Standardizing ordinal subadult age indicators: testing for observer agreement and consistency across modalities

L.K. Corron[1*], M.K. Stock[2], S.J. Cole[1], C.N. Hulse[1], H.M. Garvin[3], A.R. Klales[4], K.E. Stull[1,5]

[1]Department of Anthropology, University of Nevada, Reno, United States; [2]Department of Sociology and Anthropology, Metropolitan State University of Denver, United States; [3]Department of Anatomy, Des Moines University, United States; [4]Forensic Anthropology Program, Washburn University, United States, [5]University of Pretoria, South Africa

*Corresponding author contact information:

Dr. Louise Corron

Department of Anthropology

University of Nevada, Reno

1664 N Virginia St MS 0096

Reno, NV 89557 USA

Phone: 775-784-4834

lcorron@unr.edu

Highlights

- Inter-observer agreement is high for epiphyseal fusion and dental development scores
- There was no bias in scoring depending on observer experience
- Epiphyseal fusion scores are consistent between CT scans, x-rays and dry bone

•

Abstract

Skeletal and dental data for subadult analyses obtained from dry bones or various types of medical images, such as computed tomography (CT) scans or conventional radiographs/x-rays, should be consistent and repeatable to ensure method applicability across modalities and support combining study samples. The present study evaluates observer agreement of epiphyseal fusion and dental development stages obtained on CT scans of a U.S. sample and the consistency of epiphyseal fusion stages between CT scans and projected scan radiographs/scout images (U.S. CT sample), and between dry bones and conventional x-rays (Colombian osteological sample). Results show that both intra- and interobserver agreements of scores on CT scans were high (intra: mean Cohen's kappa = 0.757–0.939, inter: mean Cohen's kappa = 0.773–0.836). Agreements were lower for dental data (intra: mean Cohen's kappa = 0.757, inter: mean Cohen's kappa = 0.773–0.0.820) compared to epiphyseal fusion data (intra: mean Cohen's kappa = 0.939, inter: mean Cohen's kappa = 0.807–0.836). Consistency of epiphyseal fusion stages was higher between dry bones and conventional x-rays than between CT scans and scout images (mean Cohen's kappa = 0.708–0.824 and 0.726–0.738, respectively). Differences rarely surpassed a one-stage value between observers or modalities. The complexity of some ossification patterns and superimposition had a greater negative impact on agreement and consistency rates than observer experience. Results suggest ordinal subadult skeletal data can be collected and combined across modalities.

Keywords

Forensic anthropology
Epiphyseal fusion
Dental development
Dry bone
Medical imaging
Agreement
Consistency

1. Introduction

The use of medical imaging modalities such as conventional or projection radiographs (x-rays), computed tomography (CT), or magnetic resonance images (MRI), in biological and forensic anthropology is now quite common [[1], [2], [3], [4], [5], [6], [7]]. Efforts have been put into developing virtual databases [8], such as the Pediatric Radiology Interactive Atlas/PATRICIA [9], and the Subadult Virtual Anthropology Database (SVAD) (National Institute of Justice Awards 2015-DN-BX-K409 and 2017-DN-BX-0144, [10]), which are particularly helpful in addressing the lack of subadult osteological reference collections [11]. These data repositories provide researchers with opportunities to develop and validate anthropological methods and access samples that are more reflective of the diversity of modern populations than most donated subadult collections (Komar & Grivas 2008). The value of databases composed of data collected from different sources and modalities depends on data consistency, which raises a number of valid research questions: 1) is there consistency in data collection across different modalities; 2) are there systematic biases in scores; and 3) are the levels of intra- and interobserver agreement comparable across modalities? There is a need to address these concerns to understand limitations and sources of error prior to utilizing multi-modality databases or implementing a method derived from one modality onto another [12,13].

Several publications discuss errors associated with medical imaging. These include the impacts of imaging parameters on the precision of the rendered elements, distortion of anatomical elements in conventional x-rays, and the accuracy of virtual reconstructions from CT scans compared to dry skeletal elements [4,[13], [14], [15], [16], [17], [18]]. Continuous subadult data used in age estimation (*e.g.*, diaphyseal and dental dimensions) have been more commonly evaluated to quantify the comparisons [4,12,13,[16], [17], [18], [19], [20], [21]]. Evaluations into subadult ordinal data scoring consistencies are more limited. Therefore less is known about these variables in terms of consistency across modalities. Research that evaluates ordinal developmental data across modalities generally limits the comparisons to one region of interest or anatomical element, such as epiphyseal fusion of the medial clavicle or development of the third molar. Moreover, they are primarily focused on older subadults without considering a wider age range of skeletal or dental development [[22], [23], [24],25,26,[27], [28], [29],25,26,[30], [31], [32], [33]].

Most of the seminal dental age estimation methods using ordinal dental development stages are based on panoramic radiographs (PAN) [[34], [35], [36], [37]]. Although numerous publications have collected dental development stages from different modalities [27,29,[38], [39], [40], [41]], only Franco et al. [41] and Baumann et al. [39] have evaluated scoring consistency across modalities. Franco et al. [41] applied Gleiser and Hunt's (1955) scoring system for third molar development on PANs, extracted teeth, and cone beam CT (CBCT) scans of 102 individuals (36 males, 66 females, aged 16–50 years). They concluded that 63.3% of scores were the same across all three modalities (*i.e.*, perfect agreement). Baumann et al. [39] compared third molar development stages scored on MRIs and PANs of 27 individuals (19 females, 8 males, aged 13.6–23.1 years) using Demirjian et al.'s (1973) system. They found that stages tended to be slightly lower on MRIs than on PANs. Despite these inconsistencies, both Franco et al. [41] and Baumann et al. [39] reported only one-stage differences across the modalities tested with no statistical differences between stages

scored on different modalities. However, these studies did not always use the same statistics to calculate intra- and interobserver agreement rates of scores, thus the results are not directly comparable. Franco et al. [41] reported weighted Cohen's kappa (K) values for intraobserver agreement (K = 0.91 for both PAN and CBCT and K = 0.93 for extracted teeth) and interobserver agreement (K = 0.69 for CBCT, K = 0.8 for PAN, and K = 0.86 for extracted teeth) of third molar development. Baumann et al. [39] presented lower unweighted K values for interobserver agreement for both dental mineralization (K = 0.45 for PAN and 0.51 for MRI) and dental eruption (K = 0.76 for PAN and 0.57 for MRI).

[25,26] used an 11-stage system originally developed by Schmeling et al. [42] and adapted by Wittschieber et al. [43] on CT scans to evaluate intra- and interobserver agreement rates of sternal clavicle epiphyseal fusion scores from MRIs of 524 individuals aged 11–30 years. Mean weighted K values equaled 0.82 for intraobserver agreement and ranged between 0.60 and 0.64 for interobserver agreement. Kendall's coefficient of concordance was high (0.80), but a symmetry test revealed observer bias; one observer's scores were systematically lower than the other three. Other researchers have evaluated the consistency of epiphyseal fusion scores between CT scans and MRIs using Schmeling et al.'s (2004) scoring system. Reported observer errors between these two modalities varied greatly, from acceptable [22,30], to high [44,45], and comparable [22,28,30,33] to significantly different [32] per the Landis and Koch (1970) scale. None of the studies cited above tested both agreement among observers and consistency of staging across modalities within the same study.

There is an ostensibly greater concern that scores will vary more significantly between modalities for epiphyseal fusion compared to dental development. Epiphyseal fusion studies primarily focus on comparing radiographic and dry bone specimens and usually warn users that different modalities may lead to different interpretations (*e.g.*, [46,22,47,48]). In early stages of fusion, the small bony bridging at the center of the epiphysis may be more easily detected on radiographs than dry bones [46], especially since the bony connection may be taphonomically affected or dismissed as connective tissue on dry specimens. Issues with orientation and superimposition can also impact interpretation of stages in conventional x-rays [47,48]. In later stages of fusion, two-dimensional radiographic images may show epiphyses as completely fused while dry bones may appear only partially fused [22,46]. In cases of complete fusion, two-dimensional x-rays may show a persistent line of radiodensity that can be mistaken for active fusion when in fact fusion seems macroscopically complete [46]. Fojas et al. [49] compared epiphyseal scores from two-dimensional x-rays of fleshed bones to scores collected from the same dry bones. In contrast to logical assumptions, results indicated that observers assigned higher scores (*i.e.*, more advanced fusion) to dry bones than two-dimensional x-rays. Beyond differences across modalities, the authors also noted substantial differences among observers with different experience levels [49].

Studies evaluating subadult scoring consistency between MRIs and CT scans are rare. The authors could not find any study that evaluated long bone epiphyseal fusion beyond the clavicle [33]. This is likely because of challenges faced when accessing CT scans of young children, as examinations are limited to specific anatomical regions to minimize exposure. However, advanced imaging modalities are becoming more common in medical examiner's offices, and their ability to visualize internal structures and produce high resolution images

may provide more sensitive information about maturation processes than dry bones or conventional x-rays. Three-dimensional imaging such as CT scans and MRIs have the advantages of radiography (*i.e.*, ability to see beyond the external surfaces), with the additional benefit of alleviating issues of orientation and superimposition faced with two-dimensional imaging. By scrolling through slices of an element, practitioners can confirm the presence of bony bridging across the epiphyseal surface, identify the smallest trace of enamel, or visualize the beginning of dental root bifurcation, which allows for more precise definitions and evaluation of stages. An additional advantage of CT scans is the possibility to reconstruct external bone surfaces, making them visually comparable to dry bones (*e.g.*, [17]). This can be particularly helpful when evaluating the final stages of epiphyseal fusion and offers a more direct comparison to dry bone specimens.

There is a clear need for research evaluating both intra- and interobserver agreement of epiphyseal fusion and dental development scoring systems, and their consistency across imaging modalities. Therefore, the first goal of this study was to assess intra- and interobserver agreement of two dental development and epiphyseal fusion staging systems developed on postmortem CT scans of fully-fleshed individuals. If observer agreement proved sufficiently high on CT scans, epiphyseal fusion scores between 1) dry bone specimens and conventional x-rays and 2) projected scan radiograph images (two-dimensional scout images) and CT images would be evaluated. It is beyond the scope of this research to directly evaluate how reliability in staging impacted age at death estimations. However, the findings will further inform practitioners on where error is introduced in age estimations and impact future method developments.

