Journal of Forensic Science & Criminology
Volume 8 | Issue 1
ISSN: 2348-9804

Research Article Open Access

Intraobserver and Interobserver Error in Osteological Analysis as an Indicator for Non-Expert Skeletal Analysis

Henson K*[1], Harding T[1], Starcher K[1], Smith A[1] and Seccurro D[2]

[1]Forensic Science Program, Natural Sciences Department, Fairmont State University, WV, USA
[2]Daria Seccurro, School of Exercise Science and Athletic Training, West Virginia Wesleyan College, WV, USA

***Corresponding author:** Henson K, Forensic Science Program, Natural Sciences Department, Fairmont State University, 1201 Locust Ave, Fairmont, WV, 26554, USA, Tel: 3043674877, E-mail: Kristy.Henson@fairmontstate.edu

Citation: Henson K, Harding T, Starcher K, Smith A, Seccurro D (2020) Intraobserver and Interobserver Error in Osteological Analysis as an Indicator for Non-Expert Skeletal Analysis. J Forensic Sci Criminol 8(1): 104

Received Date: May 11, 2020 **Accepted Date:** June 09, 2020 **Published Date:** June 11, 2020

Abstract

Accurate skeletal analysis is needed in order to properly identify skeletal remains. Cases of misidentification occur all over the United States and may be avoided with an understanding of observer error and the appropriate level of training. In this study, four novice observers with limited osteological training conducted osteological analyses on three different skeletal remains to assess intraobserver and interobserver reliability. All observers collected data using *The Standards for Data Collection from Human Skeletal Remains and The Human Bone Manual* as well as the same measuring devices. The osteological data were collected over multiple days of observation for each observer. Data were analyzed using ANOVA and Kappa coefficient statistical models to determine reliability. Results indicate that intraobserver error was not significant. Interobserver error, however, had differing results. ANOVA results showed no significant variation while the Kappa coefficient showed low observer agreement. Anthroposcopic analyses were relatively consistent with some variation. To avoid misidentification of skeletal remains, the forensic investigator must consider interobserver error and receive appropriate osteological training.

Keywords: Forensic Anthropology; Skeletal Remains; Osteometric Analysis; Observer Error

Introduction

Forensic scientists, pathologists, and law enforcement personnel regularly encounter human remains during an investigation. Typically, human remains are discovered before they enter into a putrification state, making identification relatively simple. In instances when the remains are heavily decomposed or skeletonized, identification can be extremely challenging if genetic information is unavailable [1,2]. Skeletons can provide critical information about an individual's identification, nutrition, and trauma [1]. Error or misinterpretation of these remains affects the accuracy of a case or an individual's identity [3].

For example, in 1986 a severely decomposed individual was misidentified using dental records [3]. In North Carolina, age and ancestry of 130 cases were misidentified [4]. These cases illustrate why interobserver reliability must be considered [3,4]. According to Crowder C *et al.* [5] there has been a 30% increase in skeletal samples requiring analysis in Texas alone.

When conducting analyses on human remains, osteologists and forensic anthropologists rely on standardized osteometric measurements known as *Standards for Data Collection from Human Skeletal Remains* [6] or the adapted *Data Collection Procedures for Forensic Skeletal Material 2.0* [7]. Osteometric analyses are essential when determining sex, age, and ancestry of a skeleton. Observer errors in these measurements may alter individual characteristic estimations resulting in misidentification of an individual [1,3,7,8]. Osteometric analysis may have some variability in the measurements due to the inherent differences in the physical characteristics of the individuals, but such variability can be avoided with experience [1,5,7,8,9].

Variations due to the measurement error of the examiner can be avoided or controlled to some extent by considering intraobserver and interobserver error. Intraobserver reliability or error is the difference between multiple interpretations of one individual at different times. Interobserver reliability or error is the difference between multiple individuals performing the same task. Lynnerup *et al.* [10] conducted a study of intra and interobserver error comparing experts and novices. Both groups were able to correctly identify the age of 126/159 forensics cases using a very specific Greulich-Pyle aging method. Langley *et al.* (2018) [7] observed the reliability of osteometric data by examining experts' intra and interobserver reliability. Along with misidentification, interobserver comprehension of osteometric data points may affect large datasets and the paleoepidemiology of skeletons [7,11].

