Int J Legal Med (2009) 123:277–283
DOI 10.1007/s00414-009-0349-4

REVIEW ARTICLE

Statistical methods to assess the reliability of measurements in the procedures for forensic age estimation

L. Ferrante · R. Cameriere

Received: 21 October 2008 / Accepted: 8 April 2009 / Published online: 2 May 2009

Abstract In forensic science, anthropology, and archaeology, several techniques have been developed to estimate chronological age in both children and adults, using the relationship between age and morphological changes in the structure of teeth. Before implementing a statistical model to describe age as a function of the measured morphological variables, the reliability of the measurements of these variables must be evaluated using suitable statistical methods. This paper introduces some commonly used statistical methods for assessing the reliability of procedures for age estimation in the forensic field. The use of the concordance correlation coefficient and the intraclass correlation coefficient are explained. Finally, some pitfalls in the choice of the statistical methods to assess reliability of the measurements in age estimation are discussed.

Keywords Reliability · Age estimation · Forensic odontology · Concordance correlation coefficient · Intraclass correlation coefficient

Introduction

Generally, when medical examiners are asked to estimate the chronological age of a person, they must ask themselves two questions: which is the most appropriate method and, above all, how reliable is a particular method? In forensic science, anthropology, and archaeology, many studies deal with the estimation of chronological age in humans, but few are devoted to studying the accuracy and reliability of the results. The most widespread methods for age estimation are based on skeletal indicators such as epiphysial fusion [1–3], hand–wrist bones [4–8], changes in the pubic symphysis [9–12], fusion of cranial sutures [12, 13], dental maturation [14–22], and combined method [23].

Restricting analysis to forensic odontology, several techniques have also been developed to estimate chronological age in both children and adults, using the relationship between age and morphological changes in the structure of teeth.

Examples (among others) are the atlas method of Schour and Massler [24], specific standards from assessment of radiographic stages according to Demirjian et al. [25], the diagram of Gustafson and Koch [26], and the length and weight regression equations of Deutsch et al. [27].

More recently, some revisions of previous methods have been made, for example, by Liversidge [28] and Willems et al. [29], who proposed a revised Demirjian method. Another recent method is measurement of the amplitude of the open apex of incompletely developed teeth according to Cameriere et al. [30]. All these methods share three phases in predicting the age of an individual:

- Information gathering and data collection
- Formulation and identification of a statistical model to describe age as a function of the measured morphological variables
- Validation of the statistical model

In the first step, variables related to developing teeth, such as secondary dentine apposition, tooth length, area of tooth and pulp, distance between the inner sides of the open

L. Ferrante
Dipartimento di Medicina Clinica e Biotecnologie Applicate,
Facoltà di Medicina e Chirurgia,
Via Tronto 10/A,
60020 Ancona, Italy

R. Cameriere (✉)
AgEstination project, Institute of Legal Medicine,
University of Macerata,
Via Don Minzoni, 9,
62100 Macerata, Italy
e-mail: r.cameriere@unimc.it

apex of an immature tooth, etc, are measured from a digitalized dental X-ray and entered in a database for use as predictive variables for age estimation in subsequent statistical analysis.

To assess the extent to which the given measurement of a feature of a tooth is reliable, we must measure a number of teeth more than once. This may occur, for example, when measurements are carried out by two different observers or by the same observer on two different occasions. Subsequently, after formulating the statistical model and estimating the dental age (second step), we need to estimate the accuracy of the method by comparing the dental age of each individual with the chronological age (third step).

A fundamental problem is how to keep observational errors under control, i.e., how to take accurate and precise measurements.

If we look up the definition of these two words, *accuracy* and *precision*, in an English dictionary, we find that *Accurate* means "in exact conformity to truth or to a standard or rule, or to a model; free from error or defect; *precise*." Hence, in everyday language, accuracy and precision have the same meaning. But, in the field of statistics, and above all in applied statistics, they have two different meanings:

> *Accuracy*, also called *validity*, is the degree of conformity of a measured or calculated quantity to its (actual) true value. Accuracy means without bias.
> *Precision*, also called *reliability* or average deviation, is the degree to which further measurements or calculations give the same or similar results. Precision means small error.

