

ICS 11.040.99

C30/49

YY

中华人民共和国医药行业标准

YY/T XXXX—××××

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

征求意见稿

本稿完成日期：2021.7.21

××××-××-××发布

××××-××-××实施

发 布

目 次

| | |
|-------------------------------------|----|
| 前 言 | II |
| 1 范围 | 1 |
| 2 规范性引用文件 | 1 |
| 3 术语和定义 | 1 |
| 4 测试要求 | 2 |
| 5 算法性能测试方法 | 5 |
| 附录 A （资料性） 胸部 CT 肺结节测试数据集描述样例 | 15 |
| 附录 B （资料性） 测试指标及统计分析的一般思路 | 19 |
| 参考文献 | 2 |

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。

本文件由全国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：

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。辅助分析产品的预期用途包括辅助诊断、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，不包括影像前处理及过程优化。算法分析对象包括X射线、CT、内窥镜、MRI、超声等数据模态。

注：本标准为检测方法标准，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/Txxxxx.1 人工智能医疗器械质量要求和评价 第1部分：术语

YY/Txxxxx.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/Txxxxx.1《人工智能医疗器械 质量要求和评价 第1部分：术语》、YY/Txxxxx.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》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。

注：改写GB/T 9386-2008 定义3.6

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

以一定周期重复进行的筛查。

3.6

征象 sign

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

3.7

影像征象 signs in radiology

通过影像学手段获取的征象。

3.8

压力测试 stress test

使用具有挑战性的用例或测试集开展测试的过程。

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、测试记录和测试报告。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，应形成客观定量、结构化的测试报告，对试验结果与产品声称性能指标的符合性给出判定。

为避免针对性调优，如测试过程需要复测，复测次数应少于算法标签种类。

4.2 测试环境

4.2.1 硬件环境

硬件环境是指测试使用的服务器、客户端、网络连接设备、辅助硬件设备所构成的环境。

4.2.2 软件环境

软件环境指被测软件运行时使用的操作系统、数据库、云平台与应用系统的软件等构成的环境。

4.2.3 测试环境配置

- a) 宜在软件用户文档集中规定的最低硬件及软件环境下进行测试，如运行环境中在最低环境之外还指定了“推荐环境”、“部分功能受限环境”、“最优运行环境”等，宜在这些环境下进行必要的测试或理论分析。
- b) 如存在多个软件环境，且软件环境中规定的运行库/框架等差异对算法性能可能存在影响的，应当在所有存在疑问的环境中分别测试。

- c) 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。
- d) 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。
- e) 如按要求部署测试环境后软件无法运行（这通常是软件环境规定得不全导致的），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），应当予以记录并在结果表达中明示。
- f) 测试环境应在结果表达中完整记录。

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 xxxxx.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》。测试集应独立于算法研发、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给出测试集描述的示例。

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，计算公式如下：

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

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；

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。

测试数据集的样本总量应不低于单次测试样本总量的 n 倍， n 由制造商定义。

4.3.3 测试集配置

- a) 测试集应考虑多中心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，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。

- a) 白盒扩增方式，其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) 扩增数据集的配置应符合 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，测试平台应符合如下要求：

- 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 etc。
- 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。对算法阈值锁定的产品，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。

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、感兴

趣区域所在图像的编号、感兴趣区域的位置、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，建立测试报告。

4.7 测试报告

测试报告对测试结果进行客观、定量的描述，内容至少应包含：

- a) 软件环境；
- b) 硬件环境；
- c) 测试平台描述（如适用）；
- d) 测试集描述；
- e) 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；
- f) 算法错误分析。

5 算法性能测试方法

5.1 算法应用场景与测试方法

5.1.1 目标检测场景

5.1.1.1 标记与匹配

具有目标检测功能的产品应输出算法标记目标，测试集的参考标准应包含对应的感兴趣区域。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：匹配阈值不同于算法的检出概率。

常见标记匹配方式举例：

- 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感兴趣区域中心是否落入参考标准区域范围内来确定匹配结果。

注：中心点的选择与感兴趣区域有关。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 测试方法

- a) 算法以测试集作为输入，输出与参考标准格式兼容的结果；
- 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 5.1.1.2.1-5.1.1.2.6 描述的公式计算性能指标。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。

5.1.1.2.1 召回率——病变定位率

召回率的计算公式见公式（2）：

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

5.1.1.2.2 精确度

精确度的计算公式见公式（3）：

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

5.1.1.2.3 F1 分数

F1分数的计算公式见公式（4）：

$$F_1 = \frac{2 \times \text{Pre} \times \text{Recall}}{\text{Pre} + \text{Recall}} \dots\dots\dots (4)$$