2. Material and methods

2.1. Samples

Two samples from the Subadult Virtual Anthropology Database (SVAD) (NIJ Awards 2015-DN-BX-K409 and 2017-DN-BX-0144) were included in this study. The first sample (n = 20), acquired from the University of New Mexico Health Sciences Center, Office of the Medical Investigator in Albuquerque, New Mexico, United States (U.S.), consists of full-body postmortem CT images and scan projection radiographs (scout images) of individuals between birth and 15 years of age (Table 1). Postmortem CT scanning was conducted using a Phillips Brilliance Big Bore 16-slice multi-detector scanner prior to autopsy, with a 512 × 512 pixel matrix, 1.0 mm slice thickness and 0.5 mm slice overlap, using a soft tissue reconstruction algorithm. CT imaging parameters [6,50] and image processing settings [17,20,51] have the greatest impact on image quality, resolution, and reconstruction [52]. This is especially true for slice thickness and slice interval. An overlap reconstruction uses a slice interval smaller than slice thickness, meaning consecutive tomographic slices overlap with each other, ensuring all anatomical structures are captured by several slices instead of one, increasing precision and accuracy of the rendered images [53]. Both CT scans and scout images were visualized in Amira™ imaging software (Amira™ v.6.5.0, Thermo Fischer Scientific). Observers could scroll through the slices in all three planes (coronal, sagittal and transverse), whereas scout images showed the bodies in the sagittal and coronal planes only.

Table 1. Age and sex distributions of the two samples used in the study

U.S. sample (CT scans and scout images)								
Age (years)	0	1	3	4	5	11	13	14
Sample size	9	1	2	3	1	2	1	1
Sex	3F / 6M	M	M	2F / 1M	F	F	M	F
Colombian sample (dry bones, conventional x-rays)								
Age (years)	14	15	16	17	22			
Sample Size	2	3	3	3	1			
Sex	M	1F / 2M	M	M	M			

The second sample (n = 12) acquired from a contemporary skeletal collection housed at the Anthropology Laboratory of the Universidade de Antioquia, in Medellin, Colombia [54], is the only dry bone sample of the SVAD. It includes individuals aged 14–22 years (Table 1) for which standardized conventional x-rays of proximal and distal epiphyses of long bones were taken on-site using a hand-held x-ray machine (Nomad™ Portable X-ray System).

2.2. Data collection protocols

2.2.1. Epiphyseal fusion stages

The stage definitions in the epiphyseal fusion (henceforth referred to as EF) scoring system were based on the CT scans of the individuals from the U.S. sample. EF was scored for the left appendicular long bones, innominate, carpals, tarsals, patella, and calcaneal tuberosity (Fig. 1). A seven-stage scoring system (Fig. 2) was used for the six left long bones and calcaneal tuberosity and is defined as follows: 0 - the epiphysis has not ossified or appeared; (*i.e.*, absent); 1 - the epiphysis has appeared, but is characterized by the lack of any bony attachments (*i.e.*, present); 1/2 - early active union is used when bony bridging exists, but is between 1 and 25% of the entire surface; 2 - active union is used when bony bridging is between 25 and 50% of the length of the epiphyseal growth; 2/3 - active/advanced union is used when bony bridging is approximately at 50%; 3 - advanced union is characterized by bony bridging greater than half the length of the growth plate, with no, or only minor, radiolucent gaps retained throughout; and 4 - complete fusion, as demonstrated by homogenous radiodensity.

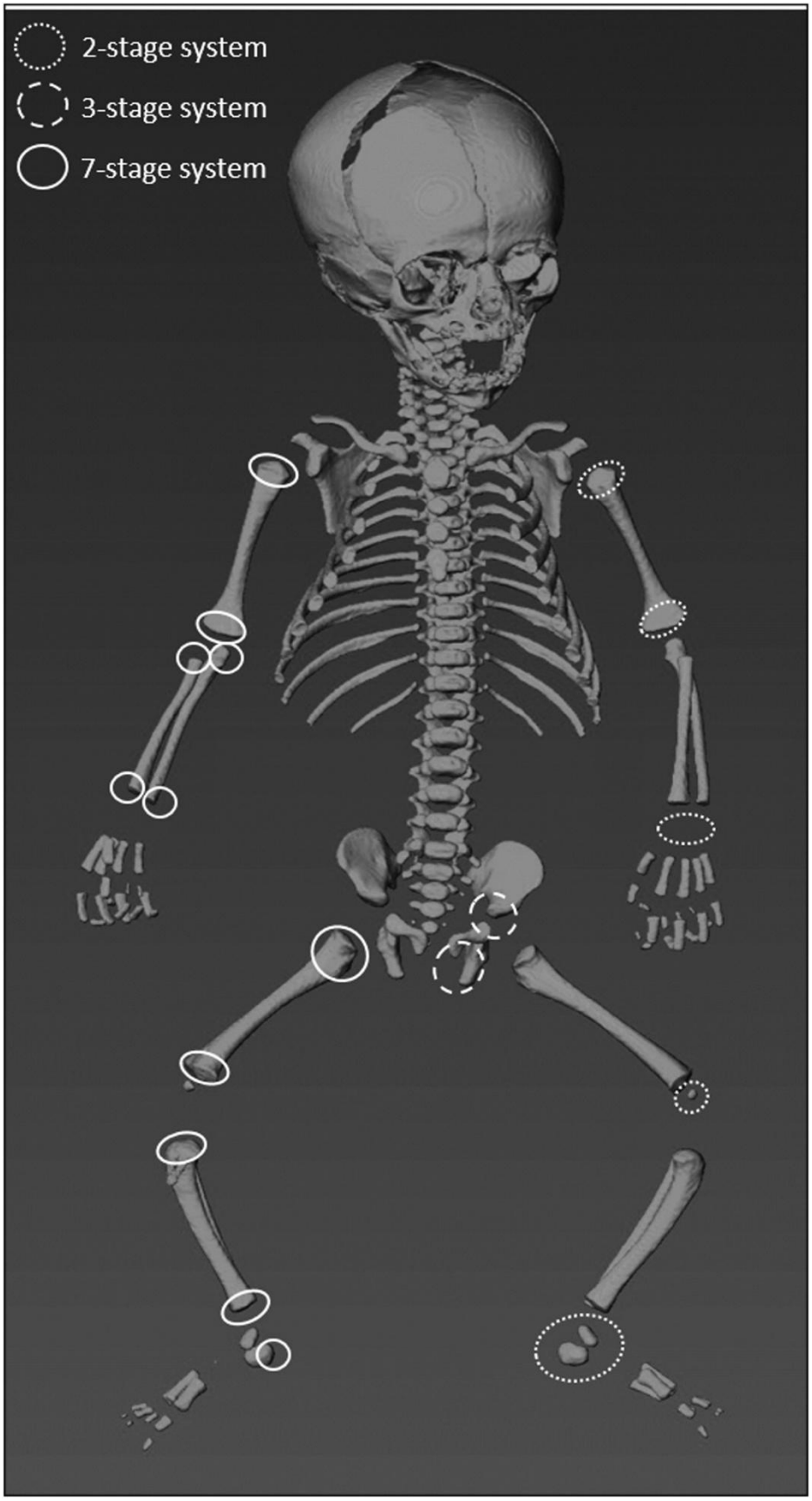

Fig. 1. Epiphyses and ossification centers scored using the three different scoring systems.