The analogy usually used to explain the difference between accuracy and precision is the target comparison. In this analogy, repeated measurements are compared with arrows fired at a target (Fig. 1). Accuracy describes the closeness of the average position of the arrows to the bull's-eye. Arrows which strike closer to the bull's-eye are considered to be more accurate. The closer the average of the measurements of the variable to the actual value, the bull's-eye, the more accurate the method is considered to be. When all the arrows are grouped tightly together, the cluster is called *precise*, since they all struck close to the same spot, if not necessarily near the bull's-eye. The measurements are precise, although not necessarily accurate. Vice versa, measurements may be accurate but not necessarily precise.

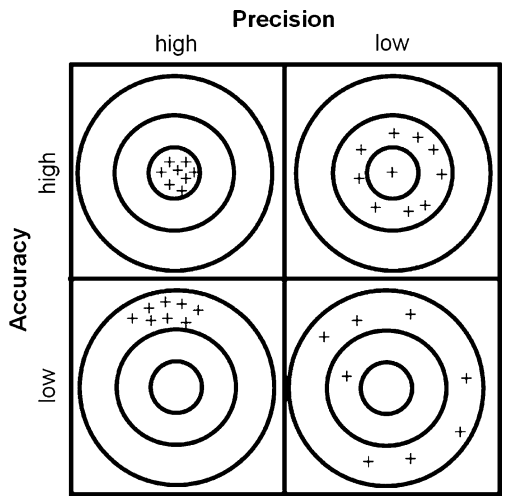

Fig. 1 Target analogy of accuracy and precision

The aims of this paper are (1) to analyze the statistical methods to assess the reliability of procedures for age estimation in the forensic field and (2) to discuss some pitfalls in the choice of these methods.

Reliability of measurements

In the first phase of data collection, we are more interested in controlling the variability of the measures rather than their closeness to the true value. Consequently, precision in this phase is a more important concept than accuracy. Before implementing a statistical model to describe age (dependent variable) as a function of the measured morphological variables, we must evaluate the reliability of the measurements of these variables. As previously noted, to assess the reliability of a measurement, we must measure a number of teeth, more than once, either when measurements are carried out by two different observers or by the same observer on two different occasions. As a consequence, the *precision* (*reliability*) of measurements is usually classified as:

> *Repeatability*—maximum effort is made to keep conditions constant by using the same observer (intra-observer reliability);
> *Reproducibility*—the same measurement process is successfully carried out by different observers (inter-observer reliability).

When the measurement of a feature of a tooth is repeated twice on the same sample by two different observers or by the same observer on two different occasions, the result may be two readings that differ from each other. The cause may be small differences in how the observer or a possible second observer uses the measurement device (instrument). However, they may also be due to small random changes in the morphological variable itself, while being measured. Whatever the cause, it is of interest to evaluate measurement precision, that is, inter-observer and intra-observer reliability. We usually implement such evaluation processes, for both intra-observer and inter-observer agreement, by means of various statistical methodologies, depending on

Table 1 2×2 table for evaluation of precision for categorical variables

		Obs.1		
		Cat_1	Cat_2	
Obs.2	Cat_1	a	b	p_1
	Cat_2	c	d	q_1
		p_2	q_2	1

whether the measured morphological variable is qualitative or quantitative.

Evaluation of reliability for qualitative variables: Cohen's kappa coefficient

To measure the reliability of qualitative (categorical or ordinal) variables, as with Demirjian stages, we calculate Cohen's kappa coefficient [31, 32].

Kappa measures the percentage of data values in the main diagonal of the table and then adjusts these values by the amount of agreement that could be expected due to chance alone. Let us suppose that two observers are asked to classify a qualitative characteristic of a tooth into categories 1 and 2 (Table 1). Consider the 2×2 table, in which each cell entry is the proportion of teeth classified into one of the two diagnostic categories by observer 1 and into another by observer 2.

