Archaeol Anthropol Sci (2016) 8:689–704
DOI 10.1007/s12520-015-0251-6

ORIGINAL PAPER

Estimating body mass from postcranial variables: an evaluation of current equations using a large known-mass sample of modern humans

Marina Elliott[1,2] · **Helen Kurki**[3] · **Darlene A. Weston**[4,5] · **Mark Collard**[1,6]

Received: 22 December 2014 / Accepted: 12 May 2015 / Published online: 29 May 2015

Abstract Many inferences in palaeoanthropology and bioarchaeology rely on estimates of body mass from skeletal material. Body mass estimation is also becoming an area of interest for forensic anthropologists. The most common approach to estimating body mass from the skeleton involves measurements of the postcranium, and a number of equations have been developed for femoral head size and stature plus bi-iliac breadth. These equations have become standard in biological anthropology, but they have rarely been tested on individuals of known mass. In addition, the effects of several assumptions involved in the application of the equations have not been rigorously investigated. Accordingly, this study employed CT scans from a sample of 253 adult modern humans of known body mass to test the accuracy of the most widely used postcranial body mass estimation equations. The results were then used to evaluate several claims concerning the performance of the equations relative to one another. Most of the equations that were tested met the criteria for acceptance as reliable estimators with the male and the combined-sex samples. However, females were not estimated as reliably. In addition, the equations did not always perform consistently or as expected. Overall, our results suggest that estimating body mass with the postcranial equations that are currently available requires more caution than is usually exercised.

Keywords Osteology · Biological anthropology · Paleoanthropology · Fossil hominin · Forensic anthropology

Electronic supplementary material The online version of this article (doi:10.1007/s12520-015-0251-6) contains supplementary material, which is available to authorized users.

✉ Mark Collard
mcollard@sfu.ca

[1] Human Evolutionary Studies Program and Department of Archaeology, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada

[2] Evolutionary Studies Institute, University of the Witwatersrand, Johannesburg, South Africa 2008

[3] Department of Anthropology, University of Victoria, Victoria, BC V8W 2Y2, Canada

[4] Department of Anthropology, University of British Columbia, Vancouver, BC V6T 1Z1, Canada

[5] Department of Human Evolution, Max Planck Institute of Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

[6] Department of Archaeology, University of Aberdeen, St Mary's Building, Elphinstone Road, Aberdeen AB24 3UF, UK

Introduction

Estimating body mass from skeletal remains is an important aspect of biological anthropology research. In palaeoanthropological and bioarchaeological contexts, body mass estimates offer one of the few ways to access key biological and behavioral information (Ruff 2002; Plavcan 2012). Estimates of mass are also often required to determine the ecological and evolutionary significance of differences between individuals living at different times and/or in different places (Smith 1996). In addition, because body mass is a conspicuous individualizing feature and a potentially significant influence on taphonomic processes, it is becoming an area of interest for modern forensic identification and multiple fatality investigations (Rainwater et al. 2007; Agostini and Ross 2011; Moore and Schaefer 2011; Byard 2012).

Body mass is most commonly estimated from the postcranium. Currently, there are two broad approaches: the "mechanical" approach and the "morphometric" approach

(Ruff 2002; Auerbach and Ruff 2004). Mechanical methods rely on the functional relationship between body mass and the skeletal elements that bear weight, such as the femur (Ruff et al. 1991; McHenry 1992; Grine et al. 1995). In contrast, morphometric methods estimate mass by reconstructing overall body shape from measures of stature and bi-iliac breadth (Ruff 1994; Ruff et al. 1997, 2005). Although interspecies analyses have been used to estimate mass in fossil hominins (e.g., Steudel 1980; McHenry 1992; Hartwig-Scherer 1993), the most commonly used methods are based on what Konigsberg et al. (1998) call the "inverse regression" of body mass on skeletal dimensions, and employ reference populations of modern humans, either individuals within a population (Ruff et al. 1991, 2012) or sex-specific means for multiple groups (McHenry 1992; Ruff 1994; Grine et al. 1995; Ruff et al. 2005).

Use of regression equations of this type to estimate body mass has become standard in biological anthropology. They have been employed to estimate the mass of numerous fossil hominin groups (Ruff et al. 1997; Churchill et al. 2012), individual hominin specimens (Ruff and Walker 1993; Ruff 1994; Arsuaga et al. 1999; Trinkaus and Jelinek 1997; Rightmire 2004; Rosenberg et al. 2006; Ruff et al. 2006; Melton et al. 2010; Ruff 2010; Walker et al. 2011), and archaeological *Homo sapiens* populations (Kurki et al. 2010; Myszka et al. 2012; Pomeroy and Stock 2012; Ruff et al. 2012). The equations have also been used to explore body mass estimation as a tool in modern forensic contexts (Lorkiewicz-Muszyńska et al. 2013).

Despite their widespread use, several questions about the equations remain unanswered. Most importantly, it is not clear how accurate they are. Currently, belief in the accuracy of the equations is based largely on the similarity of their results when compared to each other (e.g., Auerbach and Ruff 2004; Kurki et al. 2010; Pomeroy and Stock 2012). This approach is not unreasonable, but it suffers from the obvious problem that the similarities demonstrate only that the equations yield congruent results, not that they are accurate. The accuracy of some of the equations has been tested but there are reasons to be concerned about the findings of both of the relevant studies (Ruff et al. 1991; Lorkiewicz-Muszyńska et al. 2013). Ruff et al. (1991) not only used indirect measures for certain key variables but also relied on patient recall to generate values for body mass, which has obvious shortcomings. Moreover, the sample used to test the equations was small, comprising just eight individuals. Lorkiewicz-Muszyńska et al.'s (2013) study was better controlled but failed to explore a number of important results. For example, the authors reported "significant inaccuracy" (pg. 405) in prediction competence in over- and underweight groups but did not investigate this. A further problem concerning the accuracy of the equations is that new ones have been published since the Lorkiewicz-Muszyńska et al. (2013) study went to press (Ruff et al. 2012), and these have not been tested. At the moment, then, the accuracy of the currently available equations is uncertain.

There are at least three other unanswered questions concerning the use of the equations. One is whether or not the biomechanical method is better than the morphometric method. Most researchers assume that the mechanical method is more accurate because the femur bears the majority of the body's weight (Jungers 1988; Aiello and Wood 1994; Churchill et al. 2012). However, the morphometric method has been argued to be more reliable than the mechanical method because it encompasses greater geographic diversity and uses larger sample sizes (Auerbach and Ruff 2004). In fact, body masses estimated using the morphometric method are now being used as "true masses" in studies deriving new predictive equations from archaeological material (Ruff et al. 2012). For this practice to be appropriate, however, the reliability of the morphometric method must be explicitly demonstrated.

Another question relates to the choice of specific equation for the target specimen. Some authors have argued that variations in body size and proportion necessitate specific mechanical equations for smaller or larger bodied groups (McHenry 1992; Grine et al. 1995; Kurki et al. 2010). If a specimen does not fit into one of these "target" groups, the use of "generalized" equations has been recommended (Ruff et al. 1991, 2012). Alternatively, Auerbach and Ruff (2004) recommend averaging the results of multiple mechanical equations for non-specific specimens. However, it is not always possible to know how well a specimen matches a reference sample (Smith 2009). In addition, there is some question as to whether averaging estimates is statistically appropriate (Smith 2002; SWGANTH 2012). Either way, establishing the extent to which such claims are valid is an important task. In a similar vein, Ruff et al. (2005:390) argued that, when choosing which morphometric equation to use, their new equations should be preferred because they are based on a larger sample than previous equations and therefore are "more broadly applicable, particularly to tall and wide-bodied males" (Ruff 1994; Ruff et al. 1997). As a result, the new equations have largely replaced earlier ones in practice (Ruff et al. 2012; Lorkiewicz-Muszyńska et al. 2013). However, the new equations were designed specifically to increase the representation of high-latitude populations in the sample (Ruff et al. 2005). Consequently, applying them to other test groups may not be appropriate and the assumption of their superiority to previous equations requires evaluation.

Lastly, the advantage of using sex-specific equations over combined-sex equations has not been clearly established. Of the most commonly cited studies, two provide both sex-specific and combined-sex equations (Ruff et al. 1991, 2012). Two others provide only sex-specific equations (Ruff et al. 1997, 2005) and two provide only combined-sex

equations (McHenry 1992; Grine et al. 1995). In general, sex-specific equations are considered to be more accurate than combined-sex equations because of systematic differences in body size between males and females (Ruff et al. 1991). However, sex is not always easy to attribute and the level of sexual dimorphism in the reference sample may not be the same as that of the target specimen (Niskanen and Junno 2009). As a result, Henneberg et al. (2005) have suggested that combined-sex equations may be more accurate because they can employ larger reference samples and encompass a broader range of variation. As an alternative, Ruff (2000) has recommended averaging the results of the male and female morphometric equations to create the equivalent of a "combined-sex" equation for situations when sex cannot be assigned. So far, the relative merits of these different practices have not been evaluated with a sample of known body mass and sex.

