ICS 11.040.99
C30/49

GB

中 华 人 民 共 和 国 医 药 行 业 标 准

GB/TXXXX—××××

人工智能医疗器械 肺部影像辅助分析软件 算法性能测试方法

Artificial intelligence medical device—Computer assisted analysis software for pulmonary images—Algorithm performance test methods

工作组讨论稿

本稿完成日期：

××××-××-××发布 ××××-××-××实施

发 布

目　次

前　言

本文件按照GB/T 1.1-2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由国家药品监督管理局提出。

本文件由人工智能医疗器械标准化技术归口单位归口。

本文件起草单位：

本文件主要起草人：

引　言

人工智能算法在肺部影像辅助分析软件当中的应用较多，对产品的有效性与安全性影响较大。算法性能测试是产品质量评价的重要环节。本文件作为方法标准，面向辅助诊断、辅助检测等常见场景，对算法性能指标的定义、计算方式、测试过程进行规范，旨在加强相关产品的质量评价。

人工智能医疗器械 肺部影像辅助分析软件 算法性能测试方法

1 范围

本文件规定了对采用人工智能技术的肺部影像辅助分析软件的算法性能测试方法。本文件适用于采用人工智能技术、具有辅助诊断、辅助检测、辅助筛查、辅助分诊、优先级评定、随访跟踪等后处理功能的肺部影像辅助分析软件。本文件不适用于影像前处理及过程优化产品。

注：本文件为检测方法标准，不对任何功能做要求。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本部分必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本部分；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本部分。

YY/T 1833.1 人工智能医疗器械 质量要求和评价 第1部分：术语

YY/T 1833.2 人工智能医疗器械 质量要求和评价 第2部分：数据集通用要求

3 术语和定义

YY/T 1833.1、YY/T 1833.2界定的以及下列术语和定义适用于本文件。

3.1

通过准则 pass criteria

判断一个软件项或算法功能的测试是否通过的判别依据。

[来源：GB/T 9386-2008，3.2，有修改]

3.2

测试计划 test plan

描述预定测试活动的范围、方法、资源和进度的一种文档。它确定测试项、要测试的特征、测试任务、执行每一任务的人员以及需要应急对策的任何风险。

[来源：GB/T 9386-2008，3.13]

3.3

基线扫描 baseline scan

患者接受的首次影像扫描。

3.4

随访扫描 follow-up scan

患者在随访阶段接受的影像扫描。

3.5

重复筛查 repeat screening

以一定周期进行的多次筛查。

3.6

征象 sign

在进行身体检查或病理检查时，通常可由客观测度得到的、能够提供医疗进展及疾病状况的迹象及指标。

3.7

影像征象 signs in radiology

通过影像学手段获取的征象。

3.8

压力样本 stress sample

在某算法模型的标定范围内，特征容量极大或者极小的样本。

3.9

压力测试 stress test

使用压力样本开展测试的过程。

注：该定义区别于软件测试中的压力测试。

4 测试要求

4.1 通则

算法性能测试是肺部影像辅助分析软件验证与确认的重要环节，一般基于测试集对算法进行评估，对算法输出结果和参考标准进行定量比较，实现假阳性与假阴性、重复性与再现性、鲁棒性/健壮性、效率等具体指标的评估。

本文件描述了独立性能测试的方法，测试人员应建立完整的测试文档，包括测试计划、测试记录和测试结果。在测试开始前，测试人员应根据产品预期用途、临床使用场景和目标人群特征确定测试的通过准则，编写测试计划。在测试过程中，应形成测试记录，保证测试过程的可追溯。测试完成后，应对测试结果进行客观定量的描述，对试验结果与产品声称性能指标的符合性给出判定。

如测试过程需要复测，应限定复测次数的上限，例如不超过算法分类结果或检测目标的种类数量，以避免算法对参考标准进行推测或针对性调优。

4.2 测试环境

a) 宜在软件用户文档集中规定的最低运行环境下进行测试；如在最低环境之外还指定了典型运行环境，宜在该环境下进行必要的测试或理论分析。
b) 测试环境中的其他软件如影响待测产品的部署、运行和测试，测试时应进行控制。
c) 在产品临床应用环境下具备测试条件时，也可直接选择在临床应用环境下进行测试。
d) 如按要求部署测试环境后软件无法运行，或按要求部署测试环境后产品出现重大运行缺陷（如界面无法正常展示、频繁崩溃、内存泄漏等），应在结果中完整记录。
e) 测试环境应在结果中完整记录。

注：测试环境包括硬件环境和软件环境，硬件环境一般是指测试使用的服务器、客户端、网络连接设备、辅助硬件等设备所构成的环境；软件环境指被测软件运行时使用的操作系统、数据库、云平台、支持软件等构成的环境。

4.3 测试资源

4.3.1 测试集通用要求

测试集的质量应满足YY/T 1833.2。测试集应独立于算法训练、调优过程，保证封闭性和安全性。肺部影像辅助分析软件的制造商可根据产品预期用途和临床应用场景，对测试数据进行限定。

注：附录A给出测试集描述的样例。

4.3.2 测试集样本量

测试人员宜结合测试的置信度、算法主要指标的允差、阳性样本在测试集中的比例，计算单次测试的样本量要求。对预期用于分类的产品，可采用灵敏度计算单次测试中阳性样本的样本量，用特异度计算单次测试中阴性样本的样本量，计算公式见公式（1）：

$$N = \frac{Z_{1-\alpha/2}^{2} P(1-P)}{\Delta^{2}} \quad (1)$$

式中：

N——单次测试中阳性样本/阴性样本样本量；

$Z_{1-\alpha/2}$——标准正态分布的分位数；

α——显著性水平，常用取值为0.05；

P——灵敏度或特异度的预期值；

Δ——P的允许误差大小，一般取P的95%置信区间宽度的一半，常用的取值为0.05—0.10。

对预期用于检出的产品，可采用召回率计算单次测试中阳性样本的样本量。对其他预期用途的产品，制造商宜描述单次测试样本量选取的依据。

使用单次测试的阳性样本量除以阳性样本的比例（患病率），得到单次测试的样本总量。制造商宜提供患病率的数值和来源。

4.3.3 测试集配置

测试开始前，测试人员宜对测试集进行配置，考虑以下要求：

a) 测试集应考虑产品适用的临床使用场景在人群特征、疾病分布、数据质量要求、数据标注标准、数据采集设备与场所方面的统计学差异,确保数据容量与多样性。
b) 根据不同的测试目标，应组建不同的测试集和测试流程。
c) 应记录测试集的版本、标识、制造责任方、总体样本量、样本构成、使用日期、存储位置。
d) 测试人员宜根据测试集的数据层次，从设备、人群、地区、机构、数据质量、成像参数等方面抽取子测试集，开展分层测试，评估不同场景、不同配置下的算法性能。
e) 测试数据如包含同一病例在不同时间的数据，如基线扫描、随访扫描、重复筛查，宜应记录数据采集、数据标注的时间、地点、人员；如适用，对采集、标注过程的差异进行分析，对测试数据进行筛选。

