·诊治分析·

人工智能 TW3-RUS 与 TW3-Carpal 法对儿童骨龄评测的准确性比较

王柏森，王莹

【摘要】目的　探讨人工智能（AI）骨龄评测系统在 Tanner-Whitehouse 3（TW3）RUS（桡骨、尺骨、掌指骨）与 Carpal（腕骨）标准下的测定准确性，并分析其在不同年龄段及生长异常儿童中的适用性差异。**方法**　本研究为回顾性队列研究。纳入 2023 年 1 月至 2025 年 1 月于宁波市第六医院行左手腕正位 X 线检查的儿童 326 例，分别采用 AI 系统与两名高年资医师按 TW3-RUS 和 TW3-Carpal 标准评估骨龄，两名医师在双盲条件下独立评估，采用加权 *Kappa* 分析两名医师评价的一致性，以医师评估均值为“金标准”。按性别分男、女组，按骨龄“金标准”分为 4 ～ 9 岁组与≥10 岁组，根据临床资料分为正常组及疾病组。采用组内相关系数（*ICC*）评估分析 AI 与“金标准”的一致性，采用 *Pearson* 相关法分析 AI 与“金标准”之间的线性相关性，并计算 AI 骨龄的平均绝对误差（MAE）、误差≤0.5 年的符合率，并比较 TW3-RUS 和 TW3-Carpal 骨龄之间的差异。绘制 Bland-Altman 图，计算 AI 与“金标准”之间的 95%一致性界限。**结果**　两名医师一致性良好（$Kappa_{RUS}$=0.89，$Kappa_{Carpal}$=0.87）。整体样本中，AI 与“金标准”评估结果高度一致（均 $P < 0.05$），MAE 分别为 0.27 岁、0.33 岁。在不同性别分组中，TW3-Carpal、TW3-RUS 评估准确度相近（均 $P < 0.05$）。在不同的年龄分组中，4 ～ 9 岁组 TW3-Carpal 准确性更高，≥10 岁组 TW3-RUS 更优（均 $P < 0.05$）。正常组 TW3-RUS、TW3-Carpal 评估准确度均高于疾病组（均 $P < 0.05$）。**结论**　AI 骨龄系统在健康儿童中准确性高、评估效率显著优于人工。4 ～ 9 岁建议优先采用 TW3-Carpal 法，≥10 岁建议优先 TW3-RUS 法。对于生长异常儿童需结合临床指标进行修正。

【关键词】　儿童；骨龄；人工智能；生长发育障碍；Tanner-Whitehouse 法

DOI:10.3969/j.issn.1671-0800.2026.06.016

【中图分类号】　R455　**【文献标志码】**　A　**【文章编号】**　1671-0800(2026)06-0653-04

骨龄评测是对生长发育障碍相关疾病诊断与儿童生长监测的一种重要方法，通过骨龄能够判断出儿童骨骼的成熟度和生长潜能[1]。目前，临床上最常用到的是 Tanner-Whitehouse 3（TW3）法，其通过对尺骨、桡骨及短骨（radius，ulna and short bone，RUS）和腕骨（carpal bone）各骨化阶段做出量化评分，最后综合计分实现骨龄评估[2]。该方法准确度高，但人工评估耗时费力。近年来，人工智能（artificial intelligence，AI）在医学领域的应用与发展取得显著的成果，在骨龄评测方面亦有广泛普及与应用[3]。由于 TW3 法中包含 Carpal 和 RUS 两种骨龄标准，AI 辅助评估系统会同时提供这两种骨龄评测结果。但其在不同年龄段及生长发育异常儿童中的评估表现，仍缺乏系统研究。因此，本研究探讨基于 AI 系统 TW3-RUS 与 TW3-Carpal 法骨龄评估的临床价值及差异性，现报道如下。

作者单位：　315040 宁波，宁波市第六医院

通信作者：　王莹，Email：807995272@qq.com

1　资料与方法

1.1　一般资料　本文为回顾性队列研究。收集 2023 年 1 月至 2025 年 1 月 326 例于宁波市第六医院就诊且行左手腕 X 线正位片检查儿童的临床及影像资料。纳入标准：（1）年龄≤18 岁；（2）具有完整的临床资料，包括年龄、性别及病史等信息；（3）完成标准左手腕部 X 线正位片检查。排除标准：（1）不符合 TW3 法评估要求的 X 线图像；（2）存在手腕部骨折史；（3）存在先天性畸形病史。本研究获得宁波市第六医院伦理委员会批准（甬六医伦审 2025 论第 104 号），免除知情同意。