With the foregoing issues in mind, we used virtually reconstructed skeletal elements derived from CT scans for a large sample of modern humans of known body mass to systematically evaluate the predictive ability of the most widely used post-cranial body mass estimation equations. We tested four sets of "mechanical" equations (Ruff et al. 1991; McHenry 1992; Grine et al. 1995; Ruff et al. 2012) and two sets of "morphometric" equations (Ruff 1994; Ruff et al. 1997, 2005). The resulting accuracy rates were then used to evaluate several claims that have been made about the performance of the equations: (1) morphometric equations are more reliable than mechanical equations (Auerbach and Ruff 2004); (2) "matched-target" equations are more accurate than "generalized" equations, and "mismatched-target" equation are less accurate than either (McHenry 1992; Grine et al. 1995); (3) when using the mechanical method, if a specimen does not match a "target" equation, averaging the results obtained with several equations will estimate mass reliably (Auerbach and Ruff 2004); (4) sex-specific equations are more accurate than combined-sex ones (Ruff et al. 2012); and (5) when applying the morphometric method, if sex cannot be determined, averaging sex-specific equations yields reliable estimates (Ruff 2000).

Materials and methods

The sample

This study used archived CT scan data from a sample of 253 deceased modern human adults. The sample consisted of 128 males and 125 females, between 18 and 90 years of age (M mean=48.1 years, F mean=51.2 years). The data were obtained from the Institute of Forensic Medicine (IFM) at the University of Zurich, Switzerland where whole-body CT scans are routinely taken for all individuals entering the facility for forensic evaluation (Thali et al. 2003, 2007). The scans are maintained on the IFM's secure server and were accessed with the approval of the IFM in accordance with its protocols.

Sample individuals were selected through query searches of the IFM's database, record review, and visual inspection of the CT scans. Individuals with skeletal abnormalities, trauma, or implants were excluded, as were individuals who were processed more than 3 days after death. Sex, age at death (in years), body mass at death (in kg), and stature (in cm) were recorded for each individual. Body mass index (BMI) was calculated from body mass and stature using the standard equation (mass/stature^2) to provide an indication of overall body condition. As population affinity is not recorded in post-mortem documentation in Switzerland, it was not included as a variable in the present study. However, as more than 80 % of the Swiss population is of European descent (SFSO 2012), we consider the sample to be European. Table 1 provides summary data for the sample.

Imaging and 3D reconstruction protocols

CT imaging was conducted using a 128-slice, Siemens SOMATOM® Definition Flash, Dual-source CT scanner (Siemens Healthcare; Forchheim, Germany). Scans were taken according to IFM protocols at 120 kilovoltage (kV) with milliampere-second (mAs) automatically optimized using the Siemens CareDose® option, and slice thicknesses of 0.75 mm (0.375 mm overlap), using bone convolution kernels (Thali et al. 2007). Scans were accessed from the IFM's archives, and the regions of interest were volume rendered using OsiriX imaging software (http://www.osirix-viewer.com). The skeletal elements were oriented in a consistent plane and measured on the right side to the nearest 0.1 mm using OsiriX tools. In 13 cases, where the right femur was unusable due to a fracture or prosthetic, the left side was measured on the grounds that directional asymmetry in the lower limbs is usually small and inconsequential for the purposes of estimating body mass (Auerbach and Ruff 2004; Ruff et al. 2012). The accuracy of reconstructing virtual skeletal elements from CT data has been demonstrated for a number of applications (Cavalcanti et al. 2004; Lopes et al. 2008; Decker et al. 2011; Kim et al. 2012; Smyth et al. 2012). We also verified it in a previous study by physically measuring, scanning, virtually reconstructing, and then virtually remeasuring an archaeological skull from the IFM's collection (Elliott et al. 2014). In the latter study, measurement differences between the physical and virtual skulls were less than 3 % for all variables.

Skeletal variables

Two skeletal measurements were taken for this study (Table 2, Fig. 1): superior-inferior femoral head breadth (FHB) for use

Table 1 Summary data for sample

	Females (n=125)			Males (n=128)			Combined-sex sample (n=253)		
Variable	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Weight (kg)	69.5	19.3	31.8–146.0	81.6	16.4	40.5–142.2	75.6	18.8	31.8–146.0
Stature (cm)	166.3	8.2	149.0–195.0	177.5	7.9	154.0–193.0	171.9	9.8	149.0–195.0
Age (years)	51.2	16.5	18.0–90.0	48.1	14.1	18.0–80.0	49.6	15.3	18.0–90.0
BMI[a]	25.1	6.4	14.3–46.5	25.8	4.6	15.4–46.9	25.4	5.6	14.3–46.9

[a] *BMI* body mass index, calculated as mass(kg)/stature(m)^2

with the mechanical equations and bi-iliac breadth (BIB) for use with the morphometric equations. For the morphometric equations, stature (ST) was taken from patient documents, as measured by IFM staff at the time of processing. Intra-observer repeatability was tested by remeasuring both variables on six randomly selected individuals, with a 3-week time lapse (mean percent errors<1 %). It is important to note that the morphometric equations use "living BIB", a measure that includes the cartilaginous soft tissue between the innominates (Ruff 1994). When dealing with skeletal remains, Ruff (1994) recommends applying a conversion factor to skeletal BIB to obtain the "living" value needed for the equations. However, because the cartilage was still intact in the cadaveric individuals used here, this conversion was not necessary. Table 3 provides the summary data for the variables.

Analyses

Table 4 lists the published equations tested in this study, along with the composition of the reference samples and the regression method used. Equations for the mechanical and morphometric equations are designated by the abbreviation for the variable/s used—FHB for the femoral head breadth-based equations and STBIB for the stature plus bi-iliac breadth equations. Two studies provide sex-specific, as well as combined-sex, FHB equations (Ruff et al. 1991, 2012). Two others provide only combined-sex FHB equations (McHenry 1992; Grine et al. 1995). Only sex-specific STBIB equations have been published (Ruff et al. 1997, 2005[1]). As noted previously, these equations regress body mass on skeletal dimensions in what Kongisberg et al. (1998) refer to as "inverse regression". Although there is considerable debate about the best regression method to employ for predictive analyses (Smith 1996; Konigsberg et al. 1998; Smith 2009; Sokal and Rohlf 2012), most of the equations that have become standard in biological anthropology were derived using least squares regression (LSR). The FHB-4 equations are exceptions to this. These equations were derived using reduced major axis (RMA) regression (Ruff et al. 2012).

Analyses were conducted by entering the skeletal measurements into the appropriate equation and calculating an estimated mass. Raw and percent differences, percent prediction errors (PPE), and absolute percent differences (|PPE|) were calculated for each individual in three test groups: males, females, and combined sexes. Raw differences were calculated as (known − estimated mass), while percent prediction errors were calculated as (known − estimated mass)/known × 100 (Wu et al. 1995). PPEs indicate the directional difference of the error: positive PPE values indicate an underestimate (estimated mass < known mass), while negative values denote an overestimate (estimated mass > known mass). Absolute percent differences (|PPE|) assess the magnitude of the difference between the estimated and known masses (Dagosto and Terranova 1992; Aiello and Wood 1994). Differences between known and estimated mass were plotted, and the Wilcoxon signed-rank test was used to establish their significance. The percentage of individuals whose estimated body mass fell within ±20 % of the known mass was also calculated (see below). Assessment of reliability was based primarily on the |PPE| and the percentage of estimated masses that fell within ±20 % known mass. The |PPE|s were also used to compare the equations in relation to the assumptions previously discussed, with the Wilcoxon signed-rank test once again being used to determine the significance of the differences.

Table 2 Skeletal variables

No.	Abbreviation	Description	Reference
1	FHB	Femoral head breadth: superior-inferior breadth perpendicular to the cervical axis—in millimeters	Ruff et al. 1991
2	BIB	Bi-iliac breadth: maximum pelvic breadth taken across the iliac crests, taken as "Living BIB"—in millimeters	Ruff 1991, 1994

[1] The female equation in Ruff et al. (1997) differs from that originally provided in Ruff (1994) as a result of correcting a previous data point error.

