Two new methodological approaches for assessing skeletal maturity in archeological human remains based on the femoral distal epiphysis.

Rebeca García-González, José Miguel Carretero, Laura Rodríguez, Juan Luis Arsuaga.

Abstract

This study presents two new methodological approaches for estimating skeletal age from maturational changes in the femoral distal epiphysis. In the first approach, five maturity stages were coded based on morphological changes in the epiphysis that encompass the overall developmental process. Data were presented as age ranges for the different maturity stages in the reference sample. As this approach has a number of shortcomings for age assessment, a probabilistic approach was also used. Cross-validation was then used to compare the accuracy of the age estimation from the maturity stages with that from Pyle and Hoerr's atlas. This study's findings showed that Pyle and Hoerr's atlas is more precise than our qualitative method in the oldest age categories. Nonetheless, results from the test of agreement between methods showed that skeletal age estimates from both methods are interchangeable. In the second approach, the overall shape of the femoral distal epiphyses was first analyzed based on elliptical Fourier descriptors (EFDs). Since the number of EFDs is excessively large, a principal component analysis (PCA) of these EFDs was carried out. PC1 scores were used to model the relationship between age and overall shape in a sample of 110 cases of the femoral distal epiphysis. Inverse and classical regression methods of calibration were used to explore the relationship. Based on our results, we recommend the use of a classical calibration model for those cases in which we suspect that the growth and development of the target individual is advanced or delayed relative to those of the Portuguese sample. Otherwise, the inverse calibration model is preferable. Both, quantitative and qualitative methods presented herein notably improve our abilities to estimate skeletal age using incomplete femora from skeletal samples.

Keywords Skeletal age estimation · Femoral development · Maturity stages

Introduction

Among the fundamental biological parameters in anthropo- logical and bio-archeological studies, one of the most important is the age of the individual under study (Lewis 2007). It is particularly important when analyzing subadult individuals, since an accurate age at death estimation increases how reliable estimates are for other parameters such as sex, height, or weight (Sutter 2003; Ruff 2007; Cardoso and Saunders 2008; Vlak et al. 2008). All methods for age estimation in subadult individuals rely on the relationship between dental and skeletal development with chronological age. Both dental and skeletal developments are divisible into two different yet closely integrated processes: growth and maturity (Roche 1992). Growth is the continuous process that implies a progressive incremental change in size, while maturity involves the achievement of specialized and highly organized adult status (Roche 1992; Bogin 1999; Himes 2004). Dental development shows a close relationship with chronological age and thus has proved to be useful to estimate age in developing individuals (Liversidge 2008). For example in a comprehensive study about the performance of several dental age methods, Liversidge et al. (2010) showed that most methods can esti- mate age with errors less than 1 year.

However, when working with archeological specimens that do not have preserved teeth, age estimation should be based on the development of other elements, such as the post-cranial skeleton. Traditionally, physical anthropologists mainly used growth standards based on the relationship between long bone lengths and chronological age. Some of these studies stemmed from longitudinal samples of male and female youth up to 18 years old, so they can be applied to a wide age range of subadult individuals (Maresh 1955; Anderson et al. 1964). Moreover, some specific studies exist for estimating the age at death in fetal and perinatal individuals from diaphyseal lengths (Fazekas and Kósa 1978; Adalian et al. 2002; Carneiro et al. 2013). One of the main problems with these standards is that they are not suitable for age estimation since they do not provide the range of variation in age per diaphy- seal length (Stull et al. 2014). Some recent studies have par- tially overcome this problem via the generation of regression equations to predict age using long bone lengths (Rissech et al. 2008, 2013; Lopez-Costas et al. 2012; Cardoso et al. 2014b; Stull et al. 2014). Unfortunately, these methods suffer from problems that can affect the accuracy of age estimation. For example, limb bone lengths are susceptible to stunted growth attributable to starvation, malnutrition, or poor health that contribute to significant differences in body sizes among juveniles of the same age, which in turn may lead to under- or overestimating age when they are applied. Moreover, to ensure the high accuracy of age estimates, these formulae should only be used on populations from which they were derived (Stull et al. 2014).

In addition to standards for establishing age based on the growth of the long bones, there are also some based on skeletal maturation. An important aspect of skeletal maturity or bone age is its association with parameters such as body shape and size, bone cortical thickness, the percentage of adult size achieved and the timing of both, puberty, and growth spurt (Bayley 1946; Demirjian et al. 1985; Roche 1992). In clinical practice, maturational assessment has proved to be useful to evaluate the way in which children grow (Bayley 1946), and in order to quantify skeletal maturation taking into account the normal variation of this process, clinical studies are based on various maturity indicators. The maturity indicators are discrete events or stages recognizable within the continuous maturational process (Cameron 2004). Several methods have been developed based on maturity indica- tors that describe the sequence of the onset of ossification in the epiphyses of the long bones, the changes in shape and size of the epiphyses, the chronology of the epiphyseal union, and the percentage of adult size achieved (Eveleth and Tanner 1990; Humphrey 2003). Among the most used are charts and standards that describe the relationship between age and maturity indicators of the knee and wrist (Todd 1937; Greulich and Pyle 1950; Pyle and Hoerr 1955; Roche et al. 1975). These charts, derived from radiographic studies conducted on living individuals, give rise to norms or mean ages for the different degrees of the maturity stages, and their appli- cation is usually restricted to clinical evaluations (Roche 1992; Cameron 2004).

In contrast, in bioarcheological contexts, the assessment of skeletal maturity has received less attention. Traditionally, standard charting the chronology of unions of epiphyses to diaphyses (Stevenson 1924; Todd 1930; Stewart 1934; Coqueugniot and Weaver 2007; Schaefer and Black 2007; Cardoso 2008a, b; Cardoso and Ríos 2011; Cardoso et al. 2014a) are used to estimate skeletal age, and the disagreement between skeletal and dental ages is usually attributed to the effect of environmental factors (Lewis 2007). Another application for the assessment of maturity in skeletal remains was provided by Shapland and Lewis (2013, 2014), who showed the relationship between maturity in skeletal remains and the progress of pubertal growth spurts. The standards of epiphyseal union and developmental markers are nevertheless restricted to a concrete period of life, and their utility is limited to pre-adolescents and adolescents. In younger individuals, assessing skeletal maturity via the onset of the ossification of the epiphyses and their subsequent changes in size and shape is preferable. Nonetheless, as previously mentioned, maturity indicators are defined based on radiographic changes in the epiphyses, which are usually difficult to duplicate using direct dry bone observations (Krogman and Iscan 1986). However, because these maturity indicators are able to offer a fair maturity index of the entire skeleton, it could be reasonable to adapt or develop a method based on these indicators that may be applicable to direct observations on dry bones. This was previously done by Conceição and Cardoso (2011), who adapted the stages of the atlas of Pyle and Hoerr (1955) for use in skeletal remains. The main contribution of this adaptation is that the radiographic features used by Pyle and Hoerr's atlas for each stage have been eliminated since they cannot be observed in dry bones. Nonetheless, in our opinion, these maturational stages are not always applicable to archeological samples. As in the original Pyle and Hoerr's atlas, Conceição and Cardoso's adaptation relies on the preservation of the complete knee (the femoral distal end plus the tibial proximal end). When dealing with archeological specimens, complete knees (or bones) are rarely recovered, so criteria such as differences in the relative size between the femoral and distal epiphyses or between the metaphyseal surfaces and their corresponding epiphyses cannot be used. Thus, it could be fairly reasonable to develop a method based on maturity indicators that may be applicable to usually incomplete archeological specimens and when not all knee elements are recovered. In this sense, femoral distal epiphysis could be the best option for two main reasons. First, it is the fastest growing epiphysis in the body undergoing important and visible shape changes throughout its development (Scheuer and Black 2000). Second, due to its ossification starts between 1 month and 2 weeks before birth (Flecker 1932; Roche et al. 1975), it is recognizable as dry bone even in the first year of life.

Thus, this study has different, but related, objectives. The first objective is to identify and define the maturity stages of femoral distal epiphysis as a function of its morphology that are suitable for estimating age. The definition of these maturity stages relies on features of both articular and metaphyseal surfaces, bearing in mind that archeological skeletal specimens are, most times, eroded. This erosion usually affects more the metaphyseal than the articular surface of the epiphysis and can hide some important features that could be used to distinguish maturity stages. Thus, the second objective of this study is to explore to what extent changes in the shape of the femoral distal epiphysis alone are sufficient to estimate the age. To do that, we rely in a regression analysis to establish the relationship between shape and age. We focus on shape rather than size because we are dealing with maturation and not growth (as previously mentioned, changes in size are primarily a response to the growth process). In sum, we present here one qualitative method (based on the morphology of the femoral distal epiphyses) and one quantitative method (based on changes in shape and regression analysis) for skeletal age assessment in subadult individuals.