Kappa measures the agreement (precision of the categorical variable) between the observers, adjusted by the amount of agreement that could be expected due to chance alone.

The value of Kappa is defined by the following formula:

$$K = \frac{a + d - (p_1 p_2 + q_1 q_2)}{1 - (p_1 p_2 + q_1 q_2)}$$

The numerator represents the discrepancy between the observed probability of agreement between the two observers and the probability of agreement between them, under the assumption that the agreement was by chance. The denominator represents the maximum of the numerator. As the observed probability of agreement decreases, so does the numerator. Hence, Kappa is always for less than or equal to 1. A value of 1 implies perfect agreement, and values less than 1 imply less than perfect agreement. In rare situations, Kappa may be negative. This is a sign that the two observers agreed less than would be expected just by chance. The interpretation of Kappa values may be summarized as follows:

K less than 0.4: poor agreement between observers
K greater than or equal to 0.40 and *K* less than 0.60: moderate agreement
K greater than or equal to 0.60 and *K* less than 0.80: good agreement
K greater than or equal to 0.80 and *K* less than 1.0: very good agreement

Table 2 2×2 table for evaluation of inter-observer reproducibility

		Obs.1		
		H	$\overline{H}$	
Obs.2	H	0.375	0.025	0.40
	$\overline{H}$	0.00	0.60	0.60
		0.375	0.625	1.00

Table 3 *S* values obtained by two raters

Rater 1			Rater 2		
1.011	1.055	0.398	1.005	0.967	0.323
0.897	0.378	1.063	0.919	0.462	1.132
0.687	0.867	0.532	0.739	0.880	0.534
0.870	0.303	0.083	0.838	0.291	0.136
1.325	0.580	0.687	1.333	0.524	0.598
0.584	1.069	0.687	0.615	1.087	0.659
1.287	0.399	0.278	1.397	0.339	0.331
1.078	0.385	0.311	1.090	0.395	0.313
1.070	0.412	1.492	1.123	0.440	1.472
0.712	0.699	0.388	0.669	0.598	0.349

Example 1 Following the grade scheme developed by Demirijan, two observers classified the stage of third molar development into two categories, depending on whether it had reached stage *H* or not. To test inter-observer

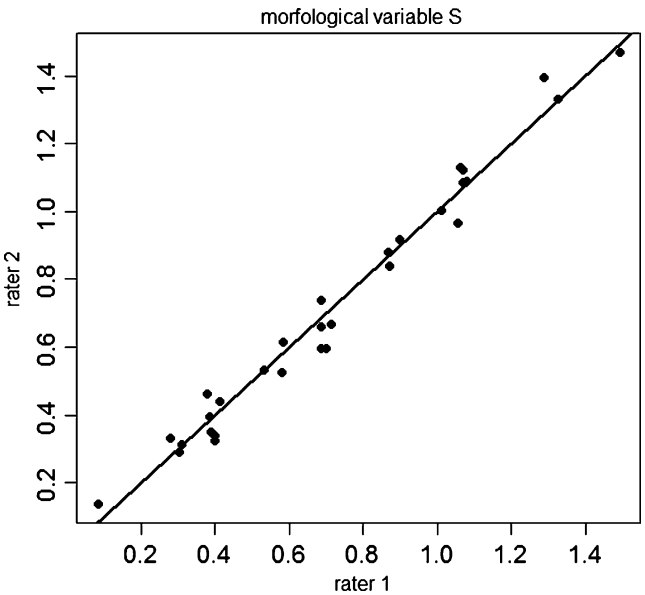

Fig. 2 Plot of measurements obtained by rater 1 against corresponding measures obtained by rater 2

Table 4 Simulated data

Rater 1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.9
Rater 2	0.4	0.47	0.5	0.55	0.58	0.65	0.7	0.81	1.27
Rater 1	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8
Rater 2	0.8	0.85	0.89	0.92	1	1.05	1	1.15	1.2

reproducibility, a random sample of 40 orthopantomographs were examined by two observers. The relative frequencies of agreement or disagreement of the two observations are reported in Table 2.

