J Forensic Sci, July 2003, Vol. 48, No. 4
Paper ID JFS2002189_484
Available online at: www.astm.org

John E. Byrd,[1] *Ph.D. and Bradley J. Adams,*[1] *Ph.D.*

Osteometric Sorting of Commingled Human Remains

ABSTRACT: This paper describes the method of osteometric sorting. Osteometric sorting is the formal use of size and shape to sort bones from one another. The method relies heavily on measurement data and statistical models and is designed to maximize objectivity. The goal of this paper is to validate the use of osteometric sorting and to provide case examples of its utility. Selected regression models are also presented for use with osteometric sorting. We advocate this technique as one tool among many in the anthropologist's toolkit for sorting commingled skeletal remains.

KEYWORDS: forensic science, forensic anthropology, commingling, human osteology, osteometric sorting, regression analysis, CILHI

The methods of forensic anthropology are most effective when applied to the complete skeleton of an individual. Age estimates rely on combined data from different points in the skeleton, such as the teeth and various long bone epiphyses, which must be synthesized and interpreted as indicative of a specific age range. Population affinity (e.g., race) is best determined through reference to a complete cranium and mandible, but additional support can be obtained through observation of the postcranial skeleton. As part of their analysis, forensic anthropologists are often confronted with the problem of commingled remains, where the remains of multiple individuals are part of a single assemblage with no readily apparent indication as to which bone specimens belong to which individual. This condition obstructs the personal identification process by frustrating attempts to synthesize data from multiple elements and, until resolved, may preclude the release of remains to next of kin for final disposition.

Nowhere has the problem of commingled human skeletal remains been more aggressively addressed than at the U.S. Army Central Identification Laboratories (CILs). The CILs have been in charge with identifying the remains of the nation's war dead for extended periods of time following World War II, the Korean War, and the Vietnam War. Naturally, these cases often included commingled remains from aircraft crashes and mass graves. Charles Snow, who worked at the CIL in Hawaii in 1947, published a paper summarizing the work being done in the Pacific CIL during his one-year tenure (1). Approximately 25% of Snow's paper was dedicated to the topic of sorting commingled remains. Snow devised a logical series of steps for sorting commingled remains that are as relevant today as they were then, and these guidelines serve as the basic framework of the methods we advocate. Some of the steps advocated by Snow use the size and shape of bones as a basis for segregating them into individuals.

[1] Forensic Anthropologist, US Army Central Identification Laboratory, Hawaii.

Received 29 May 2002; and in revised form 13 Jan. and 9 Feb. 2003; accepted 8 Mar. 2003; published 22 May 2003.

The Method of Osteometric Sorting

For sorting commingled remains, Snow took advantage of the fact that the human physique varies in predictable ways. Experienced osteologists have long recognized that an individual with long, slender femora should have long, slender humeri. Conversely, short and thick femora are found with short and thick humeri. Bones can be correctly segregated using this principle so long as the variation within the assemblage is marked and the number of commingled individuals is small. But how large must the size variation be in order for it to be accurately recognized? How much confidence can be placed in the results? To what extent is accuracy idiosyncratic to the anthropologist? Should greater emphasis be placed on the size or the shape of the bones? Answers to these questions are not readily forthcoming. What is needed is a method that incorporates objective measurements and formal arguments. Osteometric sorting has these qualities, and it is a powerful addition to the package of gross methods that are available to anthropologists for sorting commingled remains.

Osteometric sorting uses measurements on bone specimens as the basis for comparison. The measurement data from respective case specimens are simultaneously compared with reference data and with one another. Segregation decisions are made by explicitly testing the null hypothesis that two specimens, given their size and shape, could have belonged to a single individual. With its reliance on bone measurements and statistical hypothesis testing, this approach virtually eliminates subjectivity. We recommend that osteometric sorting be used as one step in the process of sorting commingled skeletal remains. Other methods, such as articulation of adjacent bones and pair-matching, are often more effective when they can be properly applied.

Previous Studies

Osteometric sorting has been attempted on a limited basis in previous studies (2–5). Buikstra et al. (2) reported the results of an osteometric sorting study designed to evaluate the likelihood that two corresponding cervical vertebrae originated in the same individual. The aim was to be able to formally test the null hypothesis of "congruence" in the size and/or shape of the corresponding vertebrae. A

Copyright © 2003 by ASTM International, 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428-2959.

series of measurements were taken on cervical vertebrae in Terry Collection skeletons at the Smithsonian Institution. The variable used in the statistical test was derived by subtracting a given measurement value of a vertebra from the equivalent measurement value of the corresponding vertebra in the caudal direction. This value was then formally compared with the Terry sample mean for the variable by way of a t-test. While Buikstra et al.'s results from osteometric sorting did not reverse the conclusions drawn from more traditional analysis (i.e., evaluating the fit of articulating elements), they did provide a more objective means of demonstrating how poor the congruence was in one case, and that the vertebrae sizes were well within the expected size range in a second case. Thus, they were able to support expert opinions with hard data and formal test results.

