The repeatability of standard cranial measurements on dry bones and MSCT images

Jerković, Ivan; Bašić, Željana; Bareša, Tina; Krešić, Elvira; Hadžić, Anita Adamić; Dolić, Krešimir; Ćavar Borić, Marija; Budimir Mršić, Danijela; Čavka, Mislav; Šlaus, Mario; ...

Source / Izvornik: **Journal of Forensic Sciences, 2022, 67, 1938 - 1947**

Journal article, Accepted version
Rad u časopisu, Završna verzija rukopisa prihvaćena za objavljivanje (postprint)

https://doi.org/10.1111/1556-4029.15100

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:227:286651

The repeatability of standard cranial measurements on dry bones and MSCT images

Abstract

The present study examined if the cranial measurements from Data Collection Procedures for Forensic Skeletal Material 2.0 are repeatable when measured in dry bones and MSCT images and if the virtual measurements correspond to the physical ones. The sample included 33 dry crania imaged by MSCT. Two observers measured dry bones, two placed landmarks on 2D and 3D MSCT reconstructions, and one conducted measurements/landmarking on both media. One of the observers for each media repeated the measurements.

Technical and relative technical error of measurement (TEM and rTEM) and percentage differences were calculated to examine the repeatability of measurements and compare measuring modalities.

Intraobserver rTEM was above 1.5% for six bone measurements: FOB, ZOB, OBB, NLH, DKB, MDH (1.51%-4.87%) and for seven MSCT measurements: OBH, FOB, OBB, MDH, NLB, ZOB, DKB (1.57%-5.55%). The interobserver rTEM was above the acceptable level (>2%) for 11 measurements: PAC, NLH, OBB, EKB, MAL, FOB, NLB, OBH, ZOB, DKB, and MDH (2.01%-9.34%). The percentage differences were not systematically larger for measurements taken by the same user on both modalities than those obtained by different users on the same modality. When physical and MSCT measurements were tested on sex classification standards, the proportion of crania classified as male or female did not significantly differ ($P > 0.05$).

The study showed that physical and virtual cranial measurements could be interchangeable for developing or applying sex estimation standards. However, clarifications and adaptations are necessary for measurements of mastoid, nasal, and orbital regions that did not meet the standard criteria.

KEYWORDS: Osteometry, MSCT, Virtual Skeletal Collection, Intraobserver error, Interobserver error, Forensic Anthropology

Highlights:

- The study examined the repeatability of standard cranial measurements on dry bones and MSCT images.
- Intraobserver rTEM >1.5% were observed for six dry bone and seven MSCT measurements (1.5%-5.5%).
- Interobserver rTEM >2% were observed for 11 variables (2.01%-9.34%).
- Most nonrepeatable measurements were those of the orbital, nasal, and mastoid regions.
- Differences in dry bone and CT measurements did not affect sex estimation accuracy.

Introduction

The development of forensic anthropology methods was till recently dependable on the existence of well-documented skeletal collections (e.g., remains of known sex, age at death, cause of death, and other determinants of the biological profile). Most of these collections are curated in the European countries and the USA, and, although they have an enviable number of curated remains, their composition often does not represent well the populations of particular periods or could be imbalanced regarding the available anatomical parts (sometimes only skulls are curated) [1,2]. Also, the most recent period specimens (from forensic cases or recently donated bodies) are commonly underrepresented in the collections. These features of the curated collections restrict the forensic anthropology community from detecting secular changes and other population changes that often stem from the increasing population admixtures in globalization. The possible solution to this problem by creating virtual collections resulted in a new trend in modern forensic anthropology. The advantages of the virtual approach using the MSCT images are numerous: the age and sex of the patient are known, pathology as well, and the specimen can be reexamined numerous times without the danger of destruction. It can be examined in real-time by scientists in different parts of the world, and the regions of interest can be 3D printed or analyzed classically and used for teaching. Lastly, unlike dry specimens that usually allow direct linear measurements only, 3D methods enable the extension of data using the existing osteological landmarks and more advanced analyses. Such benefits contributed to more dynamic advances in forensic anthropology. Most importantly, for the first time in history, it is possible to build virtual skeletal collections of every population. Therefore, it enables us to test and validate methods for sex and age estimation as well as to design the new standards for biological profiling.

Among different variables of the biological profile construction, osteometry plays a crucial role in estimating ancestry, sex, and stature. Although various skeletal measurements have been developed, they usually follow the Data Collection Procedures for Forensic Skeletal Material 2.0(DCP 2.0), modified in 2016 [3] after the previous version was used for more than two decades [4]. The DCP 2.0 standards provide detailed descriptions of landmarks and measurements but also error rates on intra- and interobserver levels. In DCP updates, some of the landmarks in the skull region were removed (alare, nasospinale), some added (asterion, mastoidale, porion, zygomaxillare anterior, zygoorbitale), and some replaced (auriculare replaced with radiculare). For some, definitions changed (basion, dacryon, prosthion), expanded (bregma, ectoconchion, lambda), or clarified (nasion). Considering the cranial

measurements, some of them changed (BBH, BNL, BPL, NLB, MDH), some of them were added (NOL, ASB, ZMB, ZOB), and some of them underwent changes in the definition of the landmark (AUB, NPH, NLH, OBB, DKB, FOL) [3,5]. Although the modifications were necessary as the definitions of landmarks were vague or subjective or showed larger interobserver and intraobserver errors [6,7], the updated standards opened some new issues in current research. First, previously measured and documented collections will probably have to be remeasured as comparing populations and implementing updated standards may not be possible without the amendments to those measurements. On the other hand, we would need to validate those measurements, not only for physical bones but also for comparison of virtual and physical measurements. To be more precise, if transferring from physical to virtual collections for the implementation and development of osteometric standards for ancestry, sex, and stature estimation, we need to examine if physical measurements correspond to the virtual ones and if the standards created on virtual measurements apply to the actual bones.