In the literature, intra and interobserver reliability when analyzing skeletal remains is lacking. The purpose of this study is to compare the intra and interobserver reliability of non-experts, or novices, when conducting osteological analyses to determine if observer error is contingent on osteological education received. We calculated the interobserver and intraobserver reliability to determine if error reliability can minimize the misidentification of skeletal remains when an expert is unavailable.

Materials and Methods

Four undergraduate students (Table 1) of varying class rank, major, and osteological background analyzed three sets of antique teaching skeletons (referred to as CHS 209, CHS 211, and CHS 212). Observer A analyzed all three skeletons over the course of two months as a summer research experience while the remaining three observers each analyzed one skeleton over two months as a hands-on laboratory experience. All observers shared the same faculty mentor and used an identical osteometric protocol. The faculty mentor showed the observers proper use of the equipment. Then, the observers demonstrated standardized techniques prior to beginning data collection.

Observer	Training	Rank	Skeleton analyzed
A	Advanced Kinesiology, Human Anatomy and Physiology, Undergraduate research in osteology	Senior	CHS 209 CHS 212 CHS 211
B	Advanced Osteology, Forensic Biology	Senior	CHS 209
C	Advanced Osteology, Forensic Biology	Junior	CHS 212
D	Advanced Osteology, Forensic Biology	Junior	CHS 211

Table 1: Rank and status of each observer

The skeletal data and osteological landmarks were analyzed using *Standards for Data Collection from Human Skeletal Remains* (1995) and *The Human Bone Manual* (2005). The observers repeated the osteological measurements of their assigned skeleton five times over the course of multiple weeks. The number of osteological landmarks varied slightly, up to 109 points, dependent on skeletal completeness (e.g. one skeleton was missing the skull). The observers were instructed to complete one full skeletal analysis at a time. Observers took all measurements using Vernier digital calipers and a Ward's osteometric bone board.

The observers' osteometric data were compared to themselves for intraobserver error and to each other for interobserver error using ANOVA (significant variability $p < 0.05$; F crit > F) and Kappa coefficient (Kc) (significant difference < 0.5) statistical models. A Bland-Altman plot was used to provide an illustration of interobserver osteometric differences. We then compared anthroposcopic variables such as the calculated height, sex, ancestry, and trauma data between the observers.

Results

Intraobserver Reliability

The intraobserver error for each observer was very low (Table 2). Though the observers were novices, there was no statistical significance or variability in their measurements when compared to the self. The same measurements were also analyzed using a Kappa coefficient of the means (Table 3). Again, all observers had low variation when comparing the results to themselves.

Specimen	Observer	*F-calc*	*p-value*	*F-crit*	*df*	*Variance*
CHS 209	A	0.001	0.999	2.921	560	0.492
CHS 209	B	0.149	0.964	2.021	560	0.249
CHS 212	A	0.003	0.999	2.396	370	0.700
CHS 212	C	0.015	0.999	2.396	370	0.222
CHS 211	A	0.034	0.999	2.228	497	301.066
CHS 211	D	6.004E-06	1.000	2.117	497	0.087

Table 2: One-way ANOVA of intraobserver error for each observer; both *p-value* and *F-value* indicate the measurements did not vary significantly (set at 95% confidence). Observer A had a particularly high intraobserver variance when measuring CHS 211. All other variances were low

Specimen	Observer	*Kc*
CHS 209	A	0.85
CHS 209	B	0.93
CHS 212	A	0.92
CHS 212	C	0.97
CHS 211	A	0.75
CHS 211	D	0.98

Table 3: Kappa coefficient comparing mean measurements, allowing a difference of 1.0 mm. Intraobserver reliability is high (Kc > 0.7)

