Received: 5 November 2020 | Revised: 28 January 2021 | Accepted: 24 February 2021

DOI: 10.1002/ajpa.24270

RESEARCH ARTICLE

American Journal of PHYSICAL ANTHROPOLOGY WILEY

Revisiting metric sex estimation of burnt human remains via supervised learning using a reference collection of modern identified cremated individuals (Knoxville, USA)

Marta Hlad[1,2] | Barbara Veselka[1] | Dawnie Wolfe Steadman[3] | Baptiste Herregods[4] | Marc Elskens[5] | Rica Annaert[1,6] | Mathieu Boudin[7] | Giacomo Capuzzo[2] | Sarah Dalle[1,8] | Guy De Mulder[8] | Charlotte Sabaux[2,8] | Kevin Salesse[2,9] | Amanda Sengeløv[2,8] | Elisavet Stamataki[1,2] | Martine Vercauteren[2] | Eugène Warmenbol[10] | Dries Tys[1] | Christophe Snoeck[1,5,11]

[1]Maritime Cultures Research Institute, Department of History, Archaeology, Arts, Philosophy and Ethics, Vrije Universiteit Brussel, Brussels, Belgium

[2]Research Unit Anthropology and Human Genetics, Faculty of Science, Université Libre de Bruxelles, Brussels, Belgium

[3]Department of Anthropology, University of Tennessee, Knoxville, Tennessee, USA

[4]Independent researcher, Brussels, Belgium

[5]Research Unit Analytical, Environmental and Geo-Chemistry, Department of Chemistry, Vrije Universiteit Brussel, AMGC-WE-VUB, Brussels, Belgium

[6]Flemish Heritage Agency, Brussels, Belgium

[7]Radiocarbon Dating Laboratory, Royal Institute for Cultural Heritage, Brussels, Belgium

[8]Department of Archaeology, Ghent University, Ghent, Belgium

[9]UMR 5199: "PACEA - De la Préhistoire à l'Actuel: Culture, Environnement et Anthropologie", University of Bordeaux, Pessac cedex, France

[10]Center de Recherches en Archéologie et Patrimoine, Department of History, Arts, and Archaeology, Université Libre de Bruxelles, Brussels, Belgium

[11]G-Time Laboratory, Université Libre de Bruxelles, Brussels, Belgium

Correspondence
Marta Hlad, Department of History, Archaeology, Arts, Philosophy and Ethics, Maritime Cultures Research Institute, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium.
Email: marta.hlad@vub.be

Funding information
Fonds Wetenschappelijk Onderzoek, Grant/Award Numbers: 12W2118N, 198613, K232919N; FWO/F.R.S.-FNRS EoS, Grant/Award Number: 30999782

Abstract

Objectives: This study aims to increase the rate of correctly sexed calcined individuals from archaeological and forensic contexts. This is achieved by evaluating sexual dimorphism of commonly used and new skeletal elements via uni- and multi-variate metric trait analyses.

Materials and methods: Twenty-two skeletal traits were evaluated in 86 individuals from the William M. Bass donated cremated collection of known sex and age-at-death. Four different predictive models, logistic regression, random forest, neural network, and calculation of population specific cut-off points, were used to determine the classification accuracy (CA) of each feature and several combinations thereof.

Results: An overall CA of ≥ 80% was obtained for 12 out of 22 features (humerus trochlea max., and lunate length, humerus head vertical diameter, humerus head transverse diameter, radius head max., femur head vertical diameter, patella width, patella thickness, and talus trochlea length) using univariate analysis. Multivariate analysis showed an increase of CA (≥ 95%) for certain combinations and models (e.g., humerus trochlea max. and patella thickness). Our study shows metric sexual

dimorphism to be well preserved in calcined human remains, despite the changes that occur during burning.

Conclusions: Our study demonstrated the potential of machine learning approaches, such as neural networks, for multivariate analyses. Using these statistical methods improves the rate of correct sex estimations in calcined human remains and can be applied to highly fragmented unburnt individuals from both archaeological and forensic contexts.

KEYWORDS
calcined human bones, metric traits, multivariate analysis, sexual dimorphism

1 | INTRODUCTION

Analysis of burnt human remains from archaeological and forensic contexts can be challenging due to their fragmentary and often incomplete nature. The process of burning causes extensive micro- and macroscopic changes in bones and teeth. Bones may fragment, shrink, expand, and otherwise deform (Dokladál, 1971; Piontek, 1976; Rösing, 1977; Shipman et al., 1984; Snoeck et al., 2014; Thompson, 2002). In addition, remains from archaeological settings carry traces of funerary rituals that can further complicate their understanding. Funerary practices vary throughout time and space, and depend on various factors, such as beliefs and customs. Pyre technology (e.g., temperature, fuel, and duration), handling of the remains during and after burning (e.g., selective deposition, further fragmentation, combining remains from different individuals) may influence the final quality and quantity of the cremated remains (McKinley, 2016; Oestigaard, 2013; Thompson, 2005). Additionally, post-depositional processes, such as taphonomic degradation, may further affect the bones. The intensity of diagenetic alteration depends on the time that passed since deposition and on how the remains were buried. The preservation varies depending on whether they were buried in an urn or organic container, scattered on the ground of a burial pit, or incorporated in different features and burial structures at the archaeological sites (Williams, 2008). Burning of human remains also occurs in forensic settings, which may be deliberate or accidental. The degree of burning will depend on the proximity to the fire, its intensity and duration, and other contextual circumstances (Mayne Correia & Beattie, 2002). Usually, burnt human remains are highly fragmented, regardless of their context, and contain a varying proportion of identifiable bone fragments, with most of them having only limited value in establishing the biological profile of the deceased.

Due to the variety of processes involved in the preservation of burnt human remains (Depierre, 2014), establishing a biological profile may be a challenging task with the currently available sexing methods. Most researchers working with archaeological cremations use morphological parameters that are routinely used for sex estimation in unburnt skeletons (Gonçalves & Pires, 2017), such as described in Buikstra and Ubelaker (1994) and the Workshop of European Anthropologists (1980). However, the fragments of the pelvis and cranium needed for sex and age estimations are often absent, and even if these diagnostic elements are present, they tend to be highly fragmented and deformed, which diminishes their potential for an accurate estimate. Bones can shrink up to 40% compared to their initial dimensions (Gonçalves et al., 2013), potentially leading to misclassification of some individuals due to morphological and dimensional changes (Thompson, 2002) and sometimes to nonnegligible disagreements in sex estimation between different scholars (Welinder, 1989). The degree of shrinkage is unknown when dealing with archaeological remains and it is closely related to pyre conditions, such as the type of fuel and temperature, as well as the skills of the cremators (Oestigaard, 2016). A solution to this issue could be the application of different correction factors, as suggested by Buikstra and Swegle (1989), Piontek (1976), and more recently by Gonçalves et al. (2020) who produced regression equations based on chemometric indices obtained via FTIR analyses.

These sexing issues are well attested in literature, where sex estimation of archaeological cremated remains is possible for as little as 20% of individuals for certain sites (Holck, 1986; Veselka & Lemmers, 2014). However, sex estimates are a key element of the biological profile. It is essential in forensic settings for the identification of victims (Gonçalves et al., 2013). In archaeology, it can be a fundamental piece of information for the study of palaeodemography and different social phenomena in the past (Brück, 2009). Over the past decades, several methods were proposed to improve the rates of sexed calcined individuals, such as assessment of the lateral angle of internal acoustic meatus (e.g., Gonçalves et al., 2011; Graw et al., 2005; Masotti et al., 2013), as well as metric analyses of teeth (Godinho et al., 2019; Gouveia et al., 2017), and bones (Cavazzuti et al., 2019; Gonçalves et al., 2013; Rösing, 1977; Schutkowski & Herrmann, 1983; Van Vark, 1975; Wahl, 1996). Despite the large number of studies on the sexually dimorphic potential of the lateral angle of internal acoustic meatus, the results are contradictory, and classification accuracy (CA) ranges from ~60% to 80%. This may be due to the strong correlation between this angle and the age of the individuals (Afacan et al., 2017; Masotti et al., 2019), implying that the accuracy would improve by eliminating elderly individuals from the studied group. However, not only is the estimation of age-at-death a challenge in highly fragmented and burnt human remains

(Veselka et al., 2020), but excluding one portion of population to correctly sex the others is not ideal if the goal is to identify an unknown individual or have a general idea of a sex ratio in a population.

Several studies confirmed that sexual dimorphism is preserved in calcined remains; modern populations of known age and sex were studied in Portugal (Gonçalves et al., 2013), Germany (Wahl, 1996), and Sweden (Gejvall, 1963; Van Vark, 1975). Archaeological collections from Poland (Piontek, 1975), Germany (Rösing, 1977), and Late Iron and Early Bronze Age Italy were sexed with metrics (Cavazzuti et al., 2019). Metric techniques for sexing bones and teeth yielded classification accuracies of over 80% for certain traits (Cavazzuti et al., 2019; Godinho et al., 2019; Gonçalves et al., 2013). However, most commonly measured and most sexually dimorphic skeletal traits (such as head of the humerus and femur) are rarely encountered in archaeological cremated remains. The measurement descriptions for metric traits that are often found in calcined human remains are not standardized to the same extent as the most common measurements, resulting in a larger inter-observer error. Furthermore, metric traits are population specific, temporally as well as geographically (Albanese, 2008; Gonçalves, 2014), which may be overcome under certain conditions, such as a balanced sex ratio and sufficient number of individuals (Albanese et al., 2005).