Material

Reference sample

The reference data come from two human skeletal collections of identified individuals. The first sample is composed of 122 individuals (adults and subadults) of known sex and age at death housed in the Bocage Museum (National Museum of Natural History, Lisbon, Portugal). The second sample is composed of 55 individuals of known sex and age at death aged from 7 to 26 years old that belongs to the collection housed in the Department of Life Sciences at Coimbra University (Coimbra, Portugal). Both collections come from modern cemetery sources and are formed by Portuguese people who lived in the nineteenth and twentieth centuries representing the middle-to- low social class of the cities of Lisbon and Coimbra (Cardoso 2006; Coqueugniot and Weaver 2007). Both samples were considered a single population for the present analysis. Findings from previous studies regarding infracranial sequences of maturation performed in these two samples did not show population differences in the ages of attainment of different stages of fusion (Coqueugniot and Weaver 2007; Cardoso 2008a, b). Moreover, these two samples were already pooled together in other growth studies, revealing similar growth patterns (Rissech et al. 2008; García-González et al. 2013). Therefore, the most plausible scenario is that there were no differences between these two samples in the maturation of the distal epiphysis.

The adults of both samples (Coimbra and Lisbon) were studied in order to evaluate the entire maturation process. Due to the inter-individual variation in maturation rates, it is not possible to assign a particular age with full maturity or the end of the maturation process. For example, Greulich and Pyle's atlas (1950) to assess skeletal maturity fails because it represented full maturity as a chrono- logical age of 18 years (Cameron 2004). Thus, in order to avoid this problem, we included adult individuals aged up to 26 years old, because this age is much higher than the age in which individuals from both collections showed a completely fused epiphysis (Coqueugniot and Weaver 2007; Cardoso 2008a). The selection of these adult individuals was carried out taking into account two criteria. First, we aimed for a balanced-sex subsample of adult individuals. Second, pathological individuals were excluded. The age and sex distribution of this reference sample is depicted in Fig. 1.

Target samples

In order to investigate the relative influence of the inter-population variation in the two methodological approaches proposed here, two additional target samples were included in this study. One of these samples is composed of 34 individuals from the archeological collection from San Pablo (Burgos) housed in the Laboratory of Human Evolution at the University of Burgos. The age at death estimate of the San Pablo individuals was based on the calcification and formation of dental crowns and roots. The mineralization stages of each tooth class were observed by conventional radiography and scored using the method of Moorrees et al. (1963).

Although these authors provided the conversion of these stages into age, the use of these data show some problems. First, Moorrees et al. (1963) gave the conversion of stages into age graphically, and it is difficult to implement because there are no accompanying numerical values for the dental stages (Simpson and Kunos 1998). Second, they describe the age of "entering" a particular stage of development and, as Smith (1991) noted, this could be inappropriate to predict age. Thus, the Moorree's stages were converted into age following the adjusted data for prediction proposed by Smith (1991). As the sex of the archeological individuals was unknown, we averaged the dental age estimates based on male and female tables. Based on these estimates, the age distribution of the San Pablo sample ranged from zero to 12 years. The second target sample is composed of eight children of known age between zero and 6 years old housed in the Faculty of Medicine at the University of Valladolid (Spain). Sex was unknown for these eight individuals.

Methods

Qualitative method for age estimation

To address the first objective (the definition of maturity stages according to morphology that are suitable for age estimation), we follow several steps explained below.

Definition of maturity stages based on epiphyseal morphology

The maturation of the distal femoral epiphysis was assessed based on its morphological and shape changes throughout the developmental process. Five maturity stages were coded to accomplish this. The first three stages encompass the period of time prior to the fusion of the femoral distal epiphysis to the diaphysis and are based on changes in the shape and features of the metaphyseal and articular surfaces. These definitions are derived from those previously provided by Pyle and Hoerr (1955), which have been summarized by Scheuer and Black (2000).

Moreover, some aspects have been added, such as the projection of lateral lip proposed by Tardieu (1998).

The three stages are depicted in Fig. 2 and their characteristics (including some new aspects added by the authors) are explained as follows:

Stage 1. The distal epiphysis is a small nodule that is usually round or kidney-shaped. The anterior edge is flat and an incipient intercondylar notch is visible on the posterior edge. The metaphyseal surface is rough and almost flat except for a central round elevation separating the medial and lateral areas (Fig. 2).

Stage 2. The shape of the epiphysis is almost rectangular and does not have an anteriorly marked projection of the lateral lip. The intercondylar notch is increased in depth and width and is covered by numerous nutrient foramina. The anterior edge is relatively sinuous due to the deepening of the trochlear groove. The articular surfaces of the condyles are relatively smooth with some pitting. On the metaphyseal surface, the trochlear and condylar areas are divided by a transverse ridge. The trochlear area is smaller than the condylar area (Fig. 2).

Stage 3. The epiphysis has already attained its distinctive shape with the lateral lip projecting more anteriorly than the medial one. Both the intercondylar notch and the trochlear groove are completely developed as is the adductor tubercle in the posteromedial border. The articular surfaces of the condyles are covered by a smooth cortical bone without pits. On the metaphyseal surfaces, the trochlear area has undergone considerable lengthening due to the anterior projection of the lateral condyle (Fig. 2).

In the other hand, stages 4 and 5 refer to partial fusion and total fusion, respectively. In these cases, the last two stages defined by Coqueugniot and Weaver (2007) and Cardoso (2008a) were utilized. Following these authors, stage 4 was assigned to cases where there were some gaps visible in the epiphyseal-diaphyseal junction, while those with a completely fused epiphysis were categorized as stage 5 (even when the epiphyseal-diaphyseal junction showed a scar or epiphyseal line).

Statistical analyses of maturity stages based on epiphyseal morphology

The maturity stage was assessed for all of the individuals and, as the age at death is known, each individual was assigned to an age class (ranked by 1-year-old classes). Data were then summarized in age categories for each of these maturity stages. The dependence between the chronological ages and degrees of maturity indicators was assessed using Spearman's rank correlation coefficient. This coefficient is herein preferable to Pearson's coefficient since our data for femoral maturity stages are ranked. Stage 5 was eliminated for this analysis because, as previously mentioned, it corresponds to a stage of complete fusion and therefore will provide a minimum age of attainment.

This approach, while useful as a first approximation, has a number of shortcomings for age estimation. The most important is that the reference sample size is small and the age coverage uneven, which could lead to age assessments that may be subject to "age mimicry." One way to partially over- come these methodological shortcomings is to perform a probabilistic approach (Coqueugniot et al. 2010). Thus, a Bayesian statistical procedure was performed to produce a distribution of probabilities that an unknown individual be- longs to a given age class based on the maturity stage it presents (posterior probability). Following Bayes' theorem, this probability can be expressed as follows:

$$P\ age_i/stage_j = Pstage_j/age_i * Pprior\ (age) /\ \text{sum of } Pstage_j/age_i * \ Pprior_{\ age=i}$$

where $i = 1, 2, 3........26$ and $j = 1, 2, 3, 4$, and 5.

The first step was to calculate the probability of observing a specific maturity stage if an individual has a specific age: P (stage$_j$/age$_i$). In large samples with many individuals in each group, this probability can be calculated from the relative frequencies of each stage and age. But this is not the case in this study. Thus, we constructed a density distribution of P (stage$_j$/ age$_i$) from the frequencies of each of the maturity stages by age as well as from the frequencies ± 2 years (that is, including 2 years older and 2 years younger) (Coqueugniot et al. 2010). Then these data were smoothed using kernel smoothing, which is a nonparametric way to estimate a real valued function as the weighted average of the neighboring observed data (Love and Muller 2002). To accomplish this, a crucial step is to choose the bandwidth parameter. The bandwidth was chosen as the 10% of the range of ages in the reference sample (Love and Muller 2002).

The density distributions were estimated using the Kern Smooth Package for R. Once the density distribution was estimated, we calculated P (stage$_j$/age$_i$), the probability that one individual in a specific stage has a particular age, as the area under the density distribution at this age. This area was found using a definite integral between two points, which was calculated following an approximation by trapezoids using Excel. The two points or limits of the definite integral were selected taking into account that in P (stage$_j$/age$_i$), age$_j$ represents all of the ages between the first and last day of each one- year class. For example, for an age of 4, the lower limit is 4.00 years and the upper limit is 4.99 years. Even so, to produce a distribution of P (age$_i$/stage$_j$), we must know not only the probabilities but also the prior probabilities for each age. Here, as in other studies, the sample was constructed by "availability sampling" and it would be inappropriate to use this information as priors. Thus, there are two options to estimate the priors: either use demographic data or assume an unbiased and uniform frequency distribution of our age categories (Braga et al. 2005). The second option was chosen here because the goal of our study is individual age assessment rather than an estimation of population age structure in a sample (Konigsberg and Frankenberg 1992).

In this way, we obtain one posterior probability distribution for each maturity stage as a function of the morphology of the femoral distal epiphyses, which are more suitable for age estimation.

Testing repeatability and reproducibility of method

In order to test the repeatability (intra-observer error) and re- producibility (inter-observer error) of this method, the overall sample was reassessed prior to data collection (Ferrante and Cameriere 2009) by the first and third authors. Intra- and inter- observer errors were calculated using Cohen's kappa value (Landis and Koch 1977).

Testing the comparative performance of method

Although the value of Spearman’s rank correlation coefficient describes the strength of the association between progression in chronological age and progression in the attainment of the different maturity stages based on epiphyseal morphology, it provides scant information regarding the magnitude and/or direction of the difference between the estimated and the ac- tual age. Bias, accuracy, and precision better express how close the chronological and estimated ages are.

From the probabilities distributions, the point estimates can be defined which in turn can be used for the calculation of the bias, precision, and accuracy of the method. The point esti- mates were calculated in two ways. First, the mode point es- timate was calculated as the age with the highest posterior probability. In cases in which more than one age showed the same posterior probability, the mode was calculated as the average of the oldest and youngest ages with this probability (Coqueugniot et al. 2010). Second, a weighted average of all ages where the posterior probabilities were different from zero was calculated. In this case, the weights were the values of the posterior probability.