4.3.4 扩增数据

在算法可靠性、鲁棒性测试中，可使用以黑盒或白盒方式扩增产生且具备参考标准的仿真数据进行附加的算法测试，研究产品性能的变化趋势，以及在极端条件下的表现。

数据扩增宜考虑以下要求：

a) 白盒扩增方式的内部环节是可理解的，如：旋转、分割、叠加噪声/伪影、叠加滤波、重建；

b） 黑盒扩增方式可忽略内部环节，集中响应输入和执行条件产生输出，如：生成对抗网络；
c） 如算法依赖的数据特征具有明确定义，可针对该特征进行针对性的扩增；
d） 测试计划应描述数据扩增的原理、方法、依据，对扩增的仿真数据与真实世界数据的异同进行比较论证，必要时进行抽样标注和验证；
e） 扩增数据集的配置宜符合 4.3.3 的要求。在标识与版本控制方面，扩增数据应与真实数据严格区分，使用记录可追溯。

4.3.5 体模与标准器

如适用，算法测试使用的体模与标准器应具备标识信息，处于计量/校准有效状态；加工精度应高于算法声称的测量精度、参考标准的精度。如适用，测试人员应在测试记录中写入体模与标准器的使用情况。

4.4 测试平台

如通过测试平台开展测试活动，测试平台宜符合如下要求：
a） 数据抽取：测试平台可按照指定条件，对测试平台可访问的测试数据进行抽取，用于组建测试集。指定的条件包括样本量、阳性样本比例、元数据字段信息、参考标准信息等。
b） 测试集管理：测试平台可记录测试集的使用与版本信息，以及数据抽取条件。
c） 可视化工具：测试平台可对算法输出结果、测试集的参考标准进行可视化的预览和比较；
d） 测试指标计算：测试平台可计算和输出算法性能指标，如检出、分类、分割等情形；
e） 网络安全：测试平台应确保测试数据、待测产品的安全性；
f） 如果测试需要在网络条件下进行，网速、传输服务质量（QoS）应不低于制造商声称的运行环境;
g） 过程记录：平台应为测试活动提供记录，包括测试人员活动记录、数据操作、待测算法运行状态、测试进度、测试结果处理等。

4.5 测试指标与通过准则

测试人员应根据产品技术特性、预期用途和使用场景，在测试计划中列出客观、定量的测试指标。制造商应给出各指标的标称值及其允差或上下限。通过准则包括单项指标和产品整体质量，测试所选取的各项指标应在测试计划中进行描述。如适用，应从病灶、部位、病例、测试集子集和测试集总体等层次开展统计分析，判断各单项指标是否通过。对于产品整体质量，测试人员应根据产品预期用途和风险分析，确定适用的整体评估指标，作为产品整体质量的判定依据。测试人员应确定各项单项指标和整体指标的通过阈值，即各项指标的预期值。

注：附录B给出测试指标及统计分析的一般思路。

4.6 测试流程要求

测试人员应根据测试计划开展测试活动，形成测试记录。

测试流程各步骤的要求如下：

a） 测试前

制造商宜提供接口，确保待测产品批量读取测试集中的数据。制造商宜提供医学影像的可视化工具，帮助测试人员预览待测产品输出的结果。待测产品输出结果的数据结构、格式应与测试集的参考标准兼容。输出结果应与输入数据唯一对应，包含测试需要的完整信息，如测试样本的编号、唯一标识、目标区域所在图像的编号、目标区域的位置、分类、边界端点坐标、算法预测的概率等。测试人员宜选用小批量数据进行预测试，避免系统偏差，评估参考标准与输出结果的可比性，包括但不限于空间位置、时序、分类、尺寸、有效数字等。上述信息宜写入测试记录。

b） 测试过程中

测试人员宜记录数据元、病例层面的测试进度，如数据读取进度、算法运行时间、运行结果，以及算法运行过程中的错误、警告、异常提示，写入测试记录。

c） 测试后

测试人员应量化比对算法输出结果与参考标准（来自测试集、体模、仿真数据、扩增数据等），汇总各指标的测试结果。

4.7 测试结果

测试人员对测试结果进行客观、定量的描述，内容至少应包含：

a） 测试环境；

b） 测试平台描述（如适用）；

c） 测试集描述；

d） 算法性能指标的符合性分析，包含性能指标的定义、测试通过准则、统计分析；

e） 算法错误统计。

注：算法错误指算法的判断、预测结果与参考标准不一致的情形。

5 算法性能测试方法

5.1 算法应用场景的测试方法

5.1.1 目标检测场景

5.1.1.1 标记与匹配

对具有目标检测功能的产品，导出算法标记目标；测试集的参考标准应包含对应的目标。测试人员应在测试计划中记录算法标记目标与参考标准目标的匹配方式和匹配阈值，匹配方式和匹配阈值由制造商声称。

注：匹配阈值是判定算法标记目标与参考标准目标匹配关系的依据，含义有别于算法的检出概率。

常见标记匹配方式举例如下：

a） 区域重叠：通过计算算法标记目标与参考标准区域重叠的程度（如 Dice 系数、Jaccard 系数）并设定匹配阈值来确定匹配结果。

b） 中心点距离：通过计算算法标记目标中心与参考标准区域中心的距离并设定匹配阈值来确定匹配结果。

c） 中心点落入：通过判断算法标记的目标区域目标中心是否落入参考标准目标内来确定匹配结果。

注：中心点的选择与目标区域、影像征象有关。以肺结节（一般为凸形状）为例，中心点为检出区域范围内长径与短径的交点。长径定义为检出区域内最大横截面空间最远两点距离。短径定义为结节内垂直于长径的最长距离。

匹配结果分为三种情形：

a）真阳性，即匹配参考标准的算法标记目标，总数记为 *TP*；

b）假阳性，即未匹配参考标准的算法标记目标，总数记为 *FP*；

c）假阴性，即未匹配算法标记目标的参考标准目标，总数记为 *FN*。

特殊情况处理：当出现多对一匹配时，匹配关系宜遵从以下优先级考虑：

a） 如采用区域重叠方式，取区域重叠的程度更大的；

b） 如采用中心点距离方式，取中心点距离更小的；

c） 如采用中心落入方式，取中心点距离更小的。

5.1.1.2 测试步骤

在标记与匹配场景下，算法测试按如下步骤进行：

a) 向待测算法输入测试集，输出算法检出的结果；该结果的格式宜与参考标准兼容，内容至少包含目标的位置、检出概率；

b) 输出结果按 5.1.1.1 规定的原则与参考标准进行匹配，得到 *TP*、*FP* 和 *FN*；

c) 按照 5.1.1.3-5.1.1.8 描述的公式计算各性能指标。可以病灶为单元，直接以整个集合的计算结果作为最终结果；如果一个病例含多个病灶，也可以病例为单元，计算每个病例内部病灶集合的计算结果，然后对病例集合的计算结果进行平均。

5.1.1.3 召回率——病变定位率

召回率的计算公式见公式（2）：

$$Rec = \frac{TP}{TP + FN} \times 100\% \quad \text{(2)}$$

式中：

Rec——召回率；

TP——真阳性个数；

FN——假阴性个数。

5.1.1.4 精确度

精确度的计算公式见公式（3）：

$$Pre = \frac{TP}{TP + FP} \times 100\% \quad \text{(3)}$$

式中：

Pre——精确度；

TP——真阳性个数；

FP——假阳性个数。

5.1.1.5 F_1度量

F_1度量的计算公式见公式（4）：

$$F_1 = \frac{2 \times Pre \times Rec}{Pre + Rec} \quad \text{(4)}$$

式中：

F_1——F_1度量；

Pre——精确度；

Rec——召回率。

5.1.1.6 平均精确度

改变算法阈值设置，计算各个阈值对应的精确度、召回率。以召回率为横坐标、精确度为纵坐标，生成精确度-召回率曲线，计算曲线下的积分面积，即平均精确度。测试记录应描述曲线是否进行平滑处理，以及所采用的方法。