Most of the metric studies in the past were conducted using statistical predictive modeling such as discriminant function analysis (DFA) and logistic regression (Bašić et al., 2013; Gama et al., 2015; Peckmann & Fisher, 2018; Sulzmann et al., 2008). While DFA is robust and useful for the purpose of sex estimation, other supervised learning algorithms showed to outperform it in univariate analysis of metric traits (Navega et al., 2015). Compared to the DFA, logistic regression reflects better the nature of sexual dimorphism as a continuous parameter, instead of dichotomous one, since its classification results are expressed in terms of probability rather than in a binary form (Bartholdy et al., 2020). Both random forests and neural networks are types of models used in machine learning. They have applications in many different fields and have previously been used to address archaeological and anthropological questions (Alunni et al., 2015; Barone et al., 2019). The decision boundaries provided by the machine learning algorithms are not linear, rather, they base their decisions directly on patterns in the data (Navega et al., 2015). Neural networks are inspired by biological nervous systems; they are composed of nodes (artificial neurons), which are linked to one another via weighed connections. The weight (strength) and direction of these connections are adjusted as the network is trained via the training dataset (in the case of sex estimation a reference metric dataset from individuals of known sex). Once the neural network is trained, a test dataset (the problem to be solved) can be introduced (Deravignone and Macchi Janica, 2006). Random forests are composed of multiple decision trees, which, at the end of the decision process "vote" for one of the proposed prediction classes. They use resampling mechanisms to grow a forest of decision trees based on the training sample (Hastie et al., 2009). One of the advantages of logistic regression, neural networks, and random forests is that they do not require assumptions, such as (multivariate) normality, to be met (Alunni et al., 2015). Since archaeological samples are usually very small, the question of an appropriate sample size arises for the applications of these predictive modeling techniques. Studies in the field of archaeology and forensic anthropology using these methods are still rare and there is so far no established "rule of thumb" for archaeological datasets in terms of sample sizes. In previous work, they range from 76 samples (du Jardin et al., 2009) to as much as 1000 (Bewes et al., 2019). However, based on previous experimentation, datasets with $n \geq 40$ individuals are considered sufficiently robust (Albanese et al., 2005).

The aim of this study is to improve the rate of correctly sexed calcined human remains. For this purpose, 22 measurements from 13 skeletal elements are taken from individuals of known sex and age-at-death to evaluate their sexual dimorphism. Four different predictive models are used to obtain the best possible CA from the available skeletal elements. Univariate and multivariate analyses are conducted, and a methodological protocol is proposed to make the method widely applicable. This will facilitate the evaluation of the reference metric datasets and make it easier to estimate sex of archaeological and forensic calcined datasets.

2 | MATERIALS AND METHODS

The William M. Bass Donated Skeletal Collection is curated at the Forensic Anthropology Center at the University of Tennessee, Knoxville. Ninety-six cremated individuals of known sex and age-at-death were available for study, of which 86 individuals were suitable for further analysis (contained at least one measurable skeletal element). The sample consisted of 35 females aged from 32 to 101 years and 51 males, aged from 32 to 85 years. All individuals were of European ancestry and of various socioeconomic statuses. The average age-at-death for the sample was 64 years, 65.1 years for females and 63.8 years for males. Elements presenting any kind of macroscopically observable pathological lesion were excluded from the study. Skeletal elements used in the study are presented in Figure 1. All cremations were conducted by commercial crematoria within the United States and all the measured individuals were completely calcined. Specific information about the temperature and duration of each individual cremation was unavailable. In general, the temperatures in commercial crematoria range from approximately 800 and 1100°C and last for approximately 1.5 hours (McKinley, 2016). More specifically for one of the crematoriums in Tennessee where a large part of the studied individuals was cremated, temperatures range from 870 to 990°C, where the average duration of each cremation ranges from 2 to 3 hours (Bass & Jantz, 2004).

Measurements were taken with a digital sliding caliper and reported in millimeters (mm). Only the internal acoustic meatus was measured with the blunt end of metal drills with increments of 0.1 mm as described by Lynnerup et al. (2006). The drill bits were inserted in the internal acoustic meatus, and the biggest drill that could tightly fit in the opening (the blunt end was completely surrounded by bone) was recorded as its dimension. All evaluated measurements were previously published (Table 1), apart from the frontal

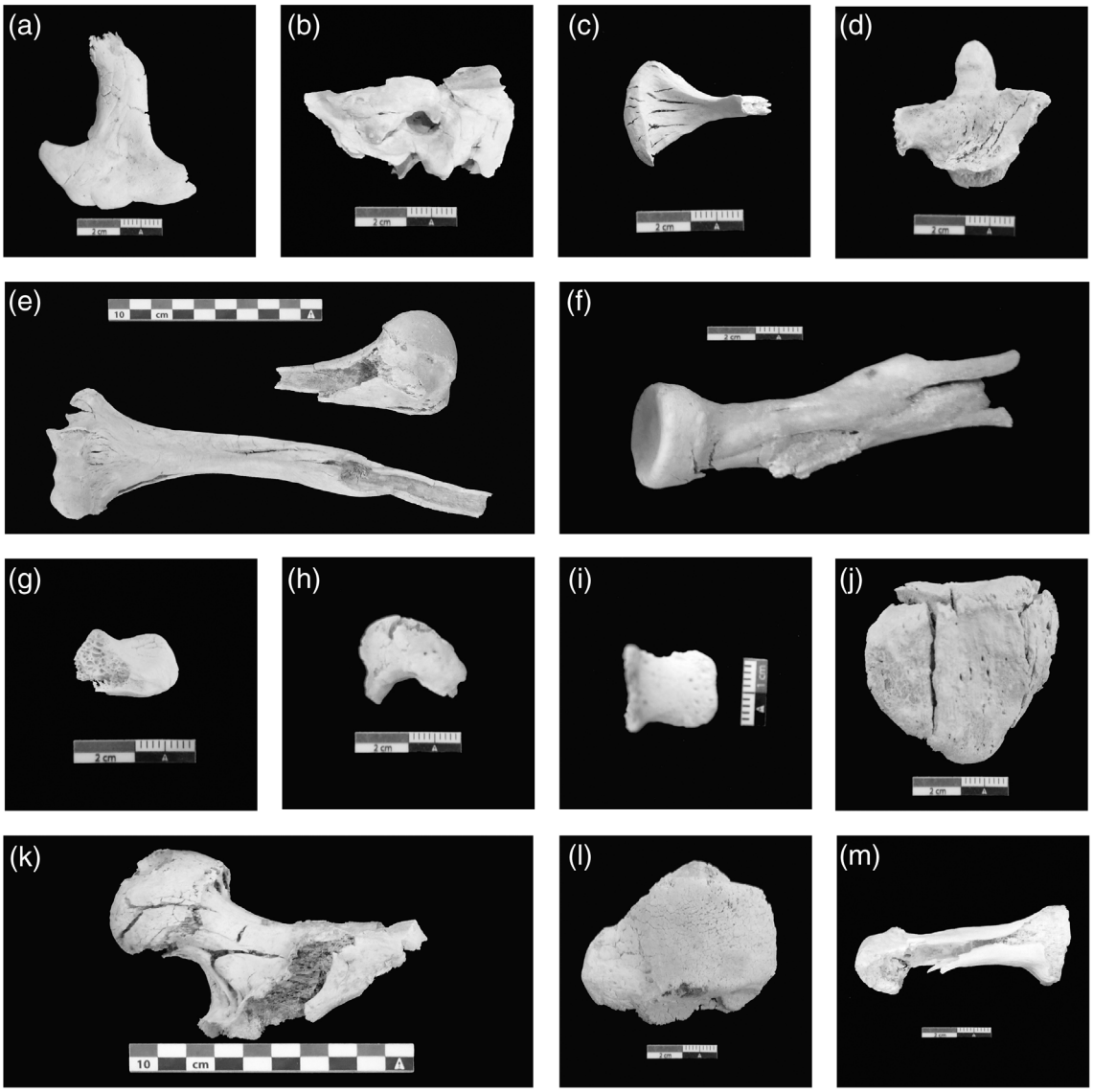

FIGURE 1 Pictures of skeletal elements that were measured in this study: (a) zygomatic bone (frontal process); (b) temporal bone (petrous part); (c) mandible (condyle); (d) axis; (e) humerus (proximal and distal); (f) radius; (g) scaphoid; (h) lunate; (i) hamate (hamulus); (j) patella; (k) femur; (l) talus; (m) first metatarsal

process of the zygomatic bone, which is a new feature that was tested in this study. This feature was added, because according to our observations it is often recovered in archaeological cremation burials. All measurements were performed by three observers with varying levels of experience. The first and second observer (M and B), had experience with osteoarchaeogical sexing methods while the third observer, C, had no experience. B remeasured 90% of individuals and C 80%. Furthermore, at least 25% of all measurements were repeated after 10 days by the first observer (M) to enable intra-observer error evaluation. Measurements were taken from the side that was available. If both left and right sides were present and because lateralization was not always possible, the values were averaged. The dataset is available in Data S1.

Results from morphological sexing methods of the skull and pelvis were compared with the metric outcomes. The following traits were assessed: occipital protuberance, nuchal crest, supraorbital margin, supraorbital ridge, mastoid process, zygomatic bone, gonial angle, mental eminence, ischial body, ventral arch, shape of the pubic bone, auricular surface, subpubic concavity, and subpubic angle (Buikstra & Ubelaker, 1994; Phenice, 1969; Workshop of European Anthropologists, 1980).

Inter- and intra-observer variability was assessed via two methods. The absolute and relative technical error of measurement, TEM and % TEM (Lyman & Van Pool, 2009) were calculated to assess the random measurement error (precision) for each measurement. Next, paired *t*-tests signed tests and Wilcoxon's signed tests were performed in IBM SPSS version 25 to reveal any systematic errors in the measurements between observers that would indicate that observers systematically measured one of the features differently. Paired *t*-test were used when the distribution of the data met the assumptions of normality, was symmetrical and without outliers. Wilcoxon signed tests were applied when the data was distributed symmetrically, but not normally. Signed tests were used if data did not meet the above assumptions, or if the sample size was

TABLE 1 List of measurements taken in the study with the descriptions and references

Trait	Reference, description
Frontal process of zygomatic	This study. Find the widest point of the upper half of the frontal process (toward the suture with the frontal bone) by moving your caliper. Care needs to be taken to place the caliper perpendicular to the lateral surface of the process.
Internal acoustic meatus	Lynnerup et al., 2006
Mandibular condyle thickness, Mandible in the two cases condyle width, Axis antero-posterior, Axis transverse, Axis height, Patella height, Patella width	Van Vark, 1975
Humerus vertical, Humerus transverse, Humerus trochlea max., Femur head vertical diameter	Martin & Saller, 1957
Humerus capitulum, Humerus trochlea min., First metatarsal dorso-plantar	Cavazzuti et al., 2019
Scaphoid max. width, Lunate length, Hamulus width	Sulzmann et al., 2008
Patella thickness	Introna et al., 1998
Radius head max. diameter, Talus max. length, Talus trochlea length, Talus trochlea width	Buikstra & Ubelaker, 1994

Abbreviations: Max., maximum; min., minimum.

below 10 observations. If %TEM was higher than 5% and if the paired *t*-tests revealed a systematic bias between the two experienced observers ($p < 0.05$), the measurement was excluded from further study, because it was deemed imprecise and/or not replicable. All features were tested for their possible correlation with age using Spearman's correlation.

