2015

Proposed method for predicting pair matching of skeletal elements allows too many false rejections

John Bowen
Central Washington University, john.bowen@cwu.edu

S. Vickers
Central Washington University

Patrick M. Lubinski
Central Washington University, pat.lubinsky@cwu.edu

L. Henebry-Deleon
Central Washington University

JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, January 2015, Vol. 60, No. 1
doi: 10.1111/1556-4029.12545
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

PHYSICAL ANTHROPOLOGY

Sara Vickers,[1] *M.S.; Patrick M. Lubinski,*[1,2] *Ph.D.; Lourdes Henebry DeLeon,*[1,2] *M.A.; and John T. Bowen,*[1,3] *Jr, Ph.D.*

Proposed Method for Predicting Pair Matching of Skeletal Elements Allows Too Many False Rejections

ABSTRACT: Byrd proposes a method for predicting pair matches in commingled remains to reduce visual comparison. The method compares differences between left and right postcranial element measurements in commingled samples with differences in known pairs from a reference sample using a *t*-score approach. We duplicated his protocol using six elements from two samples of known paired elements (n = 854 to 1063) and calculated the number of pairs correctly predicted. Time commitment was estimated by mathematically attempting matches with all left and right elements in these samples. Although the results show an 86% reduction in the number of potential pairs requiring visual matching, we do not recommend the method because (i) the normality assumption for use of a *t*-score approach is violated, (ii) no account is made for bilateral asymmetry, and (iii) the high rate of false rejections (up to 22%) undermines its ability to show true incompatibilities for potential pair matches.

KEYWORDS: forensic science, osteometric sorting, anthroposcopic, osteometric, human identification, commingling, skeletal identification, bilateral asymmetry

When human remains are commingled, in such situations as mass graves and natural disasters, separating elements into individuals is a time-consuming and delicate process. However, this is a process that must be completed to evaluate the remains for cause and manner of death, complete biological profiles (1), and return remains to descendants. The earliest detailed method for analyzing commingled remains was created by Charles Snow (2) and has continued as the primary protocol followed by physical and forensic anthropologists (1,3). Additional methods to sort commingled remains involve the process of elimination, and entail sorting the elements based on human skeletal variation (sex, chronological age, ancestry, pathology, and size), and taphonomic changes (3–10).

Anthroposcopics and anthropometrics are the usual and customary methods used to identify and separate individuals from a commingled situation. Application of anthropological field-accepted visual and metric methods, detailed by Buikstra and Ubelaker (6) in Standards for Data Collection from Human Skeletal Remains, are time intensive. Visual analysis of skeletal remains relies solely on the experience and expertise of the forensic anthropologist (7,11–13). The process of visual pair matching compares morphologic similarities between right and left sides of an element to verify an individual (14,15). Visual pair matching can be accurately performed by experienced physical anthropologists and osteologists if preservation of elements is adequate (14).

Although statistical models have been used for sex estimation, ancestry, and other characteristics for decades (16–19), models created to facilitate pair matching of commingled human remains are relatively recent (1,20). Using a statistical framework provides replicability, reliability, and removes the subjectivity inherent in analyses that are dependent on the examiner's capability to correctly identify human skeletal remains (20). The most comprehensive statistical study of commingled remains was completed by Byrd (1). This study led to an osteometric sorting method for pair matching of long bones using element measurements. The method starts with sums of a suite of 2–6 measurements on each of six long bones, taken from a reference sample of known pairs and also from the unknown, commingled sample elements. The measurement sums from each potential pair in the commingled sample are compared to provide a measurement difference *D*. This difference is then compared with the expected mean difference between sides of this element based upon the reference sample. If the commingled sample difference is significantly larger than the mean difference of the reference sample, the potential commingled pair match is rejected. These two specimens are then eliminated from the pool of possible matches that must be evaluated using visual methods in the second stage of pair matching, thereby saving time in analysis.