5.1.1.7 平均精确度均值

如适用多目标检测，应求出各类目标的平均精确度，其平均值即为平均精确度均值。

5.1.1.8 自由响应受试者操作特征曲线

改变算法阈值设置，计算各个阈值对应的召回率和非病变定位率。以召回率为纵坐标，非病变定位率为横坐标，构造的曲线为自由响应受试者操作特征曲线。横坐标的取值一般设为等比数列，即0.5、1、2、…、n，其中横坐标上限n应大于病例个体的平均病灶数量。以肺结节辅助检测为例，假设单个病例平均具有7.5个肺结节，则n取值不低于8。

其中非病变定位率用NLR表示，表达式见公式（5）：

$$NLR = \frac{NLL}{N} \times 100\% \quad \text{(5)}$$

式中：

NLR——非病变定位率；

NLL——算法检出病变位置未能正确识别出参考标准确定的病变位置的数量；

N——全体病例的数量。

注：非病变定位率也可称为单病例平均假阳个数。

5.1.2 区域分割与测量场景

5.1.2.1 测试步骤

在区域分割与测量场景下，算法测试按如下步骤进行：

a) 向待测算法输入测试集，输出算法分割的结果；该结果的格式宜与参考标准兼容，内容至少包含分割区域边界端点坐标；

b) 算法分割的目标区域与参考标准分割的目标区域的性能指标按 5.1.2.2-5.1.2.9 描述的公式进行计算。

c) 以计算结果的平均值作为最终结果。

d) 如算法合并检测功能，仅对检出结果为 *TP* 的目标进行计算。

5.1.2.2 召回率

算法分割的目标区域与参考标准分割的目标区域的交集除以参考标准分割的目标区域，用公式（6）表示：

$$Rec = \frac{S_{pr} \cap S_{gt}}{S_{gt}} \quad \text{(6)}$$

式中：

Rec——召回率；

S_{pr}——算法分割的目标区域；

S_{gt}——参考标准分割的目标区域。

5.1.2.3 精确度

算法分割的目标区域与参考标准分割的目标区域的交集除以算法分割的目标区域，用公式（7）表示：

$$Pre = \frac{S_{pr} \cap S_{gt}}{S_{pr}} \quad (7)$$

式中：

Pre——精确度；

S_{pr}——算法分割的目标区域；

S_{gt}——参考标准分割的目标区域。

5.1.2.4 目标区域重合度

当目标区域为一般实体时（如肺结节），宜采用Dice系数或Jaccard系数评价目标区域重合度。

Dice系数为算法分割的目标区域与参考标准分割的目标区域交集的两倍除以两者之和（召回率与精确度的调和平均数），用公式（8）表示：

$$Dice = \frac{2 \times |S_{pr}| \cap |S_{gt}|}{|S_{pr}| + |S_{gt}|} \quad (8)$$

式中：

$Dice$——Dice系数；

S_{pr}——算法分割的目标区域；

S_{gt}——参考标准分割的目标区域。

Jaccard系数为算法分割的目标区域与参考标准分割的目标区域交集除以两者的并集，用公式（9）表示：

$$Jaccard = \frac{|S_{pr} \cap S_{gt}|}{|S_{pr} \cup S_{gt}|} \quad (9)$$

式中：

$Jaccard$——Jaccard系数；

S_{pr}——算法分割的目标区域；

S_{gt}——参考标准分割的目标区域。

5.1.2.5 树长度检测

当目标区域为气管或其他树形结构时，宜采用树长度检测（tree length detection，TLD）评估计算正确分割的气管长度与参考标准气管长度的比例，用公式（10）表示：

$$TLD = \frac{L_{TP}}{L_{gt}} \quad (10)$$

式中：

TLD——树长度检测；

L_{TP}——正确分割的气管长度（以像素计数）；

L_{gt}——参考标准气管的长度（以像素计数）。

5.1.2.6 表面距离

表面距离为算法和参考标准给出的目标区域之间的距离，可用于评价轮廓分割的效果。

根据公式（11）计算双向豪斯多夫距离$d_H(X,Y)$：

$$d_H(X,Y) = \max\left\{\max_{x\in X}\min_{y\in Y} d(x,y), \max_{y\in Y}\min_{x\in X} d(x,y)\right\} \quad \text{(11)}$$

式中：

$d_H(X,Y)$——双向豪斯多夫距离

X——算法给出的目标区域；

Y——参考标准给出的目标区域；

$d(x,y)$——X、Y两个区域任意两点之间的距离。

5.1.2.7 密度测量

算法识别的目标区域内像素的密度值或灰度值，与参考标准目标区域的结果进行比较，以计算相对误差绝对值的平均值。根据公式（12）计算密度测量的偏差值：

$$S = \frac{\sum_{i=1}^{n}\left|\frac{L_i - L_{a,i}}{L_{a,i}}\right|}{n} \times 100\% \quad \text{(12)}$$

式中：

S——偏差值；

L_i——第i个病例的测量值；

$L_{a,i}$——第i个病例的参考标准值；

n——病例个数。

5.1.2.8 尺寸测量

尺寸测量的对象是目标区域的尺寸，如长短径、紧密包裹矩形框的长宽等。目标区域可以位于二维平面或三维立体空间内。

当算法识别的目标区域可近似为凸形状时，可对算法识别的目标区域轮廓（含边界）使用旋转卡壳法（rotating caliper）或其他方法，定位具有医学意义的关键点，计算长径、短径和平均值，与参考标准的目标区域结果进行比较，计算相对误差绝对值的平均。

5.1.2.9 体积测量

分别统计算法识别的目标区域和参考标准的目标区域内的体素数量，乘以每个体素的体积，可计算体积测量的绝对误差；也可根据体素数量，计算相对误差绝对值的平均值。

5.1.3 影像分类场景

5.1.3.1 测试步骤

在影像分类场景下，算法测试执行以下步骤：

a) 向待测算法输入测试集，输出算法分类的结果；该结果的格式宜与参考标准兼容，内容包含分类结果、分类概率（如适用）；

b) 比较算法分类与参考标准分类，计算真阳性、假阳性、真阴性、假阴性样本的数量，构造混淆矩阵。

对于分类问题，混淆矩阵的一般形式如表1所示：

表1 n 分类混淆矩阵

分类	Pred_1	Pred_2	…	Pred_n
True_1	$N_{1,1}$	$N_{1,2}$	…	$N_{1,n}$
True_2	$N_{2,1}$	$N_{2,2}$	…	$N_{2,n}$
…	…	…	…	…
True_n	$N_{n,1}$	$N_{n,2}$	…	$N_{n,n}$
注：Pred_x（x=1～n）为人工智能判断为 x 类的类别；True _x（x=1～n）为参考标准判断为 x 类的类别；$N_{i,j}$（i=1～n，j=1～n）为参考标准的判断结果为 i 类，被人工智能判断为 j 类的个数；n 为分类类型个数。				