After the assessment of inter-and intra-observer error, the dataset of the first observer (M) was used in the predictive models. The measurements were first tested for normal distribution and equality of variances. The size difference between females and males was assessed with *t*-tests, to evaluate the potential of each of the measured traits for sex estimation, based on the significance of the difference between females and males. To establish the univariate CA of each trait, all measurements were further assessed separately. Measurements of each trait were separated in training (70% of female and male measurements) and test sample (30% of female and male measurements). In each run, four different algorithms were used to predict sex – logistic regression, random forest, neural network and calculation of cut-off point as defined by Chakraborty and Majumder (1982). These algorithms used the training sample to predict the results for the test sample by means of cross-validation. The overall CA represents the percentage of all correctly classified individuals. Female CA represents the percentage of correctly classified females, while male CA represents the percentage of correctly classified males. Resampling, training, testing and calculations of CAs were performed 1000 times. Average CAs (total, female and male) were then calculated for each trait, along with the standard deviations were then gathered in a table.

To gain more insight into the classification outcomes on the level of different individuals, an additional calculation was performed. Over the 1000 runs, every time that the individual was selected as a part of the testing dataset (see above), the rate of correct classifications for each individual (number CA correct/number of runs where selected) was collected in a table and summarized in boxplots.

To assess the potential of supervised machine learning methods for multivariate analyses, several combinations of two features were tested via logistic regression, random forest and neural network algorithms. The cut-off point method is not suitable for multivariate analysis. The rest of the procedure was the same as outlined for univariate traits. Only a small number of combinations was tested in this study and only two traits were tested simultaneously in each run. This is because the dataset has a lot of missing values (just under 70% of all measurements are missing due to fragmentation and/or presence of pathology), which is why the number of individuals of both sexes with the same combinations of two traits are rare (even more so for more than two traits). All the above-mentioned analyses were conducted with Python programming language version 3.8.2. Logistic regression, neural network and random forest were used as implemented in the Orange Data Mining Library (https://orange-data-mining-library.readthedocs.io/en/latest/#tutorial; Demšar et al., 2013). The Python source code and protocol are available in the Data S2, and the algorithms are set in the same way they were used to conduct this study.

3 | RESULTS

Based on the results of paired tests, three measurements were deemed unsuitable for further analysis: mandibular condyle thickness, capitulum of the humerus and talus trochlea width (Table 2).

3.1 | Sex estimation—Univariate

Most of the traits exhibited a statistically significant difference in size between males and females. The descriptive and basic inferential statistics (*t*-tests results and their *p* values, and cut-off points) for each variable are presented in Table 3. Graphs with raw data histogram,

TABLE 2 Inter and intra-observer error results

Trait	Observer	*N*	Mean (mm)	TEM (mm)	%TEM (%)	Paired tests
Frontal process of the zygomatic	M_B	49	9.82	0.33	3.3	*0.073*
	M_C	43	9.96	0.21	2.2	0.119
	intraobserver	18	9.69	0.10	1.0	0.103
Internal acoustic meatus	M_B	62	3.55	0.13	3.5	*0.575*
	M_C	55	3.57	0.10	2.8	*0.063*
	intraobserver	20	3.50	0.12	3.5	*1.000*
Mandible condyle thickness	M_B	34	6.52	0.63	**9.7**	***0.007***
	M_C	43	6.48	0.69	**10.6**	*0.059*
	intraobserver	11	6.15	0.19	3.0	0.426
Mandible condyle width	M_B	17	16.98	0.30	1.7	0.502
	M_C	15	16.74	0.30	1.8	0.452
	intraobserver	5	17.75	0.06	0.4	0.617
Dens antero-posterior	M_B	23	9.87	0.25	2.6	*1.000*
	M_C	21	9.82	0.33	3.3	*1.000*
	intraobserver	7	9.85	0.24	2.5	*1.000*
Dens height	M_B	11	13.17	0.46	3.5	*1.000*
	M_C	9	13.14	0.41	3.1	*0.180*
	intraobserver	4	13.56	0.39	2.9	*0.625*
Dens transversal	M_B	21	9.09	0.16	1.7	*0.052*
	M_C	19	9.46	0.20	2.1	*0.143*
	intraobserver	7	9.52	0.12	1.3	*1.000*
Humerus vertical	M_B	22	42.24	0.39	0.9	0.087
	M_C	19	42.91	0.54	1.3	0.707
	intraobserver	6	42.58	0.50	1.2	0.192
Humerus transverse	M_B	14	38.10	1.24	3.2	*0.079*
	M_C	12	39.56	0.97	2.5	*1.000*
	intraobserver	6	38.57	0.67	1.7	0.068
Humerus trochlea max..	M_B	40	22.84	0.28	1.2	*0.203*
	M_C	37	22.74	0.37	1.6	*1.000*
	intraobserver	12	23.33	0.45	1.9	0.920
Humerus trochlea min.	M_B	52	13.84	0.46	3.4	*0.791*
	M_C	48	13.68	0.46	3.4	*0.171*
	intraobserver	13	13.87	0.47	3.4	*0.791*
Humerus capitulum	M_B	49	18.15	0.43	2.4	**0.000**
	M_C	40	18.37	0.54	2.9	**0.000**
	intraobserver	15	18.57	0.59	3.2	0.731
Radius head max.	M_B	15	19.50	0.19	1.0	0.853
	M_C	14	19.74	0.31	1.6	0.168
	intraobserver	5	19.76	0.19	1.0	0.450
Scaphoid	M_B	25	13.84	0.32	2.3	*0.596*
	M_C	22	14.01	0.48	3.4	*0.678*
	intraobserver	7	13.86	0.17	1.2	0.453
Lunate length	M_B	9	15.91	0.19	1.2	*0.125*
	M_C	7	15.99	0.09	0.5	0.209
	intraobserver	3	15.90	0.08	0.5	0.523

TABLE 2 (Continued)

Trait	Observer	*N*	Mean (mm)	TEM (mm)	%TEM (%)	Paired tests
Hamulus width	M_B	25	9.97	0.18	1.8	*0.074*
	M_C	22	10.07	0.43	4.2	*0.285*
	intraobserver	14	10.69	0.20	1.9	0.150
Femur head vertical diameter	M_B	34	42.19	0.30	0.7	0.378
	M_C	30	42.46	0.52	1.2	*0.890*
	intraobserver	10	42.91	0.25	0.6	*0.486*
Patella height	M_B	23	37.29	0.49	1.3	0.825
	M_C	20	36.97	0.63	1.7	*0.265*
	intraobserver	8	37.47	0.61	1.6	0.332
Patella width	M_B	23	39.08	0.33	0.9	0.307
	M_C	20	39.09	0.54	1.4	*1.000*
	intraobserver	8	37.89	0.63	1.7	0.068
Patella thickness	M_B	25	17.14	0.56	3.3	*0.096*
	M_C	20	17.16	0.70	4.1	0.178
	intraobserver	11	17.19	0.52	3.0	*0.227*
Talus max. length	M_B	16	49.31	1.03	2.1	0.441
	M_C	14	49.53	1.28	2.6	*0.332*
	intraobserver	7	48.88	1.34	2.7	0.099
Talus trochlea length	M_B	31	32.41	0.84	2.6	0.961
	M_C	27	32.52	1.09	3.4	0.655
	intraobserver	13	31.42	0.55	1.8	0.101
Talus trochlea width	M_B	41	27.13	1.27	4.7	***0.000***
	M_C	31	27.20	1.40	**5.1**	***0.000***
	intraobserver	16	25.80	0.75	2.9	0.367
First metatarsal dorso-plantar	M_B	25	17.25	0.45	2.6	0.859
	M_C	23	17.38	0.69	3.9	*0.152*
	intraobserver	10	16.52	0.21	1.3	0.404
First metatarsal medio-lateral	M_B	22	17.87	0.30	1.7	*1.000*
	M_C	15	18.20	0.63	3.4	***0.035***
	intraobserver	7	18.76	0.10	0.5	*0.125*

Note: Results in italics indicate the use non-parametric statistical tests. Significant results are in bold are in bold.
Abbreviations: B, Barbara; C, Christophe; max., maximal; *N*, number of individuals; M, Marta; min., minimal; %TEM, relative technical error of measurement; TEM, technical error of measurement.

normal density function curve and cut-off points for two traits with high and two traits with low CA are shown in Figure 2, and all the remaining graphs are available in Data S3.

Univariate analysis showed that three of the traits - humerus trochlea max., radius head, and lunate length - had an overall CA ≥ 90%, which was obtained by at least one of the four predictive models. Ten features had a CA of ≥ 80% by at least one of the four predictive models: radius head max., femur vertical head diameter, hamulus width, humerus head vertical diameter, humerus transverse diameter, dorso-plantar diameter of first metatarsal head, patella height, patella width, patella thickness, and talus trochlea length. The two traits that classified both sexes accurately in more than 90% of the cases were humerus trochlea max. and lunate length. The other traits that had a precision of ≥ 80% in at least one of the predictive models in both sexes were: humerus head vertical diameter, humerus head transverse diameter, radius head max., femur head vertical diameter, patella width, patella thickness, and talus trochlea length. These results are summarized in Table 4 and Figure 3.