1.2　方法　采用日本岛津 2P324DK85 数字 X 线摄影设备拍摄左手后前位 X 线平片：（1）左手掌面紧贴平板，五指自然分开，拇指与食指约成 30°；（2）照射野包括完全的手腕部，尺、桡骨远端包括 3 ～ 4 cm。

1.3　图像后处理及参数分析　图像资料使用 AI 骨

龄评估系统（联影骨龄检测软件 uAI Discover-Bone Age，版本号 R001）处理，其自动识别 RUS 13 块骨与 Carpal 8 块骨感兴趣区，并根据 TW3 评分表自动输出骨龄值。由两名具有 10 年以上诊断经验的影像科医师（均经过 TW3 评分标准化培训）在不知晓 AI 结果及临床信息的条件下独立评估，采用加权 *Kappa* 分析医师间一致性。取两者平均值为“金标准”，若两者评分相差大于 0.5 岁，由第三名资深专家介入，共同复核达成共识后作为最终结果。

1.4　AI 系统在不同分组中的应用效能　分组方式：按性别分为男、女组；根据相关文献共识[4]，按骨龄“金标准”分为 4 ～ 9 岁组、≥10 岁组；按临床资料分为正常组及疾病组。分析 TW3-RUS、TW3-Carpal 标准的准确性。

1.5　统计方法　采用 SPSS 26.0 及 GraphPad Prism 9 软件进行统计分析。计量资料以均数±标准差表示，采用独立样本 t 检验；计数资料采用 χ^2 检验。计算 AI 骨龄的平均绝对误差（mean absolute error，MAE）及其 95%可行度区间（*CI*）、误差≤0.5 年的符合率；采用组内相关系数（*ICC*）分析 AI 与“金标准”的一致性；采用 *Pearson* 相关系数分析 AI 与“金标准”线性相关性。判定标准：*Kappa* ＞ 0.8 提示一致性极好，*ICC* ＞ 0.75 提示一致性良好，r ＞ 0.7 提示强正相关。采用配对 t 检验比较不同分组内 TW3-RUS 与 TW3-Carpal 骨龄的差异，并计算 Cohen’s d 效应量评估其临床显著性。绘制 Bland-Altman 图，计算 AI 与“金标准”之间 95%一致性界限。P ＜ 0.05 表示差异有统计学意义。

2　结果

2.1　临床资料　326 例中女 152 例，男 174 例；年龄 4 ～ 16 岁，平均（8.7±3.2）岁。按临床资料分组：正常组 206 例，生长异常组 120 例（包括性早熟 72 例，生长激素缺乏 37 例，其他如矮小症、先天性甲状腺功能减退共 11 例）。两组一般资料差异均无统计学意义（均 P ＞ 0.05），见表 1。

2.2　AI 系统与“金标准”的一致性、相关性及准确度　两名医师一致性良好（$Kappa_{RUS}$=0.89，$Kappa_{Carpal}$=0.87）。在 TW3-RUS 法中，AI 与“金标准”分别为（8.9±3.1）岁和（8.7±2.9）岁；在 TW3-Carpal 法中，分别为（8.7±3.0）岁和（8.8±2.8）岁。AI 与“金标准”TW3-RUS 骨龄的 *ICC* 为 0.972（95%*CI*：0.968 ～ 0.976），TW3-Carpal 骨龄 *ICC* 为 0.964（95%*CI*：0.959 ～ 0.969）。*Pearson* 相关分析显示，AI 与“金标准”呈高度正相关（均 P ＜ 0.05），两种评估方法间差异有统计学意义（t=1.56，df=325，P ＜ 0.05，Cohen’s d=0.28），见表 2。

Bland-Altman 分析显示，对于 TW3-RUS，AI 系统与“金标准”平均差异为+0.02 岁，95%一致性界限为−0.45 ～+0.49 岁；对于 TW3-Carpal，AI 与“金标准”平均差异为+0.04 岁，95%一致性界限为−0.62 ～+0.70 岁，见图 1。在 TW3-Carpal 分析中，有 1 例患儿 AI 系统骨龄与“金标准”差异＞ 1.5 岁，该患儿性早熟，生活年龄为 6.4 岁，骨龄片及 AI 系统分析结果见图 2。在 TW3-RUS 骨龄评估中，所有样本 AI 测值与“金标准”差异均＜ 1 岁。

2.3　不同性别、年龄组及临床分组的 AI 骨龄测定准确度　男女两组的准确度相近；在不同年龄组中，4 ～ 9 岁组 TW3-Carpal 准确度更高，而≥10 岁组 TW3-RUS 准确度更优；在临床分组中，正常组的 TW3-RUS 和 TW3-Carpal 骨龄准确度均高于疾病组，其 MAE 更低，0.5 年符合率更高，见表 3。

