



中 华 人 民 共 和 国 医 药 行 业 标 准

YY/T XXXXX—XXXX

人工智能医疗器械 X 射线影像骨龄辅助评
估软件 算法性能测试方法

Artificial Intelligence Medical Device—Computer Assisted Analysis Software for X
ray Bone Age Assessment—Algorithm performance test methods

点击此处添加与国际标准一致性程度的标识

文稿版次选择

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

 4.1 总则..... 2

 4.2 测试场景分析..... 3

 4.3 测试环境..... 3

 4.4 测试平台..... 3

 4.5 测试资源..... 3

 4.6 测试指标..... 4

 4.7 测试流程要求..... 4

 4.8 通过准则..... 4

5 测试方法..... 4

 5.1 目标检测 4

 5.2 区域分割与测量..... 5

 5.3 影像分类..... 5

 5.4 泛化能力..... 6

 5.5 鲁棒性..... 6

 5.6 重复性..... 6

 5.7 效率..... 6

资料性附录 A 测试集模板

前言

本标准的编写遵循国家标准GB/T1.1-2009《标准化工作导则 第1部分：标准的结构和编写》的规定。

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。

本标准由浙江大学医学院附属儿童医院提出。

本标准由全国人工智能医疗器械标准化技术归口单位归口。

本标准起草单位：浙江大学医学院附属儿童医院等。

本标准主要起草人：傅君芬、周雪莲、吴蔚、李一鸣、马晓杰、宫平、袁豪磊。

人工智能医疗器械 X射线影像骨龄辅助评估软件 算法性能测试方法

1 范围

本文件规定了采用人工智能技术的X射线骨龄辅助评估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对X射线骨龄影像进行后处理的辅助分析软件、软件组件。

本文件不适用于影像前处理及过程优化。

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文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 9386-2008 计算机软件测试文档编制规范

YY/T 0106-2008 医用诊断X射线机通用技术条件

GB/T 25000.51-2016 系统与软件工程 系统与软件质量要求和评价（SQua RE） 第51部分：就绪可用软件产品（RUSP）的质量要求和测试细则

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

YY/T 1833.3 人工智能医疗器械质量要求和评价 第2部分：数据集标注通用要求

3 术语和定义

YY/T 1833.1、YY/T 1833.2、YY/T 1833.3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 骨龄 Bone Age

骨龄是骨骼的生物年龄，不依赖于“日历年龄”和生长速度的变化，直接地反映机体生长发育的成熟程度。骨龄评估是目前临床中应用最广泛地用于评价儿童青少年生物年龄及骨骼成熟度的方法。在临床医学、法医学和运动医学领域有广泛的应用。由于手腕部包含多种类型的骨化中心，且易于拍摄，X 射线照射剂量小，手腕部骨龄得到了最为广泛的应用。本标准草案中骨龄特指手腕部骨龄。

3.4 骨龄评价方法 Bone Age Assessment Methods

常用的骨龄评价方法有 G-P 图谱法、TW3 法、中华 05 法等，各种骨龄评价方法均为从特定人群中，采样构建标准化的儿童手腕骨 X 射线影像样本库，研制而成一系列标准图谱及骨发育成熟度的描述文本。不同方法评价同一儿童的骨龄可出现一定差异，骨龄评价需明确采用的具体骨龄评价方法。

3.5 G-P 图谱法

G-P 图谱参照样本为 1931~1942 年美国俄亥俄州克利夫兰地区的白人儿童。G-P 图谱共包括男 31 幅、女 29 幅标准片。每一幅标准片选自 100 名同性别、同年龄儿童的 X 线片。G-P 图谱法的参照样本是整张手腕骨 X 射线片。

3.6 TW3 法

TW3 参照样本为 20 世纪 70~90 年代的欧美儿童。不同于图谱法，TW3 方法选取 20 块手腕骨为评价对象，并将每块骨的发育过程划分为若干个等级阶段，所有骨的各等级均设置标准化图谱。TW3 法分别制定了 RUS 骨与腕骨骨龄标准。