In terms of overall CA, logistic regression, neural network, and the cut-off point method yielded similar results. Random forest almost systematically had a lower CA than the three other algorithms. Female and male CA varied for different models. The CA was generally higher for males than females. The cut-off point method performed almost equally well for both sexes for most of the traits. The other three algorithms exhibited large differences between female and male CA.

Based on individual classification (Table S4), different models did not perform equally well for the same individuals. Neural network and logistic regression were the most successful in the individual

TABLE 3 Descriptive statistics for each trait and for both sexes in anatomical order

Trait	*N* females	Mean females	SD females	*N* males	Mean males	SD males	Cut-off point	D	D SD	*t*-test	*p* value
Frontal process zygomatic	21	9.3	1.3	33	10.4	1.5	10	0.326	0.13	−2.897	**0.0055**
Internal acoustic meatus	28	3.3	0.6	41	3.7	0.7	3.6	0.282	0.12	−2.792	**0.0068**
Mandible condyle width	9	16.1	2.2	10	17.5	2.1	16.7	0.265	0.22	−1.392	0.1820
Dens height	6	12.9	2.1	5	14.7	1.4	13.4	0.418	0.26	−1.483	0.1722
Dens antero-posterior	10	9.3	1.0	15	10.6	0.9	9.8	0.506	0.18	−3.227	**0.0037**
Dens transversal	10	9.3	0.9	13	9.9	0.9	9.7	0.278	0.20	−1.598	0.1249
Humerus head vertical	8	38.8	3.0	16	44.6	2.6	41.7	0.707	0.16	−4.756	**0.0001**
Humerus head transverse	7	34.6	2.9	9	41.1	3.0	37.8	0.726	0.17	−4.058	**0.0012**
Humerus trochlea max.	17	20.0	1.3	27	24.9	1.9	22.2	0.88	0.07	−9.199	**0.0000**
Humerus trochlea min.	27	13.0	1.5	36	14.6	1.5	13.7	0.412	0.12	−4.196	**0.0001**
Radius head max.	8	18.1	0.7	9	21.0	1.1	19.4	0.89	0.11	−5.885	**0.0000**
Scaphoid	8	12.5	0.8	20	14.4	1.5	13.5	0.612	0.15	−3.320	**0.0027**
Lunate length	3	13.7	0.2	8	16.2	1.1	14.3	0.955	0.08	−3.565	**0.0061**
Hamulus width	8	8.8	0.6	20	10.6	1.1	9.6	0.741	0.13	−4.318	**0.0002**
Femur head vertical	11	38.2	2.5	27	43.8	2.8	41	0.71	0.12	−5.598	**0.0000**
Patella height	10	34.5	2.3	16	38.1	2.0	36.3	0.607	0.16	−4.109	**0.0004**
Patella width	7	36.0	2.2	18	40.2	2.1	38.1	0.679	0.17	−4.316	**0.0003**
Patella thickness	13	14.9	1.4	22	17.6	1.4	16.3	0.668	0.13	−5.429	**0.0000**
Talus max. Length	7	46.9	1.9	11	51.5	4.3	41.9	0.598	0.18	−2.517	0.0229
Talus trochlea length	13	29.2	2.1	21	34.5	2.7	31.8	0.738	0.12	−5.932	**0.0000**
First metatarsal dorso-plantar	8	16.0	1.5	20	18.2	1.5	17.1	0.548	0.18	−3.503	**0.0017**
First metatarsal medio-lateral	6	17.0	1.2	16	18.7	1.8	18.2	0.454	0.19	−2.041	0.0546

Note: Both cut-off points and the degree of sexual dimorphism (D) were calculated as defined by Chakraborty and Majumder (1982). Statistically significant *p* values are in bold.
Abbreviations: D, degree of sexual dimorphism; max., maximal; min., minimal; SD, standard deviation.

classification, closely followed by cut-off point method. Random forest performed poorly compared to the other three models (Figure 4).

From the 86 studied individuals, 10 did not have diagnostic elements available for evaluation. Of the remainder (n = 7), 14 individuals (18.6%; 14/75) had indeterminate sex, while correct sex was assigned to 54 individuals (72%; 54/75), and 7 individuals were classified incorrectly (9.3%; 7/75). This implies that only around 70% of the observable sample was correctly classified with standard anthropological methods. These results are reported in Data S4.

3.2 | Sex estimation—Multivariate

The best average CA in multivariate analysis of combinations of two traits was obtained by the neural network, followed by logistic regression and random forest. The best combination of traits resulted in a pooled CA as high as 99.1%, which was the case for the humerus trochlea max. diameter and patella thickness using the neural network model. In general, combinations of traits gave better results than each feature alone. The results for different combinations of traits for which there were enough individuals available ($n \geq 20$) are presented in Table 5.

For pooled female and male samples, only the measurement of first metatarsal medio-lateral diameter showed a statistically significant positive correlation with age (r = 0.496, p = 0.0221). In the female sample, significant positive correlations were found for the following measurements: humerus head vertical diameter, femur head vertical diameter, and the first metatarsal dorso-plantar diameter (r = 0.833, p = 0.010; r = 0.654, p = 0.029, and r = 0.809, p = 0.015 respectively).

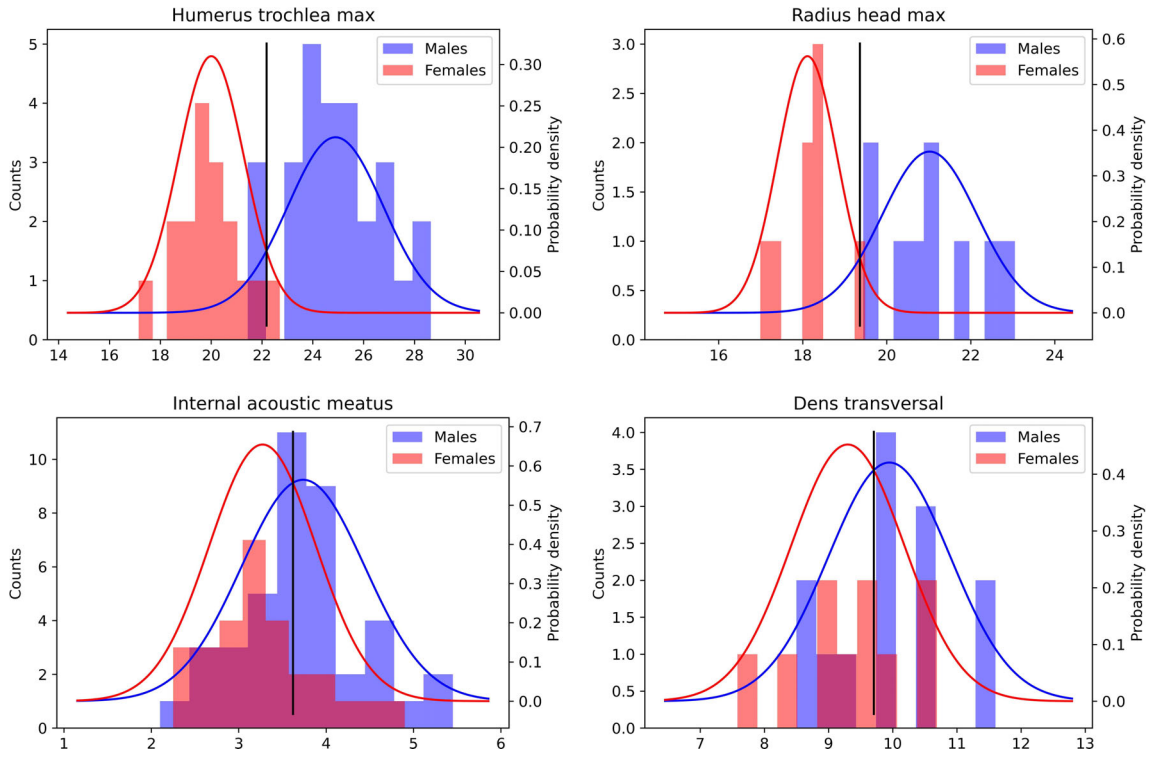

FIGURE 2 Graphs of female and male data for four traits with the calculated cut-off point (Chakraborty & Majumder, 1982) represented with a black line parallel to the *y* axis. The *x* axis represents the measurements in mm. Dark pink areas are the areas of overlap between male and female values. Top: Two traits with high classification accuracy. Bottom: Two traits with low classification accuracy. Max., maximal

4 | DISCUSSION

Sexual dimorphism appears to be preserved to a large extent in calcined remains, despite the dimensional changes that may occur as a result of the burning process. This confirms the findings of previous studies (Cavazzuti et al., 2019; Gonçalves et al., 2013). Most of the measurements (19/22; 86.4%) in this study were found to be reproducible. The inter-observer errors are similar for both experienced and inexperienced observers, suggesting that all measurements are relatively easy to learn and apply. However, osteological knowledge is required to correctly identify the fragments and landmarks of interest. Confirming the observations of Cavazzuti et al. (2019), the mandible condyle thickness, one of the three rejected measurements, was found to be particularly difficult to measure. Humerus capitulum and talus trochlea width, however, seem to yield good CAs in certain populations (Cavazzuti et al., 2019; Mahakkanukrauh et al., 2014). Further evaluation and clearer measurement descriptions are needed to avoid differences in interpretations between observers. For extensive comparison with published studies, the reader is referred to the Data S5, where CAs of the present study are compared to those of two other published calcined human remains datasets and studies on unburnt remains.