二分类的混淆矩阵可简化为表2所示：

表2 二分类混淆矩阵

分类		人工智能分类	
		阳性	阴性
参考标准分类	阳性	*TP*	*FN*
	阴性	*FP*	*TN*

多分类问题可转化为多个二分类问题，参考标准分类为第 i 类与其他类别的混淆矩阵简化形式如表3所示：

表3 多分类实际可转化为二分类混淆矩阵

分类		人工智能分类	
		阳性	阴性
参考标准分类	阳性	$TP=\sum_{i=1}^{n} N_{i,i}$	$FN=\sum_{j=1,j\neq i}^{n} N_{i,j}$
	阴性	$FP=\sum_{j=1,j\neq i}^{n} N_{j,i}$	$TN=\sum_{j=1,j\neq i}^{n}\sum_{l=1,l\neq i}^{n} N_{j,l}$

5.1.3.2 灵敏度

灵敏度用*Sen*表示，表达式见式（13）：

$$Sen = \frac{TP}{TP + FN} \times 100\% \quad \cdots\cdots (13)$$

式中：

Sen——灵敏度；

TP——真阳性样本的个数；

FN——假阴性样本的个数。

5.1.3.3 特异度

特异度用*Spe*表示，表达式见式（14）：

$$Spe = \frac{TN}{FP + TN} \times 100\% \quad \cdots\cdots (14)$$

式中：

TN——真阴性样本的个数；

FP——假阳性样本的个数。

5.1.3.4 漏检率

漏检率用*MR*表示，表达式见式（15）：

$$MR = 1 - Sen \quad \cdots\cdots (15)$$

式中：

Sen——灵敏度；

MR——漏检率。

5.1.3.5 阳性预测值

阳性预测值用*PPV*表示，表达式见式（16）：

$$PPV = \frac{TP}{TP + FP} \quad \cdots\cdots (16)$$

式中：

PPV——阳性预测值；

TP——真阳性样本的个数；

FP——假阳性样本的个数。

5.1.3.6 阴性预测值

阴性预测值用*NPV*表示，表达式见式（17）：

$$NPV = \frac{TN}{FN + TN} \quad \cdots\cdots (17)$$

式中：

NPV——阴性预测值；

TN——真阴性样本的个数；

FN——假阴性样本的个数。

5.1.3.7 准确率

准确率用*Acc*表示，表达式见式（18）：

$$Acc = \frac{\sum_{i=1}^{n} N_{i,i}}{\sum_{j=1}^{n}\sum_{i=1}^{n} N_{j,i}} \quad \cdots\cdots (18)$$

式中：

$N_{j,i}$——泛指混淆矩阵第j行、第i列的元素；

$N_{i,i}$——泛指混淆矩阵第i行、第i列的元素。

5.1.3.8 约登指数

约登指数用*Y*表示，表达式见式（19）：

$$Y = Sen + Spe - 1 \quad \cdots\cdots (19)$$

式中：

Sen——灵敏度；

Spe——特异度。

5.1.3.9 Kappa 系数

Kappa系数用*K*表示，表达式见公式（20）：

$$K = \frac{Acc - p_e}{1 - p_e} \quad \cdots\cdots (20)$$

其中：

$$p_e = \frac{\sum_{i=1}^{n}(\sum_{j}^{n} N_{i,j} \times \sum_{j}^{n} N_{j,i})}{(\sum_{a=1}^{n}\sum_{b=1}^{n} N_{a,b})^2} \quad \cdots\cdots (21)$$

式中：

$N_{j,i}$——泛指混淆矩阵第j行、第i列的元素；

$N_{i,i}$——泛指混淆矩阵第i行、第i列的元素；

Acc——准确率。

5.1.3.10 受试者操作特征曲线

制造商应给出标称的受试者操作特征曲线（receiver operating characteristics curve，ROC）的曲线下面积（area under curve，AUC）值。测试人员宜调节不同分类阈值（不宜少于1000步，可均匀设置步长），比较算法分类结果与参考标准分类结果，计算各个阈值下的灵敏度与特异度，以1减特异度为横坐标、以灵敏度为纵坐标绘制ROC曲线，并计算ROC曲线下的积分面积。

5.1.4 多功能组合场景

对于同时具有检出、分类、分割、测量等功能的产品，测试人员宜对上述功能对应的算法性能进行分步的评价，如：

——首先对目标检测场景进行评价，计算检出的指标；

——其次对标记匹配正确的目标区域，计算分类、分割的指标；

——最后计算测量相关指标。

当产品算法功能有使用限定条件或技术约束条件时（如算法仅识别大于某尺寸的病灶、算法仅适用于CT层厚2mm及以下的图像等），测试人员应对测试集、参考标准进行对应的约束，并在测试计划和测试记录中注明。

5.1.5 随访评估场景

对具有随访评估功能的产品，宜输入同一病例的基线扫描、随访扫描、重复筛查等不同时间节点的数据，比较算法对同一目标区域的匹配结果；分析结果与参考标准之间的符合性；同时，根据各时间节点的结果，可建立动态曲线，计算与参考标准曲线之间的一致性。

5.1.6 患者分诊场景

对具有患者分诊功能的产品，测试集应依据临床诊疗标准或专家共识对测试数据建立分级标签，比如阴阳性分诊或危重分诊，与算法输出的标签进行对比，建立混淆矩阵，采用5.1.3.2的方法计算灵敏度、特异度、Kappa系数等指标。

对具有患者优先级排序的产品，参照执行本条的方法。

5.2 算法质量特性与测试方法

5.2.1 泛化能力通用要求

制造商应根据产品预期用途和部署环境，对产品研发使用的训练集与真实世界陌生样本之间的差异进行分析，形成文档，作为配置测试集的依据。实际测试中，宜通过测试集的多样性与变化性，对算法的泛化能力进行验证。

5.2.2 鲁棒性

5.2.2.1 通用要求

制造商应根据产品风险分析、使用限制和临床部署环境特征，评估临床使用阶段各种可能干扰算法性能的因素，收集真实世界数据或产生仿真数据，组成专用测试集，对算法性能依据条款5.1进行扩展测试，分析各指标的变化情况，形成鲁棒性研究资料。

5.2.2.2 面向硬件变化的对抗测试

测试人员应考虑医学成像硬件设备、参数设置的多样性，收集或模拟生成更多的图像数据，作为对测试集的扩充，验证算法面对影像采集硬件设备的鲁棒性。参数设置的多样性包括：物理分辨率、像素分辨率、亮度、调焦、射线质量等。模拟生成的图像数据不应影响标注结论。

5.2.2.3 面向软件前处理的对抗测试

测试人员宜考虑软件前处理的多样性，收集或模拟生成更多的图像数据，作为测试集的扩充，验证算法面对软件前处理的鲁棒性。软件前处理的多样性包括：背景裁切、图像压缩、背景填充、平滑预处理、重建算子等。模拟生成的图像数据不应影响标注结论。

5.2.2.4 面向欺骗攻击的对抗测试

如适用，测试人员宜根据制造商提供的文档，确定欺骗攻击的类型和试验参数配置。测试人员可使用白盒或黑盒方式产生肉眼难以觉察的欺骗性扰动，然后用模型对这些添加扰动后的图像进行测试，从而验证模型是否能抵御欺骗攻击。施加扰动后的数据应通过标注人员的确认后用于测试。