3.7 中华 05 法

中华 05 法参照 TW3 计分法，参照样本为 21 世纪初中国城市汉族正常儿童样本。抽样城市为上海市、广州市、温州市、大连市、石家庄市。

3.8 桡尺短骨 RUS

RUS 是桡骨(Radius)、尺骨(Ulna)、短骨(Short bones)的缩写，其中短骨特指掌骨和指骨。RUS 骨共

计 13 块骨。

3.9 腕骨 Carpal

计分法骨龄评价中，腕骨特指头状骨、钩状骨、三角骨、月骨、舟骨、大多角骨、小多角骨共计 7 块骨。

3.10 感兴趣区域 Region of Interest(ROI)

与研究问题相关的医学图像中区域。

3.11 骨成熟度指征 Indicators of bone maturity

用来确定骨成熟度的特征称为骨成熟度指征，每块骨的成熟度指征均以不可逆的顺序规律性地出现。如骨化中心的出现、骨骺和骨干干骺端的形态变化、骨骺与骨干的融合等。

3.12 骨发育等级 bone maturity level

计分法（TW3、中华 05 等）的骨龄评价方法，根据每块手腕骨的成熟度指征，对每块骨划分出由低而高的一系列骨发育等级。骨发育等级通常以英文字母(TW3)或阿拉伯数字标记（CHN、中华 05）。所有手腕骨的骨发育等级，决定了骨龄值。

4 测试要求

4.1 总则

算法性能测试是 X 射线影像骨龄辅助评估软件验证与确认的重要环节，应建立完整的测试文档，包括测试计划、测试日志和测试报告。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编写测试计划。在测试过程中，应同步记录测试日志，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明确判定。

4.2 测试场景设计

测试计划应描述测试场景设计，评估与产品临床使用场景的差异，宜提供以下信息：

- a) 分析测试数据集与临床使用场景在人群特征、数据质量要求、标注标准、采集场所方面的差异；
- b) 分析算法测试环境与临床部署环境的差异，包括软件、硬件、网络、数据接口、存储、周境；

- c) 分析测试中的产品运行模式与临床使用模式的差异；
- d) 比较算法输出结果与实际软件界面输出结果的含义，建立映射关系；
- e) 分析算法测试过程与临床使用过程的人员操作差异。

4.3 测试环境

算法测试宜在能模拟产品临床部署场所的环境下进行。测试场地的气候环境、机械环境、电磁环境应保持稳定，不影响产品正常运行。

4.4 测试平台

—如适用，测试平台的算力、内存、存储空间等硬件配置应满足产品正常运行需要的最低要求，软件周境、操作系统、人机界面应与产品兼容。

—如果测试需要在网络条件下进行，网速、传输服务质量（QoS）应与临床部署环境一致。

—如果算法测试在云平台进行，算法的封装方式、平台调用的算力和数据应形成记录，以评估实际产品的运行效率。

4.5 测试资源

4.5.1 测试集

测试集的质量应满足 YY/Txxxxx《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。制造商宜结合测试的置信度、算法主要指标的允差，计算单次测试的样本量要求。测试数据集的总体样本量应不低于单次测量样本量的 n 倍， n 由制造商定义。

测试日志应记录测试集的版本、标识、总体样本量、实际使用的样本信息、使用日期。测试集可进行阶段性的更新，制造商宜明确测试集更新的原因、频次、比例和特征层次的变化情况，建立版本控制与变更记录，保证数据使用的可追溯。

4.6 测试指标

制造商宜根据产品技术特性、预期用途和使用场景，在测试计划中明确客观、定量的测试指标。

4.7 测试流程要求

制造商宜根据测试计划开展测试活动，实时记录测试日志。测试流程各步骤的要求如下：

a)测试前

待测产品应能读取批量数据，能对每条数据或每个病例进行预览和分析；输入数据与输出结果之间应建立唯一对应关系。制造商宜选用小批量数据进行预测试，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，避免系统偏差。

b)测试过程中

测试进度宜实现可视化。制造商应记录每条数据或每个病例的处理情况，包括数据量、处理时间、处理结果，以及算法运行过程中的错误、警告、异常提示。

c)测试后

制造商宜结构化批量导出算法结果。算法结果的数据格式应与参考标准兼容。制造商应保证算法结果的完整性与可追溯性，防止对算法结果进行误操作，从而影响算法测试结论。制造商应量化比对算法结果与参考标准（来自测试集、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 通过准则