The CA for frontal process of the zygomatic was low in this study (around 60%). Another study by Hlad et al. (in preparation) tested this feature on unburnt skeletal material from, Belgium (50 individuals: 25 females, 25 males), which yielded a CA of over 80%. Clearly, more research is needed to establish the usefulness of this feature in both unburnt and calcined human remains. Findings concerning the internal acoustic meatus agree with the conclusion of Lynnerup et al. (2006) that the feature cannot be used independently for sex estimation. As for the mandibular condyle width, the results are somewhat contradictory, since it was one of the best features in the Italian protohistoric populations (Cavazzuti et al., 2019), yet one of the worst in the modern Swedish population (Van Vark, 1975) and this study, possibly due to the high average age of the two modern samples (Ishibashi et al., 1995). However, no significant correlation of this trait with age was found in the present study. Our results also concur with previous findings concerning the features of the dens axis, concluding that they are not reliable enough to be used independently for sex classification in the populations where it was tested (Cavazzuti et al., 2019; Floyd, 2017; Van Vark, 1975). While radius head max. diameter is frequently used and consistently yields one of the best CAs (Berrizbeitia, 1989; Cavazzuti et al., 2019; Mall et al., 2001; Van Vark, 1975), the other two features that had CAs over 90%, humerus trochlea max. and lunate length, are less commonly assessed. To the knowledge of the authors, only one other study (Cavazzuti et al., 2019) used the humerus trochlea max. measurement and obtained much lower CAs for Italian protohistoric populations. It would be interesting to acquire more data for this trait in other modern and archaeological populations to establish whether it is appropriate for the use in further work. Lunate length was an extremely efficient trait for sex discrimination in the Tennessee collection (over 95%). It had a CA of 80% in Italian protohistoric collections, as well as in studies on unburnt skeletal remains from Spitalfields skeletal collection, modern

TABLE 4 Mean classification accuracy for each model in percent for females, males and overall, and standard deviations of classification accuracy

Trait	Model	Classification accuracy females (%)	Classification accuracy males (%)	Classification accuracy overall (%)	Classification accuracy overall SD (%)
Frontal process zygomatic	Logistic regression	41	83	66	8
	Random forest	32	70	55	10
	Neural network	39	83	65	8
	Cut-off point	64	61	62	10
Internal acoustic meatus	Logistic regression	41	83	66	7
	Random forest	48	66	59	9
	Neural network	49	78	66	7
	Cut-off point	72	58	64	8
Mandible condyle width	Logistic regression	62	56	59	17
	Random forest	49	56	52	16
	Neural network	56	68	62	17
	Cut-off point	67	52	59	16
Dens height	Logistic regression	67	53	60	17
	Random forest	44	43	44	18
	Neural network	59	48	53	17
	Cut-off point	62	69	66	21
Dens antero-posterior	Logistic regression	66	81	75	12
	Random forest	56	84	73	13
	Neural network	61	85	76	13
	Cut-off point	70	77	74	13
Dens transversal	Logistic regression	43	71	59	14
	Random forest	46	64	56	16
	Neural network	40	65	54	14
	Cut-off point	56	68	63	15
Humerus head vertical	Logistic regression	68	88	81	11
	Random forest	54	81	71	13
	Neural network	68	87	80	11
	Cut-off point	80	83	82	11
Humerus head transverse	Logistic regression	81	90	86	13
	Random forest	76	82	79	11
	Neural network	83	89	86	12
	Cut-off point	87	88	88	12
Humerus trochlea max.	Logistic regression	91	91	91	6
	Random forest	82	93	89	6
	Neural network	93	90	91	6
	Cut-off point	93	90	91	6
Humerus trochlea min.	Logistic regression	59	80	71	8
	Random forest	68	79	74	8
	Neural network	58	81	70	8
	Cut-off point	79	74	76	8
Radius head max.	Logistic regression	88	92	90	9
	Random forest	87	86	87	9
	Neural network	89	89	89	10
	Cut-off point	88	92	90	10

TABLE 4 (Continued)

Trait	Model	Classification accuracy females (%)	Classification accuracy males (%)	Classification accuracy overall (%)	Classification accuracy overall SD (%)
Scaphoid width	Logistic regression	49	89	76	10
	Random forest	51	75	67	13
	Neural network	50	89	76	10
	Cut-off point	76	70	72	11
Lunate length	Logistic regression	100	94	95	13
	Random forest	100	94	95	15
	Neural network	100	83	87	17
	Cut-off point	99	98	98	7
Hamulus width	Logistic regression	72	91	85	10
	Random forest	47	87	74	10
	Neural network	75	90	85	10
	Cut-off point	77	85	82	11
Femur head vertical	Logistic regression	72	92	86	8
	Random forest	57	85	76	8
	Neural network	76	91	86	8
	Cut-off point	82	83	83	9
Patella height	Logistic regression	77	88	84	12
	Random forest	80	84	83	13
	Neural network	76	88	83	12
	Cut-off point	90	78	83	12
Patella width	Logistic regression	55	90	78	9
	Random forest	58	82	74	12
	Neural network	66	87	80	10
	Cut-off point	82	84	83	10
Patella thickness	Logistic regression	69	84	78	10
	Random forest	81	85	83	10
	Neural network	69	82	77	11
	Cut-off point	73	79	77	11
Talus max. length	Logistic regression	61	82	73	13
	Random forest	56	63	60	13
	Neural network	63	78	71	12
	Cut-off point	37	80	62	10
Talus trochlea length	Logistic regression	80	87	84	9
	Random forest	66	88	80	9
	Neural network	84	83	84	9
	Cut-off point	91	81	85	9
First metatarsal dorso-plantar	Logistic regression	50	98	82	9
	Random forest	62	90	81	11
	Neural network	51	98	82	9
	Cut-off point	68	70	70	13
First metatarsal medio-lateral	Logistic regression	23	87	69	11
	Random forest	24	69	57	15
	Neural network	21	85	67	9
	Cut-off point	63	70	68	14

Abbreviations: max., maximal; min., minimal; SD, standard deviation.

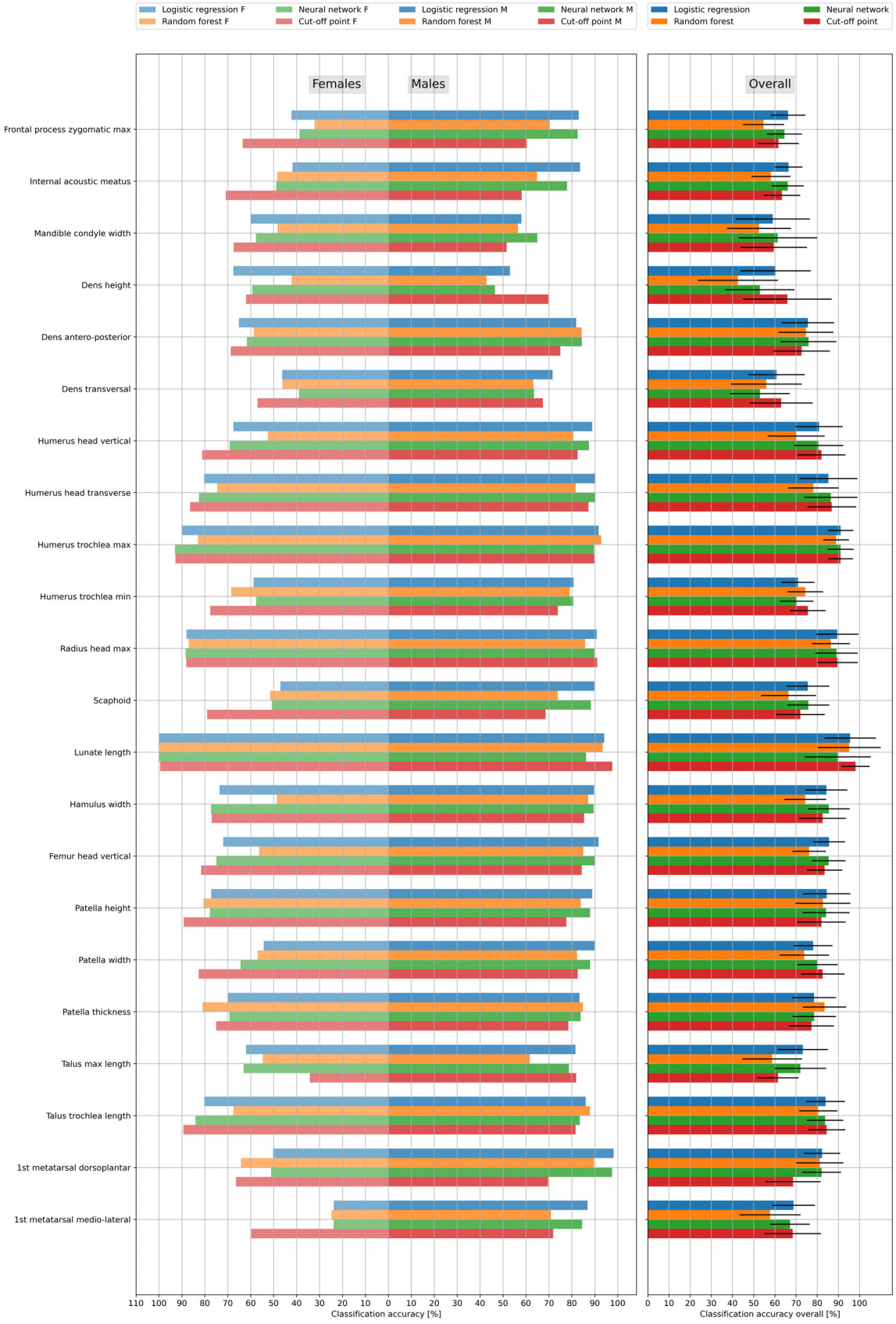

FIGURE 3 Chart showing the female, male and overall classification accuracies for each studied trait and for each of the predictive models. Black horizontal lines in the right graph are error bars. Max., maximal; min., minimal

