Forensic Sciences Research, 2024, Vol. 9, Issue 2
https://doi.org/10.1093/fsr/owae004
Research Article

Evaluation of data collection bias of third molar stages of mineralisation for age estimation in the living

Inês de Oliveira Santos[1,*], Isabel Poiares Baptista[2], Ricardo Henrique Alves da Silva[3], Eugénia Cunha[1,4]

[1]Department of Life Sciences, Laboratory of Forensic Anthropology, Centre for Functional Ecology (CFE), University of Coimbra, Calçada Martim de Freitas, Coimbra 3000-456, Portugal
[2]Dentistry Department, Faculty of Medicine, Center for Innovation and Research in Oral Sciences (CIROS), Institute of Periodontology, University of Coimbra, Coimbra 3000-075, Portuga
[3]Department of Stomatology, School of Dentistry of Ribeirão Preto, Forensic Odontology, University of São Paulo, São Paulo, Brazil
[4]Instituto Nacional de Medicina Legal e Ciências Forenses, Lisboa 1169-201, Portugal

*Corresponding author. E-mail: ines.olsantos@gmail.com

Abstract

Age assessment of the living is a fundamental procedure in the process of human identification, in order to guarantee fair treatment of individuals, which has ethical, civil, legal, and medical repercussions. The careful selection of the appropriate methods requires evaluation of several parameters: accuracy, precision of the method, as well as its reproducibility. The approach proposed by Mincer et al. adapted from Demirjian et al. exploring third molar mineralisation, is one of the most frequently considered for age estimation of the living. Thus, this work aims to assess potential bias in the data collection when applying the classification stages for dental mineralisation adapted by Mincer et al. A total of 102 orthopantomographs, of clinical origin, belonging to individuals aged between 12 and 25 years ($\bar{x} = 20.12$ years, SD = 3.49 years; 65 females, 37 males, all of Portuguese nationality) were included and a retrospective analysis performed by five observers with different levels of experience (high, average, and basic). The performance and agreement between five observers were evaluated using Weighted Cohen's Kappa and the Intraclass Correlation Coefficient. To access the influence of impaction on third molar classification, variables were tested using ordinal logistic regression Generalised Linear Model. It was observed that there were variations in the number of teeth identified among the observers, but the agreement levels ranged from moderate to substantial (0.4–0.8). Upon closer examination of the results, it was observed that although there were discernible differences between highly experienced observers and those with less experience, the gap was not as significant as initially hypothesised, and a greater disparity between the classifications of the upper (0.24–0.49) and lower third molars (>0.55) was observed. When bone superimposition is present, the classification process is not significantly influenced; however, variation in teeth angulation affects the assessment. The results suggest that with an efficient preparation, the level of experience as a factor can be overcome. Mincer and colleague's classification system can be replicated with ease and consistency, even though the classification of upper and lower third molars presents distinct challenges.

Keywords: forensic anthropology; age assessment; dental age; mincer; inter-rater reliability

Introduction

The field of forensic anthropology, like any other scientific domain, encompasses both theoretical and practical aspects [1]. In order to effectively apply practical techniques, it is crucial to thoroughly investigate the theoretical foundations and conduct empirical testing using available resources in a meticulous and ongoing manner [1, 2]. A good example of this is the research focused on developing age estimation methods applicable to living individuals [3–6].

Several obstacles exist in the age estimation process, from intrinsic variations in the biological markers available to methodological challenges, especially when dealing with the assessment of living individuals, when the well-being of the individual has to be warranted [6, 7]. For example, by the end of adolescence, few age markers are available, and the third molar encloses a great potential as an age indicator [8–11]. Thus, the analysis of the third molar for age estimation has been under scrutiny in recent decades, mainly due to the growing demand for accurate age assessments, recurring among undocumented individuals from vulnerable backgrounds such as, for example, migrants, refugees, and victims of human trafficking [12, 13].

There are several approaches available to estimate age with the third molar [14–20]. The method adapted by Mincer et al. [8], from the original by Demirjian et al. [21], is one of the most frequently considered. The experimental design of this method is in line with the guidelines by the Study Group on Forensic Age Diagnostics of the German Society of Legal Medicine [22]. In age estimation research, a method should present a series of parameters to validate its use in a medico-legal context. Those requirements encompass the accuracy, precision, and reliability of the method, but also on how adequate its application is in relation to context [23, 24]. Although all aspects are significant, research tends to

Received: August 11, 2023. **Accepted:** December 18, 2023

prioritise accuracy, which is logical. However, it is essential not to overlook the importance of developing methods that can be replicated by the scientific community. After all, an accurate method is only truly valuable if it can be utilised and validated by other researchers. For example, a review of 269 publications on age estimation of non-adults, from both bioarchaeological and forensic cases, found that <50% presented valid statistical parameters for observer error [23, 25].

An analogous situation can be found when surveying the literature related to age estimation by third molar analysis, with published works either not reporting agreement assessment or not reporting its results [26–29]. In age estimation, measuring precision involves evaluating the proximity of agreement among independent test results, with observer error serving as an indicator of precision, highlighting the importance of minimizing variability [23]. Practitioners need to be confident that the method is robust, reliable, and sufficiently accurate to be of value and admissible in a legal setting, while also being aware of the level of expertise and training required [24, 30].

Hence, this work aims to assess potential bias in the data collection process of the eight stages for dental mineralisation adapted by Mincer et al. [8] by investigating how different professional degrees of experience can affect the analysis, thus evaluating the difficulty of applying the method accordingly, and lastly, to exploit the influence of the degree of impaction of the third molar on the classification method.