通过准则包括单项指标与产品整体质量，应在测试计划中进行描述。如适用，宜从病灶、病例、测试集子集和测试集总体进行多层次统计分析，判断各单项指标是否通过。对于产品整体质量，制造商可根据产品预期用途和风险分析，确定适用的各指标权重，计算总体分数并设定阈值，作为整体通过的判定依据。

5 算法性能测试方法

5.1 各种算法应用场景的测试方法

本标准描述的典型算法应用场景有：手腕部关键点定位，骨发育等级评价，骨龄评价(计算)。

5.1.1 手腕部关键点定位

产品界面中，若体现手腕部关键点信息，应测试关键点检测成功率、关键点定位误差。该指标主要适用于逐块骨评价的方法，如 TW3 法、中华 05 法。

5.1.1.1 测试步骤

在手腕部关键点定位场景下，算法测试按如下步骤进行：

- a) 待测产品需提供其手腕部关键点的定义文档，以制定测试集的金标准。厂商定义的手腕部关键点，需覆盖所选骨龄评价方法（TW3, 中华 05 法）的全部目标骨或骨骺。
- b) 向待测算法输入测试集，输出算法定位关键点的结果，其格式宜与参考标准兼容；
- c) 关键点定位的性能按 5.1.1.2~5.1.1.3 描述的公式进行计算，得到评测指标结果。

5.1.1.2 平均径向误差(Mean Radial Error)

平均径向误差(MRE)的定义见公式(1)。其中样本总数记为N，每个样本中的关键点数记为K。第i个样本的第j个关键点，金标准坐标记为 $(\hat{x}_{ij}, \hat{y}_{ij})$ ，模型预测关键点的坐标记为 (x_{ij}, y_{ij}) 。坐标单位均为毫米 mm。

$$MRE = \frac{1}{N \times M} \sum_{i=1}^N \sum_{j=1}^K \sqrt{(x_{ij} - \hat{x}_{ij})^2 + (y_{ij} - \hat{y}_{ij})^2} \quad (1)$$

5.1.1.3 成功检测比例(Successful Detection Rate)

成功检测比例(SDR)的定义见公式(2)，其物理意义表示成功定位的关键点占全部关键点的比例。其中 $\mathbb{I}$ 表示符号函数(indicator function)，T 为一个设定的阈值。其他字符的定义同公式(1)。T 的设定非常关键。推荐分别给出 T 为 2mm, 4mm, 8mm 时的 SDR。

$$SDR = \frac{1}{N \times M} \sum_{i=1}^N \sum_{j=1}^K \mathbb{I} \left\{ \sqrt{(x_{ij} - \hat{x}_{ij})^2 + (y_{ij} - \hat{y}_{ij})^2} \leq T \right\} \quad (2)$$

5.1.2 骨发育等级评价

对具有骨发育等级自动评价功能的产品，应测试其骨发育等级评价误差。该指标主要适用于逐块骨评价的方法，如 TW3 法、中华 05 法。

5.1.2.1 测试步骤

在骨发育等级评价场景下，算法测试按如下步骤进行：

- a) 厂商需明确其模型采用的骨龄评价方法(TW3, 中华 05 等)，以决定测试集参考金标准。
- b) 向待测算法输入测试集，得到算法输出的 13 块 RUS 骨和 7 块腕骨的骨发育等级结果。该结果的格式宜与参考标准兼容。
- c) 骨发育等级评价的算法性能指标，按 5.1.2.2 描述的公式进行计算。

5.1.2.2 骨发育等级平均绝对误差(Mean Absolute Error)

骨发育等级平均绝对误差(MAE)的定义见公式(3)。其中样本总数记为N，骨龄评价涉及骨的总数记为B。对 RUS 骨龄，B 等于 13。对腕骨骨龄，B 等于 7。

$$MAE_{stage} = \frac{1}{N \times B} \sum_{i=1}^N \sum_{b=1}^B \|D_{ib} - \hat{D}_{ib}\| \quad (3)$$