The cell corresponding to (*H*, *H*) contains the value 0.375. This means that 0.375×40=15 third molars were evaluated in phase *H* by both first and second raters, and 0.025×40=1 third molar was evaluated in phase *H* by the first and in phase *H* by the second rater.

Evaluated according to the data of Table 2, Cohen's *K* is 0.947, showing high intra-observer agreement.

Evaluation of precision for quantitative variables: concordance correlation coefficient (ρ_c) and intraclass correlation coefficient (ρ_I)

To assess the reliability of quantitative (numerical) variables—for example, tooth height, tooth length/root length ratio, pulp/tooth area ratio, or the sum of normalized open apices (*S*), we evaluate the concordance correlation coefficient (ρ_c) [33]:

$$\rho_c = 1 - \frac{\mathbb{E}\left[(Y_1 - Y_2)^2\right]}{\mathbb{E}\left[(Y_1 - Y_2)^2\right] \text{when } Y_1 \text{ and } Y_2 \text{ are uncorrelated}}$$

$$= \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}$$

Where Y_1 and Y_2 are the array of measurements of each rater, μ_1 and μ_2 are the observation means, and ${\sigma_1}^2$, ${\sigma_2}^2$, and σ_{12} are observation variances and covariance. The concordance correlation coefficient evaluates the degree to which pairs of observations fall on the line through the origin with slope equal to 1, that is, the bisector or the concordance line. If each pair of readings is in perfect agreement, then ρ_c is equal to 1. The more the two observers disagree, the more ρ_c approaches zero.

Example 2 To test the inter-observer reproducibility of the morphological variable *S* used to estimate chronological age, a random sample of 30 panoramic radiographs was examined by two raters, and measurements of the variable were reported in Table 3 and Fig. 2. The concordance correlation coefficient evaluated by the two data sets, ρ_c= 0.956, indicates very good agreement between the two observers, and consequently, there were no statistically significant inter-observer differences between the paired sets of measurements carried out on the re-examined panoramic radiographs.

In order to ascertain the reliability of the quantitative data, these validation processes are often evaluated by Pearson's correlation coefficient (*r*) or Student's paired *t* test. However, there are drawbacks to both these statistical approaches, as neither of them can fully assess the desired characteristics of reliability.

Example 3 To evaluate the reproducibility of a morphological variable of a tooth, e.g. its height, we must collect duplicates of the same tooth measurement by two raters, 1 and 2.

When we plot the first measurement against the second, we hope to see that the measurements fall on the bisector of the first quadrant within a tolerable range of error. This is not what happens, for example, to the measurements reported in Table 4 and Fig. 3, below.

Pearson's correlation coefficient measures a linear relationship but fails to detect any departure from the bisector (straight line). Indeed, the correlation coefficient, *r*=0.993, indicates strong agreement, and the *t* test fails to detect poor agreement in pairs of data (*t*=1.629; *p*=0.112 is not significant). Conversely, the concordance correlation coefficient indicates poor reproducibility of variable ρ_c=0.647.