Materials and methods

The study was authorised by the Ethics Committee of the Faculty of Medicine of the University of Coimbra (Portugal) (CE-104/2018).

From a wider set of orthopantomographs (OPGs) (N = 2 000) obtained from the Area of Dental Medicine at the Faculty of Medicine of the University of Coimbra (Portugal), all acquired for medical diagnosis purposes between 2006 and 2019, the sample for this study was selected with the only criterion being the presence of at least one third molar. All identification data were anonymised before this analysis. In total, 102 OPGs from individuals aged between 12 and 25 years ($\bar{x}$ = 20.12 years, SD = 3.49 years; 65 females, 37 males; of Portuguese nationality) were included after an aleatory selection. The retrospective analysis was performed once by five observers with different levels of experience (professional and related to age estimation methods): high [a dentist (Obs. 1) and one PhD student with experience in age estimation dental methods (Obs. 2)]; average [one PhD student with experience in analysing skeletal human remains (Obs. 3)]; basic [two bachelor students with basic training in human osteology (Obs. 4 and Obs. 5)]. There was no contact between the observers at the moment of data collection. Upper and lower third molars from both quadrants, identified according to FDI (Dederation Dentaire Internationale - World Dental Federation), were analysed whenever present.

Data collection was executed following the directions provided by Mincer et al. [8] using a scale of eight stages (A to H) describing the third molar mineralisation of crown and root. An impaction assessment was performed (by Obs. 2) using the schemes by Pell and Gregory [31] and Xavier et al. [32], where the three parameters are considered: PB—position of the third molar relative to the bone (with a scale from I to III); PM—position of the third molar relative to the second molar (with a scale from 1 to 3); and W—third molar angulation (the scale used ranges from 1 to 7; however, no elements were classified between 4 and 7, thus these values were not taken into account in the analysis). All stages descriptions are available in Supplementary Tables S1 and S2.

All data were analysed using IBM SPSS Statistics 26.0 (IBM Corp., Armonk, NY, USA). The alphabetical scales used were coded to numerical scales to homogenise the database. Descriptive analysis of the sample was performed with chronological age, sex, and third molars analysed.

The agreement of the five observers was evaluated through the calculation of the Weighted Cohen's Kappa [33], interpreted according to Landis and Koch [34], where 0.00–0.20 represents poor strength of agreement, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, 0.81–1 represents an almost perfect agreement, and through the calculation of the intraclass correlation coefficient (ICC) based on a mean-rating (k = 3), absolute-agreement, two-way random model. The ICC results were interpreted according to the guideline proposed by Koo and Li [35] where ICC lower than 0.5 refers to poor reliability, between 0.5 and 0.75 refers to moderate reliability, between 0.75 and 0.9 indicates good reliability, and values greater than 0.90 indicate excellent reliability. Comparisons were made by dividing the observers by experience: high/high, high/average, basic/average.

To evaluate if the third molar impaction stages influenced third molar classification, we employed an ordinal logistic regression performed using Generalised Linear Models which applies a maximum likelihood estimation approach to estimate the coefficients (B) and odds ratios (ORs) associated with the predictor variables. In the present study, PM, PB and W were predictors, and the Obs. 2 third molar classification was the outcome: OR = 1 implies that the predictor does not affect the odds of an outcome, OR > 1 implies that the predictor is associated with higher odds of presenting the outcome, and OR < 1 implies that the predictor is associated with lower odds of exhibiting the outcome. The confidence interval (CI) was set at 95%, and it estimates the precision of the OR. A low level of precision is shown by wide CI, whereas narrow CI suggests higher precision. Significance was set at a P-value <0.05. The proportional odds [18: χ^2 (24) = 23.149, P = 0.511; 28: χ^2 (24) = 33.329, P = 0.097; 38: χ^2 (30) = 23.554, P = 0.792] and noncollinearity (VIF < 3) assumptions were verified. The data for Teeth 48 were the only instance where the proportional odds assumption was not followed (χ^2 (30) = 44.820, P < 0.05); hence, a multinomial logistic regression was performed.

Results

All available third molars were observed in the sample: 48 (lower right, 25.78%), 18 (upper right, 24.65%), 28 (upper left, 24.93%), and 38 (lower left, 24.65%). The majority of the individuals (63.7%) presented all four third molars and only four presented one third molar. Table 1 exhibits the number of teeth identified by each observer, with the minimum (stage A by Obs. 1) and maximum (H) stage of classification attributed. The frequency of stages classified, by teeth and by observer, is available in Supplementary Table S3. The stages more frequently observed were H and G (between 20% and 41%), with the exception of the classifications performed by Obs. 4 (where stages E and F were the most common).

Using weighted Kappa to measure the agreement between observers, it was possible to perceive overall moderate to strong agreement (Table 2). In fact, it is possible to verify that in all observers, the classification of the lower third molars (38

Table 1. Number of teeth (*n*) observed with the lower and maximum stage third molar mineralisation [8], attributed by each observer.

	Tooth (FDI) (N = 660)	*n* observed	Lower stage	Maximum stage
Obs. 1	18	87	A	H
	28	88	B	H
	38	87	B	H
	48	91	B	H
Obs. 2	18	88	C	H
	28	85	C	H
	38	87	B	H
	48	86	B	H
Obs. 3	18	90	C	H
	28	89	C	H
	38	90	C	H
	48	90	C	H
Obs. 4	18	85	C	H
	28	80	C	H
	38	87	C	H
	48	88	C	H
Obs. 5	18	79	C	H
	28	81	C	H
	38	87	C	H
	48	86	C	H

Table 2. Agreement analysis by weighted Cohen's Kappa.

