J Forensic Sci, May 2008, Vol. 53, No. 3
doi: 10.1111/j.1556-4029.2008.00715.x
Available online at: www.blackwell-synergy.com

Erin H. Kimmerle,[1] *Ph.D.; Debra A. Prince,*[2] *Ph.D.; and Gregory E. Berg,*[2] *M.A.*

Inter-Observer Variation in Methodologies Involving the Pubic Symphysis, Sternal Ribs, and Teeth*

ABSTRACT: For the skeletal age of a victim to be useful in victim identification, the methods on which it is based must be reliable, accurate, and the results easily duplicated. The ability of multiple investigators to duplicate results is an interesting and complex issue. The purpose of this study is to investigate how consistently multiple investigators assign skeletal traits to rib, pubic symphyseal, or tooth "phases" and measure teeth. The skeletal data from identified individuals in Kosovo are used to test inter-observer variation for a variety of skeletal and dental aging techniques. Two hundred and ninety-six (n = 296) pubic symphyses were scored in the manners of the Todd's ten-phase system and the Suchey-Brooks six-phase system. Six hundred and twenty-two (n = 622) sternal rib ends were scored in the manner of İşcan and co-author's nine-phase system. Four hundred and twelve (n = 412) single-rooted teeth were measured in the manner of Lamendin and colleagues and scored for the amount of tooth wear using Smith's nine-phase system. Repeat measures were taken by multiple observers. There appears to be a wide range of variation, even among experienced investigators in the assignment of phase or metric data. Inter-observer variation, investigated through Pearson's r correlation coefficients, the Wilcoxon signed ranks test, and paired samples t-tests demonstrate significant differences using all methods. How this variation affects the accuracy of age estimation is subject to further investigation, but what is clear is that even with collaboration among investigators to calibrate with one another, the repeatability of numerous aging methodologies is difficult to achieve. Through this investigation it appears the problem lies in the qualitative nature of broad descriptive phase categories, which contain multiple skeletal features and traits that are open to interpretation.

KEYWORDS: forensic science, repeatability, aging methods, identification, Balkans

The repeatability and accuracy of skeletal aging methods have been widely investigated and have produced varied results due to small sample sizes and different statistical methods (1–11). Generally, the consensus has been that aging methods are reliable and that accuracy increases with the investigator's level of experience and the combination of multiple age indicators used together. For example, Baccino and co-workers (11) investigated the reliability of various aging methods, the effect of inter-observer error, and the role of each investigator's experience in producing accurate results. They studied a variety of methods including dental metric techniques, phase methods based on the morphology of the fourth rib and pubic symphyseal face, and femoral cortical thickening. Their sample consisted of a modern forensic sample (n = 19) from France. Baccino and colleagues found that the level of experience among observers plays an important role in the success of accurately estimating the age-at-death and that the need to incorporate a composite of many techniques in the determination of age, rather than a single method, was recommended.

The purpose of this study was to investigate how consistently multiple investigators assign dental and skeletal traits to "phases" and consistently take linear measurements of teeth. A large reference sample (n = 540) was used to investigate five age-at-death methods. Subsamples of males and females were also created to analyze possible sex difference in estimating skeletal age. While this study highlights areas for further methodological refinement, the objective was to investigate how multiple observers assign skeletal traits to phases or repeat linear measurements. Therefore, the focus is specific and does not necessarily inform as to the overall accuracy of individual age estimation.

[1]Department of Anthropology, University of South Florida. 4202 E. Fowler Avenue SOC 107, Tampa, FL 33620.
[2]Joint POW/MIA Accounting Command-Central Identification Laboratory (JPAC-CIL), 310 Worchester Ave, Building 45, Hickam AFB, HI 96853.
*Presented at the 56th Annual Meeting of the American Academy of Forensic Sciences, Dallas, TX, February 18, 2004.

Received 17 Feb. 2007; and in revised form 14 Oct. 2007; accepted 4 Nov. 2007.

Materials and Methods

Data used in this study comes from the UT-ICTY research project that came from evidence collected by the ICTY during its investigation, via The Hague following Chain of Custody. Permission to use this data was given by the ICTY to UT who entered into a working relationship with the expressed goal of sharing data and results that would aid OTP in their investigations as well as other agencies working on human identification in the region. An essential component of this effort was the publication of scientific findings to ensure the admissibility of any new method or revised biological parameters for existing methods in court. The data presented here comes from Kosovo and consists of pubic symphyses scored in the manners of the Todd (12,13) ten-phase system and the Suchey-Brooks (4,14) six-phase system; sternal rib ends scored in the manner of İşcan and co-authors' (15,16) nine-phase system; dental metrics in the manner of Lamendin et al. (17); and Smith's (18) nine-phase system for dental wear. Both the Todd and Suchey-Brooks methods were examined because both have been used in the former Yugoslavia. Only single-rooted teeth were utilized with the dental methods. Repeat measures were collected by the three independent authors of this article. Data collected by a fourth observer (L. Jaimeson Stuart) were also available for categorical methods. Table 1 lists the sample sizes for each tooth or skeletal element by sex. The sample sizes vary among observers who omitted cases they deemed too damaged or incomplete to