London and colleagues (3,4) have experimented with osteometric sorting, as reported in two presented papers. This research has concentrated on associating the femoral head with the acetabulum. The original study, involving sample data collected from 100 individuals from the Maxwell Museum, University of New Mexico, found a significant correlation between the femur head diameter and measurements of the acetabulum. While London and Curran (3) indicate success in their osteometric sorting, no details were available as to how the correlations were exploited in the sorting protocol. Later, London and Hunt (4) revised the acetabulum measurements to be applicable to the Smithsonian Institution's Terry Collection with its many individuals exhibiting arthritic lipping on the acetabulum. In an experimental application to commingled remains of the Huntington Collection, they found that the use of osteometric sorting in tandem with visual sorting improved their ability to resolve the commingling.

Rösing and Pischtschan (5) reported an experiment with osteometric sorting applied to archaeological samples. The sample data included 16 measurements taken on 32 individuals from archaeological contexts (sample individuals were not commingled). The measurements included long bone lengths and circumferences along with two skull measurements. Correlation coefficients were calculated. A 98% confidence ellipse was calculated for a bivariate model comparing the radius and ulna since it showed the highest correlation ($r = 0.963$). Five pairs of specimens were plotted against the model and its associated ellipse along with all possible pairings of specimens including mismatched pairs. The assessment of success in the method was made by the closeness of a true match to the regression model line, as opposed to mismatches, which according to their method should be relatively further from the centroid. Because they found that mismatched specimens were often closer to the model line than true matches, they concluded that mathematical models do not offer much hope for sorting commingled remains. Part of the blame for the lack of success was attributed to the reliance on measurement data. Measurements, they argue, provide a "harsh reduction of the available information" inherent in bone specimens. Rösing and Pischtschan conclude that re-individualization of commingled bones is best done by subjective assessment.

The study by Rösing and Pischtschan should be examined in more detail since their conclusions suggest osteometric sorting offers little to the process of sorting commingled remains. Their experiment can be criticized on several grounds. First, the sample size in their study ($N = 32$) was too small to support more than a pilot study. The most serious problems with the Rösing and Pischtschan study, however, relate to their statistical procedures. Recall that Buikstra et al. (2) advocated the use of a formal statistical hypothesis test to determine the strength of congruence between two specimens proposed as a match. Rösing and Pischtschan used a stricter criterion of the closeness of the plotted measurements to the regression model line. By this method, the specimen closest to the "average" (i.e., the regression model line), using the size of the second specimen as the independent variable value, is the match despite the possibility that other close matches are present as well. This criterion is unrealistic in that it ignores the reality of human variation. Variation in bone size within the skeleton is broad enough that most true matches will not lie on the regression model line; thus, variation must be anticipated in any new method reliant upon bone measurements.

The authors included a confidence ellipse around the bivariate centroid in their example, but did not explain its use, if any, nor justify their selection of 98% as the confidence level. Confidence ellipses are typically used to represent the bivariate sampling distributions of sample centroids and to facilitate comparison of the bivariate centroids of two or more samples (6). A superior approach to comparing bone specimens (as opposed to samples) is to construct a prediction interval (7,8) on the regression model and to test the hypothesis that one specimen matches a second specimen given their measurement values. The differences in the approaches go well beyond splitting statistical hairs. The prediction interval is the sampling distribution of a single predicted value (i.e., the predicted value of a bone measurement, given the size of the bone it is compared with) when the true population model parameters must be estimated using a sample of reference data (8). Prediction intervals differ from confidence ellipses in their geometry: a prediction interval is a hyperbola whose upper and lower boundaries are *closest* near the sample centroid, while the boundaries of the confidence ellipse bulge in the vicinity of the centroid. The confidence ellipse has boundaries that intersect the regression model line, so that large bones and small bones will be excluded from the distribution. These differences relate to the fact that the confidence ellipse represents the sampling distribution of the centroid, not single predicted values. The prediction interval becomes increasingly broad at growing distances from the centroid along the regression model line, which reflects the fact that we have less confidence in the accuracy of the regression model as we move away from the centroid (due to sampling error at the upper and lower ends of the data distribution). An important ramification of these differences is that osteometric sorting using the prediction interval will have its greatest discriminatory power when case specimen measurement values are relatively close to the sample centroid.

The poor results of their study led Rösing and Pischtschan to conclude that anthropologists should rely upon subjective judgment to sort commingled skeletal remains. They note (p. 40) that morphological sorting operates on the basis of "broad personal experience" and is "sufficiently successful" so long as "the number of commingled skeletons is not too high." No evidence is provided to support their claim of success in this subjective approach. Nor do they indicate how many remains are too many. While pair-matching of the same element has been formally evaluated (9), it is unlikely that any studies have been conducted to evaluate error rates in morphological sorting—as noted above, these approaches do not lend themselves to the determination of error rates.

The Reference Sample

The method of osteometric sorting requires comparison with a large reference sample with numerous measurements that represent aspects of both size and shape of skeletal elements. We view the development of reference data as an on-going project, whereby the

sample size and the composition of the reference sample is improved on a continual basis. This approach follows the philosophy of the data banking concept (10). The reference sample used in this study consists of data collected from the Central Identification Laboratory, Hawaii (CILHI) cases, various anatomical collections, and the Forensic Data Bank (FDB) (see 10).

The list of measurements taken in support of this study originally included over 140 different observations. The core measurements are derived from the Forensic Data Bank list (11), which includes standard osteometric measurements with published definitions familiar to most forensic anthropologists. It is the unfortunate reality that most standard measurements were defined such that they must be taken on complete bones, as with the maximum lengths and with mid-shaft diameters. For this reason, the standard set was supplemented by new measurements, largely defined by the authors, designed to be applicable to fragmented bone specimens. Appendix 1 provides an abridged list of only the measurements used in this paper, along with their formal measurement definitions. The measurement numbers are consistent with the FDB numbers with the exception of the new measurements. New measurements are given a number and letter combination so that measurements on the same element will cluster together in the list, while not disrupting the original FDB measurement numbering scheme.

