ORIGINAL ARTICLE

What we see is what we touch? Sex estimation on the pelvis in virtual anthropology

Sandra Braun[1] · Nicole Schwendener[2] · Fabian Kanz[3] · Sandra Lösch[1] · Marco Milella[1]

Abstract

Background Computed tomography (CT) scans are a convenient means to study 3D reconstructions of bones. However, errors associated with the different nature of the observation, e.g. visual and tactile (on dry bone) versus visual only (on a screen) have not been thoroughly investigated.

Materials and methods We quantified the errors between modalities for sex estimation protocols of nonmetric (categorical and ordinal) and metric data, using 200 dry pelves of archaeological origin and the CT reconstructions of the same bones. In addition, we 3D surface scanned a subsample of 39 pelves to compare observations with dry bone and CT data. We did not focus on the sex estimation accuracy but solely on the consistency of the scoring, hence, the interchangeability of the modalities.

Results Metric data yielded the most consistent results. Among the nonmetric protocols, ordinal data performed better than categorical data. We applied a slightly modified description for the trait with the highest errors and grouped the traits according to consistency and availability in good, intermediate, and poor.

Discussion The investigated modalities were interchangeable as long as the trait definition was not arbitrary. Dry bone (gold standard) performed well, and CT and 3D surface scans performed better. We recommend researchers test their affinity for using virtual modalities. Future studies could use our consistency analysis and combine the best traits, validating their accuracy on various modalities.

Keywords Forensic anthropology · Nonmetric · Metric · DSP2 · Computed tomography · Surface scan

Introduction

The estimation of sex is a crucial parameter during the anthropological analysis of human skeletal remains of forensic or archaeological origin [1–4]. In osteoarchaeological research, sex is a relevant variable when addressing osteobiographies, case studies, and palaeodemographic research questions [1, 5–8]. Conversely, in forensic contexts, biological sex is of primary relevance for individual identification

purposes [9]. Although less problematic than other tasks (e.g. estimation of adult age-at-death), the estimation of sex from skeletal remains considers various potential sources of error. This concern is reflected in the currently applied terminology: while in the past, sex was *determined* [10, 11] or *assessed* [12], nowadays the preferred term is the *estimation* of sex [13]. Osteological sex estimation is based on the observation and evaluation of both quantitative (metric) and qualitative (nonmetric) dimorphic traits [13, 14]. Throughout this work, we will refer to these two types of features as metric and nonmetric.

The morphological overlap in sexual dimorphism [1] is accounted for by the ordinal nature of many of the sex estimation methods, e.g. scores 1 to 5, or very gracile to very robust [9], thus accommodating the continuum of sexual dimorphism [1, 9]. Due to its role in reproduction and the correlated selective pressures, the pelvis is the most sexually dimorphic human skeletal structures [1, 5, 6, 12, 15, 16].

✉ Sandra Braun
  sandra.braun@irm.unibe.ch

1 Department of Physical Anthropology, Institute of Forensic Medicine, University of Bern, Murtenstrasse 24-28, 3008 Bern, Switzerland

2 Department of Forensic Medicine and Imaging, Institute of Forensic Medicine, University of Bern, Bern, Switzerland

3 Forensic Anthropology Unit, Center for Forensic Medicine, Medical University of Vienna, Vienna, Austria

Qualitatively, female pelves are characterized by a series of features, including their more outwardly flared geometry, relatively wider inlet, and smaller and more gracile coxal bones [17–20]. These features and the overall lack of substantial differences in pelvic sexual dimorphism across populations make this skeletal structure the primary target for osteological sex estimation [21, 22].

Thus far, various methodological approaches have been published on the quantification of pelvic sexual dimorphism. Despite the increasing availability of molecular techniques, nonmetric and metric protocols are still preferred due to their low economical cost and low invasiveness [3]. Some of the methods most commonly applied by anthropologists, and the focus of the present study, are those of İşcan and Derrick [10], Bruzek [23], Klales et al. [15], and the Diagnose Sexuelle Probabiliste (DSP) method [24, 25]. These methods exemplify nonmetric [10, 15, 23] and metric [24, 25] and accuracy levels have been reported ranging from 86.2 [15] to 100% [26].

Sex estimation on dry bone and virtual modalities

Due to ethical concerns, the traditional way of anthropological research on human osteological collections is more and more under scrutiny [27–29]. These osteological collections represent cultural and historical documentation that should be preserved [30, 31]. At the same time, virtual alternatives (e.g. tomographic data, surface scans by means of structured light or laser, photogrammetry) have been established [32–34], and reflective imaging technologies (i.e. surface scans) require dry bones for scanning [35]. This kind of data adds value to the estimation of sex from forensic and clinical radiological images of corpses and patients [36–39] as well as from a bioarchaeological background [8, 40, 41]. In a forensic context, virtual images are nowadays permissible in courtrooms [42–45]. 3D prints of bones are usually preferred over the presentation of macerated dry bones in courtrooms, due to the potential evocation of emotions among the family present [46]. Further fields of application are paleoanthropology and osteoarchaeology [47–52] and evolutionary phylogenetic research [53]. Advantages of the virtual modalities are, for example, the worldwide access to data and consequently the facilitation of long-distance research collaborations [54], the non-invasive investigation of a present-day context [55, 56], and the possibility of advanced shape quantification methods such as geometric morphometrics [57]. However, these approaches are potentially afflicted by at least two issues: (a) the possible presence and extent of deviation between the same observations in the virtual and the physical environment, and (b) the consistency between observations on 3D models obtained by means of different imaging techniques (e.g. CT scans versus 3D surface scans). The quantification of the error affecting analogous versus

virtual observations seems especially relevant for nonmetric traits, the scoring of which is typically affected by subjectivity [23, 58–60]. By quantifying the errors associated with each of the sex estimation traits, our understanding of the methods could be impacted. Dissimilarities in the application of a variety of sex estimation traits to different modalities may be owing to the influence of the tactile sensation when handling dry bone versus the visual assessment of 3D models on a screen [36, 58, 60–62].

Previous studies already raised such concerns, while performing tests aimed at quantifying possible discrepancies: Grabherr and colleagues (2009) investigated age and sex estimation on the cranial and the pelvic regions on CT data of 22 individuals. Decker et al. (2011) used CT scans of 100 individuals to compare obstetric measurements and morphological traits on the pelvis. Both studies conclude that CT data are an excellent substitute for dry bone [63, 64], however without directly comparing the modalities and quantifying the associated errors. Another study compared the degree of closure of cranial sutures on ten crania observable on dry bone, 3D models obtained from surface scans, and CT data [61]. In contrast to Grabherr et al. (2009) and Decker et al. (2011), the results indicated differences between the analogous and the 3D surface modalities, while the observations on the CT modality were similar to those on dry bone [61]. The comparison of dry bone and CT images for the nonmetric estimation of age-at-death resulted in recommendations that scanning protocols should be uniform, with a special focus on slice thickness [62]. In another study, the direct comparison of 'dry' versus 'clinical' CT as well as 3D surface data for the measurement of inter-landmark distances on 14 female and 13 male pelves resulted in the superiority of the 'dry' CT data [65]. In contrast, the 3D surface scans of a fractured skull allowed more precise measurements than the multi-detector CT images of the same bone [66]. The authors do, however, add that the statement is limited by the fact that only one specimen was scanned [66]. The DSP method was tested for the interchangeability of the modalities dry bone and CT scans (individual os coxa and pelvic girdle) of 49 pelves [26]. In another study comparing the measurements and geometric variables of the ilium width and the ilium length, as well as the ilium module and the ilium area [67], it was found that the relative technical error of measurement (rTEM) was within the acceptable threshold of 5%, both in the intra- and the interobserver agreement tests. The authors concluded that both modalities could be used with confidence for the tested variables [67]. Although these studies contributed relevant data to discussions about the reliability of virtual observations in anthropological tasks, many of them were based on small sample sizes [61, 62, 64–66].

Since the general trend in anthropology points towards a growing interest in the virtual modalities [52, 62, 64, 68,

69], a comprehensive study on the interchangeability of the analogous and the virtual modalities in sex estimation methods is still lacking [58, 70, 71]. This study attempted to address this research gap and focus on the interchangeability of dry bone and CT on a large sample of pelves (n = 200), concentrating on the scoring and measuring consistency, and disregarding the accuracy of the employed methods in estimating sex. Therefore, the use of unidentified archaeological human remains was appropriate for our purpose. In fact, any sample of bones would have been suitable, provided their relative good state of preservation to test as many sex estimation traits as possible. In addition, for a subset of specimens (n = 39), we compared observations on digital 3D models obtained by means of 3D surface scans and CT data with observation of the dry bone specimen. We focus on the following points:

a) Are analogous pelves and their virtual counterparts interchangeable when trying to estimate sex using non-metric and metric protocols?
b) Is there a difference between observations of dimorphic traits taken on virtual models extracted from CT versus 3D surface scans?
c) Which role do the nature of observed variables (nonmetric and metric) and the scoring protocols play on the two former aims?

Materials and methods

Materials

Since our institutional forensic database consists of postmortem CT (PMCT) scans taken from fresh bodies, we had no dry bones at our disposal (e.g. [37, 39, 64, 72]). Thus, we selected 200 relatively well-preserved pelves (Table 1) from different archaeological sites and chronologies in Switzerland on which anthropological sex estimation was done following excavation [73–75]. While bone fragmentation cannot be avoided in any anthropological context, we selected the specimens according to a low degree of fragmentation. We chose individuals in order to obtain a balance between the sexes (100 females, 100 males), according to anthropological estimation based on the methods published by Buikstra and Ubelaker [5], Herrmann et al. [76], and White and Folkens [77]. Our intention was to include a wide spectrum of sexual dimorphism; hence, we did not verify sex with proteomic analysis nor DNA analysis (e.g. [78–80]). The estimated age-at-death of all individuals ranged between 18 and 80 years. We did not include subadult individuals (< 18 years), nor individuals presenting pathological changes to the pelvis.