The accuracy, bias, and precision of the femoral distal epiphyses morphological stages were compared to those of Pyle and Hoerr’s atlas. To accomplish this, we estimated the ages obtained with these two methods in 33 individuals from our reference sample (Coimbra plus Lisbon), which preserves complete knees. Only the stages prior to fusion (partial or total) were tested since they were the new stages defined in this study. The evaluation of the accuracy, bias, and precision of the femoral stages was based on point estimates. In the case of Pyle and Hoerr’s atlas, the stage of skeletal maturation of the knee following Conceição and Cardoso’s adaptation was first assessed. Then these stages were converted into age based on sex-appropriate plates provided by Pyle and Hoerr (1955). The bias was calculated as the difference between the esti- mated age by each method and the actual age. The measure of precision used herein was the standard deviation of the bias (SD) (Walther and Moore 2005). The accuracy can be defined as the overall distance between the estimated age and the actual age and was calculated as the average of the absolute difference between the estimated and the actual age and the percentage of the individuals aged within 15%, 20%, and 30% of the actual age.

Quantitative method to estimate age from the subadult femoral distal epiphysis

The protocol addresses the second objective of this study (the generation of a regression equation to predict age using the overall shape of the distal femoral epiphysis) also involves several steps.

Quantifying shape changes of the distal femur during maturation

There are two kinds of approaches to quantify the shape of biological objects: landmark-based methods and outline analysis. Here, we perform an outline analysis based on elliptical Fourier descriptors (EFDs) (Baltanás 2016; Salazar et al. 2017). The coefficients of EFDs can be used to characterize and analyze complex closed contours because they are invari- ant with the rotation, dilation, and translation of the contour (Kuhl and Giardina 1982; Rohlf and Archie 1984). An analysis based on elliptical Fourier descriptors requires available closed contours of femoral distal epiphyses. We have drawn this contours for 152 femoral distal epiphyses from our human skeletal samples (110 from identified Coimbra-Lisbon samples and 42 from Valladolid and San Pablo samples). The number of distal epiphyses from the Portuguese samples is fewer than in the previous morphological analysis, since to perform a shape analysis, the first step is to obtain closed contours from digitized images taken orthogonally to the articular surfaces. In some cases, the digitized images were not available (overall in the case of adult specimens) and thus those specimens were eliminated from the analysis. The samples of San Pablo and Valladolid were introduced into the analysis in order to consider as much shape variation as possible. Out of the 152 distal femoral epiphyses, 45 (29.6%) are in stage 1, 26 (17.1%) are in stage 2, 33 (21.7%) are in stage 3, and 48 (31.6%) are in stages 4–5.

Digitized images were taken using a Nikon Coolpix 995 in manual mode placed in a tripod to ensure that the camera is totally orthogonally to the articular surfaces. In all cases, a scale was placed together with the photographed specimen.

In all cases, closed contours were drawn from digitized images using Adobe Photoshop. First, we scaled the digitized image to the real size. Then, we use the "pen tool," to create different anchor points shaping the epiphysis. Once the anchor points were created, we closed the selection and filled the selection in black with the "fill tool." With the closed contours at hand, the analysis based on EFDs was performed using the Shape software (Iwata and Ukai 2002). This software calculate the coefficients of EFDs from a chain coder that describes a continuous contour using a sequence of piecewise linear fits consisting of several standardized line segments (from 0 to 7) (Freeman 1974). The number of coefficients of EFDs is very large and the morphological meaning of each coefficient is difficult to interpret separately (Iwata and Ukai 2002). In our case, the contour shape was described in the first 20 harmonics of the Fourier coefficients, which provided 77 coefficients of the EFDs. In order to summarize the information on the variations in the coefficients of EFDs, we performed a principal component analysis (PCA) of these coefficients based on a variance–covariance matrix (Rohlf and Archie 1984). This PCA enables the visualization of the shape variation explained by the effective PCs (those that explain a proportion of the variance larger than 1 divided by the number of total principal components). As we analyzed the variance–covariance matrix of the 77 coefficients of the EFDs, the number of total principal components is 77, and any effective PCs must explain a variance larger than 1.2987. The scores of the effective PCs were used as input data for the subsequent analyses.

Choosing the scores of the effective PCs for age estimation

We graphically explored the relationship between the scores of each effective PC and age. To accomplish this, we only used data from the individuals whose age and sex were known. Based on this relationship, we calculated Pearson's correlation coefficient between the chronological age of the individuals and the PC scores. Stages 4 and 5 were excluded from the analysis because the objective of this approach is to estimate age previous to epiphyseal fusion. If Pearson's correlation coefficient is sufficiently high (> 0.90) and statistically significant, we used the scores of this PC to estimate the age. We considered Pearson's correlation coefficient larger than 0.90 because lesser coefficients would lead to values of the determination coefficient less than 0.80 and, in turn, this would mean that less than 80% of the femoral distal epiphysis shape is explained by the age. As the objective herein is to develop a regression equation to estimate the age, this would be an unacceptable scenario.

Performing regression analysis for age estimation

The regression model most commonly used for the age estimation is the least squares regression and inverse calibration. In this calibration model, the age is regressed on another variable (in this case, the shape of the femoral distal epiphysis) and thus, the age would be the dependent variable. However, it is shape that depends on age, rather than the contrary. The alternative to inverse calibration is classical calibration. Based on this model, the shape (dependent) should be regressed on age (independent) followed by a solution for the age. While inverse calibration is preferred when the age distribution for the reference sample forms a reasonable prior, classical calibration is preferred if it is suspected that the estimated ages will be an extrapolation beyond the useful limits of the reference sample ages (Konigsberg et al. 1998). Thus, in order to take into account all of the possible scenarios, we used both inverse and classical calibration.

Before performing both models, a multiple regression analysis was carried out to determine whether the variation in the shape of the femoral distal epiphyses was related to both the age and sex of the individuals. In this regression, the shape is the independent variable, while the sex and age are dependent variables. This is necessary because if sex proved to be a significant source of variation in the shape of the epiphyses, the age estimation formulae will be separately developed for the sexes. In contrast, if sex proved to be an insignificant source of variation in the shape of the epiphyses, the age estimation formulae will be determined for the sexes combined.

Testing repeatability and reproducibility

The regression equations to predict age proposed here are based on the PC scores, and the exact PC values are entirely sample-dependent (clearly differing between samples). Fortunately, we can calculate the PC score of a problem specimen if it is introduced in the PCA analysis following the algebraic description of PCA. In a PCA, the covariance matrix is decomposed into eigenvectors and eigenvalues (Klingenberg 1996; Zelditch et al. 2012). The matrix of eigen- vectors (A) is used to transform the original data (X) into a set of new variable: the principal components ($PC = A \times X$). Any PCA can be interpreted geometrically as a rotation of the coordinate system, since the PCs are aligned with the directions of the axes of the multidimensional scatter ellipsoid (Klingenberg 1996). The position of any data relative to these new axes is given by its PC score. Because the PCs intersect at the sample mean, the values of the PC scores represent the distances of the specimen from the mean in the directions of the PCs. Thus, following Klingenberg (1996) and Zelditch et al. (2012), we can compute an individual's score on a PC from the values of the new case (Y), the initial sample mean (X_m) and the value of the eigenvector (A) with the formula:

$$PC_{score} = A \times (Y - X_m)$$

In our case, the variables are the EFDs. To ensure the predictive equations proposed in this study can be used by other re- searchers, the first step is to obtain the EFDs of the new case. Information about how to obtain these new EFDs is provided in Supplementary Information 1. The second step is to obtain the PC scores. As explained above, these can be calculated based on the algebraic description of PCA. To do that, the vectors of the means of the EFDs of our sample and the eigenvector for our PC1 are provided in Supplementary Information 2 and 3 respec- tively. Moreover, a detailed explanation to implement this calculation in Excel is given in Supplementary Information 4.

With this information, other researchers can obtain the PC1 score for new specimens based on the previously provided equation. Nonetheless, in order to ensure the repeatability of our method, the chain coder for each individual used in this regression is provided in Supplementary Information 5. The cases in which the ages were known are provided in Supplementary Information 6. Thus, other researchers can perform the analysis from the beginning and obtain the new scores for their samples and derive their new regression equations for predicting age.

The reliability of the measured contours must be evaluated in order to assess the appropriateness of the method. Thus, intra- and inter-observer variations in predicted skeletal age were examined in a subset of 15 distal femoral epiphyses (10% of the total sample), selecting five specimens randomly for each of the three maturity stages. Closed contours of the 15 specimens were drawn again from digitized images using the Adobe Photoshop by RGG (intra-observer error) and Gymp software by LR (inter-observer error). The new closed contours were introduced in the analysis, so new PC scores were obtained for each of them. These new PC scores were used to predict age from inverse and classic regression models. Age estimates were tested using a Wilcoxon signed rank test to see if there were statistically significant differences.