注：本条款提到的“白盒”与条款4.3.4中的含义不同，描述攻击手段是否基于模型内部架构、参数等信息。

5.2.2.5 压力测试

5.2.2.5.1 通用要求

测试人员宜从测试集中选取压力样本，依据条款5.1开展压力测试。压力样本不应影响医生判断。

5.2.2.5.2 压力样本的选取

压力样本的选取可考虑，但不限于以下特征：

a) 受试者年龄偏大的影像；
b) 特定疾病的影像；
c) 有伪影但满足数据质量要求的影像；
d) 影像的层厚极大或者极小；
e) 影像序列包含的图像数量极大；
f) 有植入物（干扰项）的；
g) 有并发症的；
h) 多发、弥散性病变。

5.2.3 重复性

如适用，测试人员应对同一版本的算法使用相同的样本依据条款5.1进行多次测试，测试次数不宜低于三次，观察测试结果是否变化。

5.2.4 一致性

如适用，测试人员应对算法输出的中间结论与产品输出的最终结论之间的一致性进行评估。

如中间结论具备参考标准，应使用参考标准对中间结论进行验证。对于预期用于目标检测的模型，可参照5.1.1的方法衡量输出结果与参考标准标记的匹配关系；对于预期用于分类的模型，可参照5.1.3的方法建立混淆矩阵，计算Kappa系数。

5.2.5 效率

测试人员应评估临床典型病例的处理时间，宜以数据开始导入的时刻作为起点，以算法导出全部结果的时刻作为终点。临床典型病例需约定图像数量、图像特征、成像参数、图像格式、成像方式等要素，如包含300张图像、分辨率为512×512、层厚为1mm、格式为Dicom3.0的CT平扫病例。

辅助分诊类、优先级排序类产品应以生成算法通知作为终点。

5.2.6 算法错误统计

测试人员应对算法的错误进行统计分析，如下列情形：

a) 在标记-匹配场景下，测试人员宜对假阴性结果进行分组统计，考虑未达到匹配阈值（部分重叠）、完全未匹配（零重叠）两种情形在假阴性样本中的比例；
b) 在多分类场景下，测试人员宜对每一种分类的假阴性、假阳性结果进行分组统计；
c) 在分割场景下，测试人员宜根据目标区域的尺寸，对分割结果进行分组统计；
d) 测试人员宜根据每个病例的算法性能指标，评估算法对个体的偏倚；

e） 测试人员宜在对抗测试、压力测试中采用 a）-d）的方法开展错误统计。

附 录 A
（资料性）
胸部 CT 肺结节测试集描述样例

A.1 概述

本样例以胸部CT肺结节测试集为具体案例，给出测试集描述的举例，仅作为参考信息。

A.2 数据集适用范围

数据集适用于声称能对胸部 CT 肺结节进行辅助分析的人工智能医疗器械软件产品，如肺结节辅助检出、分类、分割、测量等。

A.3 数据采集

数据采集需考虑患者人群、采集场所、采集设备、数据格式、采集人员等方面的多样性，具有合规性证明，如伦理审批。表 A.1 给出了数据来源的多样性统计举例，可进一步细化。

表A.1 数据来源的多样性统计

统计维度	范围	病例数量
年龄	40-60岁	xx例
	61-80岁	xx例
	大于80岁	xx例
性别	男	xx例
	女	xx例
地域	华东地区	xx例
	华南地区	xx例
	华中地区	xx例
	华北地区	xx例
	西北地区	xx例
	西南地区	xx例
	东北地区	xx例
CT机型	XX公司XX型号	xx例
	XX公司XX型号	xx例
扫描方式	平扫	xx例
	增强	xx例
	低剂量	xx例
管电流	小于50mA·s	xx例
	大于50mA·s	xx例
管电压	小于110kV	xx例
	大于110kV	xx例
层厚	小于1.5mm	xx例
	大于1.5mm	xx例
采集场所	体检	xx例
	门诊	xx例

	住院	xx例

A.4 数据集的分布构成

肺结节的分布构成描述模板如表A.2所示，其中尺寸的范围划分可根据具体临床指南调整：

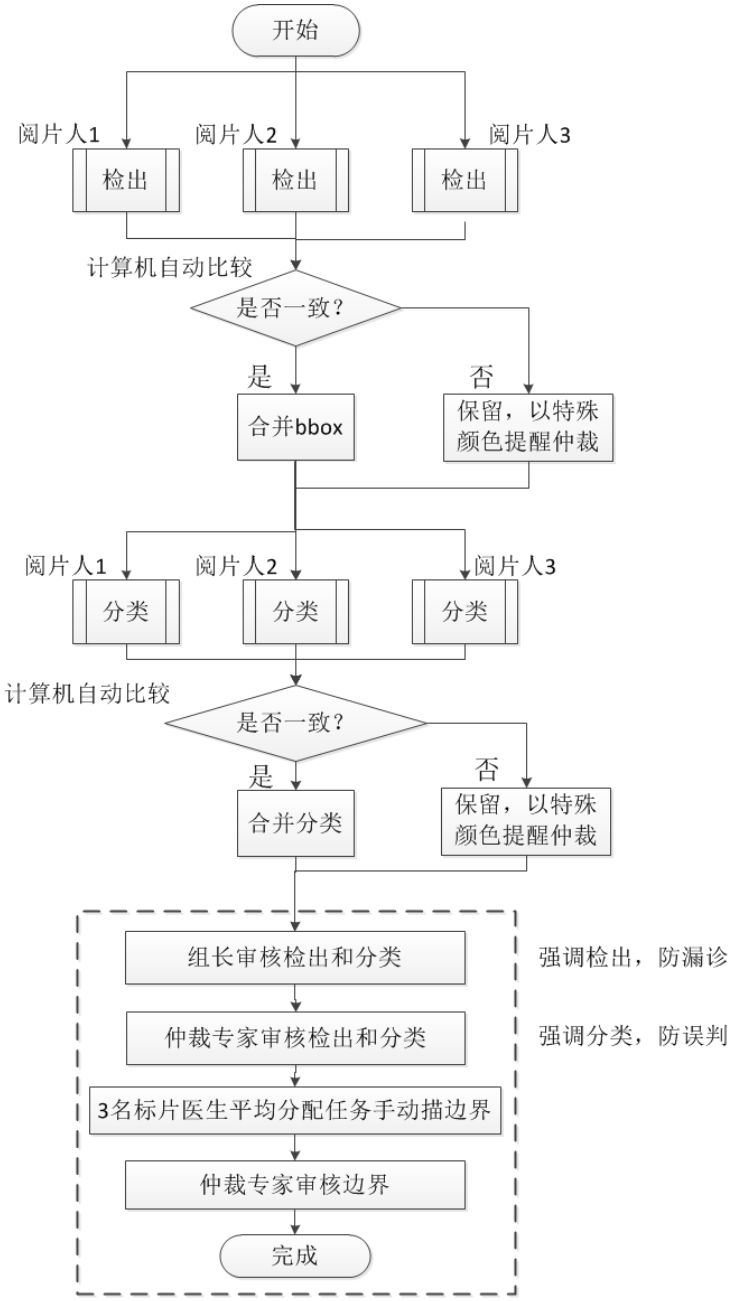
表 A.2 肺结节分布统计

肺结节种类	尺寸（mm）	病例
肺内实性结节	小于 4	xx 例
	4-8	xx 例
	大于 8	xx 例
肺内部分实性结节	小于 4	xx 例
	4-8	xx 例
	大于 8	xx 例
肺内纯磨玻璃结节	小于 4	xx 例
	4-8	xx 例
	大于 8	xx 例
肺内钙化结节	/	xx 例
胸膜实性结节	/	xx 例
胸膜钙化结节	/	xx 例
其他疾病	/	xx 例
总计	/	xx 例