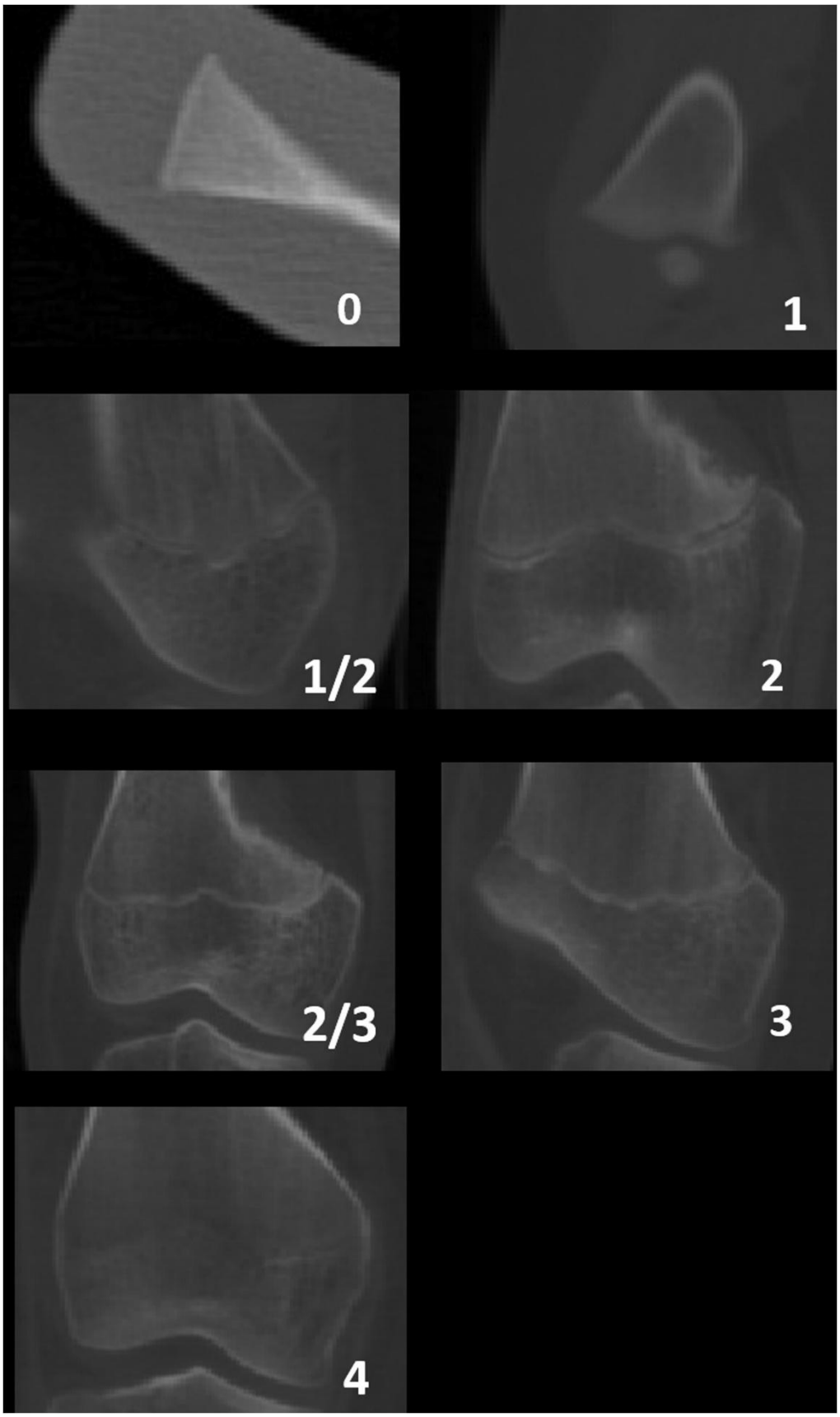

Fig. 2. Seven-stage system used to score fusion for long bone epiphyses and the calcaneal tuberosity. The distal femur is used to illustrate the transition from an absent to a completely fused state (viewed in the coronal plane).

A three-stage scoring system (0 = absence, 1 = active union, 2 = complete union) was used for the pelvis, specifically the ischiopubic ramus and the ilio-ischiatic acetabular epiphysis (Fig. 3). The appearance of ossification centers for the carpals, tarsals, patella, and different centers that comprise the proximal and distal humeral epiphyses were scored with a binary system (presence = 1/absence = 0) (Fig. 4). Observers were instructed to scroll through all planes of the slices to assess any bony bridging and consider the entire 3D metaphyseal surface to estimate the proportion of fused bone.

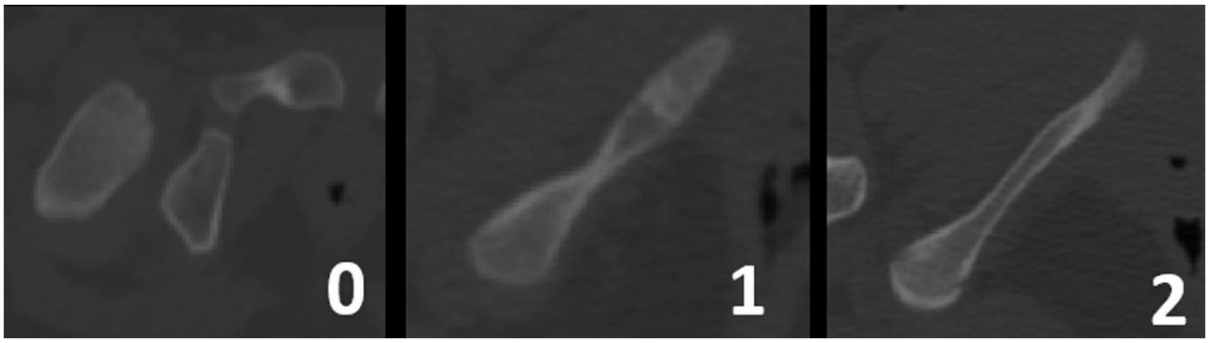

Fig. 3. Three-stage system used to score fusion for the ischio-pubic ramus and the ilio-ischiatic acetabular epiphysis; the ischio-pubic ramus is used in this illustration (viewed in the transverse plane).

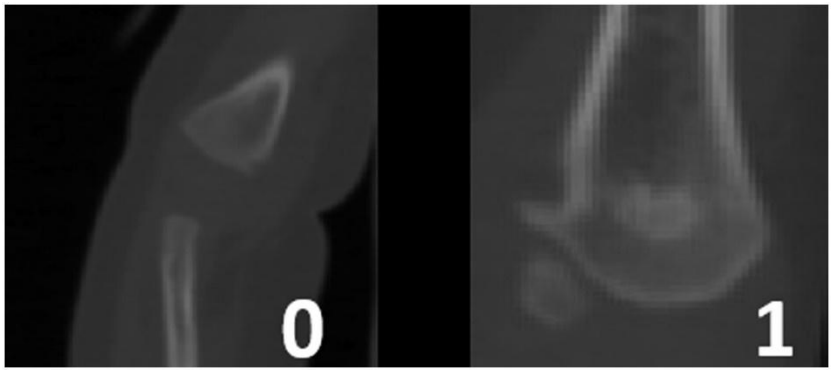

Fig. 4. Two-stage system used to score appearance of the components of the proximal and distal epiphyses of the humerus and the patella; in the current figure the capitulum is imaged (viewed in the coronal plane).

2.2.2. Dental development stages

Dental development of all 32 permanent teeth was assessed on CT slices following AlQahtani et al.'s (2010) 13-level staging system modified from Moorrees et al.'s (1963) stages, which were defined on panoramic dental radiographs [55]. As it was developed both on a skeletal collection (*i.e.*, dry bone) and PANs of living children from a clinical setting, it is presumed there would be few to no differences in developmental stages across modalities. The observers were instructed to scroll through the slices in all three planes to decide the appropriate stage. In case of asymmetrical root development for a given tooth, the more advanced stage was recorded.

Unfortunately, the four modalities evaluated here (CT scan, scout images, dry bone, and conventional x-rays) could not be obtained for a single sample. Furthermore, dental development could only be scored on the CT scans (Fig. 5, 1a), as the scout images presented with insufficient resolution and high superimposition that prevented a clear identification of developmental dental stages (Fig. 5, 2a). Therefore, consistency was evaluated only for EF data between CT scans and scout images in the U.S sample and between dry bone and conventional x-rays in the Colombian sample.

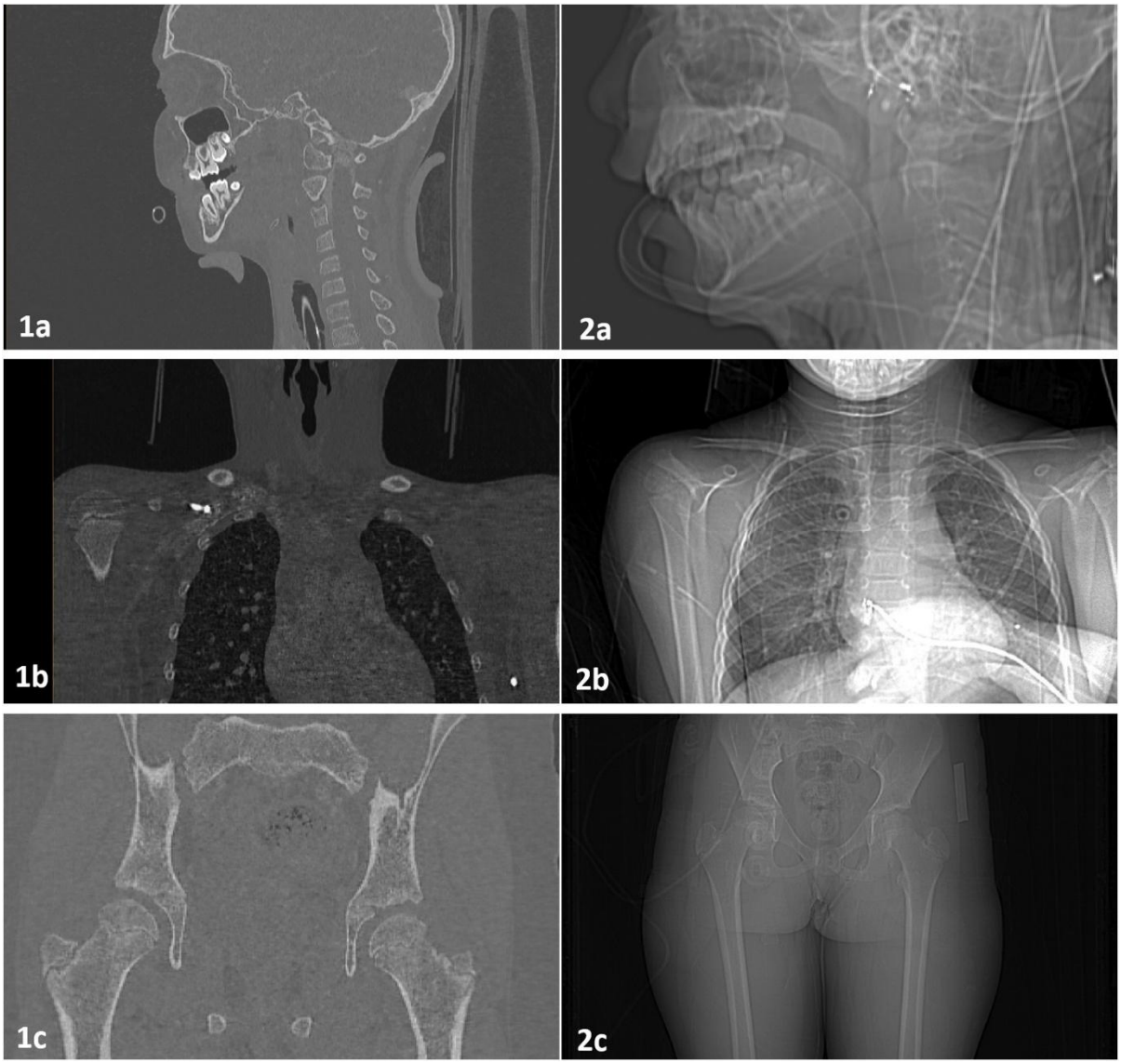

Fig. 5. Side by side comparison of CT scans (left, images 1a to 1c) and scout images (right, images 2a to 2c) used for scoring dental development (1a, 2a for illustration of the corresponding scout image) and epiphyseal fusion (humeral head 1b and 2b, and femoral head 1c and 2c) for the individuals from the U.S. sample. Images 1a and 2a are viewed in the sagittal plane and 1b to 2c in the coronal plane.