Testing the comparative performance of inverse and classic calibration models

The bias, accuracy, and precision of both models (inverse and classical calibration) were compared to those obtained using the other quantitative methods. These quantitative methods were proposed by Rissech et al. (2008) and Cardoso et al. (2014b). The former authors used an inverse calibration and the latter a classical calibration. Cardoso et al. (2014b) pro- posed equations to predict the age from all long bones, but we used those proposed by the femoral diaphyseal length. The reference sample used by Rissech et al. (2008) comprised the Coimbra and Luis Lopes (Lisbon) collections, while that used by Cardoso et al. (2014b) comprised only the Luis Lopes collection.

The bias was calculated in the same way as in the first approach based on the morphology of the femoral distal epiphyses. In this case, we used two measurements of the accuracy. First, we calculated the mean of the absolute value of the differences between the known chronological age and the estimated age. Second, we calculated the percentage of individuals whose chronological ages fell within the 95% confidence interval (95% CI) of the estimated age. In the inverse calibration models, the 95% CI was calculated by multiplying the standard error of the estimate (SEE) by the appropriate value from the t distribution for $n-2$ degrees of freedom (Sokal et al. 1979). In the classical calibration models, the standard error of the estimate (SEE) cannot be obtained.

Thus, we calculated the mean standard error (MSE). The MSE was calculated as the average of the standard error for each individual observation (Lucy 2005). The 95% CI were calculated in the same way as in inverse calibration.

Independent test of performance and test of agreement among predictive equations based on femoral distal epiphysis shape, dental age methods, and methods based on femoral length

As in the case of the qualitative method, we performed an independent test of performance and a test of agreement using the same two target samples (San Pablo and Valladolid). For the Valladolid sample, we calculated bias, accuracy and precision when inverse and classical models are applied. For the San Pablo sample, we have taken into account that the ages predicted from the two regression models are also skeletal ages. Thus, as in the case of maturity stages, these ages may be uncorrelated to dental ages in San Pablo sample. To explore the effect of this independence between dental and skeletal development, we first compared the skeletal ages derived from the inverse and classical models proposed in this study to ages based on femoral length derived from formulae proposed by Cardoso et al. (2014b) and Rissech et al. (2008). In this way, we are evaluating the agreement between different methods for age estimation based on this post-cranial skeletal element. Second, we compared the dental ages to the ages predicted by both femoral length and femoral distal epiphyseal shape. The statistical treatment is the Bland– Altman method mentioned above.

Controlling potential factors of variability in skeletal ages

Individuals with the same chronological age may show differences in the developmental stages of different biological systems (Demirjian et al. 1985). For example, Mani et al. (2008) found that among boys, the body mass index (BMI) was statistically correlated with the difference between the dental age and the chronological age. As we have mentioned above, skeletal ages are associated with body size and shape. Thus, it is possible that differences in body size and shape of individuals of the same age affect age estimations. In order to control for this potential variability, we calculate the correlation of skeletal ages predicted from inverse and classical calibration models with body mass and body size. We use the femoral distal metaphyseal breadth (DMB) as a surrogate of body mass and femoral length (FL) as a surrogate of body size (Ruff 2007). The subsample used in this analysis is composed of 42 individuals (18 males and 24 females) from the Lisbon collection. The age distribution of this subsample ranged from 1 to 16 years old.

Skeletal ages, DMB, and FL are strongly correlated with an individual's age, as they change over the course of the development. Thus, it is first necessary to remove the unwanted effects of age on DMB and FL to examine the relationship between these variables and skeletal ages. In this way, we avoid considering the variation related to age that may obscure any signal related to skeletal age, DMB and FL (Cowgill 2010; Child and Cowgill 2017). To accomplish this, we first corrected DMB and FL for age by regressing each variable on age. The most appro- priate order of regression analysis was selected based on (1) the strength of the coefficient of determination, (2) the significance of the function, and (3) the significance of the coefficients of the functions. In this way, DMB and FL were regressed on age using cubic ordinary least squares formulae with the software PAST. These fitted models age are:

$$\mathrm{FL} = 0.083 \times \mathrm{age}^3 - 2.36 \times \mathrm{age}^2 + 35.66 \times \mathrm{age} + 92.49;\ \mathrm{SEE} = 22.93;\ R^2 = 0.94$$

$$\mathrm{DMB} = 0.025 \times \mathrm{age}^3 - 0.71 \times \mathrm{age}^2 + 7.62 \times \mathrm{age} + 23.84;\ \mathrm{SEE} = 4.87;\ R^2 = 0.86$$

Based on these models, we calculated the age- standardized residuals of DMB and FL, which were used to calculate the correlation with the standardized residuals of the skeletal ages obtained from inverse and classic models. A negative standardized residual of the skeletal ages means that the estimated age is higher than the real age (advanced development) and vice versa.

Results

Qualitative method

Age assessment from maturity stages based on epiphyseal morphology

Table 1 shows the summary of the age intervals for epiphyseal maturation of the distal femur by sex. In the column corresponding to stage 1, the age of the oldest individual in this maturational stage is indicated. The last column (stage 5) represents the age of the youngest individual in this stage.

The rest of the columns provide information regarding the age interval at which a specific stage occurs.

The results show a sequential pattern in the attainment of these stages. Indeed, the value of Spearman's rank correlation coefficient was 0.90 ($p < 0.05$), which indicates that there is a strong and positive relationship between the age and these maturity stages. Figures 3a, b shows the posterior probability distributions (the probabilities of observing a specific stage if an individual has a specific age) for the maturity stages as defined in this study when they were estimated using kernel smoothing.

The most marked difference between the probability distributions depicted in Figs. 3a, b and the age ranges shown in Table 1 is for the overlapping among stages. Since kernel smoothing leads to more reasonable posterior probability distributions with unevenly distributed samples, a greater overlap between the maturity stages than that reported in Table 1 should be considered.

The mode and weighted average point estimates for males and females, and for the three new maturity stages proposed in this study, are given in Table 2.

The mode point estimate tended to be less than the weight- ed average point estimate overall in the case of stage 1, because in this stage, the age from zero to three has a posterior probability equal to 1 (Fig. 3a, b, Table 2).

Repeatability and reproducibility of qualitative method

The kappa values were 0.98 for the intra-observer errors and 0.96 for the inter-observers error, showing excellent repeat- ability and reproducibility for this method.

Comparative performance of the qualitative method

Table 3 shows the accuracy of Pyle and Hoerr's atlas and the maturity stages based on the morphology of the femoral distal epiphysis in the reference sample.

When the sexes are combined, the best overall bias is shown by our method when the weighted average point estimate is used. This method also showed the highest percentage of individuals aged to within 20% and 30% of age. In contrast, the method with the highest percentage of individuals aged to within 15% of age was Pyle and Hoerr's atlas. The smallest SD (highest precision) and smallest mean absolute difference was Pyle and Hoerr's atlas. The results separated by sex indicated that all of the methods underestimated ages in males, while the weighted average point estimate overestimated ages in females. The values of SD indicate that the precision for both the males and females is better using Pyle and Hoerr's atlas than using the weighted average point estimate. Nonetheless, high precision is obtained when the mode point estimate is used in males.

Independent test and test of agreement between methods

The use of mode point estimates in the Valladolid sample underestimates actual age with a bias equal to −0.65. In contrast, the use of weighted mean estimates yields a mean bias of 0.58. The standard deviation (precision) was in both cases equal to 1.29, which indicates that the dispersion around the age estimates is equivalent when both the mode and the weighted mean estimates were applied in the reference sample.

Table 4 shows the results of the Bland–Altman method in the San Pablo sample for assessment of agreement between age estimates from our qualitative method and dental method. Although the limits of agreement are not narrow enough, the null hypothesis that the mean differences between Pyle and Hoerr's atlas and mean and mode point estimates did not significantly differ from zero cannot be rejected (in no case was the slope significantly different from zero). Thus, these results provide support for the fact that Pyle and Hoerr's atlas and our femoral maturity stages are interchange- able, at least for the San Pablo sample. When dental age esti- mates were compared to skeletal age estimates, only the mean differences between dental age and mode point estimates derived from the Bayesian approach did not significantly differ from zero. Nonetheless, the limits of agreement between the three skeletal age estimates and dental age estimates are ex- tremely wide and even wider with Pyle and Hoerr's age estimates. This illustrates the lack of correlation between dental and skeletal development mentioned above.

Quantitative method

Shape changes throughout the development

Regarding the changes in the femoral distal epiphysis shape throughout the developmental process, the analysis of the coefficients of the EFDs of the entire sample shows that 94.39% of the total variance is explained by 4 principal components. The first component (PC1) is associated with 70.64% of the variability while the other three (PC2, PC3, and PC4) explain 17.99%, 3.85%, and 1.89%, respectively (Table 5).

PC1 accounts for the variation in the total epiphyseal shape of the distal femoral epiphysis during the developmental process. Negative values represent the rounder and more kidney-shaped distal epiphyses and positive values represent epiphyses with an anteriorly projected lateral lip, a clear intercondylar notch, and a completely developed trochlear groove.

PC2 depicts the variation in the intercondylar depth. Changes along PC3 concern the degree of anteroposterior lengthening of the lateral condyle. Finally, PC4 accounts for the variation in the shape of the trochlear groove (Fig. 4).

Figure 5 shows a scatter plot of the scores of each principal component versus the age.

