



Analysis of interrater reliability in age assessment of minors: how does expertise influence the evaluation?

Lorenzo Franceschetti^{1,2} · Vera Gloria Merelli¹ · Susanna Corona¹ · Francesca Magli¹ · Lidia Maggioni¹ · Marco Cummaudo¹ · Stefania Tritella³ · Danilo De Angelis¹ · Cristina Cattaneo¹

Received: 1 May 2021 / Accepted: 16 September 2021
© The Author(s) 2021

Abstract

Nowadays, the clinical forensic medical management of migration flows comprises the age assessment of unaccompanied minors. The process of age estimation is a fundamental pillar for legally ensuring the minors' rights and their protection needs. The procedure is complex and involves different phases and actors, from medical doctors to law enforcement officers. The present study aimed to investigate the performance of Greulich and Pyle, Demirjian, and Mincer methods when performed by raters both trained and without training. Also, the interrater reliability within groups of raters from different areas of expertise was evaluated. A total of 36 participants were enrolled for this study, divided in two groups according to their level of experience with age estimation methods. Each rater was asked to allocate stages and standards for age assessment, evaluating ten orthopantomograms and ten hand-wrist roentgenograms. The interrater reliability expressed through the Fleiss Kappa coefficient and the agreement with the reference standard were calculated. The results showed that none of the categories analyzed could reach a good interrater reliability ($\kappa > 0.8$) for both methods. The study results highlighted variation and disagreement in the interpretation of the sample among raters and in the subsequent stages and standards allocation. In conclusion, the results of this study highlight that expertise does influence the reliability of the most utilized methods of age estimation of living individuals and stress the importance of proper training and practice, which could greatly increase the accuracy of age assessments.

Keywords Age assessment · Clinical forensic medicine · Forensic anthropology · Unaccompanied minors · Interrater reliability

Introduction

In the global phenomenon of migration, age assessment has gained increasing importance as many countries are obliged to regulate, sort, and “place” the great numbers of

individuals crossing their borders without identification documents. The management of migration flows comprises the identification and the protection of the most vulnerable among migrants, such as unaccompanied minors and asylum seekers. From a humanitarian perspective, the detention and treatment of minors as adults may have a negative impact on the individuals given their more susceptible condition to mental and emotional distress [1–3]. Today, age assessment not only does entail the legal guarantee of rights to minors, but also determines the criminal liability or conviction of adults involved in child pornography.

With regard to unaccompanied foreign minors, age estimation represents a fundamental step to ensure the fulfillment of their protection needs, and it is a complex procedure involving different phases and actors [4–9]. In the Italian context, the so-called Zampa Law (Law n.47/17) [10] provided comprehensive legislation aimed at filling existing gaps in the protection of unaccompanied minors arriving

✉ Lorenzo Franceschetti
lorenzo.franceschetti@unibs.it

¹ LABANOF, Laboratorio Di Antropologia E Odontologia Forense, Department of Biomedical Sciences for Health, University of Milan, Via Luigi Mangiagalli 37, 20133 Milan, Italy

² Forensic Medicine Unit, Department of Medical and Surgical Specialties, Radiological Sciences and Public Health, University of Brescia, Piazzale Spedali Civili, 1, 25123 Brescia, Italy

³ Unit of Radiology, IRCCS Policlinico San Donato, Via Morandi 30, San Donato Milanese, 20097 Milan, Italy

in Italy. Moreover, it introduced important provisions on age assessment procedure, which should only be performed when there is a reasonable doubt concerning a child's age and by using the least invasive methods possible. Nevertheless, the law does not establish a specific trained professional figure responsible for age estimation. Different professionals can be in charge of the process, from medical doctors to law enforcement officers, often with inter-regional differences. In 2008, the *Study Group on Forensic Age Diagnostic* (AGFAD) provided recommendations for age estimation in the living [11], consisting in a three-step procedure which includes a physical examination (anthropometric data, assessment of sexual maturation and identification of potential age-relevant developmental disorders) and an evaluation of the dental status, along with X-ray examinations of the dentition, the left hand, and the clavicle. This latter method is used when the bones of the hand and wrist have completed their development. Indeed, the clavicle has an extended developmental period of its medial epiphysis, thereby providing accurate age estimates of young adults [12]. With a simple three-phase scoring system, the analysis of the medial clavicular epiphysis proved to be the least subjective, while retaining accuracy levels [13].