3　讨论

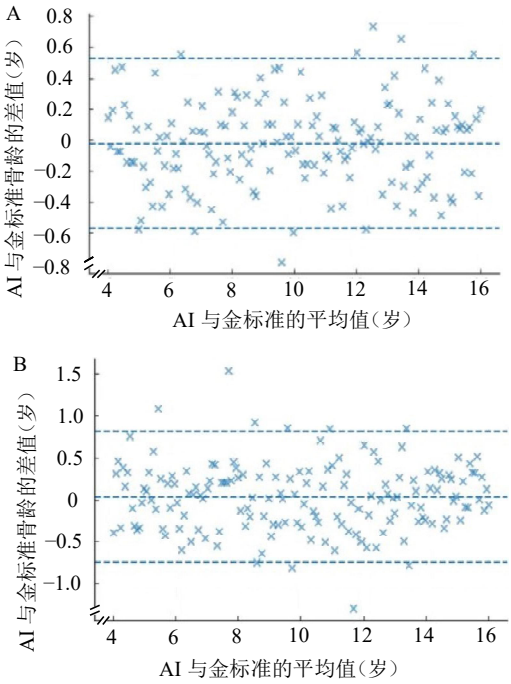
骨龄是儿童生长发育的重要指标之一，临床上主要通过非常用手和腕的 X 线片来判读。骨龄评估主要用于判断儿童发育是否正常、预测骨骼未来发育趋势及进行发育障碍的诊断、干预与随访。临床上骨龄评估主要方法可分为 3 类：计数法、图谱法和计分法[5]。前两者的方法较为简单，但准确度不高。计分法较为准确，但操作相对复杂，人工操作中需要投入较多时间与精力。近年来应用最广泛的计分法是由 Tanner 和 Whitehouse 提出的 TW3 法，其针对腕骨（Carpal）和尺骨、桡骨及短骨（RUS）的骨龄分别进行测量[6]。伴随着 AI 技术在医学领域的应用与发展，国内外已有多款 AI 骨龄评测软件上市并得到广泛应用。但是基于 TW3 法的 AI 辅助系统通常会

表 1　正常组与疾病组患儿基线特征对比

指标	正常组（n=206）	疾病组（n=120）	t（χ^2）值	P 值
性别（男/女，例）	110/96	64/56	（0.001）	＞ 0.05
生活年龄（岁）	8.6±3.1	9.0±3.4	0.812	＞ 0.05
金标准骨龄（岁）	8.72±3.05	9.18±3.38	1.258	＞ 0.05

在报告中同时提供 TW3-Carpal 和 TW3-RUS 两种骨龄结果，同一样本中两者常存在一定差异[7]。临床工作中如何更好理解应用这两个结果还有待进一步探索。因此，本研究回顾性分析了我院 AI 系统 TW3-RUS 与 TW3-Carpal 骨龄在男女儿童组、不同年龄段及是否存在发育相关疾病分组中的差异，以期为临床评估方法的选择提供参考。

本研究将年龄分组定为 4 ～ 9 岁与≥10 岁，主要基于文献中共识的骨化节律差异 [4]。结果显示，TW3-RUS 在≥10 岁组的 MAE 显著降低（0.24 岁），验证了该年龄切点的合理性。AI 系统 TW3-RUS 与 TW3-Carpal 法在整体样本组水平上均达到与“金标准”高度符合程度，Bland-Altman 图也提示两者一致性较好，与国内相关学者报道的结果相仿 [8]。AI 评估整体准确度较高，表明 AI 骨龄评估具有良好的可靠性和临床应用价值。此外，TW3-Carpal 法在 4 ～ 9 岁组儿童中准确度更高，TW3-RUS 法在≥10 岁组中更具优势。这可能与儿童生长过程中骨化模式相关[9]，低龄儿童腕骨处于活跃骨化阶段，骨化中心大且形态变化大，影像特征容易被 AI 模型识别出来，因而 TW3-Carpal 评分更敏感；而≥10 岁后步入青春期，长骨远端骺板快速发育、骨骺明显改变，RUS 评分所包含的信息量大幅增加。同时在青春期后期腕骨开始逐渐接近融合、形态变化趋于稳定，TW3-Carpal 区分度降低，与外国学者相关研究结果基本吻合[10]。性别分组结果现实，AI 评估男女儿童骨龄的准确度差异无统计学意义（$P > 0.05$），提示其在性别层面具有较高普适性。国内有学者基于深度学习的骨龄评估研究中也得到相同结论[11]，其结果显示 AI 模型在男女儿童都取得相关好的表现，且结果无差异。