	Tooth	Kappa	95%CI
Obs 1 × Obs 2	18	0.630	0.499–0.762
	28	0.734	0.614–0.854
	38	0.816	0.742–0.889
	48	0.784	0.705–0.864
Obs 2 × Obs 3	18	0.494	0.362–0.627
	28	0.607	0.489–0.725
	38	0.550	0.438–0.661
	48	0.595	0.494–0.695
Obs 2 × Obs 4	18	0.268	0.136–0.400
	28	0.332	0.197–0.466
	38	0.597	0.485–0.709
	48	0.577	0.452–0.703
Obs 1 × Obs 4	18	0.243	0.128–0.357
	28	0.459	0.346–0.572
	38	0.610	0.515–0.704
	48	0.644	0.552–0.735
Obs 2 × Obs 5	18	0.365	0.250–0.480
	28	0.497	0.384–0.610
	38	0.632	0.529–0.734
	48	0.672	0.575–0.769
Obs 4 × Obs 5	18	0.536	0.410–0.663
	28	0.394	0.255–0.533
	38	0.686	0.590–0.781
	48	0.610	0.485–0.735

and 48) presented agreement values above 0.55. Conversely, the weakest results were observed in the classification of the upper third molars (0.24–0.49: fair to moderate), and between the less experienced observers.

Table 3 demonstrates disparities in observations, where it is possible to verify that the majority occur between one level (−1 and 1), which is indicative that discordant classifications occur mainly between subsequent mineralisation stages.

ICC values, which compare the classifications made by each observer in the same tooth, varied from good (Teeth 18 and 28) to excellent (Teeth 38 and 48), with values ranging between 86.2% and 96.7% (Table 4). Cronbach's α values were greater than 0.9, which is typically indicative of excellent reliability.

When examining the performance of the observers in relation to each other, a balanced assessment was observed. In Table 5, the correlation matrix is presented, indicating that,

Table 3. Differences between levels of classification.

Levels	Obs 1 × Obs 2				Obs 2 × Obs 3				Obs 2 × Obs 4				Obs 1 × Obs 4				Obs 2 × Obs 5				Obs 4 × Obs 5			
	18	28	38	48	18	28	38	48	18	28	38	48	18	28	38	48	18	28	38	48	18	28	38	48
−5.00	-	-	-	-	-	1	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	1	-	-
−4.00	1	-	-	-	1	-	-	-	-	-	-	1	-	-	-	1	-	1	-	-	-	1	-	1
−3.00	-	1	-	-	1	1	1	1	1	1	-	1	-	-	-	-	1	-	-	-	2	1	-	-
−2.00	-	-	-	-	7	5	4	3	2	2	3	1	-	-	1	1	-	-	2	1	5	5	1	-
−1.00	6	3	6	5	19	17	30	33	6	7	25	25	5	5	20	20	11	12	17	18	19	26	11	12
0	57	63	63	57	37	41	32	34	22	27	36	39	20	23	39	34	20	28	43	44	36	27	43	41
1.00	13	10	15	23	13	12	16	14	20	14	13	10	18	18	17	22	23	22	18	17	11	11	23	22
2.00	7	3	3	1	6	3	2	2	20	19	2	2	20	19	2	4	22	16	5	3	2	3	2	2
3.00	1	2	-	1	-	-	1	-	6	6	-	1	12	11	-	1	-	-	1	1	-	1	-	1
4.00	2	1	-	-	-	-	-	-	1	1	1	1	2	2	1	1	-	-	-	-	-	-	-	1
5.00	-	1	-	-	-	-	-	-	-	1	-	-	1	1	-	-	-	-	-	-	-	-	-	-

Table 4. Intraclass correlation coefficient (ICC) for the observations executed by the five observers, for each third molar.

Teeth	Intraclass correlation	95%CI		F test with true value 0			
		Lower bound	Upper bound	Value	df1	df2	Sig
18	0.862	0.758	0.919	10.627	72	288	0.000
28	0.882	0.811	0.926	11.072	73	292	0.000
38	0.967	0.953	0.977	30.470	78	312	0.000
48	0.959	0.942	0.971	25.008	78	312	0.000

Table 5. Intraclass correlation coefficient (ICC) inter-item correlation matrix between the five observers, for each third molar.

Teeth		Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5
18	Obs. 1	1.000	0.750	0.724	0.477	0.644
	Obs. 2		1.000	0.702	0.474	0.661
	Obs. 3			1.000	0.646	0.786
	Obs. 4				1.000	0.709
	Obs. 5					1.000
28	Obs. 1	1.000	0.775	0.836	0.417	0.757
	Obs. 2		1.000	0.757	0.482	0.728
	Obs. 3			1.000	0.495	0.826
	Obs. 4				1.000	0.510
	Obs. 5					1.000
38	Obs. 1	1.000	0.934	0.847	0.844	0.864
	Obs. 2		1.000	0.795	0.803	0.835
	Obs. 3			1.000	0.891	0.883
	Obs. 4				1.000	0.875
	Obs. 5					1.000
48	Obs. 1	1.000	0.926	0.894	0.757	0.881
	Obs. 2		1.000	0.833	0.749	0.851
	Obs. 3			1.000	0.756	0.865
	Obs. 4				1.000	0.769
	Obs. 5					1.000

overall, there is a stronger consensus among observers regarding the mandibular third molars (>0.7). Furthermore, it is evident that Obs. 4 exhibited, overall, a relatively weaker performance compared to the other observers, including the one with a similar level of experience.