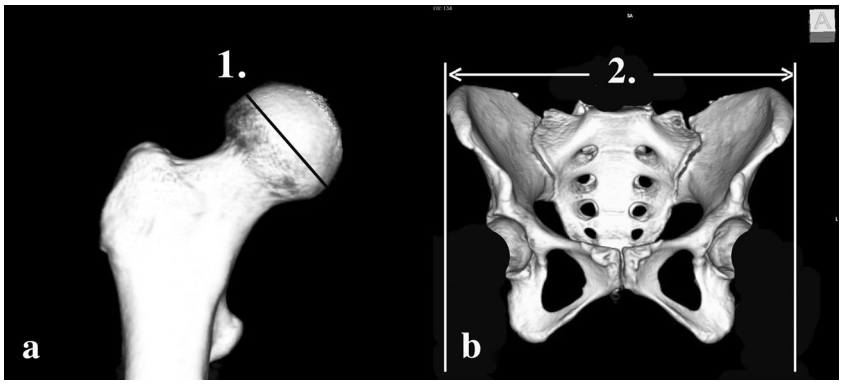

Fig. 1 Post-cranial elements three-dimensionally rendered from CT data, showing linear measurements used for this study: **a**) proximal femur, **b**) pelvis, **1**) superior-inferior femoral head breadth (FHB), **2**) bi-iliac breadth (BIB)

Expectations

There are few clear criteria for the acceptance of a skeletally-derived body mass estimate. In theory, the predictive ability of a regression equation developed and tested on a single species (e.g., humans) should be higher than one that was developed on multiple species and tested on one (Smith 2002). Since studies using interspecies comparisons have considered percent prediction errors of up to 19 % to be acceptable (Aiello and Wood 1994), this would argue in favor of a lower threshold for the postcranial intraspecific equations under consideration here. Nevertheless, because it provided a lenient baseline, we used this threshold and accepted estimates as reliable when the absolute percent prediction errors (|PPE|) fell below 19 %. In addition, studies vary in terms of their expectations for the number of individuals that can be estimated with ±20 % of their known mass. Interspecific studies tend to use relatively high thresholds—accepting equations that place 50 % of a sample within ±20 % of known mass (Dagosto and Terranova 1992; Aiello and Wood 1994). In contrast, intraspecific studies suggest that an effective equation should estimate the majority of individuals within 10 or 15 % of their known mass (Ruff et al. 2005; Lorkiewicz-Muszyńska et al. 2013). Again, we chose a conservative approach and accepted estimates when 50 % or more of the specimens fell within ±20 % of their known mass. In addition to these broad acceptance criteria, we made specific predictions for the equations' performance based on the original reference samples and intended purpose.

Mechanical (FHB) equations

Ruff et al.'s (1991) three FHB equations (FHB-1a-1c) were derived from individual data for a modern North American sample and were intended for general application to both fossil and modern humans. The reference sample is roughly contemporaneous with our test sample, and the characteristics are similar (mean body mass=76.7 kg versus our 75.6 kg). In addition, Ruff et al. (1991) provide both sex-specific and combined-sex equations. As a result, we expected each of the equations to estimate mass accurately in their respective test groups.

McHenry's (1992) single FHB equation (FHB-2) was derived from mean, combined-sex data for four modern samples (US European and African, Khoisan and African "Pygmy").[2] Because it was designed specifically to estimate mass in smaller-bodied hominins, this equation was not expected to be accurate when applied to our sample, which consists of relatively large, modern Europeans. In light of this mismatch, FHB-2 was also expected to estimate mass relatively more poorly than the other equations.

The combined-sex FHB equation provided by Grine et al. (1995) (FHB-3) was derived from mean data for ten sex-specific modern and archaeological samples, with the goal of estimating large-bodied hominins. Initially, this suggested to us that the equation might overestimate mass in the current sample, as Pleistocene hominins are considered to have been heavier and more robust than modern humans (Ruff et al. 2005; Churchill et al. 2012). However, as a Western, industrialized, modern group, our sample might also be carrying more fat than the original reference populations (Ruff et al. 1991) and the equation would underestimate mass. On the grounds that these two factors would effectively cancel each other out (Ruff 2000), we expected the FHB-3 equation to predict mass well here.

Ruff et al.'s (2012) recently derived FHB equations (FHB-4a-4c) provide sex-specific, as well as combined-sex equations for Holocene European samples. Because they were derived from a large and diverse group of individual skeletal remains, these equations have been argued to be "broadly applicable across different geographic regions and temporal periods" (Ruff et al. 2012:615). Following this, the equations were expected to estimate mass well in our sample.

[2] The equation was derived by Ruff et al. (1997) from raw data in McHenry's study.

Table 3 Summary data for each variable (in mm)

	Females (n=125)			Males (n=128)			Combined-sex sample (n=253)		
Variable	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
FHB	45.5	2.3	39.8–55.5	50.9	2.8	41.5–57.5	48.2	3.73	39.8–57.5
BIB	277.9	18.5	223.1–337.1	283.9	16.8	211.2–324.2	280.9	17.83	211.2–337.1

Morphometric (STBIB) equations

The two sex-specific STBIB-1 equations tested here were developed using population-mean anthropometric data taken from living individuals belonging to 56 different populations (Ruff 1994; Ruff et al. 1997). As associated data were not available, mean body masses for each group were gleaned from the literature (Ruff 1991). Designed to encompass a wide range of body sizes, these equations have been argued to "provide the most generally reliable" estimates of mass when bi-iliac breadth and stature can be measured or estimated with some confidence (Auerbach and Ruff 2004:340). As we were able to measure both features directly in our study sample, we expected the equations to perform well.

As noted, the STBIB-2 equations result from efforts to improve body mass estimates by expanding the range of variation in the reference sample to include more high-latitude (>46° N) populations (Ruff et al. 2005). As another high-latitude, tall, and broad population (Zurich is at 47.4° N and all means either met or exceeded those for the Inupiat and Finnish groups), we expected body mass to be estimated well with these equations, particularly for the males.

Relative performance

We also tested specific expectations for the relative performance of the equations based on the claims previously discussed. The first hypothesis states that morphometric equations are more reliable than mechanical ones (Auerbach and Ruff 2004). Thus, we expected STBIB-1 and STBIB-2 to outperform all of the FHB equations.

According to the second hypothesis, "matched-target" equations are more accurate than "generalized" or "mismatched-target" equations. Here, we expected our modern European sample to be estimated better using the large-bodied FHB-3 equation than the more "generalized" FHB-1 and FHB-4 equations (at least with the combined-sex equations). However, all three of these equations were expected to perform better than the "mismatched" FHB-2 equation, which was designed for small-bodied individuals. On the same grounds, because our test sample was derived from a relatively high-latitude population, we expected the "matched-target" STBIB-2 equation to be more accurate than the more "generalized" STBIB-1 equation, particularly for men.

Table 4 Published regression equations for estimating body mass from femoral head breadth (FHB) or stature plus bi-iliac breadth (STBIB)

Method	Female	Male	Combined-sex	Source	Sample composition and method
FHB-1	2.43*FHB-35.1	2.74*FHB-54.9	2.16*FHB-24.8	Ruff et al. 1991, 1997	80 living individuals (US whites and blacks), LSR regression
FHB-2	n/a	n/a	2.24*FHB-39.9	McHenry 1992, see Ruff et al. 1997	Mean data from four samples (US European and African, Khoisan and African Pygmy), LSR regression
FHB-3	n/a	n/a	2.27*FHB-36.5	Grine et al. 1995	Mean data from 10 sex-specific samples (African American, European American and Native American), LSR regression
FHB-4	2.18*FHB-35.8	2.80*FHB-66.7	2.30*FHB-41.7	Ruff et al. 2012	Archaeological sample of 1145 individuals (European Holocene), RMA regression
STBIB-1	0.52*STAT+ 1.81*LBIB-75.50	0.37*STAT+ 3.03*LBIB-82.5	Average of male and female[a]	Ruff 1994; Ruff et al. 1997; Ruff 2000	Sex-specific mean data for 56 samples (Worldwide), LSR regression
STBIB-2	0.50*STAT+ 1.80*LBIB-72.60	0.42*STAT+ 3.13*LBIB-92.9	Average of male and female[a]	Ruff et al. 2005	Same data as for STBIB-1, with addition of two Finnish groups (1 male, 1 female), LSR regression

All equations are for raw (non-logged) data. FHB in millimeters, stature (ST) and living bi-iliac breadth (LBIB) in centimeters. Resulting BM in kilograms.

[a] As recommended by Ruff (2000).

A third hypothesis claims that the results of multiple mechanical equations can be averaged when a specimen is not specifically large- or small-bodied (Auerbach and Ruff 2004). Thus, we expected the average of FHB-1, FHB-2, and FHB-3 to estimate mass well, and at least as well as the generalized equations. We also expected the average of four mechanical equations (FHB-1-4) to perform well.

The fourth hypothesis assumes that sex-specific equations will perform better than combined-sex equations (Ruff et al. 2012). As a result, we expected FHB-1, FHB-4, STBIB-1, and STBIB-2 to return lower error rates than FHB-2 or FHB-3.

Lastly, the fifth hypothesis argues that it is appropriate to average the results of the sex-specific equations if sex is uncertain (Ruff 2000). Although we still expected males and females to be estimated best with their respective sex-specific morphometric equations, averaging the two results was expected to produce reliable estimates.

Results

Table 5 and Fig. 2 summarize the results for each of the FHB and STBIB equations. Table 6 provides the raw mean predicted masses, difference from known mean, and 95 % confidence intervals for the predicted masses.

Mechanical (FHB) equations

Using Ruff et al.'s (1991) sex-specific FHB-1a and FHB-1b equations, respectively, the male sub-sample was estimated within acceptable limits, but the female sub-sample was not. For males, the mean |PPE| was below 16.1 % and 71.1 % of the individuals were estimated within ±20 % of their known mass. In contrast, for females the mean |PPE| exceeded 19 % and less than half the sample (48 %) was estimated within ±20 % of known mass. The combined-sex equation (FHB-1c) only partially met the acceptance criteria. More than 50 % (59.7 %) of the individuals were estimated within ±20 % of known mass, but the mean |PPE| exceeded 19 % (20.2 %). In terms of the direction of error, all three equations overestimated mass on average in their respective samples. Thus, the FHB-1 equations did not consistently estimate mass within acceptable limits in the test samples.