The accuracy of translating the virtual to physical model and vice versa has been studied previously on some bones or skeletal structures, but mostly in the form of pilot studies. One study, for example, has shown that the combination of segmentation and landmark recognition errors can be substantial and that it is questionable if MSCTs can be a good substitute for physical skeletal collections [8]. Colman et al. tested the virtual bone modeling precision per polygon mesh point on CT scans reconstructions of one cadaver. They discovered that more than 97% of pelvic locations showed point-to-point distance variation of less than 2 mm (CI = 95%) and 91% less than 1 mm [9]. Corron et al. tested the physical measurements of clavicles and the CT scans of the same bones and reported differences smaller than ±1 mm and showed more than 95% reliability [10]. Carew et al. tested three human bones and their 3D virtual and 3D printed models (n=6). They found the mean differences ranging from -0.4% to 12.0%, and interobserver error ranging from -5.3% to 0.7% [11]. Another pilot study comparing physical and virtual bone measurements on four human cadavers showed that the measurement errors were comparable in both methods and could be used to gather the population-relevant osteometric data from individuals of known sex and age [12]. Simmons-Ehrhardt et al. [13] conducted a large-scale study that examined the sample of 303 people from three ancestry groups: African, Asian, and European. Although they did redefine some measurements, they showed that using CT measurements in methods derived from dry-bone specimens did not decrease accuracy in ancestry

classification. They also proposed further studies to evaluate differences between the CT and dry bone measurements and standardize and validate CT data [13].

As previously stated, most of these studies were pilot studies with a smaller number of specimens and used different standards for measuring bones. Since the revised osteometric standards were published in 2016 [3], a study testing both repeatability of skull measurements on physical and virtual models and the applicability of standards derived from virtual measurements on physical ones has not been yet conducted.

Therefore, the first aim of this study was to test the repeatability of standard measurements defined by DCP 2.0 [3] on virtual and physical crania. The second aim was the comparison of virtual and physical measurements to test the reliability of the MSCT measurements. In the study, we examined two hypotheses:

H1: Cranial measurements show an acceptable degree of measurement error across different means (dry bones vs. CT) and observers.

H2: Measurements conducted on CT scans correspond to those taken on dry crania.

Materials and Methods

Materials

The sample included 33 dry crania (20 male, 13 female) from the Early Medieval site (7^{th} - 9^{th} century) Velim-Veliš tak, located near Benkovac, southern Croatia [14]. The sex and age composition of the crania is enlisted in Table 1.

Dry crania were scanned at the University Hospital Dubrava, Department of Diagnostic and Interventional Radiology, Zagreb, Croatia, with MSCT device Sensation 16, Siemens AG Medical Solutions, Erlangen, Germany. Scanning parameters were 120 kV and 320 mA, respectively, with isometric slices using 16 × 0.75 mm collimation. Images were reconstructed to the same slice thickness with a soft and bone tissue convolution kernel.

Settings

Four observers took part in the study. O1 has 5-year experience, and O2 has 4-year experience in the biological anthropology laboratory, where they have been conducting osteometric methods. O3 has 7-year experience with osteometric analyses on dry bones and

MSCT images, while O4, as a radiology resident (5 years), is experienced with virtual measuring techniques. Two observers (Observer 1 and 2, O1 and O2) measured dry bones, while the other two (Observer 3 and 4, O3 and O4) placed landmarks on scanned files. All observers conducted the first round of measurements, while O1 and O3 also conducted the second round. Finally, one of the dry bone observers (O2) conducted an additional round of measurements on CT scans. Twenty-eight cranial measurements were taken according to Data Collection Procedures for Forensic Skeletal Material 2.0 (4) using a spreading caliper and digital sliding caliper (Table 2). Measurements on MSCT images were calculated as distances between cranial landmarks (interlandmark distances).

Virtual cranial measurements

DICOM files obtained by scanning were imported into Stratovan Checkpoint Software Version 2020.10.13.0859 (Stratovan Corporation, Davis, CA) [15]. Bone-tissue reconstructions were loaded, and files were viewed in 2D (axial, coronal, and sagittal view) and 3D using semi-transparent 3D volume rendering. Each observer first checked the alignment in 2D views for each specimen and adjusted it in midsagittal and coronal planes where necessary. Alignments were done considering the line fitting occipital protuberance, the middle of the sella turcica and nasion, and the line connecting internal auditory meati. In the next step, observers loaded the template landmarks that were reordered compared to the original data collection procedures to enable greater precision and speed of work (Supplementary Table 1).