Additionally, when evaluating the impact of hypothetically removing individual observers (Table 6), the Cronbach's α values show no significant improvement. On the contrary, when analysing the Cronbach's α after excluding an observer, a decrease of reliability in the analysis of the upper third molars is observed in all raters, with the exception of Obs. 4. This finding reinforces the notion of overall homogeneity among the observers.

For the ordinal logistic regression, the goodness-of-fit was assessed for the three models: for Tooth 18, the −2 Log-Likelihood (−2LL) value was 127.447 [χ^2 (6) = 18.480, P = 0.011], Akaike's information criterion (AIC) = 149.447 and Bayesian information criterion (BIC) = 175.921, Pearson [χ^2 (89) = 79.798, P = 0.747], and deviance [χ^2 (89) = 73.333, P = 0.885] are indicative of a good fit; for Tooth 28, the −2LL value was 118.239 [χ^2 (6) = 33.243, P = 0.000], the AIC = 140.239 and BIC = 166.713, Pearson [χ^2 (84) = 68.580,

Table 6. Intraclass correlation coefficient (ICC) item-total statistics.

Teeth		Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Squared multiple correlation	Cronbach's α if item deleted
18	Obs. 1	28.66	22.201	0.758	0.641	0.886
	Obs. 2	29.11	22.377	0.754	0.630	0.887
	Obs. 3	28.84	20.195	0.846	0.722	0.867
	Obs. 4	30.03	25.027	0.651	0.526	0.907
	Obs. 5	29.78	22.174	0.821	0.708	0.873
28	Obs. 1	28.45	24.908	0.832	0.752	0.876
	Obs. 2	28.74	24.961	0.808	0.666	0.882
	Obs. 3	28.51	23.87	0.873	0.795	0.867
	Obs. 4	29.62	32.184	0.520	0.298	0.934
	Obs. 5	29.27	25.926	0.836	0.720	0.876
38	Obs. 1	29.90	33.169	0.928	0.906	0.956
	Obs. 2	30.06	35.060	0.890	0.876	0.962
	Obs. 3	29.86	35.173	0.901	0.848	0.960
	Obs. 4	29.94	36.086	0.901	0.838	0.960
	Obs. 5	30.11	35.461	0.916	0.847	0.958
48	Obs. 1	29.82	32.430	0.938	0.911	0.942
	Obs. 2	30.10	34.349	0.906	0.866	0.947
	Obs. 3	29.85	34.772	0.900	0.832	0.948
	Obs. 4	29.95	37.869	0.795	0.639	0.965
	Obs. 5	30.15	35.592	0.906	0.824	0.948

$P = 0.889$], and deviance [χ^2 $(84) = 65.308$, $P = 0.935$] are indicative of a good fit; and lastly for Tooth 38, the −2LL value was 135.180 [χ^2 $(6) = 33.462$, $P = 0.000$], the AIC = 159.180 and BIC = 188.350, Pearson and deviance [χ^2 $(120) = 88.748$, $P = 0.985$] are indicative of a good fit. The model coefficients, significances and ORs are presented in Table 7.

Observing the models, for the upper right third molar (18), the likelihood ratio tests showed statistically significant values for the PM ($P = 0.013$) parameter. For Teeth 28, the likelihood ratio tests showed statistically significant values for the three parameters, PB ($P = 0.002$), PM ($P = 0.00$), and W ($P = 0.019$). And for Teeth 38, the likelihood ratio tests showed statistically significant values for the PM ($P = 0.009$) and W ($P = 0.004$) parameters.

More specifically, it is possible to verify that for Tooth 18 classification, PB 1 and PM I are significant predictors, with a predicted increase of 1.490 and 2.145, respectively, of the classification being a higher scale, and ORs higher than 1 (Table 7). It is worth mentioning that PM II, although not statistically significant, presents OR = 3.516. For Tooth 28, PB 1, PB 2, and PM I are significant predictors, with OR of 14.212, 5.433, and 46.415, respectively. W shows statistical significance in W 1; however, with B = −1.478 and OR values below 1, it indicates a decreasing probability of the classification being a higher scale. For Tooth 38, PM I is a significant predictor, with OR of 5.004. It is worth to mention that PB 1, although not statistically significant, presents OR = 4.103. Again, W 1 and W 2 present negative coefficient and odd ratios below 0, suggesting a decreasing probability of the classification being a higher scale.

To access if the impaction stages had an influence on the lower right third molar (48) classification, a multinomial logistic regression was performed because the proportional odds assumption was violated. The fit of the model was verified: the −2LL value was 77.647 [χ^2 $(36) = 73.78$, $P = 0.000$], AIC = 161.647, BIC = 263.239, Pearson [χ^2 $(72) = 25.438$, $P = 1.000$], and deviance [χ^2 $(72) = 29.122$, $P = 1.000$] are indicative of a good fit. Overall, the multinomial logistic regression analysis revealed statistically significant values for the PM [χ^2 $(12) = 22.776$, $P = 0.030$] as a predictor for third molar classification stages.

Discussion

The accuracy of a research study, particularly in the context of an age estimation method, is influenced by multiple factors. Among these, a critical element shaping overall confidence is the reliability of data collection [35–37]. Thus, it was the aim of this work to assess potential bias in the data collection process in third molars according to the scale proposed by Mincer et al. [8].