Interobserver Reliability

The interobserver reliability between observers varied with each statistical method. The one-way ANOVA results for CHS 209, CHS 211, and CHS 212 showed no significant variability between observers (Table 4). The ANOVA results indicate that there was no significant variation between observers but according to the Kappa coefficient, the observers had poor agreement (Table 5). CHS 209 Bland-Altman plot (Figure 1) shows six measurements located outside of the lower limits of agreement while the remaining measurements are within the levels of agreement, clustered around the mean but the levels of agreement are high. CHS 211 (Figure 2). Bland-Altman plot reveals agreement between observer with more measurement landmarks around the mean and lower levers of agreement compared to CHS 209 but are still high. There is one outlier in which Observer C omitted results of maximum length of a long bone but included all other measurements for that bone. CHS 212 (Figure 3). Bland-Altman plot has smaller limits of agreement, but the data is scattered with landmarks outside of the LOA.

Specimen	*F-calc*	*p-value*	*F-crit*	*Df*	*Variance*
CHS 209	0.001	1.000	1.759	1120	2581.560
CHS 212	0.008	0.999	1.893	740	2.097
CHS 211	0.050	0.999	1.759	994	269.919

Table 4: One-way ANOVA of interobserver error between observers; both *p-value* and *F-value* state the measurements have no significant differences between observers (set at 95% confidence). Variances between observers measuring CHS 209 and CHS 211 were high

Specimen	Kc
CHS 209	0.23
CHS 212	0.20
CHS 211	-0.34

Table 5: Kappa coefficient comparing mean measurements, allowing a difference of 1.0 mm. Interobserver reliability is low