The measurements were taken on American Whites, American Blacks, and Asians. Both sexes are represented. The majority of the White males with the complete set of measurements (FDB and new measurements developed for fragmentary remains) were from CILHI cases. These individuals were military personnel identified by the CILHI over the past few years. CILHI cases are an ideal source of data (provided preservation is good) since these were healthy individuals of known race, age, and stature at their time of death. The statures are measured statures. Other data were obtained through visits to the Terry Collection of the Smithsonian Institution, the Hamann-Todd Collection of the Cleveland Museum of Natural History, and the Bass Collection of the University of Tennessee. Finally, postcranial measurement data from the FDB were made available to the authors by Dr. Richard Jantz. The reference sample composition is summarized in Table 1.

We have made the attempt to develop a high quality reference data sample that is generally applicable in forensic anthropology casework in the United States and beyond. Quality control measures were taken during the data gathering process. For example, CILHI cases were measured only when preservation was good enough that the measurement values would not be altered by taphonomic factors. We avoided individuals in the anatomical collections who died after prolonged periods of illness since there was potential for extreme atrophy in those skeletons. In all cases, measurements of traumatized or pathological areas of bones were not taken. The FDB data were carefully scanned for outliers. Outliers that could be attributed to data entry or measurement errors were corrected when possible (see 12) and deleted when clearly in error and uncorrectable. Interobserver error in the data collected by the authors was controlled by periodically repeating measurements to ensure that both osteologists were getting the same values for the given specimens. While there was no means of directly addressing interobserver variation in the FDB, the results of a study by Adams and Byrd (12) suggest that it is minimal for most measurements.

TABLE 1—*Reference sample broken down by collection, race, and sex.*

Collection	Sex	Black	White	Asian	Total
CILHI	F	0	1	0	1
	M	5	42	4	51
CMNH-HT	F	2	2	0	4
	M	7	7	0	14
SI-TERRY	F	14	10	0	24
	M	14	2	0	16
UT-BASS	F	3	9	0	12
	M	4	7	0	11
FDB	F	12	46	0	58
	M	17	108	0	125
Total		78	234	4	316

CILHI, US Army Central Identification Laboratory, Hawaii.
CMNH-HT, Cleveland Museum of Natural History Hamann-Todd collection.
SI-TERRY, Smithsonian Institution Terry collection.
UT-BASS, University of Tennessee Bass collection.

Osteometric Sorting Procedures

The bases for osteometric sorting are the relationships that exist among bone sizes as represented by measurements. Large humeri, for example, are associated with large femora and the strength of the association can be measured. Trotter and Gleser (13) reported correlation coefficients for the numerous long bone measurements incorporated into their classic paper on stature estimation. Trotter and Gleser (13:486) found that strong correlations exist among the long bone lengths, most ranging between 0.80 and 0.98. Their results suggest that size relationships among long bones are strong enough to support size-based sorting. Correlation coefficients have been calculated for selected measurements. Our results, too extensive to be reported here, are consistent with those of Trotter and Gleser in that they show high correlations among the long bone length measurements. The reference data include considerably more measurements than used by Trotter and Gleser and permit the examination of relationships of bone lengths with diameters, and diameter measurements with one another. It is clear that some relationships are stronger than others, such that length measurements show considerably stronger correlations than do breadth measurements (including diameters).

The statistical approach to osteometric sorting advocated here is to test the null hypothesis that two bone specimens are of sizes consistent with having originated from the same individual. The bivariate statistical models, calculated from the reference sample data, serve as the basis for testing the hypothesis. For example, a measurement can be selected from each of two specimens to be compared. An ordinary least squares regression model, with associated prediction interval, is calculated with one of the measurements as the independent variable and the other as the dependent variable. We wish to point out that there are numerous valid statistical methods other than regression that could be applied to this problem. These include, but are not limited to, using simple bone measurement ratios (13–15), reduced major axis regression (14,16), principal components analysis (6), and canonical correlation (6). Objections to the use of least squares regression models can include the seemingly arbitrary selection of independent versus dependent variables in the models, and the need for assumptions regarding the error variance in the independent variables (16). Though we advocate the use of models set up in multiple directions (i.e., humerus on femur or femur on humerus), it is important to note that the direction is not arbitrary but determined by the question asked, which is determined by the circumstances of the case. This is analogous to regressing stature on bone lengths (7,14) because one wishes to estimate stature from a case specimen. We fol-

low Trotter and Gleser (13) in comparing bone sizes with the use of ordinary least squares regression, and point out, following Jungers (17), that when the correlations among the variables are high and the goal of the analysis is prediction, least squares regression is an appropriate method.