TABLE 1—*General sample size by element and sex.*

Skeletal Element	Male *(n)*	Female *(n)*	Total *(n)*
Pubic symphysis	212	84	296
Sternal rib end	540	82	622
Tooth	364	48	412

score. Data comes from evidence collected from autopsy by the United Nations, International Criminal Tribunal for the Former Yugoslavia who gave permission to use and publish this body of work (refer to Kimmerle et al., *Skeletal estimation and identification in American and East European populations,* this volume for more details on the sample history and reliability).

Considerable effort was made prior to analysis among the four observers to calibrate with one another for each method. Several test samples were scored and the results compared. The observers discussed definitions and what features defined each phase. Once this initial calibration took place, no discussion was made among the observers regarding scoring so that all methods were scored independently.

Two tests were used to investigate inter-observer variation for phase data, Pearson's *r* correlation coefficients and the Wilcoxon signed ranks test. Pearson's *r* correlation coefficients were used to investigate the presence and pattern of inter-observer variation. Pearson's *r* correlation coefficients were used as a measure of the linear association between the four observers and as a measure of how the ranked order of phases by each of the observers were related for each phasing method (19). Further, the Wilcoxon signed ranks test measured directional differences among observers by ranking them from low to high and placing a sign based on the direction of that difference (20). These ranked differences were then used to analyze the distributions of scores among observers (20). Additionally, to test for observer variation in dental metric data, paired sample *t*-tests and a repeated measures analysis of variance (ANOVA) were used.

Results

Pubic Symphysis Inter-Observer Variation, Todd Method

Table 2 lists the Pearson's *r* correlation coefficients of the Todd method (12,13) between pairs of observers. The total Kosovo sample, as well as male and female subgroups are analyzed separately. All of the correlations are significant at the $p < 0.01$ level.

TABLE 2—*Pearson r correlation coefficients between pairs of observers for pubic symphysis, Todd method.*

	Observer 1	Observer 2	Observer 3
Total sample (n = 206)			
Observer 1			
Observer 2	0.785*		
Observer 3	0.738*	0.316*	
Observer 4	0.674*	0.549*	0.811*
Males (n = 124)			
Observer 1			
Observer 2	0.738*		
Observer 3	0.538*	0.398*	
Observer 4	0.493*	0.309*	0.735*
Females (n = 82)			
Observer 1			
Observer 2	0.843*		
Observer 3	0.929*	827*	
Observer 4	0.906*	0.812*	0.906*

*Correlation is significant at the 0.01 level.

TABLE 3—*Wilcoxon signed ranks test statistics, Todd method.*

	Obs. 1–2	Obs. 1–3	Obs. 1–4	Obs. 2–3	Obs. 2–4	Obs. 3–4
Z	−0.791	−4.828	−1.995	−4.468	−1.907	−3.906
Asymp. Sig.	0.429	<0.001	0.046	<0.001	0.057	<0.001

Correlation coefficients ranged from $r = 0.316$–0.811 for the total sample; from $r = 0.309$–0.738 among the male subgroup; and from $r = 0.812$–0.906 among the female subgroup. Note that the correlations are the strongest among the female sample.

Table 3 presents the results of the Wilcoxon signed ranks test. Observer 3 is significantly ($p < 0.05$) different from the other three observers. Observer 4 is also marginally significantly different from observers 1 and 2 and highly significantly different from observer 3. The two observers who appear to be in agreement, with no statistical differences among their scores, are observers 1 and 2. Therefore, out of four observers only two are statistically consistent.

Pubic Symphysis Inter-Observer Variation, Suchey-Brooks Method

Table 4 lists the Pearson's *r* correlation coefficients of the Suchey-Brooks method (4,14) among the four observers. Again, the total Balkan sample and male and female subgroups are analyzed separately. All of the correlations are significant at the $p = 0.01$ level. Correlations among observers are higher for Suchey-Brooks method than the Todd method. Correlations among the total sample range from $r = 0.787$–0.857; males range from $r = 0.710$–0.844; and females range from $r = 0.866$–0.939. Similar to the Todd method, the correlations are stronger in females.

The Wilcoxon signed ranks test for observer variation using the Suchey-Brooks method shows a slightly different pattern than that observed for the Todd method, although significant differences are again present among observers (Table 5). A significant difference

TABLE 4—*Inter-observer variation for pubic symphysis, Suchey-Brooks method.*

	Observer 1	Observer 2	Observer 3
Total sample (n = 206)			
Observer 1			
Observer 2	0.845*		
Observer 3	0.843*	0.837*	
Observer 4	0.787*	0.857*	0.821*
Males (n = 124)			
Observer 1			
Observer 2	0.765*		
Observer 3	0.771*	0.742*	
Observer 4	0.715*	0.844*	0.710*
Females (n = 82)			
Observer 1			
Observer 2	0.924*		
Observer 3	0.907*	0.935*	
Observer 4	0.867*	0.866*	0.939*

*Correlation is significant at the $p < 0.01$ level.