Byrd's (1) reference sample used in the method was a compilation of measurements from seven different collections totaling 376 individuals mostly of European, African, and Asian ancestry. The reference sample included individuals from the Central

[1]Resource Management Program, Central Washington University, Ellensburg, WA 98926-7544.
[2]Department of Anthropology, Central Washington University, Ellensburg, WA 98926-7544.
[3]Department of Geography, Central Washington University, Ellensburg, WA 98926-7420.

Received 22 May 2013; and in revised form 13 Sept. 2013; accepted 6 Oct. 2013.

Identification Laboratory, Hamann-Todd collection, Terry collection, Bass collection, Forensic Data Bank, International Commission on Missing Persons, and Peabody Museum. This sample intentionally included individuals of multiple ancestries and both sexes because many commingled collections have specimens of unknown ancestry and sex. Future users of the method were cautioned to ensure that reference samples are "representative of the same population as the case specimens in question" (1, p. 201).

Byrd (1, p. 203) states "this method has performed well in all test applications but could benefit from a larger sample." Tests of the method have yet to be published. The purpose of this paper is to determine the effectiveness of the osteometric sorting method, as recommended by its author, using two reference samples of individuals with known biological profiles.

Materials and Methods

Data used in this study were derived from osteological measurements reported from two datasets of postcranial remains, the Forensic Anthropology Data Bank (FDB) as of June 10, 2011, and the Goldman Osteological Data Set (GDS) as of October 29, 2012. Measurements in the FDB (21) were taken from remains in the Terry Collection at the Smithsonian's National Museum of Natural History, Hamann-Todd Collection at the Cleveland Museum of Natural History, as well as multiple smaller collections throughout the United States (1,22). The FDB dataset includes male and female remains from individuals in the United States of European, African, and Asian descent who died between 1892 and 1987 and is available by request from Dr. Richard Jantz (21). Measurements in the GDS (23,24) were taken from internationally curated collections housing remains of African, Asian, Australian, European, Native American, and Pacific Islander descent. The GDS dataset is mostly preindustrial, including male and female remains from individuals worldwide who died between 5500 years ago and 1900. It is available from Dr. Benjamin Auerbach online (24). Measurements in the FDB were recorded to the nearest mm and measurements in the GDS were recorded to the 0.5 mm (e.g., humerus maximum length) or 0.01 mm (e.g., humerus head diameter). Each element in both datasets has been identified to a known individual; therefore, this test does not include any unidentified remains that may skew the results. Additionally, all specimens in both datasets are fused adult elements free of any pathology that would affect measurements used in this test (R. Jantz personal communication; 2011; B. Auerbach personal communication; 2012). Measurements used here (Table 1, Fig. 1) are from the standards used by the FDB (22) and are the same as used in Byrd's method.

Byrd's protocol was replicated identically in this test, except that the same sample was used for both the reference data set and the unknown set. The rationale for using the same data for both the reference and the unknown was to minimize the opportunity for the protocol to perform poorly due to variation between chosen datasets. Using the same dataset as both the unknown and the reference should permit Byrd's method to perform optimally. The measurements in Table 1 were summed for each element in each sample. The difference (D) between the sums of the measurements for each element pair was calculated, subtracting each left side sum from its corresponding right side sum ($D = R_i - L_i$). Next, the standard deviation of D for each element was calculated (Table 2). Then, following Byrd, the difference between measurement sums for each potential element pair was calculated and tested against the null hypothesis of no difference (0) using a two-tailed t-distribution in the form $t = (D_i - 0/s_{ref})$ where D_i is the observed measurement sum difference for a potential element pair and s_{ref} is the standard deviation for the same element in the reference sample. As recommended by Byrd, a 0.10 significance level was used, so all element comparisons in which $p < 0.10$ resulted in rejection of the null hypothesis, meaning the measurements were too different to accept as an anatomical pair.

TABLE 1—*Measurements Taken for Each Element.*

Element	Measurement (Forensic Anthropology Data Bank measurement number)
Humerus	Maximum length (40), epicondylar breadth (41), maximum vertical diameter of the head (42)*
Radius	Maximum length (45), sagittal diameter at the midshaft (46)*, transverse diameter at the midshaft (47)*
Ulna	Maximum length (48), dorsovolar diameter (49), transverse diameter (50)
Femur	Maximum length (60), epicondylar breadth (62), maximum diameter of the head (63), anterior/posterior subtrochanteric diameter (64)†, transverse subtrochanteric diameter (65)†
Tibia	Condylo-malleolar length (69)†, maximum breadth at the proximal epiphysis (70), maximum breadth at the distal epiphysis (71)
Fibula	Maximum length (75), maximum diameter at midshaft (76)

*These three measurements are given slightly different names by the FDB and GDS but are the same measurements.

