



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

征求意见稿

XXXX—XX—XX 发布

XXXX—XX—XX 实施

发 布

目 次

前 言 II

引 言 III

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 1

5 算法性能测试方法 3

附 录 A （资料性） X 射线骨龄测试集描述样例 6

附 录 B （资料性） 常用骨龄评估方法及缩略词..... 9

参 考 文 献 10

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

本文件作为方法标准，面向骨龄辅助评估的常见场景（骨发育等级评价、骨龄评价），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 X 射线骨龄辅助评估软件 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、具有骨发育等级评价、骨龄评估等后处理功能的手腕部骨龄辅助评估软件。本文件中骨龄特指手腕部骨龄。算法的输入为手腕部X射线影像。

本文件不适用于影像前处理及过程优化产品。

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。

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3部分：数据标注通用要求

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。

注：本文件中，骨龄特指手腕部X射线骨龄。

3.4

骨成熟度指征 *Indicators of bone maturity*

X射线影像中，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、骨骺的形态变化、骨骺与骨干的融合等。

3.5

骨发育等级 *bone maturity level*

骨龄评价方法（TW3、中华05等），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（A/B/C等）或阿拉伯数字（1/2/3等）标记。

4 算法性能测试要求

4.1 概述

算法测试过程宜参照YY/T 1858-2022中4.1的要求，建立测试文档。

4.2 测试环境

测试环境宜参照YY/T 1858-2022中4.2的要求。

4.3 测试资源

4.3.1 测试集要求

测试集的质量应满足 YY/T 1833.2的要求。测试集应独立于算法训练、调优过程，保证封闭性和安全性。测试人员可根据产品预期用途和临床应用场景，对测试集进行限定。

4.3.2 数据采集要求

4.3.2.1 影像质量要求

X射线骨龄影像应具备以下要求：

- 测试集采用的影像数据预先开展去标识化，可参考 GB/T 39725-2020 条款 10.2 的方法，例如去除姓名、证件号、联系方式、医保账户、银行账号；
- X 射线影像应具备足够分辨率和对比度，以清晰显示骨化中心、骨骺、骨干等手腕部的骨成熟度指征。
- 测试集中的影像数据满足待测产品的适用范围。测试数据应包含体现数据多样性的字段，包括但不限于患者年龄、患者性别、厂商型号、管电压等信息；
- 拍摄时，应避免手腕部运动、佩戴金属饰品等情形引起的 X 射线伪影；
- 拍摄时，应优先选用非利手（通常左手），若非利手有伤残，可更换利手拍摄；
- 拍摄时，手掌面向下，紧贴探测器，五指应自然分开。除手腕部关节的各块骨，还应包括桡、尺骨远端骨干 2cm-4cm。

注：附录A.1给出了测试集描述的样例。图A.1给出了X射线骨龄影像示例。

4.3.2.2 测试集样本量要求

骨龄评价属于回归任务，骨龄预测值一般为连续数值输出。若待测软件输出骨龄存在文字表述情形，如“成人”、“骨发育成熟”、“超出骨龄下限/上限”，产品制造商应给出该文字表述对应的骨龄数字界值。测试集样本量的主要决定因素为临床可接受的骨龄评价容许误差和总体标准差。本文件推荐采用置信区间法进行测试集样本量估算，以确保模型性能评估的可靠性和泛化能力。样本量估算公式为[12]：

$$N = \left(\frac{Z \times \sigma}{\delta} \right)^2 \dots\dots\dots (1)$$

式中：

- Z —— 标准正态分布的界值，95%置信水平对应的双侧Z值为1.96；
- σ —— 总体标准差，可根据样本标准差估计。如根据样本计算出骨龄值的标准差；
- δ —— 临床可接受的骨龄评价容许误差，不宜超过0.5年；