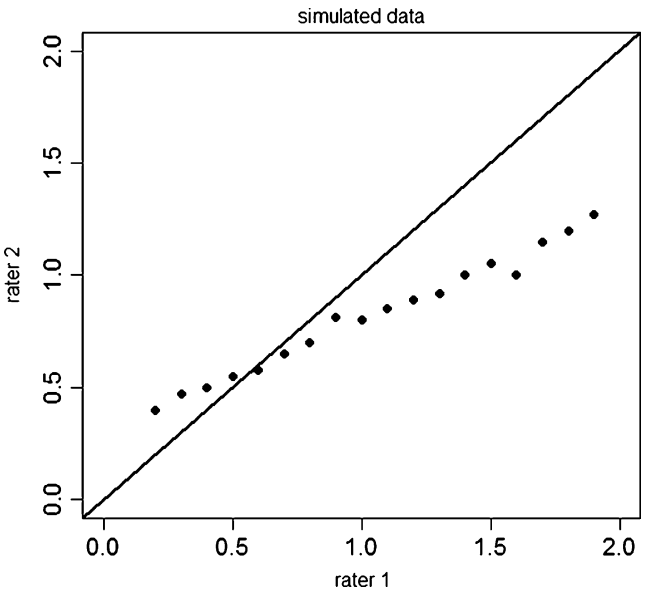

Fig. 3 Plot of simulated data shown in Table 4

Table 5 ANOVA summary

Source	*df*	Sum of Sq	Mean sum of Sq	Expected MS
Between subjects	$n-1$	SS_b	MS_b	$k\times\sigma_b^2+\sigma_e^2$
Within subjects	$n\times(k-1)$	SS_e	MS_e	σ_e^2

Consequently, in order to assess the reproducibility of measurement of a morphological variable, we suggest using the concordance correlation coefficient.

Besides the concordance correlation coefficient, another way of ascertaining inter- and intra-observer reliability is the intraclass correlation coefficient (ρ_I).

Intraclass correlation coefficients are alternative statistics for measuring reproducibility not only for pairs of measurements but also for larger sets of measurements.

There are various forms of intraclass correlation coefficient, and the classic reference is Shrout and Fleiss [34]. The form that we suggest in this paper is the intraclass correlation coefficient based on the random-effects model for one-way analysis of variance (ANOVA).

In terms of ANOVA, the interpretation of an intraclass correlation coefficient is the degree of absolute agreement among measurements made on randomly selected objects.

Let $Y_1=(y_{11}, y_{21},\ldots,y_{n1})$, $Y_2=(y_{12}, y_{22},\ldots,y_{n2})$,…, and $Y_k=(y_{1k}, y_{2k},\ldots,y_{nk})$ be k vectors of measurements of a feature of n teeth ("subjects" in terms of ANOVA) carried out by k different observers or by the same observer on k different occasions (rater 1, rater 2,…., rater k). The resulting ANOVA may be summarized as shown in Table 5.

Where:

$$SS_b = \sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2$$

with $\bar{y}_i = \frac{1}{k}\sum_{j=1}^{k} y_{ij}, i = 1, 2, \ldots, n$; the mean value of the measurements of the different raters on the same tooth ("subject") and $\bar{y} = \frac{1}{k\times n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}$ is the general mean or the mean of $y_i, i = 1, 2, \ldots, n$; and:

$$SS_e = \sum_{i=1}^{n}\sum_{j=1}^{k} \left(y_{ij} - \bar{y}_i\right)^2$$

Lastly, MS_b and MS_e can be obtained by dividing SS_b and SS_e by their respective degrees of freedom (*df*). If each measurement of each rater on the same subject had exactly the same value, there would be no within-subject variance, and all the variance in the experiment would be due to differences between subjects (remember, we are using ANOVA terms *between subjects* and *within subjects* to refer to what we would really think of as "between teeth" and "within teeth"). We can therefore obtain a measure of the degree of reliability by asking what proportion of the variance is between-subjects variance. Thus, we define our estimate of the correlation as the intraclass correlation coefficient:

$$\rho_I = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2} = \frac{MS_b - MS_e}{MS_b + (k-1)\times MS_e}$$

In the case of $k=2$ observations, the $(k-1)$ term drops out. The derivation shown above leads to the following formula for the intraclass correlation coefficient:

$$\rho_I = \frac{MS_b - MS_e}{MS_b + MS_e}$$

Example 4 To test the use of ρ_I in order to ascertain the reliability of the quantitative variables, we reconsider morphological variable *S*, used to estimate chronological age as considered in Example 2 and shown in Table 3. A good program for calculating intraclass correlation coefficients is given in *R* [35] and can be found in the "psy" package, which can be downloaded from the *R*-Project site (http://www.r-project.org). We carried out the analysis in *R* using the function "icc" and obtained the following ANOVA table (Table 6).