5.1.1.2.4 平均精确度 average precision

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。

5.1.1.2.5 平均精确度均值 mean average precision

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。

5.1.1.2.6 fROC 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fROC曲线。横坐标的取值一般设为等比数列，即0.5、1、2、…、N，其中上限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）：

$$NLR = \frac{FP}{N} \times 100\% \dots\dots\dots (5)$$

式中：

NLR ——非病变定位率；

FP ——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：也可称为单病例平均假阳个数。

5.1.2 区域分割与测量

5.1.2.1 测试方法

- a) 输入测试集，输出约定格式的结果，至少包含分割区域边界端点坐标；
- a) 算法分割的感兴趣区域与参考标准分割的感兴趣区域的性能指标按 5.1.2.1.1-5.1.2.1.4 描述的公式进行计算。
- b) 以计算结果的平均值作为最终结果。
- c) 如算法合并检测功能，仅对 TP 进行计算。

5.1.2.1.1 召回率

算法分割的感兴趣区域与参考标准分割的感兴趣区域的交集除以参考标准分割的感兴趣区域，用公式（6）表示：

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

式中：

Rec ——召回率；

S_{pr} ——算法分割的感兴趣区域；

S_{gt} ——参考标准分割的感兴趣区域。

5.1.2.1.2 精确度

算法分割的感兴趣区域与参考标准分割的感兴趣区域的交集除以算法分割的感兴趣区域，用公式（7）表示：

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

式中：

Pre ——精确度； S_{pr} ——算法分割的感兴趣区域； S_{gt} ——参考标准分割的感兴趣区域。

5.1.2.1.3 交并比

当感兴趣区域为一般实体时（如肺结节），宜采用Dice系数或Jaccard系数计算交并比。

Dice系数为算法分割的感兴趣区域与参考标准分割的感兴趣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:

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

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

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

5.1.2.1.4 树检测长度

当感兴趣区域为气管或其他树形结构时,宜采用树检测长度(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:

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

其中 L_{TP} 为正确分割的气管长度(以像素计数), L_{gt} 为参考标准气管的长度(以像素计数)。

5.1.2.1.5 表面距离

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

记 X 为算法给出的感兴趣区域, Y 为参考标准给出的感兴趣区域,根据公式(11)计算双向豪斯多夫距离 d_H 如下:

$$d_H(X, Y) = \max \{d_{XY}, d_{YX}\} = \max \left\{ \max_{x \in X} \min_{y \in Y} d(x, y), \max_{y \in Y} \min_{x \in X} d(x, y) \right\} \dots\dots\dots (11)$$

其中 $d(x, y)$ 为 X 、 Y 两个区域任意两点之间的距离。

5.1.2.1.6 密度测量

算法识别的感兴趣区域内像素的密度值或灰度值,与参考标准感兴趣区域的结果进行比较,以计算相对误差绝对值的平均值。

$$S = \frac{\sum_{i=1}^n \left| \frac{L_i - L_{a1}}{L_{a1}} \right|}{n} \times 100\% \dots\dots\dots (12)$$

式中:

S ——偏差值; L_i ——第 i 个病例的测量值; L_{a1} ——第 i 个病例的参考标准值; n ——病例个数。

5.1.2.1.7 尺寸测量

尺寸测量的对象是感兴趣区域的尺寸,如长短径、紧密包裹矩形框的长宽等。感兴趣区域可以是二维平面、三维立体空间。

当算法识别的感兴趣区域可近似看做凸形状时，可对算法识别的感兴趣区域轮廓（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，与参考标准的感兴趣区域结果进行比较，计算相对误差绝对值的平均。

5.1.2.1.8 体积测量

分别统计算法识别的感兴趣区域和参考标准的感兴趣区域内的体素数量，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

5.1.3.1 测量方法

- a) 向待测算法输入测试集，输出与参考标准格式兼容的结果；
- 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 | $N_{1,1}$ | $N_{1,2}$ | ... | ... | ... | ... |
| True_2 | ... | $N_{2,2}$ | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... |
| True_n | ... | ... | ... | ... | ... | $N_{n,n}$ |