The plot of PC1 versus the age shows how the different femoral maturity stages are distributed consecutively (Fig. 5a). The negative values of PC1 correspond to stages 1 and 2, while the positive values correspond to stages 3 and 4–5. The PC1 scores for stages 3 and 4–5 show some overlap, indicating that there are no differences in the total shape for the epiphyses in these three stages. Indeed, one of the morphological features of stage 3 is that the epiphysis shows a whole shape that is reminiscent of those of adult epiphyses. Therefore, the whole shape cannot be used to distinguish be- tween stage 3 and stages 4–5. For this reason, Pearson's correlation coefficient between the PC1 scores and chronological age was calculated excluding the epiphyses in stages 4–5. The correlation is statistically significant ($p < 0.05$) and high ($r = 0.91$), which indicates the strong and positive relationship between the age and the distal femoral epiphyseal shape changes.

A considerable overlap between the maturity stages occurs when PC2, PC3, and PC4 are plotted versus the chronological age. As previously mentioned, these factors explain the variation in the specific features of femoral distal epiphysis that do not seem to have a relationship with age in the entire sample (Fig. 5b–d). Indeed, Pearson's correlation coefficient is only significant in the relationship between PC4 and age, but the value is not sufficiently high ($r = 0.58$). In the other two cases, that is, the correlation between PC2 and PC3 and age, the values of the correlation coefficients are very low ($r = -0.29$ and $r = -0.25$, respectively). However, some interesting aspects are observed if particular maturity stages are considered. Interestingly, maturity stage 1 is split into two groups along PC2, with some specimens showing positive values, while others fall towards the negative end of the PC2 axis (Fig. 5b). This is because two different shapes can be assigned to stage 1 (see above): one that is rounder and another that is kidney shaped. The positive values correspond to round shapes and the negative values to kidney shapes. Therefore, the PC2 scores can distinguish between two age groups within stage 1: those with positive values correspond to chronological ages up to 3 years old and those with negative values correspond to ages ranging from 3 to 5 years old.

Predictive equations

Taking into account the aforementioned results, a multiple regression analysis was carried out in order to determine to what extent age is related to PC1 and its scores are affected by the sex of the individual. The epiphyses in stage 4 or 5 were excluded from this analysis because the aim of this study is to estimate the age in individuals prior to epiphyseo-diaphyseal fusion.

The multiple regression results showed that sex did not contribute significantly to the regression fit. The results of the t test for a single sample ($t = 1.34$; $p = 0.18$) showed that the confidence interval of the sex parameter did not differ from zero. This implies that this coefficient may be removed from the model without substantially modifying the fit. Thus, we did not develop a different equation for each sex. Table 6 shows the inverse and classical models for predicting ages from the PC1 scores. Each model includes the N, the predictive equation, the MSE or SEE, and the determination coefficient (R^2).

The raw residual distribution of classical (Fig. 6a) and in- verse calibration (Fig. 6b) models show no obvious pattern, since the residuals are randomly scattered above and below the line at $y = 0$. The observed versus predicted values show that both, the classic (Fig. 6c) and inverse (Fig. 6d) regressions fit the trend of the data reasonably well. Only one observation appeared to be an outlier in the inverse calibration model (standard residual > 2 × SEE). Hence, both equations can be considered good predictors of age based on the shape of distal femoral epiphyses in skeletal remains.

Repeatability and reproducibility of the quantitative method

Table 7 includes the results from the Wilcoxon ranked test performed to evaluate the intra- and inter-observer variation based on contour drawings. Based on p values, we can assert that statistically significant differences were not found in any comparison and thus the error in contour drawings does not have a significant effect on the age estimation.

In Supplementary Information 7, we provide the photo- graphs of the 15 femoral distal epiphyses randomly selected to perform the intra- and inter-observer tests of variation as well as the different contour drawings.

Comparative performance of quantitative method

Table 8 shows the bias, accuracy, and precision of the methods proposed by Rissech et al. (2008) and Cardoso et al. (2014b) based on the diaphyseal lengths and the inverse and classical models proposed in this study.

All of the methods overestimate the ages, but the method with the lowest bias was the inverse calibration model based on the femoral distal epiphysis shape. This method also had the highest percentage of individuals whose chronological ages fell within the 95% confidence interval of the estimated ages. However, the two models proposed in this study had the lowest accuracy and precision.

Independent test and test of agreement between methods

Both inverse and classical calibration models overestimate age in the Valladolid sample. The bias is 1.88 and 0.98 for the inverse and classical models respectively. The standard deviation of the bias is 1.49 for the inverse calibration model and 0.58 for the classical one. These values are smaller than the SEE and MSE of both original models, which implies that the data dispersion around the estimated ages fall well within the data dispersion of the original models.

Table 9 shows the results of Bland–Altman method for assessment of agreement between different age estimates when quantitative and dental methods were used for the San Pablo sample.

Comparisons between shape models proposed here and methods based on femoral length show that only mean differences between age estimates from the classical calibration model and those estimated using the method by Cardoso et al. (2014b) differ significantly from zero. Moreover, the limits of agreement in this case are extremely wide, and the mean differences between dental age and estimates from inverse and classical calibration models differ significantly from zero. In contrast, the mean difference between dental age and estimates derived from femoral length are only statistically different from zero in the case of the Rissech et al. (2008) method. Thus, among the methods to estimate age based on post-cranial growth and maturity, only those derived from the method proposed by Cardoso et al. (2014b) are congruent with dental age estimates.

Potential factors of variation in skeletal age

Finally, no significant relationship was found between age- standardized residuals of DMB and standardized residuals of the inverse and classical models. In contrast, age-standardized residuals of FL show a low but significant correlation ($p < 0.05$) with standardized residuals of the inverse and clas- sic calibration ($r = -0.31$ and $r = -0.33$, respectively).

Discussion

Qualitative method

This study presented two methodological approaches for estimating age in juvenile skeletons based on maturational criteria. The first approach was based on the morphological changes in the femoral distal epiphysis. When these changes are assessed on dry bones, stages 1, 2, and 3 can be used to establish a lower and upper limit for the probable skeletal ages. These results notably improved our ability to estimate skeletal ages in juvenile skeletons when the teeth were not preserved, and the long bones were broken. Hitherto, only the distal femur at a stage of partial union (herein stage 4) provided an estimate of an age interval for the specimen's true age. If the specimen had an unfused femoral distal epiphysis, only an upper limit for the age interval could be estimated. This upper limit varied between 16 and 19 years in females and between 18 and 20 years in males (McKern and Stewart 1957; Buikstra and Ubelaker 1994; Coqueugniot and Weaver 2007; Schaefer and Black 2007; Cardoso 2008a). However, using the data provided in this study for unfused femoral distal epiphysis, skeletal age can be estimated with more precision. Based on the results depicted in Table 1, we estimated the ages within a maximum interval of 5 years for epiphysis at stage 1, within a maximum interval of 7 years for epiphysis at stage 2, and within a maximum interval of 6 years for epiphysis at stage 3. These ranges could have been wider if we used the results obtained from the Bayesian approach (Fig. 3a, b). Following this approach, age is estimated within a maximum interval of 7 years for the epiphysis at stage 1, within a maximum interval of 11 years for the epiphysis at stage 2 and within a maximum interval of 8 years for the epiphysis at stage 3 (Fig. 3a, b). The upper limits for the probable skeletal ages are 7, 13, and 17 years old for stages 1, 2, and 3, respectively. The suitability of the qualitative method proposed herein for age estimation was assessed by comparing it to Pyle and Hoerr's atlas. Results for this comparison are not easy to interpret, at least at first glance. A good method should include low bias, low mean absolute differences, and high proportion of individuals aged to within a given percentage of real age (Liversidge et al. 2010). This is due to the fact that the more biased and less precise one method is, the worse is its overall ability to make an accurate estimation (Walther and Moore 2005). However, our findings showed that while our qualitative method based on weighted mean point estimates had the lowest bias, the Pyle and Hoerr's atlas stands out as performing better for accuracy measured as the mean of absolute differences and as the percentage of individuals aged to within 15%, as well as the best precision (Table 3). Otherwise, the accuracy expressed as the proportion of the individuals aged to within 20% and 30% of age was better when our stages were applied.

Thus, we should select which of the different performance measures is the best to assess if our qualitative method is more or less suitable for age estimation that Pyle and Hoer's atlas. For some authors, the lack of bias is the best indicator of the performance of a method (Liversidge et al. 2010). However, the methods tested here were not unbiased and, thus, if we only assess the bias we will evaluate the over- or under- estimation of these methods. Fortunately, measures of accuracy expressed as the percentage of individuals aged to within 15%, 20%, and 30% of real age combine bias and precision in their mathematical definitions and, therefore, are a good way to assess the overall performance (Walther and Moore 2005). However, our results showed that the method with the highest percentage of individuals aged to within 15% of the real age did not have the highest percentages of individuals aged to within 20% and 30% of the real age.

A percentage of 15% corresponds to an absolute error ranging from 1 year in either direction for the youngest age cate- gory to 2.5 years in either direction for the oldest age category. In contrast, with percentages of 20% and 30%, the absolute error can be up to 3 years in either direction for the oldest age category. Thus, the method with the highest percentage of individuals aged to within 15% is more precise than those with the highest percentages of individuals aged to within 20% or 30% of real age. Based on this, we can assert that the Pyle and Hoer's atlas is slightly more precise and accurate than our qualitative method, especially in the oldest age groups.

