International Journal of Legal Medicine
https://doi.org/10.1007/s00414-026-03946-3

RESEARCH

Does increased staging complexity influence expert agreement in dental age estimation?

Nikolaos Angelakopoulos[1,2] · Rizky Merdietio Boedi[3,4] · Stefano De Luca[5] · Nikita Polukhin[6] · Akiko Kumagai[7] · Ivan Galic[8] · Jeta Kelmendi[9] · Israel Soriano Vázquez[10] · Aida Hadzic Selmanagić[11] · Jorge Pinares Toledo[12] · Roberto Scendoni[13] · Daniel Franklin[14] · Roberta Araujo Pinto[2] · Lourdes Gómez Jiménez[15] · Arjeta Neziri[16] · Emerson Rodrigues[2] · Pisha Pittayapat[17] · Pedro Manuel Garamendi-González[18] · Joe Adserias-Garriga[19] · Bernhard Knell[20] · Ademir Franco[2,21]

Received: 11 May 2026 / Accepted: 1 August 2026

Abstract
Dental age estimation (DAE) is a well-established procedure in current practice, and it is frequently required in forensic and legal investigations. This study aimed to evaluate the intra- and inter-observer reliability of dental development staging among forensic practitioners using two widely applied staging systems, according to Demirjian et al. 1973 and Moorrees et al. 1963. The study included 250 anonymized panoramic radiographs of individuals between 6 and 14 years (median age, 10 years), comprising 51.2% males and 48.8% females. Eighteen observers with prior training in DAE independently assessed tooth development on panoramic radiographs using the Demirjian and Moorrees methods, with each method evaluated in two separate rounds. Intra-observer reliability was evaluated using weighted and unweighted Cohen's Kappa, while inter-observer agreement was assessed using Gwet's AC1 and AC2 coefficients. Weighted analyses accounted for the ordinal structure of developmental stages, whereas unweighted analyses reflected exact agreement. Intra-observer reliability was consistently high across both systems, with median weighted Cohen's Kappa values ranging from 0.897 to 0.957 and median unweighted Cohen's Kappa values ranging from 0.836 to 0.944. Inter-observer agreement values averaged across rounds assessed with Gwet's AC2 ranged from 0.914 to 0.987, whereas mean Gwet's AC1 ranged from 0.306 to 0.908, indicating that disagreement was primarily due to adjacent-stage differences. Moorrees staging showed slightly higher reliability under weighted conditions, while Demirjian demonstrated higher exact agreement under unweighted conditions. Agreement increased with the number of observers but reached a plateau at approximately 8–10 observers, suggesting an optimal number of observers for forensic consensus. The results indicated that difficulty in distinguishing adjacent stages, unclear stage boundaries, and overlapping anatomical structures were the main sources of uncertainty. The Moorrees system was perceived as more complex, while Demirjian showed difficulty focused on the root development stages. These findings highlight the importance of methodological considerations in dental age estimation and suggest that refinement of staging systems may further improve the reliability of expert assessment.

Keywords Age · Forensic Dentistry · Radiograph · Reliability · Teeth

Introduction

Dental age estimation (DAE) for living individuals is of considerable importance in forensic, legal, humanitarian, and clinical contexts. Its use, within multidisciplinary age estimation procedures for asylum seekers and undocumented minors without reliable documentation, carries substantial ethical and legal implications, as age estimation outcomes may directly influence an individual's legal status, access to protection, and treatment under immigration law [1]. In such high-stakes contexts, the reliability and reproducibility of age estimation methodologies are crucial, as they form the basis for age estimation, the outcome of which can have life-altering consequences for individuals alleged to be minors [2].

Extended author information available on the last page of the article

A range of different methodologies has been developed over the years, each characterized by distinct staging criteria, a varying number of developmental stages, and selection of targeted teeth [3, 4]. Although international standards, guidelines and recommendations, such as those issued by the European Asylum Support Office (EASO) [5], the Study Group on Forensic Age Diagnostics of the German Society of Legal Medicine (AGFAD) [6], the International Organization for Forensic Odonto-Stomatology (IOFOS) recommendations [7], the American Board of Forensic Odontology (ABFO) guidelines [8] and the Organization of Scientific Area Committees (OSAC) standards [9], do exist, their implementation has been gradual. In practice, this broad spectrum of methodological approaches has been used among medicolegal institutions [2, 10–15], even within the same geographic regions. Moreover, heterogeneity in expert training, professional experience, and adherence to standardized assessment protocols may further exacerbate variability, raising concerns regarding the overall reproducibility of DAE across forensic odontologists.

Previous research has highlighted that methods with fewer, broader stages, such as Demirjian's 8-stage method [16], generally yield higher inter-observer agreement due to their ease of application and reduced subjectivity in stage assignments. In a recent study [17], Demirjian's method applied to the seven left mandibular and maxillary teeth exhibited overall high inter- and intra-observer reliability, although assessment variability was observed for the maxillary teeth, particularly the permanent maxillary left lateral incisor.

In contrast, methods with more detailed staging systems, such as Nolla's 10-stage method [18] and the Moorrees system [19] (which describes up to 13–14 stages), offer greater categorical granularity, which can theoretically improve the description of dental development. However, this increased granularity may come at the cost of greater subjectivity and complexity, particularly when distinguishing between adjacent stages, thereby potentially reducing inter-observer agreement [20]. In this context, the observer's familiarity with the technique must also be considered [21]. This trade-off between detail and reliability is especially important to investigate by means of scientific research.

A high-quality reference standard for DAE should be based on large, population-based, age-balanced samples, using clear and reproducible staging criteria. However, many of the foundational studies used to develop these staging systems vary in quality, and few provide robust evaluations of observer agreement across different levels of expertise [20, 22, 23].

The present study aims to evaluate inter- and intra-observer agreement among eighteen qualified forensic odontologists and legal age estimation experts in the staging of dental development using the widely applied Demirjian eight-stage method [16] and the multi-stage Moorrees method [19]. Specifically, the study investigates whether the greater number of developmental stages in these methods affects the reliability of assessments among observers and examines the influence of factors such as observer experience on agreement levels. In addition, the study seeks to explore observers' perceived challenges and experiences through a structured post-assessment questionnaire, to further elucidate the impact of staging complexity on confidence, usability, and reproducibility in DAE.

Materials and methods

This study was conducted to test the null hypothesis (H0) that expert consistency is not affected by system complexity, namely between Demirjian's and Moorrees' staging systems. The alternative hypothesis (H1) is that increased staging complexity can reduce expert consistency for stage allocation in DAE.

Data collection

This observational cross-sectional study was conducted using panoramic radiographic images obtained from a previously established, anonymized dataset that has been used in prior scientific publications. No new data collection, database access, or patient contact was involved. No subjects were exposed to radiation solely for the purpose of this research. The selected sample consisted of 250 anonymized panoramic radiographs from subjects aged 6–14 years (median age, 10 years), consisting of 51.2% males and 48.8% females. Anonymization of the panoramic radiographs was achieved by extracting only the data relevant to this study, including the radiograph identification number, sex, date of birth, and date of exposure. All panoramic radiographs were assessed in their exported .jpeg format. As the images were retrospectively collected from the institutional archive, the exported image dimension/resolution varied according to the original acquisition and export settings.

Panoramic radiographs were included if all seven left mandibular permanent teeth (Fédération Dentaire Internationale [FDI] #31–37) [24] were present. When visible, the left mandibular third molar (FDI #38) was also included in the assessment to capture its early developmental stages in molars. Only teeth exhibiting typical morphological characteristics and a continuous, uninterrupted pattern of mineralization consistent with established standards of dental development were considered eligible. Hence, radiographs exhibiting dental anomalies, supernumerary teeth, or any conditions that could potentially affect normal dental development, including tumors, fractures, or pathological lesions, were excluded. In addition, children with a history of orthodontic treatment were excluded. Furthermore, radiographs with poor image quality, including significant distortion or blurring that could impair accurate staging of tooth development, were also excluded. This study received ethical approval from the Health Research Ethics Committee of the Faculty of Medicine, Universitas Diponegoro (Approval No. 212/EC/KEPK/FK-UNDIP/VIII/2025). The study was carried out under the ethical standards laid down by the Declaration of Helsinki (Finland) and its later amendments [25].