McHenry's (1992) single, combined-sex FHB-2 equation estimated mass within acceptable limits in the male and combined-sex test groups. In both cases, mean |PPE|s were below 19 % and more than 50 % of the individuals were estimated within ±20 % of their known mass. In contrast, 58.4 % of the female sample was estimated within ±20 % of known mass, but the |PPE| exceeded 19 % (albeit by a small margin at 19.9 %). The direction of error was again consistent across the three test groups, but masses were underestimated. Thus, this equation also failed to estimate mass reliably in all test groups.

Grine et al.'s (1995) equation (FHB-3) also resulted in mean estimates that met both criteria for acceptance in the male and combined-sex samples. Males showed the lowest mean error and estimated the highest number of individuals within ±20 % of known mass (>75 %). However, only one of the two acceptance criteria was met in the female sample (more than 50 % were estimated within ±20 % of known mass, but the |PPE| exceeded 19 %). In terms of directional error, FHB-3 overestimated mass on average in all three groups. This equation did not estimate mass consistently, or within acceptable limits, across the three test samples.

Of Ruff et al.'s (2012) three sample-specific equations, the ones for males (FHB-4b) and combined-sexes (FHB-4c) met the criteria for acceptance. However, the female-only equation

Table 5 Differences between known and estimated body masses for each equation

Method	Female (n=125)			Male (n=128)			Combined-sex (n=253)		
	PPE Mean[a] (SD)	|PPE| Mean (SD)	20 % (%)	PPE Mean[a] (SD)	|PPE| Mean (SD)	20 % (%)	PPE Mean[a] (SD)	|PPE| Mean (SD)	20 % (%)
FHB-1	−15.5 (29.7)	25.4 (21.7)	48.0	−7.2 (20.2)	*16.1 (14.2)*	*71.1*	−10.4 (25.0)	20.2 (18.1)	59.7
FHB-2	5.0 (24.6)	19.9 (15.2)	58.4	6.1 (17.7)	*15.0 (11.2)*	*73.4*	5.5 (21.3)	*17.4 (13.5)*	*66.0*
FHB-3	−2.3 (26.4)	20.7 (16.4)	56.0	−0.1 (18.9)	*14.7 (11.9)*	*75.8*	−1.2 (22.9)	*17.6 (14.6)*	*66.0*
FHB-4	2.8 (25.1)	19.9 (15.4)	57.6	4.0 (18.1)	*14.7 (11.3)*	*72.7*	4.0 (21.7)	*17.2 (13.7)*	*66.4*
STBIB-1[b]	6.2 (22.1)	*18.6 (13.3)*	*64.8*	11.8 (15.7)	*16.1 (11.3)*	*66.4*	8.6 (19.7)	*17.5 (12.4)*	*62.9*
STBIB-2[b]	6.5 (22.0)	*18.6 (13.4)*	*63.2*	10.7 (15.8)	*15.4 (11.2)*	*68.8*	9.1 (19.6)	*17.6 (12.4)*	*62.1*

PPE percent prediction error (known − estimated)/known × 100, |PPE| absolute percent prediction error, *20 %* percent of individuals whose estimated body masses fall within ±20 % of known mass.

[a] Directional differences (positive values indicate underestimation, negative values indicate overestimation);

[b] combined-sex values are the average of the male and female estimates as recommended by Ruff (2000);

Italics values indicate the analyses that achieved mean |PPE|s below 19 % and estimated more than 50 % of the sample within ±20 % of known mass.

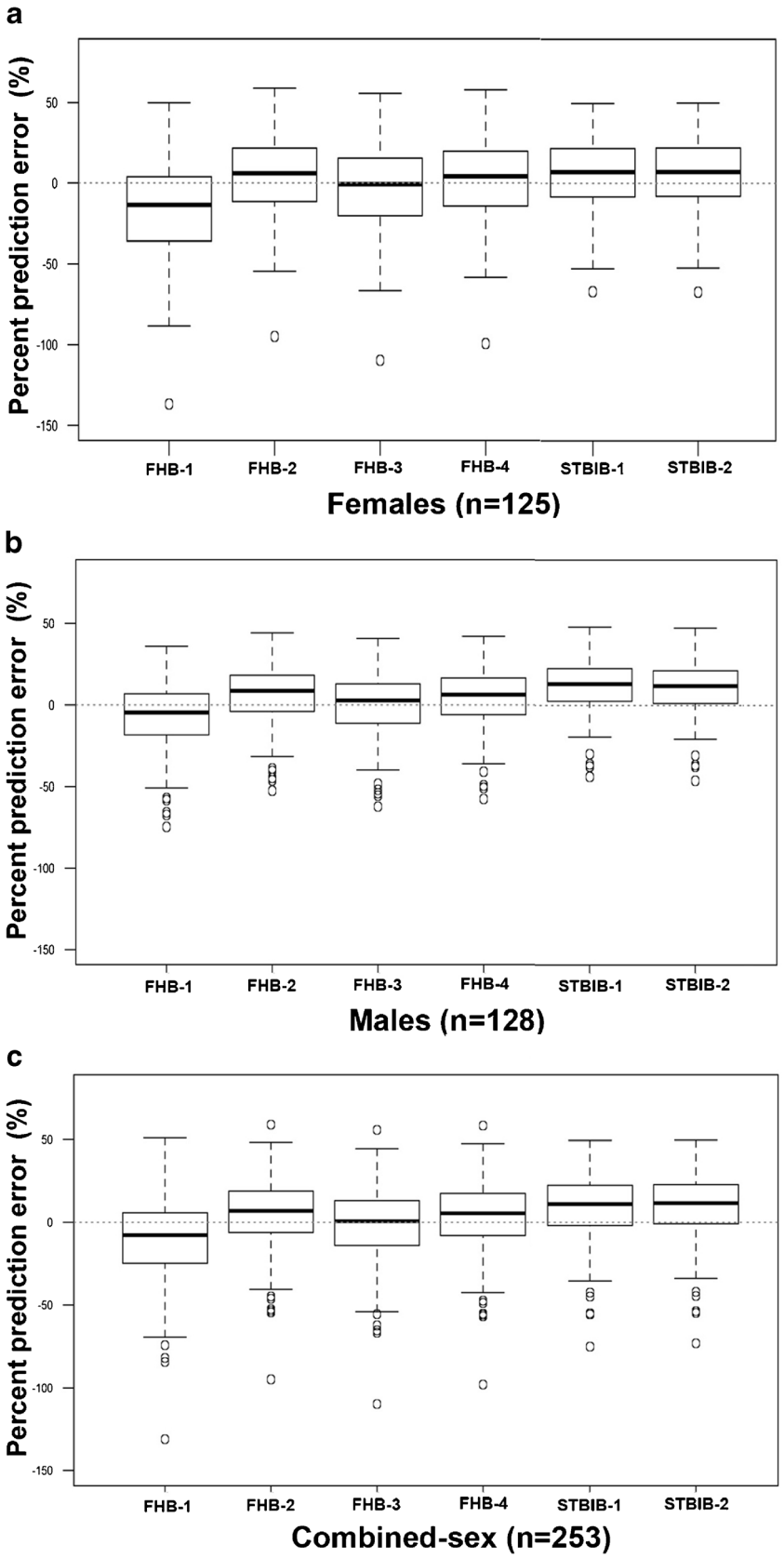

Fig. 2 Percentage prediction errors for each of the FHB and STBIB equations

Table 6 Predicted mass (kg), mean difference (kg), and confidence intervals for each test sample

Method	Female (n=125)			Male (n=128)			Combined-sex (n=253)		
	Predicted BM mean (SD)	Raw difference[a] mean (SD)	95 % CI	Predicted BM mean (SD)	Raw difference[a] mean (SD)	95 % CI	Predicted BM mean (SD)	Raw difference[a] mean (SD)	95 % CI
Known mass	69.5 kg	–	–	81.6 kg	–	–	75.6 kg	–	–
FHB-1	75.2 (5.4)	−5.7 (19.3)*	74.3–76.2	84.6 (7.7)	−3.1 (15.0)*	83.3–86.0	79.4 (8.1)	−3.7 (17.2)*	78.4–80.4
FHB-2	61.9 (5.1)	7.6 (19.3)*	61.0–62.8	74.1 (6.3)	7.5 (14.9)*	73.0–75.2	68.1 (8.4)	7.6 (17.2)*	67.0–69.1
FHB-3	66.6 (5.2)	2.9 (19.3)	65.7–67.6	79.0 (6.3)	2.6 (14.9)	77.9–80.1	72.9 (8.5)	2.8 (17.2)	71.8–73.9
FHB-4	63.3 (5.0)	6.2 (19.2)*	62.5–64.2	75.8 (7.8)	5.7 (15.0)*	74.5–77.2	69.2 (8.6)	6.4 (17.2)*	68.2–70.3
STBIB-1	61.6 (6.0)	8.0 (17.7)*	60.5–63.6	69.8 (6.8)	11.8 (14.4)*	68.6–71.0	66.0 (7.2)	9.6 (16.2)*	65.1–66.8
STBIB-2	61.3 (5.9)	8.2 (17.7)*	60.3–62.4	70.7 (7.3)	10.8 (14.3)*	69.5–72.0	65.6 (7.3)	10.0 (16.2)*	64.7–66.5

[a] Positive values indicate that the predictive equation underestimates true mass, negative values indicate that the equation overestimates true mass.

