人工智能医疗器械辅助诊断及探测性能评估参数的讨论

孟祥峰，王浩，张超，任海萍*（中国食品药品检定研究院，北京 100050）

摘要　目的：人工智能医疗器械的应用越来越广泛，但目前并没有对其性能的评价标准。希望通过本文研究为人工智能医疗器械的客观评估提供帮助。**方法：**从不同的应用角度对人工智能医疗器械评估参数进行了梳理，比较了各个参数的特点和使用场景。**结果与结论：**不同的评估参数所适用的场景不同，评估结果也存在差异，在进行人工智能产品评价时应根据产品特性合理选择。

关键词： 人工智能医疗器械；性能评价；分类；分割

中图分类号：TP181；TH77　**文献标识码：**A　**文章编号：**1002-7777(2019)09-1026-06
doi:10.16153/j.1002-7777.2019.09.011

Discussion on Evaluation Parameters of Auxillary Diagnosis and Detection Performance of Artificial Intelligence Medical Devices

Meng Xiangfeng, Wang Hao, Zhang Chao, Ren Haiping* (National Institutes for Food and Drug Control, Beijing 100050, China)

Abstract　Objective: The application of artificial intelligence medical devices is more and more extensive, but there is no evaluation standard for their performance currently. It is hoped that this paper will provide some help for objective evaluation of artificial intelligence medical devices. **Methods:** The evaluation parameters of artificial intelligence medical devices were analyzed from different application perspectives, and the characteristics and application scenarios of each parameter were compared. **Results and Conclusion:** Different evaluation parameters can be used for different scenarios, leading to different evaluation results. Therefore, in the evaluation of artificial intelligence products, selection should be made according to the characteristics of the products.

Keywords: artificial intelligence medical devices; performance evaluation; classification; segmentation

人工智能医疗器械作为一种新兴的医疗器械，在辅助诊断、辅助筛查等诸多领域实现了突破。基于神经网络的深度学习可以帮助医生识别CT影像、病理切片、皮肤损伤、视网膜图像、心电图、内窥镜检查、面部和生命体征[1-3]。人工智能产品一般是对样本数据进行分类或对样本数据的异常特征进行标记或提取。对于它们的评估多为算法的评估结果与参考标准（临床“金标准”或有经验临床医生的诊断结果）进行比较[4]，使用召回率、特异性、准确度等参数的数值大小或曲线关系来表示产品的质量水平。对于人工智能产品不同的功能，如分类、分割、检出；或者不同的应用场

基金项目：中国食品药品检定研究院中青年发展研究基金课题“人工智能医疗器械软件性能评价方法研究”（编号 2018C5）
作者简介：孟祥峰；研究方向：生物医学工程、光学医疗器械检测、人工智能
通信作者：任海萍；研究方向：生物医学工程、医疗器械检定；E-mail：renhaiping@nifdc.org.cn

景，如体检应用、门诊应用；或者不同的评价目的如产品研发过程的评价、迭代后性能的评价、不同产品的比较评价，都应该依据自身特点合理地选择评价参数，因为不同的参数所体现产品的能力是不一样的。

目前，我国虽组建了人工智能医疗器械归口单位，相应的标准也在不断的规划中，但现阶段国内外尚未建立人工智能医疗器械的评价标准与方法规范。本文对工智能产品的评估参数进行了分析，简述了各个参数的特点，这将有助于进一步明晰影像类人工智能产品的评价工作，为人工智能产品的研发和质控提供指导。

1 分类评估参数

人工智能（Artificial Intelligence，AI）医疗器械的辅助筛查、辅助识别或辅助诊断等功能多是给出患者数据的状态分类，如二分类的阴性（非患病）、阳性（患病），或多分类如糖尿病视网膜病变筛查的0期~VI期[5]。对于分类问题可采用混淆矩阵的方法[6]，见表1，进而计算灵敏度、特异性、准确率等参数。

表 1 多分类混淆矩阵（n 为分类种类）

分类	Pred_1	Pred_1	…	…	…	Pred_n
True_1	$N_{1,1}$	$N_{1,2}$	…	…	…	…
True_2	…	$N_{2,2}$	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
True_n	…	…	…	…	…	$N_{n,n}$