Midsagittal landmarks were initially placed in sagittal view, and their position was checked in 3D and other 2D reconstructions (Figure 1). This included glabella (g), opisthocranion (op), nasion (n), opisthion (o), alveolon (alv), bregma (b), lambda (l), and basion (b). The endobasion (Eba) was added to correspond to the Foramen Magnum Length (FOL) description. Instrumentally determined landmarks perpendicular to the midsagittal plane, euryon (eu), zygion (zy), and ectomolare (ecm), were detected using the following procedure (Figure 2). When viewing the cranium in 3D view, the sagittal plane was moved in the same window to find a point where the plane comes in first contact with the anatomical part of interest (i.e., most lateral point from the midsagittal plane). The axial plane was then shifted through the 3D model until the same point was reached. Lastly, the second point was marked on the opposite side, in the plane parallel to the coronal plane.

The specific workflow was also defined for several landmarks. Landmarks radiculare (ra) were placed in axial and coronal views. Since asterion (ast) can be more of a complex structure, all three planes were intersected on the 3D model to find the meeting point of the temporal, parietal, and occipital bones, and the landmark was added in the coronal plane. Prosthion (pr) was placed in sagittal view. The lowest point on the inferior border of the nasal aperture (not named in DCP 2.0) was marked in 3D and controlled in the sagittal view. Unnamed landmarks that define nasal breadth were identified by moving the axial plane through the 3D model. When most lateral points were detected, landmarks were placed in axial view. The same approach was used to define the lateral margins of the foramen magnum that define Foramen Magnum Breadth. Landmarks defined by zygomaxillary sutures: zygomaxillare anterior (zma) and zygoorbitale (zo) were defined in axial view and simultaneously controlled on the 3D model. Landmarks frontotemporale (ft) were defined directly on a 3D model and controlled in 2D. To identify porion (po), the sagittal plane was moved through the 3D model until the margin of the external acoustic meatus was reached. Then, the coronal plane was moved until the superior point was found, and the landmark was marked in sagittal view. To identify mastoidale (ms), the most inferior point on the mastoid process was first identified by the first axial slice where it becomes visible. The landmark was then defined when corresponding sagittal and coronal slices with the most inferior point were located.

To find ectoconchion (ec), the axial plane was first moved through the 3D model/coronal view and rotated until the plane was aligned with the superior orbital border. This plane was lowered until it bisected the orbit into two halves. The landmark was finally marked in axial view. Unnamed landmarks on superior and inferior orbital borders were located by moving the sagittal plane on a 3D model until it bisected the orbit into equal medial and lateral halves. The landmarks were finally placed in sagittal view.

Landmarks collected on each specimen were exported as .nts files and loaded into R (version 3.6.2), and Rstudio (version 1.2.5033) using geomorph package [16]. The same package was then used to define 31 interlandmark distances as standard cranial measurements.

Statistical analysis

The normality of data was tested using the D'Agostino-Pearson test. Intraobserver differences were assessed using a paired samples t-test or Wilcoxon test for non-normal distribution data.

Interobserver differences were tested using the repeated measurements analysis of variance (ANOVA) or Friedman test when test assumptions were not met. For measurements that showed significant differences in ANOVA, pairwise comparison with a Bonferroni correction was used to compare differences between individual observers.

Technical error of measurements (TEM) and relative technical error of measurements (rTEM) were calculated to test intra and intraobserver variability. To assess if the error level is acceptable, we used criteria employed by Langley et al. [6], where rTEM values ≤ 1.5% were considered acceptable for intraobserver error, while values ≤ 2 were acceptable for interobserver error.

TEM was calculated using the following equation [6]:

$$TEM = \sqrt{\frac{\sum_1^N \left[\sum_1^K M(n)^2 - \frac{(\sum_1^K M(n))^2}{K}\right]}{N(K-1)}}$$

where N is the number of specimens (N=33), K is the number of observers (K=4), and M is a measured value. Relative technical error of measurement is then obtained by dividing the obtained TEM value by the mean value of measurements of all the observers and multiplying it by 100 [6].

To compare differences between measuring modalities, we calculated percent differences [8] for mean values of the following combinations: differences between the same observer (O2) in physical and MSCT measurements, differences between the O2 and O1 in dry bone measurements, and differences between O2 and O3 in MSCT measurements.

Finally, to assess the practical implications of differences in measurements obtained by different modalities in sex estimation, we tested sex estimation standards developed for the medieval Croatian population [7] on dry bone and CT measurements of the same observer (O2). This included nine univariate and two multivariate discriminant functions. Sex estimation accuracy was calculated for each measuring modality of O2, and proportions of skulls classified as male or female in both groups were compared using a McNemar's Chi-squared test or Exact McNemar's test.

Statistical analyses were conducted in R Studio and MedCalc Statistical Software version 19.2.6 (MedCalc Software Ltd, Ostend, Belgium; https://www.medcalc.org; 2020) with a level of statistical significance set at $P \leq 0.05$.

Results

Intraobserver error on dry bones and MSCT images

Mean differences between the first and the second round of dry bone measurements (O1) ranged from -0.38 to 0.72. For the first and second rounds of CT measurements (O3), the mean differences were between -0.35 and 0.44. Statistically significant differences between the first and the second rounds were detected for OBB (R), EKB, DKB, and ZOB in dry bone and for GOL, BPL, and OBB (R) in MSCT measurements (Supplementary Table 2).