The lower precision of our method than Pyle and Hoerr's atlas may be due to at least two unrelated factors: differences in the number of maturational stages and differences in the reference samples. The number of stages is fewer in Pyle and Hoerr's atlas than in our method. In the Pyle and Hoerr (1955) radiographic atlas, the information concerning the maturation of the knee is presented by 25 representative plates. Each plate represents two skeletal ages: one for males and one for females. The number of plates dedicated to each skeletal age is different, since there are differential maturational rates during the developmental process. For instance, developmental timing during the first 2 years of life is represented by radiographies taken at intervals of 3 months. The equivalence among the representative plates of Pyle and Hoerr's atlas, as well as the maturity stages defined in this study, are depicted in Fig. 7. As Fig. 7 shows, the distal femoral epiphysis at birth is an oval nodule. During the first 2 years of life, the main shape changes are related to the enlargement of this epiphysis and the appearance of the intercondylar notch (first radiographic evidence). These changes are compatible with those described for our maturity stage 1 (Figs. 2 and 7). Between the ages of 3 and 11 in males and roughly2 and8 in females, both condyles become rounder, the metaphyseal margins of the epiphysis are curved, and the intercondylar notch is now well marked. These morphological modifications match well our maturity stage 2 (Figs. 2 and 7). From the age of 12 in males and 9 in females and until the onset of epiphyseo-diaphyseal fusion, the most important shape changes defined in the Pyle and Hoerr's atlas that can be detected in dry bones are related to the development of the medial and lateral epicondyle and changes in the angle between the metaphyseal and lateral mar- gins of the epiphysis. This angle reflects the anterior projec- tion of the lateral condyle and, thus, these shape changes en- compass those occurring during our maturity stage 3 (Figs. 2 and 7). In Pyle and Hoerr's atlas, fusion begins at 14.5 years in females and 17 years in males (maturity stage 4 in this study). The growth plates are totally replaced by lines of fusion at 15.5 years in females and 18 years in males (maturity stage 5 in this study). At this point, it is important to note that Pyle and Hoerr's atlas documents modal development, not the extent of possible variation in the timing of different developmental events. Thus, it is possible that the age intervals for each maturity stage were larger. Fortunately, Hackman and Black (2013) and Schaefer et al. (2015) have rectified this issue by documenting variation observed when utilizing Pyle and Hoerr's atlas. Hackman and Black (2013) reported standard deviations ranging from 9.86 months in females to 10.75 months in males for any age group in a modern Scottish population. The standard deviations reported by Schaefer et al. (2015) are 2.5 months for females and 2.3 for males at birth, 5.2 months for females between 1 and 3.8 years, and 7.0 months for males between 1 and 4.5 years in a sample mainly composed of European American, African American, and Hispanic individuals. But even considering these wider ranges for Pyle and Hoerr's atlas, the age interval for each stage was larger in our study, and the larger the age interval, the lower the precision.

Regarding differences related to reference samples, while the Pyle and Hoerr atlas comprises 4483 radiographs, the size of our reference sample is small, and the age coverage is uneven. The reference sample used in this study is positively biased towards young children up to 5 years of age (Fig. 1). Moreover, there are fewer individuals at maturity stages 2 and 3 than at stage 1 (Table 1). Due to inter-individual variation in skeletal variation, a larger sample would be desirable to capture all the variation. This shortcoming has been partially overcome with the Bayesian approach.

As mentioned previously, age ranges obtained from the Bayesian approach are wider than those assessed from the age ranges for each maturity stage, but, with this approach, we enlarged age ranges of the three maturity stages. However, the extent of variability that occurs throughout different developmental periods is not static; the developmental variation is wider during the adolescent years than in younger ages (Schaefer et al. 2015). Based on this, narrower age intervals are expected in maturity stages encompassing infancy and early childhood than those encompassing adolescence. Thus, it would be preferable to have a more evenly balanced age distribution and greater representation of adolescent individuals to capture the majority of skeletal variation yet narrow enough for the estimates to be meaningful. However, it is important to note that this study combines two of the largest skeletal collections of documented immature skeletal remains and it is currently not possible for us to add more identified skeletal material. It is hoped that with access to other skeletal collections by the authors will get a bigger and more evenly distributed sample to solve these problems. It may also be possible, given the low inter-observer error for assessing these maturity stages, to pool together data collected by multiple researchers which could be shared, as suggested by Coqueugniot et al. (2010).

Beside these considerations about size and age distribution of the samples, there is another important difference. The reference sample of Pyle and Hoerr's atlas is composed of white individuals of higher socioeconomic status deliberately chosen for their good health and nutritional status. In contrast, most individuals in the Portuguese collections (our reference sample) represent the lower-to-middle social classes in the city of Lisbon (Cardoso 2006). Maturational rate is known to de- pend on a large number of factors such as ancestry, nutritional intake, and health status (Eveleth and Tanner 1990). Nonetheless, ancestry has been shown to be less influential than nutritional intake and health status (Schmeling et al. 2000).

These environmental factors can affect the rate of ossification, leading to a delay in the skeletal development (Roche 1992). This has been noted by Conceição and Cardoso (2011) in their study of the Lisbon sample. They found that the socioeconomic differences in skeletal maturation assessed from Pyle and Hoerr's atlas range from 1.20 to 1.22 years. Since most of our reference sample is the same as that studied by Conceição and Cardoso (2011), we can assert that our sample shows a delay in skeletal maturation compared with Pyle and Hoerr's original standard. This could be the reason for some differences between two methods. The first notable discrepancy between the current study and Pyle and Hoerr (1955) is the age at which the change from a mature stage 1 to stage 2 and stage 2 to stage 3 occurs. There is a delay of approximately 2 years in the age of change from stage 1 to stage 2 and approximately 4 years between stage 2 and stage 3 in our reference sample compared to that of the Pyle and Hoerr's (1955) study (Table 1, Fig. 7). This delay is even more pronounced when data derived from the probabilistic approach are used. If we consider the upper limits of each stage (see above), there could be individuals in stage 1 up to 7 years of age, in stage 3 up to 13 years of age and in stage 4 up to 17 years of age.

This delay in skeletal development could be also related to sexual differences in the tempo of maturation. Based on the results depicted in Table 1, we did not detect that females matured earlier than males prior to epiphyseo-diaphyseal fusion. These data showed that the largest difference between males and females is found at the onset age of epiphyseo- diaphyseal fusion (stage 4), while Pyle and Hoerr (1955) not- ed a relatively advanced maturity in females starting at the beginning of the maturational process. Undoubtedly, these differences can be attributed to the particular characteristics of the reference sample, as previously mentioned. The larger the sample and the more evenly distributed the ages, the larger the detected variation would be, and thus, earlier maturation in females could be better evaluated. To understand how these shortcomings are responsible for the differences between the two methods, we can focus on the results obtained from the Bayesian approach. As a result, we detected relative advances in the age of onset of stage 3 in females. Thus, although the findings from the probabilistic approach detected a relatively advanced maturation for females earlier than the data depicted in Table 1, we found no cases of earlier maturation in females than in males from the beginning of the maturational process as Pyle and Hoerr (1955) asserted.

These findings have obvious implications for the use of the qualitative method presented herein as an indicator of skeletal age. Based on this, our reference sample may be considered representative of individuals with a delay in skeletal development mainly due to environmental factors. Thus, it is possible that the estimation of skeletal age in individuals with a normal or advanced skeletal development was less accurate and more biased. Nonetheless, neither the results from the independent test of performance nor those from the test of agreement between methods confirmed this possibility. Although the size of the Valladolid sample is small and only covers ages corresponding to stage 1, we have included it in this study for one reason. While our reference sample is formed for Portuguese who lived in the 19th and 20th centuries, the Valladolid sample is composed of contemporary Spanish individuals. Studies conducted during the last several years have identified a positive secular trend in growth and development during the last 25 years (Liversidge et al. 1999). For this reason, today children are maturing earlier than they did at the beginning of the twentieth century (Holtgrave et al. 1997). Thus, it can be assumed that the Valladolid sample represents individuals with a more advanced skeletal development than that of Portuguese individuals. Nonetheless, none of the performance measures calculated in the Valladolid sample differ from those estimated in our reference sample. Otherwise, as demonstrated in the San Pablo sample, the qualitative method based on femoral distal epiphyses and Pyle and Hoerr’s atlas are interchangeable. Moreover, age estimates from both Pyle and Hoerr’s atlas and the weighted mean point estimates statistically differ from the dental age estimation. This indicates that our qualitative method based on weighted mean point estimates is able to detect a delay or advance in skeletal development relative to dental development in the same way as Pyle and Hoerr’s atlas. In contrast, age estimates from mode point estimates did not differ from age estimates from dental development. This discrepancy between the agreement of the two point estimates based on the Bayesian approach and the dental age estimates can be attributed to many variables, including differences between the age structure of the reference sample and that of the target sample.

The most important difference between mode point estimates and the weighted mean estimates is that the former biases the age to be younger than the latter (Tables 1 and 2). It is more accentuated for maturity stage 1, where mode point estimates are equal to 1.5 years old for both sexes and the weighted mean estimate is equal to 3.3 years old. The lack of a significant dif- ference between mode point estimates and dental ages is likely a result of the fact that most (51%) of the San Pablo sample is assigned to stage 1, and the dental ages of individuals in this maturity stage are mainly younger than 3 years. This should be taken into account for the future applicability of this qualitative method. Thus, we recommend the use of the two point estimates to assess the similarity between dental ages and these point esti- mates before establishing whether an individual has a delayed or an advanced skeletal development relative to dental development.