All data were collected in the KScollect graphical user interface (GUI), which automatically saves the data in an R data format (RDS) and preserves the data structure; the GUI is available for download at https://github.com/geanes/KScollect [56]. The abbreviations for

the indicators and the scoring systems used for EF and dental development are available in the Supplementary Material.

2.3. Intra- and interobserver agreement – EF and dental development stages

All intraobserver agreement rates for EF and dental development stages were first calculated from CT scans (Table 2) of the U.S. sample. Two observers rescored the same individuals a few weeks after initial data collection: MKS for EF stages and CNH for dental stages. For interobserver agreement, EF was scored by three observers (MKS, LKC, and SJC). Dental development of all permanent teeth was also scored by three observers (LKC, CNH and KES). To evaluate the impact of experience in scoring, observers demonstrated a range of experience (from none to experienced) in the data collection protocols and/or the use of Amira™ software (Table 2).

Table 2. Indicators, observer experience and number of observers per type of indicator analyzed for *intra- and interobserver agreement on CT scans in the U.S. sample*

Agreement	Number of sites scored by observer	Number of observers	Experience in data collection and/or with the modality
Epiphyseal fusion stages			
Intraobserver	1160	1 (MKS; 2 trials)	Experienced in data collection and interpreting CT scan images
Interobserver		3 (MKS, LKC, SJC*)	• MKS and LKC: experienced in data collection and in interpreting CT scan images • SJC: no experience in data collection or in interpreting CT scan images
Dental development stages			
Intraobserver	320	1 (CNH; 2 trials)	• No experience in data collection • Moderate experience in interpreting CT scan images
Interobserver		3 (LKC*, CNH*, KES)	• LKC and CNH: no experience in data collection; experience in interpreting CT scan images • KES: experienced in data collection and interpreting CT scan images
* Indicates observers with less experience in data collection on CT scans than the other observers.			

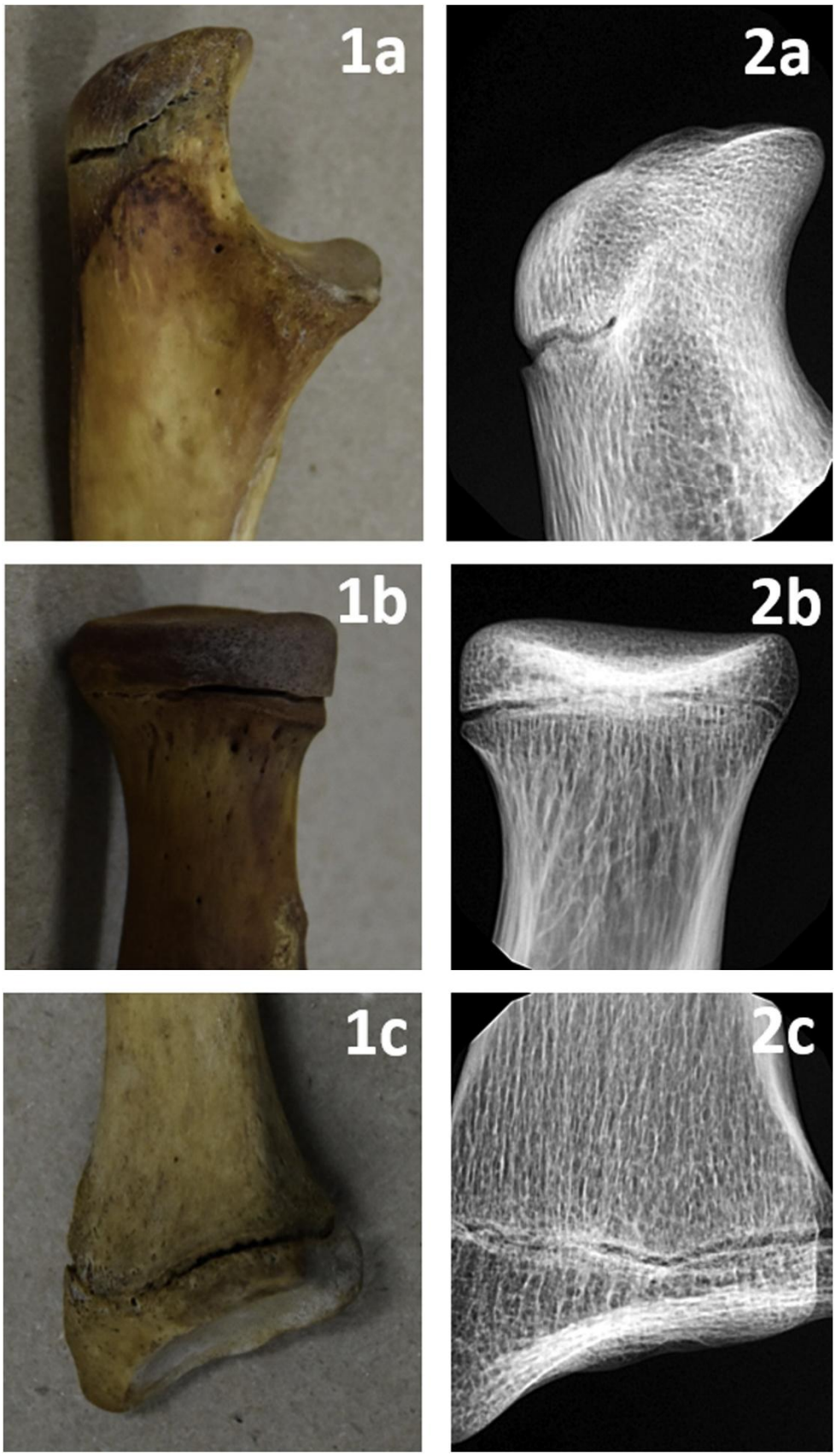

Fig. 6. Side by side comparison of dry bones (left, images 1a to 1c) and conventional x-ray images (right, images 2a to 2c) of epiphyseal fusion sites scored on both modalities in the Colombian sample.

2.4. Consistency across modalities – EF stages

EF was scored by two observers (MKS and LKC) on the U.S. sample scout images and CT images (Fig. 5 and Table 3) and by two observers (LKC and KES) on the Colombian sample of dry bone and conventional x-rays (Fig. 6 and Table 3). Observers were familiar with the initial scoring system developed on CT scans when scoring conventional x-rays, dry bones, or scout images. Because the Colombian sample was dry bone, preservation, damage, or trauma to the skeletons limited the number of sites that could be scored for each individual. There was no way of knowing with certainty whether the absence of an epiphysis (stage 0) was physiological (a true 0) or due to *post-mortem* loss of an unfused epiphysis (stage 1). To avoid biasing the results, comparisons were only done on sites with active or complete fusion: stages 1/2–4 on the seven-stage scoring system. This resulted in the Colombian sample having a narrower age range with higher ages compared to the U.S. sample, but also ensured these more complex, and therefore potentially more error-prone stages, were emphasized in the evaluation.

Table 3. Description of the samples used to assess consistency of epiphyseal fusion across modalities

Sample origin	Modality	Number of sites scored by observer	Number of observers
Colombia	Dry bone	61	2 (LKC and KES)
	Conventional x-ray		
U.S.	CT scan	1160	2 (MKS and LKC)
	Scout image		

2.5. Complexity of staging systems

Several authors have modified staging systems when applying them to a different anatomical element or imaging modality than the one they were initially developed on in an effort to increase precision or reliability [57,58]: this can be done by subdividing existing stages or by collapsing existing stages, respectively. To check if the complexity of the staging systems affected agreement rates in our study, both the seven-stage EF system and the 13-stage dental scoring system were collapsed to create simplified approaches. Specifically, EF stages 1/2 (early active) and 2/3 (active advanced) of the seven-stage scoring system were collapsed as 2 (active) to create a more classic five-stage system (stages 0–4). For dental development, stages 12 (apex closed, root ends converge with wide periodontal ligament (PDL) and 13 (apex closed with normal PDL width) were collapsed into one single stage. These collapsed systems were tested on all available modalities that the full systems were tested on: CT scans, scout images, dry bones, and conventional x-rays for EF, and CT scans for dental development.

2.6. Statistical analyses

The statistical parameters used to quantify intra- and interobserver agreement and assess consistency across modalities were selected based on their suitability for ordinal scales with high numbers of values. They include:

1. Linearly weighted Cohen's kappa (K) for two observers [59], where all levels of disagreement among raters were rated equally between observers. This parameter is best suited in this situation as it is less skewed by a high number of ordinal categories [60]. Cohen's kappa values were qualified as poor (<0.00), slight (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), or almost perfect (0.81–1.00) following the Landis and Koch [61] scale;
2. Kendall's (1948) non-parametric coefficient of concordance (W). This parameter measures interobserver reliability when there are three or more observers, as was the case here and provides an overall estimation of observer agreement. It ranges from 0 (no agreement) to 1 (complete agreement);
3. Percent agreement (percentage of similar or comparable stages scored by two observers), reported both with no tolerance (*i.e.*, absolute agreement) and with a one-stage tolerance to evaluate the extent of scoring differences. Percent agreement was selected as it is easily interpretable, and can be compared with results in other studies;
4. A symmetry test, to assess any systematic biases in observer scores and across modalities, *i.e.*, if scores were consistently higher or lower compared to the other observer(s) or between modalities. A one-sided exact Fisher-Pitman test for paired observations was used to compare the distributions of paired scores. The alternative hypothesis was set to one observer systematically giving higher scores than the other.

The higher number of comparisons available in the U.S. sample allowed evaluations of these statistical parameters per EF site and tooth. To facilitate the evaluation of observer agreement and consistency, the mean K values obtained for each EF site and each tooth were averaged to provide an overall mean K value for all EF scores and all dental scores, respectively. The mean K values for each EF site and tooth can be found in the Supplementary Material. Because fewer EF sites were evaluated in the Colombian sample (61, compared to 1160 in the U.S. sample), EF scores were combined rather than being parsed by anatomical element/location, and the K mean and 95% confidence interval were recorded. All statistical analyses were conducted using the R programming environment [62].