A.5 数据集标注规则

标注工作主要依据文献[8]进行。

标注流程：为提高标注的准确性和敏感度，降低假阳性率，避免记忆偏倚，标注流程建议多轮次分组交叉进行，优化人力资源，主要包含肺结节的检出、分类、边界分割和尺寸测量（图A.1所示）。考虑不同环节的工作量和人员资质的差异，标注工作需要标注医师、标注组长和仲裁专家3种级别的医师参加。标注组长由工作经验10年以上的副主任医师担任，仲裁专家由工作经验15年以上的副主任医师或主任医师担任。每一批标注任务由标注组长带领两名标注医师承担，分为4个主要环节。



图A.1肺结节标注流程图

a) 检出环节：3 名标注医师背靠背独立标注，然后用计算机自动判断检出的一致性，以所有人标注结果的并集作为结果。

b) 分类环节：3 名标注医师背靠背进行分类，分类结果同样由计算机自动判断一致性和进行合并，同时保留不同意见。

c) 审核环节：由其他标注组长和仲裁专家各自独立对检出和分类结果进行审核与修改，纠正漏诊、误诊和误判。如果遇到疑难问题，仲裁专家可以进行集体讨论与确认。本环节过后，每个病例至少由 5 名医师进行过阅片，其中至少由两名具有高级职称的医生进行过审核。

d) 边界分割与尺寸测量：在检出与分类完成之后，由于边界分割相对简单，建议普通病例的边界分割由 1 名标注医师执行，由 1 名审核专家进行审核。遇到复杂征象时，可酌情增加审核人数，以保证标注质量。结节的尺寸根据手工边界由计算机自动生成，标注医师和仲裁专家可以手动修改。

A.6 样本量的估计

为保证灵敏度的抽样误差不大于允差，总体样本量应不低于公式（A.1）的计算结果：

$$N_1 = \frac{Z_{1-\alpha/2}^2 P_{sen}\left(1 - P_{sen}\right)}{d^2 \times P_{re}} \quad \text{(A.1)}$$

式中：

N_1——总体样本量；

$Z_{1-\alpha/2}$——标准正态分布的分位数；

α——显著性水平，常用取值为0.05；

P_{sen}——估计的灵敏度；

d——灵敏度的允差；

P_{re}——测试集中的患病率。

为保证特异度的抽样误差不大于允差，总体样本量应不低于公式（A.2）的计算结果：

$$N_2 = \frac{Z_{1-\alpha/2}^2 P_{spe}\left(1 - P_{spe}\right)}{d^2 \times \left(1 - P_{re}\right)} \quad \text{(A.2)}$$

式中：

N_2——总体样本量；

$Z_{1-\alpha/2}$——标准正态分布的分位数；

α——显著性水平，常用取值为0.05；

P_{spe}——估计的特异度；

P_{re}——测试集中的患病率；

d——特异度的允差。

基于上述考虑，单次测试的总体样本量不低于N_1、N_2的最大值。

A.7 测试集偏倚分析

测试数据集选择可能是诊断性能评价中偏差的一个主要来源，这是许多诊断测试共同具有的风险。当选择的病例不能代表目标人群时，测试集偏倚的概念被引入。来自病例选择的一些偏倚不可避免，但应给出可能造成偏倚的风险分析，以便使用者认识潜在偏差对于测试结果的影响。依据YY/T 1833.2，对测试集的选择偏倚、覆盖偏倚、验证偏倚等风险进行分析，具体内容见数据集风险分析文档。

附 录 B
（资料性）
测试指标及统计分析的一般思路

B.1 概述

测试指标的选取和统计分析的思路，对确定测试通过准则具有重要的影响。本附录对测试指标及统计分析的一般思路进行补充说明，作为参考信息。在制定测试计划时，测试人员宜明确统计检验的类型、检验假设、判定界值等。判定界值的确定应有依据，需提供相应的置信区间结果，置信区间通常取 95%。

下面针对几种常见的测试情形介绍相关的单项和总体指标，并介绍对应的统计检验方法。鉴于统计理论模型的多样性，本附录仅属于推荐性内容。

B.2 情形一：测试结果为二分类变量

若测试的目的为对影像的阴性或阳性进行判断，测试结果为二分类，例如根据肺部影像对该受试者是否患有某种疾病进行判断。此时建议构造混淆矩阵（如表 B.1 所示），采取灵敏度（Sensitivity）和特异度（Specificity）作为测试指标。灵敏度表示的是当测试影像为阳性样本时，产品正确将该影像判别为阳性的概率；特异度表示的是测试影像为阴性时，产品正确将该影像判别为阴性的概率。

表 B.1 二分类混淆矩阵

分类		算法分类	
		阳性	阴性
参考标准分类	阳性	$N_{1,1}$	$N_{1,2}$
	阴性	$N_{2,1}$	$N_{2,2}$

灵敏度的估计值为：

$$Sen = \frac{N_{1,1}}{N_{1,1} + N_{1,2}} \tag{B.1}$$

式中：

Sen——灵敏度的估计值；

$N_{1,1}$——真阳性样本数量；

$N_{1,2}$——假阴性样本数量；

特异度的估计值为：

$$Spe = \frac{N_{2,2}}{N_{2,1} + N_{2,2}} \tag{B.2}$$

式中：

Spe——特异度的估计值；

$N_{2,1}$——假阳性样本数量；

$N_{2,2}$——真阴性样本数量。

特别地，测试者应明确灵敏度或特异度是以病例为单位的或者是以病灶为单位的。

关于灵敏度的$\left(1-\alpha\right)\times 100\%$置信区间的表达式为

$$\left(Sen - {}^{z_{1-\frac{\alpha}{2}}}\sqrt{\frac{Sen\left(1-Sen\right)}{N_{1,1}+N_{1,2}}}, Sen + {}^{z_{1-\frac{\alpha}{2}}}\sqrt{\frac{Sen\left(1-Sen\right)}{N_{1,1}+N_{1,2}}}\right)$$

式中：

$Z_{1-\frac{\alpha}{2}}$——正态分布的$1-\frac{\alpha}{2}$分位数，通常建议取$\alpha = 0.05$；

Sen——灵敏度的估计值。

关于特异度的$\left(1-\alpha\right)\times 100\%$置信区间的表达式为：

$$\left(Spe - {}^{z_{1-\frac{\alpha}{2}}}\sqrt{\frac{Spe\left(1-Spe\right)}{N_{2,1}+N_{2,2}}}, Spe + {}^{z_{1-\frac{\alpha}{2}}}\sqrt{\frac{Spe\left(1-Spe\right)}{N_{2,1}+N_{2,2}}}\right)$$