注：Pred_x（x=1～n）为被AI医疗器械分为x类的个数；True _x（x=1～n）为真实分类为x类的个数。

灵敏度：

$$P_{sen,i}=\frac{N_{i,i}}{\sum_{j=1}^{6}N_{i,j}}\times 100\% \quad (1)$$

特异性：

$$P_{spe,i}=\frac{\sum_{m=1}^{6}\sum_{n=1}^{6}N_{m,n}-(\sum_{j=1}^{6}N_{i,j}+\sum_{j=1}^{6}N_{j,i})+N_{i,i}}{\sum_{m=1}^{6}\sum_{n=1}^{6}N_{m,n}-\sum_{j=1}^{6}N_{i,j}}\times 100\% \quad (2)$$

准确率：

$$p_0=\frac{\sum_{k=1}^{6}N_{k,k}}{\sum_{i=1}^{6}\sum_{j=1}^{6}N_{i,j}} \quad (3)$$

$N_{i,j}$（i=1～n，j=1～n）为真实分类为i类，被AI产品判为j类的个数；$P_{sen,i}$为第i类为阳性，其他类为阴性的灵敏度；$P_{spe,i}$为第i类为阳性，其他类为阴性的特异性。

灵敏度指参考标准中实际的阳性样本被正确判断的比率，见式（1），用来评估人工智能产品对目标疾病的识别能力。相反，特异性是指参考标准中实际的阴性样本被正确判断的比率，见式（2），用来评估人工智能产品对非目标疾病的识别能力。而准确度是指所有样本被正确判断的比率，见式（3）。这些参数都是0~1的数值，越接近1表示算法的性能越好。

单一参数很高并不能说明产品的优劣。比如准确度，其数值与发病率有一定相关性，当某一类数据的样本量远大于另一类时，即使另一类全部判断错误也不会对准确度产生太大影响，所以即使分数很高，也无法对于特定类别的识别能力进行判断。所以大部分情况下可用多个参数同时用于产品性能的评估，比如用灵敏度和特异性两个参

数来评价产品的性能。一般成熟的产品算法的评估阈值是一定的，也就是灵敏度和特异性是唯一的。特定阈值下的参数只能体现产品应用性能的优劣，并不能评价产品算法的优劣，比如一个优质算法在一个存在偏倚的数据集上进行训练，产品出厂时并没有选择最优的阈值，这导致召回率等参数没达到预期。所以为了进一步评价算法的好坏通常采用ROC、Precision-Recall（P-R）曲线等来对产品进行评价[7]。在医用范畴，多数情况下正负样本比例差距较大（与发病率相关，Precision-Recall曲线在正负样本不均衡的情况下会出现较大波动），且ROC曲线包含混淆矩阵的所有信息（Precision-Recall曲线缺少混淆矩阵的真阴性个数信息），因此ROC曲线更常见。它通过调节算法的阈值来计算不同阈值下的灵敏度和特异性，以1减特异性为横坐标，灵敏度为纵坐标，绘制ROC曲线，ROC曲线不仅能体现算法在不同阈值下的泛化能力，同时，还通过计算曲线下面积（AUC）对不同的AI产品用一个参数进行比较。

2 分割评估参数

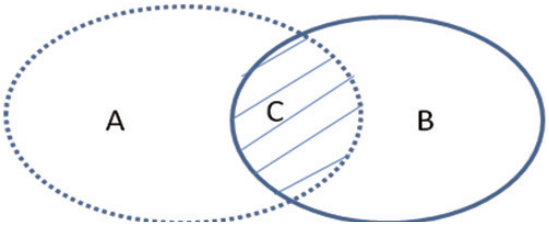
人工智能辅助检测功能多应用于影像识别类产品，其作用在于准确地识别图像中的病灶位置并进行边界分割，其分割性能多采用分割结果与参考标准比较，比如在FDA发布的计算机辅助探测（Computer-assisted Detection Devices）的510（k）提交指导原则[8]中提到了用分割区域的位置关系进行评价计算。目前比较算法中被广泛应用的评价方式有两种：Jaccard系数[也称之为交并比（IoU）]和Dice系数[9-10]。