3. Results

3.1. Intra- and interobserver agreement – EF and dental development stages

Intraobserver agreement associated with EF was almost perfect (mean K = 0.939). The percent agreement with no tolerance was 93.7% and increased to 99.2% when a +/- one-stage tolerance was accepted (Table 4). Symmetry tests did not identify any systematic trends or biases in the EF scores by the same observer. Intraobserver agreement was lower for dental development stages, but still conveyed a substantial agreement rate (mean K = 0.757). Percent agreement rate with no tolerance was 53.3% but increased to 86.7% when a one-stage tolerance was incorporated (Table 4). There was a tendency for one observer to give a higher score for dental development of five mandibular teeth (I_1, I_2, C, M_2 and M_3) in the second round (Table 4). However, the percent agreement values indicate that

discrepancies rarely exceeded one stage (only 13.3% of the 320 teeth scored) between rounds (Table 4).

Table 4. Intraobserver agreement results on CT scans

Indicator	Cohen's kappa value ranges*			Percent agreement (%)		Symmetry test ($p < 0.05$)
	Lowest	Mean	Highest	0-stage tolerance	1-stage tolerance	
Epiphyseal fusion stages	0.741	0.939	1	93.7	99.2	Z = -1.6865 p-value = 0.9542
Dental development stages	0.585	0.757	0.898	53.3	86.7	p-value < 0.05 for mandibular I_1, I_2, C, M_2, and M_3
* Lowest and highest kappa and W values are reported here as the lowest and highest results across all epiphyseal sites. Mean kappa and W values are averaged across all sites.						

Interobserver agreement associated with EF was almost perfect (mean K = 0.807–0.836) (Table 5). The lowest mean K value reported for the EF data was 0.429 (moderate). All or part of the ossification centers and/or epiphyses of the distal humerus (HC_Oss, HT_Oss, HLE_Oss, HCE1_EF, HCE2_EF, HDE_EF and/or HME_EF) and/or the proximal radius and ulna (RPE_EF and UPE_EF, respectively) systematically showed lower interobserver agreement rates (see Supplementary Material for detailed results). Mean Kendall's W was 0.792 when incorporating the scores of all three observers. Even with the lowest minimum K values for some of the epiphyses, absolute percent agreement of the least experienced observer was higher than 86.0% and went up to 94.1% with a one-stage tolerance.

Interobserver agreement for dental development data was substantial to almost perfect, with a mean K between 0.773 and 0.820, and mean Kendall's W of 0.90. Percent agreement ranged from 48.2% to 62.5% with a zero-stage tolerance but increased to 77.2% and 91.5% with a one-stage tolerance. The same teeth were found to be problematic between observers and within each observer: I_1, I_2, C, M_2 and M_3 (see Supplementary Material for details). However, paired symmetry tests showed no systematic bias in EF or dental development scores between any pair of observers.

Table 5. Interobserver agreement results on CT scans

Indicator	Observers	Cohen's kappa*			Percent agreement (%)		Symmetry test (p < 0.05)
		Lowest	Mean	Highest	0-stage tolerance	1-stage tolerance	
Epiphyseal fusion stages	MKS & LKC	0.771	0.815	0.918	100	-	Z = 0.013909, p-value = 0.4945
	MKS & SJC**	0.429	0.807	1	87.4	94.1	Z = -4.7643, p-value = 1
	LKC & SJC**	0.467	0.836	1	87.9	94.8	Z = -4.564, p-value = 1
	MKS & LKC &SJC**	Kendall's W*			-	-	-
		Lowest	Mean	Highest	-	-	-
		0.501	0.792	1	86.0	94.1	-
Dental development stages	CNH** & KES	Cohen's kappa*			-	-	-
		Lowest	Mean	Highest	-	-	-
		0.684	0.820	1	62.5	85.3	Z = -4.5092, p-value = 1
	CNH** & LKC**	0.632	0.797	1	60.7	84.8	Z = -3.0984, p-value = 0.999
	KES & LKC**	0.614	0.773	1	61.6	91.5	Z = 1.0871, p-value = 0.1385
	CNH** & KES & LKC**	Kendall's W*			-	-	-
		Lowest	Mean	Highest	-	-	-
		0.875	0.900	1	48.2	77.2	-

* Lowest and highest kappa and W values are reported here as the lowest and highest results across all epiphyseal sites. Mean kappa and W values are averaged across all sites.
** Indicates observers with less experience in data collection from CT scans than the other observers.

3.2. Consistency across modalities – EF stages

Consistency was substantial across all modalities per Landis and Koch's (1977) scale, with mean K values greater than 0.708 for all pairwise modality comparisons (Table 6). The lowest K values (< 0.5) were systematically found between the scout and CT images. However, the lower K values were associated with five sites of the elbow and two sites of the foot: distal humerus (HDE_EF), humeral composite epiphyses (HCE1_EF and HCE2_EF),

proximal radius (RPE_EF) and ulna (UPE_EF), calcaneal tuberosity (CT_EF), and number of tarsals (TC_Oss). The lowest K values (< 0.5) between dry bones and conventional x-rays were associated with the proximal and distal epiphyses of the ulna (UPE_EF, UDE_EF respectively) and the distal end of the humerus (HDE_EF).

Table 6. Consistency across modalities for epiphyseal fusion stages

Modalities	Observer	Cohen's kappa			Percent agreement (%)		Symmetry tests ($p < 0.05$)
		Lowest	Mean	Highest	0-stage tolerance	1-stage tolerance	
CT scan & scout image	MKS	0.429*	0.738*	1*	85.7	97.4	Z = -3.0442, p-value = 0.9988
	LKC	0.301*	0.726*	1*	85.2	95.9	Z = -4.7923, p-value = 1
Dry bone & conventional x-ray	LKC	0.573**	0.708**	0.843**	68.4	98.2	Z = -1.0911, p-value = 0.8624
	KES	0.722**	0.824**	0.927**	80.7	100	Z = -2.1106, p-value = 0.9826

* Lowest and highest kappa and W values are reported here as the lowest and highest results across all epiphyseal sites. Mean kappa and W values are averaged across all sites.
**Lowest, mean and highest kappa values represent the 95% confidence interval for all sites combined for analysis.

For both modality comparisons, the percent agreement was greater between CT and scout images (85.2–85.7%) than between dry bone and conventional x-rays (68.4–80.7%) when no tolerance was accepted (Table 6). The paired symmetry tests did not identify any systematic trends or biases in EF scores to indicate if one modality consistently yielded higher or lower scores.

3.3. Complexity of staging systems

Switching to a five-stage scoring system for EF did not change percent agreement rates for CT scans or scout images, but substantially improved the percentage agreements for dry bones and conventional x-rays (Table 7). Collapsing the last two dental stages into one did not significantly change percent agreement rates (Table 7). Similarly to previous findings, a one-stage tolerance significantly improved agreement rates for all modalities.

Table 7. Percent agreement rates between observers of the uncollapsed and collapsed scoring systems of epiphyseal fusion and dental development for different modalities

Indicator	Observers	Modality	Percent agreement (%) – uncollapsed staging systems		Percent agreement (%) – collapsed staging systems	
			0-stage tolerance	1-stage tolerance	0-stage tolerance	1-stage tolerance
Epiphyseal fusion	MKS – LKC	CT scan	78.1	97.8	78.1	97.9
		Scout image	75.7	96.9	75.7	98.3
	LKC – KES	Conventional x-ray	75.4	75.4	84.2	100
		Dry bone	78.9	78.9	94.7	100
Dental development	LKC – CNH	CT scan	48.2	91.5	54.9	94.6

4. Discussion

4.1. Intra- and interobserver agreement – EF and dental development stages

Following the thresholds defined by Landis and Koch [61], all ordinal score data exhibited substantial to perfect intra- and interobserver agreement (Table 4, Table 5). Values were slightly lower for the observers who were inexperienced in scoring EF stages on CT images (K = 0.429–0.467). Fojas and collaborators (2015) found that experience greatly affected the ability to accurately assess EF, with K values reported between 0.028 and 0.36 when the less experienced observer was involved. Despite some slight differences with experience, the agreement rates in the present study were still fairly good. Moreover, the symmetry tests did not detect any bias in scores between observers (Table 5).

Although these results are encouraging, some anatomical areas emerged as problematic, namely, the epiphyses of the elbow. In previous studies, the distal epiphysis of the humerus (HDE_EF) is generally scored as a single epiphysis and not as the successive union and maturation of several composite epiphyses (HCE1_EF and HCE2_EF here) [47,63]. Therefore, the complexity of its ossification and fusion patterns is rarely appreciated. The ossification of the capitulum is followed by the ossification of the trochlea and the lateral epicondyle. Subsequently, the capitulum and trochlea fuse to form the first composite epiphysis (HCE1_EF); which then fuses with the lateral epicondyle of the humerus to form the second composite epiphysis (HCE2_EF); finally, the medial epicondyle of the humerus (HME_EF) fuses to form the distal humeral epiphysis (HDE_EF) [64,65]. Because some EF sites (such as the distal humerus) are interdependent and mature either simultaneously or successively to form different ossification complexes, under- or over-scoring one of the sites can potentially lead to mis-scoring the composite site to which it contributes. The higher inconsistencies in the distal humerus are thus more likely to be linked to the complexity of the ossification pattern for a site than observer experience.

Dental development stages generally presented with slightly lower agreement rates than EF stages. However, K values achieved for dental stages in the current study generally align with other publications. AlQahtani et al. [55] reported intraobserver K values at 0.81 (n = 50 PANs) and 0.90 (n = 150 extracted teeth). Using a 10-stage scale to score dental development across different modalities, Franco et al. [41] found that PANs and extracted teeth had the highest observer agreement rates (K = 0.8 and 0.86, respectively), but the K value decreased to 0.69 when CT slices were evaluated. Similar to Franco et al. [41], intra- and interobserver differences between developmental stages in the current study did not surpass one stage 84.8% of the time (Table 4, Table 5).