式中：

$Z_{1-\frac{\alpha}{2}}$——正态分布的$1-\frac{\alpha}{2}$分位数，通常建议取$\alpha = 0.05$；

Spe——特异度的估计值。

其他构建置信区间的方法可以参见参考文献[7]中的第四章。此外，阳性预测值和阴性预测值也可以作为备选的指标，其选择条件以及相应的分析方法可以参见参考文献[7]的第二章和第四章。

B.3 情形二：测试结果为有序型或连续型变量

B.3.1 有序型变量

当诊断试验的结果是有序变量，例如依据报告和数据系统（reporting and data system，简称 RADS）进行分级时，建议绘制 ROC 曲线，并以 ROC 曲线下面积（area under the curve, AUC）作为统计评价指标。诊断试验结果的数据结构如表 B.2 所示。利用所有有序评分尺度绘制 K 个散点，即可得到经验 ROC 曲线。

表 B. 2 诊断试验有序数据结构（斜体问题）

疾病状态（*D*）	试验结果（*T*）				
	1	2	…	K	合计
D = 1	S_1	S_2	…	S_K	N_1
D = 0	R_1	R_2	…	R_K	N_0
合计	M_1	M_2	…	M_K	N
注：*T* 表示诊断试验结果，包括 *K* 类有序结果；*D* 表示个体的真实疾病状态, *D* =1 和 0 分别表示患病和未患病；*N* 表示个体总数；N_1 表示患病个体总数；N_0 表示未患病个体总数；S_i 表示患病个体被诊断为第 *i* 类有序结果的数量；R_i 表示未患病个体被诊断为第 *i* 类有序结果的数量。					

对于经验 ROC 曲线上的第 *m* 个散点（取值范围为 1～*K*），

其横坐标表示为：

$$1 - Spe\left(m\right) = \frac{1}{N_0}\sum_{j=m}^{K} R_j \tag{B.3}$$

式中：

$Spe\left(m\right)$——对第 *m* 个散点的特异度的估计值；

R_j——未患病个体被诊断为第 *j* 类有序结果的数量。

其纵坐标表示为：

$$Sen\left(m\right) = \frac{1}{N_1}\sum_{j=m}^{K} S_j \tag{B.4}$$

式中：

$Sen\left(m\right)$——对第 *m* 个散点的灵敏度的估计值；

S_j——患病个体被诊断为第 *j* 类有序结果的数量。

ROC 曲线下面积可由非参数方法估计，表示为：

$$AUC = \frac{1}{N_0 N_1}\sum_{i=1}^{N_1}\sum_{j=1}^{N_0}\varphi\left(T_{1i}, T_{0j}\right) \tag{B.5}$$

式中：

AUC——曲线下面积估计值；

T_{0i}——第 i 个未患病个体的诊断试验结果；

T_{1i}——第 i 个患病个体的诊断试验结果；

$\varphi\left(T_{1i},T_{0j}\right)$——示性函数，如果$T_{0j}>T_{1i}$，$\varphi\left(T_{1i},T_{0j}\right)=0$；如果$T_{0j}=T_{1i}$，$\varphi\left(T_{1i},T_{0j}\right)=\frac{1}{2}$；如果$T_{0j}<T_{1i}$，$\varphi\left(T_{1i},T_{0j}\right)=1$。

AUC 的渐进方差估计值为

$$VAR\left(AUC\right)=\frac{AUC\left(1-AUC\right)+\left(N_1-1\right)\left(Q_1-AUC^2\right)+\left(N_0-1\right)\left(Q_2-AUC^2\right)}{N_0N_1} \tag{B.6}$$

式中：

Q_1——含义为$AUC/(2-AUC)$；

Q_2——含义为$2AUC^2/(1+AUC)$；

N_1——患病个体总数；

N_0——未患病个体总数。

AUC 的$\left(1-\alpha\right)\times 100\%$置信区间可表示为：

$$\left(AUC-{}^{Z_{1-\frac{\alpha}{2}}}\sqrt{VAR\left(AUC\right)},AUC+{}^{Z_{1-\frac{\alpha}{2}}}\sqrt{VAR\left(AUC\right)}\right)$$

B.3.2 连续型变量

当诊断试验的结果是连续型变量（如患病概率）时，同有序型变量时一样，建议绘制 ROC 曲线，并以 ROC 曲线下的面积作为评价指标。用 T_0 表示未患病者连续型试验结果的随机变量，其累积分布函数 (cumulative distribution function)为 F_0；用 T_1 表示患病者连续型试验结果的随机变量，其累积分布函数为 F_1。将散点连接起来建立经验 ROC 曲线。曲线上第 i 个散点坐标的横坐标表示为：

$$1-F_0\left(C_i\right)=\frac{1}{N_0}\sum_{j=1}^{N_0}\mathrm{I}\left(T_{0j}>C_i\right) \tag{B.7}$$

式中：

F_0 ——对 F_0 的估计；

C_i ——观测到的第 i 个试验结果值，i 取值范围为 1~N，N 为样本总量；

N_0——未患病个体总数；

T_{0j}——第 j 个未患病者的连续型试验结果；

$\mathrm{I}\left(T_{0j}>C_i\right)$——示性函数，即 $T_{0j}>C_i$ 时，$\mathrm{I}\left(T_{0j}>C_i\right)=1$，$T_{0j}\le C_i$ 时，$\mathrm{I}\left(T_{0j}>C_i\right)=0$。

曲线上第 i 个散点坐标的纵坐标表示为：

$$1-F_1\left(C_i\right)=\frac{1}{N_1}\sum_{j=1}^{N_1}\mathrm{I}\left(T_{1j}>C_i\right) \tag{B.8}$$

式中：

F_1 ——对 F_1 的估计；

N_1——患病个体总数，等于 N-N_0；

T_{1j}——第 j 个患病者的连续型试验结果；

$\mathrm{I}\left(T_{1j}>C_i\right)$——示性函数，即 $T_{1j}>C_i$ 时，$\mathrm{I}\left(T_{1j}>C_i\right)=1$，$T_{1j}\le C_i$ 时，$\mathrm{I}\left(T_{1j}>C_i\right)=0$。

ROC 曲线下的面积（AUC）的估计值可根据非参数方法得到，如下：

$$AUC=\frac{1}{N_0N_1}\sum_{i=1}^{N_1}\sum_{j=1}^{N_0}\mathrm{I}\left(T_{1i}>T_{0j}\right) \tag{B.9}$$

式中：

$\mathrm{I}\left(T_{1i}>\mathrm{T}_{0j}\right)$——示性函数，即 $T_{1i}>\mathrm{T}_{0j}$ 时，$\mathrm{I}\left(T_{1i}>\mathrm{T}_{0j}\right)=1$，$T_{1i}\le \mathrm{T}_{0j}$ 时，$\mathrm{I}\left(T_{1i}>\mathrm{T}_{0j}\right)=0$。

可应用 Bootstrap 抽样方法得到 AUC 估计值的置信区间，Bootstrap 方法的详细介绍参见参考文献[7]附录 B。

在实际的产品测试和比对中，可通过直接进行数值积分的方式，对 AUC 进行近似计算，从而避免理论模型的差异造成的影响。当不同产品的 ROC 曲线存在交叉时，可计算横坐标特定区间内的部分曲线下面积（partial AUC，简称 pAUC），扩展产品比对的维度。当测试人员关注算法的某个总体指标变化区间时（例如灵敏度高于某数值），也可计算该区间对应的 pAUC。