Minimum sample size calculation

Sample size was estimated according to Gwet [26] for a fully crossed design with a 0.1 margin of error at 95% confidence, yielding a minimum of 168 radiographs for the inter-observer and 175 for the intra-observer evaluation. As Gwet's tabulated parameters cap raters and categories at five, this represents a conservative estimate given the present study's 18 observers and 8–14 stage categories; precision of the agreement estimates is expected to improve further with the larger number of observers and categories employed [27]. The final sample of 250 radiographs therefore ensured robust estimates for reproducibility analyses.

Data management

Each panoramic radiograph was anonymized and assigned a unique identification number, according to ethical standards. To calculate the chronological age (CA) of each subject, the date of birth was subtracted from the date of radiograph exposure, and CA was expressed in decimal years. Detailed information for each radiograph, including identification number, sex, date of birth, date of radiograph exposure, and CA, was meticulously recorded using Microsoft Excel 2016 for data management.

Assessment systems and the rationale for their selection

Tooth development was assessed using two staging frameworks selected to span the extremes of developmental resolution. The Moorrees et al. [19] system relies on fractional crown- and root-length estimates supported by morphological landmarks (i.e., cusp tip appearance, crown completion, half-root length, and apex closure) and resolves crown and root development into 13 stages for anterior and 14 stages for posterior teeth. The Demirjian et al. [16] system, in contrast, defines eight stages per tooth based on discrete morphological criteria and relative crown/root proportions, supported by schematic drawings and radiographic exemplars.

Selecting both frameworks therefore enabled assessment of the same dentition across the minimum and maximum spectrum of staging detail. The Moorrees system subdivides crown and root development into a larger number of intermediate (transitional) substages — that is, additional fractional steps between crown completion and apex closure (e.g., between root initiation and one-quarter root length, and between three-quarter root length and apex closure) — whereas Demirjian aggregates these intervals into broader, discretely defined categories. All stages were recorded in a structured Excel spreadsheet for downstream analysis.

Participants and calibration

Eighteen observers participated in this study, consisting of fifteen qualified forensic odontologists and three forensic experts with backgrounds in forensic anthropology and legal medicine, all of whom had documented DAE experience in both clinical and research domains. Eligibility was based on academic publications with DAE as a main topic within the preceding two years and experience with real-world DAE casework for the clinical component. To minimize order and fatigue effects, the 250 radiographs were presented in an independently randomized sequence for each observer, and observers were advised to limit individual sessions to 30–60 min. The observers were blinded to the CA and all identifying information of the individuals throughout the staging process.

All observers were given detailed instructions to assess the radiographs and apply the staging process according to the Demirjian 8-stage [16, 28] and Moorrees multi-stage systems [19]. To ensure independent evaluation of the two staging systems, each observer completed two separate staging rounds for each method, resulting in a total of four staging rounds per observer. The Demirjian and Moorrees assessments were performed independently, and observers were not aware of the stage assigned using one method while evaluating the same image with the other method. An Excel workbook containing four separate worksheets, corresponding to the four staging rounds, was used to record the observations. Each spreadsheet listed an identification number for each tooth radiograph, along with a column for recording the Demirjian [16] and Moorrees stages [19]. Furthermore, calibration instructions were provided to each observer to ensure a standardized application of each method.

First, when there was uncertainty about stage assignment, or a tooth fell between two stages, the observer was instructed to select the lower stage. This conservative "benefit of the doubt" rule is a recognized good practice in age estimation based on growth and development, as it avoids overestimation of maturity and aligns with the minimum-age principle that should guide every practitioner.

Second, when a tooth exhibited two different stages, particularly in multi-rooted teeth (e.g., when the distal root of a molar was classified as R1/4 and the mesial root as R1/2), the lower stage was assigned. Third, the observation was made on the third quadrant (mandibular left teeth, FDI 31–38) only; hence no changes were made to the opposite side when a tooth could not be staged. Following this, if a tooth could not be staged, Not Applicable (NA) was recorded in the form. Finally, observers were permitted to use any image-processing software (e.g., ImageJ) to assist with the assessment, including adjusting image contrast and brightness according to their individual preferences. Image regions could also be magnified to facilitate detailed assessment.

Post-assessment questionnaire

Following completion of the second round of staging assessments, all participating observers were invited to complete a questionnaire designed to evaluate their experience during the evaluation process. This was conducted using a self-administered questionnaire created with Microsoft Forms (https://forms.office.com/Pages/). The survey aimed to capt ure perceived difficulties, sources of uncertainty, and overall usability of the applied staging systems, with particular emphasis on the impact of staging complexity on confidence and decision-making. The questionnaire included both closed-ended (Likert-scale and multiple-choice) and open-ended questions addressing factors such as stage differentiation, image interpretation, fatigue, and system preference. The complete questionnaire used in this study is provided as Supplementary Questionnaire S1. The collected responses were analyzed descriptively and used to complement the quantitative assessment of inter- and intra-observer agreements. All data were collected anonymously and stored securely, ensuring that participants' personal information was protected. Any data that could potentially reveal their identity was handled with the utmost confidentiality.

Statistical analysis

Because Demirjian [16] and Moorrees [19] tooth staging systems differ in their number of developmental stages, intra- and inter-observer reliability was assessed with both weighted and unweighted coefficients so that the systems could be compared fairly. Weighted analyses accounted for the ordinal nature of tooth developmental stages, whereas unweighted analyses were used to assess exact agreement, without reducing the contribution of adjacent-stage disagreements that may be less influential in more detailed staging systems such as Moorrees [19].

Intra-observer reliability was assessed using Cohen's Kappa (both weighted and unweighted) [29, 30] for each observer by comparing staging consistency between Round 1 and Round 2.

Inter-observer reliability was evaluated using Gwet's agreement coefficients (unweighted AC1 and weighted AC2) [31]. These coefficients were chosen due to their robustness to the "Kappa paradox". Gwet's AC1 and AC2 were preferred over Fleiss' Kappa because chance-corrected Kappa statistics behave paradoxically when the marginal distribution of categories is highly skewed. This paradox occurs when there is a very high observed agreement that can nevertheless yield unexpectedly low Kappa values. It is common in dental staging within a narrow age range, where a few developmental stages predominate for each tooth and the stage distribution is markedly unbalanced. Gwet's coefficients estimate the probability of chance agreement in a way that is far less sensitive to this prevalence and marginal imbalance and therefore provide more stable and interpretable estimates of true observer agreement under the conditions of the present study.

For clarity and readability, intra-observer reliability was summarized as the median across observers, reflecting the distribution of agreement estimates obtained from all observers. Inter-observer reliability was summarized as the mean of the two assessment rounds, providing a single representative estimate per tooth. Individual observer-specific estimates and round-specific results are provided in the Supplementary Material.

"Not applicable", where observers were not able to assign a tooth developmental stage, and blank responses, where the tooth was not visible in the image, were excluded from analysis. Agreement on NA responses was analyzed separately using Gwet's AC1.

The interpretation of Cohen's Kappa and Gwet's AC was based on the benchmarking scale proposed by Landis and Koch [32]: values < 0 indicated poor agreement, 0.00–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect agreement.

To assess statistical uncertainty and compare agreement across conditions, a non-parametric bootstrap approach was applied. The dataset was resampled with replacement (1000 iterations) maintaining the original sample size. Differences in agreement (Δ) were evaluated using bootstrap resampling for differences between rounds (Round 2 - Round 1) and between staging systems (Demirjian - Moorrees). Results were summarized as mean differences with 95% confidence intervals.

To assess how inter-observer reliability depends on the number of observers, observer subsets ranging in size from 2 to 18 were evaluated. All possible combinations of observers were analyzed when their total number did not exceed 1000. Otherwise, a random sample of 1000 combinations was used to limit the computational burden associated with combinatorial growth.

Descriptive statistics included an initial assessment of normality of distribution of continuous variables using the Shapiro–Wilk test. Continuous variables were described as medians with interquartile ranges (Me (IQR)). Categorical variables were summarized as counts and proportions (%).