Although developed several decades ago, the Greulich and Pyle "Radiographic Atlas of Skeletal Development of the Hand and Wrist" [14] and the methods on dental development by Demirjian and colleagues [15] and Mincer and colleagues [16] are still commonly utilized in forensic practice and have been tested in different populations across the world [17]. However, at present, there are no studies that have examined how the level of expertise of the rater can influence the performance of such methods in the age assessment.

In this perspective, the present study has a twofold objective: On the one hand, it aims to investigate the performance of Greulich and Pyle, Demirjian, and Mincer methods when performed by raters trained on age assessment and raters without training; on the other hand, it aims to assess the reliability of agreement within six groups of raters belonging to different areas of expertise (forensic physicians, odontologists, anthropologists, radiologists, non-forensic physicians, and medical students).

Materials and method

The sample sent to the participants consisted of ten orthopantomograms and ten hand-wrist roentgenograms belonging to twenty subjects from a database owned by the University Institute of Legal Medicine of Milan. The full database, obtained from the Sesto San Giovanni Hospital (Milan), consists of 385 orthopantomograms and 55 hand-wrist roentgenograms belonging to individuals ranging from 8 to

25 years. For the present study, the best radiographs in terms of image quality were chosen from subjects aged between 8 and 19 years (mean = 13.19 with SD = 3.11), with an equal ratio of boys and girls (1:1). No individuals presented any congenital or acquired malformation. An example of the questionnaire is provided in Supplementary Materials 1 (a 12-year-old male) and 2 (a 17-year-old female).

A total of 36 participants were enrolled for this study, divided in two main groups according to their level of experience with age estimation methods. The first group included six forensic physicians, six odontologists, six anthropologists, and six radiologists. Among them, four had sporadic experience in age assessment in the living (two anthropologists, a radiologist, and a forensic physician), whereas the remaining raters had only a basic training in age estimation during their academic education. The second group included six non-forensic physicians and six medical students without any training nor experience in age estimation methods.

Each rater received a survey comprising ten orthopantomograms and ten hand-wrist roentgenograms from twenty subjects, and they were asked to allocate stages and standards for age assessment of the individuals by applying the Demirjian method [15] and Greulich and Pyle atlas [14]. Whenever the highest Demirjian score was reached (98.4 for males and 100 for females), the Mincer method—based on wisdom teeth—was applied. This latter method explores the maturation stage of the third molar, which is typically the only tooth still in development during the young adult age [16]. The third molar is usually considered the most variable tooth of our dentition. The Mincer method is useful in age estimation in those cases in which the Demirjian score has reached its maximum discriminatory potential. A total of 740 determinations were performed.

The statistical analysis was conducted by calculating the interrater reliability expressed through the Fleiss Kappa coefficient [18]. Data obtained from surveys were recorded and entered in a digital data set and subsequently analyzed using Excel® software. In order to calculate the Fleiss Kappa coefficient, samples were organized according to age groups, e.g., 8–10 and 11–13, for a total of six categories. The Fleiss kappa was calculated for each group of raters (forensic physicians, odontologists, anthropologists, radiologists, non-forensic physicians, and medical students) and for groups according to the level of experience in age assessments (sporadic experience vs no experience). In addition, the allocations made by two professionals with over 10 years of experience in age assessment (an anthropologist and an odontologist) were used as the "reference standard" for the hand and wrist and dental age estimates respectively. These data were utilized in order to test the agreement between the "reference standard" and the allocations of stages and standards made by each of the other categories (e.g., radiologists vs "reference standard," forensic physicians vs "reference

standard,” anthropologists vs “reference standard”). For the dental age estimate, the agreement was analyzed not among estimated ages but among the stages assessed for each tooth of every radiograph, converting stages A–H to numbers (from 1 to 8).

Results

Interrater reliability

The results of the statistical analysis for the interrater reliability are shown in Fig. 1. Overall, the highest Fleiss Kappa for the dental age estimates was obtained by forensic physicians (0.54), followed by odontologists (0.49), anthropologists (0.41), radiologists (0.36), non-forensic physicians (0.35), and medical students (0.34). Concerning the Greulich and Pyle atlas (hand-wrist), the highest interrater reliability was attained by anthropologists (0.72), followed by forensic physicians (0.41), radiologists (0.33), medical students (0.32), non-forensic physicians (0.14), and odontologists (0.07).