TABLE 5—*Wilcoxon signed ranks test statistics, Suchey-Brooks method.*

	Obs. 1–2	Obs. 1–3	Obs. 1–4	Obs. 2–3	Obs. 2–4	Obs. 3–4
Z	−2.141*	−1.224*	−3.265*	−0.994*	−1.812*	−2.427*
Asymp. Sig.	0.032	0.221	0.001	0.320	0.070	0.015

*Based on negative ranks.

is present among observers 1 and 2 ($p = 0.032$), whereas with the Todd method no difference is noted. As with the Todd method, observer 4 significantly differs from the other three observers. Differences among observers 1 and 4 ($p = 0.001$), observers 2 and 4 ($p = 0.015$), and observers 3 and 4 ($p = 0.015$) are evident. Observer 3 is consistent with observers 1 and 2 in that no statistical differences are noted.

Sternal Rib Inter-Observer Variation, İşcan et al. Method

Table 6 lists the Pearson's r correlation coefficients of the İşcan and co-workers (15,16) methods among the four observers. All of the correlations are significant at the $p = 0.01$ level. Correlations among the total sample range from $r = 0.807$–0.844; males range from $r = 0.839$–0.864; and females range from $r = 0.715$–0.872. The correlations tend to be the stronger among males, which is converse to the pubic symphyseal methods. The Wilcoxon signed ranks test for observer variation demonstrates significant differences are present among all observers (Table 7).

Dental Inter-Observer Variation, Smith Method

Table 8 lists the Pearson's r correlation coefficients among the Smith method (18) for tooth wear and age-at-death by tooth type. The sample size by tooth type is also listed in this table. Table 9 lists the Pearson's r correlation coefficients of the Smith method among the four observers. The correlations among observers for repeatability are shown to be high for this method. All of the correlations are significant at the $p = 0.01$ level. Correlations among the total sample range from $r = 0.781$–0.869; males range from $r = 0.781$–0.849; and females range from $r = 0.751$–0.855. The Wilcoxon signed ranks test for observer variation shows significant differences are again present among all observers (Table 10).

TABLE 6—*Inter-observer variation for sternal rib phasing method.*

	Observer 1	Observer 2	Observer 3
Total ($n = 580$)			
Observer 1			
Observer 2	0.840*		
Observer 3	0.836*	0.824*	
Observer 4	0.834*	0.844*	0.807*
Males ($n = 497$)			
Observer 1			
Observer 2	0.850*		
Observer 3	0.846*	0.864*	
Observer 4	0.839*	0.853*	0.851*
Females ($n = 83$)			
Observer 1			
Observer 2	0.825*		
Observer 3	0.798*	0.715*	
Observer 4	0.872*	0.799*	0.749*

*Correlation is significant at the 0.01 level.

TABLE 7—*Wilcoxon signed ranks test statistics for sternal ribs, İşcan et al. method.*

	Obs. 1–2	Obs. 1–3	Obs. 1–4	Obs. 2–3	Obs. 2–4	Obs. 3–4
Z	−10.736*	−7.107*	−14.233*	−3.811**	−9.247*	−10.606*
Asymp. Sig.	<0.000	<0.000	<0.000	<0.000	<0.000	<0.000

*Based on negative ranks.
**Based on positive ranks.

TABLE 8—*Distribution of tooth type by sex used in analysis of teeth.*

Tooth Type	Male (n)	Female (n)	Total (n)
Maxillary incisors	99	11	110
Maxillary canines	44	10	54
Maxillary pre-molars	2	0	2
Mandibular incisors	85	12	97
Mandibular canines	121	14	135
Mandibular pre-molars	13	1	14
Total (n)	364	48	412

TABLE 9—*Inter-observer variation for tooth wear, Smith method.*

	Observer 1	Observer 2	Observer 3
Total sample ($n = 412$)			
Observer 1			
Observer 2	0.781*		
Observer 3	0.843*	0.832*	
Observer 4	0.825*	0.833*	0.869*
Males ($n = 364$)			
Observer 1			
Observer 2	0.781*		
Observer 3	0.849*	0.842*	
Observer 4	0.823*	0.842*	0.871*
Females ($n = 48$)			
Observer 1			
Observer 2	0.751*		
Observer 3	0.810*	0.761*	
Observer 4	0.835*	0.772*	0.855*

*Correlation is significant at the 0.01 level.

