657.
[15] https://stanfordmlgroup.github. io/projects/mura.
[16] Investigational Device Exemptions(IDEs) for early feasibility medical device clinical studies, including certain first in human(FIH) studies[Z]. (2013-10-01)[2019-03-01].
[17] Precertification pilot program working model version 1.0[Z]. (2019-01-07)[2019-03-01].
[18] 魏强, 陆平. 人工智能算法面临伦理困境[J]. 互联网经济, 2018(5): 26-31.

上接第364页

械产品生产企业与协议委托的灭菌机构共同合作完成的，产品的设计、生产和质量人员必须全程参与产品灭菌确认过程研究。建议监管部门根据医疗器械产品的风险等级和企业的质量管理体系自我评估和核查评估情况，主动对灭菌机构进行医疗器械注册质量管理体系延伸检查，切实保证无菌医疗器械的灭菌过程的有效性和稳定性。

参考文献

[1] 张秀丽, 薛玲, 王晨. 医用手套产品技术审查关注点探讨[J]. 首都食品与医药, 2016, 23(12): 4.
[2] 朱南康, 王春雷, 滕维芳. 医疗器械辐射灭菌的现状及进展[J]. 核技术, 2003, 26(3): 189-196.
[3] 翁辉, 胡昌明, 林玉清. 环氧乙烷灭菌确认时短周期运行的意义[J]. 中国医疗器械信息(医疗器械包装特刊), 2018: 54-55.
[4] 储云高, 钱虹, 朱颖峰. 医疗器械注册资料技术审评探讨[J]. 中国医疗器械杂志, 2017, 41(4): 286-288.