*Indicates predicted vs. known mass difference significance at p=0.05.

(FHB-4a) failed to estimate the test sample within the 19 % |PPE| criterion (again, by a small margin; 19.9 %). Directionally, the three equations all underestimated mass in their respective test groups. Thus, only two of the three equations estimated mass in the test samples within acceptable limits.

Morphometric (STBIB) equations

Ruff et al.'s (1997) two sex-specific equations (STBIB-1a and 1b) resulted in estimates that met the criteria for acceptance in their respective test groups. Ruff et al.'s (2005) sex-specific equations (STBIB-2a and 2b) also resulted in estimates that fell within acceptable limits for each test group. All four equations underestimated mass on average. Thus, both sets of equations estimated mass reliably in the test groups.

Hypotheses

1. *The morphometric/STBIB method estimates body mass more reliably than the mechanical/FHB method.* Table 7 summarizes the results for this comparison. As can be seen, not all the results are consistent with the prediction. In females, the STBIB-1a and STBIB-2a equations estimated mass more reliably than most of the FHB equations, to a statistically significant level (p<0.05). However, neither was significantly better than that of the FHB-4 equations. In males, neither the STBIB-1b nor the STBIB-2b equation estimated mass more reliably than that of the FHB equations.

   The results for the sample as a whole were mixed. As noted previously, because separate equations were not provided, the combined-sex STBIB-1c or 2c results are the product of averaging the male and female estimates. However, there is still value in considering these results in the context of comparing the mechanical and morphometric methods. Here, the STBIB-1c or STBIB-2c averages were only more reliable than the FHB-1c equation, at a significant level. The remaining combinations either went against the prediction or were statistically insignificant. Overall, the claim that the morphometric method produces more reliable estimates than the mechanical method was not consistently supported.
2. "*Matched-target*" *equations are more accurate than* "*generalized*" *equations and* "*mismatched-target*" *equations are less accurate than either the* "*matched-target*" *or* "*generalized*" *equations.* The results of this test are summarized in Table 8. This claim was only partially supported in the test sample. In females, three equations performed as expected. However, one matched-target equation (FHB-3) went against the prediction and was significantly *less accurate* than a generalized equation (STBIB-1). The expectation that the mismatched FHB-2 equation would estimate mass less accurately than the matched equations, held only in relation to STBIB-2 and not to FHB-3. The mismatched FHB-2 equation estimated mass significantly worse than the generalized STBIB-1 equation and went against expectations by estimating mass significantly better than that of the generalized FHB-1 equation. In the male sample, two matched-target equations met expectations and estimated mass significantly more accurately than the general equations. However, none of the other combinations met the prediction to statistically significant levels. For the combined-sex sample, the matched-target FHB-3 equation and STBIB-2 average were significantly more accurate than the general FHB-1 equation. Against expectations, the general FHB-1 equation was not more accurate than the mismatched FHB-2 equation in this group. Thus, the results do not support the assertion that body mass will be estimated most reliably by using a "matched-target" equation.

Table 7 Expectation that morphometric (STBIB) equations will achieve greater accuracy than the mechanical (FHB) equations

Expectation	Female (n=125)		Male (n=128)		Combined-sex (n=253)	
	P value	Expectation[a] met?	P value	Expectation[a] met?	P value	Expectation[a] met?
STBIB-1 more accurate than FHB-1	0.00*	Y	0.95	=	0.02*	Y
STBIB-1 more accurate than FHB-2	0.05*	Y	0.12	N	0.78	N
STBIB-1 more accurate than FHB-3	0.04*	Y	0.17	N	0.65	Y
STBIB-1 more accurate than FHB-4	0.08	Y	0.10	N	0.57	N
STBIB-2 more accurate than FHB-1	0.00*	Y	0.65	Y	0.03*	Y
STBIB-2 more accurate than FHB-2	0.05*	Y	0.51	N	0.62	N
STBIB-2 more accurate than FHB-3	0.04*	Y	0.44	N	0.98	=
STBIB-2 more accurate than FHB-4	0.08	Y	0.36	N	0.44	N

[a] Based on |PPE| values, equal sign "=" indicates |PPE|s were the same for both formulae;

*Indicates the difference between methods is significant (p=0.05).

3. *When using the mechanical method, if a specimen does not fit with one of the "target" equations, taking the average of the results from other equations produces reliable estimates*. Table 9 summarizes these results. This hypothesis was also partially supported. In males, both the three and four-average estimates resulted in final estimates that met the criteria for acceptance. In this group, the averaged results estimated mass as well as, or better than, using a single equation. The same pattern held for the combined-sex samples—both the average of three and the average of four estimates produced mean estimates that met the criteria for acceptance, and performance was better than most of the single FHB equations. In females, averaging the results of three (FHB-1-3) or four (FHB-1-4) estimates did not result in a final estimate that met the acceptance criteria for reliability. However, the estimation accuracy remained similar by averaging estimates in this group. Using the average of four equations was better than using the average of three equations in the female and combined sex samples (p<0.05), but in not the male group. These results partially support the practice of averaging different equations for a "generalized" specimen.

Table 8 Expectation that "matched-target" equations will achieve greater accuracy than "generalized" equations, but that the "mismatched-target" equation will achieve lower accuracy than the "generalized" equations

Expectation	Female (n=125)		Male (n=128)		Combined-sex (n=253)	
	P value	Expectation[a] met?	P value	Expectation[a] met?	P value	Expectation[a] met?
Matched target vs. generalized						
FHB-3 more accurate than FHB-1	0.00*	Y	0.02*	Y	0.00*	Y
FHB-3 more accurate than FHB-2	0.22	N	0.56	Y	0.58	N
FHB-3 more accurate than FHB-4	0.08	N	0.93	=	0.22	N
FHB-3 more accurate than STBIB-1	0.04*	N	0.17	Y	0.65	N
STBIB-2 more accurate than FHB-1	0.00*	Y	0.65	Y	0.03*	Y
STBIB-2 more accurate than FHB-2	0.05*	Y	0.51	N	0.62	N
STBIB-2 more accurate than FHB-4	0.08	Y	0.36	N	0.44	N
STBIB-2 more accurate than STBIB-1	0.57	=	0.00*	Y	0.00*	N
Mismatched target vs. generalized						
FHB-2 less accurate than FHB-1	0.00*	N	0.31	N	0.00*	N
FHB-2 less accurate than FHB-4	0.95	=	0.26	Y	0.09	Y
FHB-2 less accurate than STBIB-1	0.05*	Y	0.12	N	0.78	N

[a] Based on |PPE| values, equal sign "=" indicates |PPE|s were the same for both formulae;

*Indicates difference between methods is significant (p=0.05).

Table 9 Differences between known and estimated body masses when multiple mechanical methods are averaged

Method	Female (n=125)			Male (n=128)			Combined-sex (n=253)		
	PPE Mean[a] (SD)	|PPE| Mean (SD)	20 % (%)	PPE Mean[a] (SD)	|PPE| Mean (SD)	20 % (%)	PPE Mean[a] (SD)	|PPE| Mean (SD)	20 % (%)
Mean of FHB 1-3[b]	−4.3 (26.9)	21.1 (17.0)	58.4	−0.4 (18.9)	*14.7 (11.9)*	*75.8*	−2.0 (23.1)	*17.8 (14.8)*	*66.4*
Mean of FHB 1-4[c]	−2.5 (26.4)	20.7 (16.5)	56.8	0.7 (18.7)	*14.6 (11.7)*	*75.0*	−0.5 (22.7)	*17.5 (14.4)*	*66.4*

Italic numbers indicate the variables that achieved |PPE|s below 19 % and estimated more than 50 % of the sample within ±20 % of known mass;

[a] Directional differences (positive values indicate underestimation, negative values indicate overestimation);

[b] As recommended in Auerbach and Ruff (2004);

[c] As calculated in this study.

4. *Sex-specific equations are more accurate than the combined-sex equations*. Table 10 summarizes these results. This hypothesis was also only partially supported. In males, most of the comparisons went against expectations, but the differences were not statistically significant. The differences between FHB-4 and the two combined-sex equations were also not statistically significant. However, in females, the sex-specific FHB-1 equation estimated mass significantly worse than the combined-sex FHB-2 or FHB-3 equations ($p<0.05$). In the context of this assertion, the sex-specific STBIB-1 and STBIB-2 equations were also expected to estimate mass better than that of the combined-sex FHB-2 and FHB-3 equations. This was true for the female test sample (all statistically significant at $p<0.05$), but not for the males. Thus, sex-specific equations are not necessarily more reliable than combined-sex equations.
5. *When applying the morphometric method, if sex cannot be determined, mass will be estimated reliably by taking the average of the sex-specific equations.* The results, summarized in Table 4, support this hypothesis. For both STBIB-1 and STBIB-2, taking the mean of the male and female results produced estimates that met both criteria for acceptance. Not surprisingly, using the sex-averaged equations resulted in error rates that fell between those of the two sex groups, and overall, did not significantly reduce accuracy. In this context, the results support the hypothesis that averaging the results of the sex-specific equations will estimate mass reliably.