4.2. Consistency across modalities – EF stages

Our results equate to or surpass those of previous studies [49,46,25]: consistency in EF stages between dry bones and conventional x-rays and between CT scans and scout images was similarly high, with rates ranging from substantial to perfect (Table 6). Although stages never crossed the boundaries of unfused, fusing, and fused, there was a tendency to score higher (more advanced fusion) on radiographs than dry bones (18.0% of cases for KES and 14.8% for LKC) and lower on scout images than CTs (9.5% of cases for MKS and 7.6% for LKC. When differences were observed between EF stages taken on dry bones and conventional x-rays, scores were generally higher on the x-rays than the dry bones (11/18 cases for KES, 9/11 cases for LKC). These differences always involved stages 2, 2/3, and 3 but never deviated beyond a consecutive stage. This is most likely due to the high quality of the x-ray images, allowing the observers to see the trabecular patterns and bony bridges very clearly in order to evaluate the state of fusion between 2 (active fusion) and 3 (advanced fusion) more precisely.

Observer experience did not skew the results in the present study, as all symmetry tests failed to reject the null hypothesis of symmetry in paired sets of scores. This is in contrast to Fojas et al. [49] who compared EF scored on conventional x-rays and dry bone and found that observer experience played a substantial role in score discrepancies, with linearly weighted K values ranging from 0.028 to 0.36 for the most novice observer, and from 0.45 to 0.80 for the two more experienced observers.

Differences in scores were slightly higher and more dispersed between CT scans and scout images than between dry bones and conventional x-rays: they generally ranged from two stages lower to three stages higher for MKS with one occurrence at six stages lower and another at three stages lower, and between two stages lower or higher for LKC. Similarly, the lowest consistency between scout images and CT scans and most of the divergence (90/160 for MKS and 44/124 for LKC) were found for the calcaneal tuberosity (CT_EF), the number of tarsals (TC_Oss), and the epiphyses of the elbow. This is likely due to the superimposition of these elements, which prevents them from being clearly visualized (Fig. 7). Superimposition would be particularly problematic for the teeth (Fig. 7, 1a and 2a), which is why dental development was not scored on scout images. With scout images and CT slices, contrast and brightness can be increased or decreased by varying the colormap in the Amira™ software. Contrary to CT slices, scout images do not permit navigation slice by slice in the three different planes. Therefore, there were cases where the position of the anatomical structures and/or of medical equipment prohibited a clear view of all structures

in scout images. The low consistency is because the sites themselves were scored differently and/or because some of the sites were sometimes not scored at all for lack of properly discerning bone contours on scout images. This particular point proves how important a standard body position and adequate acquisition parameters are for two-dimensional data collection – none of which were met for scout images – and calls for a conservative approach in data collection when ideal images are not available.

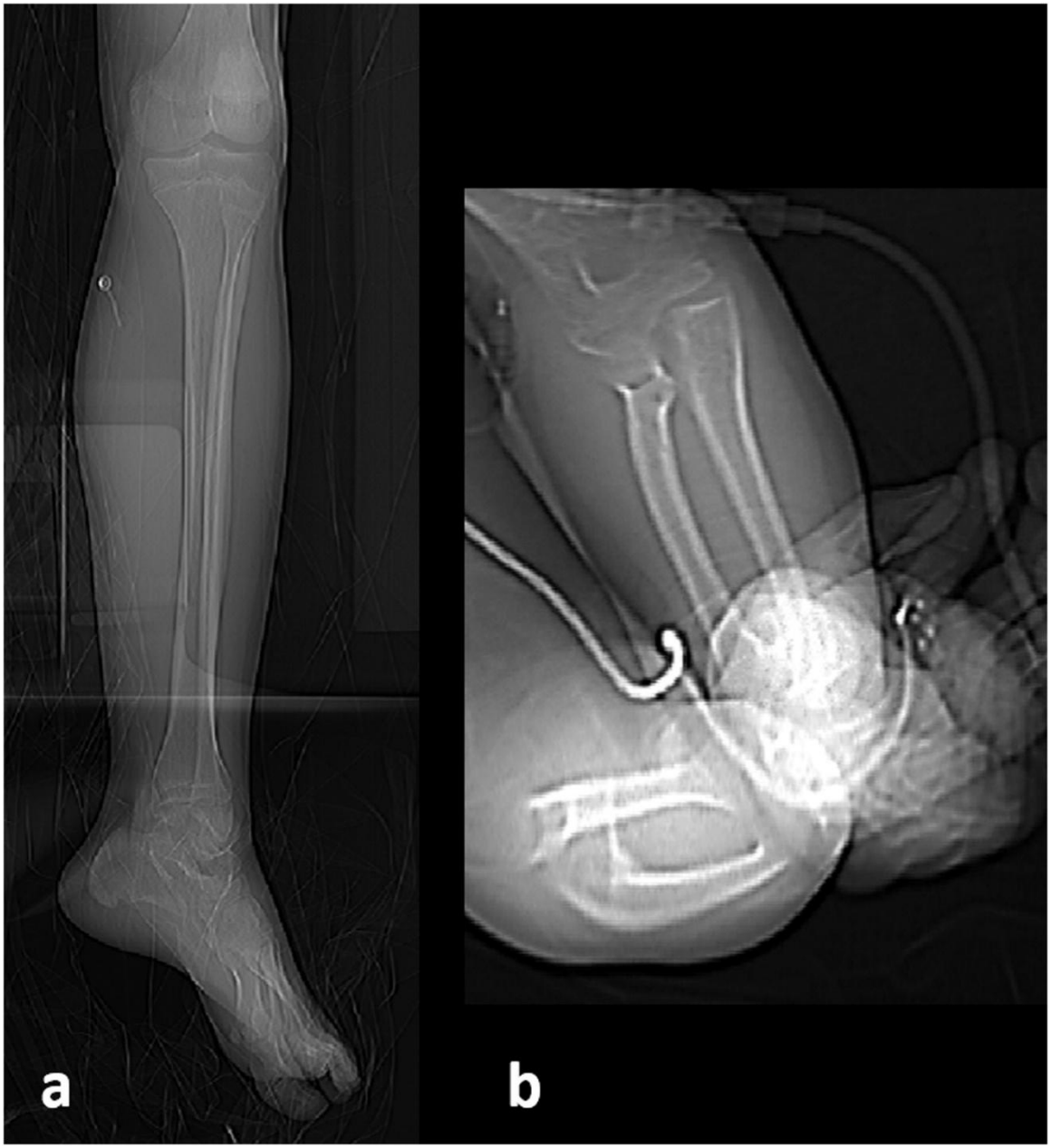

Fig. 7. Examples of superimposition of anatomical structures and medical equipment on scout images preventing the visualization of several epiphyseal fusion sites of the foot (a, sagittal plane) and the elbow (b, coronal plane). Only the epiphyses of the knee (a) are clearly visible.

Also because of superimposition, more EF sites were noted as unscorable on scout images compared to their equivalent on CT scans. Unsurprisingly, the distal humerus epiphyses (lateral epicondyle, HCE1_EF, HCE2_EF, HDE_EF, ossification of the capitulum, trochlea,

medial epicondyle), the proximal and distal radius and ulna epiphyses, carpal and tarsal count, and the calcaneal tuberosity were sites with the most discrepancies. Stages scored on both CT scans and scout images diverged for 12.07% cases for MKS and 13.8% for LKC. When considering a one-stage tolerance, agreement was met in 84.3% sites for MKS and in 72.5% sites for LKC. EF was not scored consistently higher or lower on either modality, nor was there a pattern of certain sites or stages being more frequently discordant than others.

These limitations show the importance of high quality standardized radiographs of elements placed in the anatomical position to ensure an accurate rendition of these elements to score ordinal (and continuous) data. In the field for example, the same type of hand-held portable x-ray as the one used on the Colombian sample would yield usable images for such purposes. As illustrated by scout images, conventional x-rays with low resolution, superimposition, and/or non-standard positioning of anatomical structures, should be avoided.

4.3. Complexity of staging systems

When switching to the simplified EF staging system, percent agreement rates did not improve significantly (Table 7) for CT scans, but they did for dry bones and conventional x-rays. These findings indicate that CT resolution is high enough to discriminate between three levels of active EF compared to two-dimensional images and dry bones. This is particularly striking for dry bones, as it is not possible to visualize the center of the metaphyseal plate where active fusion commences and limits its precise evaluation [46]. Superimposition on conventional x-rays can also hinder the visibility of internal structures. For dry bones and conventional x-rays, the switch to a five-stage system might be more appropriate; although it decreases precision in terms of scoring, it improves agreement and makes it more applicable.

The lower K values associated with dental development could be linked to the higher number of stages in the dental development scoring system compared to the lower number of stages in the EF scoring system. A high number of stages means each stage is only slightly different than the preceding or the following stage. This makes it harder to distinguish between adjacent stages, especially early (stages 7 and 8) and late root development (stages 12 and 13). However, collapsing the last two stages characterizing late root formation and closure (12 and 13) only slightly improved the percentage agreement rates between the three observers (Table 7). This means that substantial improvements to the percentage agreement may require additional collapsing of stages and that the majority of the discrepancies did not occur between these stages. Precise scoring of dental development using a high number of stages relies on high image resolution [25]. The CT scans of fully-fleshed individuals with 1.0 mm slice thickness and 0.5 mm slice overlap used here still present with lower agreement rates than those found on PAN radiographs used in forensic odontology or clinical practice in other studies [39,41]. Indeed, the present results show that resolution and the presence of soft tissue could impact the scoring of dental stages as the contrast between dental and surrounding tissue is not as clear compared to that of standardized panoramic radiographs. For dental data specifically, the combination of image resolution, observer experience, and the high number of dental stages could explain discrepancies in scores between observers and lower Kappa values.