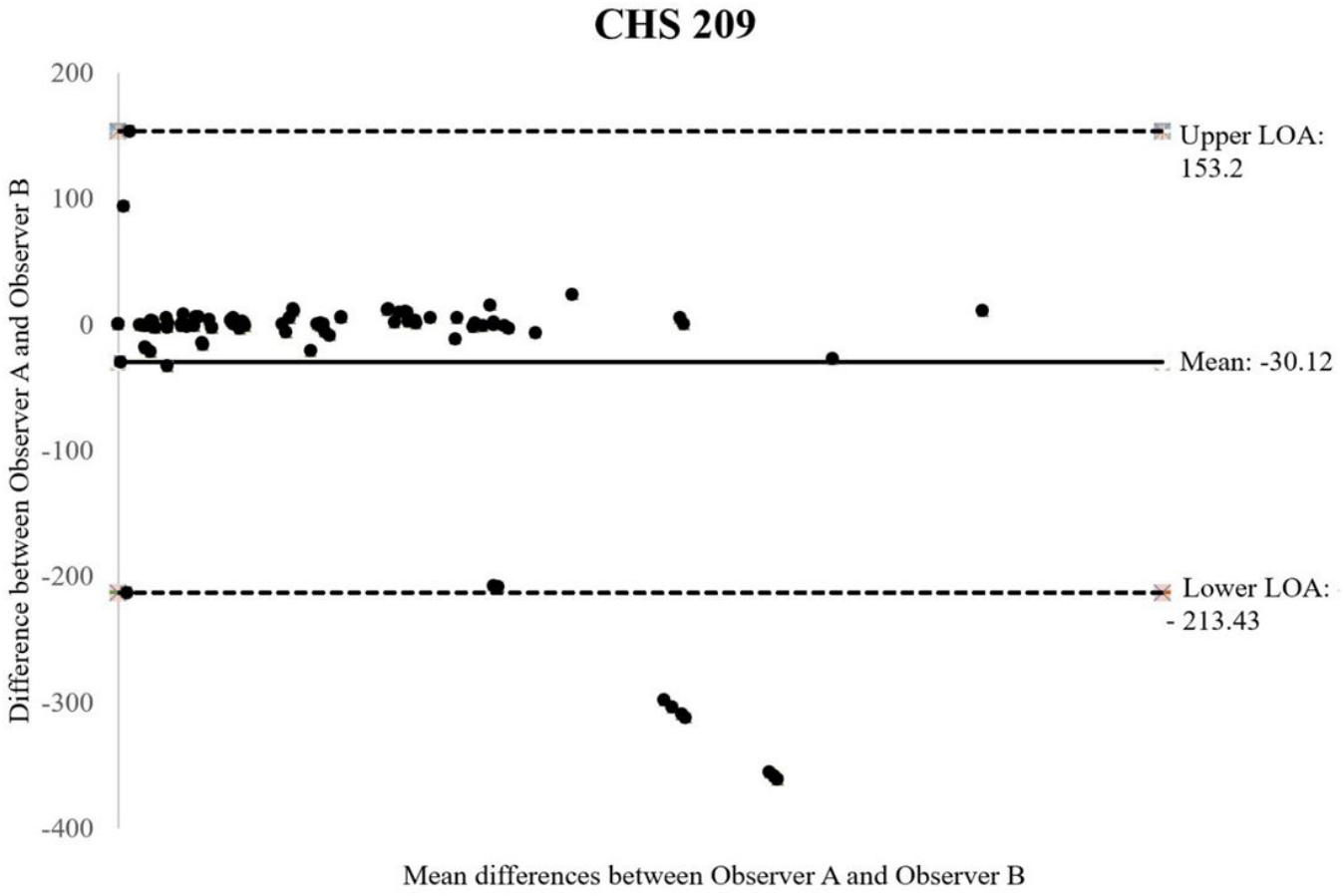

Figure 1: Bland-Altman plot on measurements between observers

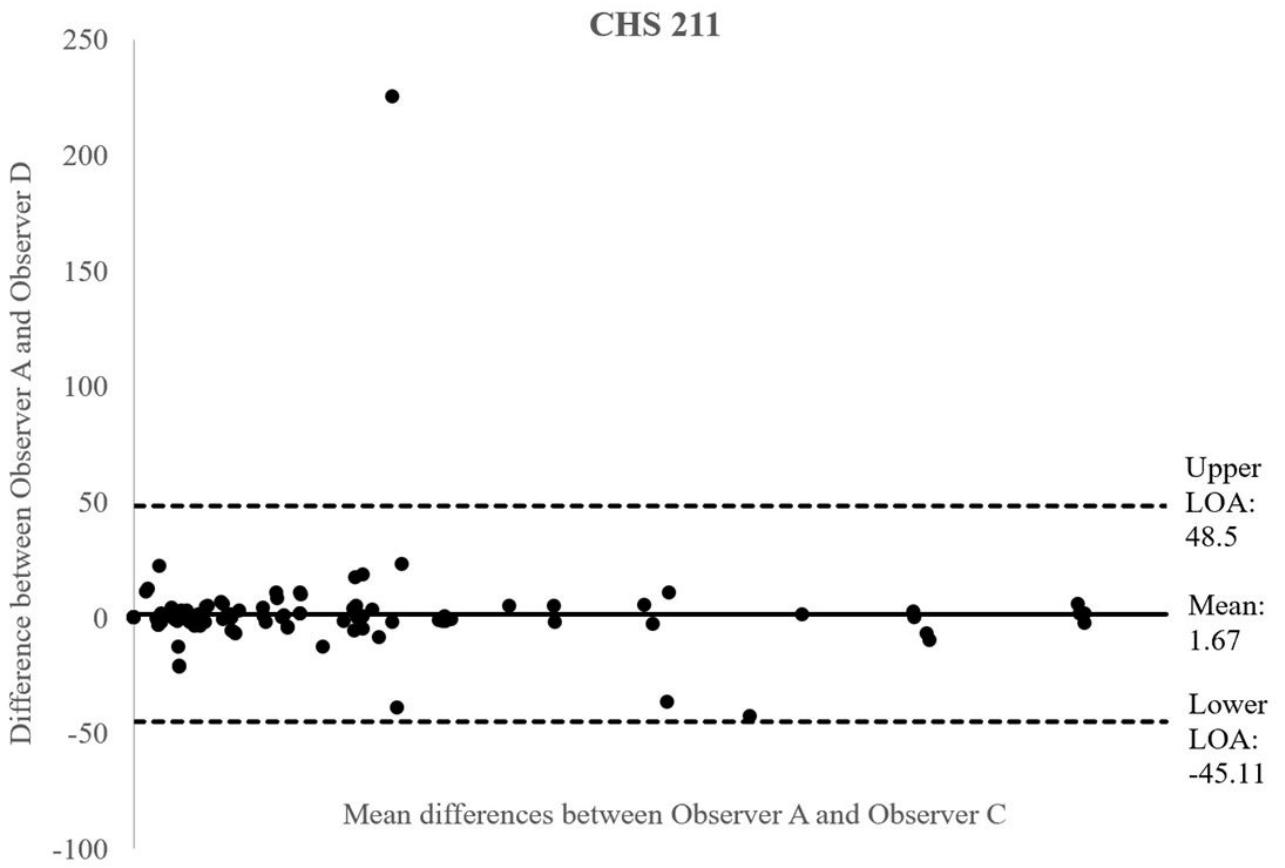

Figure 2: Bland-Altman plot on measurements between observers. The outlier is a maximum long bone length that Observer C omitted from measurement