Gathering all five raters' information, it was possible to detect that the number of teeth identified varied among the observers (Table 1). These differences, although not extravagant, can be explained by some difficulties in identifying the third molars in the OPGs. There are several factors that can influence the correct identification of third molars, from image quality to the absence of adjacent teeth as well as the morphology of this type of tooth [24, 38–40]. Considering that there was no contact between the observers at the moment of data collection, it is predictable to expect that some differences may arise in situations where other tooth types are absent (either due to clinical extraction or result of pathology) or where tooth migration happens, for example. This effect may be influenced by the level of experience of the observers; therefore, this study worked toward examining whether levels of expertise significantly affect the classification of the third molar or if the method is user-friendly and accessible even for individuals without extensive experience.

In general, the agreement levels among the five observers ranged from moderate to substantial (0.4–0.8, see Table 2). These findings align with those reported in previous studies

Table 7. Parameter estimates and significance of the ordinal logistics regression models for upper third molars (18; 28) and lower left third molar (38). Parameters with value 0 were considered redundant as they did not add information to the model.

Teeth	Parameters	B	Std. Error	Wald 95%CI		Hypothesis test			Exp(B)	Wald 95%CI for Exp(B)	
				Lower	Upper	Wald χ^2	df	Sig.		Lower	Upper
18	PB 1	1.490	0.7578	0.005	2.975	3.866	1	0.049	4.437	1.005	19.591
	PB 2	1.257	0.7504	−0.213	2.728	2.808	1	0.094	3.516	0.808	15.307
	PB 3	0							1.000		
	PM I	2.145	0.7879	0.601	3.689	7.412	1	0.006	8.543	1.824	40.020
	PM II	0.382	0.5075	−0.613	1.376	0.565	1	0.452	1.465	0.542	3.960
	PM III	0							1.000		
	W 1	0.035	0.5254	−0.995	1.065	0.004	1	0.947	1.036	0.370	2.900
	W 2	0.070	0.5655	−1.039	1.178	0.015	1	0.902	1.072	0.354	3.248
	W 3	0							1.000		
28	PB 1	2.654	0.7955	1.095	4.213	11.132	1	0.001	14.212	2.989	67.571
	PB 2	1.692	0.7693	0.185	3.200	4.840	1	0.028	5.433	1.203	24.539
	PB 3	0							1.000		
	PM I	3.838	1.1727	1.539	6.136	10.709	1	0.001	46.415	4.661	462.236
	PM II	0.869	0.4742	−0.061	1.798	3.355	1	0.067	2.384	0.941	6.038
	PM III	0							1.000		
	W 1	−1.478	0.5597	−2.575	−0.382	6.978	1	0.008	0.228	0.076	0.683
	W 2	−0.460	0.6333	−1.701	0.781	0.527	1	0.468	0.631	0.183	2.184
	W 3	0							1.000		
38	PB 1	1.412	0.7338	−0.026	2.850	3.702	1	0.054	4.103	0.974	17.288
	PB 2	0.186	0.5636	−0.919	1.290	0.109	1	0.742	1.204	0.399	3.634
	PB 3	0							1.000		
	PM I	1.610	0.7218	0.196	3.025	4.977	1	0.026	5.004	1.216	20.594
	PM II	−0.173	0.5486	−1.248	0.903	0.099	1	0.753	0.842	0.287	2.466
	PM III	0							1.000		
	W 1	−2.061	0.7775	−3.585	−0.537	7.028	1	0.008	0.127	0.028	0.584
	W 2	−2.356	0.7662	−3.858	−0.854	9.456	1	0.002	0.095	0.021	0.426
	W 3	0							1.000		

[41]. However, comparing this parameter can often be challenging due to variations in the reporting methods employed. For instance, Arany et al. [42] reported 81.5% interobserver agreement for maxillary third molar and 85% for mandibular third molar, assessing inter- and intraobserver agreements through the Wilcoxon matched-pairs signed-ranks test in a sample of 100 OPGs. Amanullah et al. [43] employed the Wilcoxon matched-pairs signed-rank test to determine intra- and interobserver agreements, finding no significant differences in the observations of 30 OPGs. Cameriere et al. [44] evaluated interobserver reliability in 30 OPTs, using Kappa statistics, and reported good interobserver repeatability, with a Cohen's Kappa of 0.93 ± 0.07, indicating substantial agreement between operators [45]. The reported Kappa scores of 0.869 for 8 stages and 0.863 for 10 stages of third molar development in an undisclosed number of OPGs, indicating good interrater agreement. Slightly lower, Alsaffar et al. [46] demonstrated a substantial agreement of 0.767 for the interobserver Kappa score in the observation of 10 OPGs. Carneiro et al. [47] assessed reproducibility by examining the agreement between 20 randomly selected OPGs, used Cohen's Kappa statistic, and found an almost perfect agreement of k = 0.802. On the other hand, May et al. [48] reported inter-examiner reliability of 0.866, indicating "almost perfect" intra- and inter-examiner reliability scores for the analysis of 37 OPGs, but did not provide details about the employed methodology. It is also possible to find published papers that either do not report agreement assessment, e.g. the works by Berkvens et al. [26], Olze et al. [27–29], or do not report the agreement values. An example of the latter, Qing et al. [49], indicated the assessment of 250 OPGs; however, the results of the evaluation are not described. Hence, it is crucial to provide a systematic and transparent report detailing the methodology employed and the specific parameters within which it was conducted [23]. This ensures a clear understanding of the study's procedures and allows for meaningful comparisons with other research activities. Such transparency enhances the interpretation of the findings and promotes scientific rigour in the field, ultimately facilitating informed decision-making in research and practice.

Meanwhile, these differences contribute to the difficulty of method selection, as precision and reproducibility are key factors for conscious and informed decision-making [35, 50]. Observer reliability is essential because it determines the accuracy of the collected data as a true reflection of the evaluated variables. In any research project, numerous sources of error can arise, but by minimizing these errors, researchers can have confidence in the validity of their study's results [35, 37].