Dental Inter-Observer Variation

Four dental metrics were analyzed for inter-observer variation by the four observers (crown height, root height, periodontal recession, and translucency of the root). A repeated measures analysis of variance (ANOVA) produced no significant differences for crown height (Table 11). Significant differences among two of the observers were detected for root height, periodontal recession, and translucency of the root (Tables 12–14).

Discussion and Conclusions

The purpose of this study was to test how different observers compare in assigning phases or collecting metric data. Method

TABLE 10—*Wilcoxon signed ranks test statistics, tooth wear method.*

	Obs. 1–2	Obs. 1–3	Obs. 1–4	Obs. 2–3	Obs. 2–4	Obs. 3–4
Z	−6.420*	−4.601*	−7.345*	−1.948*	−1.533*	−3.458*
Asymp. Sig.	<0.000	<0.000	<0.000	0.051	0.125	0.001

*Based on negative ranks.

TABLE 11—*Paired sample t-test for crown height.*

	Mean Diff.	Std. Dev.	Std. Error	t
Observers 1–2	−0.0428	2.48759	0.13551	−0.316
Observers 1–3	0.2140	2.31519	0.11739	1.823
Observers 2–3	0.2583	2.95544	0.16123	1.602

None significant, $p < 0.05$.

TABLE 12—*Paired sample t-test for root height.*

	Mean Diff.	Std. Dev.	Std. Error	*t*
Observers 1–2	−0.4229	2.90493	0.15572	−2.716*
Observers 1–3	0.832	2.39856	0.12099	0.688
Observers 2–3	0.4263	3.01805	0.16344	2.608*

*$p < 0.05$.

TABLE 13—*Paired sample t-test for periodontal regression.*

	Mean Diff.	Std. Dev.	Std. Error	*t*
Observers 1–2	0.0689	2.85737	0.15339	0.449
Observers 1–3	0.9099	2.33615	0.11784	7.721*
Observers 2–3	0.8631	2.75850	0.14960	5.770*

*$p < 0.05$.

TABLE 14—*Paired sample t-test for translucency of the root.*

	Mean Diff.	Std. Dev.	Std. Error	*t*
Observers 1–2	−0.5242	3.60132	0.19389	−2.703*
Observers 1–3	−0.1013	3.03717	0.15320	−0.662
Observers 2–3	0.4691	3.62290	0.19706	2.380*

*$p < 0.05$.

repeatability among observers may have significant effects on age estimation. However, how multiple traits are used in combination to construct actual age intervals is a different question. The results for ICTY anthropologists as reported in the overview study (Kimmerle et al., *Skeletal estimation and identification in American and East European populations,* this volume) demonstrate that the age estimates of multiple observers are highly accurate. It is important to point out that each of these methods has been looked at independently of one another, that is, each trait was scored in isolation to investigate observer repeatability. When estimating age of a skeleton in practice, the whole skeleton is considered and many factors are weighed when estimating a specific age interval. Practice, repeatability tests, and calibration among investigators are useful tools to ensure the most accurate age estimation possible. Therefore, this exercise was undertaken to better understand how observers differ in interpreting and applying phase definitions and the patterns of variation among practitioners.

The Wilcoxon signed ranks test for observer variation demonstrates significant differences using all categorical phase methods. For the Todd method, two of the four observers are statistically consistent. Overall, correlations among observers vary from low to high but are generally lower than correlations among observers using the Suchey-Brooks method.

Testing observer variation for the Suchey-Brooks method shows a slightly different pattern from that of the Todd method, with more observer variation noted for the former. This is in contrast to previous reports. For example, Galera and colleagues (9) have previously reported substantial, but not statistically significant, differences among observers and that the Suchey-Brooks method was more reliable than the Todd method.

In this investigation, significant differences were also noted among all observers for the sternal rib and dental methods. Although the ANOVA was utilized to determine the dental metric variation among observers, it does not inform us of who agrees with whom. Therefore, the results of this test could be significant because one observer is an outlier, rather than general disagreement among all observers.

Baccino and co-workers (11) argued that observer experience is an important factor and recommended that investigators conduct repeatability tests. In spite of substantial efforts among the authors of this study to calibrate with another in the collection of these data, the tests demonstrate variation for all methods. The differences among observers may have been greater without prior collaboration and repeatability tests. We believe the level of investigator experience is only part of the picture. While all of the observers in this study were graduate students at the time of data collection, all worked on active forensic cases and signed official reports. Further, each of the authors had extensive practical experience, and had been employed by the Smithsonian Institution, the United Nations International Criminal Tribunal for the former Yugoslavia, and the JPAC Central Identification Laboratory. So, what may explain these findings?