Methods

Scoring protocols

We scored the dry bone pelves following the nonmetric sex estimation protocols by İşcan and Derrick [10], Bruzek [23], and Klales et al. [15] and collecting the ten linear DSP2 measurements by means of sliding and spreading calipers [25].

The method by İşcan and Derrick [10] outlines three traits (iliac tuberosity [IT], postauricular space [Pspace] and postauricular sulcus [PS]) to be scored categorically as female (F) or male (M). The postauricular space is regarded as the '*most reliable*' of the three structures [10, 81]. In this study, we used the method due to its simple application for individuals with an intact sacroiliac joint. Bruzek's method is more complex and contains a variety of conditions for each of the five traits [23]. The composite arch (CA) and ischiopubic proportion (IP) are scored categorically as female

Table 1 Archaeological sites, chronologies, number of female (F) and male (M) pelves, and modalities (dry bone, CT and 3D surface scans)

Site	Period	F	M	dry bone	CT	3D surface scans
Bern französische Kirche	13th-19th cent.	6	0	x	x	
Bern Grosse Schanze	18th-19th cent.	6	10	x	x	
Biel-Mett Kirche	7th-9th/13th-14th cent.	9	8	x	x	
Büren aA Chilchmatt	8th-16th cent.	8	6	x	x	
Ins Kirchgemeindehaus	**6th-10th cent.**	**18**	**21**	**x**	**x**	**x**
Kallnach Bergweg 95	6th-10th cent.	13	11	x	x	
Köniz Kirche	6th-14th cent.	12	13	x	x	
Miscellaneous		6	15	x	x	
Nidau St. Nikolauskapelle	16th-17th cent.	10	0	x	x	
Twann St. Petersinsel	8th-14th cent.	6	16	x	x	
Zweisimmen	8th-16th cent.	6	0	x	x	
	Total	100	100			

Boldface: site used for intra- and interobserver, as well as intermodality agreements on three modalities

(F), indeterminate (I) or male (M). The other three traits (preauricular surface [PSurf], greater sciatic notch [SN] and ischiopubic proportion [IP]) consist of three conditions each (e.g. F, I, M). The final sex estimation per trait is based on the mean score (e.g. female, if at least two conditions are rated female). The ventral arc (VA), the subpubic concavity (SC) and the medial aspect of the ischio-pubic ramus (IR) are comprised in the method that is ordinally scored from 1 (female) to 5 (male) [15]. The publication by Bruzek et al. [25] depicts each variable (acetabulo-symphyseal pubic length [PUM], cotylo-pubic width [SPU], innominate length [DCOX], greater sciatic notch height [IIMT], ischium post-acetabular length [ISMM], iliac breadth [SCOX], spino-sci-atic length [SS], spino-auricular length [SA], cotylo-sciatic breadth [SIS], and vertical acetabular diameter [VEAC]). It also includes the linear discriminant analysis formula used for the workable download. The measurements are entered into the database, with the resulting sex estimation given with immediate effect.

All trait scores and linear measurements for each individual and separated per observer, method and modality were entered in an excel sheet. Moreover, we scored and measured the left coxal bone where available and intact. If that was not the case, we observed the traits on the right coxal bone [59, 82].

CT and 3D surface scanning

The dry pelves (Fig. 1a) were CT scanned with a Somatom Definition AS 64 (Siemens, Berlin/Munich, Germany). Scanning parameters were 140 kV, 118 to 216 mAs, and a slice thickness of 0.6 mm (increment 0.3 mm). Reconstruction parameters of the matrix of 512×512 pixels was a field of view between 200 and 400 mm. We exported all raw tomographic scans as DICOM data from PACS IDS 7 v. 20.2.8.3353 (Sectra, Linköping, Sweden) and reconstructed them in Avizo (Thermo Fisher Scientific Inc., Waltham, MA, USA). We exported the reconstructed bone models again in DICOM format and imported them into Artec Studio software (Artec 3D, Luxembourg). In addition, we obtained 3D surface scans (Fig. 1b) of the 39

pelves (18 female, 21 male) from the Ins Kirchgemeinde-haus site (Table 1) with an Artec Spider scanner (Artec 3D, Luxembourg), using the handheld device with the maximum of eight frames per second. We aligned, segmented and reconstructed the 3D surface scans with Artec Studio software, and carried out the subsequent scoring and measuring of both the CT (Fig. 1c) as well as the Artec 3D reconstructions in the Artec Studio software (Artec 3D, Luxembourg). We scanned the pelvic girdles for the assessment of the Pspace trait, and individual coxal bones.

Data analysis

Intra- and interobserver agreement

For the evaluation of the intraobserver agreement, one observer (SB) scored the 39 pelves (female $n = 18$, male $n = 21$) from Ins Kirchgemeindehaus twice on dry bone, on CT as well as on 3D surface scan reconstructions (Table 2). For the interobserver agreement, we compared the first observations of the dry bone, the CT and the 3D surface scan modalities with those of a second observer (MM). We performed the scoring sessions at least two weeks apart from each other. We reported our findings separately for the categorical, ordinal and metric data. The two observers had three and 13 years of experience with imaging techniques, respectively.

Table 2 Intra- and interobserver, and intramodality agreement tests

Agreement test	Modalities	Abbrev.	N pelves
Intraobserver	3D surface scans - 3D surface scans	AA	39
Intraobserver	Dry bone - dry bone	BB	39
Intraobserver	CT - CT	CC	39
Interobserver	3D surface scans - 3D surface scans	AA*	39
Interobserver	Dry bone - dry bone	BB*	39
Interobserver	CT - CT	CC*	39
Intermodality	3D surface scans - dry bone	AB	39
Intermodality	3D surface scans - CT	AC	39
Intermodality	Dry bone - CT	BC	200

Fig. 1 Left coxal bone of Ins Kirchgemeindehaus individual 3543 on **a** dry bone, **b** Artec 3D surface scan, and **c** CT scan

Intermodality agreement

We evaluated the intermodality agreement by comparing the deviations in the scores assigned on the same pelvis between the modalities dry bone, CT and 3D surface scans. For the comparisons between the modalities, we used the first scoring sessions of the first observer. We used AB to refer to the comparison between 3D surface scans (Artec) and dry bone ($n = 39$), AC for the comparison between 3D surface scans (Artec) and CT ($n = 39$) and BC for the comparison of dry bone and CT ($n = 200$) (Table 2).

For the categorical data [10, 23], we employed Cohen's kappa κ [83] tests, for the interpretation of which we followed Landis and Koch [84]. The interpretation states that $\kappa < 0$ indicates a less than chance agreement; $\kappa = 0.01$ to 0.2 slight agreement; $\kappa = 0.21$ to 0.4 fair agreement; $\kappa = 0.41$ to 0.6 moderate agreement, $\kappa = 0.61$ to 0.8 substantial agreement; and $\kappa = 0.81$ to 1 almost perfect to perfect agreement [84]. For the ordinally scored method by Klales and colleagues [15], we applied Cohen's weighted κ [83] tests. For all Cohen's kappa tests, we assumed κ-values > 0.6 as acceptable agreement [36, 85]. For the DSP2 method, we analyzed the errors with the rTEM, in %. Although some publications refer to an acceptable rTEM threshold of $\leq 10\%$ [1, 86], we assumed an acceptable rTEM threshold of $\leq 5\%$ [25, 87]. Moreover, we assumed the dry bone modality as the gold standard [36].

Trait performance, availability, and scoring consistency

We investigated all traits by depicting the κ-value and rTEM ranges across all tests. In addition, we analyzed the DSP2 measurement correlations between the modalities by means of Pearson tests [88].

Osteological remains, especially from a burial context, may be fragmented and badly preserved. Therefore, some sexually dimorphic traits may be unavailable for scoring. We established three scores for the trait availability: score 1 for 80% to 100%, score 2 for availability between 60 and 80%, and score 3 for availability below 60%. Furthermore, consistency scores were either 1 ($\kappa > 0.6$, rTEM $< 5\%$) or 2 ($\kappa < 0.6$, rTEM $> 5\%$). Thus, a minimum score of 2, and a maximum of 5 could be reached per trait. We approached the central question of our research, e.g. the interchangeability of modalities by combining trait consistency and availability, resulting in a score between 2 and 5.

We performed all analyses and created all figures in R (version 4.0.4) using the packages *psych* [89], *DescTools* [90], *irr* [91], *fmsb* [92], *ggplot2* [93], and *BlandAltmanLeh* [94].

Results

Intra- and interobserver agreements

For the intraobserver agreement tests, we obtained the highest κ-values on the CC modality (mean 0.896, SD = 0.078) for the categorical traits (Fig. 2a). This agreement was closely followed by the AA comparison (mean 0.827, SD = 0.106). The BB comparison yielded a mean of 0.714 (SD = 0.153). In the interobserver agreement test (Fig. 2b), the CC* modality performed best again (mean 0.731, SD = 0.129), followed by the AA* (mean 0.582, SD = 0.135) and the BB* (mean 0.576, SD = 0.147) comparisons, both below the acceptable agreement threshold of 0.6. Cohen's weighted kappa tests for the ordinal traits (Fig. 2c and d) showed that the AA comparison in the intraobserver agreement test performed best (mean 0.803, SD = 0.021). The agreement of the CC comparison was also acceptable (mean 0.728, SD = 0.047), but the BB comparison and all interobserver tests yielded results below 0.6. The mean rTEM (Fig. 2e and f) were below 5% on all comparisons.