Looking further into the results presented above, although there were discernible differences between the highly experienced observers and the less experienced, the disparity was not as significant as initially hypothesised. Examining the frequency table (Supplementary Table S3), it was noticeable that there is only one observer (Obs. 4) who presents a classification tendency contrary to the others, displaying a lower percentage of G and H stages. Nonetheless, the ICC analysis informed that although one of the observers (Obs. 4) did not perform as well as the others, the agreement values between the other observers would not improve significantly upon its hypothetical exclusion. However, considering the scarcity of related studies on this comparisons, conducting further research would be beneficial to add perspective to this work.

Furthermore, upon examining the extent of differences between observations, it becomes clear that the majority of classification mismatches consistently occur by only one degree. This pattern can be attributed to the inherent challenge of distinguishing between adjacent degrees in tooth development, rather than solely to the difficulty of comprehending the classification method. In fact, this type of discrepancy has been reported in other works either where teeth classification was manually or automatically executed [51–53]. If there were a lack of understanding regarding the method, one would expect more significantly divergent outcomes in the classifications.

This does not mean that there are no differences in the performance of observers based on experience [39, 53], but that this factor might be easily overstepped with a solid and quick preparation and may not be the most relevant when considering this method.

Through the observation of agreement values, it becomes apparent that there is a greater disparity between the classifications of the upper and lower third molars. Specifically, the agreement in observations of Teeth 18 and 28 consistently tends to be lower among all the observers. This issue is not often mentioned in the literature. Generally, the works elect to work only with the mandibular third molars [54–57], or use both upper and lower third molars without addressing any differences in method performance [58–61]. Conversely, Uys et al. [62] have reported a better agreement on the observation of maxillary third molars than of mandibular.

Consequently, a decision was made to explore whether varying degrees of third molar impaction had any influence on these observations. One potential explanation for these observations is that the visibility of the third molar in the upper jaw is compromised due to overlapping bone structures.

For this, third molar positions were evaluated according to three parameters: PB, PM, and W (Table 7). Overall, the results of the ordinal logistic regression analysis indicate the significant influence of PB 1, PB 2, and PM I on Tooth 28 classification, PB 1 and PM I on Tooth 18 classification, and PM I on Tooth 38 classification. However, the W variable demonstrates negative coefficients and ORs below 1 for Tooth 28 and Tooth 38 classifications. In the case of Teeth 48, only the PM variable showed a significant correlation.

It was observed that only a "correct position", not impacted, of the third molar in relation to the bone and second molar possibly correlates with the classification. Additionally, a "worse position", superimposed by bone, will not benefit the classification stage. However, it is also possible to verify that angulation presents a reverse influence on the third molar classification. This is indicative that although superimposition of structures and angulation has an influence, these two factors do not impact the classification in the same way. Although authors usually mention the difficulty in evaluating the upper third molars [39, 55], not much information is provided about such difficulties, and although it may be easier to simply exclude teeth from an experimental observation, that may not be a possibility in a real case. Thus, these observations reinforce the need for a case-by-case analysis for subjects under evaluation and the necessity of experimental work reporting results by type of teeth in the clearest possible way.

Conclusion

The mineralisation of the third molar is influenced by various factors such as genetics, population origin, environment, among others, which poses a constant challenge for age estimation research. Therefore, it is crucial for studies to adhere strictly to appropriate scientific and methodological guidelines in order to improve methodologies and ensure replicability.

In this study, it was possible to access that the Mincer and colleague's classification is easy and consistent to apply, regardless of the different levels of experience of the observers, but still requires training for all observers. It was also observed that the classification between upper and lower third molars presents different challenges, and the position and angulation of the third molar both influence this classification, however in different forms.

The nature of the research protocol employed in this study distinguishes itself by incorporating an adequate sample size and a diverse range of observers, thus allowing to focus and provide a new perspective on the challenges associated with the method, which are often cited as reasons for its dismissal as a valid age assessment method.

Although the Mincer method, as applied here, yielded good results, it is evident that clear and succinct guidelines and unequivocal reports of methods and results are necessary to facilitate the application of knowledge in the fields of Forensic Anthropology and Legal Medicine.

Acknowledgements

The authors would like to thank Constanze Schattke, Giovanna Matteussi, Ema Ferreira, and Inês Abreu for contributing to this work.

Authors' contributions

Inês de Oliveira Santos was responsible for the conceptualisation, methodology, and formal analysis. Inês Oliveira-Santos, Isabel Poiares Baptista, and Eugénia Cunha wrote the original draft. All authors contributed to the review and editing of the article, the final text and approved it.

Compliance with ethical standards

Not applicable.

Disclosure statement

Not applicable.

Funding

This work was supported by the R&D Unit Centre for Functional Ecology—Science for People and the Planet (CFE) [UIDB/04004/2020], financed by FCT/MCTES through national funds (PIDDAC). The author Inês Oliveira-Santos was financed by FCT [grant number SFRH/BD/139158/2018].