Discussion

Many of the equations for estimating body mass from post-cranial material met the criteria for acceptance in the male and the combined-sex samples. However, this was not the case for the combined-sex FHB-1 equation. In addition, none of the mechanical/FHB equations estimated mass reliably in the female sample. The equations also did not consistently perform as expected given their reference samples and target groups. For example, Ruff et al.'s (1991) FHB-1 equations were the least accurate estimators of mass in the test sample. This was

Table 10 Expectation that sex-specific equations will achieve greater accuracy than mixed-sex equations

Expectation	Female (n=125)		Male (n=128)	
	P value	Expectation[a] met?	P value	Expectation[a] met?
FHB-1 more accurate than FHB-2	0.00*	N	0.31	N
FHB-1 more accurate than FHB-3	0.00*	N	0.02*	N
FHB-4 more accurate than FHB-2	0.95	=	0.26	Y
FHB-4 more accurate than FHB-3	0.08	Y	0.93	=
STBIB-1 more accurate than FHB-2	0.05*	Y	0.12	N
STBIB-1 more accurate than FHB-3	0.04*	Y	0.17	N
STBIB-2 more accurate than FHB-2	0.05*	Y	0.51	N
STBIB-2 more accurate than FHB-3	0.04*	Y	0.44	N

[a] Based on |PPE| values, equal sign "=" indicates |PPE|s were the same for both formulae;

*Indicates difference between methods is significant (p=0.05).

despite being derived from modern individuals with characteristics similar to those of the test sample and including sex-specific equations. The FHB-4 equations did not estimate mass as well as expected despite being derived from a much larger sample, providing sex-specific equations and being described as "superior" to all other methods (Ruff et al. 2012:601). In contrast, McHenry's (1992) FHB-2 equation estimated mass better than anticipated given the single combined-sex equation and "mismatched" test and reference samples. Even with the lenient criteria used in the present study, none of the FHB equations estimated mass to the level of acceptance for reliability in females.

We also found mixed support for the claims that have been made regarding the way the equations should perform relative to one another. The morphometric/STBIB equations did not estimate mass more reliably than the mechanical/FHB methods in all groups. Females were estimated better with the STBIB equations, but males were estimated better with the FHB equations. When the sexes were combined, the morphometric/STBIB equations estimated mass less reliably than all but one of the mechanical/FHB equations (FHB-1c). In light of these results, it is not appropriate to use the morphometric/STBIB equations in preference to the mechanical/FHB equations.

Similarly, using "matched-target" equations did not consistently improve estimation accuracy over "generalized" equations. In keeping with the predictions, the FHB-3 (Grine et al. 1995) equation designed for large-bodied hominins estimated mass better than the general FHB-1 equation (Ruff et al. 1991) in the combined-sex group. However, it did not perform as well as the other general FHB-4 equation (Ruff et al. 2012). McHenry's (1992) FHB-2 equation met expectations and estimated mass less reliably than the general FHB-4 equation. However, as noted previously, it performed better than the generalized FHB-1 equation despite being designed to estimate mass in small-bodied hominins. It also unexpectedly estimated males better than it did females. For the morphometric method, the "matched-target" STBIB-2 equation (Ruff et al. 2005) estimated mass more reliably than the more general STBIB-1 equation (Ruff et al. 1997) in males. However, it did not estimate mass better in the female test sample or when the sexes were combined. Although the results support Ruff et al.'s (2005) suggestion that the newer equations may be more appropriate for estimating large, high-latitude males, they also suggest that broadening the reference sample does not necessarily provide better estimation and the STBIB-2 equations should not simply replace the STBIB-1 equations uncritically.

In partial support of the claim that averaging the results of multiple equations produces reliable body masses (Ruff et al. 1991; McHenry 1992; Grine et al. 1995), the male and combined-sex groups were estimated slightly more accurately using the average of three FHB equations than using a single equation. In females, although the error rates decreased slightly compared to most of the single FHB equations, averaging the results of FHB-1a, FHB-2a, and FHB-3a still did not produce mean estimates that met the criteria for acceptance. Averaging the results of four FHB equations (Ruff et al. 1991; McHenry 1992; Grine et al. 1995; Ruff et al. 2012) also resulted in a modest improvement compared to using a single equation in the male and combined-sex samples. But again, the four-estimate average failed to estimate females to an acceptable level. In general, averaging multiple FHB equations has a neutral or slightly positive effect on predictive accuracy.

Sex-specific equations were not consistently more accurate than those designed for combined sexes. The sex-specific FHB-1 or FHB-4 equations should have achieved greater accuracy than either of the combined-sex FHB equations (McHenry 1992; Grine et al. 1995). However, this was not the case, particularly in females. This result suggests that the use of sex-specific equations may not be critical, at least in species like humans, who exhibit relatively low levels of sexual dimorphism (Ruff 2002). In fact, it may be that the greater number of individuals in the combined-sex reference group is driving the improved accuracy and that large sample sizes are more important than group-specificity for developing predictive equations.

Lastly, averaging the male and female morphometric/STBIB estimates in the absence of known sex, produced results that met the acceptance criteria. |PPE|s for the averaged values were slightly lower than those for females, but higher than for males. Fewer individuals were estimated within ±20 % of known mass using the average of the sexes than either sex alone. However, in both cases, the differences were small. The results of this test suggest that there does not appear to be a significant cost to accuracy by averaging sex-specific equations.

Several of these results require further consideration. First, although they were surprising in relation to the claims in the literature, there may be a simple explanation for the relatively poor performance of Ruff et al.'s (1991) FHB-1 equations—namely the use of indirect measures for the key variables. Specifically, Ruff et al. (1991) used patient-recalled weight as the measure for body mass. While recall information may be useful in some contexts (Olivarius et al. 1997), self-reported weights are notoriously inaccurate, particularly in women (Perry et al. 1995; Bayomi and Tate 2008). The fact that female mass was incorrectly estimated more often than males in the current sample supports this as a possibility. Ruff et al. (1991) also measured femoral head breadth indirectly from conventional radiographs, compensating for variations in magnification caused by differences in the distance from the radiographic plate, with a single correction factor (19 %). While the actual variation between individuals may not have been large, the inability to measure each element directly introduces an additional source of error to the method.

Combined with the subjective assessment of weight, this could explain why the equations did not perform well, particularly in females.

The results in relation to the female sub-sample as a whole are more difficult to explain. As noted earlier, none of the FHB equations resulted in mean estimates that met the criteria for acceptance in this group, even when sex-specific equations were used. Apart from the biases discussed above in relation to FHB-1, this initially suggested that the test group females were too different from the females in the reference samples and the equations failed because they were extrapolating beyond their range (Konigsberg et al. 1998). However, the actual samples are not consistent with this explanation. The reference sample for McHenry's (1992) equation had a FHB mean of 41.5 mm and a range of 33–47.5 mm, while Grine et al.'s (1995) reference sample ranged from 38.4 to 50.5 mm (Ruff 2010). Our female sample had a FHB mean of 45.5 mm and a range of 39.8–55.5 mm. Thus, it was more similar to Grine et al.'s (1995) reference sample and "fit" the appropriate range for their equation better than McHenry's (1992). As result, Grine et al.'s (1995) FHB-3 equation should have estimated mass better than McHenry's (1992) FHB-2 equation. However, this was not the case and FHB-3 returned higher |PPE|s and estimated fewer individuals within ±20 % of known mass than FHB-2. This indicates that a "match" between the reference sample and target specimen does not ensure a reliable estimate.

Despite this, a reference-target sample "mismatch" does not fully explain why all the FHB equations estimated females relatively poorly. One possibility relates to the level of adipose tissue in females. In addition to carrying more fat and less muscle mass than men, women carry their mass differently and are more prone to fluctuations in weight than men (Shen et al. 2004; Power and Schulkin 2008). As a result, it is possible that the failure of the postcranial equations to estimate mass as well in females as they did in males has to do with a "looser" relationship between femoral head morphology and body mass in the former. Correlation coefficients in the present sample support this: the relationship between FHB and body mass in females (r=0.15) was considerably lower than that for males (r=0.42). However, on these grounds, the sex-specific FHB-1a (Ruff et al. 1991) and FHB-4a (Ruff et al. 2012) equations should have estimated mass better than the combined-sex equations of McHenry (1992) and Grine et al. (1995) because they were developed from exclusive female reference groups. This was not the case. Thus, further research is needed to determine the cause of the differential results for males and females.