注：Pred_x (x=1~n) 为人工智能诊断为 x 类的类别；True_x (x=1~n) 为参考标准诊断为 x 类的类别； $N_{i,j}$ (i=1~n, j=1~n) 为参考标准的诊断结果为 i 类，被人工智能诊断为 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 | $TP = N_{i,i}$ | $FN = \sum_{j=1, j \neq i}^n N_{i,j}$ |
| | 阴性 | $FP = \sum_{j=1, j \neq i}^n N_{j,i}$ | $TN = \sum_{j=1, j \neq i}^n \sum_{l=1, l \neq i}^n N_{j,l}$ |

5.1.3.1.1 灵敏度

灵敏度用 Sen 表示，表达式见式（13）：

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

式中：

Sen ——灵敏度；

TP ——真阳性样本的个数；

FN ——假阴性样本的个数。

5.1.3.1.2 特异度

特异度用 Spe 表示，表达式见式（14）：

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

式中：

TN ——真阴性样本的个数；

FP ——假阳性样本的个数。

5.1.3.1.3 漏检率

漏检率用 MR 表示，表达式见式（15）：

$$MR = 1 - Sen \dots\dots\dots (15)$$

式中：

Sen ——灵敏度；

MR ——漏检率。

5.1.3.1.4 阳性预测值

阳性预测值用 PPV 表示，表达式见式（16）：

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

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。

5.1.3.1.5 阴性预测值

阴性预测值用NPV表示，表达式见式（17）：

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

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。

5.1.3.1.6 准确率

准确率用Acc表示，表达式见式（18）：

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

5.1.3.1.7 约登指数

约登指数用Y表示，表达式见式（19）：

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

5.1.3.1.8 Kappa 系数

Kappa系数用K表示，表达式见公式（20）：

$$K = \frac{Acc - P_e}{1 - P_e} \dots\dots\dots (20)$$

其中：

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

5.1.3.1.9 ROC 曲线

制造商应给出标称的ROC曲线与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性，以1-特异性为横坐标、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

对于同时具有检出、分类、分割、测量等功能的产品，测试人员对算法性能宜进行分步、渐进的评价，避免各性能指标在计算时互相干扰，如：

1) 首先对标记-匹配场景进行评价，计算检出的指标；

- 2) 其次对正确检出的感兴趣区域，计算分类的指标；
- 3) 再次对分类正确的感兴趣区域，计算分割的指标；
- 4) 最后计算测量相关指标。

当产品算法功能有约束条件时（如算法仅识别大于某尺寸的病灶），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

对具有随访评估功能的产品，应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，比较算法对同一感兴趣区域的分析结果与参考标准之间的符合性，计算绝对误差；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

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5.1.3.1的方法计算灵敏度、特异性、Kappa系数等指标。

对具有患者优先级排序的产品，参照执行本条款的方法。

5.1 算法质量特性与测试方法

5.1.1 泛化能力

泛化能力是指算法对陌生样本的适应能力。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。

5.1.2 鲁棒性

制造商应根据产品风险分析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，获取或模拟相关数据，组成专用测试集，对算法性能进行对抗测试，分析各指标的变化情况，形成鲁棒性研究资料。

5.1.2.1 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。

5.1.2.2 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。

5.1.2.3 面向欺骗攻击的对抗测试

欺骗攻击是一种加入人员难以觉察的扰动从而骗过模型的攻击手段，测试人员可使用白盒攻击（Projected Gradient Descent, PGD）产生最大范数有限（如不到8/256）的扰动，并将扰动插入到原始图像中，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恶意欺骗攻击。

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。

测试人员宜根据产品的网络安全能力及风险分析文档，确定欺骗攻击的适用性和试验参数配置。施加扰动后的数据应通过标注人员的确认后用于测试。

5.1.2.4 压力测试

压力测试是在模拟实际应用中可能遇到的长时间极端输入或者环境下（不同负载、极限值、边界值、大容量数据、错误数据、稀有数据等），测试某算法模型的性能、可靠性、稳定性等。

5.1.2.4.1 压力样本的定义

压力样本是指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。压力样本不应影响医生的正常判断。

5.1.2.4.2 压力样本的选取

压力样本的选取可遵循以下原则：

- 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。

5.1.3 重复性

测试人员应对同一测试集进行重复测试，测试次数不宜低于三次。

5.1.4 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。

5.1.5 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。

5.1.6 错误分析

测试人员应对算法的错误进行定量分析，如下列情形：

- 1) 在标记-匹配场景下，测试人员宜对假阴性结果进行分解，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，写入测试报告；
- 2) 在多分类场景下，测试人员宜对每一种分类的假阴性、假阳性结果进行分解，写入测试报告；
- 3) 在分割场景下，测试人员宜根据感兴趣区域的尺寸，对分割结果进行分解，写入测试报告；
- 4) 测试人员宜针对每个病例计算算法性能指标，评估算法对个体的偏倚，写入测试报告；