†These three measurements are given slightly different names by Byrd (1,4) and the FDB but are the same measurements.

Results

Table 3 shows percentages of FDB known pairs that were accepted or rejected as possible matches using the FDB sample as the reference set. The method falsely rejects 7–17% of known pairs in the sample. Table 4 shows the results for GDS known pairs using the GDS sample as the reference set; the method falsely rejects 14–22% for these two elements. For both of these tests, if each standard deviation was generated from the *absolute value* of sum difference ($D = |R_i - L_i|$) as implied by Byrd's example (1, Table 10.2), the method would perform more poorly, with false rejection rates of 19–31% for the FDB sample and 27–36% for the GDS sample.

The error rates for some of the FDB samples and both of the GDS samples are higher than expected, given that a 0.10 significance level was used. If the measurement sum difference were normally distributed and if the mean measurement sum difference were 0, a 0.10 significance level should have excluded no more than 10% of the valid pairs. In fact, neither condition is supported by the data. First, the distributions are not normal when evaluated with the Shapiro–Wilk test, probably due to both kurtosis and especially skewness (Table 5). The skewness indicates asymmetry in the data. There is statistically significant asymmetry ($p \leq 0.05$) in seven of the eight samples (all but the FDB fibula sample, for which $p = 0.25$). Such pronounced, statistically significant skewness is not unexpected due to the fact that humans are known to commonly exhibit bilateral asymmetry in long bone dimensions (23,25,26). As the data from both of these large samples are not normally distributed, it is likely that this problem will exist for other data sets. Second, given the expected asymmetry in long bones, the implicit assumption in Byrd's formula of no difference (i.e., the zero in the numerator, $D_i - 0$) in the measurement sums between left and right element pairs is unrealistic. As shown in Table 5, the mean measurement

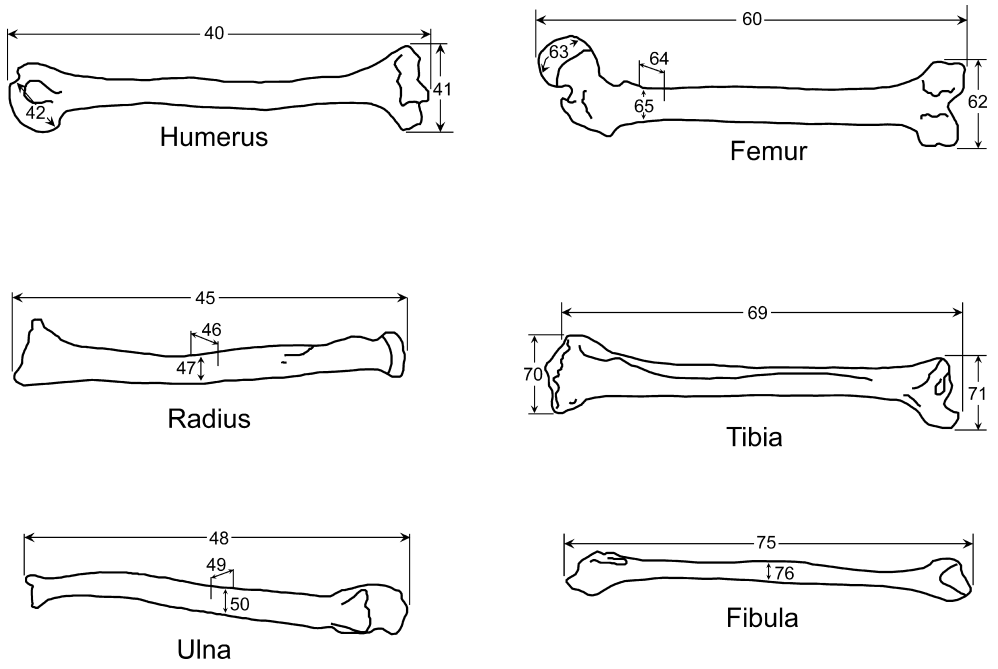

FIG. 1—*Long bone measurements employed in this study (after 6,22). The numbers refer to standardized Forensic Anthropology Data Bank measurements provided in Table 1.*