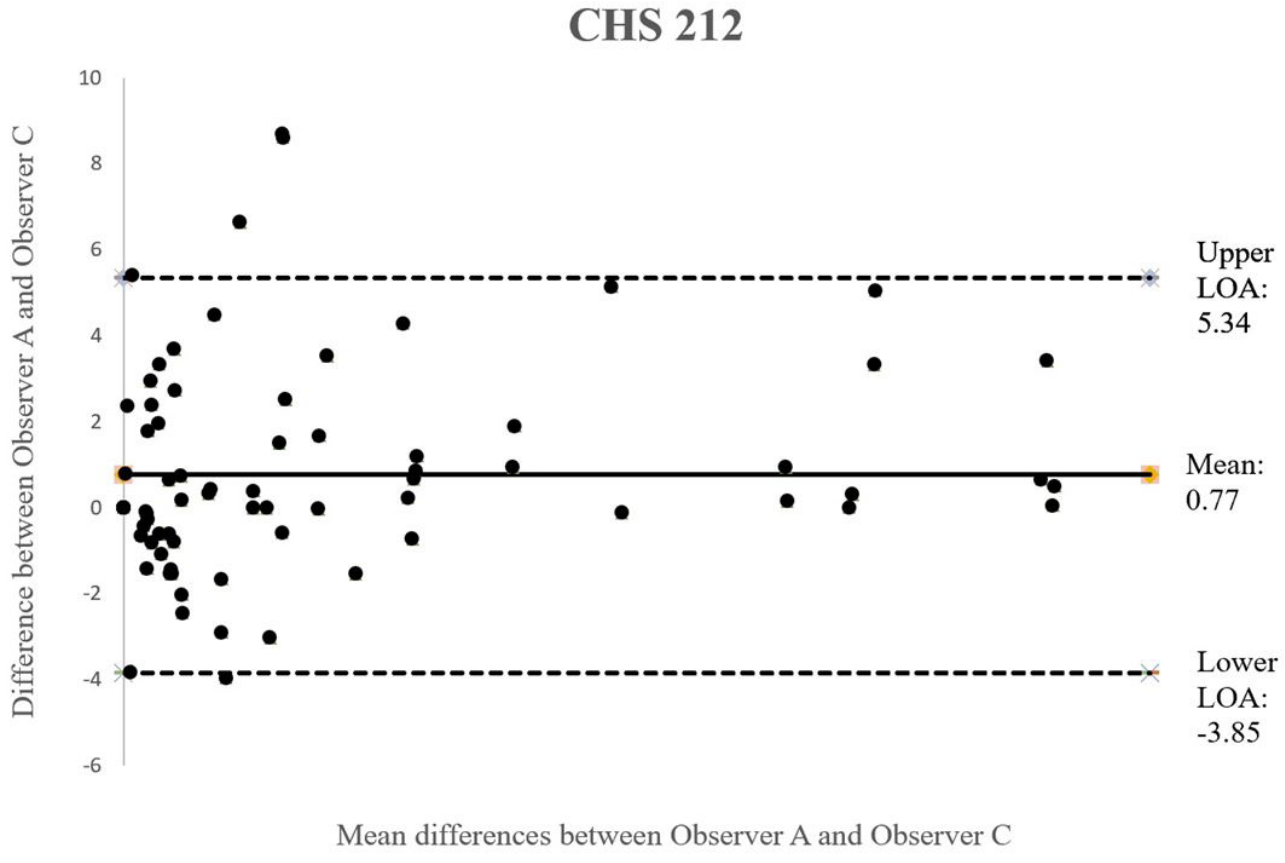

Figure 3: Bland-Altman plot on measurements between observers

Anthroposcopic Analysis

CHS 209: Observer A and B agreed on the sex and ancestry of the individual (Table 6). The relative age range and the height estimations were within the accepted standard deviations. The trauma analyses were relatively consistent with both observers describing the presence of scoliosis and differentiating between postmortem and antemortem skeletal modifications. Observer A did acknowledge more antemortem skeletal pathology than observer B.

	Observer A	Observer B
Height	Humerus: 156.75 cm +/- 4.45 cm Femur: 153.18 – 153.39 cm +/- 3.72 cm	Humerus: 153.39 cm +/- 4.45 cm Femur: 152.41 – 152.90 +/- 3.72 cm
Sex	Probable female	Probable female
Age	17-23	16-20
Ancestry	Asian	Asian

Table 6: Observer results for CHS 209

CHS 211: Observer A and D did not agree on the sex of the individual. Observer D could not determine the sex of the individual (Table 7). Ancestry and estimated height of the individual were consistent. The age range was inconsistent, with observer A believing the individual was over 30 and observer D stating the individual was below 30. The trauma analysis of observer D was superficial, observing antemortem scoliosis and postmortem modification. Again, observer A conducted a detailed antemortem skeletal analysis describing osteoarthritis and cervical fusion not mentioned by observer D.

	Observer A	Observer D
Height	Humerus: 155.91 - 156.08 cm +/- 4.45 cm Femur: 156.11 - 156.61 cm +/- 3.72 cm	Humerus: 156.66 - 156.93 cm +/- 4.45 cm Femur: 156.51 - 157.35 +/- 3.72 cm
Sex	Probable female	Ambiguous
Age	+30	20 - 30
Ancestry	Asian	Asian

Table 7: Observer results for CHS 211

CHS 212: Observer A and C did not agree on the sex, age, or height of the individual (Table 8). Ancestry estimation of the individual was consistent. Observer A and C exhibited some overlap of the estimated age of the individual but 16–35 years old is not an acceptable age range. Observer C acknowledged only postmortem modification. Observer A included antemortem trauma as well as postmortem skeletal modifications. CHS 212 had the largest source of morphological variation; one observer misidentified the individual.

	Observer A	Observer D
Height	Humerus: 155.75 cm +/- 4.25 cm Femur: 154.75–154.88 cm +/- 3.80 cm	Humerus: 161.01 cm +/- 4.05 cm Femur: 160.22 +/- 3.27 cm
Sex	Probable female	Probable male
Age	16-26	20-35
Ancestry	Asian	Asian

Table 8: Observer results for CHS 212

Discussion

The intraobserver error was low, and variability was not statistically significant. Hypothetically, where an observer interprets a skeletal landmark should not vary when conducting measurements. It is important to note that the faculty mentor was informed that observer D measured each landmark five consecutive times before moving on to the next location. This means there was no variation between measurements and retracts the validity of the intraobserver reliability. If an examiner makes repeated consecutive measurements over a brief amount of time, it may make those observations invalid as it eliminates potential error and thus could lead to misinterpretation and subsequent misidentification. This could impact the interobserver error between observer A and D.

The interobserver error data was quite surprising. The ANOVA results of the interobserver measurements did not indicate significant variability between results, yet the Kappa coefficient shows that there was poor agreement between observers. This lack of agreement can be most easily seen in the height analyses. Paradoxically, bone length may be thought to be easier to record than many other measurements, yet challenges were present and height discrepancies are not rare [7,12]. Height variation can be problematic for the identification of an individual and could lead to misidentification of unknown remains. The long bone equations for height estimations are different for males and females; this may explain why CHS 212's height discrepancy was so large. According to Klepinger, [12] height estimations with a 95 percent confidence interval can still have an error rate of 12–20 cm but be accurate enough 19/20 times.