It has been suggested that age estimation for younger individuals is more accurate and has higher repeatability than estimation of older aged adults. Observers in this study were more consistent among earlier phases, rather than late phases. However, the "age" of the individual does not offer the whole explanation either. The morphological characteristics, used by investigators to assign each trait to a particular "phase" are diverse. Moreover, the range of morphology, at any given age, varies substantially and is likely a contributing factor to the level of observer variation. This observation has been discussed in the literature, particularly by investigators proposing new methods for use (21). For example, since Brooks' model (22) first developed from Todd's ten phases, casts or line drawings have been used to represent the typical morphology for each symphyseal phase. However, most pubic symphyses do not appear as the prototype due to individual variation or the timing of morphological change (refer to the photographic essay in Kimmerle et al., *Analysis of age-at-death estimation through the use of pubic symphyseal data,* this volume). Additionally, there are a variety of features to consider when assigning a phase, such as the formation of the ventral rim or surface porosity. As a result, there is a wide range of variation in the morphology of the pubic symphysis at any particular age and is further evident by the large age intervals reported for each phase. The same variation in morphological features present at any given age is consistent in all aging methods. For example, Figs. 1 and 2 demonstrate the variation observed among males (known age, 60 years) and females (known ages 50–52) for sternal rib morphology.

A problem particular with the pubic symphysis was noted by observers who collected data used in this study over confusion between the formation and degeneration of the ventral rampart of some pubic symphyses. Interestingly, this problem was also noted by Suchey and Katz (21; 221) who wrote:

> Phase III is somewhat problematic; it peaks in the mid to late twenties, but there are outlying cases trailing into the sixties. These outlying cases are probably interpretative errors on the part of Suchey and Brooks who may have confused buildup of the ventral rampart with breakdown of the ventral rampart.

In addition to error, transition analysis (23, also refer to the other articles in this volume by Kimmerle, Berg, Prince, Konigsberg, and co-workers) demonstrates the variation is widely distributed among each phase, with various forms of traits expressed at any given point in time, which is typical of the human skeletal aging process. Observers appear to vary in their application of phase definitions, each with an emphasis on different aspects of morphology. Compounding this issue is the nature of changing morphology over time—the categories are not discrete but are in a constant state of

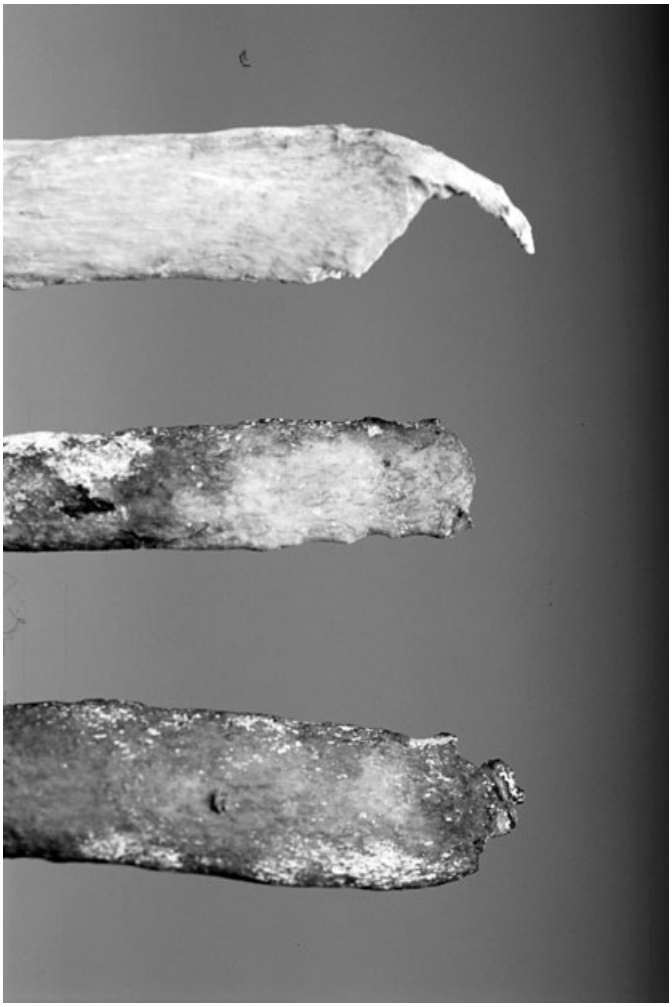

FIG. 1—*The fourth rib from three different female individuals who were all between the ages 50 and 52 years at the time of their deaths. The figure illustrates the range of morphology present at this age.*