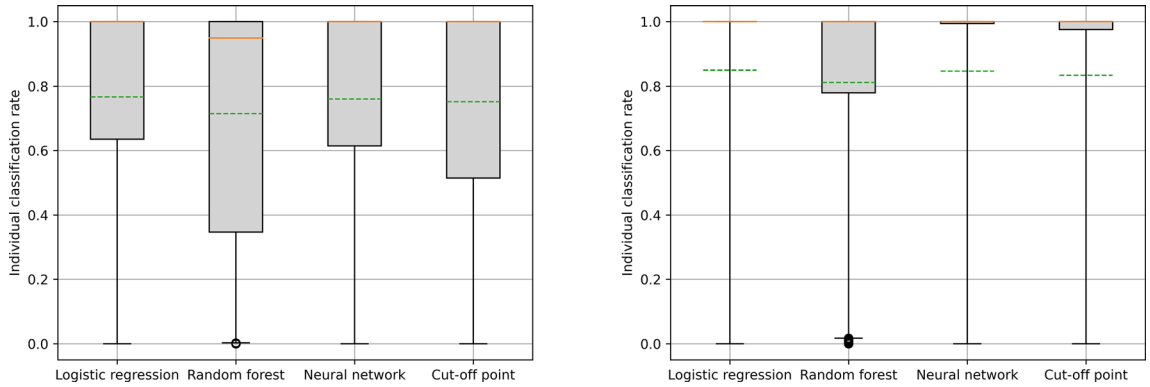

FIGURE 4 Left: Boxplot showing the individual classification rate in percent of different models used across all features. Right: Boxplot showing the individual classification rate in percent of different models used across 12 features that had classification accuracy over 80%. Orange lines represent median values, green dashed line represent mean values

Mexican, and modern Spanish populations (Mastrangelo, De Luca, Alemán, & Botella, 2011; Mastrangelo, De Luca, & Sánchez-Mejorada, 2011; Sulzmann et al., 2008). However, in our study, the lunate subsample might not be representative of the whole population because it is very small. The rest of the features had CAs between 70% and 80%. This suggest they may not applicable as independent features for sex estimation, but they could be of use when incorporated in multivariate models. Nevertheless, more research is needed on larger and more complete datasets and with clearer and more standardized measurement instructions.

While logistic regression and neural network methods yielded similar overall CAs out of all the used algorithms, the cut-off point method yielded the most balanced CAs for each separate sex. Generally, the CA was found to be higher for males than for females in the studied sample, which is related to the fact that the two samples are not well balanced for most of the traits (males being better represented), with the sex ratio males: females ranging from 0.75:1 to 3:1. The neural network yielded better CAs than the random forest. This could be due to settings that were used to run the algorithms, such as number of trees for the random forest and the number of neurons and iterations for the neural network. Another explanation may be related to sample size. Based on the above, cut-off point method might be the best choice when dealing with the univariate analyses of small samples that have an unbalanced sex ratio. The summary of the individual classification (Figure 4) underlines that random forest is especially sensitive to the quality and size of the training sample that is used. As suggested by Alunni et al. (2015), more research on the applications of supervised machine learning using different settings is needed. While a thorough evaluation of the sample size effect for each of the features is necessary, the sample size guidelines from previous work ($n \geq 40$) are recommended (Albanese et al., 2005). Consideration of the individual classification rates (Data S4) for different models is promising for a better understanding of their advantages and limitations and could be used in the future to narrow down their selection to one or two most appropriate models for the univariate analysis. The real potential of supervised machine learning techniques such as neural networks for the purpose of sex estimation lies in the multivariate analysis, since certain combinations of traits (e.g., humerus trochlea max. and patella thickness) correctly classified more than 98% of the sample. Neural network yielded best CAs in the multivariate analysis than the two other algorithms tested, which agrees with previous studies (du Jardin et al., 2009). The major advantage of the methodology presented in this study is that different metric datasets can be studied quickly and the most pertinent trait combinations for each population can easily be tested by following the steps described in the methods section using the Python script provided in the Data S2. It is, however, necessary to confirm these findings with larger and more complete datasets.

Another advantage of supervised learning is that it does not require (multivariate) normality and other common assumptions for data distribution to be valid. Although most of the population metric data tends toward a normal distribution (Thompson, 2002), archaeological, and in particular cremated samples are usually relatively small and incomplete, which makes many statistical tests, such as DFA less suitable for this type of samples, especially for multivariate analyses. Comparisons between metric and morphological sexing methods are difficult, because of the indeterminate category used in morphological sexing. This category was not used in this study, because it would imply to arbitrarily choose indeterminate cut-off points. It is clear, however, that even if the indeterminate individuals are ignored, the CA of the morphological methods amount to around 87%. While this is a good result regarding the fragmentation and preservation of most of calcined human remains, pelvic fragments are reported to be rare in many cases (Depierre, 2014; Gonçalves & Pires, 2017). A multivariate supervised machine learning approach showed that this result can be improved with the multivariate metric methods. Another issue linked to the fragmentation of calcined human remains is that both sides are rarely present from the same individual and often not preserved well enough to measure. Since there were few individuals with both sides present/measurable, the comparison between left and right elements was not possible. Averaging of the measurements from both sides or using either of them may lower the CA. On the other hand, fragments from both sides can be used.

Body size in different populations is subject to many different factors, such as environment, genetics, and secular change (Albanese,

TABLE 5 Classification accuracy in percent for different combinations of two traits analyzed with logistic regression, random forest, and neural network algorithms

Traits	Model	Classification accuracy females	Classification accuracy males	Classification accuracy pooled	Classification accuracy SD	Training set *N*	Test set *N*
Dens transverse, Dens antero-posterior	Logistic regression	65.7	75.8	71.4	16.8	15	7
	Random forest	52.7	78	67.1	14.6	15	7
	Neural network	59	80	71	14.4	15	7
Humerus trochlea max., Patella thickness	Logistic regression	98.7	95.5	96.9	8.2	13	7
	Random forest	99.7	97.5	98.4	5.7	13	7
	Neural network	100	98.5	99.1	3.9	13	7
Humerus trochlea max., Talus trochlea length	Logistic regression	98.5	91.2	93.3	10.4	13	7
	Random forest	100	97.4	98.1	5.2	13	7
	Neural network	100	97.6	98.3	5.1	13	7
Humerus trochlea max., Femur head vertical	Logistic regression	83.3	98.2	93.2	9	15	9
	Random forest	88	93.3	91.6	7.2	15	9
	Neural network	87.7	88	87.9	9.4	15	9
Humerus trochlea min, Internal acoustic meatus	Logistic regression	71.6	80.6	76.6	8.6	37	16
	Random forest	67.3	81.3	75.2	9	37	16
	Neural network	69.3	87.7	79.6	9.6	37	16
Talus trochlea length, Internal acoustic meatus	Logistic regression	81.5	90.7	87	8.5	22	10
	Random forest	82.5	88.8	86.3	9.7	22	10
	Neural network	81.3	91.8	87.6	9.8	22	10
Patella thickness, Internal acoustic meatus	Logistic regression	66.8	89.5	80.4	11.5	20	10
	Random forest	70.8	92	83.5	10.8	20	10
	Neural network	73.3	94.2	85.8	8.7	20	10
Patella height, Patella thickness	Logistic regression	65.3	84.6	77.4	13.2	14	8
	Random forest	70	87.8	81.1	11.2	14	8
	Neural network	70.3	88.8	81.9	10.7	14	8

Abbreviations: max., maximal; min., minimal; *N*, number of individuals; SD, standard deviation.

2008), which was confirmed when comparing different cremated datasets with the one collected in this study (Cavazzuti et al., 2019; Van Vark, 1975). Thus, the metric sexing methods developed on one population may not be suitable to sex other populations, which is why it is important to have an easily applicable protocol to establish population-specific cut-off points, if the universal applicability cannot be achieved. Once a metric dataset is obtained it can be tested using the algorithms proposed. This provides a solid basis for sex estimation in collections of unknown sex.

While the number of individuals in the studied dataset, especially females, is relatively limited, it is one of the largest datasets that is currently available for cremations and its findings could be used to help with identification of individuals of European ancestry from Southeast United-States in forensic cases. There are different ways in which the applicability of the method can be explored in future research in forensic as well as archaeological contexts. Any metric dataset of known sex can be tested with the protocol developed for this study and can thus produce reference datasets for metric sex estimation for similar populations. A dataset with any number of features can be inserted in the Python script and run any number of times to obtain population specific results, by following the workflow in Data S2. Individuals (burnt or unburned) that can be confidently sexed with morphological methods may be used as a reference for more fragmented individuals and partially preserved individuals from the same collection. Additional steps are required for the collections or individuals for which reference populations are not available. Applying chemosteometric indices, as suggested by Gonçalves et al. (2020), may solve the problem with the level of shrinkage in cremations, since the original (unburned) size of bones can be obtained and therefore unburned reference collections could be used. This, however, involves an additional step of FTIR analysis. The applicability of the method presented in this study would be further enhanced by developing a user-friendly interface where the potential of different combinations of features and algorithm settings could be tested by scholars for reference skeletal datasets (burnt or unburnt) relevant to the populations they want to assess.

5 | CONCLUSIONS

More than half of studied metric traits are highly sexually dimorphic, which makes them suitable for improving sex estimation rates of highly

fragmented and/or calcined individuals. Combinations of different traits and predictive models showed the potential of raising the CA over 95% for both sexes when using a multivariate machine learning approach. This points to the utility of this protocol for the evaluation of the potential of different combinations of traits for population specific sex estimation of cremated human remains as proposed in this study. More research and more data are now needed to further support these findings.

In future work, the emphases should lie on the measurement standardization of certain traits with strong sexually dimorphic characteristics that are complicated to measure to avoid measurement errors and increase the choice of useful traits. Additionally, collecting large quantities of metric data on different populations from different periods and regions will give a better idea about the potential of different traits in different populations. Lastly, further increasing the potential of multivariate statistics and different machine learning algorithms for predicting sex is necessary to improve the possibilities of sex estimation for highly fragmented and calcined human remains from forensic as well as archaeological contexts.

ACKNOWLEDGMENTS

This research is part of the FWO – FNRS F.R.S. EoS Project CRUMBEL (30999782). MH's research is supported by the FWO doctoral grant (198613) and the VUB Doctoral School travel grant. CS would like to thank the FWO for his post-doctoral fellowship (12W2118N) and his short research stay (K232919N) to carry out the research in Tennessee. The authors want to thank Melanie Beasley who facilitated us access to the collection and for her warm welcome in Knoxville and Caroline Znachko for her assistance during the research stay at the Forensic Anthropology Center of the University of Tennessee, Knoxville, USA.