TABLE 2—*Byrd, Forensic Anthropology Data Bank (FDB), and Goldman Osteological Data Set (GDS) Sample Sizes and Standard Deviations.*

Element	Byrd Pairs*	Byrd *s**	FDB Pairs	FDB *s*	GDS Pairs	GDS *s*
Humerus	113	5.28	1063	4.59	1058	5.28
Radius	100	3.56	981	3.40	1004	3.58
Ulna	93	3.60	934	3.48	NA	NA
Femur	67	3.99	1001	4.42	NA	NA
Tibia	87	3.68	933	4.67	NA	NA
Fibula	71	2.99	855	3.67	NA	NA

*Pairs and standard deviation as provided by Byrd (1, p. 203).

TABLE 3—*Evaluation of Byrd's Method with Forensic Anthropology Data Bank Sample.*

	Humerus	Radius	Ulna	Femur	Tibia	Fibula
Sample Pairs	1063	981	934	1001	933	855
Predicted Paired	970	868	775	917	857	793
Correct, %	91	88	83	92	92	93
Predicted Not Paired	93	113	159	84	76	62
Incorrect, %	9	12	17	8	8	7

TABLE 4—*Evaluation of Byrd's Method with Goldman Osteological Data Set Sample.*

	Humerus	Radius
Sample Pairs	1058	1004
Predicted Paired	827	864
Correct, %	78	86
Predicted Not Paired	231	140
Incorrect, %	22	14

sum difference for most elements was positive and in some cases (e.g., GDS humerus) markedly so.

An alternative to Byrd's method still employing his measurement approach is simply to use the range of the absolute value of sum differences from a large, known reference source like the FDB or GDS sample (Table 6). If the difference between the measurement sums of a possible pair of commingled elements lies within the range of differences from the large reference sample, the pair can be considered a possible match. If not within the range, the pair is rejected as a possible match. Using the range is an uncomplicated mathematical process that employs osteometric measurements already gathered when completing a biological profile and should significantly reduce the number of false rejections. Naturally, a test of the FDB or GDS sample using the difference range generated from itself would result in perfect performance. However, a test of the GDS sample as an unknown using the smaller FDB ranges also shows a near-perfect performance (Table 7).

Although the range method provides a very low false-negative error rate, this is at the expense of time savings. To model time commitment, we used the FDB ranges with the GDS sample as if it were commingled, comparing every left with every right element to see how many possible pairs would be eliminated and how many would need to be checked visually. This exercise provides 1058×1058 (1,119,364) possible humerus pairs and 1004×1004 (1,008,016) possible radius pairs. The range method would eliminate the need to visually check 32.6% of these possible humerus matches and 31.5% of these possible radius matches. For comparison, Byrd's method (using the FDB standard deviations) would eliminate the need to visually check 85.9% of possible humerus matches and 85.5% of possible radius matches.

Discussion and Conclusions

No analytical technique is perfect. All potentially produce some number of false-positive or false-negative results. The best methods are those that provide a reasonable number of false results in light of the research goals. For pair matching of commingled remains, a compromise must be reached between the

TABLE 5—*Descriptive Statistics and Normality Tests for Difference Scores.*

	FDB Sample						GDS Sample	
Element	Humerus	Radius	Ulna	Femur	Tibia	Fibula	Humerus	Radius
Pairs (*n*)	1063	981	934	1001	933	855	1058	1004
Mean (mm)	1.56	1.84	2.06	−0.16	0.50	0.36	4.95	2.50
Kurtosis	8.22	8.82	4.52	1.19	7.43	0.99	6.93	5.41
Skewness	−0.27	−0.73	−0.73	−0.30	0.16	−0.10	−0.49	0.41
Shapiro–Wilk score (*w*)	0.9163	0.9232	0.9575	0.9850	0.9401	0.9866	0.9439	0.9600
Probability (*p*)	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