All statistical analyses were performed in R version 4.5.2. Key packages included "vcd" for Cohen's Kappa, "irrCAC" for Gwet's agreement coefficients, "dplyr" and "tidyr" for data manipulation, and "ggplot2" for visualization.

Results

Eighteen observers participated in the study. The distribution of assigned tooth developmental stages was skewed and is presented in Supplementary Tables S1 and S2. The proportion of teeth that were reported to be not visible on the radiographs was below 0.1% for all teeth except FDI38, for which approximately one-third of responses were blank, reflecting the expected absence of the tooth in a substantial proportion of cases within the studied age range.

Intra-observer reliability was high across both staging systems (Table 1), with median weighted Cohen's Kappa values ranging from 0.897 to 0.957 and median unweighted Cohen's Kappa values ranging from 0.836 to 0.944. The weighted Cohen's Kappa was slightly higher for the Moorrees staging system, whereas the unweighted Cohen's Kappa was slightly higher for the Demirjian system; under both weighting schemes the median values remained within the almost-perfect agreement range and the differences between systems were small.

Supplementary Tables S3, S4, S5 and S6 show the full Cohen's Kappa estimates with 95% confidence intervals across observers, teeth, and methods. Weighted Cohen's Kappa values generally exceeded the unweighted estimates, particularly where intra-observer reliability was lower, and variability was greater. Although almost perfect agreement was observed, some instances of substantial agreement (0.61–0.80) were noted. However, confidence intervals largely overlapped across methods, indicating comparable robustness and intra-observer reliability of both the Demirjian and the Moorrees staging systems.

Gwet's AC showed different inter-observer reliability depending on the weighting technique.

Mean Gwet's AC2 ranged from 0.914 to 0.987, reaching almost perfect agreement across all teeth (Table 2). The highest Gwet's AC2 values were observed for FDI31, FDI32, and FDI36 in both staging systems. However, the bootstrapped-derived differences demonstrated in Supplementary Table 7 were small, and their confidence intervals mostly included zero, indicating no consistent advantage of either method. The largest and significant negative difference values (favoring Moorrees) were observed for FDI37 and FDI34, although difference sizes remained small and ranged between −0.022 and −0.027.

Mean unweighted Gwet's AC1 values were lower overall, ranging from 0.306 to 0.908. Demirjian consistently showed higher Gwet's AC1 values than Moorrees across all teeth in both rounds. The values of difference in reliability between staging systems were significantly positive (favoring Demirjian) for all teeth, ranging from 0.066 to 0.165 in round 1 and from 0.078 to 0.147 in round 2 (Supplementary Table S8).

Both weighted and unweighted Gwet's AC remained stable across rounds with only minor significant bootstrap-derived differences between repeated measurements for FDI33 measured with AC2 and for FDI37, FDI36, and FDI33 measured with AC1; in each case, agreement was slightly higher in the first round with differences not exceeding −0.037.

Table 1 Intra-observer reliability for each tooth assessed using weighted and unweighted Cohen's Kappa, presented as median (IQR) across observers

Tooth	Weighted		Unweighted	
	Demirjian	Moorrees	Demirjian	Moorrees
FDI38	0.926 (0.845–0.958)	0.938 (0.898–0.955)	0.876 (0.743–0.932)	0.836 (0.738–0.877)
FDI37	0.947 (0.860–0.960)	0.945 (0.867–0.971)	0.903 (0.786–0.934)	0.860 (0.690–0.917)
FDI36	0.931 (0.873–0.965)	0.897 (0.807–0.958)	0.922 (0.870–0.963)	0.849 (0.713–0.929)
FDI35	0.929 (0.865–0.960)	0.938 (0.881–0.974)	0.882 (0.768–0.937)	0.853 (0.713–0.932)
FDI34	0.932 (0.888–0.960)	0.946 (0.873–0.966)	0.898 (0.806–0.933)	0.867 (0.670–0.923)
FDI33	0.930 (0.880–0.964)	0.928 (0.853–0.967)	0.895 (0.810–0.946)	0.840 (0.651–0.915)
FDI32	0.957 (0.888–0.979)	0.934 (0.896–0.977)	0.944 (0.864–0.972)	0.897 (0.817–0.960)
FDI31	0.926 (0.844–0.954)	0.903 (0.814–0.982)	0.904 (0.826–0.948)	0.876 (0.743–0.972)

Table 2 Inter-observer reliability for each tooth assessed using weighted and unweighted Gwet's AC, presented as mean across two assessment rounds

Tooth	Weighted (AC2)		Unweighted (AC1)	
	Demirjian	Moorrees	Demirjian	Moorrees
FDI38	0.919	0.914	0.452†	0.306†
FDI37	0.928†	0.953†	0.533*,†	0.402†
FDI36	0.977	0.970	0.802*,†	0.662†
FDI35	0.938	0.950	0.565†	0.452†
FDI34	0.923†	0.952†	0.596†	0.465†
FDI33	0.926*	0.945	0.584*,†	0.439†
FDI32	0.959	0.971	0.829†	0.718†
FDI31	0.982	0.987	0.908†	0.837†

* Significant difference between assessment rounds for the corresponding tooth and staging system

† Significant difference between the Demirjian and Moorrees staging systems

Individual coefficients, bootstrap-derived differences, and corresponding 95% confidence intervals are reported in Supplementary Tables S7 (Gwet's AC2) and S8 (Gwet's AC1)

Table 3 Inter-observer reliability on NA responses for each tooth assessed using unweighted Gwet's AC, presented as mean across two assessment rounds

Tooth	Demirjian	Moorrees
FDI38	0.984	0.987
FDI37	0.998	0.997
FDI36	0.996	0.996
FDI35	0.995	0.994
FDI34	0.990*	0.992
FDI33	0.948*	0.950†
FDI32	0.907	0.910
FDI31	0.944	0.943

* Significant difference between assessment rounds for the corresponding tooth and staging system

† Significant difference between the Demirjian and Moorrees staging systems in the first round

Individual coefficients, bootstrap-derived differences, and corresponding 95% confidence intervals are reported in Supplementary Table S9

Overall, weighted agreement (Gwet's AC2) showed negligible differences between staging systems, whereas exact agreement (Gwet's AC1) was consistently higher for the Demirjian staging system across all teeth in both rounds.

The overall proportion of NA responses in two rounds was 1.7% for both staging systems. The highest proportion of NA responses in both Demirjian and Moorrees staging systems was observed for FDI32 (5.6% and 5.9%, respectively), FDI31 (3.0% and 3.2%), FDI33 (2.9% and 2.8%), and FDI38 (1.3% and 1.0%). For other teeth, the proportion of NA responses did not exceed 0.6% in both staging systems (Supplementary Tables S1 and S2).

As shown in Table 3, mean inter-observer reliability on NA responses ranged from 0.907 to 0.998, approaching almost perfect agreement across all teeth and both systems. Bootstrap-derived differences between Demirjian and Moorrees staging systems were relatively small with confidence intervals consistently including zero, indicating no significant difference. A small but statistically detectable increase between rounds was observed for FDI34 and FDI33 under the Demirjian system.

As shown in Fig. 1, Gwet's AC2 increased with the number of observers across all teeth, with a rapid initial rise followed by a relative plateau. An exception to this overall trend was observed for FDI38, where agreement did not consistently increase with the number of observers. Gwet's AC2 demonstrated almost perfect agreement even within smaller observer subsets. The Moorrees staging system generally showed a tendency toward higher agreement values than the Demirjian system across subsets. However, for FDI38, the Demirjian staging system tended to yield higher Gwet's AC2 values than the Moorrees system. A similar tendency was observed for FDI36. The levels of inter-observer agreement were comparable across assessment rounds.

Gwet's AC1 was consistently lower than AC2 but followed a similar overall pattern of increasing agreement with more observers (Fig. 2). The increase in agreement with a greater number of observers was less evident, and agreement remained moderate for FDI33-35 and FDI37-38. In contrast to AC2, the Demirjian staging system demonstrated higher agreement than Moorrees across all teeth and both rounds. Differences between rounds were minimal as in Gwet's AC2.

Inter-observer disagreement occurred predominantly between adjacent stages, as reflected by consistently higher agreement values for weighted Gwet's AC2 compared to unweighted Gwet's AC1.