AUTHOR CONTRIBUTIONS

Marta Hlad: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; resources; software; validation; visualization; writing-original draft; writing-review & editing. **Barbara Veselka:** Conceptualization; formal analysis; methodology; supervision; validation; writing-original draft; writing-review & editing. **Dawnie Wolfe Steadman:** Project administration; resources; writing-review & editing. **Baptiste Herregods:** Formal analysis; methodology; software; validation; visualization; writing-review & editing. **Marc Elskens:** Formal analysis; methodology; software. **Rica Annaert:** Writing-review & editing. **Mathieu Boudin**: Writing-review & editing; funding acquisition. **Giacomo Capuzzo:** Writing-review & editing. **Sarah Dalle:** Writing-review & editing. **Guy De Mulder:** Funding acquisition; project administration; writing-review & editing. **Charlotte Sabaux:** Writing-review & editing. **Kevin Salesse:** Writing-review & editing. **Amanda Sengeløv:** Writing-review & editing. **Elisavet Stamataki:** Writing-review & editing. **Martine Vercauteren:** Funding acquisition; project administration; supervision; writing-review & editing. **Eugène Warmenbol:** Funding acquisition; project administration; writing-review & editing. **Dries Tys:** Funding acquisition; project administration; supervision; writing-review & editing. **Christophe Snoeck:** Conceptualization; formal analysis; funding acquisition; investigation; project administration; supervision; writing-original draft; writing-review & editing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

ORCID

Marta Hlad https://orcid.org/0000-0002-9263-4048
Barbara Veselka https://orcid.org/0000-0002-2692-9577
Dawnie Wolfe Steadman https://orcid.org/0000-0003-0812-0739
Marc Elskens https://orcid.org/0000-0002-8862-9422
Rica Annaert https://orcid.org/0000-0001-8976-6325
Mathieu Boudin https://orcid.org/0000-0002-3991-1026
Sarah Dalle https://orcid.org/0000-0003-3338-0700
Guy De Mulder https://orcid.org/0000-0002-4180-502X
Charlotte Sabaux https://orcid.org/0000-0002-2805-2529
Kevin Salesse https://orcid.org/0000-0003-2492-1536
Amanda Sengeløv https://orcid.org/0000-0003-3884-9600
Elisavet Stamataki https://orcid.org/0000-0002-7010-2585
Dries Tys https://orcid.org/0000-0001-9202-2685
Christophe Snoeck https://orcid.org/0000-0003-3770-4055

REFERENCES

Afacan, G. O., Onal, T., Akansel, G., & Arslan, A. S. (2017). Is the lateral angle of the internal acoustic canal sexually dimorphic in non-adults? An investigation by routine cranial magnetic resonance imaging. *HOMO - Journal of Comparative Human Biology*, *68*(5), 393–397. https://doi.org/10.1016/j.jchb.2017.09.001

Albanese, J. (2008). A critical review of the methodology for the study of secular change using skeletal data. *Ontario Archaeology*, *85-88*, 139–156.

Albanese, J., Cardoso, H. F. V., & Saunders, S. R. (2005). Universal methodology for developing univariate sample-specific sex determination methods: An example using the epicondylar breadth of the humerus. *Journal of Archaeological Science*, *32*(1), 143–152. https://doi.org/10.1016/j.jas.2004.08.003

Alunni, V., du Jardin, P., Nogueira, L., Buchet, L., & Quatrehomme, G. (2015). Comparing discriminant analysis and neural network for the determination of sex using femur head measurements. *Forensic Science International*, *253*, 81–87. https://doi.org/10.1016/j.forsciint.2015.05.023

Barone, G., Mazzoleni, P., Spagnolo, G. V., & Raneri, S. (2019). Artificial neural network for the provenance study of archaeological ceramics using clay sediment database. *Journal of Cultural Heritage*, *38*, 147–157. https://doi.org/10.1016/j.culher.2019.02.004

Bartholdy, B. P., Sandoval, E., Hoogland, M. L. P., & Schrader, S. A. (2020). Getting rid of dichotomous sex estimations: Why logistic regression should be preferred over discriminant function analysis. *Journal of Forensic Sciences*, *65*(5), 1685–1691. https://doi.org/10.1111/1556-4029.14482

Bašić, Ž., Anterić, I., Vilović, K., Petaros, A., Bosnar, A., Madžar, T., Polašek, O., & Andelinović, Š. (2013). Sex determination in skeletal remains from the medieval eastern Adriatic coast - discriminant function analysis of humeri. *Croatian Medical Journal*, *54*(3), 272–278. https://doi.org/10.3325/cmj.2013.54.272

Bass, W. M., & Jantz, R. L. (2004). Cremation weights in East Tennessee. *Journal of Forensic Sciences*, *49*(5), 1–4. https://doi.org/10.1520/jfs2004002

Berrizbeitia, E. L. (1989). Sex determination with the head of the radius. *Journal of Forensic Sciences*, *34*(5), 12754J. https://doi.org/10.1520/jfs12754j

Bewes, J., Low, A., Morphett, A., Pate, F. D., & Henneberg, M. (2019). Artificial intelligence for sex determination of skeletal remains: Application of a deep learning artificial neural network to human skulls. *Journal of Forensic and Legal Medicine*, *62*(January), 40–43. https://doi.org/10.1016/j.jflm.2019.01.004

Brück, J. (2009). Women, death, and social change in the British bronze age. *Norwegian Archaeological Review*, *42*(1), 1–23. https://doi.org/10.1080/00293650902907151

Buikstra, J., & Ubelaker, D. H. (1994). Standards for data collection from human skeletal remains. Paper presented at: Proceedings of a seminar at the field museum of natural history. Fayetteville: Arkansas Archaeological Survey.

Buikstra, J. E., & Swegle, M. (1989). Bone modification due to burning: Experimental evidence. In R. Bonnichsen & M. H. Sorgh (Eds.), *Bone modification* (pp. 247–258). USA: Center for the Study of the First Americans.

Cavazzuti, C., Bresadola, B., D'Innocenzo, C., Interlando, S., & Sperduti, A. (2019). Towards a new osteometric method for sexing ancient cremated human remains. Analysis of late bronze age and iron age samples from Italy with gendered grave goods. *PLoS One*, *14*(1), e0209423. https://doi.org/10.1371/journal.pone.0209423

Chakraborty, R., & Majumder, P. P. (1982). On Bennett's measure of sex dimorphism. *American Journal of Physical Anthropology*, *59*, 295–298. https://doi.org/10.1002/ajpa.1330590309

Demšar, J., Curk, T., Erjavec, A., Gorup, C., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, *14* (August), 2349–2353.

Depierre, G. (2014). *Crémation et archéologie: nouvelles alternatives méthodologiques en ostéologie humaine*. Université de Dijon.

Deravignone, L., & Macchi Jánica, G. (2006). Artificial neural networks in archaeology. *Archeologia e Calcolatori*, *17*, 121–136.

Dokladál, M. (1971). A further contribution to the morphology of burned human bones. In N. Novotny (Ed.), Proceedings of the Anthropological Congress Dedicated to Aleš Hrdlička, (pp. 561–568). Prague: Czechoslovak Academy of Science.

du Jardin, P., Ponsaillé, J., Alunni-Perret, V., & Quatrehomme, G. (2009). A comparison between neural network and other metric methods to determine sex from the upper femur in a modern French population. *Forensic Science International*, *192*, 127.e1–127.e6. https://doi.org/10.1016/j.forsciint.2009.07.014

Floyd, E. N. (2017). Sex estimation utilizing dimensions from the occipital bone, atlas and axis (Unpublished Master's thesis). Middle Tennessee State University, Murfreesboro.

Gama, I., Navega, D., & Cunha, E. (2015). Sex estimation using the second cervical vertebra: A morphometric analysis in a documented Portuguese skeletal sample. *International Journal of Legal Medicine*, *129*, 365–372. https://doi.org/10.1007/s00414-014-1083-0

Gejvall, N. G. (1963). Cremations. In D. Brothwell & E. Higgs (Eds.), *Science in archaeology* (pp. 468–479). London: Thames and Hudson.

Godinho, R. M., Oliveira-Santos, I., Pereira, M. F., Mauricio, A., Valera, A., & Gonçalves, D. (2019). Is enamel the only reliable hard tissue for sex metric estimation of burned skeletal remains in biological anthropology? *Journal of Archaeological Science: Reports*, *26*(June), 101876. https://doi.org/10.1016/j.jasrep.2019.101876

Gonçalves, D. (2014). Evaluation of the effect of secular changes in the reliability of osteometric methods for the sex estimation of Portuguese individuals. *Cadernos Do GEEvH*, *3*(1), 53–65.

Gonçalves, D., Campanacho, V., & Cardoso, H. F. V. (2011). Reliability of the lateral angle of the internal auditory canal for sex determination of subadult skeletal remains. *Journal of Forensic and Legal Medicine*, *18*(3), 121–124. https://doi.org/10.1016/j.jflm.2011.01.008

Gonçalves, D., & Pires, A. E. (2017). Cremation under fire: A review of bio-archaeological approaches from 1995 to 2015. *Archaeological and Anthropological Sciences*, *9*, 1677–1688. https://doi.org/10.1007/s12520-016-0333-0

Gonçalves, D., Thompson, T. J. U., & Cunha, E. (2013). Osteometric sex determination of burned human skeletal remains. *Journal of Forensic and Legal Medicine*, *20*(7), 906–911. https://doi.org/10.1016/j.jflm.2013.07.003

Gonçalves, D., Vassalo, A. R., Makhoul, C., Piga, G., Mamede, A. P., Parker, S. F., Ferreira, M. T., Cunha, E., Marques, M. P. M., & Batista de Carvahlo, L. A. E. (2020). Chemosteometric regression models of heat exposed human bones to determine their pre-burnt metric dimensions. *American Journal of Physical Anthropology*, *173*(4), 734–747. https://doi.org/10.1002/ajpa.24104

Gouveia, M. F., Oliveira Santos, I., Santos, A. L., & Gonçalves, D. (2017). Sample-specific odontometric sex estimation: A method with potential application to burned remains. *Science and Justice*, *57*(4), 262–269. https://doi.org/10.1016/j.scijus.2017.03.001

Graw, M., Wahl, J., & Ahlbrecht, M. (2005). Course of the meatus acusticus internus as criterion for sex differentiation. *Forensic Science International*, *147*(2-3 SPEC.ISS), 113–117. https://doi.org/10.1016/j.forsciint.2004.08.006

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning. Data mining, inference, and prediction*. Springer Series in Statistics, (2nd ed.). New York: Springer.