5) 测试人员宜在对抗测试、压力测试中采用1) -4) 的方法开展错误分析。

附 录 A
(资料性)

胸部 CT 肺结节测试数据集描述样例

本文件以胸部CT肺结节测试数据集为具体案例，给出测试数据集描述的举例，仅作为参考信息。

A. 1 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. 2 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，可根据实际掌握的信息进一步细化。

表A. 1 数据来源的多样性统计

| | | |
|------|----------|-----|
| 年龄 | 40-60岁 | xx例 |
| | 61-80岁 | xx例 |
| | >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 | <50mAs | xx例 |
| | >50mAs | xx例 |
| 管电压 | <110kV | xx例 |
| | >110kV | xx例 |
| 层厚 | <1. 5mm | xx例 |
| | >1. 5mm | xx例 |
| 采集场所 | 体检 | xx例 |
| | 门诊 | xx例 |
| | 住院 | xx例 |

A.3 数据集的分布构成

以胸部CT肺结节测试集的构成为例，描述模板如表A.2所示：

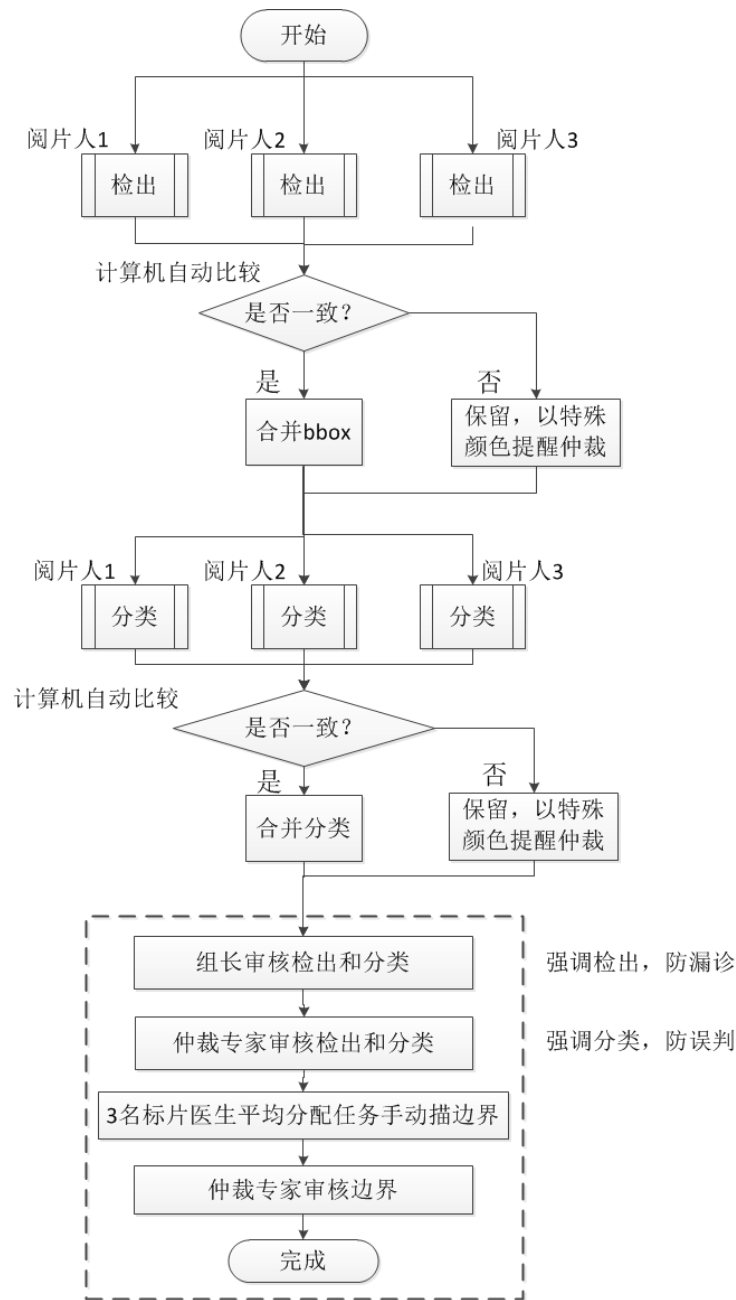
表 A.2 肺结节分布统计

| 肺结节种类 | 尺寸（mm） | 病例(个) |
|----------|--------|-------|
| 肺内实性结节 | <4 | xx |
| | 6-8 | |
| | >8 | |
| 肺内部分实性结节 | <4 | xx |
| | 6-8 | |
| | >8 | |
| 肺内纯磨玻璃结节 | <4 | xx |
| | 6-8 | |
| | >8 | |
| 肺内钙化结节 | / | xx |
| 胸膜实性结节 | / | xx |
| 胸膜钙化结节 | / | xx |
| 其他疾病 | / | xx |
| 总计 | / | xx |