Influence of the expertise of the participants

As illustrated in Fig. 2, according to the expertise, the “expert group” achieved a higher interrater reliability compared to the group without experience. In fact, the analysis of Fleiss kappa coefficient for raters with sporadic or continuous experience in age estimation resulted in a Fleiss kappa of 0.37 for the dental methods and 0.70 for the hand-wrist method. The raters without experience obtained a Fleiss kappa of 0.34 for the dental methods and 0.32 for the hand-wrist method.

Fig. 1 Fleiss kappa by professional category

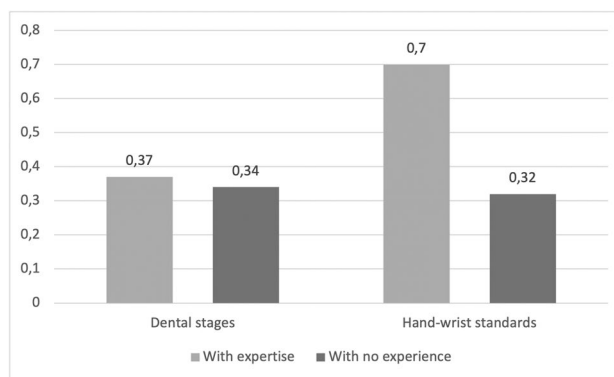
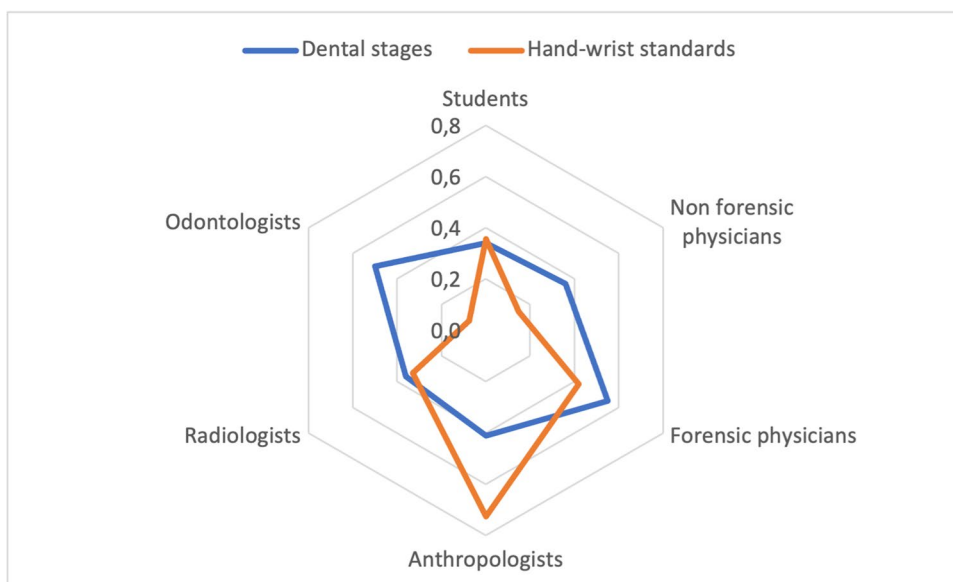


Fig. 2 Fleiss kappa by experience

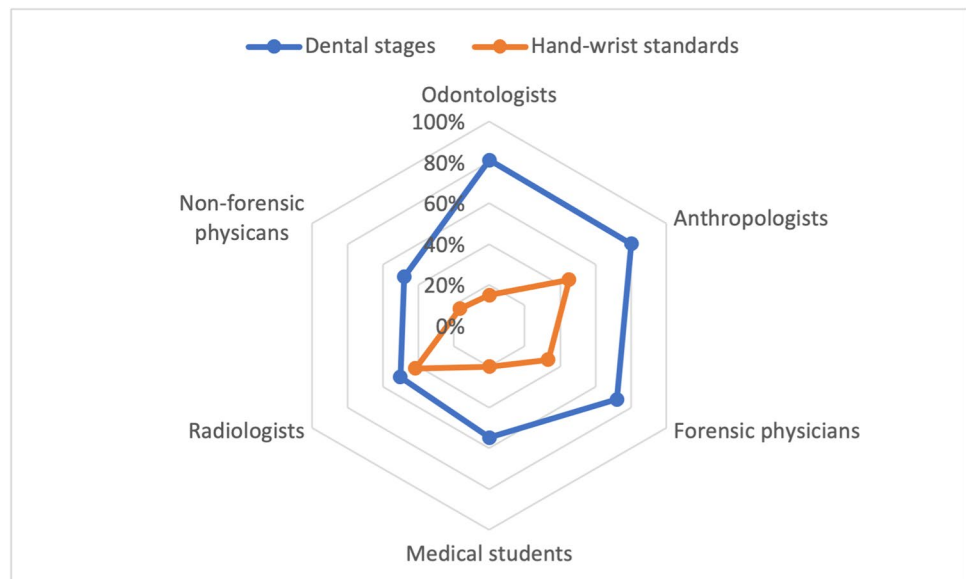
Comparison with the “reference standard”