交并比是指参考标准和人工智能算法区域交集与并集（见图1）的比率，见式（4）：

$$IoU = \frac{C}{A + B - C} \qquad (4)$$

Dice系数是指参考标准和人工智能算法区域交集与二者区域平均值的比率，见式（5）：

$$Dice = \frac{2C}{A + B} \qquad (5)$$



A. 参考标准的分割面积；B. 人工智能算法的分割面积；C. 参考标准与人工智能算法分割面积的重叠部分。

图1 尺寸分割评价参数举例

从公式（4）和（5）可以看出，虽然二者都是在0～1变化的数值，但相同情况下IoU数值要低于Dice系数，见式（6）：

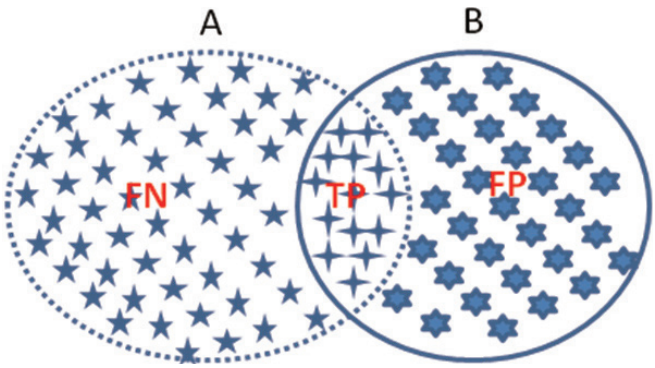
$$\frac{Dice}{IoU} = 2 - \frac{2C}{A + B}$$
$$C \le \min(A, B), \frac{Dice}{IoU} \ge 1 \qquad (6)$$

IoU比Dice系数提供了更宽的数值评估范围，尤其是在分割结果较差的情况下，IoU能更好地评估分割图像中的微小变化，对于不同产品的评价提供了更好的区分度；如果一个AI产品的分割性能进行了提升，随着重叠面积的增加，Dice系数呈线性变化，而IoU呈非线性变化，这对于同一产品分割性能的评价，尤其在算法整改后的评价上，Dice更为直观。

对于分割性能，在检出类AI产品如肺结节识别上，有可能会通过区域分割指标来判断真阳性（TP）和假阳性（FP），进而计算灵敏度、特异性等参数，就是要确定分割性能参数阈值的大小，这涉及了标记匹配的内容[11]。比如交并比不低于某一小于1的数，这个数值直接决定了是否被命中，进而影响灵敏度、精确度等参数。我们能判断越接近于1，算法是越优秀的，但是我们无法确切定义哪个百分比对于临床医生的使用是足够了，也就是检出来了，这部分还有待进一步研究。

此外，还可以把算法分割结果与参考标准当成两个像素集，二者重叠像素点记为TP，参考标准去掉TP部分为FN，算法分割结果去掉TP部分为FP，这样可以用召回率[见式（7）]和精确度[见式

（8）]两个参数对分割结果进行评价，见图2。这类似于对于病灶检出的评价方式。



A. 参考标准的分割区域；B. 人工智能算法的分割区域。

图2 像素点集合进行分割参数评价

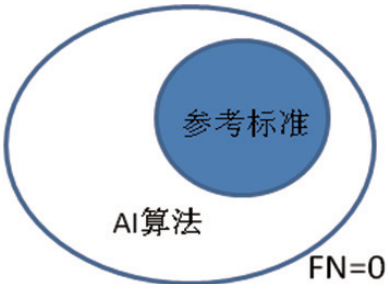
召回率：

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (7)$$

精确度：

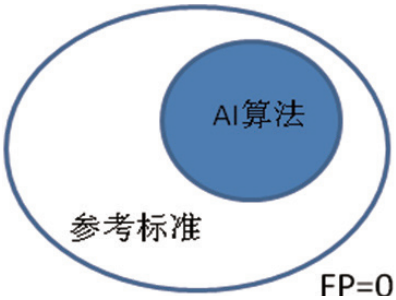
$$Prc = \frac{TP}{TP + FP} \times 100\% \quad (8)$$