Interobserver sex estimations were consistent between observer A and B but inconsistent between observer A and C, D. Sex estimation can be determined using the pelvis, skull, and long bones, but metabolic bone disease, age, and size of a population can affect sexing an individual. Sexing using the size of the bone can result in misidentification when the individual is from a population of smaller individuals (e.g., Asian populations are smaller than Caucasians) Klepinger, Langley *et al.* (2018) [7,12] found that observers also had inconsistencies when using various hard-to-determine pelvic landmarks. According to Klipinger (2006), [12] when sexing a skeleton, only 70% of males have a masculine skeletal structure and 78% of females have a feminine skeletal structure. The remaining individuals exhibit ambiguous skeletal traits which may account for the differences in sex between observer A and D. Observer D could have been overly cautious or unable to state whether the skeleton was more feminine. Observer A and observer C did not agree on their sex estimations. CHS 212 did not have a skull associated with the skeleton and sex would be determined using the pelvis. Upon professional observations, CHS 212 is a female based on the large sciatic notch.

The age of a skeleton affects an investigator's ability to accurately determine the sex. As the skeleton matures, secondary sex hormones affect the shape of various bones. From birth to adolescence, epiphyseal ossification sites are the primary landmarks used to estimate age. Dental age is not affected by the sex of an individual but is affected by individual health and genetics. There are some discrepancies between dental age and skeletal age [12]. Error in age estimations tends to go in all directions: young individuals are usually overestimated while older individuals are usually underestimated. Mid-adult skeletons are open to bias in either direction [12]. Observer A and B agreed on the age-range of CHS 209. Observer A and D may have agreed on the age of CHS 211. Observer A stated CHS 211 was 30 years old or older and observer D stated CHS 211 was between 20-30 years old. Using more skeletal landmarks may have corrected their discrepancy. Observer A and C had some agreement on the age range of CHS 212.

Misidentification or omission of metabolic bone disease was present in all observers' analyses. Observer A did mention osteoarthritis, but the other observers only noted scoliosis without indicating any other diseases or antemortem skeletal trauma. This is presumably due to the inexperience of the observers. If the observers had more experience with trauma and how it affected the skeleton, they would have been able to mention such findings in their reports. Even well-trained investigators may not be aware of the subtle macroscopic effects that specific diseases leave on bone [11,13].

Our research shows that relatively novice observers presented with discrepancies in their analyses that may lead to misidentification of skeletal remains, but the sample size is too small to draw any firm conclusions. Similar to findings from Langley *et al.* and Jamaiya *et al.* [7,14] observer discrepancies may be the result of landmark locating. Investigators may avoid identification discrepancies by using multiple observers, uniform standards, and compiling this data to assist in the skeletal analysis. Lynnerup *et al.* [10] believed their novices and experts were able to agree heavily due to the use of explicit and simple standards. A similar study by Langley *et al.* [7] analyzed expert interobserver error and found regular interobserver disagreement when the skeletal landmarks were difficult to locate. The largest inconsistency discovered by Langley *et al.* [7] was the way in which the examiners understood the definition of the landmark. Studies such as Ramsthaler *et al.* and Langley *et al.* [10,15] still showed expert observer disagreement. In contrast, Lynnerup *et al.* and Davis *et al.* [10,16] had similar disagreement between experts and non-experts. Inexperienced investigators may lack the specialization in analyzing skeletal remains, but other studies found with explicit direction error can be minimized [7,10]. Other mechanisms to avoid misidentification would include accounting for interobserver error.

Limitations of this Study

This study includes limitations such as the small skeletal sample. Additionally, all investigators were not able to examine all skeletons, although this would have been highly desirable. Due to these limitations we are unable to provide an explicit conclusion.