The comparison between each category of raters with the “reference standard” (Fig. 3) have shown that the highest rate of agreement for dental methods was obtained by forensic anthropologists (81%), followed by odontologists (76%), forensic physicians (63%), radiologists (38%), non-forensic physicians (30%), and medical students (26%). With regard to Greulich and Pyle atlas, the highest rate of agreement was achieved by forensic anthropologists (45%), followed by radiologists (30%), forensic physicians and medical students (20%), non-forensic physicians (16%) and odontologists (15%).

Discussion

Age assessment of the living has become an important pillar of the forensic practice [17]. One of the main problems regarding age estimation is the impossibility to get an

Fig. 3 Percentages of agreement between each category of raters with the reference standard



accurate and definitive result [11, 19–25]. The main reason for this is the intrinsic biological variability of development among different individuals, depending on both genetic and environmental conditions. Each child has a specific and individual pattern of growth that can vary widely depending on the socio-economical, nutritional, and social status of the child as well as his/her level of physical activity [26–32].

Despite a number of methods for age estimation of living individuals having been developed and tested on several populations across the world [17], it is still unclear whether these methods can be applied successfully by raters belonging to different areas of expertise [33–36]. Moreover, no study has yet investigated how the rater's experience in applying those methods can influence the reliability of the age estimate. Although population studies are needed, studies on intrinsic method precision may be even more meaningful than testing on continuously changing populations, especially with the current increase in migration flows. In fact, a study by Thevissen and colleagues [6] showed that country-specific databases hardly increased the mean absolute difference. Consistently, a meta-analysis of published data from retrospective studies of dental maturity from eight countries showed no major statistical differences in the timing of tooth formation stages [34]. These studies were both on dental development, which is known to be less affected by environmental factors than skeletal development. Even regarding skeletal development, however, today's populations are far from constant and uniform. A sample of a given population at a certain time is not necessarily representative of the same population at another time, as a period of time has passed characterized by emigration or immigration flows. Moreover, as the interval of time extends, environmental conditions such as nutrition habits or basic quality of life also could have changed.

In this regard, the interobserver reliability of the most commonly utilized methods for skeletal and dental age estimation was evaluated. This was done on the grounds that the replication of results means that a method and the results obtained from it are valid. Indeed, the calculation of the interobserver error is critical for an objective comparison of different methods, independently of the sample type. The Fleiss kappa coefficient represents the most popular index to evaluate interrater agreement: In particular, a κ value > 0.8 is generally considered the minimum satisfactory result [18, 37]. Although Cohen's Kappa coefficient is also frequently used in forensic anthropology, it is not without problems that can yield misleading conclusions under certain conditions [38, 39].

The present study demonstrated how none of the categories analyzed (forensic physicians, forensic anthropologists, odontologists, radiologists, non-forensic physicians, and medicine students) could reach a good interrater reliability ($\kappa > 0.8$) for both dental and skeletal methods. The highest kappa coefficient was reached by forensic anthropologists (0.73) for the Greulich and Pyle method and by forensic physicians for dental methods (0.55). Overall, the dental methods obtained a higher degree of congruence compared to the Greulich and Pyle atlas. It should be considered that with the exception of the six experts, all the other raters received only a basic training (or no training at all) on skeletal and dental age estimation methods during their academic education. None of them has ever had a direct experience in age estimation in the living. Age assessment requires a specific set of skills (both theoretical and practical) that cannot be fully acquired during the work experience of the different medical and non-medical professionals who can deal with age estimation, such as those considered in this study. In this respect, as expected, the "expert group" achieved a higher interrater reliability compared to the group without experience (non-forensic

physicians and medical students). It is interesting to note that higher interrater reliability could be achieved by trained personnel using the hand/wrist standards and that the dental stage method was overall less reliable and less sensitive to training. This may be due to the fact dental methods consist of the sum of different stages on many teeth and therefore could be subjected to more variability, compared to the qualitative and more general hand/wrist methods which are less structured. However, the possible disagreement between raters concerns mainly the second and thirds molars, hence reducing the number of variables observed in the dentition making it more comparable to the hand-wrist assessment.