4.3.2.3 测试集多样性要求

测试数据集宜体现待测产品适用范围和临床使用场景内的样本多样性，包括成像设备维度、患者维度。

成像设备维度应考虑：

- 设备厂商，型号的多样性；
- 成像参数的多样性，包括但不限于层厚，管电压等；

患者维度应考虑以下内容：

- 性别、年龄多样性；
- 不同地域，不同医疗机构；
- 采集场景多样（如体检，门诊，急诊等场景）；

4.3.2.4 扩增数据要求

宜参照YY/T 1858-2022中4.3.4对测试集进行扩增；扩增数据的参考标准通过人工确认后交付使用。

4.3.3 测试工具

如适用，算法测试使用的体模，宜参照YY/T 1858-2022中4.3.5的要求。

4.4 测试平台

如使用测试平台进行算法测试，测试平台宜满足YY/T 1858-2022中4.4的要求。

4.5 测试指标与通过准则

测试指标与通过准则的选取，宜参照执行YY/T1858-2022中4.5的一般要求，根据产品的适用功能选择测试指标。

测试计划应根据产品的预期用途制定，按照产品对骨龄的评价方式（TW3法、中华05法）适用的信息，作为开展测试的依据。

4.6 测试流程要求

测试人员宜参照YY/T1858-2022中4.6的一般要求，根据测试计划开展测试活动，建立测试记录。

4.7 测试结果

测试人员宜参照YY/T1858-2022中4.7的一般要求，描述测试结果。

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 骨发育等级评价场景

具有骨发育等级评价功能的产品，应测试其骨发育等级评价误差。该指标适用于支持TW3法中华05法等逐块骨给出发育等级评价的产品。发育等级的具体范围，随手腕骨类型、方法规定有差异，需遵循对应骨龄评价方法资料的规定。例如，中华05法的骨发育等级为1级至14级。TW3法骨发育级标记为A, B, C, ..., H, I共9级，在评测前，应映射为阿拉伯数字1至9级。

5.1.1.1 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：

- 厂商需明确其产品所遵循的骨龄评价方法（附录 B），以匹配参考标准采用的骨龄评价方法。
- 向待测算法输入测试集，得到算法输出的骨发育等级结果。该结果的格式宜与参考标准兼容。
- 骨发育等级评价的算法性能指标，按 5.1.1.2 和 5.1.1.3 描述的公式进行计算。

5.1.1.2 平均绝对误差

平均绝对误差用MAE (Mean Absolute Error)表示，表达式如下

$$MAE = \frac{1}{N} \sum_{i=1}^N |Y - \hat{Y}| \dots\dots\dots (2)$$

式中：

N —— 样本量

Y —— 骨发育等级参考标准

$\hat{Y}$ —— 骨发育等级预测值

5.1.1.3 准确率

骨发育等级的评价也可理解为分类任务：算法输出等级与参考标准骨发育等级一致，为正确；否则为错误。准确率用Accuracy表示，表达式如下

$$Accuracy = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{Y = \hat{Y}\} \dots\dots\dots (3)$$

式中：

N —— 样本量

Y —— 骨发育等级参考标准

$\hat{Y}$ —— 骨发育等级预测值

$\mathbb{1}$ —— 指示函数，此处预测骨发育等级等于参考标准，为 1，否则为 0。

5.1.2 骨龄评价场景

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：

- 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。
- 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。
- 骨龄评价的算法性能指标，按 5.1.2.2 和 5.1.2.3 描述的公式进行计算。

5.1.2.2 平均绝对误差

平均绝对误差用 MAE (Mean Absolute Error) 表示，表达式如下

$$MAE = \frac{1}{N} \sum_{i=1}^N |Y - \hat{Y}| \dots\dots\dots (4)$$

式中：

N —— 样本量

Y —— 骨龄参考标准

$\hat{Y}$ —— 骨龄预测值

5.1.2.3 均方根误差

均方根误差用 RMSE (Root Mean Square Error) 表示，表达式如下

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y - \hat{Y})^2} \dots\dots\dots (5)$$