TABLE 6—*Measurement Difference Ranges in Forensic Anthropology Data Bank (FDB) and Goldman Osteological Data Set (GDS) Samples.*

	|FDB Range| (mm)	|GDS Range| (mm)
Humerus	0–31	0.01–41.70
Radius	0–23	0.01–30.56
Ulna	0–25	NA
Femur	0–24	NA
Tibia	0–36	NA
Fibula	0–17	NA

TABLE 7—*Evaluation of Forensic Anthropology Data Bank Ranges used on Goldman Osteological Data Set Sample.*

	Humerus	Radius
Sample Pairs	1058	1004
Predicted Paired	1056	1003
Correct, %	99.7	99.9
Predicted Not Paired	2	1
Incorrect, %	0.2	0.1

time-saving advantages of methods that reduce the number of possible matches to be examined and the accuracy needs for the goal of the analysis.

Any method used in forensic anthropology should be judged in light of current best practices in the field and the Daubert criteria. The Scientific Working Group for Forensic Anthropology (SWGANTH) provides best practice recommendations for resolving commingled remains, emphasizing the greater confidence gained by approaches that show incompatibilities (exclusions) over consistencies that do "not mean with certainty that they originated from the same person" (10, p. 3). The former allow segregation with confidence, whereas the latter are "not sufficient evidence for association" (10, p. 6). For the use of forensic data in court cases, the Daubert criteria hold that only testable, replicable, reliable, and scientifically valid methods are to be used to justify an expert's opinion (20,27).

The purpose of this study was to test Byrd's (1) proposed method on two samples of known individuals to determine whether this simple and replicable technique provided a reliable and time-efficient method to reduce the number of possible matches of elements that needed to be checked visually. In light of the SWGANTH best practices and Daubert criteria, there are three concerns about the method. First, the use of the *t*-score approach in the method is unwarranted since the right versus left long bone measurement distribution is not likely to be normal for case samples. This non-normal distribution is indicated in all of our test samples and is likely for other human populations given the prevalence of bilateral asymmetry in the limbs (23,25,26). Second, the method makes an unwarranted assumption that the expected difference between right and left measurement sums is zero, when in fact this is unlikely due to the same bilateral asymmetry. Third, the method permits the false rejection of up to 22% of true pairs, which undermines its ability to show true incompatibilities. The goal of the proposed method is to save time by reducing the number of possible matches that need to be checked visually, but if one cannot have a high degree of confidence in the rejections, then it is not helpful, as these would all need to be verified visually. If false rejections are taken as correct with no additional visual analysis, it might well result in an inability to re-associate individuals that could be critical to a case.

Use of an alternative like the measurement sum range would provide far fewer false negatives and thereby provide results in which analysts can have more confidence. However, this alternative also rejects fewer possible matches and so does not result in the same time savings as the originally proposed method. Additionally, the method is still sensitive to possible differences in ancestral population, and users should ensure they use a sufficiently diverse reference sample for their case specimens, as noted by Byrd (1) for the original method. Nonetheless, the alternative reduces the number of possible pairs that would need to be checked, while maintaining a higher degree a confidence than the original method. Visual analysis is not replaced but is still reduced with the range method.

Acknowledgments

Dr. Richard Jantz provided the Forensic Data Bank's postcranial data. Dr. Benjamin Auerbach provided the Goldman Osteological Data Set. Helpful comments were provided by two anonymous reviewers.