The average intraobserver TEM for dry bone measurements was 0.65 mm (sd = 0.27 mm). TEM ranged from 0.28 mm to 1.36 mm. The highest level of error of measurements (TEM > 1 mm) was noted for MDH, PAC, and BBH. The average relative TEM was 1.15%, with a standard deviation of 0.97%, while rTEM values ranged from 0.27% to 4.87%. rTEM above acceptable levels was noted for eight variables (six measurements): FOB, ZOB, left and right OBB, NLH, DKB, and left and right MDH (Supplementary Table 3).

For CT measurements, the average intraobserver TEM was 0.67 mm (sd = 0.26 mm) and ranged between 0.37 mm and 1.47 mm. The highest TEM (> 1 mm) was observed for ZMB, PAC, DKB, and ZOB. RTEM ranged from 0.22% to 5.55% with mean value of 1.25% (sd = 1.07). Ten variables (7 measurements) achieved rTEM higher than acceptable: OBH, FOB, OBB, MDH, NLB, ZOB, and DKB (Supplementary Table 3).

Interobserver error across means and observers

The first round of all observers (two dry bone and two MSCT observers) was used to assess interobserver differences. Supplementary Table 4 shows the results of the comparison between the first round of measurements that included all the observers. The differences were statistically significant for 21 of 31 variables. It included length measurements like NOL and chord and basion-based measurements (BBH and BNL); breadth measurements like XCB, ZYG, WFB, and UFBR; nasal, mastoidal, and foramen magnum measurements; and three orbital measurements (EKB, DKB, and ZOB). Supplementary Table 5 demonstrates the results of pairwise comparisons of measurements that previously showed significant differences among observers. Differences that were significant were grouped into three categories: differences between dry bone observers only, observer-specific differences regardless of the media, and differences between dry bone and CT measurements. According

to the results, errors that occurred were sporadic rather than systematic, i. e., one observer or one measuring means did not show directional discrepancies.

Measurements showing significant differences between dry bone observers were NOL and AUB. In those cases, measurements of O1 were greater than those of O2.

For ZYG, differences were significant only between O2 and all others. The mean difference measured by O2 and other observers ranged between -0.778 and -1.167. For BBH and WFB, significant differences were not observed only between the O2 and O3, meaning that O1 and O4 showed differences between each other and all other observers. Measurements taken by O1 were, on average greater than those of others, while measurements taken by O4 were smaller than those of others. BNL showed differences between O1 and other observers, where O1 measurements were greater. A significant difference was also observed between O2 and O3, with greater values of O2. UFBR showed significant differences between O1 and the other two observers (O2 and O4), as well as between O2 and the other two observers (O1 and O3).

NLH, EKB, and PAC showed differences in all combinations except for dry bone observers. Measurements of CT observers (O3 and O4) were greater than those of O1 and O2, and measurements of O4 were greater than O3. FRC and OCC showed significant differences between O4 and all other observers and between O1 and O2. Measurements of O4 were smaller than those of other observers, while measurements by O1 were greater than those of O2.

XCB, NLB, and MDH were significantly greater when measured by CT observers than in dry bones. DKB was significantly larger in O1 than in O3. Foramen magnum measurements showed differences between CT observers (O3 and O4) and O2 and differences between O1 and O4. ASB showed significant differences between O1 and CT observers (O3 and O4), while ZOB differences were also significant between O1 and CT observers (O3 and O4) and between O2 and O4.

The average interobserver TEM was 1.31 mm (sd = 0.56 mm), with minimum and maximum values between 0.60 mm and 2.77 mm (Supplementary Table 6). MDH measurement had a maximum error level, while PAC, ZOB, and ECB also achieved TEM > 2 mm (Figure 4). Relative TEM ranged between 0.35% and 9.34, with a mean value of 2.36% (sd = 1.87). A total of 14 variables (11 measurements) had an rTEM above the acceptable limit (2%). This included mastoid measurements (MDH), orbital measurements (DKB, ZOB, OBB, OBH,

EKB), nasal measurements (NLH, NLB), and MAL, while PAC was around threshold level (2.01%).

Differences between dry bone and MSCT measurements

When the same observer (O2) measured specimens using both modalities, the average percentage differences between dry bone and MSCT measurements were -0.82%. When dry bone measurements of O2 were compared to dry bone measurements of O1, this difference was -0.58%. The same comparison on MSCT measurements (O2 vs. O3) revealed an average difference of -0.52% (Supplementary Table 7). Although average differences were greatest for bones vs. CT comparison and smallest for CT vs. CT comparison, comparison of percentage differences according to individual measurements did not show that only one modality contributes to the variation in measurements.

Practical implications of differences between dry bone and CT measurements in sex estimation

When the dry bone and CT measurements of O2 were included in nine univariate and two multivariate discriminant functions developed for the medieval Croatian population, no statistically significant differences were found in the proportions of males and females obtained by physical and CT measurements (Table 3). In univariate functions, the accuracy of sex estimated using physical and CT measurements was the same for 3/9 measurements, while for other variables, slight differences were detected. Both multivariate functions (F1 and F2) showed no differences in sex estimation accuracy on dry bones and CT images.

Discussion

The present study showed that most standard cranial measurements from DCP 2.0 were sufficiently repeatable when measured on dry bones or MSCT images. Despite some variations detected when different measuring modalities were used, the study showed that such differences had no practical significance when applying measurements to estimate sex. Therefore, skeletal measurements from MSCT images could be a valid source for developing standards for craniometric sex or population affinity estimations. However, results suggest that several measurements like those of mastoid, orbital, and nasal region should be additionally analyzed and adapted as they showed a degree of error higher than acceptable by current standards [6].