Intermodality agreements

In the intermodality tests for the ordinal traits (Fig. 3a), we obtained the highest agreement on the AC comparison (mean 0.756, SD = 0.105). The comparison between virtual modalities was followed by the BC comparison (mean 0.647, SD = 0.074), while the AB comparison (mean 0.562, SD = 0.132) was below the acceptable threshold of 0.6. The Cohen's weighted κ-value ranges (Fig. 3b) showed the highest agreement (mean 0.657, SD = 0.080) in the AC comparison. All three comparisons were above 0.6. All intermodality tests for the DSP2 method resulted in rTEM below 5% (Fig. 3c).

Analysis per trait

Regarding the categorical traits, we found the smallest range of κ-values in the IT trait, followed by Iprop, which also yielded the highest κ-values (mean 0.849, SD = 0.111). We saw the lowest mean κ-values in the Psurf and the IP traits (Fig. 4a). For the ordinal traits in Fig. 4b, we obtained the best results (mean 0.665, SD = 0.094) for the SC trait, followed by the IR and the VA trait. We detected a high rTEM for the IIMT measurement when analyzing the performance of the DSP measurements (Fig. 4c).

We display an in-depth evaluation of the κ-values and rTEM in Table 3. For the traits IT, Iprop, PUM, DCOX,
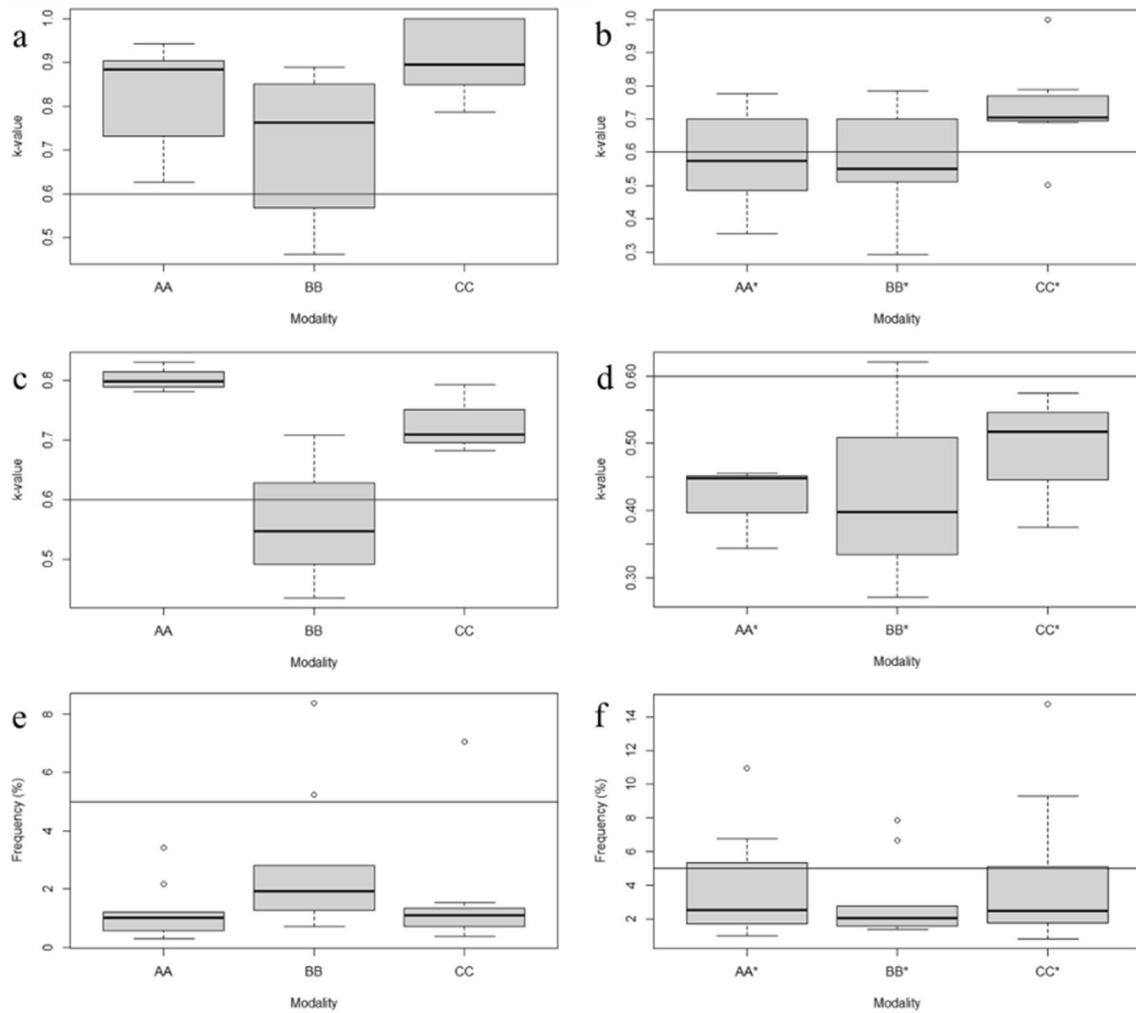
Fig. 2 Intra- and interobserver κ-value ranges of categorical traits (**a** and **b**), ordinal traits (**c** and **d**) and rTEM of metric traits (**e** and **f**). Horizontal lines indicate the acceptable thresholds (κ: 0.6; rTEM: 5%). Intraobserver agreements: AA = surface scan-surface scan; BB = dry bone-dry bone; CC = CT scan-CT scan. Interobserver agreements: AA* = surface scan-surface scan; BB* = dry bone-dry bone; CC* = CT scan-CT scan
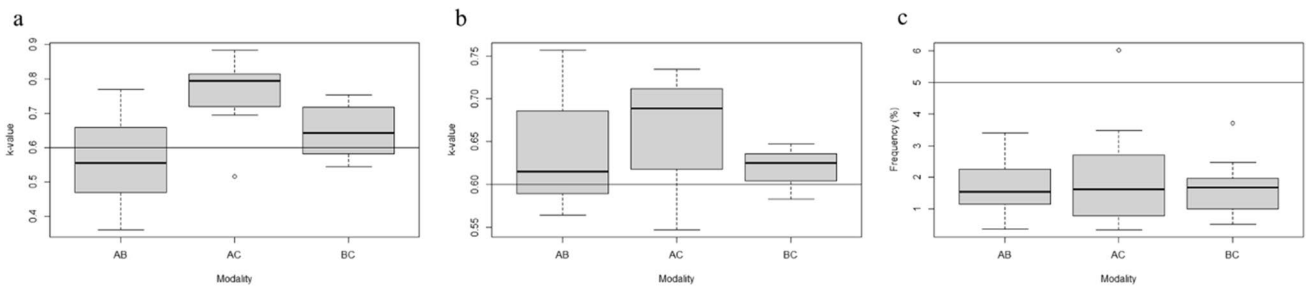

Fig. 3 Intermodality agreement κ-value ranges of categorical (**a**) and ordinal traits (**b**); rTEM of metric traits (**c**). Horizontal lines indicate the acceptable thresholds (κ: 0.6; rTEM: 5%). Intermodality agreements: AB = surface scan-dry bone; AC = surface scan-CT scan; BC = dry bone-CT scan
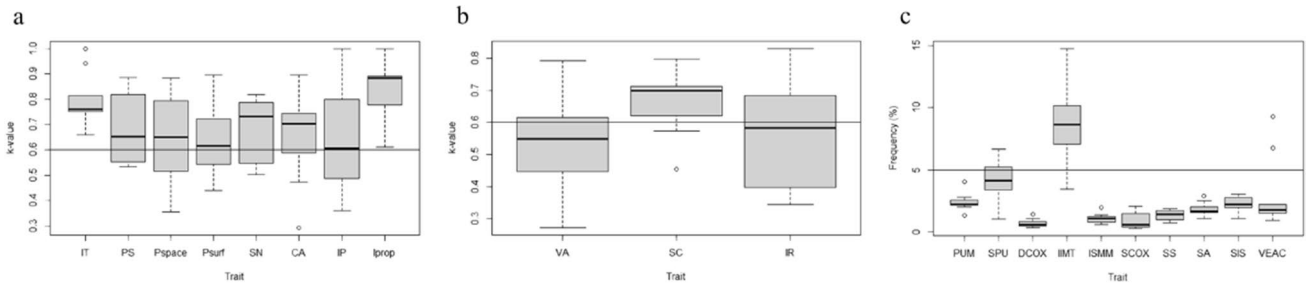
Fig. 4 Trait performance for categorical (**a**), ordinal (**b**), and metric traits (**c**). Horizontal lines indicate the acceptable thresholds (κ: 0.6; rTEM: 5%). IT = iliac tuberosity; PS = postauricular surface; Pspace = postauricular space; Psurf = preauricular surface; SN = sciatic notch; CA = composite arch; IP = inferior pelvis; Iprop = ischio-pubic proportion; VA = ventral arc; SC = subpubic concavity; IR = medial aspect of the ischio-pubic ramus; PUM = acetabulo-symphyseal pubic length; SPU = cotylo-pubic width; DCOX = innominate length; IIMT = greater sciatic notch height; ISMM = ischium post-acetabular length; SCOX = iliac breadth; SS = spino-sciatic length; SA = spino-auricular length; SIS = cotylo-sciatic breadth; VEAC = vertical acetabular diameter

Table 3 Trait consistency: κ-values and rTEM in descending/ascending order per trait; *n* refers to the number of available traits for BC (*n* = 200) comparison, in bold. Italic font indicates κ-values < 0.6 and rTEM > 5%