With the results listed in Table 6, we evaluated the intraclass correlation coefficient:

$$\rho_I = \frac{0.26013 - 0.00137}{0.26013 + 0.00137} = 0.9895$$

According to the value of ρ_c, estimated in Example 2, this value of ρ_I indicates very good agreement between the two observers, and consequently, there were no statistically significant inter-observer differences between the paired sets of measurements carried out on the re-examined panoramic radiographs.

Discussion

As in any estimation process, including age determination, statistics play an essential role in producing good-quality

Table 6 ANOVA summary for data in example 2

Source	*df*	Sum of Sq	Mean sum of Sq
Between subjects	29	7.544	0.26013
Within subjects	30	0.041	0.00137

results. In every phase of the process of age estimation, various statistic techniques are used, ranging from simple description of data to the formulation of an interpretative model.

When measurements of a morphological variable are collected, we are interested in their reliability, but usually, we cannot know their accuracy because the true value of the measured variable is unknown. In fact, the concept of accuracy, and consequently the concept of bias, is dependent on the actual knowing of the true value of the morphological variable.

In the age estimation processes, bias measures typically take into account the difference between predicted age, obtained using a statistical model, and chronological age [36]. When we have identified a statistical model to estimate the chronological age as function of the measured morphological variables, we can evaluate the performance of the proposed statistical model assessing the bias of the predicted age. For example, one common bias measure, called mean prediction error (ME), is the mean of the absolute value of the differences between the estimated and the true values [37].

If the statistical method applied is not suitable for the researchers' purpose, the quality of the results will be negatively influenced. In particular, as regards the precision (reproducibility) of the measurements of biological variables for age estimation, if not completely suitable statistical techniques are used, the consequence will be the production of false results, and indeed, in Example 3, analysis of the precision of measurements repeated by means of the correlation coefficient or the *t* test indicates good reproducibility, which is not in fact so (Fig. 3). In this case, the most suitable statistical instrument would be the concordance correlation coefficient or, alternatively, the intraclass correlation coefficient.

In conclusion, we suggest devoting the same care and attention applied to the choice of biological techniques for estimating age to the choice of statistical methods applied to data analysis.

Acknowledgments This work was supported by a grant from the Polytechnic University of Marche. Part of this research was presented at the FASE Meeting 2007, Macerata, Italy.