Mean differences for repeated measurements of an observer that measured dry bones and an observer that used MSCT images were small in both cases (-0.38 mm–0.72 mm and -0.35 mm–0.44 mm). Such differences were also not statistically significant except for OBB, EKB, DKB, and ZOB for dry bone; and GOL, BPL, and OBB for MSCT measurements.

Average TEM and rTEM values on dry bone and CT measurements achieved a relatively low intraobserver error level, with only slightly higher values in CT measurements (0.65 mm vs. 0.67 mm; 1.15% vs. 1.25%). Unacceptable rTEM was detected in both measuring modalities for FOB, ZOB, OBB, DKB, and MDH. For dry bone measurements, NLH, and for CT measurements, OBH and NLB also had relative errors above 1.5%.

Mastoid measurements had the highest TEM and rTEM on dry bones (1.36 mm and 4.87% for the right side), which could also be a consequence of the small average size of this variable. This confirms around-average TEM values on CT measurements (0.76 mm) which provide unacceptable rTEM (2.42%). However, differences between the two measuring modalities might imply that it is easier to identify porion and mastoidale in radiological views than using a standard caliper. DKB had an rTEM of 2.45% on dry bones, despite having below-average TEM (0.59 mm). As DKB was on average the smallest (24.3 mm for the first round), this finding could suggest that a 1.5% threshold is not realistic to achieve using standard measuring tools. On CT, this measurement demonstrated above-average TEM (1.29 mm) and the greatest rTEM of 5.5%. Except for its small dimension, rTEM in this case could have been influenced by the lower visibility of dacryon on MSCT images. Intraobserver rTEM was also highest for those two variables in the original standardization study by Langley et al. [6], where MDH and DKB were the only cranial variables that did not meet the criteria. Other studies also detected above threshold intraobserver error for those measurements [17,18].

Higher error in OBB measurements partly stems from measurement size, but the precision on CT measurements could have also been affected by difficulties in finding dacryon. Similar problems when locating landmarks, especially on MSCT images, are probably also a major source of error for ZOB. FOB also did not meet rTEM criteria despite having below-average TEM values (0.46 mm and 0.51 mm), suggesting that the intraobserver threshold might not be appropriate for such cases. The size can be a source of error of NLB on CT, considering that NLB had the second smallest value. Errors in OBH for CT measurements could be directly related to the errors in OBB measurement as the OBH is measured perpendicular to

the OBB. A slightly higher error in NLH was probably caused by inconsistencies in identifying landmarks on the border of the nasal aperture because other nasion-based measurements performed well. Previous research also supports present findings, e.g., the study based on MSCT measurements among 12 common cranial variables found OBB, OBH, DKB, and FOB to have intraobserver errors above 1.5% [18].

When considering interobserver variabilities for the first round of measurements (two dry bone and two MSCT observers), we detected statistically significant differences for 21 of 31 measurements. However, pairwise comparisons showed no observable patterns in the errors by either observer or modality, so it was impossible to systematically attribute them to the modality (dry bone/CT) or specific observer (his/her experience or background).

Interobserver technical error calculated for all observers from the first round yielded about two times higher average TEM (1.33 mm vs. 0.65 and 0.67 mm) and rTEM (2.36% vs. 1.15 and 1.25%) than for intraobserver errors. Most measurements that demonstrated interobserver rTEM higher than acceptable (>2 %) were those previously identified in intraobserver errors. This included MDH, DKB, ZOB, OBB, OBH, EKB, NLH, NLB, MAL, and PAC.

Mastoid Length provided the highest degree of error (9.34% and 6.17%), 2-3 times higher than in previous studies in dry bones [6,17] and on CT images [19]. This error was probably caused by several factors, including small size and high intraobserver inconsistencies that were additionally increased when multiple observers and two measuring modalities were used. This implies that the attempt to define the landmark in revised standards better did not have an effect and that additional modifications are necessary. This was the reason for proposing a new measurement definition by Langley et al. [6]. However, original [6] and further studies that compared old and new measurement definitions found that they had an unaccepted degree of error [17].

All orbital measurements (DKB, ZOB, OBB, OBH, EKB) also yielded unacceptable rTEM errors ranging from 2.72% for OBB to 4.17% for DKB. Those errors are also partly the product of their smaller size (e.g., DKB and OBH), but inconsistencies in identifying dacryon probably contributed to a greater extent, primarily since measurements like OBB and OBH depend on each other. There is also a potential contribution of the updated ectoconchion definition, which requires tracing the superior orbital border until finding a point on the lateral orbital border that bisects orbits into equal parts. Considering inconsistencies detected on the intraobserver level, such measurements and landmarks could also be modified,

particularly for the virtual environment. For example, the study by Simmons-Ehrhardt et al. [13] reported the inability to detect dacryon on MSCT images. They reintroduced maxillofrontalle, defined by White [20], which is undoubtedly a landmark easier to locate. Ectoconchion could also be easily redefined and standardized, especially in a virtual environment. For example, instead of tracing the superior orbital border, the axial plane could be moved through the orbits until bisecting them into equal parts. Then, landmarks could be placed and checked in the coronal and sagittal planes. Zygoorbitale breadth also had rTEM above 2%. This could result from poor landmark visibility, which is probably more pronounced on MSCT measurements demonstrating higher intraobserver error. Visibility is necessary not only related to the measuring modality but can also be affected when zygomaxillary suture cannot be traced appropriately and when the landmark can be confused for anatomical variations like infraorbital sutures.