Nonmetric traits

Mod.	IT (n=172)	Mod.	PS (n=181)	Mod.	Pspace (n=121)	Mod.	Psurf (n=192)	Mod.	SN (n=195)	Mod.	CA (n=198)	Mod.	IP (n=111)	Mod.	Iprop (n=110)	Mod.	VA (n=108)	Mod.	SC (n=107)	Mod.	IR (n=109)
CC	1	AA	0.886	BB	0.883	CC	0.895	BB	0.819	AA	0.897	CC	1	CC	1	CC	0.793	AA	0.798	AA	0.831
AA	0.943	CC	0.881	CC	0.818	CC*	0.789	CC	0.788	CC	0.896	AA	0.913	CC*	1	AA	0.781	AB	0.757	AC	0.735
BB	0.814	AC	0.820	BC	0.794	AA	0.722	AC	0.788	AC	0.743	AC	0.800	BB	0.890	AB	0.615	BC	0.712	CC	0.683
AC	0.808	BB	0.711	CC*	0.706	AC	0.694	BC	0.771	AA*	0.743	CC*	0.700	BC	0.890	*BC*	*0.583*	CC	0.71	**BC**	**0.647**
BB*	0.765	CC*	0.706	**BC**	**0.674**	BB*	0.638	AA	0.743	**BC**	**0.715**	**BC**	**0.612**	AA	0.883	*BC*	*0.548*	BB	0.708	BC	0.601
AB	0.758	*BC*	*0.597*	AA	0.628	*BC*	*0.593*	**BC**	**0.718**	CC*	0.690	AA*	0.600	AC	0.883	*AC*	*0.547*	AC	0.689	*AB*	*0.564*
BC	**0.754**	*AB*	*0.560*	BB*	0.551	*AB*	*0.554*	AB	0.557	BC	0.636	*BB*	*0.548*	BB*	0.784	*CC**	*0.517*	**BC**	**0.625**	*BB*	*0.548*
CC*	0.750	**BC**	**0.553**	AC	0.516	**BC**	**0.543**	AA*	0.548	BB	0.588	BB*	0.486	AA*	0.776	*AA**	*0.448*	BB*	0.621	*BB**	*0.398*
BC	0.742	BB*	0.551	AB	0.465	BB	0.462	BB*	0.539	*AB*	*0.474*	*BC*	*0.469*	AB	0.769	*BB*	*0.436*	CC*	0.574	*CC**	*0.375*
AA*	0.659	AA*	0.533	AA*	0.356	AA*	0.439	CC*	0.503	*BB**	*0.292*	*AB*	*0.361*	**BC**	**0.612**	*BB**	*0.272*	AA*	0.455	*AA**	*0.345*

Metric traits

Mod.	PUM (n=98)	Mod.	SPU (n=138)	Mod.	DCOX (n=148)	Mod.	IIMT (n=189)	Mod.	ISMM (n=179)	Mod.	SCOX (n=113)	Mod.	SS (n=192)	Mod.	SA (n=192)	Mod.	SIS (n=191)	Mod.	VEAC (n=197)
CC	1.347	AA	1.057	AC	0.332	AA	3.427	AA	0.583	AA	0.3	CC	0.734	CC	1.085	CC	1.098	AA	0.944
AC	2.021	CC	1.535	CC	0.379	*AC*	*6.03*	CC	0.719	AB	0.355	AA	0.97	AA	1.221	AA	1.131	CC	1.166
BC	2.168	AB	3.409	AA	0.475	*CC*	*7.073*	AC	0.78	CC	0.395	**BC**	**0.992**	BB*	1.574	**BC**	**1.96**	AC	1.502
CC*	2.181	AC	3.476	**BC**	**0.506**	*BB**	*7.867*	AA*	1.005	BC	0.5	BB	1.397	**BC**	**1.663**	AB	2.131	AB	1.537
AA	2.185	**BC**	**3.719**	AB	0.527	*BB*	*8.39*	**BC**	**1.027**	AC	0.501	BB	1.397	AB	1.668	BB	2.14	BB	1.694
AB	2.248	BC	4.579	BC	0.7	*BC*	*8.853*	BC	1.131	**BC**	**0.652**	BC	1.439	AC	1.719	AB	2.343	**BC**	**1.857**
BC	**2.468**	CC*	5.084	BB	0.709	*BC*	*9.222*	AB	1.153	BB	0.889	BC	1.455	AC	1.847	AC	2.695	BC	2.107
BB*	2.543	*BB*	*5.24*	CC*	0.83	*AB*	*10.174*	BB	1.254	CC*	1.464	AA*	1.712	AA*	2.018	BB*	2.765	BB*	2.216
BB	2.822	*AA**	*5.353*	AA*	1.081	*AA**	*10.941*	BB*	1.399	BB*	1.713	CC*	1.792	BB	2.532	CC*	2.835	*AA**	*6.744*
AA*	4.022	*BB**	*6.664*	BB*	1.427	*CC**	*14.75*	CC*	1.968	AA*	2.052	BB*	1.865	CC*	2.908	AA*	3.056	*CC**	*9.262*

IT, iliac tuberosity; *PS*, postauricular surface; *Pspace*, postauricular space; *Psurf*, preauricular surface; *SN*, sciatic notch; *CA*, composite arch; *IP*, inferior pelvis; *Iprop*, ischio-pubic proportion; *VA*, ventral arc; *SC*, subpubic concavity; *IR*, medial aspect of the ischio-pubic ramus; *PUM*, acetabulo-symphyseal pubic length; *SPU*, cotylo-pubic width; *DCOX*, innominate length; *IIMT*, greater sciatic notch height; *ISMM*, ischium post-acetabular length; *SCOX*, iliac breadth; *SS*, spino-sciatic length; *SA*, spino-auricular length; *SIS*, cotylo-sciatic breadth; *VEAC*, vertical acetabular diameter; A, 3D surface scans; B, dry bone; C, CT scans

ISMM, SCOX, SS, SA and SIS, the κ-values on all modalities were above 0.6 or the rTEM below 5%, respectively. For the IIMT measurement, only the AA comparison was within the 5% threshold.

In the Pearson analysis, the IIMT measurement resulted in the lowest mean correlation coefficient *r* (0.65), while the mean of the other nine measurements ranged from 0.92 to 0.99. While all Pearson correlations were significant ($P < 0.001$), the IIMT measurement was on the border or beyond the standard deviation (SD) in the Bland–Altman plots (Fig. 5).

Since the high rTEM of the IIMT measurement could have been due to our interpretation of the location of the postero-inferior iliac spine according to White and Folkens
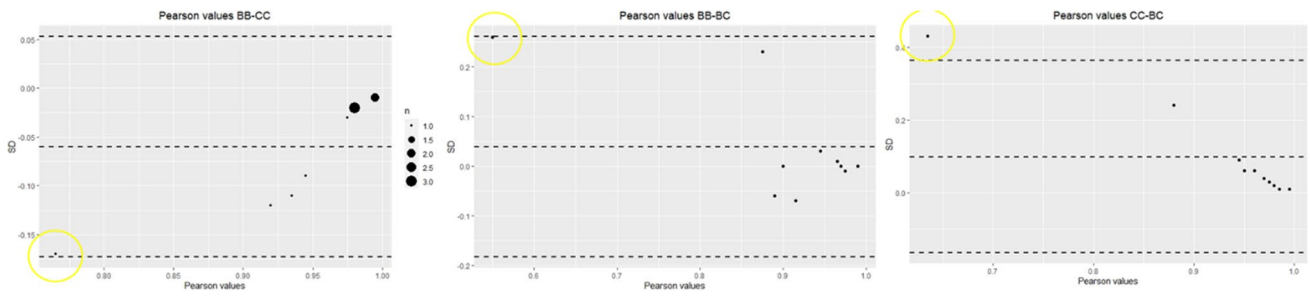
Fig. 5 Bland–Altman plots of Pearson correlations of DSP2 measurements of BB-CC, BB-BC and CC-BC (*n* = 200). Yellow circle indicates IIMT measurement