The decision as to which of the two specimens is to be the independent variable is always determined by the circumstances of the specific case. An example would be where one has begun with a partial skeleton consisting of a lower body, and must determine whether or not an isolated humerus (or multiple humeri) could have originated from the same individual. Here, the size of the femur is a logical independent variable that is used to predict the size of a humerus originating from the same body. We typically start with lower bodies in our sorting regimen, following the systematic steps advocated by Snow (1). The case specimen measurement value (independent variable) is entered into the regression model formula to produce a predicted value for the other bone measurement. If the actual measurement value of the second bone specimen falls within the prediction interval surrounding the predicted value, then the null hypothesis is accepted. Only where the null hypothesis is rejected do we sort bones into separate individuals. The method supports a sound argument that, where the null hypothesis is rejected, it is unlikely that the two specimens originated from the same individual. However, *the reverse is not necessarily true*. In most instances, failure to reject the null hypothesis is not sufficient evidence to conclude that the elements are from the same individual. This argument must be supported by independent evidence.

For osteometric sorting, the hypothesis tests do not need to be based upon a single bone measurement from each specimen. Principal components analysis of the reference data consistently shows that size accounts for the majority of variation in the data. We have experimented with a variety of ways to combine multiple measurements from a single element into a single variable, including the use of allometric coefficients from principal components analysis as measurement weights prior to the summation of the measurement values. Improvement of the statistical model characteristics was used as the criterion for evaluation of the respective approaches. It was found that a simple summation of the available measurements on a bone element provides an effective variable for use in a bivariate model. Summation of multiple measurements leads to significantly higher correlations between bone sizes, especially where two or more diameter measurements are combined into a single variable (many of the length measurements already show high correlations with little room for improvement). The linear combinations we have used as variables in our analyses, whether using bone lengths or not, have shown correlation coefficients of 0.80 or higher (see test applications below). Aside from the marked improvement in the statistical models, it appears that the use of multiple measurements has the advantage of incorporating more information (such as pertaining to shape) into one model. It is also possible to combine measurements from multiple elements into a single variable if warranted (i.e., in situations when the elements are known to originate from the same individual based on articulation or other means). Canonical correlation analysis (6) is a more sophisticated way of combining multiple measurements into a single variable. This procedure weights the measurement values so that the overall correlation between two sets of measurements is maximized, and arguably makes more use of shape characteristics than the simple linear combination we advocate here. However, applications of canonical correlation to our data have revealed that any improvement is negligible (e.g., there is not a substantial improvement in the use of shape) and the additional computational complexity is not justified.

Data transformations can improve the characteristics of the regression models used in the hypothesis tests. A common transformation is the logarithmic transformation of measurement values (16). Logarithmic transformations were an essential step in the classic allometric studies (16,18,19) because many growth relationships among organs of the body are non-linear. The allometric models become linear following transformation and often show improved homoscedasticity (see 16). Other transformations are possible as well. In this study we experimented with numerous transformations and settled on the natural logarithm as a reasonable choice. We transform the summed measurement values from an element into the natural logarithm of the variable value. Note that some bias must be corrected when transforming model estimates back to raw numbers (see 20).

The following procedures provide a basic summary of our method. We wish to stress that the particulars of the case will determine the form of the hypothesis to be tested, such that if one is beginning with lower body portions and attempting to associate (or segregate) an upper body bone to the individual, then the size of the lower body portions (or a selected bone such as the femur) should serve as the independent variable. Next, the appropriate measurements are taken. The measurement values for each element are summed. The summed total is then converted to a natural logarithm for each element (the logarithmic value should be expressed with a minimum of two decimal places). Next, the desired regression model is calculated from the reference data (numerous regression formulae are presented below), and the second element is regressed on the first. The value obtained for the second case specimen is compared to the desired prediction interval obtained from the regression model. Giles and Klepinger (7) detail the procedure for calculating a prediction interval based on linear regression from summary statistics. We provide the necessary statistics throughout this paper for the calculation of predication intervals by the reader. If the value falls outside of the prediction interval, then the null hypothesis is rejected and the specimens are sorted. Figure 1 shows a graphic example of the application of the method to known individuals using a 90% prediction interval. In this example, two humeri, known to originate from different individuals, were compared with two femora, also known to originate from different individuals. Thus, beginning with two lower bodies represented by the femora, we simultaneously tested four hypotheses regarding the possible association of each humerus to the respective femora. Presentation of the results as a plot (Fig. 1) provides simple interpretation: we must reject the null hypotheses for humerus

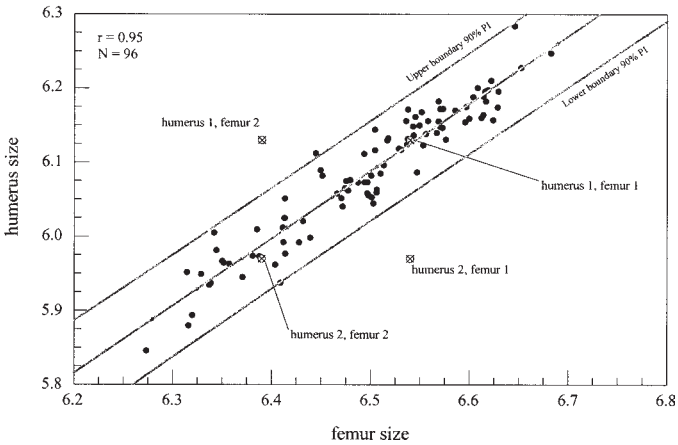

FIG. 1—*Plot showing segregation based on osteometric sorting.*

1 with femur 2 and for humerus 2 with femur 1. If these remains were commingled, we would sort them, leaving humerus 1 with femur 1 and humerus 2 with femur 2. This evidence alone does not prove the latter associations, but provides a sound, objective argument for the sorting. Applications of the method to cases of known individuals follow.