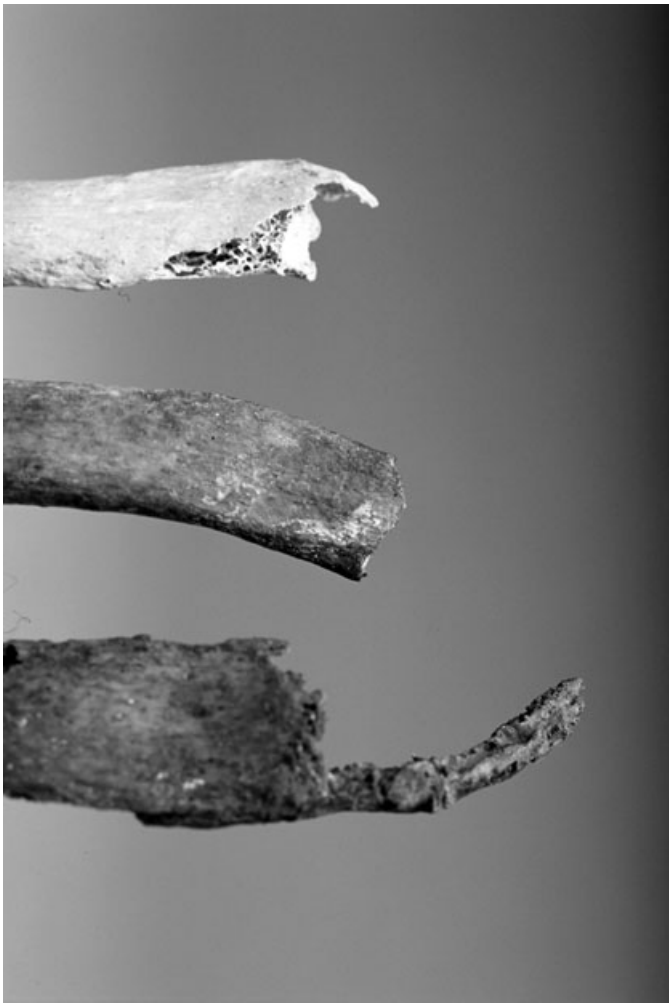

FIG. 2—*The fourth rib from three different male individuals who were all 60 years old at the time of their deaths. The figure illustrates the range of morphology present at this age.*

transition and therefore may reflect characteristics from more than one phase. Investigators should distinguish between those traits that are most important as present or absent to be assigned to a particular phase, which may reduce the amount of observer variation. One strategy for investigators to deal with this issue may be through the use of detailed decision-making trees to supplement phase descriptions, in conjunction with "transition" analysis (refer to Berg, *Pubic bone age estimation in adult women*, this volume).

When the scores of each observer are plotted against one another, consistent patterning among the observers for each method occurs. For the Suchey-Brooks method, there is a tendency for observers to differ by one phase and there tends to be agreement among observers in assigning the earliest and latest categories. However, there is little agreement among observers in assigning phase III of the Suchey-Brooks method, which yielded the most dispersion. Among observers 1 and 4, scores are evenly split between phases IV and VI. However, the largest discrepancies occur from one phase to the next, for example, observers 1, 2, and 4 were divided between assigning phases V and VI, whereas observer 1 compared to the observer 3 were divided between phases IV/V and V/VI (Table 15).

Cross-tabulating the scores of each of the observers shows a similar pattern for rib phases, in that observations tend to disagree by only one phase (Table 16). The most discrepancies occur among

TABLE 15—*Frequency of Suchey-Brooks phases comparing observer 1 scores to those of observers 2–4.*

	Observer 1					
	Phase I	Phase II	Phase III	Phase IV	Phase V	Phase VI
Observer 2						
Phase I	15	0	0	0	0	0
Phase II	0	4	2	0	0	1
Phase III	0	0	5	4	0	0
Phase IV	0	0	4	23	6	4
Phase V	0	0	0	11	22	7
Phase VI	0	0	2	8	17	61
Observer 3						
Phase I	13	0	0	0	0	0
Phase II	2	4	2	0	0	0
Phase III	0	0	3	3	0	1
Phase IV	0	0	1	26	13	1
Phase V	0	0	5	8	25	15
Phase VI	0	0	2	8	7	55
Observer 4						
Phase I	14	0	0	0	0	0
Phase II	1	4	2	0	1	1
Phase III	0	0	5	9	0	2
Phase IV	0	0	1	13	4	3
Phase V	0	0	0	10	12	4
Phase VI	0	0	5	14	27	63

TABLE 16—*Frequency of sternal rib phases comparing observer 1 scores to those of observers 2–4.*