Fig. 6 **a** the IIMT measurement from the postero-inferior iliac spine according to White and Folkens (2005), perpendicular to the anterior margin of the sciatic notch. Note the relatively longer distance (60.27 mm versus 38.72 mm); **b** our assumed IIMT measurement; 1 is the intersection between the posterior margin of the sciatic notch and the auricular surface; 2 the point intersecting the anterior margin of the sciatic notch perpendicularly.
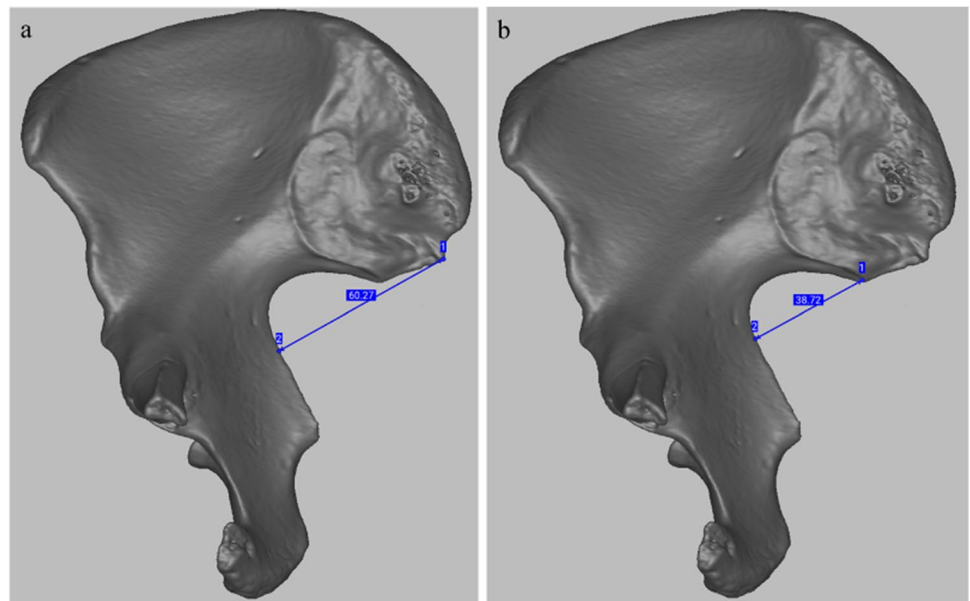
Fig. 7 rTEM of IIMT according to our definition. Line indicates rTEM 5%. Agreement tests: AA = surface scan-surface scan; AC = surface scan-CT scan; BB = dry bone-dry bone; BC = dry bone-CT scan
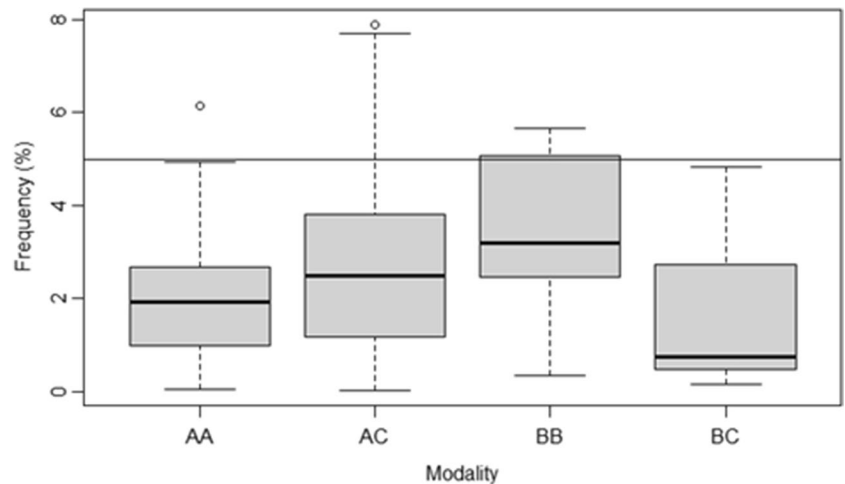
(2005), we tried to focus on the intersection of the auricular surface and the posterior portion of the sciatic notch as a starting point (Fig. 6a and b). This resulted in acceptable rTEM of this measurement (Fig. 7).

Trait availability

The traits Psurf, SN, IT, PS, CA, IIMT, ISMM, SS, SA, SIS and VEAC were mostly available (> 80%), while SPU
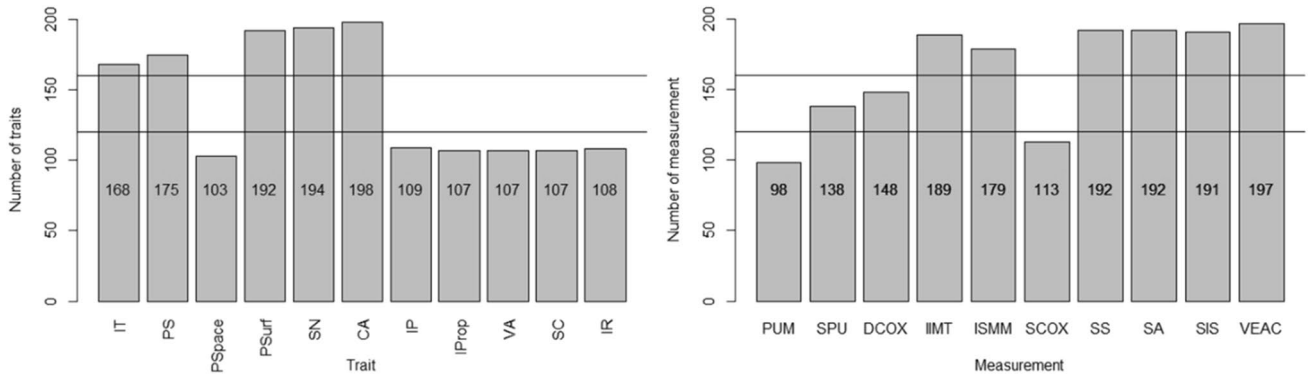
Fig. 8 Trait availability: nonmetric (left) and the metric (right) traits, with lines at 60% and 80% availability. Lines indicate 60% and 80% trait availability. IT = iliac tuberosity; PS = postauricular surface; Pspace = postauricular space; Psurf = preauricular surface; SN = sciatic notch; CA = composite arch; IP = inferior pelvis; Iprop = ischio-pubic proportion; VA = ventral arc; SC = subpubic concavity; IR = medial aspect of the ischio-pubic ramus; PUM = acetabulo-symphyseal pubic length; SPU = cotylo-pubic width; DCOX = innominate length; IIMT = greater sciatic notch height; ISMM = ischium post-acetabular length; SCOX = iliac breadth; SS = spino-sciatic length; SA = spino-auricular length; SIS = cotylo-sciatic breadth; VEAC = vertical acetabular diameter

and DCOX were available in 60% to 80% of observations. The traits Pspace, IP, Iprop, VA, SC, IR and PUM were available in less than 60% of cases (Fig. 8).

Consistency analysis

Based on the above results, we created spiderwebs (Fig. 9), categorizing the trait consistency and availability on the modality pairs AB, AC and BC. The traits IT, IIMT (adapted definition), ISMM, SS, SA, SIS and VEAC obtained the minimal score of 2 in all three comparisons. They are thus the traits with the highest consistency as well as availability. The traits PS, Psurf, SN, CA, Iprop, SC, PUM, SPU,

DCOX and SCOX fall into an intermediate class of consistency and availability. Finally, the traits Pspace, IP, VA and IR are poorest in both consistency and availability. Overall, metric traits fared best in terms of consistency, followed by the categorical traits and, finally, by the ordinal traits.

Discussion

The first research question addressed the degree of agreement between observations of nonmetric and metric sex estimation traits on the pelvis performed on analogous versus
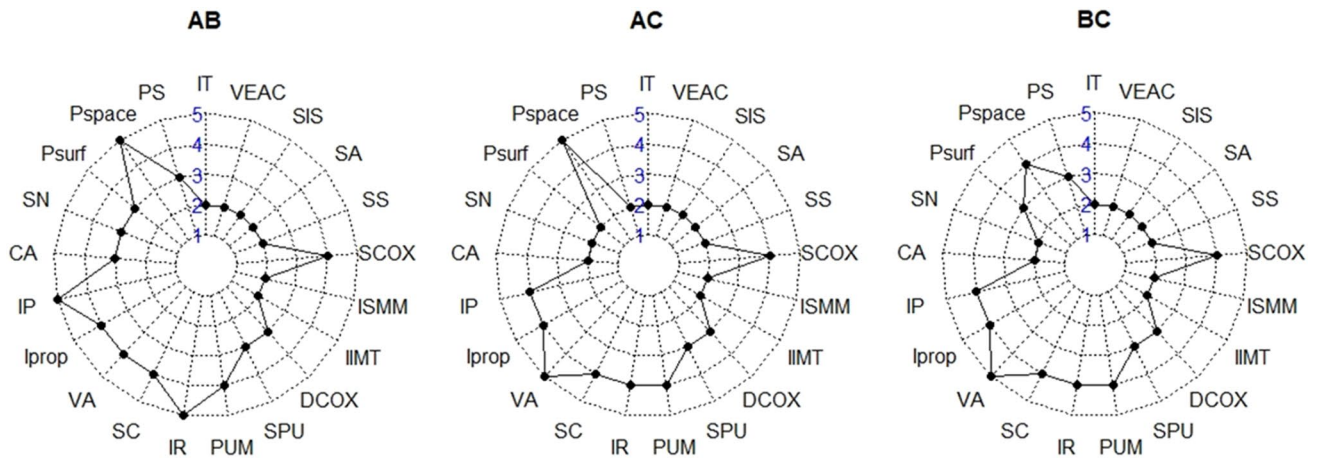

Fig. 9 Spiderwebs showing trait consistency and availability in the comparisons between modalities. Scores range from 2 (best) to 5 (poorest) IT = iliac tuberosity; PS = postauricular surface; Pspace = postauricular space; Psurf = preauricular surface; SN = sciatic notch; CA = composite arch; IP = inferior pelvis; Iprop = ischio-pubic proportion; VA = ventral arc; SC = subpubic concavity; IR = medial aspect of the ischio-pubic ramus; PUM = acetabulo-symphyseal pubic length; SPU = cotylo-pubic width; DCOX = innominate length; IIMT = greater sciatic notch height; ISMM = ischium post-acetabular length; SCOX = iliac breadth; SS = spino-sciatic length; SA = spino-auricular length; SIS = cotylo-sciatic breadth; VEAC = vertical acetabular diameter

virtual models, for which we found satisfactory agreement. The second research question concerned the degree of agreement between virtual models obtained from tomographic data versus 3D surface scans. For this, we found acceptable agreement for all intermodality comparisons. In the following, we discuss our results in detail.

The authors of a previous study [36] analyzed dry pelves and CT scans of the same bones of 14 female and 13 male individuals, and obtained weighted κ-values for the intra- and interobserver agreements between 0.61 and 0.9. The κ-values of the intermodality agreement tests ranged between 0.58 and 0.84 [36]. Comparing their study to ours, we obtained the highest agreements between modalities, while theirs was within observer. We found intermediate agreement within observer and the lowest agreement between observers, while their study indicated intermediate agreement between observers and the lowest agreement between modalities [36].