Both nasal measurements also had unacceptable rTEM. Higher rTEM was pronounced for NLB (3.47%), although TEM was below average (0.85 mm). So, measurement size probably mainly contributed to the error, while some instrumental restrictions, users' inconsistencies, and differences in measuring modalities might be responsible for the rest of the variability. For example, in a virtual environment, small measurements and the inability to detect a "sharp border" could be affected by minor adjustments during window leveling. The second measurement, nasal height, was not so influenced by the measuring scale, but probably by other mentioned factors and possible difficulties in detecting inferior border on the floor of the nasal cavity, as other nasion-based measurements performed well. Maxillo-Alveolar Length also provided high interobserver rTEM (3.09%), but considering the extremely low sample size (n = 12), this result should be considered cautiously.

Among the remaining measurements, only PAC had rTEM slightly above the threshold (2.01%). Although rTEM was not greatly affected due to its larger size, the overall interobserver TEM (2.64 mm) was the third greatest. This error can be attributed to multiple factors. First, except for suture visibility, sutural variations can sometimes be complex, especially around lambda. This could be more pronounced in present research on the interobserver level due to the differences in experience and professions but also due to the measuring modalities. Similar findings were also reported for intraobserver errors by Simmons-Ehrhardt et al. [13], where chord variables revealed the highest TEMs. Except for landmark visibility issues, the authors have attributed this error to the specific position of landmarks that could be displaced during the 3D reconstructions and processing. However,

since basion-based measurements performed relatively well in our study, it is possible that methodology which combined standard 2D views and volume rendering had a better result than just 3D reconstructions.

Since the study detected various degrees of discrepancies in cranial measurements, we directly compared the differences between dry bone and MSCT measurements, demonstrating that measuring modality was not the primary source of error. This was done by showing that the percent differences of cranial measurements between modalities did not systematically exceed the differences between CT or dry bone measurements. Instead, average percent differences in CT vs. CT (0.52%) and dry bone vs. dry bone (0.58%) comparisons were almost the same, while the average intermodality difference was only slightly higher (0.82%). Differences in individual measurement also supported the same observation, as for some measurements percentage differences were largest for CT vs. dry bone, for some for dry bone vs. dry bone, and for some for CT vs. CT comparison. To further examine the impact of these differences, we examined their practical implications by testing dry bone and MSCT measurements on discriminant functions for sex classification. Accuracies of sex estimation showed slight or no differences, and proportions of crania classified as male or female did not significantly differ. Moreover, no differences were detected when applying the multivariate discriminant function, which would be the most realistic case in practice considering the limited importance of single cranial variables in sex estimation [21]. These findings also concur with Simmons-Ehrhardt et al. [13], that tested the craniometric data from medical CT scans on previously developed dry bone standards for ancestry and sex estimation and showed that some dry bone-CT differences detected in the study also did not hamper the craniometric biological profiling.

The present study is one of the relatively larger sample size studies that examined the repeatability of cranial measurements on dry bones and in the virtual environment. The measurements' repeatability was previously examined separately or as a part of the study’s internal validity. Such studies were often restricted to the limited number of specimens [9,12,13,22,23]. More importantly, the present study was the first to test the cranial landmarks and measurements presented in DCP 2.0 [3] both on dry bones and MSCT images. The specific protocol was developed to mark cranial landmarks in the virtual environment using the simultaneously 3D volume-rendered model and multiplanar reconstructions. Such a combination has rarely been employed (e. g., in [13] for instrumentally determined measurements), while many previous studies were mainly based on 3D model examinations

[18,19,23] that can be less precise than working with standard radiological planes. The proposed approach met the current standards, except in regions more prone to inconsistencies in dry bone measurements. The last part of the analysis showed that most of the observed differences could not be attributed to the specific modality, indicating that they could result from multiple sources like observer experience and background, but also definitions of cranial landmarks and measurements. Most importantly, no statistical or practical significance was observed when using MSCT instead of dry bone measurements for craniometric sex estimation. Therefore, we showed that, in present setting, cranial measurements obtained from MSCT images could be used for developing craniometric standards for biological profiling.

The present study has several drawbacks that may limit the generalizability of the results. Firstly, despite a relatively large sample size for such study type, not all crania were preserved entirely. So, results for some measurements like MAB and MAL (with n = 12) could not be completely representative. Sample size differences seen in physical and virtual measurements could be mainly caused by the damage to the crania that occurred after imaging. This damage resulted from the magnitude 5.5 earthquake in 2019 [24] that caused direct damage to the crania as well as the consequent transportation of material to a more adequate and safe place. The second limitation could stem from the study design in which different observers, and observers of different profiles and experiences, conducted physical and CT measurements. It was, therefore, not possible to definitely interpret the sources of each error, but their levels could be indicative. The present design actually represents the real-life situation where professionals and researchers of different profiles and experiences would analyze human skeletal material, some using imaging technologies and some dry bones, where measurement and landmark definitions should be clear for all those trained in human anatomy and osteology. Lastly, the present study used images of dry skeletal remains, which might not completely represent the situation where medical scans of living patients are used. In such cases, working with soft tissues might result in some differences, and the process of thresholding and segmentation might be more difficult. The previous research that compared pelvic images of clinical and dry bone CT 3D models and optical 3D models as a gold standard showed that differences were more pronounced in clinical CTs than in dry CT models [8]. The study also showed that differences between modalities were larger than differences between observers, thus indicating a need for such study type for different skeletal elements. Still, unlike our study, the named study [8] and many previous ones