References

1. Byrd JE. Models and methods for osteometric sorting. In: Byrd JE, Adams BJ, editors. Recovery, analysis, and identification of commingled human remains. Totowa, NJ: Humana Press, 2008;199–220.
2. Snow CE. The identification of unknown war dead. Am J Phys Anthropol 1948;6:323–8.
3. Adams BJ, Konigsberg LW. Estimation of the most likely number of individuals from commingled human skeletal remains. Am J Phys Anthropol 2004;125:138–51.
4. Byrd JE, Adams BJ. Osteometric sorting of commingled human remains. J Forensic Sci 2003;48:717–24.
5. Ubelaker DH. Approaches to the study of commingling in human skeletal biology. In: Haglund W, Sorg M, editors. Advances in forensic taphonomy: method, theory, and archaeological perspectives. New York, NY: CRC Press, 2002;332–46.
6. Buikstra J, Ubelaker DH, editors. Standards for data collection from human skeletal remains. Proceedings of a seminar at the Field Museum of Natural History organized by Jonathan Haas. Fayetteville, AR: Arkansas Archeological Survey, 1994; Research Series Report No. 44.
7. Iscan M. Forensic anthropology of sex and body size. Forensic Sci Int 2005;147:107–12.

8. Garcia S. Is the circumference at the nutrient foramen of the tibia of value to sex determination on human osteological collections? Testing a new method. Int J Osteoarchaeol 2010;22:361–5.
9. Patil K, Mody R. Determination of sex by descriminant function analysis and stature by regression analysis: a lateral cephalometric study. Forensic Sci Int 2005;147:175–80.
10. http://swganth.startlogic.com/Commingling%20Rev2.pdf.
11. Bruzek J, Murail P. Methodology and reliability of sex determination from the skeleton. In: Schmitt A, Cunha E, Pinheiro J, editors. Forensic anthropology and medicine; complementary sciences from recovery to cause of death. Totowa, NJ: Humana Press, 2006;225–42.
12. Walrath DE, Turner P, Bruzek J. Reliability test of the visual assessment of cranial traits for sex determination. Am J Phys Anthropol 2004;125:132–7.
13. Del Alamo A. Assessment of a simple model and method for osteometric sorting. Proceedings of the American Academy of Forensic Sciences; 2010 Feb 22–27; Seattle, WA, Denver, CO: American Academy of Forensic Sciences, 2010;369.
14. Adams BJ, Byrd JE. Resolution of small-scale commingling: a case report from the Vietnam war. Forensic Sci Int 2006;156:63–9.
15. L'Abbe EN. A case of commingled remains from rural South Africa. Forensic Sci Int 2005;151:201–6.
16. Dibernnardo R, Taylor JV. Multiple discriminant function analysis of sex and race in the postcranial skeleton. Am J Phys Anthropol 1983;61:305–14.
17. Dittrick J, Suchey JM. Sex determination of prehistoric central California skeletal remains using discriminant analysis of the femur and humerus. Am J Phys Anthropol 1986;70:3–9.
18. Meadows L, Jantz RL. Estimation of stature from metacarpal lengths. J Forensic Sci 1992;37:147–54.
19. Safont S, Malgosa A, Subira ME. Sex assessment on the basis of long bone circumference. Am J Phys Anthropol 2000;113:317–28.
20. Dirkmaat DC, Cabo LL, Ousley SD, Symes SA. New perspectives in forensic anthropology. Am J Phys Anthropol 2008;137(Suppl 47):33–52.
21. http://fac.utk.edu/databank.html.
22. Moore-Jansen PH, Jantz RL, Ousley SD. Data collection procedures for forensic skeletal material. Knoxville, TN: Forensic Anthropology Center, Department of Anthropology, University of Tennessee, 1994.
23. Auerbach BM, Ruff CB. Limb bone bilateral asymmetry: variability and commonality among modern humans. J Hum Evol 2006;50:203–18.
24. http://web.utk.edu/~auerbach/GOLD.htm.
25. Plochocki JH. Bilateral variation in limb articular surface dimensions. Am J Hum Biol 2004;16:328–33.
26. Schultz AH. Proportions, variability and asymmetries of the long bones of the limbs and clavicles in man and apes. Hum Biol 1937;9:281–328.
27. Christiansen AM, Crowder CM. Evidentiary standards for forensic anthropology. J Forensic Sci 2009;54:1211–6.

Additional information and reprint requests:
Patrick Lubinski, Ph.D.
Department of Anthropology, Central Washington University
400 E. University Way
Ellensburg, WA 98926-7544
E-mail: lubinski@cwu.edu