To the best of our knowledge, the methods by İşcan and Derrick (1984) and Bruzek (2002) have not been previously tested for consistency on different modalities. Hence, we cannot compare our results to earlier findings, which suggest best agreement within observer, intermediate agreement between modalities and lowest agreement between observers.

Analogous and virtual application of the DSP2 method has been previously compared on 49 pelvic bones [26]. They reported perfect agreement between observers and between modalities, possibly owing to the software used in the study (lhpFusionBox), allowing the 'virtual palpation' of anatomical landmarks [26]. Our study does not corroborate the findings of perfect agreement between observers and modalities for the DSP2 method; while our rTEM were acceptable for nine of the ten DSP variables, that was not the case for IIMT. When we adapted the description of this variable slightly for our purposes focusing on the intersection of the auricular surface and the posterior portion of the sciatic notch, we were able to reduce the rTEM below 5% on the different modalities. Thus, irrespective of modality, a precise definition of the traits under analysis is essential. Confounding or imprecise descriptions can lead to inaccuracies in the estimation of sex, and subsequently to a low degree of consistency on any modality [65]. Consequently, we consider the nature of an observation (tactile or visual) less influential on the consistency than the trait description.

Generally, the 3D surface scan and CT (AC) comparison yielded an agreement slightly superior to that of 3D surface scan and dry bone (AB), and dry bone and CT (BC). With reference to the fact that both former modalities are virtual, the greater difference between the analogous and the virtual modalities is self-evident. It seems logical to obtain better agreement between two visual modalities, suggesting a bigger gap between tactile and visual sex estimation than

between two visual modalities. Consequently, we summarize that in comparison to dry bone, considered the gold standard, the virtual modalities were even more consistent, thus, interchangeability was superior. This held true for all the different sex estimation protocols used in this study, with better results for the metric than for the nonmetric traits.

In general, and relating to our third research question, a greater error in nonmetric as opposed to metric methods can be ascribed to the inherent subjectivity of qualitative assessments [23, 58, 59]. We were therefore not surprised that the metric traits with a lower degree of subjectivity were more consistent than the nonmetric traits. Moreover, while with categorical traits the scoring options are limited (female or male), ordinal traits are more subtle in the scoring process, allowing the observer a gradual assessment of five scores [36, 59]. These expectations were reflected in our findings, as the source of error was lower with a gradual assessment across the modalities, as opposed to only two categories for selection. However, the categorical data was superior within observer, contrasting with ordinal data.

The challenging fragmentation of human remains from forensic or bioarchaeological contexts due to taphonomic processes [1, 95, 96] is especially pronounced in relation with the pubic bone [97, 98]. To that end, the IP, Iprop, VA, SC, IR and PUM traits were often unavailable for assessment. Other traits with low availability rates (below 60%) were SCOX, owing to eroded margins of the ilium, and Pspace, due to absent or fragmented sacra. Even though we minimized fragmentation in our sample due to a selective sampling process impossible in a forensic context [99], it was still an issue.

For the traits IT, IIMT (adapted definition), ISMM, SA, SS, SIS, and VEAC we found the greatest consistency (interchangeability of modality) as well as availability. The traits Iprop, SN, CA, PS, Psurf, SC, PUM, SPU, DCOX, and SCOX had intermediate qualities of modality interchangeability and availability. The four traits with the poorest consistency and availability in our study were Pspace, IP, VA, and IR. Our explanation for the poor performance of the Pspace trait relates to the fact that the observer needs to articulate the ilium and sacrum. If the two bones are slightly misaligned, results can differ. The trait IP refers to the characterization of the margo inferior ossis coxae, the absence or presence of the phallic ridge and the ischio-pubic ramus aspect [23]. The degree of inferior margin eversion could have been challenging to assess as the orientation on the virtual modalities could easily lead to a misjudgment of the eversion. In the VA trait, it could have been confusing to distinguish a '*true ventral arc*' occurring only in females from a '*ridge of bone*' present in males [100]. The subjectivity of the assessment could have led to the discrepancies in observations on the different modalities. In addition, the absence of tactile sensation on the virtual modalities could have inhibited the precise reproduction of assessing the ridge of bone. The same assessment of a '*ridge*' in females could

have caused the inconsistency of the IR trait [100]. Hence, traits with a description involving a ridge could present a pitfall in the scoring on a non-tactile modality.

To the best of our knowledge, this is the first study to analyze metric and nonmetric methods, the latter including categorical and ordinal data, at the same time encompassing a large dataset of 200 pelves of dry bone and CT data, with a subsample of 3D surface scans. This paper deals with the modality consistency of sex estimation traits. Our study could lay the foundation for future research, focusing on the best and the intermediate group of traits (IT, IIMT [adapted definition], ISMM, SA, SS, SIS, VEAC, Iprop, SN, CA, PS, Psurf, SC, PUM, SPU, DCOX, and SCOX). These traits could be used for a validation study investigating their combined accuracy for estimating sex, thus possibly creating a new set of sex estimation traits [see 62]. The subsequent data would thus encompass not only accuracy but also the trait consistency across modalities.

Limitations

Limitations to the general applicability of our results include the number of imaging approaches compared, and the specific state of preservation of the skeletal material used in the study. Here, we included two virtual approaches (CT and surface scans) chosen due to their frequent use in anthropological research. However, it would be interesting to test additional approaches (e.g. micro-focus X-ray computed tomography, photogrammetry, laser scanning). The relatively good preservation of the osteological material used in this study likely played a role on our results, especially those from the comparisons between modalities. A useful extension of this study would therefore be the comparison between our results and those obtained by performing the same analyses on highly degraded skeletal remains. Moreover, the state of preservation could be a potential limitation relating to the quality of scans as opposed to a dataset of unfragmented bones as we would need a direct comparison for the depiction of fragments. The participation of an observer without any previous experience with the virtual modalities can be recommended for a future study as levels of observer confidence could vary [101].

As with any sample of bones, the 3D surface scans are challenging on shiny surfaces such as teeth, but also on small foramina and orifices [49].

Conclusions

Our study confirmed that nonmetric methods are more prone to errors than metric methods, due to the inherent subjectivity of the former. The modalities dry bone and CT are interchangeable within observer, even if on a slightly lower level than the two virtual modalities (CT and 3D surface scans). Irrelevant of their accuracy in estimating sex, IT, IIMT (adapted definition), ISMM, SS, SA, SIS, and VEAC achieved best consistency and availability scores. We conclude that the modalities dry bone, 3D surface scans and CT are interchangeable for the pelvic nonmetric and metric sex estimation techniques applied in this study.

We recommend that future researchers test their affinity for applying a method to the virtual modalities before using it routinely. We support the conclusion of previous work that anthropological curricula should put more weight on developing and improving proficiency in imaging technologies [37, 102]. More pertinent training in this area could encourage the establishment of best practice manuals.

Future research investigating large samples for the consistency of sex estimation techniques on dry bone and virtual modalities will encompass a large sample of skulls. Validating the best and the intermediate traits for sex estimation accuracy could be the objective of another future study, leading to a set of sex estimation traits on the human pelvis and skull that are accurate as well as consistent between the modalities. Moreover, the traits under analysis in this work could be applied to a forensic PMCT dataset of present-day context to test for any possible bias regarding the use of an archaeological sample as in the present research.

Declarations

Competing interests The authors declare no competing interests.