Test Applications of Osteometric Sorting

The method of osteometric sorting has been applied to artificially commingled sets of remains and actual case examples. Through these tests it is possible to evaluate the overall effectiveness of the approach and to verify that the error rates are at the expected levels or lower. In all test applications, the data are from known individuals. Note that the same two specimens are never compared twice in a test application. We utilize the 90% prediction interval in all of the tests. In the following examples, the method is applied as an essentially stand-alone procedure though we do not recommend its normal use in this manner. As stated previously, osteometric sorting is incorporated as one step among many in the process of sorting commingled remains. Although numerous test cases were performed using osteometric sorting, only two are presented in this paper that are illustrative of the method. While not exhaustive, the regression models presented in the following examples can be applied to cases so long as the same core measurements are used to construct the variables.

Test Application 1

The first test application (Test 1) consists of partial remains of six individuals as listed in Table 2. These data were taken from a combination of four CILHI cases and two Terry Collection skeletons. Test 1 can be considered an ideal situation for osteometric sorting in that there are a variety of body sizes in the group and it is assumed that the represented bones are complete. Table 3 provides the variables as defined for the test and Table 4 contains the regression model statistics used in the hypothesis tests. Calculations were performed using the SAS PROC REG routine (21). Table 5 illustrates how the results are read and shows the hypothesis test results by regression model for a set of comparisons involving the radius and humerus.

These results allow us to quantify the performance of the method. In total, there were 113 comparisons of case specimens made. Of these, 58% ($n = 65$) led to rejection of the null hypotheses in situations where the specimens originated from different individuals. Another 17% ($n = 19$) of the comparisons were from the same individual and had the null hypotheses accepted. Type I errors, where comparisons involving specimens from the same individual had the null hypotheses rejected, occurred in 3% ($n = 3$) of the comparisons. The remaining 22% ($n = 26$) of the comparisons involved specimens from different individuals, but the null hypotheses were accepted. Regarding the last circumstance, we would like to remind the reader that this method does not claim to

TABLE 2—*Test 1 sample and relevant skeletal elements.*

Individual	Year of Death	Race	Sex	Age	Stature (cm)	Elements
A	1934	Black	F	32	156	Left ulna, right radius, left humerus
B	1925	Black	F	24	171	Left humerus
C	1943	White	M	20	178	Left humerus, right ulna, left femur, left tibia
D	1941	White	M	17	177	Right humerus, left radius, left femur, right tibia
E	1994	Asian (Korean)	F	40	157	Left humerus, right radius, left ulna, left femur, right tibia
F	1943	White	M	22	182	Right ulna, right femur

TABLE 3—*Size measurements used as part of Test 1 including their means and standard deviations (STD). The variable is calculated as LN(a+b+c . . .).*

Variable	Element	Measurements	Mean	STD
TIB	TIBIA	69, 71, 72, 73, 74A, 74B	6.32	0.09
FEM	FEMUR	60, 62, 63, 64, 65, 68A, 68B	6.56	0.08
ULN*	ULNA	49, 50, 51A, 51C	4.20	0.11
RAD	RADIUS	45, 47A, 47B, 47C, 47D	5.77	0.09
HUM	HUMERUS	40, 41, 41A, 42, 42A, 44B	6.31	0.08

* Includes no length measurement.

TABLE 4—*Regression model statistics from Test 1.*

Model	*N*	*r*	Root MSE	*p*
TIB=1.08(FEM) − 0.78	103	0.96	0.03	0.0001
TIB=0.65(ULN) + 3.60	95	0.80	0.06	0.0001
TIB=0.96(RAD) + 0.77	95	0.94	0.04	0.0001
TIB=1.09(HUM) − 0.54	94	0.93	0.04	0.0001
FEM=0.59(ULN) + 4.08	93	0.82	0.05	0.0001
FEM=0.84(RAD) + 1.74	94	0.91	0.04	0.0001
FEM=1.0(HUM) + 0.28	93	0.96	0.03	0.0001
ULN=1.03(RAD) − 1.78	97	0.84	0.07	0.0001
ULN=1.23(HUM) − 3.58	94	0.89	0.06	0.0001
RAD=1.04(HUM) − 0.81	94	0.93	0.04	0.0001

TABLE 5—*Results from Test 1. The top row includes the specimen used as the independent variable, the predicted value (of dependent variable) from the regression model, and the 90% PI. The left column includes the specimens that are compared with the top row and their respective values.*

Element	HUM A 5.62 5.58–5.67	HUM B 5.69 5.65–5.74	HUM C 5.79 5.73–5.82	HUM D 5.79 5.73–5.82	HUM E 5.63 5.56–5.66
RAD A 5.68	Reject*!	Accept	Reject	Reject	Reject
RAD D 5.78	Reject	Reject	Accept	Accept*	Reject
RAD E 5.65	Accept	Accept	Reject	Reject	Accept*

* Specimens are from the same individual.
! Type I error committed.

sort individuals of the same general size, nor is acceptance of the null hypothesis sufficient proof of an association. An error rate of 3% (Type I errors) is encouraging as it is lower than the 10% expected when using a 90% prediction interval.