注：A 为 TW3-RUS，B 为 TW3-Carpal

图 1　AI 系统与金标准骨龄间的 Bland-Altman 图



注：金标准骨龄 7.1 岁，AI 系统 TW3-Carpal 骨龄 8.7 岁，误差＞ 1.5 岁

图 2　性早熟女性患儿左手腕 X 线正位片及 AI 评测结果

本研究结果显示，在 TW3-RUS 和 TW-Carpal 法中，疾病组 AI 评测准确度均低于正常组（均 $P < 0.05$），与国内学者相关研究报道一致[12]。即性早熟

表 2　人工智能与“金标准”的一致性、相关性及准确度比较

方法	平均绝对误差（岁，95%*CI*）	0.5 年符合率（%，95%*CI*）	*ICC*（95%*CI*）	*r* 值	*P* 值
TW3-RUS	0.27±0.20（0.24 ～ 0.30）	89.4（86.8 ～ 91.9）	0.972（0.968 ～ 0.976）	0.923	＜ 0.05
TW3-Carpal	0.33±0.25（0.29 ～ 0.37）	82.5（79.0 ～ 85.6）	0.964（0.959 ～ 0.969）	0.938	＜ 0.05

表 3　不同性别、年龄及临床分组下人工智能骨龄测定的平均绝对误差与 0.5 年符合率比较

分组	男（*n*=174）	女（*n*=152）	4 ～ 9 岁（*n*=180）	≥10 岁（*n*=146）	正常组（*n*=206）	疾病组（*n*=120）
TW3-RUS MAE（岁，95%*CI*）	0.28（0.25 ～ 0.31）	0.26（0.23 ～ 0.29）	0.30（0.27 ～ 0.33）	0.24（0.21 ～ 0.27）	0.23（0.20 ～ 0.26）	0.34（0.30 ～ 0.38）
0.5 年符合率（%，95%*CI*）	89.1（85.0 ～ 92.3）	89.9（86.1 ～ 93.2）	86.0（81.0 ～ 89.9）	92.9（89.1 ～ 95.6）	91.2（87.6 ～ 93.9）	82.4（76.5 ～ 87.1）
TW3-Carpal MAE（岁，95%*CI*）	0.34（0.31 ～ 0.37）	0.32（0.29 ～ 0.35）	0.29（0.26 ～ 0.32）	0.38（0.34 ～ 0.42）	0.32（0.28 ～ 0.36）	0.35（0.31 ～ 0.39）
0.5 年符合率（%，95%*CI*）	81.3（77.3 ～ 85.6）	80.5（76.1 ～ 84.9）	89.6（85.4 ～ 92.8）	74.0（68.5 ～ 79.0）	87.2（82.3 ～ 91.5）	77.9（73.6 ～ 81.5）

注：MAE 为平均绝对误差

及矮小症患儿骨龄AI评估误差相对偏大，其原因可能是骨龄发育节律出现偏差异常，而现有AI模型的训练资料中未涵盖此类特殊发育规律。本研究中出现 1 例 AI 系统 TW3-Carpal 骨龄误差＞ 1.5 岁的性早熟患儿。经专家组复核发现，由于该患儿性激素提前分泌，导致头状骨与钩骨骨化中心不仅体积增大，且边缘出现病理性的假性“成熟”切迹。AI 模型捕捉到这些异常纹理特征后将其判定为成熟等级 G 级（实际应为 E 级），从而导致严重过高评估。相比之下，其 TW3-RUS 误差仅为 0.6 岁，显示长骨受此类激素波动的形态干扰较小。这说明 AI 系统整体表现虽然可靠，但在生长发育异常儿童中，AI 评估结果有可能存在较大偏差。国外有学者提出的 AI（Bone Xpert）系统在大样本训练后，对患有生长发育障碍疾病儿童骨龄评估具有较高的准确性[13]。Bland-Altman 图显示 AI 评测骨龄误差大多介于±0.5 岁之间，符合临床精度需求，与国内有学者研究报告的平均误差约 0.32 岁相近[14]，说明国内研发 AI 系统已接近国际先进水平。但是，和部分国外研究报道的 MAE ＜ 0.2 岁结果相比还存在一定差距[15]，原因可能与训练数据量不足、算法迭代不够及影像质量等多方面相关。因此，对于临床诊断为生长发育障碍的儿童，AI 结果应仅视为预判结果。医师应重点观察 Carpal 法下各骨块是否存在非典型骨化，并结合 RUS 结果进行综合修正，必要时回归原始图谱复核。