A.4 数据集标注规则

标注参考依据：胸部CT肺结节数据标注与质量控制专家共识（2018）

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.5 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：

$$N1 = \frac{Z_{a/2}^2 P_{sen} (1 - P_{sen})}{d^2 \times P_{re}} \dots\dots\dots (A.1)$$

式中：

$N1$ ——总体样本量；

$Z_{a/2}$ ——标准正态分布的 Z 分数（置信度可由制造商宣称，一般为95%； α 为显著性水平）；

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。

为保证特异性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：

$$N2 = \frac{Z_{a/2}^2 P_{spe} (1 - P_{spe})}{d^2 \times (1 - P_{re})} \dots\dots\dots (A.2)$$

式中：

$N2$ ——总体样本量；

$Z_{a/2}$ ——标准正态分布的 Z 分数（置信度可由企业宣称，一般为95%； α 为显著性水平）；

P_{spe} ——特异性；

P_{re} ——测试集中的患病率；

d ——特异性的允差。

A.6 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 XXXX.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》，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

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% 置信区间。

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，本部分仅属于推荐性内容。

B.1 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。使用表 B.1 中的符号，灵敏度和特异度的估计值为

$$\widehat{Sen} = \frac{N_{1,1}}{N_{1,1} + N_{1,2}} \quad (B.1)$$

$$\widehat{Spe} = \frac{N_{2,2}}{N_{2,1} + N_{2,2}} \quad (B.2)$$

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。关于灵敏度和特异度的 $(1 - \alpha) \times 100\%$ 置信区间的表达式分别为

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

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

其中， $z_{1-\frac{\alpha}{2}}$ 为正态分布的 $1 - \frac{\alpha}{2}$ 分位数。通常，建议取 $\alpha = 0.05$ 。其他构建置信区间的方法可以参见参考文献[6]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[6]的第二章和第四章。

B.2 情形二：测试结果为有序型或连续型变量

a) 有序型变量

当诊断试验的结果是有序变量（如 RADS 分级）时，建议绘制 ROC 曲线，并以 ROC 曲线下的面积，即 AUC，作为统计评价指标。这里，我们用 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。用 T 表示诊断试验结果，为 K 类有序结果。诊断试验结果的数据结构如表 2.2 所示。利用

所有有序评分尺度绘制散点 $\left(1 - \widehat{Spe}(k), \widehat{Sen}(k)\right)$, $k = 1, \dots, K$, 即可得到经验 ROC 曲线，其中

$$1 - \widehat{Spe}(k) = \frac{1}{N_0} \sum_{j=k}^K r_j, \quad \widehat{Sen}(k) = \frac{1}{N_1} \sum_{j=k}^K s_j. \quad \text{用 } T_{0i} \text{ 表示第 } i \text{ 个未患病个体的诊断试验结果, } T_{1i} \text{ 表示第 } i \text{ 个患病个体的诊断试验结果, ROC 曲线下的面积估计可由非参数方法得到, 即}$$

$$\widehat{AUC} = \frac{1}{N_0 N_1} \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} \varphi(T_{1i}, T_{0j}) \quad (\text{B.3})$$

其中，如果 $T_{0j} > T_{1i}$, $\varphi(T_{1i}, T_{0j}) = 0$ ；如果 $T_{0j} = T_{1i}$, $\varphi(T_{1i}, T_{0j}) = \frac{1}{2}$ ；如果 $T_{0j} < T_{1i}$,

$\varphi(T_{1i}, T_{0j}) = 1$ 。 \widehat{AUC} 的渐进方差估计值为

$$\widehat{VAR}(\widehat{AUC}) = \frac{\widehat{AUC}(1 - \widehat{AUC}) + (N_1 - 1)\left(\widehat{Q}_1 - \widehat{AUC}^2\right) + (N_0 - 1)\left(\widehat{Q}_2 - \widehat{AUC}^2\right)}{N_0 N_1} \quad (\text{B.4})$$