Finally, our method is easier to apply than Pyle and Hoerr's atlas showing a high repeatability and reproducibility. The method developed by Pyle and Hoerr (1955) and its adaptation proposed by Conceição and Cardoso (2011) describe changes occurring in the distal end of the femur and in the proximal end of the tibia and fibula throughout the developmental process. Therefore, they not only evaluate the shape changes in the femoral, tibial, and fibular epiphysis but also assess other maturity indicators, such as those changes related to the ratio of the epiphyseal width vs. the metaphyseal width and the morphological modifications of the metaphyseal surfaces. Thus, in order to assess these stages, the complete knee must be preserved, which is difficult in archeological specimens.

Quantitative method

For the second approach, we used equations to estimate the skeletal ages from the shape of the femoral distal epiphysis. Sex proved to be an insignificant source of variation in the shape of the epiphyses and thus the age estimation formulae were determined for the sexes combined. This is very useful in archeological contexts in which sex is difficult to estimate prior to age estimation. The equations were developed following inverse and classical calibration in order to cover all of the possible scenarios. Both equations can be considered good predictors of age based on the shape of distal femoral epiphyses in skeletal remains. Moreover, the error in contour drawings does not significantly affect the age estimation. The inverse calibration model had lower bias than the classical calibration model. Moreover, the inverse calibration model had higher accuracy and precision than the classical calibration model. These results were not surprising. As mentioned in the "Methods" section, while inverse calibration is preferred when the age distribution for the reference sample forms a reasonable prior, classical calibration is preferred if it is suspected that the estimated ages represent an extrapolation beyond the useful limits of the reference sample ages. As both models were applied to the same sample that was used to develop them, it is expected that the inverse calibration model performs better than the classic calibration model.

The performance of the formulae provided in this study was compared to those proposed by Rissech et al. (2008) and Cardoso et al. (2014b). The inverse calibration model based on the femoral distal epiphysis shape had the highest percentage of individuals whose chronological ages fell within the 95% confidence interval of the estimated age and the lowest bias. In the classical calibration model, the percentage of individuals whose chronological ages fell within the 95% confidence interval of the estimated ages was comparable to those obtained from methods based on the femoral diaphyseal lengths. In contrast, the accuracy expressed as the mean of the absolute differences between the estimated and chronological ages was lower in our models and both methods based on the diaphyseal lengths. However, in all of the cases, the mean difference between the estimated and chronological ages was less than 2 years.

Based on these, we can assert that the use of the inverse calibration model presented herein provided age estimates comparable to, if not better than, those obtained for the methods based on the diaphyseal lengths.

Results from the test of performance of the inverse and classic calibration models in the Valladolid sample showed that the overall bias and precision were better when the classic calibration model was used. As mentioned above, it is expected that the skeletal development in the Valladolid sample is advanced in relation to our reference sample, and in these cases the classical calibration models provide better results.

Age estimates from both the inverse and classical calibration models statistically differ from dental age estimates in the San Pablo sample (Table 9). In both cases, dental age estimates were higher than skeletal age estimates. As in the case of the qualitative method based on morphological changes of the femoral distal epiphysis, these results suggest that both quantitative models presented herein are able to detect an advance or delay in the skeletal development relative to dental development. This clear relationship between age estimates from the femoral distal epiphysis shape and dental ages is not found between age estimates from femoral length and dental ages. Only age estimates from femoral lengths using the equation proposed by Rissech et al. (2008) differed from dental age estimates. This could be due to the fact that the predictive equation proposed by Rissech et al. (2008) was developed using conventional least squares regression, while Cardoso et al. (2014b) modeled age on femoral length using the classic calibration model. Moreover, based on the results presented in Table 9, age estimates from femoral length using the equation provided by Rissech et al. (2008) and those from the femoral distal epiphysis shape using the inverse calibration model are interchangeable. A preliminary study about skeletal growth in the San Pablo sample showed that the femoral growth seems retarded compared to that of the reference sample used in this study (García-González 2013). Thus, the lack of statistically significant differences between dental ages and ages based on femoral length using equations provided by Cardoso et al. (2014b) indicates that age estimates from this classical calibration model are less affected for a delay or advancement in the skeletal growth than the inverse calibration model. If this is true, it is puzzling that age estimates from our classical calibration model statistically differ from age estimates for the classical calibration model based on femoral length. However, environmental factors seem to affect more to skeletal development more than skeletal growth (Cardoso 2007; Conceição and Cardoso 2011). Indeed, ages estimated from fem- oral length are older than those estimated from femoral distal epiphysis shape. Thus, it is likely that the lack of statistical agreement between age estimates from femoral growth and those from femoral maturity is due to a delay of skeletal development relative to skeletal growth in the San Pablo sample.

In sum, femoral distal epiphysis shape seems to be a good maturity indicator from which to predict skeletal age. Although these skeletal age estimates are slightly less accurate than those obtained from femoral lengths, they are suitable for determining if the skeletal development is advanced or delayed.

This implies that, as in the case of skeletal ages predicted from other skeletal regions, they are much more informative about the stage of development of a growing child than chronological age alone (Bayley 1946). Indeed, skeletal ages are applied to predict adult stature and craniofacial growth rather than chronological age (Roche 1992).

The statistically significant negative correlation between femoral length and skeletal ages predicted from femoral distal epiphysis shape also indicates that these ages can be used to assess the developmental status of an individual. Femoral length is the best surrogate of stature. Generally, individuals showing an advanced skeletal development, while tall for their chronological age, tend to be short for their skeletal ages and, individuals showing a delayed skeletal development, are short for their chronological age but tall for their skeletal age (Bayley 1946). Thus, this negative correlation suggests that individuals with a delay in skeletal development are taller than other individuals with the same skeletal age.

Conclusions

This study presented two methodological approaches for age estimating in juvenile skeletons, one qualitative and one quantitative. Although the qualitative method based on femoral distal epiphyses is slightly less precise and accurate than Pyle and Hoerr's atlas, skeletal age estimates from both methods are inter- changeable. Both quantitative methods can be considered good predictors of age based on the shape of the distal femoral epiphyses in skeletal remains. The inverse calibration model is preferable when the skeletal growth and development of the target sample is similar to those of the Portuguese sample. In contrast, if we suspect that our target sample or individual have an advance or delay in the skeletal development, the classical calibration model should be used. In sum, we conclude that both approaches can improve the ability to estimate the skeletal ages of juvenile skeletal remains in cases in which the dental remains are not preserved, and the long bones are broken.

References

Adalian P, Piercecchi-Marti M-D, Bourlière-Najean B, Panuel M, Leonetti G, Dutour O (2002)Nouvelle formule de détermination de l'âge d'un fœtus. CR Biologies 325:261–269

Anderson M, Messner MB, Green WT (1964) Distribution of lengths of the normal femur and tibia in children from one to eighteen years of age. JBJS. 46:1197–1202

Baltanás A (2016) "Outline analysis", the Cinderella of geometric mor- phometrics. Abstract book II Iberian symposium on geometric mor- phometrics. pp 8

Bayley N (1946) Tables for predicting adult height from skeletal age and present height. J Pediatr 28:49–64

Bland JM, Altman DG (1999) Measuring agreement in method compar- ison studies. Stat Methods Med Res 8:135–160

Bland JM, Altman DG (2010) Statistical methods for assessing agree- ment between two methods of clinical measurement. Int J Nurs Stud 47:931–936

Bogin B (1999) Patterns of human growth. Cambridge University Press

Braga J, Heuze Y, Chabadel O, Sonan NK, Gueramy A (2005) Non-adult dental age assessment: correspondence analysis and linear regression versus Bayesian predictions. Int J Legal Med 119:260–274

Buikstra JE, Ubelaker DH (1994) Standards for data collection from human skeletal remains. Archaeological Survey, Fayetteville

Cameron N (2004) Measuring maturity. In: Hauspie RC, Cameron N, Molinari L (eds) Methods in human growth research. Cambridge University Press, pp 108–140

Cardoso HFV (2006) Brief communication: the collection of identified human skeletons housed at the Bocage Museum (National Museum of Natural History), Lisbon, Portugal. Am J Phys Anthropol 129: 173–176

Cardoso HFV (2007) Environmental effects on skeletal versus dental development: using a documented subadult skeletal sample to test a basic assumption in human osteological research. Am J Phys Anthropol 132:223–233

Cardoso HFV (2008a) Epiphyseal union at the innominate and lower limb in a modern Portuguese skeletal sample, and age estimation in adolescent and young adult male and female skeletons. Am J Phys Anthropol 135:161–170

Cardoso HFV (2008b) 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

Cardoso HFV, Ríos L (2011) Age estimation from stages of epiphyseal union in the presacral vertebrae. Am J Phys Anthropol 144:238–247

Cardoso HFV, Saunders SR (2008) Two arch criteria of the ilium for sex determination of immature skeletal remains: a test of their accuracy and an assessment of intra-and inter-observer error. Forensic Sci Int 178:24–29

Cardoso HFV, Pereira V, Rios L (2014a) Chronology of fusion of the primary and secondary ossification centers in the human sacrum and age estimation in child and adolescent skeletons. Am J Phys Anthropol 153:214–225

Cardoso HFV, Abrantes J, Humphrey LT (2014b) Age estimation of immature human skeletal remains from the diaphyseal length of the long bones in the postnatal period. Int J Legal Med 128:809–824