B.4 情形三：测试结果涉及影像位置

当测试目的为测试产品定位目标区域的准确程度时，测试数据蕴含了位置信息，此时测试者应该比较金标准中的目标区域位置和测试产品标记的位置，分辨标记正确和标记错误的位置，只有标记在阳性

影像上且位置正确的标记才能算作正确标记。

此时，建议采用自由响应受试者操作特征曲线（free-response receiver operating characteristic curve，fROC curve）或候选自由受试者操作特征曲线（alternative free receiver operating characteristics curve，AFROC curve）的曲线下面积作为评价指标。fROC 曲线和 AFROC 曲线的详细定义和计算方法可以参见参考文献[9]和[10]，以下为简略介绍。

fROC 曲线的纵坐标为不同算法阈值下对目标区域的召回率，横坐标为不同算法阈值下的非病变定位率（non-localization fraction, NLF），即每个病例上假阳性标记数量的平均值，可以估计为：

$$NLF\left(\zeta\right) = \frac{NLL\left(\zeta\right)}{N} \times 100\% \tag{B.10}$$

式中：

ζ——算法阈值；

$NLF\left(\zeta\right)$——在算法阈值 ζ 下假阳性病灶总数；

N——全体病例数量。

通过改变算法阈值设置，计算各个阈值对应的目标区域的召回率和非病变定位率，以目标区域的召回率为纵坐标，非病变定位率为横坐标，连接不同阈值下的点，可以得到经验 fROC 曲线，并可以计算该曲线下面积，即为 fROC-AUC。

fROC 曲线可以转化为 AFROC 曲线。AFROC 曲线的纵坐标与 fROC 曲线相同，横坐标为不同算法阈值下的假阳性发现率（false positive fraction, FPF），即通过对每个阴性病例上所有的标记（如果存在的话）的置信度取最大值，在给定阈值下将阴性病例错误判别为阳性病例的比例,可以估计为：

$$FPF\left(\zeta\right) = \frac{NFP\left(\zeta\right)}{N_{neg}} \times 100\% \tag{B.11}$$

式中：

$FPF\left(\zeta\right)$——在给定阈值下将阴性病例错误判别为阳性病例的比例估计值；

ζ——算法阈值；

$NFP\left(\zeta\right)$——在算法阈值 ζ 下假阳性病例总数；

N_{neg}——阴性病例总数。

通过改变算法阈值设置，计算各个阈值对应的病灶水平的灵敏度（召回率）和假阳性发现率，以病灶水平的灵敏度（召回率）为纵坐标，假阳性发现率为横坐标，连接不同阈值下的点以及(1,1)这一曲线

终点，可以得到经验 AFROC 曲线，并可以计算其曲线下面积，即为 AFROC-AUC。可应用 Bootstrap 抽样方法得到曲线下面积估计值的置信区间，具体抽样方法详见参考文献[11]附录 B。

在实际测试中，FROC、AFROC 等曲线的数值积分计算量可能远大于 ROC 曲线的数值积分计算量。为缩短测试周期，测试人员可对横坐标进行抽样，计算其中一组节点对应的纵坐标，用于进行产品比对。

当测试结果涉及影像位置时，存在一种特殊场景，即“首选”场景，满足如下限定条件：算法在每一幅图像上仅对其认为最可疑的目标区域给出提示；当且仅当算法检出的目标区域符合标记-匹配规则且概率最大时视为真阳性结果；不同图像的检出结果互不影响。对“首选”场景，可采用 LROC 曲线（localization receiver operating characteristic curve）的曲线下面积作为评价指标。具体定义及计算方法详见参考文献[14]。LROC 曲线的纵坐标为真阳性病例定位率（true positive localization fraction， TPLF），即被正确检出且目标区域定位准确的阳性病例占全体病例样本的比例；横坐标为不同算法阈值下的假阳性发现率（false positive fraction, FPF）。

B.5 主要指标的假设检验

假设 p 为主要评价指标（根据上述中描述的不同情形，p 可以是灵敏度、特异度、曲线下面积等），关于 p 的假设检验可表示为H_0：$p \le p_0$，H_1：$p > p_0$。

其中，p_0是测试人员根据产品实际需要提前选定的目标值。对于构造的关于 p 的$\left(1-\alpha\right)\times 100\%$置信区间，如果其置信区间下限大于目标值$p_0$，则说明此产品的统计指标优于预期值，满足统计有效性。

此外，当存在多个统计假设检验时，应考虑检验的多重性问题。试验目标中涉及多个指标，建议在方案设计阶段对潜在的多重性问题予以考虑，必要时应对统计检验的显著性水平进行控制，保障试验整体假阳性风险程度不超过α的水平（常用的α可取0.05）。

参 考 文 献

[1] GB/T 9386-2008 计算机软件测试文档编制规范.

[2]ISO/IEC TR 29119-11:2019, Software and systems engineering—Software testing—Part 11:Guidelines on the testing of AI-based systems.

[3]国家药品监督管理局医疗器械技术审评中心.《人工智能医疗器械注册审查指导原则（征求意见稿）》[Z]，2021

[4]国家药品监督管理局医疗器械技术审评中心. 深度学习辅助决策医疗器械软件审评要点[Z]. 北京：国家药品监督管理局医疗器械技术审评中心, 2019.

[5]国家药品监督管理局医疗器械技术审评中心. 肺炎CT影像辅助分诊与评估软件审评要点（试行）[Z]. 北京：国家药品监督管理局医疗器械技术审评中心, 2020.

[6] 国家食品药品监督管理总局. 医疗器械临床评价技术指导原则[Z]. 北京：国家食品药品监督管理总局，2015

[7]侯艳, 李康, 宇传华, 周晓华. 诊断医学中的统计学方法[M]. 第二版.北京:高等教育出版社, 2016.

[8] 中国食品药品检定研究院, 中华医学会放射学分会心胸学组. 胸部CT肺结节数据标注与质量控制专家共识(2018)[J]. 中华放射学杂志, 2019, 53(1):7.

[9]Bunch, P. C., & GH, S. (1978). A free-response approach to the measurement and characterization of radiographic-observer performance[C]. Proc. SPIE 0127, Application of Optical Instrumentation in Medicine VI, (27 December 1977).

[10] Toussaint G T. Solving geometric problems with the rotating calipers[C]. ProcIEEEMelecon, 1983.

[11]Zhou, X. H., Obuchowski, N. A., &McClish, D. K. (2014). Statistical Methods in Diagnostic Medicine[M]. John Wiley & Sons.

[12]Chakraborty, D. P., & Winter, L. H. (1990). Free-response methodology: alternate analysis and a new observer-performance experiment[J]. Radiology, 174(3), 873-881.

[13] He B, Di Dong Y S, Zhou C, et al. Predicting response to immunotherapy in advanced non-small-cell lung cancer using tumor mutational burden radiomic biomarker[J]. Journal for immunotherapy of cancer, 2020, 8(2).

[14] Zwanenburg A, Vallières M, Abdalah M A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping[J]. Radiology, 2020, 295(2): 328-338.

[15]Swensson, Richard G . Unified measurement of observer performance in detecting and localizing target objects on images[J]. Medical Physics, 1996, 23(10):1709.