References

1. Cardoso HF (2008) Age estimation of adolescent and young adult male and female skeletons II, epiphyseal union at the upper limb and scapular girdle in a modern Portuguese skeletal sample. Am J Phys Anthropol 137:97–105
2. Schmeling A, Schulz R, Reisinger W, Mühler M, Wernecke KD, Geserick G (2004) Studies on the time frame for ossification of the medial clavicular epiphyseal cartilage in conventional radiography. Int J Legal Med 118:5–8
3. Banerjee KK, Agarwal BB (1998) Estimation of age from epiphyseal union at the wrist and ankle joints in the capital city of India. Forensic Sci Int 98:31–39
4. Cameriere R, Ferrante L, Mirtella D, Cingolani M (2006) Carpals and epiphyses of radius and ulna as age indicators. Int J Legal Med 120:143–146
5. Schmidt S, Baumann U, Schulz R, Reisinger W, Schmeling A (2008) Study of age dependence of epiphyseal ossification of the hand skeleton. Int J Legal Med 122:51–54
6. Castriota-Scanderbeg A, Sacco MC, Emberti-Gialloreti L, Fraracci L (1998) Skeletal age assessment in children and young adults: comparison between a newly developed sonographic method and conventional methods. Skeletal Radiol 27:271–277
7. 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
8. Büken B, Safak AA, Yazici B, Büken E, Mayda AS (2007) Is the assessment of bone age by the Greulich-Pyle method reliable at forensic age estimation for Turkish children? Forensic Sci Int 173:146–53
9. Jackes MK (1985) Pubic symphysis age distributions. Am J Phys Anthropol 68:281–99
10. Sinha A, Gupta V (1995) A study on estimation of age from pubic symphysis. Forensic Sci Int 75:73–8
11. Kimmerle EH, Konigsberg LW, Jantz RL, Baraybar JP (2008) Analysis of age-at-death estimation through the use of pubic symphyseal data. J Forensic Sci 53:558–568
12. Berg GE (2008) Pubic bone age estimation in adult women. J Forensic Sci 53:569–577
13. Dorandeu A, Coulibaly B, Piercecchi-Marti MD, Bartoli C, Gaudart J, Baccino E, Leonetti G (2008) Age-at-death estimation based on the study of frontosphenoidal sutures. Forensic Sci Int 177:47–51
14. Galera V, Ubelaker DH, Hayek LA (1998) Comparison of macroscopic cranial methods of age estimation applied to skeletons from the Terry Collection. J Forensic Sci 43:933–939
15. Ubelaker DH, Parra RC (2008) Application of three dental methods of adult age estimation from intact single rooted teeth to a Peruvian sample. J Forensic Sci 53:608–611
16. Meinl A, Huber CD, Tangl S, Gruber GM, Teschler-Nicola M, Watzek G (2008) Comparison of the validity of three dental methods for the estimation of age at death. Forensic Sci Int 178:96–105
17. Cameriere R, De Angelis D, Ferrante L, Scarpino F, Cingolani M (2007) Age estimation in children by measurement of open apices in teeth: a European formula. Int J Legal Med 121:449–453
18. Knell B, Ruhstaller P, Prieels F, Schmeling A (2009) Dental age diagnostics by means of radiographical evaluation of the growth stages of lower wisdom teeth. Int J Legal Med. 2009 Feb 25
19. Thevissen PW, Fieuws S, Willems G (2009) Human dental age estimation using third molar developmental stages: does a Bayesian approach outperform regression models to discriminate between juveniles and adults? Int J Legal Med Feb 24
20. Alt KW, Rösing FW, Teschler-Nicola M (1998) Dental Anthropology. Springer, Vienna
21. Cameriere R, Ferrante L, Belcastro MG, Bonfiglioli B, Rastelli E, Cingolani M (2007) Age estimation by pulp/tooth ratio in canines by peri-apical X-rays. J Forensic Sci 52:166–170
22. Prince DA, Ubelaker DH (2002) Application of Lamendin's adult dental aging technique to a diverse skeletal sample. J Forensic Sci 47:107–116
23. Acsàdi G, Nemeskéri J (1970) History of Human span and mortality. Akademiai Kiadò, Budapest
24. Schour I, Massler M (1941) Development of human dentition. J Am Dent Assoc 28:1153

25. Demirjian A, Goldstein H, Tanner JM (1973) A new system of dental age assessment. Hum Biol 45:211–227
26. Gustafson G, Koch G (1974) Age estimation up to 16 years based on dental development. Odont Revy 25:297–306
27. Deutsch D, Pe'er E, Gedalia I (1984) Changes in size, morphology and weight of human anterior teeth during the fetal period. Growth 48:74–85
28. Liversidge HM (1999) Dental maturation of 18th and 19th century British children using Demirjian's method. Int J Paediatr Dent 9:111–115
29. Willems G, Van Olmen A, Spiessens B, Carels C (2001) Dental age estimation in Belgian children: Demirjian's technique revisited. J Forensic Sci 46:893–895
30. Cameriere R, Ferrante L, Cingolani M (2006) Age estimation in children by measurement of open apices in teeth. Int J Legal Med 120:49–52
31. Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20:37–46
32. Fleiss JL (1981) Statistical methods for rates and proportions, 2nd edn. Wiley & Sons, New York, pp 212–236
33. Lin LI (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255–268
34. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 2:420–428
35. Dalgaard P (2008) Introductory Statistics with R, 2nd edn. Springer, New York
36. Maber M, Liversidge HM, Hector MP (2006) Accuracy of age estimation of radiographic methods using developing teeth. Forensic Sci Int 159S:S68–S73
37. Cameriere R, Ferrante L, Liversidge HM, Prieto JL, Brkic H (2008) Accuracy of age estimation in children using radiograph of developing teeth. Forensic Sci Int 176:173–177