式中：

MAE_{stage} ——骨发育等级评价的平均绝对误差；

D_{ib} ——第 i 个样本，第 b 块手腕骨的算法预测骨发育等级

$\hat{D}_{ib}$ ——第 i 个样本，第 b 块手腕骨的金标准骨发育等级

5.1.3 骨龄评价

该指标适用于所有骨龄评价方法，包括但不限于 G-P 图谱、TW3 法、中华 05 法。

5.1.3.1 测试步骤

在骨发育等级评价场景下，算法测试按如下步骤进行：

- d) 厂商需明确其模型采用的骨龄评价方法(TW3, 中华 05 等)，以决定测试集参考金标准。
- e) 向待测算法输入测试集，得到算法输出的 13 块 RUS 骨和 7 块腕骨的骨发育等级结果。该结果的格式宜与参考标准兼容。
- f) 骨发育等级评价的算法性能指标，按 5.1.2.2 描述的公式进行计算。

5.1.3.2 骨龄平均绝对误差(Mean Absolute Error)

骨龄平均绝对误差(MAE)的定义见公式(4)。其中样本总数记为N

$$MAE_{BA} = \frac{1}{N} \sum_{i=1}^N \|Y_i - \hat{Y}_i\| \quad (4)$$

式中：

MAE_{BA} ——骨龄的平均绝对误差

Y_i ——第 i 个样本的算法预测骨龄

$\hat{Y}_i$ ——第 i 个样本的金标准骨龄

5.2 泛化能力

可根据测试集的数据质量层级，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，评估算法在不同子集的性能，作为对泛化能力的评估。

5.3 鲁棒性

制造商可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，对算法性能进行测试，分析各指标的变化情况，形成鲁棒性研究资料。

5.4 重复性

制造商宜对同一测试集进行重复测试，可选用不同运行环境和软硬件配置。为避免对某个测试集进行针对性调优，测试次数应小于算法标签种类。

5.5 效率

—制造商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，进行多次计时取平均。

—辅助评估类、优先级排序类产品应以医生收到算法通知作为终点。

附录 A

人工智能医疗器械 X 射线影像骨龄辅助评估软件测试数据集描述样例（依据 YY/T xxx 标准）

A-1 数据集适用范围

数据集适用于声称能对 X 射线影像骨龄辅助评估的软件产品，这些产品预期用途为骨龄评测、骨发育等级评价、手腕骨关键点定位等。

A-2 数据的获取

患者人群：数据来源于 X 个省份的 X 家医院，年龄 0-20 岁之间，男性占 yy%，女性占 xx%。

采集场所：x 家三甲医院，y 家 2 级医院，门诊占 m%，住院部占 n%；z 家体检机构；

采集设备：主流厂家的 DR 设备

数据格式：Dicom 3.0

采集人员：临床医生

伦理：如适用，测试集使用的数据已获得医院伦理委员会批准，患者的隐私保护应当满足法规的要求。

注：建议给出不同维度的统计表格。

年龄	0-2 岁	xx 例
	2-4 岁	xx 例
	4-6 岁	xx 例
	6-8 岁	xx 例
	8-10 岁	xx 例
	10-12 岁	xx 例
	12-14 岁	xx 例
	14-16 岁	xx 例
	16-18 岁	xx 例
	>18	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
DR 机型	XX 公司 XX 型号	xx 例
	XX 公司 XX 型号	xx 例
管电流	<4mAs	xx 例
	>=4mAs	xx 例
管电压	<50kV	xx 例