[18,19,23] employed 3D models only, ignoring the fact that such volume or surface rendering methods are usually auxiliary means in radiology [25] and that differences in reconstructions algorithms and software could be an additional source of error [22]. Anthropologists, mainly used to working with physical bones that correspond to 3D models, should therefore receive additional training required for operating with standard radiological views.

Overall results showed that some measurements might require redefinition and additional adaptations for the virtual environment. Still, some small-scaled measurements showed below-average TEM and above 1.5% intra- and/or 2% interobserver rTEM, which are currently set as the standard for the highest acceptable error level. This might imply that those rTEM limits are not realistic in all cases and could bring into question the convenience of these criteria.

Future studies should also be conducted to completely standardize standard osteological measurements in virtual environments. This would require different studies and perspectives to consider all sources of inconsistencies, starting from sample types (dry bone or clinical CTs), variations in imaging parameters and reconstructions, software differences, and user-specific preferences (e. g. 2D or 3D views, thresholding, etc.). Although there should be future research and discussion of this topic, the results of our study imply that in the present setting, virtual collections could be used as population-referent data that could replace the gaps of physical collections adequately.

References

1. Forensic Anthropology Society of Europe. Map of identified osteological collections. http://forensicanthropology.eu/osteological-collections/#page-content. Accessed November 2, 2022.

2. Petaros A, Caplova Z, Verna E, Adalian P, Baccino E, de Boer HH, et al. The Forensic Anthropology Society of Europe (FASE) Map of Identified Osteological Collections. Forensic Sci Int. 2021;328:110995. doi: 10.1016/j.forsciint.2021.110995.

3. Langley NR, Jantz LM, Ousley SD, Jantz RL, Milner G. Data collection procedures for forensic skeletal material 2.0. Knoxville, TN: Forensic Anthropology Center, Department of Anthropology, University of Tennesse, 2016; 107.

4. Moore-Jansen PH, Jantz RL, Ousley SD. Data collection procedures for forensic skeletal material. Knoxville, TN: Forensic Anthropology Center, Department of Anthropology, University of Tennessee, 1994. 1–78.

5. Langley NR, Jantz LM, McNulty S, Maijanen H, Ousley SD, Jantz RL. Data for validation of osteometric methods in forensic anthropology. Data Br. 2018;19:21–8. doi: 10.1016/j.dib.2018.04.148.

6. Langley NR, Meadows Jantz L, McNulty S, Maijanen H, Ousley SD, Jantz RL. Error quantification of osteometric data in forensic anthropology. Forensic Sci Int. 2018;287:183–9. doi: 10.1016/j.forsciint.2018.04.004.

7. Bašić Ž. Determination of anthropological measurements and their ratios that are significant for sex determination on skeletal remains from medieval population of Eastern Adriatic Coast [dissertation]. Split, Croatia: University of Split, School of Medicine, 2015.

8. Colman KL, de Boer HH, Dobbe JGG, Liberton NPTJ, Stull KE, van Eijnatten M, et al. Virtual forensic anthropology: The accuracy of osteometric analysis of 3D bone models derived from clinical computed tomography (CT) scans. Forensic Sci Int. 2019;304:1–18. doi: 10.1016/j.forsciint.2019.109963.

9. Colman KL, Dobbe JGG, Stull KE, Ruijter JM, Oostra RJ, van Rijn RR, et al. The geometrical precision of virtual bone models derived from clinical computed tomography data for forensic anthropology. Int J Legal Med. 2017;131(4):1155–63. doi: 10.1007/s00414-017-1548-z.

10. Corron L, Marchal F, Condemi S, Chaumoître K, Adalian P. Evaluating the consistency, repeatability, and reproducibility of osteometric data on dry bone surfaces, scanned dry bone surfaces, and scanned bone surfaces obtained from living individuals. Bull Mem Soc Anthropol Paris. 2017;29(1–2):33–53. doi: 10.1007/s13219-016-0172-7.

11. Carew RM, Morgan RM, Rando C. A Preliminary Investigation into the Accuracy of 3D Modeling and 3D Printing in Forensic Anthropology Evidence Reconstruction. J Forensic Sci. 2019;64(2):342–52. doi: 10.1111/1556-4029.13917.

12. Verhoff MA, Ramsthaler F, Krähahn J, Deml U, Gille RJ, Grabherr S, et al. Digital forensic osteology-Possibilities in cooperation with the Virtopsy® project. Forensic Sci Int 2008;174(2–3):152–6. doi:10.1016/j.forsciint.2007.03.017.

13. Simmons-Ehrhardt TL, Ehrhardt CJ, Monson KL. Evaluation of the suitability of cranial measurements obtained from surface-rendered CT scans of living people for estimating sex and ancestry. J Forensic Radiol Imaging. 2019;19:100338. doi: 10.1016/j.jofri.2019.100338.