4.4. Final remarks

Consistency of dental development stages was previously tested by several authors across conventional x-rays, dry bones, CT scans, or MRIs. Each showed little to no differences between modalities [29,39,41,55]. However, until protocol consistency and agreement can be tested on a sample of uniform age and sex distributions for which data can be collected and compared on dry bones and several other imaging modalities for the same individuals, studies on consistency and agreement will largely remain bimodality-specific. Indeed, despite its considerable size and the variety of samples composing the SVAD, the Colombian sample is its only subadult reference skeletal sample and one of the few for which both dry bones and conventional x-rays of dry bones are available. Similarly, only CT scans and scout images were available for the U.S. sample. Studies such as this one are therefore still necessary, as high levels of consistency, repeatability and reproducibility (intra- and inter-observer agreement, respectively) are required to combine or compare data taken on different modalities to increase sample sizes and ensure a protocol is generalizable to various case in forensic anthropology and odontology. Using medical images for data collection expands the number of available resources. This not only helps to increase sample sizes, but also augments the number of skeletal and dental indicators obtainable per individual. Moreover, scoring systems developed on medical images are often more sensitive to developmental changes than those developed on dry bones because they allow access to internal structures. Thus, more individual variation can be captured and evaluations of developmental age indicators can be more detailed. This ultimately contributes to improving precision in subadult ageing methods based on these indicators and increases their applicability to different modalities.

5. Conclusions

This study aimed to assess whether epiphyseal fusion (EF) and dental development, two skeletal indicators commonly used in anthropological studies of subadult individuals, could be consistently scored on various modalities of study by different observers with minimal discrepancies. Intra- and interobserver agreements of scores were high on CT scans, although agreements were lower for dental data compared to EF data. Consistency of EF stages was higher between dry bones and conventional x-rays than between CT scans and scout images, but differences rarely surpassed a one-stage value between observers or modalities.

The consistency of EF scores between dry bones and conventional x-rays on the one hand, and CT scans and scout images on the other hand, supports aggregating EF data from various studies and modalities into larger databases. This also indicates that epiphyseal scoring systems developed on one modality can be applied on another during research or casework. Nevertheless, the authors still advise the evaluation of consistency and intra- and interobserver agreement before starting any sort of data collection, as the resolution of conventional x-rays or CT scans may produce error rates exceeding those reported here. Even if data collection protocols are transposable across modalities, they still depend on the visibility, resolution, and definition of certain structures such as root apices or metaphyses, and the definition of bony edges, prompting the need to perform these tests.

Comparing EF between observers and imaging modalities exposed the complex fusion patterns of specific anatomical locations (*e.g.*, distal humerus) and the subsequent decreased agreement and consistency when structures were superimposed. The combination of these factors had a greater negative impact on agreement and consistency rates than observer experience.

Furthermore, the current study confirmed the persevering hypothesis that conventional x-rays would yield more advanced scores of active fusion compared to dry bones, because of the ability to visualize internal surfaces for the former. Nevertheless, scoring remained consistent across modalities when adopting a one-stage difference tolerance. This work is part of the collective effort in biological anthropology to help advance the standardization of protocols and systems for data collection using data sources beyond dry skeletal elements.

CRediT authorship contribution statement

L.K. Corron: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Investigation, Validation, Writing - original draft, Writing - review & editing. M.K. Stock: Data curation, Writing - review & editing. S.J. Cole: Data curation, Writing - review & editing. C.N. Hulse: Data curation, Writing - review & editing. H.M. Garvin: Methodology, Project administration, Supervision, Validation, Writing - original draft, Writing - review & editing. A.R. Klales: Methodology, Project administration, Supervision, Validation, Writing - original draft, Writing - review & editing. K.E. Stull: Conceptualization, Data curation, Conceptualization, Data curation, Methodology, Investigation, Project administration, Resources, Software, Supervision, Validation, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare no conflict of interest

Acknowledgments

The authors wish to thank their collaborators at the Universidad de Antioquia (UDEA), Medellin and the University of New Mexico (UNM), Albuquerque, with specific thanks to Dr. Timisay Monsalve (UDEA) and Dr Natalie L. Adolphi (UNM) for granting us access to the collections. This work was funded by the National Institute of Justice (2015-DN-BX-K409 and 2017-DN-BX-0144) and the National Science Foundation (BCS-1551913). Opinions expressed herein do not necessarily represent the official position or policies of the U.S. Department of Justice or the NIJ.

References

[1] H.M. Garvin, M.K. Stock. The utility of advanced imaging in forensic anthropology. Acad. Forensic Pathol., 6 (3) (2016), pp. 499-516

[2] H.M. Garvin, A.R. Klales, S. Furnier. Emerging technologies in forensic anthropology: the potential utility and current limitations of 3D technologies. R. Johnson (Ed.), Emerging and Advanced Technologies in Diverse Forensic Sciences, CRC Press (2018), pp. 102-131

[3] J.T. Richtsmeier, C.H. Paik, P.C. Elfert. Precision, repeatability, and validation of the localization of cranial landmarks using computed tomography scans. Cleft Palate Craniofac. J., 32 (1995), pp. 217-227

[4] K.E. Stull, M.L. Tise, Z. Ali, D.R. Fowler. Accuracy and reliability of measurements obtained from computed tomography 3D volume rendered images. Forensic Sci. Int., 238 (2014), pp. 133-140

[5] L. Corron, F. Marchal, S. Condemi, K. Chaumoitre, P. Adalian. A new approach of juvenile age estimation using measurements of the ilium and multivariate adaptive regression splines (MARS) models for better age prediction. J. Forensic Sci., 62 (1) (2017), pp. 18-29

[6] A. Brough, C. Villa, K.L. Colman, F. Dedouit, S.J. Decker. The benefits of medical imaging and 3D modelling to the field of forensic anthropology-Positional statement of the members of the Forensic Anthropology working group of the International Society of Forensic Radiology and Imaging. J. Forensic Radiol. Imaging, 18 (2019), pp. 18-19

[7] T. Uldin. Virtual anthropology – a brief review of the literature and history of computed tomography. Forensic Sci. Res., 2 (4) (2017), pp. 165-173

[8] S.D. Berry, H.J.H. Edgar. Extracting and standardizing medical examiner data to improve health. AMIA Joint Summits on Translational Science Proceedings (2020), pp. 63-70

[9] S. Ousley. A Radiographic Database for Estimating Biological Parameters in Modern Subadults. NIJ Award 2008-DN-BX-K152 - Final Technical Report. (2013) 58p

[10] K. Stull. The Subadult Virtual Anthropology Database (SVAD). (2020) https://www.unr.edu/anthropology/research-and-facilities/subadult-database

[11] J. Albanese. Identified Skeletal Reference Collections and the Study of Human Variation. PhD. McMaster University (2003)

[12] C.F. Hildebolt, M.W. Vannier, R.H. Knapp. Validation study of skull three-dimensional computerized tomography measurements.Am. J. Phys. Anthropol., 40 (1990), pp. 283-294

[13] A.L. Brough, J. Bennett, B. Morgan. Anthropological measurement of the juvenile clavicle using multi-detector computed tomography - affirming reliability. J. Forensic Sci., 58 (4) (2013), pp. 946-951

[14] A.L. Brough, G.N. Rutty, S. Black, B. Morgan. Post-mortem computed tomography and 3D imaging: anthropological applications for juvenile remains. Forensic Sci. Med. Pathol., 8 (3) (2012), pp. 270-279, 10.1007/s12024-012-9344-z

[15] C. Robinson, R. Eisma, B. Morgan. Anthropological measurement of lower limb and foot bones using multi-detector computed tomography. J. Forensic Sci., 53 (2008), pp. 1289-1295

[16] K.E. Stull, E.N. L'Abbé, S. Steiner. Measuring distortion of skeletal elements in Lodox Statscan-generated images. Clin. Anat., 26 (2013), pp. 780-786

[17] K.L. Colman, J.G.G. Dobbe, K.E. Stull, J.M. Ruijter, R.J. Oostra, R.R. van Rijn. The geometrical precision of virtual bone models derived from clinical computed tomography data for forensic anthropology. Int. J. Legal Med., 131 (4) (2017), pp. 1155-1163

[18] L. Spake, J. Meyers, S. Blau, H. Cardoso, N. Lottering. A simple and software-independent protocol for the measurement of post-cranial bones in anthropological contexts using thin slab maximum intensity projection. Forensic Imaging (2020), p. 20

[19] K.L. Colman, H.H. de Boer, J.G. Dobbe, N.P.T.J. Liberton, K.E. Stull, M. van Eijnaten, G.J. Streekstra, R.J. Oostra, R.R. Van Rijn, A.E. van der Merwe. Virtual forensic anthropology: the accuracy of osteometric analysis of 3D bone models derived from clinical computed tomography (CT) scans. Forensic Sci. Int., 304 (2019), pp. 1-10

[20] M.K. Stock, H.M. Garvin, L.K. Corron, C.N. Hulse, L.E. Cirillo, A.R. Klales, K.L. Colman, K.E. Stull. The importance of processing procedures and threshold values in CT scan segmentation of skeletal elements: An example using the immature os coxa. Forensic Sci. Int., 309 (2020) no. 110232

[21] L. Corron, F. Marchal, S. Condemi, K. Chaumoître, P. Adalian. Evaluating the consistency, repeatability and reproducibility of osteometric data on dry bone surfaces, scanned dry bone surfaces and scanned bone surfaces from living individuals. Bulletins et Mémoires de la Société d'Anthropologie de Paris, 29 (1-2) (2017), pp. 33-53

[22] R. Schulz, M. Mühler, W. Reisinger, S. Schmidt, A. Schmeling. Radiographic staging of ossification of the medial clavicular epiphysis. Int. J. Legal Med., 122 (2008), pp. 55-58

[23] M. Kellinghaus, R. Schulz, V. Vieth, S. Schmidt, H. Pfeiffer, A. Schmeling. Enhanced possibilities to make statements on the ossification status of the medial clavicular epiphysis using an amplified staging scheme in evaluating thin-slice CT scans. Int. J. Legal Med., 124 (2010), pp. 321-325