References

1. İşcan MY, Steyn M. (2013) The human skeleton in forensic medicine. 3rd edition ed. Charles C. Thomas Springfield, Illinois.
2. Klales AR (2020) Sex estimation using pelvis morphology. In: Klales AR (ed) Sex estimation of the human skeleton. Academic Press Cambridge, Massachusetts, pp 75–93
3. Krishan K, Chatterjee PM, Kanchan T, Kaur S, Baryah N, Singh RK (2016) A review of sex estimation techniques during examination of skeletal remains in forensic anthropology casework. Forensic Science International 261: 165.e1-.e8. https://doi.org/10.1016/j.forsciint.2016.02.007
4. Spradley MK (2016) Metric methods for the biological profile in forensic anthropology: sex, ancestry, and stature. Acad Forensic Pathol 6:391–9. https://doi.org/10.23907/2016.040
5. Buikstra JE, Ubelaker DH (1994) Standards for data collection from human skeletal remains. Arkansas archaeological survey research series 44.
6. Acsádi G, Nemeskéri J. (1970) History of human life span and mortality. Akademiai kiado Budapest.
7. White TD, Black MT, Folkens PA (2012) Human Osteology, 3rd edn. Academic Press Burlington, MA
8. Bethard JD, VanSickle C (2020) Applications of sex estimation in paleoanthropology, bioarchaeology, and forensic anthropology. In: Klales AR (ed) Sex estimation of the Human Skeleton. Academic Press London, UK, pp 25–34
9. Garofalo EM, Garvin H. (2020) The confusion between biological sex and gender and potential implications of misinterpretation. In: Klales AR, ed. Sex Estimation of the Human Skeleton. Academic Press
10. İşcan MY, Derrick K (1984) Determination of sex from the sacroiliac joint: a visual assessment technique. Florida Scientist 47:94–98
11. Gibbon V, Paximadis M, Strkalj G, Ruff P, Penny C (2009) Novel methods of molecular sex identification from skeletal tissue using the amelogenin gene. Forensic Sci Int Genet 3:74–79. https://doi.org/10.1016/j.fsigen.2008.10.007
12. Spradley MK, Jantz RL (2011) Sex estimation in forensic anthropology: skull versus postcranial elements. J Forensic Sci 56:289–296. https://doi.org/10.1111/j.1556-4029.2010.01635.x
13. Klales AR. (2020) Introduction to sex estimation and this volume. In: Klales AR, ed. Sex Estimation of the Human Skeleton. Academic Press.
14. Pretorius E, Steyn M, Scholtz Y (2006) Investigation into the usability of geometric morphometric analysis in assessment of sexual dimorphism. Am J Phys Anthropol 129:64–70. https://doi.org/10.1002/ajpa.20251
15. Klales AR, Ousley SD, Vollner JM (2012) A revised method of sexing the human innominate using Phenice's nonmetric traits and statistical methods. Am J Phys Anthropol 149:104–114. https://doi.org/10.1002/ajpa.22102
16. Machado MPS, Costa ST, Freire AR et al (2018) Application and validation of Diagnose Sexuelle Probabiliste V2 tool in a miscegenated population. Forensic Sci Int 290(351):e1–e5. https://doi.org/10.1016/j.forsciint.2018.06.043
17. Washburn SL (1960) Tools and human evolution. Sci Am 203:63–75
18. Schwidetzky I, Stloukal M, Ferembach D (1980) Recommendations for age and sex diagnoses of skeletons. J Hum Evol 9:517–549
19. Pavlicev M, Romero R, Mitteroecker P (2020) Evolution of the human pelvis and obstructed labor: new explanations of an old obstetrical dilemma. Am J Obstet Gynecol 222:3–16. https://doi.org/10.1016/j.ajog.2019.06.043
20. Haeusler M, Grunstra NDS, Martin RD, Krenn VA, Fornai C, Webb NM (2021) The obstetrical dilemma hypothesis: there's life in the old dog yet. Biol Rev Camb Philos Soc 96:2031–2057. https://doi.org/10.1111/brv.12744
21. Kenyhercz MW, Klales AR, Stull KE, McCormick KA, Cole SJ (2017) Worldwide population variation in pelvic sexual dimorphism: A validation and recalibration of the Klales et al. method. Forensic Sci Int 277(259):e1–e8. https://doi.org/10.1016/j.forsciint.2017.05.001
22. Betti L (2014) Sexual dimorphism in the size and shape of the os coxae and the effects of microevolutionary processes. Am J Phys Anthropol 153:167–177. https://doi.org/10.1002/ajpa.22410
23. Bruzek J (2002) A method for visual determination of sex, using the human hip bone. Am J Phys Anthropol 117:157–168. https://doi.org/10.1002/ajpa.10012
24. Murail P, Bruzek J, Houët F, Cunha E (2005) DSP: A tool for probabilistic sex diagnosis using worldwide variability in hip-bone measurements. Bulletins et Mémoires de la Société d'Anthropologie de Paris 17:167–176
25. Bruzek J, Santos F, Dutailly B, Murail P, Cunha E (2017) Validation and reliability of the sex estimation of the human os coxae using freely available DSP2 software for bioarchaeology and forensic anthropology. Am J Phys Anthropol 164:440–449. https://doi.org/10.1002/ajpa.23282
26. Chapman T, Lefevre P, Semal P et al (2014) Sex determination using the Probabilistic Sex Diagnosis (DSP: Diagnose Sexuelle Probabiliste) tool in a virtual environment. Forensic Sci Int 234(189):e1-8. https://doi.org/10.1016/j.forsciint.2013.10.037
27. Belcastro MG, Pietrobelli A, Nicolosi T, Milella M, Mariotti V (2022) Scientific and Ethical Aspects of Identified Skeletal Series: The Case of the Documented Human Osteological Collections of the University of Bologna (Northern Italy). Forensic Sciences 2:349–361. https://doi.org/10.3390/forensicsci2020025
28. Alves-Cardoso F, Campanacho V, Plens CR (2022) Topical Collection "The Rise of Forensic Anthropology and Documented Human Osteological Collections." Forensic Sciences 2:551–555. https://doi.org/10.3390/forensicsci2030039
29. Squires K, Errickson D, Márquez-Grant N. (2019) Ethical approaches to human remains - a global challenge in biolarchaeology and forensic anthropology. Springer Cham, Switzerland.
30. Mann RW, Koel-Abt K, Dhody A et al (2021) The importance of human osteological collections: Our past, present, and future. Forensic Sci Int 325:110895. https://doi.org/10.1016/j.forsciint.2021.110895
31. Petaros A, Caplova Z, Verna E et al (2021) Technical note: the Forensic Anthropology Society of Europe (FASE) Map of Identified Osteological Collections. Forensic Sci Int 328:110995. https://doi.org/10.1016/j.forsciint.2021.110995
32. Edgar H, Berry S (2019) NMDID: a new research resource for biological anthropology. Am J Phys Anthropol Suppl 168:66
33. Stull KE, Corron LK (2022) The subadult virtual anthropology database (SVAD): an accessible repository of contemporary subadult reference data. Forensic Sciences 2:20–36. https://doi.org/10.3390/forensicsci2010003
34. L'Abbé EN, Krüger GC, Theye CEG, Hagg AC, Sapo O (2021) The pretoria bone collection: a 21st century skeletal collection in South Africa. Forensic Sciences 1:220–227. https://doi.org/10.3390/forensicsci1030020

35. Friess M (2012) Scratching the Surface? The use of surface scanning in physical and paleoanthropology. J Anthropol Sci 90:7–31. https://doi.org/10.4436/jass.90004

36. Colman KL, van der Merwe AE, Stull KE et al (2019) The accuracy of 3D virtual bone models of the pelvis for morphological sex estimation. Int J Legal Med 133:1853–1860. https://doi.org/10.1007/s00414-019-02002-7

37. Braun S, Indra L, Lösch S, Milella M (2022) Perimortem skeletal sharp force trauma: detection reliability on CT data, demographics and anatomical patterns from a forensic dataset. Biology 11. https://doi.org/10.3390/biology11050666

38. Simmons-Ehrhardt T (2021) Open osteology: medical imaging databases as skeletal collections. Forensic Imaging 26. https://doi.org/10.1016/j.fri.2021.200462

39. Zech WD, Hatch G, Siegenthaler L, Thali MJ, Lösch S (2012) Sex determination from os sacrum by postmortem CT. Forensic Sci Int 221:39–43. https://doi.org/10.1016/j.forsciint.2012.03.022

40. Hassett BR (2018) Which bone to pick: creation, curation, and dissemination of online 3D digital bioarchaeological data. Archaeologies 14:231–249. https://doi.org/10.1007/s11759-018-9344-z

41. Kuzminsky SC, Gardiner MS (2012) Three-dimensional laser scanning: potential uses for museum conservation and scientific research. J Archaeol Sci 39:2744–2751. https://doi.org/10.1016/j.jas.2012.04.020

42. Ampanozi G, Zimmermann D, Hatch GM et al (2012) Format preferences of district attorneys for post-mortem medical imaging reports: understandability, cost effectiveness, and suitability for the courtroom: a questionnaire based study. Leg Med (Tokyo) 14:116–120. https://doi.org/10.1016/j.legalmed.2011.12.008

43. Errickson D, Grueso I, Griffith SJ et al (2017) Towards a best practice for the use of active non-contact surface scanning to record human skeletal remains from archaeological contexts. Int J Osteoarchaeol 27:650–661. https://doi.org/10.1002/oa.2587

44. Carew RM, Errickson D (2019) Imaging in forensic science: five years on. Journal of Forensic Radiology and Imaging 16:24–33. https://doi.org/10.1016/j.jofri.2019.01.002

45. Errickson D, Thompson TJU, Rankin BWJ (2014) The application of 3D visualization of osteological trauma for the courtroom: a critical review. Journal of Forensic Radiology and Imaging 2:132–137. https://doi.org/10.1016/j.jofri.2014.04.002

46. Kettner M, Schmidt P, Potente S, Ramsthaler F, Schrodt M (2011) Reverse engineering–rapid prototyping of the skull in forensic trauma analysis. J Forensic Sci 56:1015–1017. https://doi.org/10.1111/j.1556-4029.2011.01764.x

47. Beaudet A, Clarke RJ, de Jager EJ et al (2019) The endocast of StW 573 ("Little Foot") and hominin brain evolution. J Hum Evol 126:112–123. https://doi.org/10.1016/j.jhevol.2018.11.009

48. Licata M, Tosi A, Ciliberti R, Badino P, Pinto A (2019) Role of Radiology in the Assessment of Skeletons from Archeological Sites. Semin Ultrasound CT MR 40:12–17. https://doi.org/10.1053/j.sult.2018.10.003

49. Waltenberger L, Rebay-Salisbury K, Mitteroecker P (2021) Three-dimensional surface scanning methods in osteology: a topographical and geometric morphometric comparison. Am J Phys Anthropol. https://doi.org/10.1002/ajpa.24204

50. Lundquist AA, Jensen ND, Jørkov ML, Lynnerup N, Villa C (2022) How 3D printing and physical reconstruction of a skull helped in a complex pathological case. Anthropol Anz 72:83–94

51. Siebke I, Campana L, Ramstein M, Furtwängler A, Hafner A, Lösch S (2018) The application of different 3D-scan-systems and photogrammetry at an excavation — a Neolithic dolmen from Switzerland. Digital Applications in Archaeology and Cultural Heritage 10. https://doi.org/10.1016/j.daach.2018.e00078

52. Uldin T (2017) Virtual anthropology - a brief review of the literature and history of computed tomography. Forensic Sci Res 2:165–173. https://doi.org/10.1080/20961790.2017.1369621

53. Fruciano C, Celik MA, Butler K, Dooley T, Weisbecker V, Phillips MJ (2017) Sharing is caring? Measurement error and the issues arising from combining 3D morphometric datasets. Ecol Evol 7:7034–7046. https://doi.org/10.1002/ece3.3256