Test Application 2

The second test (Test 2) uses data obtained from a CILHI case involving 8 individuals lost in an aircraft crash at the close of World War II. The commingled remains were recovered by a CILHI team in 1999. Table 6 summarizes the test assemblage. There was extensive fragmentation of the remains due to the violence of the crash, therefore long bone lengths were not included in the variables (Table 7). The regression models are reported in Table 8. Note the relatively lower correlation coefficients for the models incorporating fewer measurements and lacking bone lengths. Of the 286 comparisons made, 30% ($n = 87$) led to successful rejections of the null hypotheses. Null hypotheses were accepted where bones were from the same individual in 13% ($n = 38$) of comparisons. Type I errors were committed in 2% ($n = 5$) of comparisons and the remaining 55% were null hypotheses that were accepted even though the specimens originated in different individuals. Thus, despite the fragmented nature of the remains, slightly under one-third of the specimens could be successfully sorted using osteometrics alone. As a validation of the technique in this case, the osteometric sorting was subsequently confirmed using mtDNA sequence data.

The error rates in these tests are surprisingly low given the use of a 90% prediction interval. Additional test applications were performed (not detailed here) and similar results were found. In total, 636 comparisons were made using known individuals and the greatest error rate in any one test was 5%, and the overall error rate was under 3%. Trotter and Gleser (13) encountered a similar phenomenon when they tested their regression models against an independent dataset. While they should have found approximately 66% of the test subjects' statures within one standard error, there were 79% in this range for the model utilizing the femur and tibia. On the other hand, their humerus model performed worse than expected at 62% (13). It is likely that the error rates in this study are smaller than expected due to noise in the reference data, possibly resulting from slight interobserver variation, data entry errors, and other problems common to pooled osteometric data (12). Note that all measurements in the test applications were taken by the authors.

In most cases, the errors in the test applications appear to be the result of normal human variation. However, some concerns have emerged from careful study of the specimens in error. Individual D in Test 1 was a 17 year-old male whose long bone epiphyses had yet to close at the time of death. The girth measurements for this individual, particularly the humerus and femur head diameters, were unusually large relative to the lengths. Caution is in order when applying the regression models, intended for the adult population, to sub adults. Handedness was possibly a cause of error in some instances where only girth measurements were used, but this does not appear to be a significant factor overall in osteometric sorting. We did not find that race or sex was a significant factor in causing error.

Discussion

Osteometric sorting shows great potential as an addition to existing procedures for sorting commingled remains. The advantages include: 1) the method is inexpensive to apply, 2) it yields results in a short period of time, 3) it has low error rates, 4) it has considerable power when applied to individuals of varying size, and 5) the statistics are simple and well-grounded in anthropology. There are disadvantages to using the method as well. These include: 1) its low power when applied to individuals of the same general size, 2) its uselessness when the measurements cannot be taken, as due to poor preservation, and 3) the effects of secular trends, handedness, race, and sex have not been formally explored. As to the latter concern, this study has found no evidence that these factors adversely affect the method. The lack of a noticeable demographic or temporal effect is possibly due to the inclusion in the reference data sample individuals of multiple races, sexes, and decades of death. Refinement of the regression models in the future, tailored to specific components of the population, could lead to greater power. In the

TABLE 7—*Size measurements used by element for Test 2 including their means and standard deviations (STD). The variable is calculated as LN(a+b+c . . .).*

Variable	Element	Measurements	Mean	STD
HUM*	HUMERUS	44B	2.88	0.13
RAD*	RADIUS	47A, 47B, 47C	3.85	0.11
ULN*	ULNA	49, 50, 51A, 51C	4.20	0.11
FEM*	FEMUR	64, 65, 68A, 68B, 68E	4.97	0.10
TIB*	TIBIA	72, 73, 74A, 74B	4.78	0.11

* Includes no length measurements. The fragmented nature of the remains precluded the use of additional measurements.

TABLE 8—*Regression model statistics for Test 2.*

Model	*N*	*r*	Root MSE	*p*
HUM=1.08(RAD) − 1.27	112	0.89	0.06	0.0001
HUM=1.04(ULN) − 1.47	99	0.89	0.06	0.0001
HUM=1.18(FEM) − 2.98	102	0.82	0.08	0.0001
HUM=1.02(TIB) + 1.97	110	0.86	0.07	0.0001
RAD=0.84(ULN) + 0.34	101	0.90	0.04	0.0001
RAD=0.96(FEM) − 0.96	101	0.84	0.06	0.0001
RAD=0.81(TIB) − 0.02	108	0.82	0.07	0.0001
ULN=1.02(FEM) − 0.87	99	0.81	0.07	0.0001
ULN=0.85(TIB) + 0.11	99	0.84	0.07	0.0001
FEM=0.74(TIB) + 1.45	105	0.89	0.05	0.0001

TABLE 6—*The assemblage for Test 2.*

Individual	Year of Death	Race	Sex	Age	Stature (cm)	Elements
A	1945	White	M	adult	180	Left humerus, right radius, right ulna, left femur, left tibia
B	1945	White	M	adult	168	Left humerus, left radius, left ulna, left femur, left tibia
C	1945	White	M	adult	175	Right humerus, right radius, left ulna, left femur, right tibia
D	1945	White	M	adult	185	Right humerus, left femur, left tibia
E	1945	White	M	adult	184	Left humerus, left femur
F	1945	White	M	adult	170	Left humerus, left radius, left ulna, left femur
G	1945	White	M	adult	185	Right humerus, right radius, right femur, left tibia