[24] L. Meijerman, G.J.R. Maat, R. Schulz, A. Schmeling. Variables affecting the probability of complete fusion of the medial clavicular epiphysis. Int. J. Legal Med., 121 (2007), pp. 463-468

[25] J. De Tobel, E. Hillewig, M. van Wijk, S. Fieuws, M.B. de Haas, R.R. van Rijn, P.W. Thevissen, K.L. Verstraete. Staging clavicular development on MRI: pitfalls and suggestions for age estimation. J. Magn. Reson. Imaging, 51 (2) (2019), pp. 377-388

[26] J.E. De Tobel, M.B. Hillewig, B. de Haas, Van Eeckhout, S. Fieuws, P.W. Thevissen, K.L. Verstraete. Forensic age estimation based on T1 SE and VIBE wrist MRI: do a one-fits-all staging technique and age estimation model apply? Eur. Radiol., 29 (6) (2019), pp. 2924-2935

[27] M.B. Bjork, S.I. Kvaal. CT and MR imaging used in age estimation: a systematic review. J. Forensic Odontostomatol., 36 (1) (2018), pp. 14-25

[28] E. Hillewig, J. De Tobel, O. Cuche, P. Vandemaele, M. Piette, K. Verstraete. Magnetic resonance imaging of the medial extremity of the clavicle in forensic bone age determination: a new four-minute approach. Eur. Radiol., 21 (4) (2011), pp. 757-767

[29] J. De Tobel, E. Hillewig, S. Bogaert, K. Deblaere, K. Verstraete. Magnetic resonance imaging of third molars: developing a protocol suitable for forensic age estimation. Ann. Hum. Biol., 44 (2) (2017), pp. 121-129

[30] S. Schmidt, M. Mühler, A. Schmeling, W. Reisinger, R. Schulz. Magnetic resonance imaging of the clavicular ossification. Int. J. Legal Med., 121 (2007), pp. 321-324

[31] F. Ufuk, K. Agladioglu, N. Karabulut. CT evaluation of medial clavicular epiphysis as a method of bone age estimation in adolescents and young adults. Diagn. Interv. Radiol., 22 (3) (2016), pp. 241-246

[32] V. Vieth, M. Kellinghaus, R. Schulz, H. Pfeiffer, A. Schmeling. Beurteilung des Ossifikations-stadiums der medialen Klavikulaepiphysenfuge. Rechtsmedizin, 20 (2010), pp. 483-488

[33] S. Tangmose, K.E. Jensen, N. Lynnerup. Comparative study on developmental stages of the clavicle by postmortem MRI and CT imaging. J. Forensic Radiol. Imaging, 1 (3) (2013), pp. 102-106

[34] C.F.A. Moorrees, E.A. Fanning, E.E. Hunt. Age variation of formation stages for ten permanent teeth. J. Dent. Res., 42 (6) (1963), pp. 1490-1502

[35] A. Demirjian, H. Goldstein, J.M. Tanner. A new system of dental age assessment. Hum. Biol., 45 (2) (1973), pp. 211-227

[36] H. Mornstad, V. Staaf, U. Welander. Age estimation with the aid of tooth development: a new method based on objective measurements. Scand. J. Dent. Res., 102 (1994), pp. 137-143

CrossRefView Record in ScopusGoogle Scholar

[37] I. Gleiser, E.E. Hunt. The permanent mandibular first molar: its calcification, eruption and decay. Am. J. Phys. Anthropol., 13 (2) (1955), pp. 253-283

[38] J. De Tobel, E. Hillewig, K. Verstraete. Forensic age estimation based on magnetic resonance imaging of third molars: converting 2D staging into 3D staging. Ann. Hum. Biol., 44 (2) (2017), pp. 121-129

[39] P. Baumann, T. Widek, H. Merkens, J. Boldt, A. Petrovic, M. Urschler, B. Kimbauer, N. Jakse, E. Scheurer. Dental age estimation of living persons: comparison of MRI with OPG. Forensic Sci. Int., 253 (0) (2015), pp. 76-80

[40] Y. Guo, A. Olze, C. Ottow, S. Schmidt, R. Schulz, W. Heindel, H. Pfeiffer, V. Vieth, A. Schmeling. Dental age estimation in living individuals using 3.0T MRI of lower third molars. Int. J. Legal Med., 129 (6) (2015), pp. 1265-1270

[41] A. Franco, F. Vetter, E. de Fatima Coimbra, A. Fernandes, P. Thevissen. Comparing third molar root development staging in panoramic radiography, extracted teeth, and cone beam computed tomography. Int. J. Legal Med., 134 (2019), pp. 347-353

[42] A. Schmeling, R. Schulz, W. Reisinger, M. Muhler, K.-D. Wernecke, G. Geserick. Studies on the time frame for ossification of the medial clavicular epiphyseal cartilage in conventional radiography. Int. J. Legal Med., 118 (1) (2004), pp. 5-8

[43] D. Wittschieber, R. Schulz, V. Vieth, M. Kuppers, T. Bajanowski, F. Ramsthaler, K. Puschel, H. Pfeiffer, S. Schmidt, A. Schmeling. The value of sub-stages and thin slices for the assessment of the medial clavicular epiphysis: a prospective multi-center CT study. Forensic Sci. Med. Pathol., 10 (2) (2014), pp. 163-169

[44] R. Cameriere, S. De Luca, D. De Angelis, V. Merelli, A. Giuliodori, M. Cingolani, C. Cattaneo, L. Ferrante. Reliability of Schmeling's stages of ossification of medial clavicular epyphyses and its validity to assess 18 years of age in living subjects. Int. J. Legal Med., 126 (6) (2012), pp. 923-932

[45] S. Tangmose, K.E. Jensen, C. Villa, N. Lynnerup. Forensic age estimation from the clavicle using 1.0 T MRI - Preliminary results. Forensic Sci. Int., 234 (1) (2014), pp. 7-12

[46] F. Introna, C. Campobasso. Biological versus legal age of living people. A. Cunha Schmitt, E. Pinheiro, J. Totowa (Eds.), Forensic Anthropology and Forensic Medicine - Complementary Sciences. From Recovery to Cause of Death, Humana Press, NJ (2006), pp. 57-82

[47] H.F.V. Cardoso. 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 (2008), pp. 97-105

[48] H.F.V. Cardoso. 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 (2008), pp. 161-170

[49] C. Fojas, S. Collins, N. Shirley. Radiographic versus dry bone assessment of sacral epiphyses. FASEB J., 29 (1) (2015), pp. 866-869

[50] M.C. Aalders, N.L. Adolphi, B. Daly, G.G. Davis, H.H. de Boer, S.J. Decker, J.J. Dempers, J. Ford, C.Y. Gerrard, G.M. Hatch, P.A.M. Hofman, M. Iino, C. Jacobsen, W.M. Klein, B. Kubat, P.M. Leth, E.L. Mazuchowski, K.B. Nolte, C. O'Donnell, M.J. Thali, R.R. van Rijn, K. Wozniak. Research in forensic radiology and imaging: identifying the most important issues. J. Forensic Radiol. Imaging, 8 (2017), pp. 1-8

[51] K.L. Colman, A.E. van der Merwe, K.E. Stull, J.G. Dobbe, G.J. Streekstra, R.R. van Rijn, O. Roelof-Jan, H.H. de Boer. The accuracy of 3D virtual bone models of the pelvis for morphological sex estimation. Int. J. Legal Med., 133 (6) (2019), pp. 1853-1860

[52] J.M. Ford, S.J. Decker. Computed tomography slice thickness and its effect on three-dimensional reconstruction of anatomical structures. J. Forensic Radiol. Imaging, 4 (2016), pp. 43-46

[53] M.A. Gavrielides, R. Zeng, K.J. Myers, B. Sahinger, N. Petrick. Benefit of overlapping reconstruction for improving the quantitative assessment of CT lung nodule. olume. Acad. Radiol. 20 (2) (2013), pp. 173-180

[54] T. Monsalve Vargas, J. Isaza. Estudio biosocial de una muestra de restos óseos provenientes de la colección osteológica de referencia de la Universidad de Antioquia. Boletãn Antropol. 29 (47) (2014), pp. 28-55

[55] S.J. AlQahtani, M.P. Hector, H.M. Liversidge, Brief communication: the London atlas of Human tooth development and eruption. Am. J. Phys. Anthropol. 142 (2010), pp. 461-490

[56] K.E. Stull, KScollect: Purpose-built App for Collecting Data for Future Inclusion in KidStats R Package Version 0.6.0.9001 (2017)

[57] S. Schmidt, U. Baumann, R. Schulz, W. Reisinger, A. Schmeling. Study of age dependence of epiphyseal ossification of the hand skeleton. Int. J. Legal Med., 122 (2008), pp. 51-54

[58] U. Baumann, R. Schulz, W. Reisinger, A. Heinecke, A. Schmeling, S. Schmidt. Reference study on the time frame for ossification of the distal radius and ulnar epiphyses on the hand radiograph. Forensic Sci. Int., 191 (1-3) (2009), pp. 15-18

[59] J. Cohen. A coefficient of agreement for nominal scales. Educ. Psychol. Meas., 20 (1) (1960), pp. 37-46

[60] H. Brenner, U. Kliebsch. Dependence of weighted Kappa coefficients on the number of categories. Epidemiology, 7 (2) (1996), pp. 199-202

[61] R.J. Landis, G.G. Koch. The measurement for observer agreement for categorical data. Biometrics 33 (1) (1977), pp. 159-174

[62] R Core Team, R: A Language and Environment for Statistical Computing. (2019) Vienna, Austria

[63] H. Coqueugniot, T.D. Weaver. Brief communication: infracranial maturation in the skeletal collection from Coimbra, Portugal: new aging standards for epiphyseal union. Am. J. Phys. Anthropol. 134 (2007), pp. 424-437, 10.1002/ajpa.20683

[64] L. Scheuer, S. Black. The Juvenile Skeleton. 1 vols, Elsevier Academic Press, London (2004)

[65] L. Scheuer, S. Black., Developmental Juvenile Osteology. 1 vols, Gray Publishing. Elsevier Academic Press, San Diego (2000)