	Observer 1 Phases								
	0	I	II	III	IV	V	VI	VII	VIII
Observer 2									
Phase 0	5	3	0	1	0	0	0	0	0
Phase I	2	11	5	1	2	1	1	0	0
Phase II	0	3	9	12	6	1	1	0	1
Phase III	0	0	3	19	13	3	4	1	0
Phase IV	0	0	0	8	39	24	8	5	2
Phase V	0	0	1	3	24	52	29	8	9
Phase VI	0	0	0	0	2	31	58	16	14
Phase VII	0	0	0	0	0	2	10	34	14
Phase VIII	0	0	0	0	0	2	2	8	67
Observer 3									
Phase 0	5	2	0	0	0	0	0	0	0
Phase I	2	11	4	2	0	0	0	0	0
Phase II	0	3	7	9	4	1	0	0	0
Phase III	0	0	3	19	8	2	1	0	0
Phase IV	0	1	1	9	39	27	6	3	3
Phase V	0	0	3	4	26	60	41	9	6
Phase VI	0	0	0	1	4	20	49	22	12
Phase VII	0	0	0	0	1	2	7	28	34
Phase VIII	0	0	0	0	0	3	9	10	52
Observer 4									
Phase 0	3	8	0	0	0	0	1	0	0
Phase I	3	9	7	5	1	1	0	0	0
Phase II	1	0	6	15	7	2	0	0	0
Phase III	0	0	2	12	11	4	5	0	0
Phase IV	0	0	1	8	34	28	3	2	1
Phase V	0	0	0	2	24	60	52	13	8
Phase VI	0	0	1	1	9	19	35	27	19
Phase VII	0	0	0	0	0	2	7	24	18
Phase VIII	0	0	0	0	0	0	10	6	60

TABLE 17—*Frequency of dental wear phases comparing observer 1 scores to those of observers 2–4.*

	Observer 1 Phases							
	I	II	III	IV	V	VI	VII	VIII
Observer 2								
Phase 0	0	0	0	0	0	0	0	0
Phase I	35	2	0	0	0	0	0	0
Phase II	21	21	3	2	0	0	0	0
Phase III	10	40	17	7	0	0	0	0
Phase IV	4	16	34	36	11	1	0	0
Phase V	1	2	15	35	41	6	0	0
Phase VI	0	1	1	2	13	17	0	0
Phase VII	0	0	0	0	1	6	1	0
Phase VIII	0	0	0	0	0	0	10	0
Observer 3								
Phase 0	0	0	0	0	0	0	0	0
Phase I	34	2	2	0	0	0	0	0
Phase II	26	28	4	3	0	0	0	0
Phase III	8	34	27	24	4	1	1	0
Phase IV	1	14	17	30	14	1	0	0
Phase V	1	3	18	19	32	5	1	0
Phase VI	1	1	2	6	15	17	2	0
Phase VII	0	0	0	0	1	6	7	0
Phase VIII	0	0	0	0	0	0	0	0
Observer 4								
Phase 0	8	0	0	0	0	0	0	0
Phase I	26	4	1	0	0	0	0	0
Phase II	24	19	1	1	0	0	0	0
Phase III	4	27	5	8	1	0	0	0
Phase IV	4	25	25	17	1	0	0	0
Phase V	4	4	32	38	13	0	0	0
Phase VI	0	1	5	15	34	5	1	1
Phase VII	1	1	1	3	17	32	4	4
Phase VIII	0	1	0	0	0	2	6	6

phases III–VII. For example, observer 1 assigned phase III (n = 56), whereas observer 2 assigned the same cases as phase III (n = 13), IV (n = 39), and V (n = 24).

The pattern of tooth wear phases shows disagreement among observers in the earliest stages rather than the later stages (Table 17). Observers usually differed from one another by one phase, among phases I–V. For example, observer 1 chose phase IV (n = 73), whereas observer 3 chose phases III (n = 24), IV (n = 30), and V (n = 19). There was less agreement for phases VI–VII. Observers 2 and 3 never assigned a category VIII, whereas observer 1 (n = 6) and observer 4 (n = 11) each assigned phase VIII.

From each of the methods investigated, it is clear that similar patterns occur among the variation in observer repeatability. First, there is more agreement among younger and older aged individuals than among middle-aged adults. Only the method for scoring tooth wear differed from this trend, in that the earliest of phases showed more disagreement than the later stages. Second, observers usually differ from one another by one phase. The exception to this finding is that observers confuse phases III and IV for phase VI, using the Suchey-Brooks method.

Understanding observer variation highlights trends that when addressed may aid investigators. It also demonstrates what we know about the continuous nature of categories and the need for "transition analysis" (23) in calculating aging methodologies. The results of this investigation highlight areas where multiple investigators vary in assigning age categories. However, the wide age ranges associated with such categories used in combination with multiple traits generally lead to accurate individual age estimation, as there is room for such variation in the process of estimating skeletal age to interpret biological age.

Disclaimer

This study does not represent in whole or in part the views of the United Nations but those of the authors.

Acknowledgments

Funding for this research was provided through a Graduate School Professional Development Award, The University of Tennessee. Permission to use and publish this data was granted by the United Nations, International Criminal Tribunal for the Former Yugoslavia, Office of the Prosecutor and Registry. We thank Mr. David Tolbert, Mr. Peter McCloskey, and Mr. Eamonn Smyth of the ICTY Office of the Prosecutor, for their collaboration in this investigation and for allowing us access to the OTP cases, reports, and evidence. We sincerely thank Jaime Stuart for collaborating and collecting data for this project. We would like to thank Dr. Andrew Kramer for his role in securing funding for this project. We thank Drs. Richard Jantz and Lyle Konigsberg for their comments on this article and their general assistance in making this project possible. Additionally, we thank all anthropologists and other team members that worked over the years with ICTY contributing to bring to justice those responsible for serious human rights abuses and the UT volunteers who assisted during the course of this investigation.