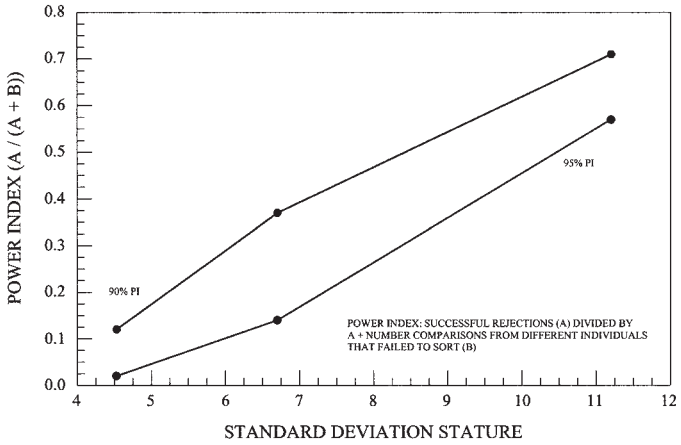

FIG. 2—*Power index showing the relationship of stature (standard deviation) and osteometric sorting.*

interim we follow the recommendations of Feldesman and Fountain (15) and developed generic models.

The test applications bring to light a number of important points regarding osteometric sorting. Measurements must be taken properly, or the statistical models are invalid. Poor bone preservation (i.e., cortical exfoliation) will preclude the attainment of accurate data. Poorly preserved specimens should be excluded from metric analysis. Measurements that are believed to be "abnormal," such as those of bone parts that are diseased, traumatized, or damaged postmortem, should be excluded.

Some statements can be made concerning the power of the method and the rate of error. The actual error rates in case applications relate directly to the size of the prediction interval and appear to be considerably smaller than is predicted by the models, though we cannot guarantee error rates lower than 10% in future applications. We believe that some tightening of the interval is justifiable and recommended. How tight is at the discretion of the analyst. The 90% PI appears to offer acceptable results given a desire to keep error rates modest.

There is a relationship between the power of the method and the variation in the overall sizes of the individuals represented in the assemblage. The *power index* of the method is defined as:

$$A / (A + B)$$

where A is the number of comparisons involving bones from different individuals that lead to the successful rejection of the null hypothesis and B is the number of comparisons involving bones from different individuals where the null hypothesis must be accepted. The power index values vary from a low of 0 in the case of no successful sorts and a high of 1.0 in the case where every bone originating in a separate individual is successfully sorted. The power index in the tests appears to relate strongly to the standard deviation of the statures of the included individuals (see Fig. 2).

The power of the method is also related to the size of the prediction interval, such that there is an increase in power as the prediction interval is narrowed (see Fig. 2 above). Thus, one must balance power against error rates when selecting the appropriate test criteria (e.g., prediction interval).

Osteometric sorting is a relatively simple, effective method for sorting commingled remains when it is used in conjunction with other methods. We have found that the power of the method varies with the size variation of the individuals being sorted, but the error rates are consistently low. The method is most effective when applied to an assemblage in conjunction with other methods.

Acknowledgments

This research could not have been completed without support from Tom Holland at the CILHI, Dave Hunt at the Smithsonian Institution, Bruce Latimer and Lyman Jellico at the Cleveland Museum, and Lee Jantz at the University of Tennessee, Knoxville. We would also like to acknowledge the individuals who read early drafts of this paper and provided comments. These include Richard Jantz, Eugene Giles, Tom Holland, Steve Ousley, Dave Hunt, Mark Leney, Chris McDermott, and Paul Emanovsky. We would also like to thank the anonymous reviewers for their comments, which have improved the paper considerably.

References

1. Snow CE. The identification of the unknown war dead. Am J Phys Anthropol 1948;6:323–8.
2. Buikstra JE, Gordon CC, St. Hoyme L. The case of the severed skull: individuation in forensic anthropology. In: Rathbun TA, Buikstra JE, editors. Human identification: case studies in forensic anthropology. Springfield: C.C. Thomas, 1984.
3. London MR, Curran BK. The use of the hip joint in the separation of commingled remains (abstract). Am J Phys Anthropol 1986;69:231.
4. London MR, Hunt DR. Morphometric segregation of commingled remains using the femoral head and acetabulum (abstract). Am J Phys Anthropol 1998;26 (suppl.):152.
5. Rösing FW, Pischtschan E. Re-individualisation of commingled skeletal remains. In: Jacob B, Bonte W, editors. Advances in Forensic Sciences. Berlin: Verlag, 1995.
6. Tatsuoka MM. Multivariate analysis: techniques for educational and psychological research. New York: Macmillan Publishing Company, 1988.
7. Giles E, Klepinger LL. Confidence intervals for estimates based on linear regression in forensic anthropology. J Forensic Sci 1988;33(5):1218–22.
8. Neter J, Wasserman W. Applied linear statistical models: regression, analysis of variance, and experimental designs. Homewood: Richard D. Irwin, Inc., 1974.
9. Adams BJ. The Use of the Lincoln/Petersen index for quantification and interpretation of commingled human remains [Unpublished Master's Thesis]. Knoxville: University of Tennessee; 1996.
10. Jantz RL, Moore-Jansen PH. A forensic data base for forensic anthropology: structure, content, and analysis. Report of investigations No. 47. Knoxville: University of Tennessee, Department of Anthropology; 1988.
11. Moore-Jansen PM, Ousley SD, Jantz RL. Data collection procedures for forensic skeletal material. Report of investigations no. 48. Knoxville: University of Tennessee, Department of Anthropology; 1994.
12. Adams BJ, Byrd JE. Interobserver variation of selected postcranial measurements. J Forensic Sci 2002;47(6):1193–202.
13. Trotter M, Gleser GC. Estimation of stature from long bones of American whites and Negroes. Am J Phys Anthropol 1952;10:463–514.
14. Konigsberg LW, Hens SM, Jantz LM, Jungers WL. Stature estimation and calibration: Bayesian and maximum likelihood perspectives in physical anthropology. Yearbook Phys Anthropol 1998;41:65–92.
15. Feldesman MR, Fountain RL. "Race" specificity and the femur/stature ratio. Am J Phys Anthropol 1996;100:207–24.
16. Sokal RB, Rohlf FJ. Biometry: The principles and practice of statistics in biological research. San Francisco: W.H. Freeman and Company, 1969.
17. Jungers WL. Aspects of size and scaling in primate biology with special reference to the locomotor skeleton. Yearbook Phys Anthropol 1984;27:73–97.
18. Huxley JS. Problems of relative growth. New York: Dover Publications, Inc., 1972.
19. Gould SJ. Allometry and size in ontogeny and phylogeny. Biol Rev 1966;41:587–640.
20. Smith RJ. Logarithmic transformation bias in allometry. Am J Phys Anthropol 1993;90:215–28.
21. SAS. SAS/STAT user's guide, 4th ed., Volume 2. Cary: SAS Institute Inc., 1990.