式中：

N —— 样本量

Y —— 骨龄参考标准

$\hat{Y}$ —— 骨龄预测值

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，对算法性能依据进行扩展测试。

5.2.2.2 面向硬件变化的鲁棒测试

测试人员应考虑X射线成像硬件设备、参数设置的多样性，收集更多样图像，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、峰值管电压、电流时间积等。

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、平滑等。

5.2.2.4 压力测试

5.2.2.4.1 通用要求

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

5.2.2.4.2 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) 受试者年龄偏小(接近或达到所支持骨龄评估方法下限)、偏大(接近或达到骨龄评估方法上限)的影像；
- b) 有伪影，但满足骨龄阅片要求的影像；
- c) 有植入物或体表附属物干扰，但满足骨龄阅片要求的影像；
- d) 图像分辨率偏高、偏低的影像；
- e) 手腕部在图像中占比较小、占比很大的影像；
- f) 存在手腕部生理性变异的影像（如第五中节指骨短指骨）；
- g) 骨头显示为低密度影，背景显示为高密度影的影像；

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 5.1 进行多次测试，测试次数不宜低于 3 次，观察测试结果是否变化。

5.2.4 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、图像分辨率、图像格式等要素，如包含 100 张图像、分辨率为1024×1024、格式为 Dicom 3.0。

5.2.5 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) 在骨发育等级评级场景下，测试人员宜对每一种骨的误差（MAE, Accuracy）进行分组统计
- b) 在骨龄评价场景下，测试人员宜对不同年龄段、不同性别、不同骨龄方法（RUS骨龄, Carpal骨龄）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)
X 射线骨龄测试集描述样例

A. 1 概述

本样例以X射线测试集为具体案例，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X射线手腕部骨龄影像辅助分析的人工智能医疗器械软件产品，如TW3法骨龄评价、中华05法骨龄评价等。

A. 3 数据采集

数据采集一般考虑样本年龄、性别、地域、采集设备、拍摄参数等方面的多样性，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，可进一步细化或合并临近分组。图A. 1给出了规范拍摄的骨龄X射线影像示例。

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	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
DR机型	XX公司XX型号	XX 例
	XX公司XX型号	XX 例
	XX公司XX型号	XX 例
采集场所	体检	XX 例
	门诊	XX 例
	住院	XX 例

图A.1 骨龄 X 射线影像示例



A.4 数据集标注规则

标注工作主要依据参考文献 [8] [9] [10]进行。

A.4.1 标注流程

整体的标注流程如图 A.1 所示，含标注人员选拔、标注培训、准入考核、初标、审核等环节。

标注医生负责骨龄初标，宜具有影像科、儿童内分泌科、儿童保健科或相关科室 3 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。

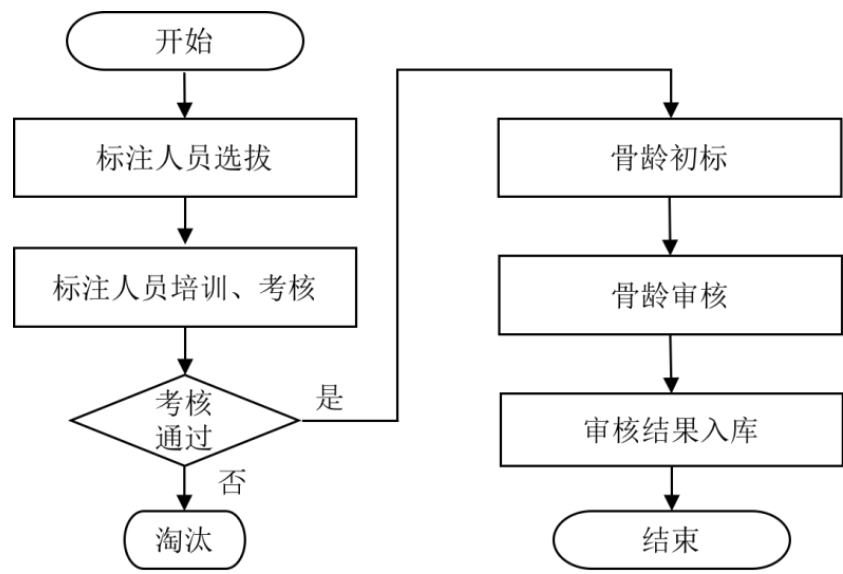
审核医生负责骨龄审核，宜具有影像科、儿童内分泌科、儿童保健科或相关科室 5 年以上骨龄阅片经验，且取得副高及以上职称。