其中， $Q_1 = \frac{\widehat{AUC}}{2 - \widehat{AUC}}$ 和 $Q_2 = \frac{2\widehat{AUC}^2}{1 + \widehat{AUC}}$ ，AUC 的 $(1 - \alpha) \times 100\%$ 置信区间可表示为

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

其中， $Z_{1-\frac{\alpha}{2}}$ 为正态分布的 $1 - \frac{\alpha}{2}$ 分位数。

b) 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，即 AUC，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，其分布函数为 F_0 ；用 T_1 表示患病者连续型试验结果的随机变量，其分布函数为 F_1 。经验 ROC 曲线将

点 $\left(1 - \widehat{F}_0(C_i), 1 - \widehat{F}_1(C_i)\right)$ 连接起来的图形, 其中 $1 - \widehat{F}_0(C_i) = \frac{1}{N_0} \sum_{j=1}^{N_0} I(T_{0j} > C_i)$,

$1 - \widehat{F}_1(C_i) = \frac{1}{N_1} \sum_{j=1}^{N_1} I(T_{1j} > C_i)$, C_i 的取值范围是观测到的试验结果值, $I(T_{0j} > C_i)$ 表示示性函数, 即 $T_{0j} > C_i$ 时, $I(T_{0j} > C_i) = 1$, $T_{0j} \leq C_i$ 时, $I(T_{0j} > C_i) = 0$ 。

ROC 曲线下的面积 (AUC) 的估计值可根据非参数方法得到,

$$\widehat{AUC} = \frac{1}{N_0 N_1} \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} I(T_{1i}, T_{0j}) \quad (\text{B.5})$$

其中, $I(T_{1i} > T_{0j})$ 是示性函数, 定义同上。可应用 Bootstrap 抽样方法得到曲线下面积估计值的置信区间, Bootstrap 方法的详细介绍参见参考文献[6]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.3 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感兴趣区域的准确程度时, 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感兴趣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 FROC 曲线 (free-response receiver operating characteristic curve) 或 AFROC 曲线 (Alternative FROC) 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[8]和[9], 以下为简略介绍。

FROC 曲线的纵坐标为不同算法阈值下对感兴趣区域的召回率, 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:

$$\widehat{NLF}(\zeta) = \frac{NLL(\zeta)}{N} \times 100\% \quad (\text{B.6})$$

其中, ζ 为算法阈值, $NLL(\zeta)$ 为在算法阈值 ζ 下假阳性病灶总数; N 为全体病例数量。通过改变算法阈值设置, 计算各个阈值对应的感兴趣区域的召回率和非病变定位率, 以感兴趣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：

$$\widehat{FPF}(\zeta) = \frac{NFP(\zeta)}{N_{neg}} \times 100\% \quad (B.7)$$

其中， ζ 为算法阈值， $NFP(\zeta)$ 为在算法阈值 ζ 下假阳性病例总数， N_{neg} 为阴性病例总数。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[7]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感兴趣区域给出提示；当且仅当算法检出的感兴趣区域符合标记-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[12]。LROC 曲线的纵坐标为真阳性病例定位率（true positive localization fraction, TPLF），即被正确检出且感兴趣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.4 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、AUC 等），关于 p 的假设检验可表示为 $H_0 : p \leq p_0$ ， $H_1 : p > p_0$

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 $(1 - \alpha) \times 100\%$ 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，满足优效性。

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）。

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 | $N_{2,1}$ | $N_{2,2}$ |

表 B.2 诊断试验有序数据结构

| 疾病状态 (D) | 试验结果 (T) | | | | 合计 |
|----------|----------|-------|-----|-------|-------|
| | 1 | 2 | ... | K | |
| 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 |

参 考 文 献

- [1] ISO/IEC TR 29119-11:2019, Software and systems engineering—Software testing—Part 11: Guidelines on the testing of AI-based systems.
- [2] 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
- [3] 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.
- [4] 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.
- [5] 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
- [6] 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.
- [7] Zhou, X. H., Obuchowski, N. A., & McClish, D. K. (2014). Statistical Methods in Diagnostic Medicine[M]. John Wiley & Sons.
- [8] 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).
- [9] Chakraborty, D. P., & Winter, L. H. (1990). Free-response methodology: alternate analysis and a new observer-performance experiment[J]. Radiology, 174(3), 873-881.
- [10] 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).
- [11] 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.
- [12] Swensson, Richard G. Unified measurement of observer performance in detecting and localizing target objects on images[J]. Medical Physics, 1996, 23(10):1709.