Post-evaluation survey

The median years of experience among all observers was 11 years (7.5–15.5), ranging from 2 to 26 years. Familiarity of observers with the Demirjian staging system was high, with a median score of 7 (7–7.25) out of 8. However, familiarity with the Moorrees system was lower, with a median of 5.5 (3.75–7), corresponding to a moderate familiarity. The median time required to complete the task was 21 h (10–63). The overall perceived difficulty of the task was 3 (2–3) out of 5, corresponding to a moderate level of difficulty.

Eleven observers (61.1%) reported the Moorrees staging system to be more difficult. One observer (5.6%) noted that the Demirjian system was more difficult, while another observer (5.6%) reported difficulty with both systems. The remaining observers (27.8%) reported no difficulty with either the Demirjian or the Moorrees system. In comparison with Demirjian, most observers reported that the Moorrees

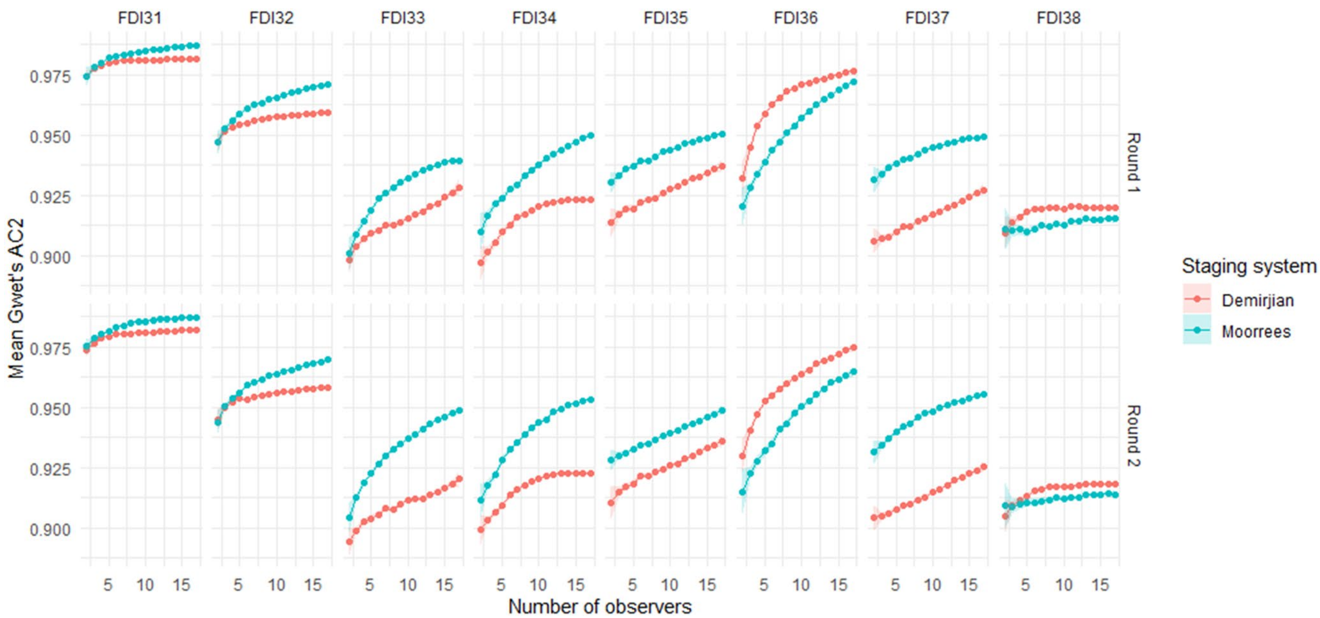

Fig. 1 Effect of the number of observers on Gwet's AC2. Each curve corresponds to one mandibular left tooth (FDI 31–38); the horizontal axis shows the number of observers included in the subset and the vertical axis the corresponding weighted Gwet's AC2, with shaded bands denoting bootstrap 95% confidence intervals

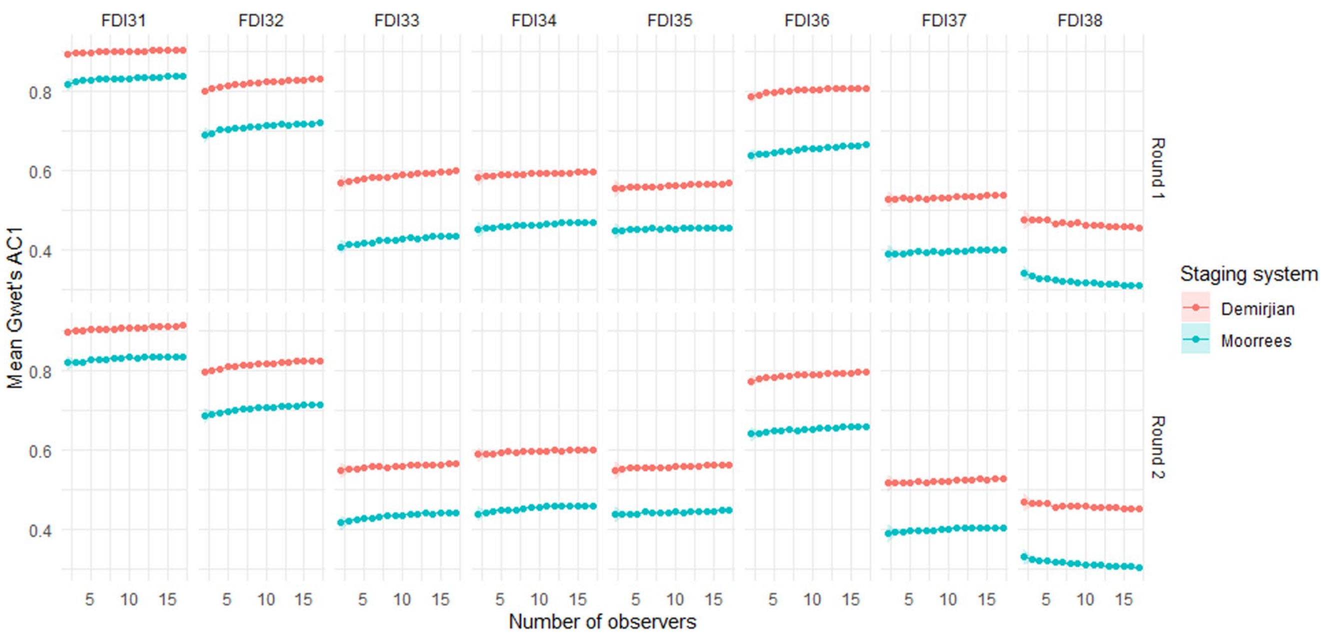

Fig. 2 Effect of the number of observers on Gwet's AC1. Each curve corresponds to one mandibular left tooth (FDI 31–38); the horizontal axis shows the number of observers included in the subset and the vertical axis the corresponding unweighted Gwet's AC1, with shaded bands denoting bootstrap 95% confidence intervals

staging system was either more difficult (44.4%) or much more difficult (11.1%). Only three observers stated that the Moorrees system was either easier (11.1%) or much easier (5.6%) for them. Five observers (27.8%) considered the two systems to be comparable. It is important to note that nearly half of the observers (44.4%) reported that the increased number of stages slightly increased confidence in assigning stages. In contrast, one-third of observers (33.3%) indicated that it slightly reduced their confidence, while one observer (5.6%) reported a significant reduction in confidence. Three observers (16.7%) stated that the larger number of stages had no effect on their confidence in assigning stages.