References

1. Boyd CC, Boyd DC. The theoretical and scientific foundations of forensic anthropology. In: Boyd CC, Boyd DC, editors. Forensic

Anthropology: Theoretical Framework and Scientific Basis. West Sussex (UK): John Wiley & Sons, Ltd., 2018. p. 1–18.
2. Langley NR, Dudzik B. The application of theory in skeletal age estimation. In: Boyd CC, Boyd DC, editors. Forensic Anthropology: Theoretical Framework and Scientific Basis. West Sussex (UK), John Wiley & Sons, Ltd., 2018. p. 99–112.
3. Baccino E, Ubelaker DH, Hayek LA, et al. Evaluation of seven methods of estimating age at death from mature human skeletal remains. J Forensic Sci. 1999;44:931–936.
4. Cunha E, Baccino E, Martrille L, et al. The problem of aging human remains and living individuals: a review. Forensic Sci Int. 2009;193: 1–13.
5. Cunha E, Ferreira MT. Antropologia Forense. In: Corte-Real F, Santos A, Cunha E, et al, editors. Tratado de Medicina Legal. Lisboa, Practor, 2022, 255–280.
6. Ubelaker DH, Khosrowshahi H. Estimation of age in forensic anthropology: historical perspective and recent methodological advances. Forensic Sci Res. 2019;4:1–9.
7. Rahman SA, Giacobbi P, Pyles L, et al. Deep learning for biological age estimation. Brief Bioinform. 2020;22:1767–1781.
8. Mincer HH, Harris EF, Berryman HE. The A.B.F.O. Study of third molar development and its use as an estimator of chronological age. J Forensic Sci. 1993;38:379–390.
9. Baccino E, Cunha E, Cattaneo C. Aging the dead and the living. In: Siegal JA, Saukko PJ, editors. Encyclopedia of Forensic Sciences 2nd edn. London (UK): Academic Press, 2013. p. 42–48.
10. Schmeling A, Black S. An introduction to the history of age estimation in the living. In: Black S, Aggrawal A, Payne-James J, editors. Age Estimation in the Living: The Practitioner's Guide. West Sussex (UK): John Wiley & Sons, Ltd., 2010. p. 1–18.
11. Tafrount C, Galić I, Franchi A, et al. Third molar maturity index for indicating the legal adult age in southeastern France. Forensic Sci Int. 2019;294:218.e1–218.e6.
12. Santiago BM, Biazevic MGH, Fernandes MM, et al. Métodos dentais Para estimativa da idade. In: Machado CEP, Deitos AR, Velho JA, et al., editors. Tratado de Antropologia Forense. Fundamentos e Metodologias Aplicadas à Prática Pericial. Campinas (Brazil): Millennium Editora, 2022. p. 415–436. Portuguese.
13. Schmeling A. Forensische Altersdiagnostik bei lebenden Jugendlichen und jungen Erwachsenen. *Rechtsmedizin*. 2011;21: 151–162. German.
14. Cameriere R, Pacifici A, Viva S, et al. Adult or not? Accuracy of Cameriere's cut-off value for third molar in assessing 18 years of age for legal purposes. Minerva Stomatol. 2014;33:111–115.
15. Haavikko K. The formation and the alveolar and clinical eruption of the permanent teeth. Suom Hammaslaak Toim. 1970;66: 103–170.
16. Harris EF. Mineralization of the mandibular third molar: a study of American blacks and whites. Am J Phys Anthropol. 2007;132: 98–109.
17. Kullman L, Johanson G, Akesson L. Root development of the lower third molar and its relation to chronological age. Swed Dent J. 1992;16:161–167.
18. Nolla CM. The development of permanent teeth. J Dentistry Child. 1960;1:254–266.
19. Nortjé CJ. The permanent mandibular third molar. Its value in age determination. J Forensic Odontostomatol. 1983;1:27–31..
20. Thevissen P, Kaur J, Willems G. Human age estimation combining third molar and skeletal development. Int J Leg Med. 2012;126: 285–292.
21. Demirjian A, Goldstein H, Tanner JM. A new system of dental age assessment. Hum Biol. 1973;45:211–227.
22. Schmeling A, Geserick G, Reisinger W, et al. Age estimation. Forensic Sci Int. 2007;165:178–181.
23. Adalian P. General considerations about data and selection of statistical approaches. In: Obertová Z, Stewart A, Cattaneo C, editors. Statistics and Probability in Forensic Anthropology. London (UK): Academic Press, 2020. p. 59–72.
24. Taylor J, Blenkin M. Age evaluation and odontology in the living. In: Black S, Aggrawal A, Payne-James J (eds.) Age Estimation in the Living: The Practitioner's Guide. West Sussex, John Wiley & Sons, Ltd., 2010, 176–201.
25. Corron L, Adalian P, Condemi S, et al. Sub-adult aging method selection (SAMS): a decisional tool for selecting and evaluating subadult age estimation methods based on standardized methodological parameters. Forensic Sci Int. 2019;304:109897.
26. Berkvens ME, Fairgrieve SI, Keenan S. A comparison of techniques in age estimation using the third molar. J Canad Soc Forensic Sci. 2017;50:74–83.
27. Olze A, Pynn BR, Kraul V, et al. Studies on the chronology of third molar mineralization in first nations people of Canada. Int J Leg Med. 2010;124:433–437.
28. Olze A, Taniguchi M, Schmeling A, et al. Comparative study on the chronology of third molar mineralization in a Japanese and a German population. Leg Med. 2003;5:S256–S260.
29. Olze A, Taniguchi M, Schmeling A, et al. Studies on the chronology of third molar mineralization in a Japanese population. Leg Med. 2004;6:73–79.
30. Kimmerle EH, Prince DA, Berg GE. Inter-observer variation in methodologies involving the pubic symphysis, sternal ribs, and teeth. J Forensic Sci. 2008;53:594–600.
31. Pell GJ, Gregory GT. Impacted mandibular third molars: classification and modified technique for removal. The Dental Digest. 1933;39:330–338.
32. Xavier C, Dias-Ribeiro E, Ferreira-Rocha J, et al. Avaliação das posições dos terceiros molares impactados de acordo com as classificações de Winter e Pell & Gregory em radiografias panorâmicas. Rev Cirurg Traumatol Bucomaxilofacial. 2010;10: 83–90. Portuguese.
33. Cohen J. Weighted Kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull. 1968;70: 213–220.
34. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. Vol. 33. 1977.
35. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15:155–163.
36. Ferrante L, Cameriere R. Statistical methods to assess the reliability of measurements in the procedures for forensic age estimation. Int J Leg Med. 2009;123:277–283.
37. McHugh ML. Lessons in biostatistics interrater reliability: the Kappa statistic. Biochem Med. 2012;22:276.
38. Bolaños MV, Moussa H, Manrique MC, et al. Radiographic evaluation of third molar development in Spanish children and young people. Forensic Sci Int. 2003;133:212–219.
39. Dhanjal KS, Bhardwaj MK, Liversidge HM. Reproducibility of radiographic stage assessment of third molars. Forensic Sci Int. 2006;159:S74–S77.
40. Fuller JL, Denehy GE, Schulien TM. Concise Dental Anatomy and Morphology. Iowa City (IA): University of Iowa College of Dentistry, 2001.
41. Banar N, Bertels J, Laurent F, et al. Towards fully automated third molar development staging in panoramic radiographs. Int J Leg Med. 2020;134:1831–1841.
42. Arany S, Iino M, Yoshioka N. Radiographic survey of third molar development in relation to chronological age among Japanese juveniles. J Forensic Sci. 2004;49:1–5.
43. Amanullah A, Ullah U, Yunus S, et al. Development stages of third-molar tooth for estimation of chronological age in children and young adult. Pak J Med Health Sci. 2016;10:750–754.
44. Cameriere R, Ferrante L, De Angelis D, et al. The comparison between measurement of open apices of third molars and Demirjian stages to test chronological age of over 18 year olds in living subjects. Int J Leg Med. 2008;122:493–497.
45. Amiroh Priaminiarti M, Syahraini SI. Comparison of age estimation between 15–25 years using a modified form of Demirjian's