Carneiro C, Curate F, Borralho P, Cunha E (2013) Radiographic fetal osteometry: approach on age estimation for the Portuguese popula- tion. Forensic Sci Int 231:397e1–397e5

Child SL, Cowgill LW (2017) Femoral neck-shaft angle and climate- induced body proportions. Am J Phys Anthropol 164:720–735

Conceição ELN, Cardoso HFV (2011) Environmental effects on skeletal versus dental development II: further testing of a basic assumption in human osteological research. Am J Phys Anthropol 144:463–470

Coqueugniot H, Weaver TD (2007) Brief communication: infracranial maturation in the skeletal collection from Coimbra, Portugal: new aging standards for epiphyseal union. Am J Phys Anthropol 134: 424–437

Coqueugniot H, Weaver TD, Houët F (2010) Brief communication: a probabilistic approach to age estimation from infracranial sequences of maturation. Am J Phys Anthropol 142:655–664

Cowgill L (2010) The ontogeny of Holocene and late Pleistocene human postcranial strength. Am J Phys Anthropol 141:16–37

Demirjian A, Buschang PH, Tanguay R, Patterson DK (1985) Interrelationships among measures of somatic, skeletal, dental, and sexual maturity. Am J Orthod 88:433–438

Eveleth PB, Tanner JM (1990) World variation in human growth. Cambridge University Press

Fazekas IG, Kósa F (1978) Forensic fetal osteology. Akadémiai Kiadó

Ferrante L, Cameriere R (2009) Statistical methods to assess the reliability of measurements in the procedures for forensic age estimation. Int J Legal Med 123:277–283

Flecker H (1932) Roentgenographic observations of the human skeleton prior to birth. Med J Aust 19:640–643

Freeman H (1974) Computer processing of line-drawing images. ACM Comput Surv 6:57–97

García-González R (2013) Estudio comparativo de los patrones de crecimiento y desarrollo corporal en humanos actuales y fósiles a partir del análisis de los huesos largos. PhD Dissertation. Burgos University. Unpublished

García-González R, Carretero JM, Rodríguez L, Arsuaga JL (2013) Skeletal growth pattern in a Portuguese sample. In: Fernandes et al (eds) Abstract book I Bioanthropological meeting. pp 51

Greulich WW, Pyle SI (1950) Radiographic atlas of skeletal development of the hand and wrist. Stanford University Press, Stanford

Hackman L, Black S (2013) The reliability of the Greulich and Pyle atlas when applied to a modern Scottish population. J Forensic Sci 58: 114–119

Himes JH (2004) Why study child growth and maturation? In: Hauspie RC, Cameron N, Molinari L (eds) Methods in human growth re- search. Cambridge University Press, pp 3–26

Holtgrave E, Kretschmer R, Müller R (1997) Acceleration in dental de- velopment: fact or fiction. Eur J Orthodon 19:703–710

Humphrey LT (2003) Linear growth variation in the archaeological re- cord. Cambridge studies in biological and evolutionary anthropolo- gy. p 144–169

Iwata H, Ukai Y (2002) SHAPE: a computer program package for quan- titative evaluation of biological shapes based on elliptic Fourier descriptors. J Hered 93:384–385

Klingenberg CP (1996) Multivariate allometry. In: Advances in morpho- metrics. Springer, Boston, pp 23–49

Konigsberg LW, Frankenberg SR (1992) Estimation of age structure in anthropological demography. Am J Phys Anthropol 89:235–256

Konigsberg LW, Hens SM, Jantz LM, Jungers WL (1998) Stature esti- mation and calibration: Bayesian and maximum likelihood perspec- tives in physical anthropology. Am J Phys Anthropol 107:65–92

Krogman W, Iscan M (1986) The human skeleton in forensic medicine.Thomas, Springfield

Kuhl FP, Giardina CR (1982) Elliptic Fourier features of a closed contour. Comput Vision Graph 18:236–258

Landis JR, Koch GG (1977) An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 33:363–374

Lewis ME (2007) The bioarchaeology of children. Perspectives from biological and forensic anthropology. Cambridge University Press

Liversidge H (2008) Dental age revisited. In: Irish JD, Nelson GC (eds) Technique and application in dental anthropology. Cambridge University Press, pp 234–251

Liversidge H, Speechly T, Hector M (1999) Dental maturation in British children: are Demirjian's standards applicable? Int J Paed Dentis 9: 263–269

Liversidge HM, Smith BH, Maber M (2010) Bias and accuracy of age estimation using developing teeth in 946 children. Am J Phys Anthropol 143:545–554

Lopez-Costas O, Rissech C, Trancho G, Turbon D (2012) Postnatal on- togenesis of the tibia. Implications for age and sex estimation. Forensic Sci Int 214:207e1–207e11

Love B, Muller HG (2002) A solution to the problem of obtaining a mortality shcedule for paleodemographic data. In: Hoppa RD, Vaupel JW (eds) Paleodemography: age distributions from skeletal samples. Cambridge University Press, Cambridge, pp 191–192

Lucy D (2005) Introduction to statistics for forensic scientists. Wiley, England, pp 75–93

Mani SA, Naing L, John J, Samsudin AR (2008) Comparison of two methods of dental age estimation in 7–15-year-old Malays. Int J Paed Dentis 18:380–388

Maresh MM (1955) Linear growth of long bones of extremities from infancy through adolescence: continuing studies. Am J Dis Child 89:725–742

McKern TW, Stewart TD (1957) Skeletal age changes in young American males analysed from the standpoint of age identification. Headquarters Quartermaster Research and development command, Technical report (EP-45). Natick, MA

Moorrees CF, Fanning EA, Hunt EE Jr (1963) Age variation of formation stages for ten permanent teeth. J Dent Res 42:1490–1502

Pyle SI, Hoerr NL (1955) A radiographic standard of reference for the growing knee. CC Thomas

Rissech C, Schaefer M, Malgosa A (2008) Development of the femur— implications for age and sex determination. Forensic Sci Int 180:1–9

Rissech C, López-Costas O, Turbón D (2013) Humeral development from neonatal period to skeletal maturity—application in age and sex assessment. Int J Legal Med 27:201–212

Roche AF (1992) Growth, maturation, and body composition: the Fels Longitudinal Study 1929-1991. Cambridge University Press

Roche AF, Thissen D, Wainer H (1975) Skeletal maturity. The knee joint as a biological indicator. Plenum Medical Book Company, New York and London

Rohlf FJ, Archie JW (1984) A comparison of Fourier methods for the description of wing shape in mosquitoes (Diptera: Culicidae). Syst Biol 33:302–317

Ruff C (2007) Body size prediction from juvenile skeletal remains. Am J Phys Anthropol 133:698–716

Salazar A, García-González R, Carretero JM (2017) Methods to quantify shape changes in the proximal metaphysis of the humerus. La Revue de Médecine Légale 8:182

Schaefer MC, Black SM (2007) Epiphyseal union sequencing: aiding in the recognition and sorting of commingled remains. J Forensic Sci 52:277–285

Schaefer M, Hackman L, Gallagher J (2015) Variability in developmental timings of the knee in young American children as assessed through Pyle and Hoerr's radiographic atlas. Int J Legal Med. https://doi.org/ 10.1007/s00414-015-1141-2:1-9

Scheuer L, Black S (2000) Developmental juvenile osteology. Academic press, London

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 Legal Med 113:253–258

Shapland F, Lewis ME (2013) Brief communication: a proposed osteo- logical method for the estimation of pubertal stage in human skeletal remains. Am J Phys Anthropol 151:302–310

Shapland F, Lewis ME (2014) Brief communication: a proposed method for the assessment of pubertal stage in human skeletal remains using cervical vertebrae maturation. Am J Phys Anthropol 153:144–153

Simpson SW, Kunos CA (1998) A radiographic study of the development of the human mandibular dentition. J Hum Evol 35:479–505

Smith BH (1991) Standards of human tooth formation and dental age assessment. Wiley-Liss Inc.

Sokal RR, Rohlf FJ, Lahoz León M (1979) Biometría: principios y métodos estadísticos en la investigación biológica. Blume Ediciones, Madrid

Stevenson PH (1924) Age order of epiphyseal union in man. Am J Phys Anthropol 7:53–93

Stewart TD (1934) Sequence of epiphyseal union, third molar eruption and suture closure in Eskimos and American Indians. Am J Phys Anthropol 19:433–452

Stull KE, L'Abbé EN, Ousley SD (2014) Using multivariate adaptive regression splines to estimate subadult age from diaphyseal dimensions. Am J Phys Anthropol 154:376–386

Sutter R (2003) Non metric subadult skeletal sexing traits: I. A blind test of the accuracy of eight previously proposed methods using prehistoric known-sex mummies from northern Chile. J Forensic Sci 48: 1–9

Tardieu C (1998) Short adolescence in early hominids: infantile and ad- olescent growth of the human femur. Am J Phys Anthropol 107: 163–178

Todd TW (1930) The anatomical features of epiphyseal union. Child Dev 1:186–194

Todd TW (1937) Atlas of skeletal maturation (hand). The CV Mosby Company, St. Louis

Vlak D, Roksandic M, Schillaci M (2008) Greater sciatic notch as a sex indicator in juveniles. Am J Phys Anthropol 137:309–315

Walther BA, Moore JL (2005) The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography. 28:815–829

Zelditch ML, Swiderski DL, Sheets HD (2012) Geometric morphometrics for biologists: a primer. Academic Press