Table 4 Perceived difficulties and sources of uncertainty during the staging ordered by frequency of responses

Question	Answer	*n* (%)
Main difficulties encountered during staging	Differentiating between adjacent stages	14 (77.8)
	Overlapping anatomical structures	10 (55.6)
	Ambiguity in stage definitions	8 (44.4)
	Crown/root proportional estimation difficulty	7 (38.9)
	Poor radiographic quality	5 (27.8)
	Fatigue	4 (22.2)
	Time pressure	2 (11.1)
	Root morphology variations (e.g., fused roots)	1 (5.6)
Primary source of uncertainty	Lack of clear stage boundaries	13 (72.2)
	Image quality	7 (38.9)
	Number of stages	5 (27.8)
	Personal experience	3 (16.7)
	Tooth-specific anatomy	3 (16.7)
	Other	1 (5.6)
The most difficult tooth groups to stage	Molars	9 (50.0)
	Canines	6 (33.3)
	Incisors	5 (27.8)
	Premolars	4 (22.2)
The most difficult developmental phase to assess	Early root formation	7 (38.9)
	Apical closure	7 (38.9)
	Crown completion	5 (27.8)
	Mid-root development	4 (22.2)
	Early crown formation	3 (16.7)
	None of the above	1 (5.6)

Table 5 The most difficult stages to distinguish in Demirjian staging system

Stage	*n* (%)
None	7 (38.9)
A	1 (5.6)
B	2 (11.1)
C	2 (11.1)
D	3 (16.7)
E	6 (33.3)
F	6 (33.3)
G	4 (22.2)
H	0 (0.0)

Table 6 The most difficult stages to distinguish in Moorrees staging system

Single-rooted teeth		Multi-rooted teeth	
Stage	*n* (%)	Stage	*n* (%)
None	2 (11.1)	None	2 (11.1)
C_i	0 (0.0)	C_i	0 (0.0)
C_{co}	0 (0.0)	C_{co}	0 (0.0)
C_{oc}	0 (0.0)	C_{oc}	0 (0.0)
$Cr_{1/2}$	4 (22.2)	$Cr_{1/2}$	4 (22.2)
$Cr_{3/4}$	4 (22.2)	$Cr_{3/4}$	4 (22.2)
Cr_c	3 (16.7)	Cr_c	4 (22.2)
R_i	4 (22.2)	R_i	6 (33.3)
$R_{1/4}$	3 (16.7)	Cl_i	3 (16.7)
$R_{1/2}$	1 (5.6)	$R_{1/4}$	3 (16.7)
$R_{3/4}$	2 (11.1)	$R_{1/2}$	1 (5.6)
R_c	2 (11.1)	$R_{3/4}$	2 (11.1)
$A_{1/2}$	6 (33.3)	R_c	3 (16.7)
A_c	4 (22.2)	$A_{1/2}$	7 (38.9)
		A_c	3 (16.7)

As summarised in Table 4, the difficulties reported most frequently during staging were differentiating adjacent stages, overlapping anatomical structures, and ambiguity in stage definitions, while the most common source of uncertainty was a lack of clear stage boundaries. Molars were regarded as the most challenging group of teeth, while early root formation and apical closure were identified as the most difficult developmental phases to assess.

In the Demirjian staging system, the greatest difficulty was concentrated in the intermediate stages, particularly stages E, F, and G, whereas early (A-C) and the final stage (H) were rarely or never considered problematic. Notably, a substantial proportion of observers reported no difficulties distinguishing any Demirjian stages (Table 5).

For the Moorrees staging system, difficulty was more widely distributed across stages. For single-rooted teeth, the highest difficulty was reported around stage $A_{1/2}$, whereas for multi-rooted teeth the greatest difficulty was observed at stages $A_{1/2}$ and R_i. Stages $Cr_{1/2}$–Cr_c also appeared to be challenging for both single- and multi-rooted teeth, although less frequently than the most problematic stages. Early crown formation stages were not reported as problematic in either tooth group. Notably, only two observers indicated no difficulties with staging using the Moorrees system (Table 6).

Eleven observers (61.1%) were confident in the consistency of their own staging, while one (5.6%) was highly confident. A notable proportion of observers was neutral (27.8%) and only one observer (5.6%) was slightly uncertain about the consistency of their own staging.

Regarding consistency with all other observers' staging, 55.6% of observers expected moderate consistency and 44.4% expected high consistency of inter-observer agreement.

Half of the observers (50.0%) indicated that the combination of both systems, Demirjian and Moorrees, would be more appropriate for forensic casework. Four observers (22.2%) chose Demirjian as the more appropriate system, while two observers (11.1%) preferred Moorrees. Three observers (16.7%) stated that the choice of staging system should depend on the case.

Discussion

Forensic DAE assessment, as part of the DAE evaluations on forensic institutes around the world [10], carry significant legal weight, directly influencing judicial decisions regarding an individual's minority or majority of age status. This assessment impacts legal protections, detention conditions, and access to social services. Moreover, calibration and consensus on stage classifications are essential in research settings. To mitigate stage misclassifications, which are particularly problematic when applying the "benefit of the doubt" or minimum age principle [33, 34], DAE studies should be conducted through a consensus approach involving experienced observers. Thus, the reliability, transparency, and reproducibility of the methods employed are of high importance.

Despite the considerable advances in artificial intelligence (AI) within forensic odontology (particularly in domains such as dental age assessment) human expert judgment remains the focal point of forensic practice. In alignment with both the FDI AI Statement [35] and the 2018 European Ethical Charter on AI in Judicial Systems (CEPEJ) [36], AI should be conceptualized as an assistance of human expert judgment rather than a replacement. Within this paradigm, the forensic odontologist retains ultimate interpretative authority and responsibility, while algorithmic systems enhance analytical precision, efficiency, reliability, and evidentiary traceability. This study aimed to evaluate and compare the inter- and intra-observer agreement among 18 forensic odontologists and forensic experts trained in DAE, using both Demirjian's eight-stage method [16] and the multi-stage dental development system proposed by Moorrees et al. [19], which includes up to 14 stages per tooth. While the Moorrees staging method offers superior anatomical and developmental precision compared to more simplified systems, such as Demirjian's 8-stage method, our findings highlight that increased staging complexity introduces significant challenges in observer consistency, particularly in applied forensic contexts.

Dental development staging has evolved from quantitative fractional approaches — illustrated by Gleiser and Hunt [37] and later refined by Nolla [18] and Moorrees et al. [19], who described development through proportional crown-height and root-length estimates supported by morphological landmarks such as cusp-tip appearance, crown completion, half-root length, and apex closure — to the morphological-criteria approach proposed by Demirjian et al. [16], which aggregates the same biological process into eight discretely defined stages with schematic and radiographic exemplars. The Demirjian and Moorrees systems were therefore selected for the present study because they bracket the spectrum of staging granularity: Demirjian represents the minimum-resolution end with eight broad stages, while Moorrees represents the maximum-resolution end with 13–14 fractional stages for anterior and posterior teeth, respectively. This pairing allowed the same dentitions to be evaluated under markedly different cognitive demands and stage-boundary definitions, directly addressing the central question of whether increased staging complexity influences expert agreement.

Agreement across the two staging systems

Fewer-stage methods such as Demirjian's tend to produce higher exact inter-observer agreement, whereas multi-stage systems like Moorrees may show comparable or higher agreement when ordinal differences between adjacent stages are accounted for. In the present study, the Moorrees system demonstrated higher inter-observer agreement under the weighted Gwet's AC2, consistent with its larger number of stages and the correspondingly smaller penalty assigned to adjacent-stage disagreements. Conversely, the Demirjian system demonstrated higher inter-observer agreement under the unweighted Gwet's AC1, reflecting its fewer and more discretely defined categories that facilitate exact stage agreement between observers. The Demirjian system may therefore be preferable when precise stage assignment is required.

Most likely, methods with fewer stages, such as Demirjian's [16], tend to produce higher exact inter-observer agreement (Gwet's AC1), likely due to their intuitive design and the "pattern recognition" acquired through long-standing professional familiarity of many forensic practitioners. Forensic odontologists who routinely perform DAE assessments often retain these stages, allowing for rapid assessment without constant reference to staging charts.

In contrast, the Moorrees staging system [19], while more descriptive, requires a higher cognitive demand, implying continuous consultation of reference tables due to subtle differences in mineralization between adjacent stages. This not only extends the assessment time but also induces observer fatigue. In forensic science, inter- and intra-observer agreement serve as a proxy for method robustness and reliability, which are essential for ensuring scientific admissibility and legal defensibility [38]. High variability between observers undermines the credibility of age estimates and can pose serious challenges in court, especially when the opinion of the forensic expert is crucial to a legal decision. Therefore, the theoretical precision of a high-resolution staging system is only valuable if it does not compromise its practical reliability.