References

1. Suchey JM. Problems in the aging of females using the os pubis. Am J Phys Anthropol 1979;51(3):467–70.

2. Lovejoy CO, Meindl RS, Mensforth RP, Barton TJ. Multifactorial age determination of skeletal age at death: a method and blind tests of its accuracy. Am J Phys Anthropol 1985;68:1–14.
3. Meindl RS, Lovejoy CO, Mensforth RP, Walker RA. A revised method of age determination using the os pubis, with a review of and tests of accuracy of other current methods of pubic symphyseal aging. Am J Phys Anthropol 1985;68:29–45.
4. Suchey JM, Wiseley DV, Katz D. Evaluation of the Todd and McKern-Stewart methods for aging the male os pubis. In: Reichs KJ, editor. Forensic osteology: advances in the identification of human remains. Springfield, IL: C.C. Thomas, 1986;33–67.
5. Klepinger LL, Katz D, Micozzi MS, Carroll L. Evaluation of case methods for estimating age from the os pubis. J Forensic Sci 1992;37(3): 763–70.
6. Saunders SR, Fitzgerald C, Rogers T, Dudar C, McKillop H. A test of several methods of skeletal age estimation using a documented archaeological sample. Can Soc Forensic Sci J 1992;25:97–118.
7. Aiello LC, Molleson T. Are microscopic ageing techniques more accurate than macroscopic ageing techniques? J Archaeol Sci 1993;20:689–704.
8. Bedford ME, Russell KF, Lovejoy CO, Meindl RS, Simpson SW, Stuart-Macadam PL. Test of the multifactorial aging method using skeletons with known ages-at-death from the Grant collection. Am J Phys Anthropol 1993;91:287–97.
9. Galera V, Ubelaker DH, Hayek L. Comparison of macroscopic cranial methods of age estimation applied to skeletons from the Terry Collection. J Forensic Sci 1995;43:933–9.
10. McKeown AH, Meadows Jantz L, Herrmann NP. Evaluation of age data from the forensic data bank. Proceedings of the 49th Annual Meeting of the American Academy of Forensic Sciences; 1997 Feb 17-22; New York, NY; Colorado Springs, CO: American Academy of Forensic Sciences, 1997.
11. Buccino E, Ubelaker D, Hayek L, Zerilli A. Evaluation of seven methods of estimating age at death from mature human skeletal remains. J Forensic Sci 1999;44(5):931–6.
12. Todd TW. Age changes in the pubic bone. Am J Phys Anthropol 1920;3(3):285–339.
13. Todd TW. Age changes in the pubic bone. Am J Phys Anthropol 1921;4(1):1–76.
14. Katz D, Suchey J. Race differences in pubic symphyseal aging patterns in the male. Am J Phys Anthropol 1986;80:167–72.
15. İşcan YM, Loth S, Wright RK. Age estimation from the rib by phase analysis: White females. J Forensic Sci 1985;30:853–63.
16. İşcan YM, Loth S, Wright RK. Age estimation from the rib by phase analysis: White males. J Forensic Sci 1984;29:1094–104.
17. Lamendin H, Baccino E, Humbert JF, Tavernier JC, Nossintchouk RM, Zerilli A. A simple technique for age estimation in adult corpses: the two criteria dental method. J Forensic Sci 1992;37(5):1373–9.
18. Smith BH. Patterns of molar wear in hunter-gatherers and agriculturalists. Am J Phys Anthropol 1984;63(1):39–56.
19. Pagano M, Gauvreau K. Principles of biostatistics. Belmont, CA: Duxbury Press, 1993.
20. Zar JH. Biostatistical analysis. 2nd ed. Englewood Cliffs: Prentice-Hall, Inc. 1984.
21. Suchey JM, Katz D. Applications of pubic age determination in a forensic setting. In: Reichs KJ, editor. Forensic osteology: advances in the identification of human remains. Springfield, IL: C.C. Thomas, 1998;204–36.
22. Brooks ST. Skeletal age at death, the reliability of cranial and pubic age indicators. Am J Phys Anthropol 1955;13:567–97.
23. Boldsen JL, Milner GR, Konigsberg LW, Wood JW. Transition analysis: a new method for estimating age from skeletons. In: Hoppa RD, Vaupel JW, editors. Paleodemography: age distributions from skeletal samples. Cambridge, UK: Cambridge University Press, 2002;73–106.

Additional information:
Erin H. Kimmerle, Ph.D.
Department of Anthropology
University of South Florida
4202 E. Fowler Avenue SOC 107
Tampa, FL 33620
E-mail: kimmerle@cas.usf.edu