Holck, P. (1986). Cremated bones. Medical-anthropological study of an archaeological material on cremation burials. In *Antropologiske skrifter nr. 1c*. Anatomisk Institutt Universitetet.

Introna, F., Di Vella, G., & Campobasso, C. P. (1998). Sex determination by discriminant analysis of patella measurements. *Forensic Science International*, *95*, 39–45. https://doi.org/10.1016/S0379-0738(98)00080-2

Ishibashi, H., Takenoshita, Y., Ishibashi, K., & Oka, M. (1995). Age-related changes in the human mandibular condyle: A morphologic, radiologic, and histologic study. *Journal of Oral and Maxillofacial Surgery*, *53*(9), 1016–1023. https://doi.org/10.1016/0278-2391(95)90117-5

Lyman, L. R., & Van Pool, T. L. (2009). Metric data in archaeology: A study of intra-analyst and inter-analyst variation. *Society for American Archaeology*, *74*(3), 485–504. https://doi.org/10.1017/S0002731600048721

Lynnerup, N., Schulz, M., Madelung, A., & Graw, M. (2006). Diameter of the human internal acoustic meatus and sex determination. *International Journal of Osteoarchaeology*, *16*, 118–123. https://doi.org/10.1002/oa.811

Mahakkanukrauh, P., Praneatpolgrang, S., Ruengdit, S., Singsuwan, P., Duangto, P., & Case, D. T. (2014). Sex estimation from the talus in a Thai population. *Forensic Science International*, *240*, 152.e1–152.e8. https://doi.org/10.1016/j.forsciint.2014.04.001

Mall, G., Hubig, M., Büttner, A., Kuznik, J., Penning, R., & Graw, M. (2001). Sex determination and estimation of stature from the long bones of the arm. *Forensic Science International*, *117*(1-2), 23–30. https://doi.org/10.1016/S0379-0738(00)00445-X

Martin, R., & Saller, K. (1957). *Lehrbuch der anthropologie, band I*. Stuttgart: Fischer.

Masotti, S., Pasini, A., & Gualdi-Russo, E. (2019). Sex determination in cremated human remains using the lateral angle of the pars petrosa ossis temporalis: Is old age a limiting factor ? *Forensic Science, Medicine and Pathology*, *15*, 392–398. https://doi.org/10.1007/s12024-019-00131-4

Masotti, S., Succi-Leonelli, E., & Gualdi-Russo, E. (2013). Cremated human remains: Is measurement of the lateral angle of the meatus acusticus internus a reliable method of sex determination? *International Journal of Legal Medicine*, *127*(5), 1039–1044. https://doi.org/10.1007/s00414-013-0822-y

Mastrangelo, P., De Luca, S., Alemán, I., & Botella, M. C. (2011). Sex assessment from the carpal bones: Discriminant function analysis in a 20th century Spanish sample. *Forensic Science International*, *206*(1-3), 216.e1–216.e10. https://doi.org/10.1016/j.forsciint.2011.01.007

Mastrangelo, P., De Luca, S., & Sánchez-Mejorada, G. (2011). Sex assessment from carpal bones: Discriminant function analysis in a contemporary Mexican sample. *Forensic Science International*, *209*(1-3), 196.e1–196.e15. https://doi.org/10.1016/j.forsciint.2011.04.019

Mayne Correia, P., & Beattie, O. (2002). A critical look at methods for recovering, evaluating, and interpreting cremated human remains. In W. D. Haglund & M. H. Sorg (Eds.), *Advances in forensic Taphonomy. Method, theory and archaeological perspectives* (pp. 436–449). CRC Press.

McKinley, J. I. (2016). Complexities of the ancient mortuary rite of cremation: An Osteoarchaeological conundrum. In G. Grupe & G. C. McGlynn (Eds.), Isotopic landscapes in bioarchaeology proceedings of the international workshop "a critical look at the concept of isotopic landscapes and its application in future bioarchaeological research", Munich, October 13–15, 2014, (pp. 17–41). Heidelberg: Springer.

Navega, D., Vicente, R., Vieira, D. N., Ross, A. H., & Cunha, E. (2015). Sex estimation from the tarsal bones in a Portuguese sample: A machine learning approach. *International Journal of Legal Medicine*, *129*(3), 651–659. https://doi.org/10.1007/s00414-014-1070-5

Oestigaard, T. (2013). Cremations in culture and cosmology. In L. Nilsson Stutz & S. Tarlow (Eds.), *The Oxford handbook of the archaeology of death and burial* (pp. 497–510). Oxford: Oxford University Press.

Oestigaard, T. (2016). Kremasjon. Etnografiske paralleller og arkeologiske perspektiver. In K. Cassel (Ed.), *Socioekonomisk mångfald. Ritualer och urbanitet* (pp. 65–77). Stockholm: Statens Historiska Museer.

Peckmann, T. R., & Fisher, B. (2018). Sex estimation from the patella in an African American population. *Journal of Forensic and Legal Medicine*, *54*, 1–7. https://doi.org/10.1016/j.jflm.2017.12.002

Phenice, T. W. (1969). A newly developed visual method of sexing the os pubis. *American Journal of Physical Anthropology*, *30*(2), 297–301. https://doi.org/10.1002/ajpa.1330300214

Piontek, J. (1975). Polish methods and results of investigations of cremated bones from prehistoric cemeteries. *Glasnik Antropološkog Društva Jugoslavije*, *12*, 23–34.

Piontek, J. (1976). Proces kremacji i jego wpływ na morfologię kości w świetle wyników badań eksperymentalnych. *Archeologia Polski*, *21*(1), 247–280.

Rösing, F. (1977). Methoden und Aussagemöglichkeiten der anthropologischen Leichenbrandbearbeitung. *Archäologie Und Naturwissenschaft*, *1*, 53–80.

Schutkowski, H., & Herrmann, B. (1983). Zur Möglichkeit der metrischen Geschlechtsdiagnose an der Pars petrosa ossis temporalis. *Zeitschrift für Rechtsmedizin*, *90*(3), 219–227. https://doi.org/10.1007/BF02116233

Shipman, P., Foster, G., & Schoeninger, M. (1984). Burnt bones and teeth: An experimental study of color, morphology, crystal structure and shrinkage. *Journal of Archaeological Science*, *11*(4), 307–325. https://doi.org/10.1016/0305-4403(84)90013-X

Snoeck, C., Lee-Thorp, J., & Schulting, R. (2014). From bone to ash: Compositional and structural changes in burned modern and archaeological bone. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *416*, 55–68. https://doi.org/10.1016/j.palaeo.2014.08.002

Sulzmann, C. E., Buckberry, J. L., & Pastor, R. F. (2008). The utility of carpals for sex assessment: A preliminary study. *American Journal of Physical Anthropology*, *135*, 252–262. https://doi.org/10.1002/ajpa.20738

Thompson, T. J. U. (2002). The assessment of sex in cremated individuals: Some cautionary notes. *Journal of the Canadian Society of Forensic Science*, *35*(2), 49–56. https://doi.org/10.1080/00085030.2002.10757535

Thompson, T. J. U. (2005). Heat-induced dimensional changes in bone and their consequences for forensic anthropology. *Journal of Forensic Sciences*, *50*(5), 1–8. https://doi.org/10.1520/JFS2004297

Van Vark, G. N. (1975). The investigation of human cremated skeletal material by multivariate statistical methods, II. Measures. *Ossa. International Journal of Skeletal Research*, *2*, 47–68.

Veselka, B., Hlad, M., Steadman, D., Annaert, R., Boudin, M., Capuzzo, G., Dalle, S., Kontopulos, I., de Mulder, G., Sabaux, C., Salesse, K., Sengeløv, A., Stamataki, E., Vercauteren, M., Tys, D., & Snoeck, C. (2020). Estimation age-at-death in burnt adult human remains using the Falys-Prangel method. *American Journal of Physical Anthropology*, Early view. https://doi.org/10.1002/ajpa.24210

Veselka, B., & Lemmers, S. A. M. L. (2014). Deliberate selective deposition of iron age cremations from Oosterhout (prov. Noord-Brabant, The Netherlands): A 'pars pro toto' burial ritual. *Lunula, Archaeologia Protohistorica*, *22*, 151–158.

Wahl, J. (1996). Erfahrungen zur metrischen geschlechtsdiagnose bei leichenbränden. *HOMO- Journal of Comparative Human Biology*, *47*(1–3), 339–359.

Welinder, S. (1989). An experiment with the analysis of sex and gender of cremated bones. *Tor*, *22*, 29–41.

Williams, H. (2008). Towards an archaeology of cremation. In C. W. Schmidt & S. A. Symes (Eds.), *The analysis of burned human remains* (pp. 239–269). London: Academic Press. https://doi.org/10.1016/B978-012372510-3.50017-4

Workshop of European Anthropologists. (1980). Recommendations for age and sex diagnoses of skeletons. *Journal of Human Evolution*, *9*(7), 517–549. https://doi.org/10.1016/0047-2484(80)90061-5

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Hlad M, Veselka B, Steadman DW, et al. Revisiting metric sex estimation of burnt human remains via supervised learning using a reference collection of modern identified cremated individuals (Knoxville, USA). *Am J Phys Anthropol*. 2021;1–17. https://doi.org/10.1002/ajpa.24270