Consequently, understanding the characteristics of dental and skeletal indicators of development and the standardization of their description remains a crucial topic to address for correct age estimation, documentation of comparable data, and accurate assessment. It is likely that specific training in these methods will increase the accuracy rates and reduce the variation observed among participants, regardless of their field of specialization or experience.

In the European context, where a proper ascertainment of age in minors is frequently badly managed [9], this is extremely important. Practical guidelines for age assessment of minors [40, 41] recommend adopting a multidisciplinary holistic approach [2]. In Italy, the daily practice does not include the existence of specifically trained personnel in hospitals nor do most physicians or policymakers know about the above-mentioned standards, apart from the smaller forensic community which is not present in most hospitals. Radiological methods are frequently left as a last resort, even if they are more quantifiable and reliable than neuropsychiatric and psychological evaluations [42]. Professionals with very different types of training may perform the age estimate, from social workers to psychologists, to medical doctors and law enforcement officers, with consequent paradoxes and poor administration of the rights of minors.

Limitations of this study include the limited number of radiographs analyzed. The number can be considered sufficient for the analysis of the interrater reliability (which is the main focus of this study), whereas, in order to test the accuracy of the age assessments when compared to the real ages, a larger sample would have been more adequate. However, this can be considered a pilot study, and further studies must be conducted on larger samples. In addition, although the present research showed how expertise plays a role in the method's applicability, it refers to stage allocation for teeth and hand/wrist, and not to the specific age estimate. Further considerations about age estimations based on the stages and standards can be made only following the examination of a larger population. Moreover, this investigation involved only Italian experts in different areas, but the recruitment of more experts, including those of the international community, is likely to take place in subsequent research. Finally, it might be interesting to evaluate how

the quality of radiographic images can influence the interrater reliability of both skeletal and dental methods.

In conclusion, the results of this study highlight that expertise does have an effect on the reliability of the most commonly utilized methods of age estimation of living individuals and show the importance of proper training and practice, which could greatly increase the accuracy of age assessments.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00414-021-02707-8>.

Availability of data and material The authors confirm that the data supporting the findings of this study are available within the article.

Code availability Not applicable.

Declarations

Ethics approval Institutional Review Board approval for the study was obtained. The participants had given specific informed consent for the use of their data for research purposes.

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Council of Europe convention on action against trafficking in human beings (n.d.) <https://rm.coe.int/168008371d>. Accessed 30 Jul 2021
2. Aynsley-Green A, Cole TJ, Crawley H, Lessof N, Boag LR, Wallace RM (2012) Medical, statistical, ethical and human rights considerations in the assessment of age in children and young people subject to immigration control. *Br Med Bull* 102:17–42. <https://doi.org/10.1093/bmb/lds014>
3. Annex II (2000) Protocol to prevent, suppress and punish trafficking in persons, especially women and children, supplementing the United Nations Convention against transnational organized crime. In: *Traffick. Hum. Beings from a Hum. Rights Perspect.* <https://doi.org/10.1163/ej.9789004154056.i-247.45>
4. Smith T, Brownlees L (2011) Age assessment practices: a literature review & annotated bibliography. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.309.2374&rep=rep1&type=pdf>. Accessed 30 Jul 2021
5. Andersen E (1971) Comparison of Tanner-Whitehouse and Greulich-Pyle methods in a large scale Danish survey. *Am J Phys Anthropol* 35:373–376. <https://doi.org/10.1002/ajpa.1330350312>