As noted earlier, Lorkiewicz-Muszyńska et al. (2013) also found significant inaccuracies in body mass estimates when using a known-mass sample. Although Lorkiewicz-Muszyńska et al. (2013) did not provide the details of their sample, it was also reasonably large (n=120), presumably European (Polish medical sample), and similar to our sample in age range (20-88 years versus our 18-90) and BMI distribution (53 % vs our 46 % within the "normal range"). Consequently, we expected the pattern of mean differences to be roughly similar between the two studies. Indeed, in both studies, Grine et al.'s (1995) combined-sex equation estimated mass better than the other equations, females were generally estimated more poorly than males, and the STBIB equations did not return lower mean differences than the FHB equations. These similarities are problematic for the postcranial methods, however, as they reinforce inconsistencies in the way the equations perform relative to their expected performance.

Two additional factors suggest that the equations may be even less accurate than what has been described here. First, although many of the equations met the acceptance criteria, the standards were extremely lenient, particularly for intraspecies regressions. As noted earlier, scaling differences between groups suggest that intraspecies regressions will be more accurate than interspecies regressions (Smith 2002). As a result, a more appropriate criterion for acceptance in an intraspecies analysis would expect the majority of individuals to fall within 10–15 % of their known mass (Ruff et al. 2005). However, if we apply this to the current sample and require 50 % of the individuals to be estimated within ±15 % of their known mass, the number of "acceptable" equations falls significantly (Supplementary Table 1). With a ±10 % criterion, none of the equations would be considered reliable. In fact, no equation estimates more than 41 % of the sample within ±10 % of the known mass, a figure that suggests the body mass estimates calculated from these equations should only be considered loose approximations.

The second point relates specifically to the use of the morphometric equations. Although they estimated mass within acceptable limits and performed marginally better than the mechanical equations, it must be emphasized that our results were obtained using *documented* stature—a condition that is rarely available in skeletal specimens, even modern ones. Importantly, this suggests that existing estimates of body mass that use the morphometric/STBIB method with stature estimated from some other skeletal feature are likely to be even less reliable because the error is compounded. Thus, our results do not support the claim that the STBIB method should be preferred over the FHB methods, even when bi-iliac breadth and stature can be reliably measured (Auerbach and Ruff 2004).

An important consideration with respect to these results relates to age. The current test sample encompassed an age range of 18–90 years, with 45 % of the individuals falling between 40 and 59 years (Supplementary Table 2). Body mass is known to fluctuate with age and tends to increase after the fifth decade, particularly in females (Holloway 1980; Ruff

et al. 1991, 2005). In addition, some have argued that past populations may not have lived much beyond the age of 60 years (Robson and Wood 2008) and may not have been as heavy as modern groups tend to be. Consequently, it is possible that age-related changes in body mass are responsible for some of the error associated with the estimations of mass using existing postcranial equations.

In a previous paper, we explored the effect of age in relation to existing cranially-based body mass equations and did not find a significant effect on estimation accuracy (Elliott et al. 2014). To consider this in relation to the six postcranial equations tested here, we conducted two additional sets of analyses. The first employed an approach used in our earlier study focused on cranial equations (Elliott et al. 2014) and reassessed the postcranial equations using only individuals between 18–60 years (n=186). Restricting the sample to an age range more in keeping with past populations resulted in modest improvements in accuracy for all equations (Supplementary Table 3). However, most differences were not significant (p<0.01) and females did not show a more marked effect than males.

In the second set of analyses, we carried out correlation analyses between the absolute PPEs for the published equations and age (Supplementary Table 4). For most of the equations, correlations were slightly positive in females and the combined-sex group, but none were significant (p<0.01). In males, correlations with age tended to be negative but were also insignificant. McHenry's (1992) equation was slightly positively correlated in males, but was again insignificant. These results suggest that age-related body mass differences between females and males may be an important consideration. However, the effect was small in the current sample and further research is needed before age adjustments could be recommended. A similar test with stature found an increase in prediction error (overestimation) with age (Ruff et al. 2012) and a more exaggerated response in females. However, as in the current study, the effect was small in both sexes and Ruff et al. (2012) concluded that an age adjustment was unnecessary. Overall, the differential performance of the equations in the present study does not appear to be related closely to the age of the individuals in the sample.

Our results have some important implications for biological anthropology. While most of the equations performed adequately on our study sample, they did not estimate mass well in the female subsample. In addition, the results were achieved under ideal conditions: closely matched test and reference groups, directly measured and associated variables, and lenient acceptance criteria. For fossil species, whose skeletal elements may require significant reconstruction or approximation and whose body proportions, muscle mass, and sexual dimorphism may differ markedly from modern groups (Churchill et al. 2012), errors in estimation are likely to be much larger than those obtained here. Thus, the existing equations may not be appropriate for the wide range of species they are often applied to (Ruff 2010) and concerns about making inferences about the physiology and behavior of fossil hominins from these estimates are still relevant (Smith 1996). Similar caveats apply to body mass estimates in archaeological and modern samples. Even in forensic contexts, complete elements are rare (Pokines et al. 2013) and in both cases variations in body proportion and muscle mass between groups make it difficult to identify the "appropriate" equation for the target specimen (Ruff et al. 2012). Even with a suitable reference group, fluctuations in individual mass in response to dietary changes, pregnancy, age, etc. mean that estimating mass on an individual level will be associated with even greater error.

Conclusions

Our results suggest that existing equations for estimating body mass from the postcranium need to be used more carefully that they typically have been. Most of the equations met the criteria for acceptance in the test sample, but the limits for acceptance were set very low and reliability dropped significantly when more realistic criteria were used. In addition, not all the equations performed equally well or within expectations given the sample characteristics. Several assumptions regarding the use of the equations were not fully supported. Specifically, the morphometric/STBIB equations are not more reliable than the mechanical/FHB equations, even when stature and bi-iliac breadth are measured directly. "Matched-target" equations do not consistently estimate mass better than equations designed for broader application. Newer equations for estimating mass from stature plus bi-iliac breadth are not generally better, or necessarily more appropriate for high latitude groups, than earlier equations. Although not consistent across all test groups, averaging the results of multiple FHB equations has a neutral or slightly positive effect on estimation accuracy. Sex-specific equations are not necessarily more accurate that equations derived for combined-sexes, but averaging the results of sex-specific equations does not significantly reduce estimation accuracy. Lastly, given the lenient acceptance criteria employed in the present study, the estimation accuracy for all the equations is likely to be even lower than that achieved here. Consequently, existing body mass estimates derived from these equations must be viewed cautiously.

Our results suggest that current postcranial body mass estimation methods need to be evaluated and applied more critically than has been the practice in biological anthropology to date. In order to do this, the issues that have been identified here must be resolved using large samples of individuals with matched biological information and skeletal data.

Acknowledgments We thank Michael Thali, Wolf Schweitzer, and the team at the Institute for Forensic Medicine in Zurich, Switzerland for data access and assistance. The comments of William Jungers, the main editor, and two anonymous reviewers greatly improved the article. For this project, ME was funded by a Social Sciences and Humanities Research Council Graduate Student Scholarship (#767-2009-1887 3) and Simon Fraser University. MC is funded by the Canada Research Chairs Program, the Canada Foundation for Innovation, the British Columbia Knowledge Development Fund, the Social Sciences and Humanities Research Council, and Simon Fraser University.

Grant sponsors Social Sciences and Humanities Research Council, Canada Research Chairs Program, Canada Foundation for Innovation, British Columbia Knowledge Development Fund, and Simon Fraser University.

References

Agostini GM, Ross AH (2011) The effect of weight on the femur: a cross-sectional analysis. J Forensic Sci 56:339–343

Aiello LC, Wood B (1994) Cranial variables as predictors of hominine body mass. Am J Phys Anthropol 95:409–426

Arsuaga J-L, Lorenzo C, Carretero J-M, Gracia A, Martinez I, Garcia N, Castro J-MB, Carbonell E (1999) A complete human pelvis from the Middle Pleistocene of Spain. Nature 399:255–258

Auerbach BM, Ruff CB (2004) Human body mass estimation: a comparison of "morphometric" and "mechanical" methods. Am J Phys Anthropol 125:331–342

Bayomi DJ, Tate RB (2008) Ability and accuracy of long-term weight recall by elderly males: the Manitoba follow-up study. Ann Epidemiol 18:36–42

Byard RW (2012) The complex spectrum of forensic issues arising from obesity. Forensic Sci Med Pathol 8:402–413

Cavalcanti M, Rocha S, Vannier M (2004) Craniofacial measurements based on 3D-CT volume rendering: implications for clinical applications. Dentomaxillofac Radiol 33:170–176

Churchill SE, Berger LR, Hartstone-Rose A, Zondo BH (2012) Body size in African middle Pleistocene *Homo*. In: Reynolds SC, Gallagher A (eds) African genesis: Perspectives on hominin evolution. Cambridge University Press, Cambridge, pp 319–346

Dagosto M, Terranova C (1992) Estimating the body size of Eocene primates: a comparison of results from dental and postcranial variables. Int J Primatol 13:307–344

Decker SJ, Davy-Jow SL, Ford JM, Hilbelink DR (2011) Virtual determination of sex: metric and nonmetric traits of the adult pelvis from 3D computed tomography models. J Forensic Sci 5:1107–1114