Additional information–reprints not available from authors:
John Byrd
US Army CILHI
310 Worchester Ave.
Hickam AFB, HI 96853
E-mail: byrdj@cilhi.army.mil

APPENDIX 1

Abridged list of post-cranial measurements used in study. Note that the numbering scheme is designed to correspond with the Forensic Data Bank.

Measurement	Definition
Humerus	
40. Max. length	*
41. Epicondylar br.	*
41A. Capitulum-trochlea br.	The breadth of the capitulum and trochlea at the distal humerus. One end of the sliding calipers is positioned parallel to the flat, spool-shaped surface of the trochlea, and the other end is moved until it comes into contact with the capitulum.
42. Max. vertical diam., head	*
42A. A-P br., head	The maximum breadth of the humeral head taken in the anterior-posterior direction on the articular surface. This measurement is taken perpendicular to the vertical diameter of the humeral head.
44B. Min. diam. diaph.	The minimum diameter of the humeral diaphysis taken in any direction perpendicular to the shaft. This measurement should be taken on the oval part of the shaft, superior to the flattening observed around the olecranon fossa and the lateral supercondylar ridge. Often it is found near midshaft.
Radius	
45. Max. length	*
47A. Max. diam. radial tub.	The maximum shaft diameter on the radial tuberosity. Position the calipers around the tuberosity and rotate the bone until the maximum distance is obtained.
47B. Max. diam. diaph. distal to radial tub.	The maximum shaft diameter distal to the radial tuberosity, *positioned along the interosseous crest*. The bone should be rotated to find the maximum distance.
47C. Min. diam. diaph. distal to radial tub.	The minimum shaft diameter anywhere distal to the radial tuberosity. The bone may be rotated to find the minimum distance.
47D. Max. diam. radial head	Position the calipers around the radial head and rotate the bone until the maximum distance is obtained.
Ulna	
49. Dorso-volar diam.	*
50. Transverse diam.	*
51A. Min. diam. along interosseus crest	Locate the minimum diameter of the diaphysis along the portion of the bone that includes the interosseous crest. This measurement may not necessarily include the interosseous crest, but should be taken on that part of the shaft that exhibits the crest.
51B. Min. diam.	This measurement will be found near the distal epiphysis of the ulna. The bone should be rotated in order to locate the minimum distance.
51C. Br. distal end of semi-lunar notch	This is a measure of only the distal surface of the semilunar notch (the base). In order to obtain the distance, one end of the calipers is positioned within the radial notch (approximate midpoint), roughly parallel to the shaft. The other end of the calipers is applied to the medial edge of the semilunar notch.
Femur	
60. Max. length	*
62. Epicondylar br.	*
63. Max. diam. head	*
64. A-P subtrochlear diam.	*
65. Trans. subtrochlear diam.	*
68A. Min. A-P diam. diaph.	The minimum anterior-posterior diameter anywhere along the diaphysis. The linea aspera should be utilized in order to orient the bone.
68B. Min. M-L diam. diaph.	The minimum medial-lateral diameter anywhere along the diaphysis. The linea aspera should be utilized in order to orient the bone.
68E. Max. diam. along linea aspera	The maximum shaft diameter at any point along the linea aspera. As the bone should be rotated to obtain the maximum distance, the measurement does not necessarily have to include the linea aspera.
Tibia	
69. Max. length	*
71. Max. br. distal epiphysis	*
72. Max. diam. nut. foramen	*
73. Trans. diam. nut. foramen	*
74A. Max. A-P diam. distal to popliteal line	This measurement should be taken at the most distal point of the popliteal line. Note that the correct location may be difficult to determine in very gracile individuals.
74B. Min. A-P diam. distal to popliteal line	Locate the smallest anterior-posterior distance at any point on the tibial shaft.

* See (Ref. 11) for measurement description.