Stage-specific difficulties and observer experience

The difference in cognitive demand is further reflected in the pattern of stage-specific difficulties observed in the present study. The Demirjian staging system showed more localized difficulty with tooth developmental stage assignments, with the most challenging stages concentrated in root formation and root–apex transition phases (stages E, F, and G). Importantly, no observer encountered difficulties in assigning Demirjian stage H for third molars, which is of considerable forensic relevance, since the identification of complete root apex closure in third molars is frequently used as a key indicator in the legal distinction between minor and adult status. However, most observers reported no difficulties with stage assignments using the Demirjian system [16] overall. Conversely, the Moorrees staging system [19] demonstrated more widely distributed difficulty across stages, without a clear bottleneck. Increased difficulty in stage assignment was particularly observed in later crown formation stages, the transition between crown and root, and advanced root and apical stages for both single- and multi-rooted teeth.

The distribution of "Not Applicable" (NA) responses was concentrated in the anterior teeth (FDI 31, 32, 33) and the third molar (FDI 38). It is reported that the reduced clarity typical of the anterior region on panoramic radiographs (superimposition of the cervical spine and ghost images), makes definitive stage assignment difficult and increases the frequency of NA [39, 40]. For the third molar (FDI 38), the high NA value was hypothesized due to the variation in anatomical position of the third molars.

Furthermore, our findings emphasize the influence of the observer experience and training. Participants with extensive exposure to DAE, especially with the specific staging method examined, generally demonstrated higher internal consistency. However, variability increased when the staging system was unfamiliar. This effect was more evident in the unweighted test (Gwet's AC1), indicating lower exact stage agreement for Moorrees, whereas the weighted test demonstrated relatively stable agreement between systems. This pattern aligns with the ABFO [8] and IOFOS [7] guidelines, which explicitly recommend that examiner familiarity with the chosen staging method should be considered alongside accuracy and applicability when selecting an approach for forensic age estimation. These results underline the importance of training programs, reference charts, and standardized calibration exercises prior to conducting assessments, particularly when using detailed multi-staging methods such as Moorrees [19] or Nolla [18]. The considerable between-observer variability in system preference revealed by the post-evaluation survey is consistent with this interpretation: it most plausibly reflects differences in observers' training and prior experience–including which staging system was emphasized during their forensic training–rather than an intrinsic superiority of either method.

Number of observers and panel size

An important methodological finding of this study is the analysis of agreement as a function of number of observers. Our bootstrap analysis shows that agreement rates increased rapidly with the addition of observers, reaching a functional plateau at approximately 8–10 observers. This identifies a threshold of diminishing returns, where the incremental gain in reliability is offset by the logistical burden of multi-observer consensus. For practical forensic casework, our data indicate that a panel of well-trained experts approaching this plateau (approximately 8 to 10 observers) maximises reliability; smaller panels may still achieve acceptable agreement, albeit with a modest reduction in reproducibility, so the final number of observers should balance diagnostic reliability against operational feasibility. This finding provides an empirical basis for determining minimum observer requirements in forensic dental research and consensus-based age estimation protocols.

Even so, it is important to note the primary statistical method used in this study, Cohen's Kappa and Gwet's AC families of coefficients, quantify agreement as a function of concordance corrected for chance, with the weighted variants further adjusting the disagreement contribution by a distance-based penalty matrix. Because these penalty weights are typically normalised by the maximum possible distance (*k-1 for k* stages), the step penalty for a single-stage discrepancy decreases as the number of stages increases. For example, in a 14-stage system such as Moorrees, an adjacent-stage disagreement is assigned a smaller penalty compared with an 8-stage system such as Demirjian, where the same discrepancy is given approximately twice the relative weight. As a result, more finely graded staging systems may yield higher weighted agreement values, even when the underlying pattern of disagreements is similar.

Implications for age estimation

Because an exact match in stage corresponds to an identical estimated age, the higher exact (unweighted) agreement of the Demirjian system implies greater reproducibility of the resulting point age estimate between observers. Conversely, the finer Moorrees system may be advantageous where legal age thresholds are decisive: a single Demirjian stage (i.e., stage G) gives several Moorrees stages, so the reference-age confidence intervals attached to a Moorrees stage would be expected to be narrower. This can also be less advantageous if observers are less certain about the many Moorrees apical stages than about Demirjian stage G, since greater uncertainty would widen the reference-age intervals for those stages.

These implications were then reflected in the findings of the survey. The post-evaluation survey offers valuable insight into the cognitive dimension of dental staging. The most frequently reported challenge, namely difficulty in distinguishing adjacent stages, unclear stage boundaries, and the presence of overlapping anatomical structures, closely mirrored the observed pattern of adjacent-stage disagreement identified in the quantitative analysis. The Moorrees staging system was generally perceived as more demanding by the majority of observers, which is likely attributable to its increased number of stages, the greater complexity of its stage definitions, and the higher cognitive load required during assessment. Interestingly, perceptions of this increased granularity varied: while some observers reported enhanced confidence due to the finer differentiation of developmental stages, others experienced reduced confidence, suggesting a trade-off between precision and usability. Furthermore, the concentration of difficulty within Demirjian stages E–G supports the interpretation that root development phases represent critical transitional periods, during which morphological features become less distinct and more challenging to classify consistently. This observation may also indicate that further methodological refinement is warranted, and that the development of an adapted staging approach with more clearly defined intermediate stages in these transition zones could potentially improve consistency and reduce observer-related variability. Very recently, a refined approach to the Demirjian staging system has been proposed [41], incorporating additional intermediate stages for the assessment of third molar mineralisation. While such refinements aim to increase chronological accuracy, the findings of the present study introduce a critical caveat regarding observer reproducibility. Our comparison between the 8-stage Demirjian framework and the high-resolution, multi-stage Moorrees method demonstrates that an increase in staging complexity inherently elevates subjective ambiguity, subsequently depressing inter-observer agreement. Consequently, while the sub-stages proposed by Timme et al. [41] may mathematically optimize age estimation models, they risk introducing the same classification friction observed herein for the Moorrees system. In forensic daily practice, a balance should be struck between achieving greater granularity and maintaining diagnostic reliability. Therefore, any extension of existing staging systems warrants cautious calibration and extensive multi-centric reliability testing before widespread implementation.

Looking ahead, the use of developmental staging methods with a higher number of stages, such as those proposed by Moorrees et al. [19] and Nolla [18], may benefit substantially from the integration of artificial intelligence (AI) and digital imaging technologies. While these methods offer enhanced anatomical resolution and the potential for more precise chronological age estimates, their inherent complexity can pose significant cognitive and visual challenges for human observers. Distinguishing between multiple closely related developmental stages often requires careful comparison with reference images or staging tables, which may prolong assessment time and increase subjectivity.

In this context, AI-assisted staging platforms and machine learning algorithms trained on large, annotated radiographic datasets could offer a promising solution. By automating or semi-automating the identification and classification of developmental stages, such tools could mitigate inter-observer variability, enhance reliability, and increase the overall efficiency of age estimation procedures. Moreover, standardizing assessments across practitioners through digital decision-support systems would not only strengthen the scientific robustness of DAE but also enhance its forensic defensibility in legal contexts. Future research should therefore explore the feasibility, accuracy, and usability of AI-driven staging models, particularly for complex multi-stage systems, as part of a broader effort to modernize and harmonize forensic dental practices.

Ultimately, the selection of a DAE method should be informed not only by its theoretical accuracy or anatomical detail, but by its reliability, clarity of staging definitions, and forensic utility. In contexts where legal outcomes hinge on dental assessments, consistency across practitioners becomes crucial. The present study contributes to the growing body of evidence indicating that staging complexity must be carefully weighed against practical reliability. These findings underscore the need for greater harmonization of training protocols, further research into observer performance across methods, and consideration of digital decision-support tools to improve consistency in DAE.