本研究的局限性如下：（1）本研究为单中心回顾性研究，可能存在选择偏倚；（2）本研究虽补充了基线比较，但疾病组内部亚组（如甲状腺功能减退、Turner 综合征等）样本量仍不足。未来需开展多中心前瞻性研究，重点针对生长发育障碍亚组进行模型迁移学习训练，进一步提升 AI 系统在真实临床环境中的评估准确性；（3）本研究基于 AI 系统仅对 TW3 评分法进行研究分析，将来应加入中华 05 骨龄法（RUS-CHN）等方法[16]进一步验证。

利益冲突　所有作者声明无利益冲突

作者贡献声明　王柏森：实验操作，论文撰写，数据整理；王莹：统计学分析，研究指导，论文修改

参　考　文　献

[1] 白凤森, 袁新宇, 马毅民, 等. 基于调查问卷分析国内儿童骨龄评估现状及发展趋势[J]. 中华放射学杂志, 2024,58(2):225-228.

[2] 袁新宇, 潘诗农, 程晓光. 构建儿童骨龄核心数据建立中国儿童骨龄标准体系[J]. 中华放射学杂志, 2023,57(4):341-343.

[3] 张义, 朱文文, 李凯, 等. 人工智能在儿童骨龄影像评估的应用进展[J]. 中华放射学杂志, 2023,57(4):420-423.

[4] Lee M, Choi Y H, Lee S B, et al. Retrospective clinical trial to evaluate the effectiveness of a new tanner-whitehouse-based bone age assessment algorithm trained with a deep neural network system[J]. Diagnostics (Basel), 2025, 15(8): 993.

[5] Suh J, Heo J, Kim S J, et al. Bone age estimation and prediction of final adult height using deep learning[J]. Yonsei Med J, 2023, 64(11): 679-686.

[6] Chang T Y, Chou T Y, Jen I A, et al. Artificial intelligence algorithm improves radiologists' bone age assessment accuracy[J]. J Chin Med Assoc, 2025, 88(7): 530-537.

[7] Liu Y Q, Ouyang L J, Wu W, et al. Validation of an established TW3 artificial intelligence bone age assessment system: A prospective, multicenter, confirmatory study[J]. Quant Imaging Med Surg, 2024, 14(1): 144-159.

[8] 杨洋, 王驰, 欧凌, 等. 真实世界中基于人工智能骨龄测量系统TW3-Carpal和TW3-RUS的临床应用差异[J]. 中华放射学杂志, 2023,57(4):359-363.

[9] Huang S R, Su Z, Liu S Y, et al. Combined assisted bone age assessment and adult height prediction methods in Chinese girls with early puberty: Analysis of three artificial intelligence systems[J]. Pediatr Radiol, 2023, 53(6): 1108-1116.

[10] Zadoo N, Tak N, Reddy A J, et al. Enhancing pediatric bone age assessment using artificial intelligence: Implications for orthopedic surgery[J]. Cureus, 2025, 17(2): e79507.

[11] Liang Y, Chen X B, Zheng R X, et al. Validation of an AI-powered automated X-ray bone age analyzer in Chinese children and adolescents: A comparison with the tanner – Whitehouse3 method[J]. Adv Ther, 2024, 41(9): 3664-3677.

[12] 常沙, 闫东, 杜霞, 等. 基于深度学习的骨龄评估系统对生长发育异常儿童骨龄评测的准确性[J]. 中华放射学杂志, 2023,57(4):364-369.

[13] Thodberg H H, Thodberg B, Ahlkvist J, et al. Autonomous artificial intelligence in pediatric radiology: The use and perception of Bone Xpert for bone age assessment[J]. Pediatr Radiol, 2022, 52(7): 1338-1346.

[14] Zhao X, Zhang M, Cheng M Y, et al. Construction of artificial intelligence system of carpal bone age for Chinese children based on China-05 standard[J]. Med Phys, 2022, 49(5): 3223-3232.

[15] Pape J, Hirsch F W, Deffaa O J, et al. Applicability and robustness of an artificial intelligence-based assessment for Greulich and Pyle bone age in a German cohort[J]. Rofo, 2024, 196(6): 600-606.

[16] Yang C, Dai W, Qin B, et al. A real-time automated bone age assessment system based on the RUS-CHN method[J]. Front Endocrinol, 2023, 14: 1073219.

收稿日期：2026-02-03

（本文编辑：吴迪汉）