54. Garvin HM, Stock MK (2016) The utility of advanced imaging in forensic anthropology. Academic Forensic Pathology 6:499–516. https://doi.org/10.23907/2016.050

55. Buck U, Naether S, Braun M, Thali M (2008) Haptics in forensics: the possibilities and advantages in using the haptic device for reconstruction approaches in forensic science. Forensic Sci Int 180:86–92

56. Teodoru-Raghina D, Perlea P, Marinescu M (2017) Forensic anthropology from skeletal remains to CT scans: A review on sexual dimorphism of human skull. Romanian Journal of Legal Medicine 25:287–292. https://doi.org/10.4323/rjlm.2017.287

57. Bytheway JA, Ross AH (2010) A geometric morphometric approach to sex determination of the human adult os coxa. J Forensic Sci 55:859–864. https://doi.org/10.1111/j.1556-4029.2010.01374.x

58. Braun S, Ridel AF, L'Abbé EN, Theye CEG, Oettlé AC (2022) Repeatability of a morphoscopic sex estimation technique for the mental eminence on micro-focus X-ray computed tomography models. Forensic Imaging 28. https://doi.org/10.1016/j.fri.2022.200500

59. Santos F, Guyomarc'h P, Rmoutilova R, Bruzek J (2019) A method of sexing the human os coxae based on logistic regressions and Bruzek's nonmetric traits. Am J Phys Anthropol 169:435–447. https://doi.org/10.1002/ajpa.23855

60. Siek T (2015) An exploration of tactile interaction in osteology and material culture. Journal of Graduate Students in Anthropology PLATFORUM: 147–64.

61. Abegg C, Balbo I, Dominguez A, Grabherr S, Campana L, Moghaddam N (2021) Virtual anthropology: a preliminary test of macroscopic observation versus 3D surface scans and computed tomography (CT) scans. Forensic Sci Res 6:34–41. https://doi.org/10.1080/20961790.2020.1817270

62. Uldin T (2016) Virtual anthropology: the forensic approach. University of Geneva

63. 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 56:1107–1114. https://doi.org/10.1111/j.1556-4029.2011.01803.x

64. Grabherr S, Cooper C, Ulrich-Bochsler S et al (2009) Estimation of sex and age of "virtual skeletons"–a feasibility study. Eur Radiol 19:419–429. https://doi.org/10.1007/s00330-008-1155-y

65. Colman KL, de Boer HH, Dobbe JGG et al (2019) Virtual forensic anthropology: The accuracy of osteometric analysis of 3D bone models derived from clinical computed tomography (CT) scans. Forensic Science International 304:109963. https://doi.org/10.1016/j.forsciint.2019.109963

66. Fahrni S, Campana L, Dominguez A et al (2017) CT-scan vs. 3D surface scanning of a skull: first considerations regarding reproducibility issues. Forensic Sci Res 2:93–99. https://doi.org/10.1080/20961790.2017.1334353

67. Corron L, Marchal F, Condemi S, Chaumoître K, Adalian P (2016) 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. Bulletins et Mémoires de la Société d'Anthropologie de Paris 29:33–53. https://doi.org/10.1007/s13219-016-0172-7

68. Rowbotham SK, Blau S (2020) The application of medical imaging to the anthropological estimation of sex. In: Klales AR (ed)

Sex estimation of the human skeleton. Academic Press Cambridge, Massachusetts, pp 351–369

69. Brough A, Rutty G, Villa C, Colman K, Dedouit F, Decker SJ (2019) The benefits of medical imaging and 3D modelling to the field of forensic anthropology positional statement of the members of the forensic anthropology working group of the International Society of Forensic Radiology and Imaging. J Forensic Radiol Imaging 18:18–19. https://doi.org/10.1016/j.jofri.2019.07.003

70. Garvin HM, Klales AR (2018) A Validation Study of the Langley et al. (2017) Decision Tree Model for Sex Estimation. Journal of Forensic Sciences 63:1243–51

71. Bertoglio B, Corradin S, Cappella A et al (2020) Pitfalls of Computed Tomography 3D Reconstruction Models in Cranial Nonmetric Analysis. J Forensic Sci. https://doi.org/10.1111/1556-4029.14535

72. Zech WD, Naf M, Siegmund F, Jackowski C, Losch S (2016) Body height estimation from post-mortem CT femoral F1 measurements in a contemporary Swiss population. Leg Med (Tokyo) 19:61–66. https://doi.org/10.1016/j.legalmed.2016.02.004

73. Ulrich-Bochsler S, Cooper C, Baeriswyl A (2016) Karies, Knochenbrüche, Infektionen. Berner Zeitschrift für Geschichte 78:1–52

74. Ulrich-Bochsler S. (2010) Die Anthropologische Sammlung des Naturhistorischen Museums Bern - Katalog der Neueingange 1985 bis 2005. Naturhistorisches Museum Bern Bern, Switzerland.

75. Schoch W, Ulrich-Bochsler S. (1987) Die Anthropologische Sammlung des Naturhistorischen Museums Bern - Katalog der Neueingänge 1956 bis 1985. Naturhistorisches Museum Bern Bern, Switzerland.

76. Herrmann B, Grupe G, Hummel S, Piepenbrink H, Schutkowski H (1990) Prähistorische Anthropologie. Springer-Verlag, Berlin Heidelberg New York

77. White TD, Folkens PA (2005) The human bone manual. Elsevier Academic Amsterdam, Boston

78. Rebay-Salisbury K, Janker L, Pany-Kucera D et al (2020) Child murder in the Early Bronze Age: proteomic sex identification of a cold case from Schleinbach. Austria Archaeol Anthropol Sci 12:265. https://doi.org/10.1007/s12520-020-01199-8

79. Stone AC, Milner GR, Pääbo S, Stoneking M (1996) Sex determination of ancient human skeletons using DNA. Am J Phys Anthropol 99:231–238. https://doi.org/10.1002/(SICI)1096-8644(199602)99:2%3c231::AID-AJPA1%3e3.0.CO;2-1

80. Furtwängler A, Rohrlach AB, Lamnidis TC et al (2020) Ancient genomes reveal social and genetic structure of Late Neolithic Switzerland. Nat Commun 11:1915

81. Krogman WM, İşcan MY. (1986) The human skeleton in forensic medicine. 2nd edition ed. Charles C Thomas Springfield.

82. Quatrehomme G, Radoman I, Nogueira L, du Jardin P, Alunni V (2017) Sex determination using the DSP (probabilistic sex diagnosis) method on the coxal bone: Efficiency of method according to number of available variables. Forensic Sci Int 272:190–193. https://doi.org/10.1016/j.forsciint.2016.10.020

83. Cohen J (1968) Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70:213–220

84. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

85. Reneman MF, Brouwer S, Meinema A, Dijkstra PU, Geertzen JHB, Groothoff JW (2004) Test-retest reliability of the Isernhagen work systems functional capacity evaluation in healthy adults. J Occup Rehabil 14:295–305

86. Rogers TL (2005) Determining the sex of human remains through cranial morphology. J Forensic Sci 50:1–8. https://doi.org/10.1520/JFS2003385

87. Weinberg SM, Scott NM, Neiswanger K, Marazita ML (2005) Intraobserver error associated with measurements of the hand. Am J Hum Biol 17:368–371. https://doi.org/10.1002/ajhb.20129

88. Schober P, Boer C, Schwarte LA (2018) Correlation Coefficients: Appropriate Use and Interpretation. Anesth Analg 126:1763–1768. https://doi.org/10.1213/ANE.0000000000002864

89. Revelle W. (2020) psychTools: Tools to accompany the 'psych' package for psychological research. 2.0.8 ed Northwestern University, Evanston, Illinois.

90. Signorell A, al. e. (2021) DescTools: tools for descriptive statistics. 0.99.40 ed.

91. Gamer M, Lemon J, Fellows I, Singh P. (2019) Various coefficients of interrater reliability and agreement.

92. Nakazawa M. (2022) Package 'fmsb'. 0.7.3 ed.

93. Wickham H, Chang W, Henry L et al. (2022) Package 'ggplot2'.

94. Lehnert B. (2022) Package 'BlandAltmanLeh'. 0.3.1 ed.

95. Walker PL (2001) A bioarchaeological perspective on the history of violence. Annu Rev Anthropol 30:573–596

96. Indra L, Lösch S (2021) Forensic anthropology casework from Switzerland (Bern): Taphonomic implications for the future. Forensic Science International: Reports 4. https://doi.org/10.1016/j.fsir.2021.100222

97. Kelley MA (1979) Sex determination with fragmented skeletal remains. J Forensic Sci 24:154–158

98. Blake KAS, Hartnett-McCann K (2018) Metric assessment of the pubic bone using known and novel data points for sex estimation. J Forensic Sci 63:1472–1478. https://doi.org/10.1111/1556-4029.13732

99. Scott S, Jantz RL (2022) Survivability versus rate of recovery for skeletal elements in forensic anthropology. J Forensic Sci 67:1758–1765. https://doi.org/10.1111/1556-4029.15087

100. Phenice T (1969) A newly developed visual method of sexing the os pubis. Am J Phys Anthropol 30:297–301. https://doi.org/10.1002/ajpa.1330300214

101. Jepps H, Carew RM, Nakhaeizadeh S (2023) Assessing the feasibility of estimating the age and sex from virtual 3D models: a pilot study into virtual forensic anthropology. Forensic Imaging. https://doi.org/10.1016/j.fri.2023.200544

102. Spies AJ, Steyn M, Prince DN, Brits D (2021) Can forensic anthropologists accurately detect skeletal trauma using radiological imaging? Forensic Imaging 24. https://doi.org/10.1016/j.fri.2020.200424