Strengths, limitations, and future directions

Nonetheless, this study has certain limitations. A limitation of the present study is the absence of radiographs representing the earliest stages of tooth development. This shortcoming is inherently linked to the retrospective design of the research and the lack of publicly available radiographic databases encompassing these stages. In accordance with established clinical guidelines and the ALARA (As Low As Reasonably Achievable) principle, panoramic radiographs are generally not indicated in children younger than approximately six years of age, given the absence of valid clinical justification at this developmental stage; accordingly, strict adherence to radiation protection principles as outlined by the International Commission on Radiological Protection remains imperative [42], particularly when imaging younger individuals. Consequently, existing radiographic collections worldwide predominantly include subjects aged six years and older, when mixed dentition begins, and justified imaging may be warranted. For this reason, the earliest developmental stages could not be represented in our dataset. Relatedly, because the sample was restricted to a narrow age range (median 10 years; IQR, 8–12 years), the developmental stages were not evenly represented across tooth types. Mid-developmental stages predominated, whereas later stages, in which apical closure is generally more difficult to distinguish, were comparatively infrequent. This uneven stage distribution may have favoured higher agreement than would be expected in a sample spanning the full developmental spectrum, and the present estimates should therefore be interpreted as applying primarily to this age window rather than to dental staging across all ages.

A further consideration is that, while all observers were experienced practitioners, their prior familiarity with the two systems was uneven: the Demirjian system was widely known, whereas familiarity with the Moorrees system was more variable among the observers. This imbalance is likely to have favored exact agreement under the more familiar Demirjian framework, as reflected in the higher Gwet's AC1 values for that system, and should be considered when generalizing the present findings.

Future studies should explore the integration of training interventions or calibration sessions to improve agreement across complex multi-staging systems. There is also a growing opportunity to leverage artificial intelligence and computer vision to automate or assist in the staging process, particularly methods that involve high visual discrimination [4, 43]. Moreover, further research is needed to correlate inter-observer agreement with actual chronological age estimation errors, to determine whether lower agreement always translates into reduced DAE accuracy.

Conclusion

Both the Demirjian and Moorrees dental staging systems demonstrated high intra- and inter-observer reliability when applied by trained observers, supporting their continued use in dental age estimation. It should be emphasized, however, that the present study quantifies precision (i.e., agreement among observers) rather than accuracy against true CA; despite demonstrating high statistical agreement, a given method may still be prone to misclassification at the level of individual ages, influenced by factors including observer expertise and imaging modality. Disagreement between observers was predominantly driven by adjacent-stage differences, particularly in transitional phases of tooth development. The findings also highlight that increasing the number of observers yields limited benefit beyond a moderate group size, emphasizing the value of well-trained expert panels. Taken together, increased staging complexity did not substantially compromise expert agreement, as both the Demirjian and Moorrees systems achieved high intra- and inter-observer reliability. The practical choice between them depends on the objective–the Demirjian system is preferable when exact stage concordance is required, whereas the finer resolution of the Moorrees system is advantageous when greater developmental granularity is needed–rather than on any material loss of reliability attributable to its larger number of stages.

Future methodological refinement, particularly in the definition of transitional stages, may further enhance consistency and reliability. Overall, the results reinforce the robustness of current staging systems while identifying areas for improvement in forensic dental age assessment.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00414-026-03946-3.

Acknowledgements None.

Author contributions NA: Conception, Data collection, data curation Writing, Critical review, Proofreading and Supervision, RMB: Conception, data curation, Critical review, supervision-SDL: Data curation, Writing, Critical review and supervisionNP : Data analysis and interpretation, writingAK: Data curation, Critical reviewIG: Data curation, Critical reviewJK: Data curationISV: Data curationAHS: Data curation JPT: Data curation RS: Data curationDF: Data curation, manuscript revisionRAP : Data curationLGJ: Data curationAN: Data curationER: Data curationPP: Data curation PMGG: Critical reviewJAG: Data curation , writing Critical reviewBK: Data curation and Critical reviewAF: Data curation, Writing, Critical review and supervision.

Funding Open access funding provided by University of Bern. This work did not receive any specific funding from public, commercial, or non-profit organizations.

Data availability The data supporting this study's findings are available on request from the corresponding author.

Declarations

Ethical approval This study received ethical approval from the Health Research Ethics Committee of the Faculty of Medicine, Universitas Diponegoro (Approval No. 212/EC/KEPK/FK-UNDIP/VIII/2025) The study used fully anonymized, retrospectively collected data from an existing dataset with no direct involvement of human participants.

Compliance with ethical standards The study was conducted in accordance with the ethical standards set out in the Declaration of Helsinki (Finland).

Research involving Human Participants and/or Animals Not applicable.

Human Ethics and Consent to Participate declarations not applicable.

Informed consent Not applicable.

Competing Interests The authors declare no competing interests.

Clinical trial number not applicable.

References

1. Mishori R (2019) The Use of Age Assessment in the Context of Child Migration: Imprecise, Inaccurate, Inconclusive and Endangers Children's Rights. Child (Basel) 6(7):85. https://doi.org/10.3390/children6070085
2. Lossois M, Cyteval C, Baccino E, Peyron PA (2022) Forensic age assessments of alleged unaccompanied minors at the Medicolegal Institute of Montpellier: a 4-year retrospective study. Int J Legal Med 136(3):853–859. https://doi.org/10.1007/s00414-022-02813-1
3. Dewana AN, Auerkari EIA (2025) Comprehensive review: update in age estimation of forensic odontology. Bull Int Assoc Paleodontol 19(1):48–54
4. Vila-Blanco N, Varas-Quintana P, Tomás I, Carreira MJ (2023) A systematic overview of dental methods for age assessment in living individuals: from traditional to artificial intelligence-based approaches. Int J Legal Med 137(4):1117–1146. https://doi.org/10.1007/s00414-023-02960-z
5. Practical guide on age assessment 2nd edn. Technical report EASO. [Internet] pp 56–59. Available from: https://reliefweb.int/sites/reliefweb.int/files/resources/easo-practical-guide-on-age-assesment-v3–2018.pdf
6. Schmeling A, Grundmann C, Fuhrmann A et al (2008) Criteria for age estimation in living individuals. Int J Legal Med 122:457–460
7. Pereira, C. P., Liversidge, H., Brkić, H., Vodanović, M., Pinchi, V., Alves da Silva, R. H., … I. Kvaal, S. (2026). I.O.F.O.S. Recommendations for Quality Assurance: Dental age assessment in living individuals. The Journal of Forensic Odonto-Stomatology - JFOS 44(1):2–10. https://doi.org/10.5281/zenodo.19689723
8. American Board of Forensic Odontology Standards and guidelines for dental age assessment 2020, Available from: https://abfo.org/wp-content/uploads/2021/02/ADA-Technical-Report-No.–1077_July_2020.–02–2021.pdf
9. Technical ADA Report 1077–2020 Human Age Assessment by Dental Analysis
10. Thicot F, Egger C, Castiglioni C et al (2024) Forensic age estimation at the University Center of Legal Medicine Lausanne-Geneva: a retrospective study over 12 years. Int J Legal Med 138:1881–1889. https://doi.org/10.1007/s00414-024-03254-8
11. Taranilla Castro AM, Pujol-Robinat A, Badía-García MA et al (2023) Forensic age estimation in Barcelona: analysis of expert reports issued between 2011 and 2018. Int J Legal Med 137:395–402. https://doi.org/10.1007/s00414-022-02926-7
12. Pradella F, Pinchi V, Focardi M, Grifoni R, Palandri M, Norelli G-AA (2017) The age estimation practice related to illegal unaccompanied minors immigration in Italy. J Forensic Odontostomatol 2:147–156
13. Hagen M, Schmidt S, Schulz R, Vieth V, Ottow C, Olze A et al (2020) Forensic age assessment of living adolescents and young adults at the Institute of Legal Medicine, Münster, from 2009 to 2018. Int J Legal Med 134:745–751
14. Metsäniitty M, Varkkola O, Waltimo-Sirén J, Ranta H (2017) Forensic age assessment of asylum seekers in Finland. Int J Legal Med 131:243–250. https://doi.org/10.1007/s00414-016-1498-x
15. Rudolf E, Kramer J, Gebauer A, Bednar A, Recsey Z, Zehetmayr J et al (2015) Standardized medical age assessment of refugees with questionable minority claim—a summary of 591 case studies. Int J Legal Med 129:595–602
16. Demirjian A, Goldstein H, Tanner JM (1973) A new system of dental age assessment. Hum Biol 45:211–227. https://doi.org/10.2307/41459864
17. Merdietio Boedi R, Angelakopoulos N, Franco A et al (2025) Evaluation of Demirjian staging method reproducibility and reliability for dental age estimation between eight forensic experts. Int J Legal Med. https://doi.org/10.1007/s00414-025-03560-9
18. Nolla CM (1960) The development of permanent teeth. J Dent Child 27:254–266
19. Moorrees CF, Fanning EA, Hunt EE Jr (1963) Age Variation of Formation Stages for Ten Permanent Teeth. J Dent Res 42:1490–1502. https://doi.org/10.1177/00220345630420062701
20. Abdul Rahim AH, Davies JA, Liversidge HM (2023) Reliability and limitations of permanent tooth staging techniques. Forensic Sci Int 346:111654. https://doi.org/10.1016/j.forsciint.2023.111654
21. Shrout PE (1998) Measurement reliability and agreement in psychiatry, Stat. Methods Med Res 7(3):301–317
22. Marconi V, Iommi M, Monachesi C (2022) Validity of age estimation methods and reproducibility of bone/dental maturity indices for chronological age estimation: a systematic review and meta-analysis of validation studies. Sci Rep 12:15607. https://doi.org/10.1038/s41598-022-19944-5
23. Dhanjal KS, Bhardwaj MK, Liversidge HM (2006) Reproducibility of radiographic stage assessment of third molars. Forensic Sci Int 159:S74–S77. https://doi.org/10.1016/j.forsciint.2006.02.020
24. Leatherman G (1971) Two-digit system of designating teeth–FDI submission. Aust Dent J 16:394
25. World Medical Association (2025) World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Participants. JAMA 333(1):71–74. https://doi.org/10.1001/jama.2024.21972
26. Gwet KL (2018) An Evaluation of the Impact of Design on the Analysis of Nominal-Scale Inter-Rater Reliability Studies; Advanced Analytics, LLC: Gaithersburg, MD, USA
27. Rotondi MA, Donner A (2012) A confidence interval approach to sample size estimation for interobserver agreement studies with multiple raters and outcomes. J Clin Epidemiol 65(7):778–784. https://doi.org/10.1016/j.jclinepi.2011.10.019
28. Acharya AB (2011) Age estimation in Indians using Demirjian's 8-teeth method. J Forensic Sci 56(1):124–127.https://doi.org/10.1111/j.1556-4029.2010.01566.x Epub 2010 Sep 20. PMID: 20854361
29. Cohen J (1960) A coefficient of agreement for nominal scales. Educational Psychol Meas 20:37–46. https://doi.org/10.1177/001316446002000104
30. Cohen J (1968) Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. Psychol Bull 70:213–220. https://doi.org/10.1037/h0026256
31. Gwet K (2014) Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters, 4th Ed. Advanced Analytics, LLC
32. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174. https://doi.org/10.2307/2529310
33. Wirtz J, Smit C, Robinson L et al (2025) The temporal course of third molar mineralization in a black South African population. Int J Legal Med. https://doi.org/10.1007/s00414-025-03664-2