当召回率为1时，参考标准被完全包裹在算法分割区域范围内，如图3所示；当精确度为1时，算法分割区域被完全包裹在参考标准范围内，如图4所示。通过两个参数不仅可以分析分割结果的相关度，还可以对分割的位置及形状进行判断，IoU和Dice虽然能分析分割结果的相关度，但通过参数不能判断分割面积的大小和相互包含关系。



参考标准被完全包裹在算法分割区域范围内，召回率为1。

图3 位置及形状示例一



算法分割区域被完全包裹在参考标准范围内，精确度为1。

图4 位置及形状示例二

3 检出算法的曲线评估参数

检出类算法一般会在一幅图像上诊断出多个异常，这种病灶检出的方式无法采用ROC曲线进行评价，因为假阳性的个数是没有限制的。这种情况一般采用FROC曲线来进行评价。其绘制方法是在不同的阈值下，计算算法的召回率和平均假阳个数（平均每个病人所含有的假阳个数）。以召回率为纵坐标，平均假阳个数为横坐标，绘制曲线。对于曲线评价，我们不仅希望从曲线的趋势图或曲线上特定点来评价算法的好坏，我们更希望通过曲线提取出一个综合参数，用这个参数对算法进行评价，比如ROC曲线的AUC。对于FROC，同样可以计算曲线下的面积，但这种方式可能需要调节多个阈值，计算量较大。且假阳结节的数量会因为产品的性能不同而不同，这导致FROC曲线横坐标（平均假阳个数）终点不一致，这样计算的面积很难进行横向比较。为解决这一问题，可以采用给横坐标一个限制，如横坐标都采用平均假阳个数8个，这样面积的理想值就进行了统一，但损失了一部分阈值下的数据考量。另一个问题，对于较好或较差的算法，平均假阳个数8个可能会太多或者太少，给评价带来一定的局限。我们还可采用曲线上召回率的平均值，这种对于线性度较好的曲线是个不错的选择，但对于线性度较差的曲线，可能会存在偏差。

Precision–Recall曲线是以召回率为横坐标，精确度为纵坐标绘制的曲线。曲线构造和ROC曲线类似[10]，曲线下面积理想值为1。该曲线的评价方式很好地继承了ROC曲线的优点，能够实现不同算法性能的评估和统计比较。

FROC与P–R曲线都包含了TP、FP、FN的信息，两种曲线有着各自的特点，见图5、图6。在FROC曲线上能更为直观地找到曲线的拐点，这对于产品研发阶段合理的阈值调节具有很大的帮助。但曲线在阈值无限小的情况下，召回率趋于定值，而平均假阳个数是在不断增加的，无法通过计算FROC曲线下的面积对AI算法进行评估，这一点上P–R曲线更具优势。P–R曲线理想状态下曲线下面积为1，随着阈值的减小精确度趋于0，可以通过曲线下面积对不同算法进行比较[12]。此外，P–R曲线的横坐标和纵坐标都和TP的数量密切相关，如果数据集阳性样本数量变化时，曲线的变化有可能会大于FROC曲线。