	>=50kV	xx 例
采集场所	体检	xx 例
	门诊	xx 例
	住院	xx 例

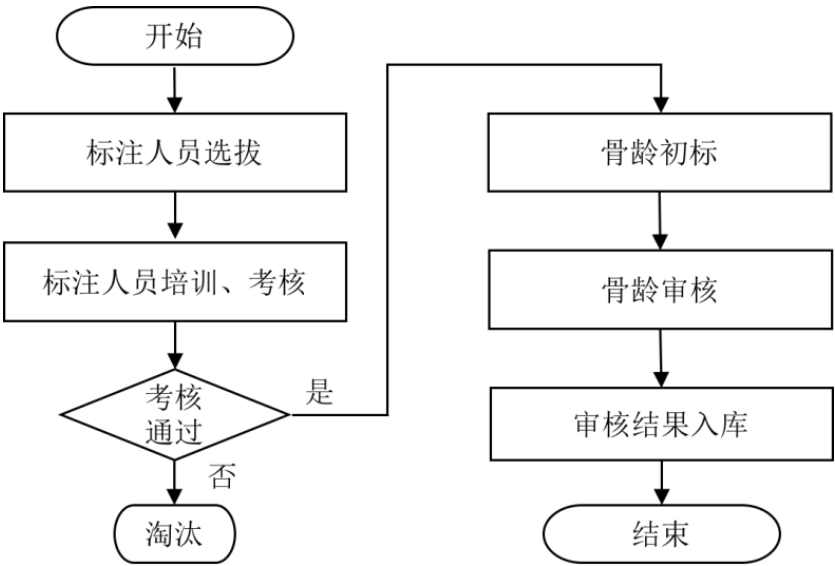
A-3 数据标定规则

标注参考依据：

G-P 图谱法标注参考依据：参考文献(1)

TW3 法标注参考依据：参考文献(2)

中华 05 法标注参考依据：参考文献(3)



标注流程质控：

(1) 标注流程。整体的标注流程如上图所示，含标注人员选拔、标注培训、准入考核、初标、审核等环节。选拔的标注医生负责“骨龄初标”，应具有影像科、儿童内分泌科、儿童保健科或相关科室 2 年以上骨龄阅片经验，获得执业医师资格证，能独立完成诊断报告。招募的审核医生负责“骨龄审核”，应

具有影像科、儿童内分泌科、儿童保健科或相关科室 10 年以上骨龄阅片经验，且取得副高及以上职称。标注医生和审核医生均需参与骨龄阅片培训，学习所选用的骨龄评价方法的权威资料(参见“**标注参考依据**”)。培训后，医生需完成“骨龄阅片考核任务”。根据其在“骨龄阅片考核任务”的表现，决定是否录用。

(2) 质控措施。鉴于骨龄阅片的主观性和部分影像的歧义性，建议采用两标一审的模式：即先由两位标注医生行盲法独立标注同一份病例，再由审核医生给出最终的骨发育等级评价。审核医生可见两位初标医生的标注结果，审核医生以“**标注参考依据**”为准，核实标准图谱和文字，并进行必要的测量，尽可能纠正标注医生的主观认知偏差以及偶然的标注失误。审核完成后的结果，录入数据库，作为参考标准使用。

A-4 样本量的估计

其样本量估算公式为：

$$n_T = n_C = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 [P_C(1 -) + P_T(1 - P_T)]}{(\Delta - |D|)^2}$$

P_T 、 P_C 分别为试验组和对照组预期事件发生率； $|D|$ 为两组预期率差的绝对值， $|D| = |P_T - P_C|$ ； Δ 为等效界值（适用于劣侧界值与优侧界值相等的情形），取正值。

H0: $|P_T - P_C| \geq \text{界值}$

H1: $0 \leq |P_T - P_C| < \text{界值}$

A-5 测试集偏倚分析

测试数据集选择可能是评估性能评价中偏差的一个主要来源，这是许多评估软件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。

参考文献

1. Greulich, W.W., Pyle, S.I.: Radiographic Atlas of Skeletal Development of the Hand and Wrist. Stanford University Press, Stanford, 2nd edition. (1959)
2. JM Tanner, MJR Healy, H Goldstein and N Cameron. Assessment of Skeletal Maturity and Prediction of Adult Height (TW3 Method) [M]. Published by WB Saunders, London, 2001. ISBN 0-7020-2511-9.
3. 中华人民共和国国家体育总局. 中国青少年儿童手腕骨成熟度及评价方法: TY/T 3001—2006[S]. 北京: 中国标准出版社, 2006。
4. 张绍岩等. 中国人手腕部骨龄标准——中华 05 及其应用[M]. 科学出版社, 2015 年 09 月
5. GB/T 9386-2008 计算机软件测试文档编制规范
6. 国家药品监督管理局医疗器械技术评审中心. 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, 2020