14. Jurić R. Ranosrednjovjekovno groblje u Velimu kod Benkovca [Early Medieval graveyard in Velim near Benkovac]. Obavijesti HAD–a 2004;36:20.

15. Stratovan Corporation. Stratovan Checkpoint [Software]. Version 2018.08.07. Aug 07, 2018. URL: https://www.stratovan.com/products/checkpoint.

16. Adams DC, Otárola-Castillo E. Geomorph: an R package for the collection and analysis of geometric morphometric shape data. Methods Ecol Evol. 2013;4(4):393–9. doi: 10.1111/2041-210X.12035.

17. Liebenberg L, Krüger GC. Standardization and quality assurance in skeletal landmark placement and osteometry. Forensic Sci Int. 2020;308. doi: 10.1016/j.forsciint.2020.110168.

18. Kranioti EF, García-Donas JG, Can IO, Ekizoglu O. Ancestry estimation of three Mediterranean populations based on cranial metrics. Forensic Sci Int. 2018;286:265.e1-265.e8. doi: 10.1016/j.forsciint.2018.02.014.

19. Franklin D, Cardini A, Flavel A, Kuliukas A, Marks MK, Hart R, et al. Concordance of traditional osteometric and volume-rendered MSCT interlandmark cranial measurements. Int J Legal Med. 2013;127(2):505–20. doi:10.1007/s00414-012-0772-9.

20. White TD, Black MT, Folkens PA. Human osteology. 3rd edn. London, U.K.:Academic Press; 2012. p. 58

21. Spradley MK, Jantz RL. Sex Estimation in Forensic Anthropology: Skull Versus Postcranial Elements. J Forensic Sci. 2011;56(2):289–96. doi: 10.1111/j.1556-4029.2010.01635.x.

22. Guyomarc'h P, Santos F, Dutailly B, Desbarats P, Bou C, Coqueugniot H. Three-dimensional computer-assisted craniometrics: A comparison of the uncertainty in measurement induced by surface reconstruction performed by two computer programs. Forensic Sci Int. 2012;219(1–3):221–7. doi: 10.1016/j.forsciint.2012.01.008.

23. Toneva D, Nikolova S, Agre G, Zlatareva D, Hadjidekov V, Lazarov N. Machine learning approaches for sex estimation using cranial measurements. Int J Legal Med. 2021;135(3):951–66. doi: 10.1007/s00414-020-02460-4.

24. Markušić S, Stanko D, Korbar T, Belić N, Penava D, Kordić B. The Zagreb (Croatia) M5 . 5 Earthquake on. Geosciences. 2020;10(7):252. doi: 10.3390/geosciences10070252.

25. Ebert LC, Franckenberg S, Sieberth T, Schweitzer W, Thali M, Ford J, et al. A review of visualization techniques of post-mortem computed tomography data for forensic death investigations. Int J Legal Med. 2021;135(5):1855–67. doi: 10.1007/s00414-021-02581-4.

TABLES

TABLE 1 Sex and age distribution of the sample

Age group	Males	Females	Total
18-35	7	4	11
35-50	10	6	16
50+	3	3	6
Total	20	13	33

TABLE 2 Standard cranial measurements included in the study

Abbreviation	Measurement	Abbreviation	Measurement
GOL	Maximum Cranial Length	NLB	Nasal Breadth
NOL	Nasio-occipital length	OBB	Orbital Breadth
XCB	Maximum Cranial Breadth	OBH	Orbital Height
ZYB	Bizygomatic Breadth	EKB	Biorbital Breadth
BBH	Basion-Bregma Height	DKB	lnterorbital Breadth
NLB	Cranial Base Length	FRC	Frontal Chord
BPL	Basion-Prosthion Length	PAC	Parietal Chord
MAB	Maxillo-Alveolar Breadth	OCC	Occipital Chord
MAL	Maxillo-Alveolar Length	FOL	Foramen Magnum Length
AUB	Biauricular Breadth	FOB	Foramen Magnum Breadth
NPH	Nasion-Prosthion Height	MDH	Mastoid Height
WFB	Minimum Frontal Breadth	ASB	Biasterionic Breadth
UFB	Upper Facial Breadth	ZMB	Bimaxillary breadth
NLH	Nasal Height	ZOB	Zygoorbitale breadth

TABLE 3 Accuracy of sex estimation standards for medieval Croatian populations tested on dry bone and CT measurements of O2

		Accuracy (%)		McNemar's test	
	n	Dry bones	CT	$\chi 2$	*P*
GOL	33	72.7	69.7	0.000	1.000
XCB	33	42.4	48.5	0.500	0.480
BNL	29	34.5	34.5	0.500	0.480
MAB	16	87.5	87.5	0.500	0.480
NPH	16	81.3	87.5	0.000	1.000
WFB	33	48.5	42.4	0.250	0.617
UFBR	29	51.7	44.8	0.500	0.480
EKB	28	71.4	67.9	0.000	1.000
DKB	30	63.3	66.7	0.000	1.000
F1 (all variables)	11	72.7	72.7	0.000	1.000
F2 (stepwise)	27	59.3	59.3	N/A	1.000*

*Exact McNemar's test

FIGURE LEGENDS

FIGURE 1 Cranial landmarks in midsagittal plane

FIGURE 2 Placing instrumentally defined landmarks (eurion)