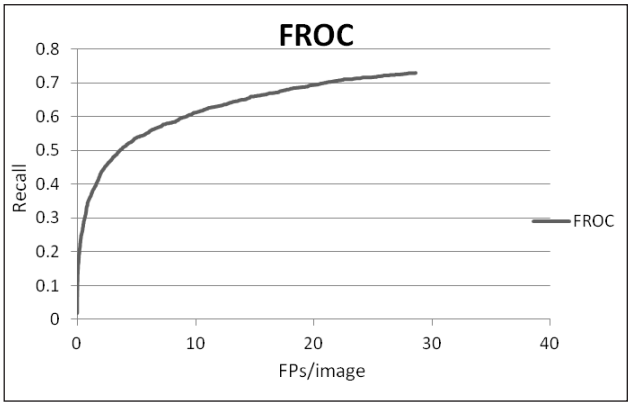


图5 FROC曲线

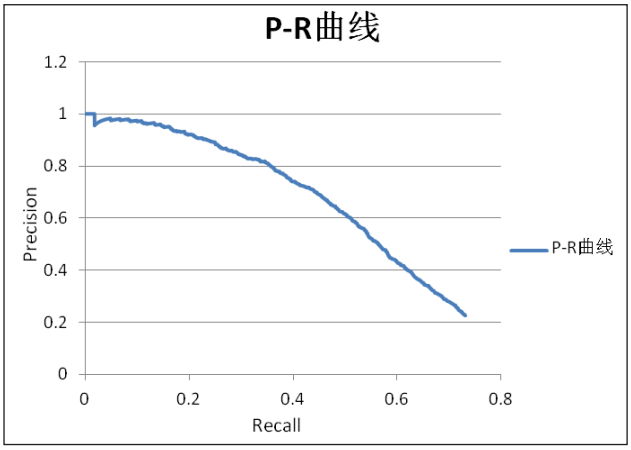


图6 P-R曲线

4 总结

统一的评价标准是人工智能算法质量评估重要的研究内容之一，这有助于实现AI算法的横向比较，使评价更为客观。本文讨论了不同评估参数的定义及适用场景，简述了它们各自的优缺点，但目前还没有形成统一的标准。相信随着人工智能在医疗领域的普及，以及临床实际应用经验与应用模式（如人+AI工作、AI单独工作）的不断进步，评价标准会逐步统一和提高。但现阶段，任何对于AI算法的评估，不仅要给出数据集的情况描述，还应给出全面的质量评估算法的描述，不能直接给出一个最终结果。

参考文献：

[1] Eric J Topol. High-performance Medicine: The Convergence of Human and Artificial Intelligence[J]. Nature Medicine, 2009, 25: 44-56.

[2] Setio A AA, Traverso A, De Bel T, et al. Validation, Comparison, and Combination of Algorithms for Automatic Detection of Pulmonary Nodules in Computed Tomography Images: The LUNA16 Challenge[J]. Medical Image Analysis, 2017, 42: 1-13.

[3] Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs[J]. JAMA, 2016, 316（22）: 2402 - 2410.

[4] Petrick N, Sahiner B, Armato SG, et al. Evaluation of Computer-aided Detection and Diagnosis Systems[J]. MedPhys, 2013, 40（8）: 87001.

[5] 中华医学会眼科学会眼底病学组. 我国糖尿病视网膜病变临床诊疗指南（2014年）[J]. 中华眼科杂志, 2014, 50（11）: 851-865.

[6] 孟祥峰, 王浩, 王权, 等. 影像类人工智能医疗器械评价方法研究[J]. 中国医疗设备, 2018, 33（12）: 23-26, 30.

[7] Jesse Davis, Mark Goadrich. The Relationship Between Precision-recall and ROC Curves[C].Appearing in Proceedings of the 23rd International Conference on Machine Learning: Pittsburgh, PA, 2006.

[8] FDA. Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions[S]. Washington DC: Food and Drug Administration, 2009.

[9] Chang H H, Zhuang A H, Valentino D J, et al. Performance Measure Characterization for Evaluating Neuroimage Segmentation Algorithms[J]. NeuroImage, 2009, 47（1）: 122-135.

[10] C á rdenes R, de LuisGarc í a R, Bachcuadra M. A Multidimensional Segmentation Evaluation for Medical Image Data[J]. Comput Methods Programs Biomed, 2009, 96（2）: 108-124.

[11] Kallergi M, Carney G M, Gaviria J. Evaluating the Performance of Detection Algorithms in Digital Mammography[J]. Medical Physics, 1999, 26（2）: 267.

[12] Sahiner B, Chen W, Pezeshk A, et al. Semi-parametric Estimation of the Area Under the Precision-recall Curve[C]. Spie Medical Imaging, 2016.

（收稿日期 2019年6月24日 编辑 王雅雯）