ten stage method and two teeth regression formula. J Phys. 2017;884:012070.

46. Alsaffar H, Elshehawi W, Roberts G, et al. Dental age estimation of children and adolescents: validation of the Maltese reference data set. J Forensic Leg Med. 2017;45:29–31.
47. Carneiro JL, Caldas IM, Afonso A, et al. Examining the socioeconomic effects on third molar maturation in a Portuguese sample of children, adolescents and young adults. Int J Leg Med. 2017;131: 235–242.
48. May LK, Shian AYM, Durward C, et al. A method of estimating age of undocumented children and young adults of different socioeconomic status in Cambodia. Heliyon. 2020;6: e03476.
49. Qing M, Qiu L, Gao Z, et al. The chronological age estimation of third molar mineralization of Han population in southwestern China. J Forensic Leg Med. 2014;24:24–27.
50. Liversidge HM, Smith BH, Maber M. Bias and accuracy of age estimation using developing teeth in 946 children. Am J Phys Anthropol. 2010;143:545–554.
51. Boedi RM, Banar N, De Tobel J, et al. Effect of lower third molar segmentations on automated tooth development staging using a convolutional neural network. J Forensic Sci. 2020;65: 481–486.
52. De Tobel J, Radesh P, Vandermeulen D, et al. An automated technique to stage lower third molar development on panoramic radiographs for age estimation: a pilot study. J Forensic OdontoStomatol. 2017;35:42–54.
53. Kullman L, Tronje G, Teivens A, et al. Methods of reducing observer variation in age estimation from panoramic radiographs. Dentomaxillofac Radiol. 1996;25:173–178.
54. Acharya AB. Accuracy of predicting 18 years of age from mandibular third molar development in an Indian sample using Demirjian's ten-stage criteria. Int J Leg Med. 2011;125:227–233.
55. de Oliveira FT, Capelozza ALÁ, Lauris JRP, et al. Mineralization of mandibular third molars can estimate chronological age—Brazilian indices. Forensic Sci Int. 2012;219:147–150.
56. Friedrich RE, Ulbricht C, Baronesse von Maydell LA. The influence of wisdom tooth impaction on root formation. Ann Anat. 2003;185:481–492.
57. Johan NA, Khamis MF, Abdul Jamal NS, et al. The variability of lower third molar development in Northeast Malaysian population with application to age estimation. J Forensic Odonto Stomatol. 2012;30:45–54.
58. Ramaswami TB, da Rosa GC, Fernandes MM, et al. Third molar development by Demirjian's stages and age estimation among Brazilians. Forensic Imaging. 2020;20:200353–200323.
59. Selmanagić A, Ajanović M, Kamber-Ćesir A, et al. Radiological evaluation of dental age assessment based on the development of third molars in population of Bosnia and Herzegovina. Acta Stomatol Croat. 2020;54:161–167. Croatian.
60. Streckbein P, Reichert I, Verhoff MA, et al. Estimation of legal age using calcification stages of third molars in living individuals. Sci Justice. 2014;54:447–450.
61. Thevissen PW, Fieuws S, Willems G. Third molar development: evaluation of nine tooth development registration techniques for age estimations. J Forensic Sci. 2013;58:393–397.
62. Uys A, Bernitz H, Pretorius S, et al. Estimating age and the probability of being at least 18 years of age using third molars: a comparison between Black and White individuals living in South Africa. Int J Leg Med. 2018;132:1437–1446.