Elliott M, Kurki H, Weston DA, Collard M (2014) Estimating fossil hominin body mass from cranial variables: an assessment using CT data from modern humans of known body mass. Am J Phys Anthropol 154:201–214

Grine FE, Jungers WL, Tobias PV, Pearson OM (1995) Fossil *Homo* femur from Berg Aukas, northern Namibia. Am J Phys Anthropol 97:151–185

Hartwig-Scherer S (1993) Body weight prediction in early fossil hominids: towards a taxon-"independent" approach. Am J Phys Anthropol 92:17–36

Henneberg M, Stephan CN, Norris RM (2005) Sources of biological variation. Is sex really important? Am J Phys Anthropol 126:114

Holloway RL (1980) Within-species brain-body weight variability: a re-examination of the Danish data and other primate species. Am J Phys Anthropol 53(1):109–121

Jungers WL (1988) Relative joint size and hominoid locomotor adaptations with implications for the evolution of hominid bipedalism. J Hum Evol 17:247–265

Kim G, Jung HJ, Lee HJ, Lee JS, Koo S, Chang SH (2012) Accuracy and reliability of length measurements on three-dimensional computed tomography using open-source OsiriX software. J Digit Imaging 25: 486–491

Konigsberg LW, Hens SM, Jantz LM, Jungers WL (1998) Stature estimation and calibration: Bayesian and maximum likelihood perspectives in physical anthropology. Yearb Phys Anthropol 41:65–92

Kurki HK, Ginter JK, Stock JT, Pfeiffer S (2010) Body size estimation of small-bodied humans: Applicability of current methods. Am J Phys Anthropol 141:169–180

Lopes PML, Moreira CR, Perrella A, Antunes JL, Cavalcanti MGP (2008) 3-D volume rendering maxillofacial analysis of angular measurements by multislice CT. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 105:224–230

Lorkiewicz-Muszyńska D, Przystańska A, Kociemba W, Sroka A, Rewekant A, Zaba C, Paprzycki W (2013) Body mass estimation in modern population using anthropometric measurements from computed tomography. Forensic Sci Int 231:405, **e1-6**

McHenry HM (1992) Body size and proportions in early hominids. Am J Phys Anthropol 87:407–431

Melton N, Montgomery J, Knusel CJ, Batt C, Needham S, Pearson MP, Sheridan A, Heron C, Horsely T, Schmidt A, Evans A, Carter E, Edwards H, Hargreaves M, Janaway R, Lynnerup N, Northover P, O'Conner S, Ogden A, Taylor T, Wastling V, Wilson A (2010) Gristhorpe man: an early Bronze Age log-coffin burial scientifically defined. Antiquity 84:796–815

Moore MK, Schaefer E (2011) A comprehensive regression tree to estimate body weight from the skeleton. J Forensic Sci 56:1115–1122

Myszka A, Piontek J, Vancata A (2012) Body mass reconstruction on the basis of selected skeletal traits. Anthropol Anz 69:305–315

Niskanen M, Junno JA (2009) Estimation of African apes' body size from postcranial dimensions. Primates 50:211–220

Olivarius NF, Andreasen AH, Loken J (1997) Accuracy of 1-, 5- and 10-year body weight recall given in a standard questionnaire. Int J Obes Relat Metab Disord 21:67–71

Perry GS, Byers TE, Mokdad AH, Serdula MK, Williamson DF (1995) The validity of self-reports of past body weights by U.S. adults. Epidemiology 6:61–66

Plavcan JM (2012) Body size, size variation, and sexual size dimorphism in early *Homo*. Curr Anthropol 53:S409–S423

Pokines J, Symes SA, Roper C (2013) Manual of forensic taphonomy. CRC Press, Boca Raton

Pomeroy E, Stock JT (2012) Estimation of stature and body mass from the skeleton among coastal and mid-altitude Andean populations. Am J Phys Anthropol 147:264–279

Power ML, Schulkin J (2008) Sex differences in fat storage, fat metabolism, and the health risks from obesity: Possible evolutionary origins. Br J Nutr 99:931–940

Rainwater C, Cabo-Perez L, Symes S (2007) Body mass estimation and personal identification. Am J Phys Anthropol 132(S44):194–195

Rightmire GP (2004) Brain size and encephalization in early to mid-Pleistocene *Homo*. Am J Phys Anthropol 124:109–123

Robson SL, Wood B (2008) Hominin life history: Reconstruction and evolution. J Anat 212(4):394–425

Rosenberg KR, Zuné L, Ruff CB (2006) Body size, body proportions, and encephalization in a Middle Pleistocene archaic human from northern China. Proc Natl Acad Sci U S A 103:3552–3556

Ruff CB (1991) Climate and body shape in hominid evolution. J Hum Evol 21:81–105

Ruff CB (1994) Morphological adaptation to climate in modern and fossil hominids. Am J Phys Anthropol 37:65–107

Ruff CB (2000) Body mass prediction from skeletal frame size in elite athletes. Am J Phys Anthropol 113:507–517

Ruff CB (2002) Variation in human body size and shape. Annu Rev Anthropol 31:211–232

Ruff CB (2010) Body size and body shape in early hominins: Implications of the Gona pelvis. J Hum Evol 58:166–178

Ruff CB, Walker A (1993) Body size and body shape. In: Walker A, Leakey RE (eds) The Nariokotome *Homo erectus* skeleton. Harvard University Press, Cambridge, pp 234–265

Ruff CB, Scott WW, Liu AYC (1991) Articular and diaphyseal remodeling of the proximal femur with changes in body mass in adults. Am J Phys Anthropol 86:397–413

Ruff CB, Trinkaus E, Holliday TW (1997) Body mass and encephalization in Pleistocene *Homo*. Nature 387:173–176

Ruff CB, Niskanen M, Junno J-A, Jamison P (2005) Body mass prediction from stature and bi-iliac breadth in two high latitude populations, with application to earlier higher latitude humans. J Hum Evol 48:381–392

Ruff CB, Holt BM, Sládek V, Berner M, Murphy WA Jr, Zur Nedden D, Seidler H, Recheis W (2006) Body size, body proportions, and mobility in the Tyrolean "Iceman". J Hum Evol 51:91–101

Ruff CB, Holt BM, Niskanen M, Sladék V, Berner M, Garofalo E, Garvin HM, Hora M, Maijanen H, Niinimäki S, Salo K, Schuplierova E, Tompkins D (2012) Stature and body mass estimation from skeletal remains in the European Holocene. Am J Phys Anthropol 148:601–617

Scientific Working Group for Forensic Anthropology (SWGANTH) (2012) Stature estimation. http://swganth.startlogic.com/Stature%20Estimation%20Rev%201.pdf. Accessed 22 May 2015

Shen W, Punyanitya M, Wang ZM, Gallagher D, St. Onge MP (2004) Total body skeletal muscle and adipose tissue volumes: Estimation from a single abdominal cross-sectional image. J Appl Physiol 97: 2333–2338

Smith RJ (1996) Biology and body size in human evolution: statistical inference misapplied. Curr Anthropol 37:451–481

Smith RJ (2002) Estimation of body mass in paleontology. J Hum Evol 43:271–287

Smith RJ (2009) Use and misuse of the reduced major axis for line-fitting. Am J Phys Anthropol 140:476–486

Smyth AM, Viner MD, Conlogue GJ, Blyth T (2012) An evaluation of medical imaging techniques for craniometric data collection. Am J Phys Anthropol 147:274–274

Sokal RR, Rohlf FJ (2012) Biometry, 4th edn. WH Freeman & Co, New York

Steudel K (1980) New estimates of early hominid body size. Am J Phys Anthropol 52:63–70

Swiss Federal Statistical Office (2012) Population size and composition and factors influencing health. http://www.bfs.admin.ch/bfs/portal/en/index.html

Thali MJ, Yen K, Schweitzer W, Vock P, Boesch C, Ozdoba C, Schroth G, Ith M, Sonnenschein M, Doernhoefer T, Scheurer E, Plattner T, Dirnhofer R (2003) Virtopsy, a new imaging horizon in forensic pathology: Virtual autopsy by postmortem multislice computed tomography (MSCT) and magnetic resonance imaging (MRI): a feasibility study. J Forensic Sci 48:386–403

Thali MJ, Jackowski C, Oesterhelweg L, Ross SG, Dirnhofer R (2007) VIRTOPSY—the Swiss virtual autopsy approach. Legal Med 9: 100–104

Trinkaus E, Jelínek J (1997) Human remains from the Moravian Gravettian: the Dolní Věstonice 3 postcrania. J Hum Evol 33:33–82

Walker MJ, Ortega J, Parmovd K, Lopez MV, Trinkaus E (2011) Morphology, body proportions, and postcranial hypertrophy of a female Neandertal from the Sima de las Palomas, southeastern Spain. Proc Natl Acad Sci U S A 108:10087–10091

Wu G, Baraldo M, Furlanut M (1995) Calculating percentage prediction error: a user's note. Pharmacol Res 32:241–248