6. Thevissen PW, Alqerban A, Asaumi J, Kahveci F, Kaur J, Kim YK, Pittayapat P, Van Vlierbergh M, Zhang Y, Fieuws S, Willems G (2010) Human dental age estimation using third molar developmental stages: accuracy of age predictions not using country specific information. *Forensic Sci Int* 201:106–111. <https://doi.org/10.1016/j.forsciint.2010.04.040>
7. UNICEF (2010) Progress for children: achieving the MDGs with equity
8. Rozzi E (2017) The new Italian law on unaccompanied minors: a model for the EU? – EU Immigration and Asylum Law and Policy, Assoc. per Gli Stud. Giuridici Sull' Immigrazione. <http://eumigrationlawblog.eu/the-new-italian-law-on-unaccompanied-minors-a-model-for-the-eu/>. Accessed 30 Jul 2021
9. Cummaudo M, De Angelis D, De Micco F, Campobasso C, Cattaneo C (2021) The “forensic paradox” of aging unaccompanied minors in the migration crisis: why medicine and forensics are a must. *J Forensic Leg Med* 79:102133
10. Gazzetta Ufficiale della Repubblica Italiana. Italian law n°47/2017 Disposizioni in materia di misure di protezione dei minori stranieri non accompagnati”. <https://www.gazzettaufficiale.it/eli/id/2017/04/21/17G00062/sg/>. Accessed 30 Jul 2021
11. Schmeling A, Grundmann C, Fuhrmann A, Kaatsch HJ, Knell B, Ramsthaler F, Reisinger W, Riepert T, Ritz-Timme S, Rösing FW, Röttscher K, Geserick G (2008) Criteria for age estimation in living individuals. *Int J Legal Med* 122:457–460. <https://doi.org/10.1007/s00414-008-0254-2>
12. Langley NR (2016) The lateral clavicular epiphysis: fusion timing and age estimation. *Int J Legal Med* 130:511–517. <https://doi.org/10.1007/s00414-015-1236-9>
13. Langley-Shirley N, Jantz RL (2010) A Bayesian approach to age estimation in modern Americans from the clavicle. *J Forensic Sci* 55:571–583. <https://doi.org/10.1111/j.1556-4029.2010.01089.x>
14. Greulich WW, Pyle SI (1959) Radiographic Atlas of Skeletal Development of the Hand and Wrist, 2nd edn. Stanford University Press, Stanford, CA
15. Demirjian A, Goldstein H, Tanner J (1973) A new system of dental age assessment. *Hum Biol* 45:211–227
16. Mincer JJ, Harris EF, Berryman HE (1993) The A.B.F.O. study of third molar development and its use as an estimator of chronological age. *J Forensic Sci* 38:379–390
17. Cummaudo M, De Angelis D, Magli F, Minà G, Merelli V, Cattaneo C (2021) Age estimation in the living: a scoping review of population data for skeletal and dental methods. *Forensic Sci Int* 320:110689. <https://doi.org/10.1016/j.forsciint.2021.110689>
18. McHugh ML (2012), Interrater reliability: the kappa statistic, *Biochem. Medica*. 22:276–282. <https://doi.org/10.11613/bm.2012.031>.
19. Thevissen PW, Kvaal SI, Dierickx K, Willems G (2012) Ethics in age estimation of unaccompanied minors. *J Forensic Odontostomatol* 30:85–102
20. Cole TJ (2015) The evidential value of developmental age imaging for assessing age of majority. *Ann Hum Biol* 42:379–388
21. Lopes LJ, Nascimento HAR, Lima GP, dos Santos LAN, de P. Queluz D, Freitas DQ. (2018) Dental age assessment: which is the most applicable method? *Forensic Sci Int* 284:97–100. <https://doi.org/10.1016/j.forsciint.2017.12.044>
22. Olze A, Reisinger W, Geserick G, Schmeling A (2006) Age estimation of unaccompanied minors. Part II. Dental aspects *Forensic Sci Int* 159:S65–S67. <https://doi.org/10.1016/j.forsciint.2006.02.018>
23. Olze A, Schmeling A, Taniguchi M, Maeda H, Van Niekerk P, Wernecke KD, Geserick G (2004) Forensic age estimation in living subjects: the ethnic factor in wisdom tooth mineralization. *Int J Legal Med* 118:170–173. <https://doi.org/10.1007/s00414-004-0434-7>
24. Ritz-Timme S, Cattaneo C, Collins MJ, Waite ER, Schütz HW, Kaatsch HJ, Borrman HIM (2000) Age estimation: the state of the art in relation to the specific demands of forensic practise. *Int J Legal Med* 113:129–136. <https://doi.org/10.1007/s004140050283>