Conclusion

Intraobserver error should be considered and documented but it is not as beneficial as interobserver error. The observers in this study had a reliable intraobserver error but demonstrated potentially consequential variation between observers. The critical question is what threshold of training is adequate to reduce interobserver error to an acceptably low level? Having one osteological-focused course is insufficient training to conduct osteological analyses; substantial training is needed as well as interdepartmental collaboration [5]. Important questions remain: what determines expertise and how do we cultivate experts?

When investigators encounter skeletal remains it may be prudent to collaborate to aid in osteological analysis. The ongoing utilization of osteometric analysis, isotope analysis, Fordisc, and DNA analysis or skeletal remains is still relevant and beneficial [1,5,7,8,11]. Future research should compare an expert and novice to determine the acceptable course requirements to conduct accurate osteological analyses and identify unknown human remains.

Acknowledgement

We would like to acknowledge Dr. Greg Popovich and Ms. Amy Rogosky of West Virginia Wesleyan College and The West Virginia Division of Science and Research SURE grant.

References

1. Dirkmaat D, Cabo L, Ousley S, Symes S (2008) New Perspectives in Forensic Anthropology. Am J Phys Anthropol 51: 33-52.

2. Langley N, Jantz L, McNulty S, Maijanen H, Ousley S, et al. (2018) Error quantification of osteometric data in forensic anthropology. Forensic Sci Int 287: 183-9.

3. Cecchi R, Cipolloni L, Nobile M (1997) Incorrect identification of a military pilot with international implications. Int J Legal Med 110: 167-9.

4. Ross A, Juarez C, Urbanova P (2016) Chapter 14 - Complexity of Assessing Migrant Death Place of Origin. Biological Distance Analysis Forensic and Bioarchaeological Perspectives. Academic Press 265-83.

5. Crowder C, Wiersema J, Adams B, Austin D, Love J (2016) The Utility of Forensic Anthropology in the Medical Examiner's Office. Acad Forensic Pathol 6: 349-60.

6. Buikstra J, Ubelaker D (1994) Standards for data collections from human skeletal remains: proceedings of a seminar at the field museum of natural history, organized by Jonathan Haas In: Arkansas Archaeological Survey, Arkansas Archeological Survey Research Report, Fayetteville, Arkansas, USA.

7. Langley N, Jantz L, Ousley S, Jantz R, Milner G (2016) Data Collection Procedures for Forensic Skeletal Material 2.0 (3rd Edn) The University of Tennessee Department of Anthropology and Forensic Anthropology Center, Knoxville, USA.

8. Galera V, Ubelaer D, Hayek L (1995) Interobserver error in macroscopic methods of estimating age at death from the human skeleton. Int J Anthropology 10: 229-39.

9. White T, Folkens P (2005) The Human Bone Manual. Elsevier Academic Press, Amsterdam, Netherlands.

10. Lynnerup N, Belard E, Buch-Olsen K, Sejrsen B, Damgaard-Pedersen K (2008) Intra- and interobserver error of the Greulich-Pyle method as used on a Danish forensic sample. Forensic Sci Int 179: 242.e1-42.e6.

11. Milner G, Boldsen J (2017) Life not death: Epidemiology from skeletons. Int J Paleopathology 17: 26-39.

12. Klepinger L (2006) Fundamentals of Forensic Anthropology, A John Wiley & Sons, Inc, USA.

13. Brickely M, Ives R (2008) The Bioarchaeology of Metabolic Bone Disease (1st Edn) Elsevier Academic Press, Oxford, UK.

14. Jamaiyah H, Safiza M, Wong N, Kee, C, Rahmah R, et al. (2010) Reliability, technical error of measurements for children under two years old in Malaysia. Med J Malaysia 65: 131-7.

15. Ramsthaler F, Kreutz K, Zipp K, Verhoff M (2009) Dating skeletal remains with luminol-chemiluminescence. Validity, intra- and interobserver error. Forensic Sci Int 187: 47-50.

16. Davis C, Shuler K, Danforth M, Herndon K (2013) Patterns of interobserver error in the scoring of entheseal changes. Int J Osteoarchaeol 23: 147-51.