标注医生和审核医生均需参与骨龄阅片培训，学习相应骨龄评价方法的权威资料。培训后，应安排骨龄阅片考核，根据医生在考核中的表现，决定是否录用。

A.4.2 标注一致性

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（或多标一审）的模式：即先由不同的标注医生标注同一份病例，再由审核医生给出该病例最终的骨龄评价。审核医生可见两位初标医生的标注结果，复核标准图谱和文字，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。



图A. 2 骨龄标注流程图

附录 B (资料性) 常用骨龄评价方法简介

B.1 概述

本附录给出常用骨龄评估方法的简介及缩略词定义，仅作为参考信息。

B.2 骨龄评价方法简介

B.2.1 骨龄评价方法 Bone Age Assessment Methods

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。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。明确采用的骨龄评价方法，是骨龄算法性能评测的前提条件。本节给出常用骨龄评估方法的简介，仅作为参考信息。

B.2.2 桡尺短骨 RUS Bones

RUS是桡骨(Radius)、尺骨(Ulna)、短骨(Short bones)的合称，其中短骨指掌骨和指骨。RUS骨共计13块骨（单侧手）。基于RUS骨的骨龄评价方法有TW3 RUS（属于TW3法），RUS-CHN(属于中华05法)。

B.2.3 腕骨 Carpal

骨龄评价方法中，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。基于腕骨的骨龄评价方法有TW3 Carpal（属于TW3法），TW3-C Carpal（属于中华05法）。

B.2.4 TW3法 Tanner and Whitehouse 3 Method

Tanner等在2001年基于欧美多国儿童的资料，制定了新的骨龄参考标准和评价图表，称为TW3法。不同于G-P图谱法，TW3法对13块RUS骨和7块腕骨，分别制定了一系列骨发育等级评估系统，分开计算RUS骨龄与腕骨骨龄，称为TW3-RUS和TW3-Carpal。

B.2.5 中华05法 CH05 Method

中华05法为张邵岩等采集21世纪初中国五城市儿童标准化样本17401名创建而成。抽样城市为上海市、广州市、温州市、大连市、石家庄市。依照TW3方法，中华05法制定了RUS和腕骨骨龄评价标准，称为TW3-Chinese RUS和TW3-Chinese Carpal。另外，在TW3-RUS骨发育等级的基础上，张邵岩等选择新的成熟度指征，提出了RUS-CHN法。

参 考 文 献

- [1] GB/T 9386-2008 计算机软件测试文档编制规范
- [2] IEEE 2802:2022, Standard for Performance and Safety Evaluation of Artificial Intelligence Based Medical Devices: Terminology
- [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》[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. 人工智能辅助检测医疗器械(软件)临床评价注册审查指导原则[Z]. 北京: 国家药品监督管理局医疗器械技术审评中心, 2023.
- [6] 原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: 原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] Greulich, W.W., Pyle, S.I.: Radiographic Atlas of Skeletal Development of the Hand and Wrist [M]. Stanford University Press, Stanford, 2nd edition, 1959
- [9] 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
- [10] 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
- [11] 张绍岩等. 中国人手腕部骨龄标准——中华 05 及其应用[M]. 北京: 科学出版社, 2015 年 09 月
- [12] 詹思延等. 流行病学[M]. 第 8 版. 北京: 人民卫生出版社, 2017 年 07 月