25. Royal College of Paediatrics and Child Health, ed., Age assessment of separated young people: proposal to develop practical guidance for paediatricians, 2012.
26. Franklin D, Flavel A, Noble J, Swift L, Karkhanis S (2015) Forensic age estimation in living individuals: methodological considerations in the context of medico-legal practice. *Res Reports Forensic Med Sci* 5:53–56. <https://doi.org/10.2147/rrfms.s75140>
27. Guo G, Mu G. Human age estimation: what is the influence across race and gender?. In: 2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Work. CVPRW 2010, 2010: pp. 71–78. <https://doi.org/10.1109/CVPRW.2010.5543609>.
28. Pinchi V, Bugelli V, Vitale G, Pradella F, Farese L, Focardi M (2018) Dental age estimation in children with chromosomal syndromes. *J Forensic Odontostomatol* 36:44–52
29. Santoro V, De Donno A, Marrone M, Campobasso CP, Introna F (2009) Forensic age estimation of living individuals: a retrospective analysis, *Forensic Sci. Int* 193:129.e1-129.e4. <https://doi.org/10.1016/j.forsciint.2009.09.014>
30. Schmeling A, Dettmeyer R, Rudolf E, Vieth V, Geserick G (2016) Forensic age estimation: methods, certainty, and the law. *Dtsch Aerzteblatt Online* 113:40–53. <https://doi.org/10.3238/arztebl.2016.0044>
31. Thevissen PW, Kaur J, Willems G (2012) Human age estimation combining third molar and skeletal Development. *Int J Legal Med* 126:285–292. <https://doi.org/10.1007/s00414-011-0639-5>
32. Cunha E, Baccino E, Martrille L, Ramsthaler F, Prieto J, Schuliar Y, Lynnerup N, Cattaneo C (2009) The problem of aging human remains and living individuals: a review, *Forensic Sci. Int* 193:1–13. <https://doi.org/10.1016/j.forsciint.2009.09.008>
33. Lynnerup N, Belard E, Buch-Olsen K, Sejrsen B, Damgaard-Pedersen K (2008) Intra- and interobserver error of the Greulich-Pyle method as used on a Danish forensic sample, *Forensic Sci. Int* 179:242.e1-242.e6. <https://doi.org/10.1016/j.forsciint.2008.05.005>
34. Maber M, Liversidge HM, Hector MP (2006) Accuracy of age estimation of radiographic methods using developing teeth. *Forensic Sci Int* 159(Suppl 1):S68-73. <https://doi.org/10.1016/j.forsciint.2006.02.019>
35. Schmidt S, Koch B, Schulz R, Reisinger W, Schmeling A (2007) Comparative analysis of the applicability of the skeletal age determination methods of Greulich-Pyle and Thiemann-Nitz for forensic age estimation in living subjects. *Int J Legal Med* 121:293–296. <https://doi.org/10.1007/s00414-007-0165-7>
36. Schmidt S, Nitz I, Ribbecke S, Schulz R, Pfeiffer H, Schmeling A (2013) Skeletal age determination of the hand: a comparison of methods. *Int J Legal Med* 127:691–698. <https://doi.org/10.1007/s00414-013-0845-4>
37. Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20:37–46
38. Flight L, Julious SA (2015) The disagreeable behaviour of the kappa statistic. *Pharm Stat* 14:74–78. <https://doi.org/10.1002/pst.1659>
39. Shirley NR, Ramirez Montes PA (2015) Age estimation in forensic anthropology: quantification of observer error in phase versus component-based methods. *J Forensic Sci* 60(1):107–111. <https://doi.org/10.1111/1556-4029.12617>
40. European Asylum Support Office (EASO). *EASO practical guide on age assessment*. second ed.; 2018. <https://www.easo.europa.eu/easo-practical-guide-on-age-assessment>

europa.eu/sites/default/files/easo-practical-guide-on-age-assessment-v3-2018.pdf. Accessed 30 Jul 2021

41. UN children's fund (UNICEF), age assessment: a technical note, 2013. <https://www.refworld.org/docid/5130659f2.html/>. Accessed 30 Jul 2021
42. Schmeling A, Reisinger W, Loreck D, Vendura K, Markus W, Geserick G (2000) Effects of ethnicity on skeletal maturation: consequences for forensic age estimations. *Int J Leg Med* 113:253–258. <https://doi.org/10.1007/s004149900102>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.