34. Wirtz J, Guo YC, Timme M et al (2025) The time frame of third molar mineralization in a northern chinese population. Forensic Sci Med Pathol. https://doi.org/10.1007/s12024-025-01131-3
35. FDI World Dental Federation (2023) FDI policy statement on artificial intelligence in dentistry. Int Dent J 73(6):601–603
36. Council of Europe (2018) European ethical charter on the use of artificial intelligence in judicial systems and their environment (CEPEJ). Council of Europe, Strasbourg
37. Gleiser I, Hunt E (1955) The permanent first molar: its calcification. Am J Phys Anthropol 13(2):253–283
38. Angelakopoulos N, Boedi M, Franco R (2025) Assessing observer-dependent dental age estimation procedures: intra- and inter-observer reliability across four well established radiographic systems for dental analysis. Int J Legal Med. https://doi.org/10.1007/s00414-025-03616-w
39. Dhillon M, Raju SM, Verma S, Tomar D, Mohan RS, Lakhanpal M, Krishnamoorthy B (2012) Positioning errors and quality assessment in panoramic radiography. Imaging Sci Dent 42(4):207–212. https://doi.org/10.5624/isd.2012.42.4.207
40. White SC, Pharoah MJ (2014) Oral Radiology: Principles and Interpretation, 7th edn. Mosby, St. Louis
41. Timme M, Olze A, Wirtz J et al (2026) Dental age assessment: A proposal for refining and extending the Demirjian stages for third molar mineralization evaluation. Int J Legal Med. https://doi.org/10.1007/s00414-026-03819-9
42. International Commission on Radiological Protection (2007) The 2007 recommendations of the International Commission on Radiological Protection. Ann ICRP 37(2–4):1–332. https://doi.org/10.1016/j.icrp.2007.10.003
43. Franco A, Murray J, Heng D, Lygate A, Moreira D, Ferreira J, Miranda E, Paulo D, Machado CP, Bueno J, Mânica S, Porto L, Abade A, Paranhos LR (2024) Binary decisions of artificial intelligence to classify third molar development around the legal age thresholds of 14, 16 and 18 years. Sci Rep 14(1):4668. https://doi.org/10.1038/s41598-024-55497-5

Authors and Affiliations

Nikolaos Angelakopoulos[1,2] · Rizky Merdietio Boedi[3,4] · Stefano De Luca[5] · Nikita Polukhin[6] · Akiko Kumagai[7] · Ivan Galic[8] · Jeta Kelmendi[9] · Israel Soriano Vázquez[10] · Aida Hadzic Selmanagić[11] · Jorge Pinares Toledo[12] · Roberto Scendoni[13] · Daniel Franklin[14] · Roberta Araujo Pinto[2] · Lourdes Gómez Jiménez[15] · Arjeta Neziri[16] · Emerson Rodrigues[2] · Pisha Pittayapat[17] · Pedro Manuel Garamendi-González[18] · Joe Adserias-Garriga[19] · Bernhard Knell[20] · Ademir Franco[2,21]

✉ Nikolaos Angelakopoulos
nikolaos.angelakopoulos@unibe.ch

1 Department of Orthodontics and Dentofacial Orthopedics, University of Bern, Bern, Switzerland

2 Division of Forensic Dentistry, Faculdade São Leopoldo Mandic, Campinas, Brazil

3 Department of Dentistry, Faculty of Medicine, Universitas Diponegoro, Semarang, Indonesia

4 Department of Forensic Medicine, Sechenov First Moscow State Medical University, Moscow, Russian Federation

5 Area of Physical Anthropology, Department of Organisms and Systems Biology, Faculty of Biology, University of Oviedo, Oviedo, Spain

6 Department of Public Health and Medical Social Sciences, Synergy University, Moscow, Russian Federation

7 Division of Forensic Odontology and Disaster Oral Medicine, Department of Forensic Science, Iwate Medical University, Iwate, Japan

8 Department of Oral Surgery, School of Medicine, University Hospital of Split, University of Split, Split, Croatia

9 Department of Orthodontics, University Dentistry Clinical Center of Kosovo, Pristina 10000, Kosovo

10 Forensic Odontologist, Independent Researcher, Mexico City, Mexico

11 Department of Dental Morphology with Dental Anthropology and Forensics Faculty of Dentistry, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

12 Departamento de Patologia y Medicina Oral, Facultad de Odontologia, Universidad de Chile, Santiago, Chile

13 Department of Law, Institute of Legal Medicine (AgEstimation Project), University of Macerata, Macerata, Italy

14 Centre for Forensic Anthropology, School of Social Sciences, The University of Western Australia, 35 Stirling Hwy, Crawley, WA, Australia

15 Instituto Nacional de Patología "Dr. Sergio Sarita Valdez", Santo Domingo, Dominican Republic

16 Independent Researcher, Pristina, Kosovo

17 Department of Radiology, Faculty of Dentistry, Chulalongkorn University, Bangkok, Thailand

18 Instituto de Medicina Legal y Ciencias Forenses de Huelva, Huelva, Spain

19 Department of Applied Forensic Science, Mercyhurst University, Erie, PA 16546, USA

20 University of Zurich, Zurich, Switzerland

21 Department of Therapeutic Dentistry, Institute of Dentistry, Sechenov First Moscow State Medical University, Moscow, Russian Federation